WORKING paper

BANQUE DE FRANCE
EUROSYSTÈME

In the Land of AKM: Explaining the Dynamics of Wage Inequality in France

Damien Babet[1], Olivier Godechot[2] & Marco G. Palladino[3]

March 2025, WP #987

Abstract

We use a newly constructed and quasi-exhaustive matched employer-employee database to study the contribution of firms to wage inequality in France. We implement a simple and tractable correction for the limited mobility bias. Our analysis, covering the period 2002-2019, reveals an increase in between-firm inequality, mainly due to the growing clustering of workers with similar market value. These phenomena are associated with increasing occupational specialization at the firm level. Our results highlight the importance of bias-corrected AKM estimates of the Abowd, Kramarz et Margolis (1999) model –hereafter AKM- in capturing the dynamics of wage inequality, and show how both observable job types and unobservable individual characteristics contribute to these patterns.

Keywords: Wage Inequality, Worker Segregation, Occupational Sorting, Employer-Employee Data

JEL classification: C23, J24, J31, J62

[1] Insee
[2] Sciences Po, CRIS-CNRS and AxPo
[3] Banque de France, MarcoGuido.PALLADINO@banque-france.fr

Non-Technical Summary

Using a comprehensive database of matched employer-employee information, this paper examines how wage and workplace inequality evolved in France between 2002 and 2019. While wage inequality increased in many developed countries during this period, France presents an interesting case where overall wage inequality remained relatively stable (Figure 1). Underlying this apparent stability, however, were important changes in the distribution of workers across firms.

First, there is increasing 'segregation' - workers with similar earning potential are increasingly likely to work together in the same firms. This means that high earners are increasingly concentrated in certain firms, while low earners are concentrated in others. In fact, this trend has been going on in France since the early 1980s. Second, there is a modest increase in 'sorting' - a growing tendency for high potential earners to be employed in firms that pay higher wages across the board.

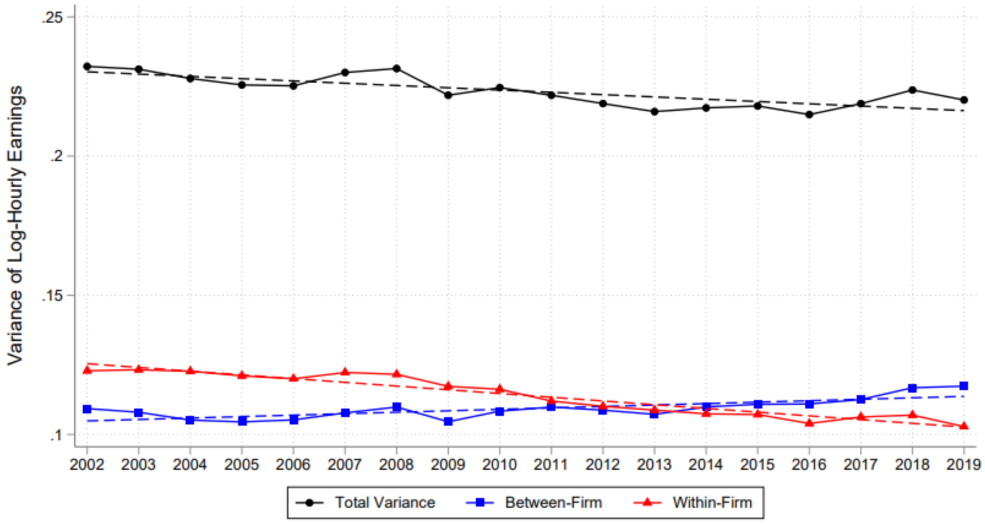
The increasing segregation between workers at different levels of earnings is mainly due to changes in the way occupations are distributed across firms. Firms are becoming more specialised in the types of occupations they employ. For example, firms are increasingly concentrating on either high-skilled occupations, such as managers and engineers, or low-skilled occupations, rather than having a mix of both.

The study also looked at whether these changes could be explained by other factors. We found only a modest role for rising returns to skill (where skilled workers command increasingly higher wages). We also found little evidence that changes in the way firms share profits with workers explain the patterns.

Rather, these trends reflect wider changes in the way work is organised across firms. Technological changes, particularly in information technology, have made it easier to coordinate work across company boundaries. At the same time, financial pressures have led firms to focus on their 'core' activities and to simplify their structures. This has led to more outsourcing and specialisation, with different occupational groups that used to work together within the same company now being spread across different companies.

This increasing segregation of workers by productivity level and occupation can have important social implications. As firms become more homogeneous in terms of the types of workers they employ, there are fewer opportunities for interaction across social and economic boundaries within workplaces. This could potentially reduce social mobility and increase inequality of opportunity, even if overall wage inequality remains stable.

Figure 1 - Evolution of wage inequality - France



This figure shows the evolution over time of the variance of log-wage, the between-firm variance of log-wage, and the within-firm variance of log-wage. All individuals employed for at least 360 days by the same firm during the year are included for a given year. Individuals and firms in public administration are not included. Source: Base Tous Salariés (2002-2019).

Retour sur AKM : expliquer la dynamique des inégalités salariales en France

RÉSUMÉ

Nous utilisons une nouvelle base de données appariées employeur-employé quasi-exhaustive afin d'étudier la contribution des entreprises à l'inégalité salariale en France. Nous appliquons une correction simple et facile à mettre en œuvre pour corriger le biais de mobilité limitée. Notre analyse, qui couvre la période 2002-2019, révèle une augmentation de l'inégalité salariale entre les entreprises, principalement due à la concentration croissante de travailleurs ayant une valeur marchande similaire. Ces phénomènes sont associés à une spécialisation professionnelle accrue au sein des entreprises. Nos résultats soulignent l'importance des estimations de Abowd, Kramarz et Margolis (1999) –ci-après AKM- corrigées des biais pour comprendre la dynamique de l'inégalité salariale et montrer comment les catégories professionnelles observables et les caractéristiques individuelles non observables contribuent à ces tendances.

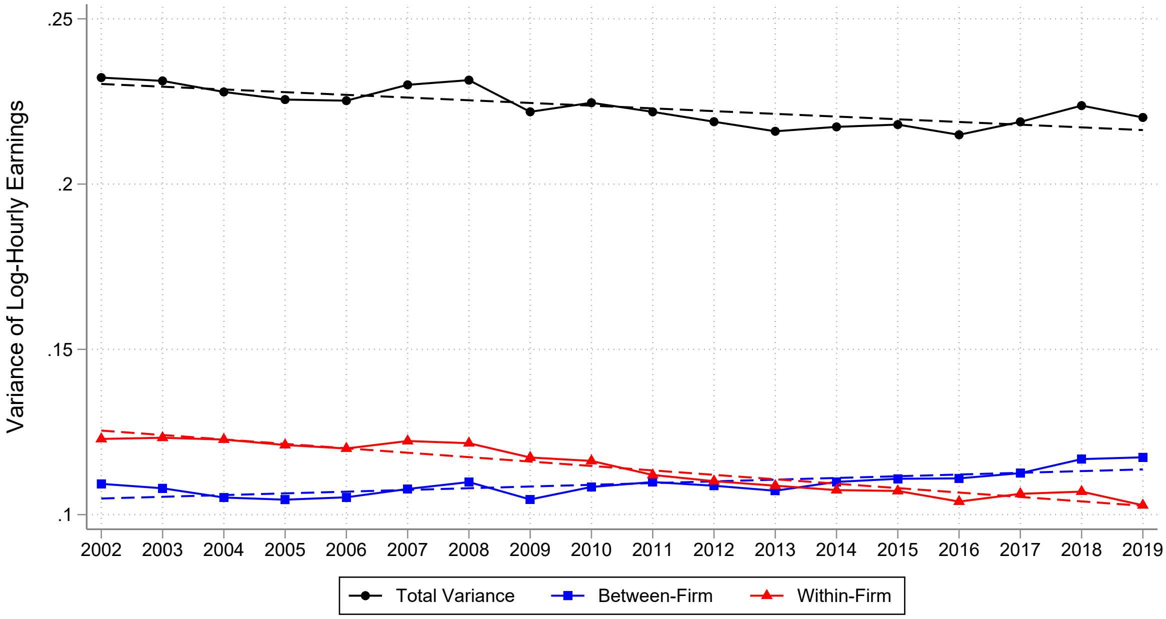
Mots-clés : inégalité salariale, ségrégation des travailleurs, tri par métier, données employeur-employé.

Les Documents de travail reflètent les idées personnelles de leurs auteurs et n'expriment pas nécessairement la position de la Banque de France ou de l'Eurosystème. Ils sont disponibles sur publications.banque-france.fr

Introduction

Wage inequality is a driving force of economic inequalities. Its rise over several decades in most rich countries is well documented[1]. Firms play a central role in driving these dynamics, as evidenced by studies conducted in Germany (Card, Heining, and Kline 2013) and the USA (Davis and Haltiwanger 1991; Barth et al. 2016; Song et al. 2019), where rising inequality is largely attributed to between-firm wage disparities. France is an interesting case. Wage inequality has remained stable or even declined in recent decades, despite polarizing dynamics similar to those observed in other countries characterized by increasing inequalities between firms (Figure 1).

FIGURE 1. Evolution of wage inequality - France



Notes: This figure shows the evolution over time of the variance of log-wage, the between-firm variance of log-wage, and the within-firm variance of log-wage. We compute the overall variance of log-earnings as $\frac{1}{N_t}\sum_i (w_{it} - \overline{w}_t)^2$, the between-firm variance as $\frac{1}{N_t}\sum_f N_{ft}(\overline{w}_{ft} - \overline{w}_t)^2$ and the within-firm variance as $\frac{1}{N_t}\sum_f \sum_{i\in f}(w_{it} - \overline{w}_{ft})^2$, where workers are indexed by i and time by t and firms by f. N_t and N_{ft} denote the number of workers in total and in each firm, respectively; w_{it}, $\overline{w}_t$ and $\overline{w}_{ft}$ are the log worker wage, the overall average log wage and the average log wage within each firm, respectively. All individuals employed for at least 360 days by the same firm during the year are included for a given year. Individuals and firms in public administration are not included.

To explore the factors contributing to the French exception, we employ the Abowd, Kramarz, and Margolis (1999) model (hereafter AKM) to decompose log-wage variance into between- and within-firm components. This model, originally developed to test

[1] Tomaskovic-Devey et al. (2020) and OECD (2021) for a recent international comparison.

efficiency wage theories, has evolved to become a powerful tool for analyzing the sources of wage inequality. This literature revealed a significant increase in *sorting*, wherein high-wage workers are increasingly employed in high-premium firms, with the correlation between worker and firm fixed effects rising steadily in both Germany (Card, Heining, and Kline 2013) and the United States (Song et al. 2019). While remaining agnostic about its relationship to firm efficiency, these studies have uncovered a key mechanism driving inequality: workers at the lower end of the wage scale are increasingly excluded from firm wage premiums.

However, the sorting of workers and firms is not the only mechanism behind the increase in the between-firm component of wage variance. Song et al. (2019) identifies another channel: *segregation*, which captures the clustering of workers with similar permanent person-specific components of wages within firms, regardless of the firm's wage premium. The two mechanisms, sorting and segregation, are distinct. Sorting can increase even if segregation remains constant or decreases (e.g., if firms with high-wage workers increase their wage premium), and conversely, segregation can increase while sorting remains stable or decreases (e.g., if firms with low wage premiums increase their concentration of high-wage workers).

In this paper, we examine the respective roles of sorting and segregation in France, the birthplace of the AKM estimates, between 2002 and 2019, using a newly created, quasi-exhaustive matched employer-employee dataset. We introduce a novel, simplified approach based on split sampling to address the limited mobility bias in AKM estimates and confirm significant differences between uncorrected and corrected estimates. Importantly, we find that the limited mobility bias is not constant over time, as previously assumed (Card, Heining, and Kline 2013; Song et al. 2019). Uncorrected estimates attribute about 60% of the increase in between-firm wage inequality to sorting, while our corrected estimates reduce this figure to about 10%. Conversely, our correction reveals a stronger contribution of segregation to the increase in the between-firm component of the wage variance (93% instead of 73% in the uncorrected estimates). This substantial difference highlights the importance of using corrected estimates to accurately capture the relative importance of the two channels in driving between-firm inequality. In addition to the between-firm divergence in average worker fixed effects, we uncover a within-firm convergence in worker fixed effects in France that accounts for 80% of the decline in within-firm wage variance. This trend complements the diagnosis of increasing segregation in France, suggesting a growing concentration of similarly valued workers within the same organisations.

To provide a broader historical context, we extend our analysis using long-term series from 1976 and rolling panel decompositions from our main dataset. This extended perspective shows that sorting started to increase in the mid-1990s and has been procyclical in recent years. Segregation has been on a steady upward trend since the early 1980s, with only a brief plateau from the mid-1990s to the mid-2000s.

An important feature of our data is the availability of high-quality longitudinal information on workers' occupations. We find that both observed occupational characteristics — the average worker fixed effect associated with a particular occupation — and unobserved worker characteristics within occupations— individual deviations from this average — play a significant role in the increase in between-firm inequality. We observe stronger segregation patterns based on both occupational and individual worker characteristics, with differences between firms increasing in both dimensions. In other words, workers in occupations with similar market values and workers with similar individual-specific wage components within their occupation are increasingly likely to work together. The sorting of individuals across firms becomes more pronounced on the basis of the unobserved, individual-specific components. Interestingly, while within-firm occupational specialization has increased, we also observe a narrowing of the distribution of average worker fixed effects across occupations. This trend is driven by high-skilled occupations such as managers and engineers, whose share of employment has increased but whose relative market value has decreased. We rule out alternative explanations for the increase in inequalities between firms. We find a modest increase in the return to skills and no change in the rent sharing behaviour of firms.

Contribution to literature. This paper contributes to the existing literature in three main ways. First, we create a new, quasi-exhaustive matched employer-employee dataset for France which allows us to provide the first decomposition of the variance of log wages and its evolution from 2002 to 2019[2].

Second, we propose a novel method to correct for the well-known "limited mobility" bias in the measurement of sorting through AKM models described in Abowd

[2] The AKM model of log wages with additive workers and firm fixed effects was originally estimated on French data: a panel sample of 1/24th of French wage earners (excluding civil servants) from 1976 to 1987. The paper inspired much subsequent work on matched employer-employee datasets, most of them from countries where exhaustive, panelized administrative data were available to researchers: this exhaustiveness turned out to be essential for the quality of estimation of these models. We construct such a dataset for France in order to bring AKM back to the state of the art in its original land. A first panelization of the BTS was developed in Godechot et al. (2020). We improved the algorithm and made it public for further use, see Appendix C.

et al. (2004), Andrews et al. (2008), and Bonhomme et al. (2023). This bias is due to the limited number of observations available for individual firm and worker parameters. The individual parameter estimates remain unbiased, but the variance of the error term is underestimated. We use a split-sampling strategy (building on Chanut (2018), Kline, Saggio, and Sølvsten (2020) and others) and provide proof that, under reasonable hypotheses analogous to Kline, Saggio, and Sølvsten, split-sampling corrects the limited mobility bias in quadratic terms. We also implement Bonhomme, Lamadon, and Manresa (2019) firm clustering method (without random effects) and find the results consistent with split-sampling. Both Card, Heining, and Kline (2013) and Song et al. (2019) acknowledge the mobility bias in sorting measurement but assume its stability over time. Our corrected results challenge this assumption, revealing the potential impact of the mobility bias on observed dynamics.

Finally, to gain a deeper understanding of the rise in segregation and sorting, we investigate the role of the occupational employment and wage structure. Card, Heining, and Kline (2013) find that occupations were important in explaining wage inequality in Germany. They show that higher-wage occupations became more prevalent in high-wage firms, while lower-wage occupations became more common in low-wage firms. Our analysis extends this work in three ways. First, we take a closer look at firm specialization and show that firms are becoming more similar in the types of occupations they employ. This may be due to changes in the division of labor across firms, possibly through outsourcing, which has been studied in other countries (Goldschmidt and Schmieder 2017; Dorn, Schmieder, and Spletzer 2018; Drenik et al. 2023) and recently in France (Bergeaud et al. 2021; Bilal and Lhuillier 2021; Godechot et al. 2024). Second, we examine the dynamics within occupations by focusing on individual-specific skills. These are measured as deviations of a worker's fixed effect from the average worker's fixed effect in his occupation. We find that workers with similar individual-specific skills are increasingly clustered together. Moreover, the sorting of workers across firms based on these individual-specific skills has intensified over time. Third, we look at how the market value – the average worker fixed effects – of different occupations has changed over time. We find that differences between occupations have actually become smaller, mainly because the wage components associated with growing high-skill occupations have declined over time. This "skill dilution" effect, where expanding high-wage occupations attract relatively less-skilled workers, is consistent with the findings of Böhm, von Gaudecker, and Schran (2024) for Germany.

Outline. Section 1 details our data construction process. Section 2 outlines our empirical approach, including the AKM model-based log-wage variance decomposition and bias correction techniques using split-sampling and firm clustering. We present our main findings in Section 3. Section 4 explores the key drivers behind the observed empirical trends, with a particular focus on the role of occupations. Section 5 concludes with a discussion of our findings and their implications.

1. Data

1.1. Building an Exhaustive Pseudo-panel

We use BTS data, which stands for "Base tous salariés", or "all wage-earners file"[3]. This is an exhaustive annual dataset built by the French national statistical institute (Insee) on the basis of tax declaration files provided by firms on their payroll employees. This dataset serves as the source for French official statistics on wage evolution.

To conduct panel analysis, we need to address the issue of pseudonymity in the data. Each individual is assigned a unique identifying code that changes every year, allowing for cross-sectional analysis but not for long panel analysis. In France, panel analysis on matched employer-employee wage data is usually carried out using the "BTS panel" or "DADS panel," which is a narrower panel constructed from a sample of 1/24th of the data before 2002 and 1/12th after. This "narrow panel" samples the same individuals as a permanent demographic panel. The sampling also facilitates additional data quality checks and corrections that would be more challenging with the exhaustive data. The oldest years of this narrow panel were the basis for the original AKM. The narrow panel remained the basis for later AKM estimations on French data, notably in Abowd, Kramarz, and Roux (2006) and Coudin, Maillard, and Tô (2018).

However, since 1999, AKM models have been estimated more accurately in countries where researchers have had access to exhaustive panel data, such as the USA, Germany, Sweden, Austria, Italy, Norway, and Denmark. The reduction in sample size and precision due to sampling increases the uncertainty in the estimation process. For instance, in the narrow panel, firms need to be about 12 times larger to have their firm fixed effects estimated with the same precision as in the exhaustive data. This decrease in precision introduces a larger "limited mobility bias" for variance and sorting estimates.

[3] formerly known as "DADS", which stands for "déclaration annuelle de données sociales" or "annual social data declaration". DADS were the main source for BTS, supplemented by other administrative sources (public sector data for instance), and has been gradually replaced by a new administrative source, the DSN ("déclarations sociales nominatives") since 2016.

In addition, the identification of AKM models relies on mobile workers moving between firms, and this is only possible within the group of firms that are linked by such workers. Sampling drastically reduces the proportion of firms belonging to the main connected component and biases the estimation towards larger and more connected firms.

Each BTS annual file contains data for both the current year and the previous year. This overlap allows for matching between annual files based on common information, such as establishment ID, gender, number of hours, job duration, start and end dates of the job, municipality of work and residence, earnings, and age. The matching procedure provides a single match for about 98% of individuals between 2002 and 2019, except for the period 2016-2018 where it drops to 91%-93%. However, there are rare cases where matching is not possible, such as when all matching variables are identical for several individuals or when individual data are changed between annual files. Workers who are not matched due to career breaks or missing matches still appear in the panel, but they are represented by multiple identification numbers. This results in an almost exhaustive pseudo-panel that we call the “wide panel”.

Before 2002, the matching procedure is not applicable as there is no link within the annual files between the different job spells of a single worker. In other words, it is not possible to follow a worker through different employers, even within the same year. Therefore, matching is only possible for workers who have had the same job for two consecutive years. As a result, the AKM estimation cannot be performed before 2002[4].

We have computed long-term series on sorting from 1976 to 2019 using the narrow panel, which provides wage and career information since 1976 with some missing years and varying data quality.

To supplement the analysis, we incorporate exhaustive firm financial data from administrative sources (FICUS/FARE files), which are matched to the wage files and provide information on value-added per worker and total employment at the legal unit level. We use legal unit identification numbers (SIREN) as our empirical units for firms[5]. Estimates based on establishment identification numbers (SIRET) are very similar[6].

[4] The matching procedure is detailed in Appendix C.

[5] Our approach is in line with other studies that have used the AKM method to analyse French data. More recently, Insee has started to provide datasets on groups based on financial links between legal units. The historical depth is not yet sufficient to measure trends at this level of observation.

[6] See Table A4.

1.2. Sample Restrictions

We exclude public employees because they are not included before 2009. We focus on ordinary jobs, excluding subsidized contracts, interns, and apprenticeships. Both men and women are included in the analysis. We limit our analysis to metropolitan France.

We divide the data into three adjacent six-year periods: 2002-2007, 2008-2013, and 2014-2019[7]. Each observation consists of a worker / firm / year triplet, where each individual worker is associated with the firm from which she earned the most during the year (or, when equal, for which she worked the most). We refer to these observations as a "wage", or a "job" for convenience. Each worker can appear up to six times in each sample period.

We have information on the number of paid hours, which is rare in this kind of data. Without it, it is common practice in the literature to set a minimum earnings threshold for inclusion and to exclude women to reduce the risk of misidentifying part-time workers. We can keep both part-time and full-time employees and avoid these exclusions. Our target earnings variable is the log hourly gross wage (including employees' social contributions but excluding those of the employer). We restrict the sample to individuals employed for the whole year to minimize the impact of annualized payments[8]. We limit the impact of possibly erroneous extreme values by some additional selection. We exclude jobs with an hourly wage below 80% of the legal minimum hourly wage for the corresponding year, or above 1000 times the minimum hourly wage, with less than 100 working hours per year, and observations with missing values for sex, age and employer. We only keep workers between the ages of 16 and 70. These restrictions are applied after matching when the wide panel is already constructed, selecting specific observations while retaining individuals who have been working during the year or receiving unemployment benefits. All sample restrictions may introduce selection effects that may change over time. For example, the restriction to workers employed for the whole year excludes workers who are more likely to be at the bottom of the wage distribution.

In our historical series computed on the narrow panel, paid hours are not reliable before 1996, but job duration in days and an indicator variable for part-time jobs are available since 1976. There have been other changes in the variables over the 44 years.

[7] Card, Heining, and Kline (2013) divide their 1985-2009 data into four overlapping seven-year panels, and Song et al. (2019) divide their 1980-2013 data into five adjacent six-year panels.

[8] We found similar results when extending the selection to all individuals whose main job during the year lasts more than 90 days (Table A8).

For example, the distinction between the public and private sectors is not consistent[9]. Because of these changes, it is not possible to precisely replicate the selection criteria used in the 2002-2019 wide panel. We use a slightly different selection based on Insee's long-term, private sector series selection.

2. Methodology

2.1. The AKM Model

We follow AKM with an additive model of log wages :

$$y_{it} = \beta x_{it} + \theta_i + \psi_{j(i,t)} + u_{it} \tag{1}$$

Here y_{it} is the logarithm of the hourly wage of worker $i = 1, 2, \ldots, N$ during year $t = 1, \ldots, T$, and $X_{i,t}$ controls for a cubic polynomial in age[10] and year dummies. This model relies on two notable assumptions. First, it assumes no interaction effect between firm and worker type, implying that the fixed effects enter the log wage additively in. This implies that a firm's wage premium ($\psi_{j(i,t)}$) is constant across all worker types, regardless of characteristics such as gender, age, or skill level. Second, the model assumes exogenous mobility, where the residual term u_{it} is strictly exogenous with respect to the variables x_{it}, i, t, and j, as is traditionally assumed. This assumption implies that, on average, wages before and after a job change remain the same as they would have been without the job change, except for the difference in firm effects ($\psi_{j(i,t+1)} - \psi_{j(i,t)}$). Although both hypotheses may seem unrealistic and have been subjected to scrutiny, they seem to provide reasonable approximations[11].

2.2. Log-wage Variance Decomposition

Following Card, Heining, and Kline (2013) and Song et al. (2019), we take $V(y) = Var(y_{it})$ as a measure of wage inequalities and observe its evolution through 3 six-year periods:

[9] Partly because of nationalization and privatization of firms, such as banks, airlines, posts, and telecommunications, etc. over the period, which is difficult to follow in the data.

[10] Following Card et al. (2018), we normalize the age term to be flat at 40 and exclude the linear term to avoid collinearity with worker and year effects.

[11] Recent studies have found only small deviations from the additive linear model (Bonhomme, Lamadon, and Manresa 2019) and limited impact of more dynamic specifications (Di Addario et al. 2023). While there are potential specification errors, such as evolving firm "fixed" effects, research suggests that firm premiums are mostly stable over time (Engbom, Moser, and Sauermann 2023; Lachowska et al. 2023), although possibly procyclical.

2002-2007, 2008-2013 and 2014-2019. Ignoring for simplicity of exposition the time-varying workers variables x_{it}, we can decompose, for each period, $V(y)$ as a sum of the variances of θ, ψ, u, and their respective covariances, estimated over all worker-years observations:

$$V(y) = V(\theta) + V(\psi) + V(u) + 2Cov(\theta, \psi) \tag{2}$$

Song et al. further distinguish within-firms and between-firms components of wage variance, and extend the law of total variance $V(y) = V[E(y|j)] + E[V(y|j)]$ to:

$$V(y) = \underbrace{V(\bar{y}_j)}_{\text{Between-firm component}} + \underbrace{\sum_j m_j \times V(y_i | i \in j)}_{\text{Within-firm component}} \tag{3}$$

$$V(y) = \underbrace{V(\psi) + 2Cov(\bar{\theta}_j, \psi) + V(\bar{\theta}_j)}_{\text{Between-firm component}} + \underbrace{V(\theta_i - \bar{\theta}_j) + V(u)}_{\text{Within-firm component}} \tag{4}$$

with $\bar{y}_j = \bar{y}_{j(i,t)}$ and $\bar{\theta}_j = \bar{\theta}_{j(i,t)}$ the respective expectations on i, t in firm j. By hypothesis, the analogous $\bar{u}_j$ is equal to 0. All moments of the distribution of firm variables are weighted by the share m_j of each firm in the total number of observations either directly as in Equation 3 or implicitely when computing variance over all (i, t) observations as in Equation 4. Our interest lies in the evolution of the sorting component of this decomposition, $2Cov(\theta, \psi)$, which is by construction entirely contained in the between-firm component of wage variance, and in the evolution of segregation, which we define in this context as $V(\bar{\theta}_j)$ as in Song et al. (2019). Segregation captures the extent to which high-wage workers tend to work with one another, and low-wage workers with one another. As with other dimensions of segregation (residential, school, etc.), workplace segregation by individual wage levels may affect patterns of social interaction and networking opportunities, but it has no direct effect on overall wage inequality, because the increase in between-firm variance comes from a decrease in within-firm variance, leaving the overall distribution unchanged[12].

[12] We also compute Song et al. (2019)'s "Segregation Index" as $Var(\bar{\theta}_j)/Var(\theta_i)$ when possible.

2.3. Limited Mobility Bias

The so-called "limited mobility bias" affects the estimation of variance and covariance terms in the previous decompositions. It has a simple cause : each fixed effect is estimated with an estimation error. The estimated variance is thus equal to the true variance plus the variance of the estimation error. In other contexts, the limited mobility can also be thought of as an incidental parameter bias, or as overfitting in a statistical learning context. The importance of the bias in our context is due to three elements: (i) the large number of parameters to be estimated, each with a small number of observations; (ii) the focus on quadratic transformations, variances and covariances, of these noisy estimates; (iii) the sparse, centralized, and clustered network structure of the design matrix, which can approach collinearity. The last aspect, resulting from the network of mobility of workers, gives the name to the bias. To illustrate intuitively, consider two large groups of firms connected by only a handful of mobile workers. The identification of the relative premiums between these groups will depend heavily on these few observations. While the parameter estimates themselves are unbiased, the large estimation errors lead to biased variance terms: the variance of individual and especially firm effects is overestimated, while their covariance, which measures sorting, is underestimated.

Several correction strategies have been proposed in the literature. Andrews et al. (2008) directly correct estimates using a bias correction factor derived from the error term variance estimate. However, their homoscedasticity assumption is unrealistic due to the networked nature of the estimation error (Jochmans and Weidner 2019). Borovičková and Shimer (2017) model heterogeneity as random effects rather than fixed effects and find much higher sorting than previous estimates. However, while fixed effects models allow for further exploration of the heterogeneity and distribution of fixed effects, this is more challenging with random effects models. Bonhomme, Lamadon, and Manresa (2019) cluster firms based on the similarities between their wage distributions, then estimate a wage model where workers' effects are treated as random effects. The clustering creates a dense mobility network with many observations per cluster, which allows for the estimation of richer models, including interaction and dynamic terms, at the cost of the additional hypothesis that clusters are correctly identified. Kline, Saggio, and Sølvsten (2020)'s leave-one-out strategy is equivalent to Andrews et al. (2008)'s bias correction factor method but is compatible with heteroscedasticity. However, it is complex and computationally expensive for large datasets[13].

[13] Kline, Saggio, and Sølvsten provide a more tractable estimation method using a large number

In Section 2.3.1, we introduce a simpler and more computationally efficient split-sampling strategy to correct for the limited mobility bias in quadratic terms. In addition, we implement Bonhomme, Lamadon, and Manresa (2019)'s clustering method (without random effects) in Section 2.3.2 and find results consistent with split sampling.

2.3.1. Split-Sampling Bias Correction

We employ a simpler split-sampling strategy than Kline, Saggio, and Sølvsten, applying only one split to our data instead of a leave-one-out method with as many splits as observations. Our strategy requires only two estimations on two half-samples, at worst doubling computing time[14]. The bias arises because the estimated effects are inconsistent with correlated estimation errors. However, they are unbiased. Thus, if one obtains two independent unbiased estimates from two different samples (i.e., split-samples), then the covariance in these estimates is informative about the underlying variance or covariance. Chanut (2018) introduced split-sampling in a similar setting with French data, and demonstrated its bias-correcting properties using a toy example. Other authors in similar settings implement split sampling either for instrumental variable estimation or to compare different groups of workers, usually without explicitly mentioning that they are correcting for the limited mobility bias: Drenik et al. (2023), Goldschmidt and Schmieder (2017), Gerard et al. (2021), Godechot, Safi, and Soener (2021), Sorkin (2018), Frederiksen, Kahn, and Lange (2020) and Schoefer and Ziv (2022). We extend these works by generalizing the concept and proving that, under reasonable hypotheses analogous to Kline, Saggio, and Sølvsten, split-sampling corrects the limited mobility bias in quadratic terms (Appendix D.1).

Split-sampling introduces additional uncertainty due to the reduced effective sample size. The sample splitting strategy must be carefully considered in this context. In each split sample, the main connected set is smaller than in the original sample, and the two are distinct. Consequently, the common sample of workers and firms belonging to the main connected set in both split samples is reduced, as is the corresponding parameter vector that can be estimated in both split samples. The simplest approach is a direct random split of observations in two equally sized samples. By balancing the sampling

of random projections (in the hundreds). Bonhomme et al. (2023) still find this method demanding and further approximates it, although they express concern that sequence of approximate estimates, combined with those typical of AKM models, may have poorly understood consequences.

[14] We also leveraged the lower computation time of the R package *fixest* (Bergé 2018), up to ten times faster than the R package *lfe* (Gaure 2013) or the Stata package *reghdfe* (Correia 2016), although we used all three packages depending on the setting.

by workers, and splitting for each worker the periods of observation, one increases the probability that each worker is present in both samples' main connected set. We call this method "period splitting". Conversely, by splitting individuals rather than observations, one increases the connectivity in each set because individual careers are preserved. By balancing this individual split by firm, we increase the probability of each firm's fixed effect being estimated in each sample. We call this method "firm splitting". Under firm splitting, each firm with two workers or more is present in both samples and belongs to each main connected set if it remains connected with each random half of its employees. Firm splitting, however, estimates each worker's fixed effect only once, which precludes direct correction for $var(\theta)$ and $var(u)$ quadratic forms through split sampling. We describe the split algorithms in detail and discuss the computation and approximations of the corrected variance and covariance components in Appendix D.2.

We show that in our data using firm splitting, the additional uncertainty due to the reduced sample size is small compared to the bias reduction effect. We provide standard deviations computed on multiple random splits (Appendix D.3) and Monte Carlo experiment results (Appendix D.4), which confirm the stability of the procedure and offer reassurance about consistency and convergence rates.

2.3.2. Firm Clustering

We also implement Bonhomme, Lamadon, and Manresa (2019) strategy. We ran a firm-clustering algorithm with 1000 clusters (explaining around 90% of the between-firm dispersion in earnings for each period) before estimating AKM on firm clusters (rather than individual firms), with the hypothesis that firms' fixed effects are discretely distributed with a small number of values. The clustering algorithm is a kmeans clustering based on quantiles of the wage distribution, as the identification of clusters can not rely on firm mean wage and must use higher moments of the distribution of wages[15]. This approach addresses the limited mobility bias by reducing the dimensionality of the problem and pooling information across similar firms. The resulting mobility network between clusters is very dense, effectively increasing the sample size and connectivity for each cluster. As a result, cluster fixed effects generally have smaller standard errors, indicating a more precise estimation compared to individual firm effects in traditional AKM models.

[15] Following the original Bonhomme, Lamadon, and Manresa (2019) specification, we do not add additional firm variables to feed the clustering algorithm. Such developments are possible.

Strengths and weaknesses of the proposed correction methods. Period splitting may not fully correct for limited mobility bias, as the error terms (u_i) are likely to be correlated across multiple observations of the same employer/employee pair. However, our specific setting mitigates this problem. We only include observations with full-year jobs, and movers are typically observed for five years or less within the six-year panel. These restrictions reduce the probability that a worker is a mover in both samples after random splitting. Because our estimation relies only on movers, individual residuals are less likely to be correlated on either side of the split. Firm splitting completely avoids the drawback of correlated individual residuals by keeping all of an individual's observations on one side of the split. Moreover, while both firm and period splitting require two moves per firm to be estimated, period splitting can only be estimated for workers who are in the sample for at least two years, which further reduces the sample size. Therefore, we prefer firm splitting for presenting baseline corrected estimates of sorting and segregation.

Firm clustering comes with the additional hypothesis that firms' fixed effects are discretely distributed. This approach carries a risk: firms may be clustered based on some combination of their own fixed effects and the average workers' fixed effects of their employees. Consequently, an AKM estimation following this procedure could potentially show higher sorting and lower cluster effect variance than is actually the case. Bonhomme, Lamadon, and Manresa (2019) acknowledge the risk, mention job-market models that satisfy the conditions for cluster identification[16], and provide in-depth robustness analysis suggesting that this risk has limited impact in practice[17].

Firm splitting, while faster due to the high computational cost of the clustering step, relies solely on AKM hypotheses. However, it further reduces the estimation sample to individuals or firms that belong to the main connected components in each split. In contrast, firm clustering allows the use of the entire dataset[18]. We view the convergence of the two methods as strong evidence of the robustness of our results. Descriptive information characterizing both the full population and the different connected sets

[16] "In some environments without firm capacity constraints, such as Postel-Vinay and Robin (2002), the upper bound of earnings in the firm is increasing in firm productivity, so firm-specific distributions are all different and firms may be consistently classified based on their earnings distributions. It is difficult to obtain similar guarantees in models with capacity constraints" (p. 217).

[17] More generally, Bonhomme, Lamadon, and Manresa (2022) provides theoretical conditions and convergence results for the *two-step grouped fixed-effects* (GFE) method, where the clusters are viewed as an approximation to the underlying continuous unobserved heterogeneity.

[18] Following BLM, we cluster firms using only the empirical distribution stayers' wages. As a consequence, only firms with at least one stayer in each period are selected. This explains the small discrepancy in Table A7's last row and Table 1's overall sample person/yr observations.

is provided in Table 1 and A1. The main difference between the connected sets and the full sample in each period is firm size: firms that belong to the main connected components in each split are on average larger.

TABLE 1. Summary statistics

				Log-Hourly Wage	
	Person*yr	Individuals	Firms	Mean	Std.Dev.
			Overall Sample		
2002-2007	65,457,069	21,356,960	1,203,830	2.67	0.46
2008-2013	68,998,598	18,595,890	1,232,452	2.81	0.45
2014-2019	67,928,369	20,894,672	1,397,646	2.91	0.46
			Largest Connected Set		
2002-2007	58,666,317	18,842,389	536,814	2.69	0.46
2008-2013	61,413,372	16,266,899	560,778	2.83	0.45
2014-2019	59,550,287	18,065,244	564,113	2.94	0.46
			Firms in Both Connected Sets		
2002-2007	52,154,249	16,704,724	230,895	2.70	0.46
2008-2013	54,628,124	14,444,257	238,110	2.84	0.46
2014-2019	53,141,843	16,098,477	228,222	2.96	0.47

Note: In the overall sample, all firms and individuals in firms with at least 1 employee are included. Only individuals employed for at least 360 days by the same firm during the year are included for a given year. Individuals and firms in public administration are not included. The largest connected set entails the group of firms connected by worker mobility. Firms in both connected sets refer to firms present in both main connected components in each split sample for the firm splitting method (Section 2.3.1).

3. A Robust Rise in Segregation and a Mild Increase in Sorting

To conduct the baseline variance decomposition, Table 2 applies two variance decompositions (Equations 2 and 4) to the estimates from AKM, corrected using the split-sampling method with firm splitting. The correction involves splitting the data at the individual level and balancing it by firm (see Section 2.3.1). However, our baseline correction method does not allow for the direct computation of corrected and distinct estimates of $Var(u)$ and $Var(\theta)$. To provide a comprehensive analysis of their evolution, we rely on several alternative specifications, which are presented in Appendix A.

As shown in Figure 1, the analysis reveals an increase in between-firm inequalities and a decrease in within-firm inequalities. The overall log-hourly wage variance increased slightly from 2002-2007 to 2014-2019, moving from 0.213 to 0.220[19]. France's quasi-stable wage inequalities during this period are atypical among developed countries. Nevertheless, the rise in between-firm wage inequalities aligns with the findings of Barth et al.; Song et al. for the US and Card, Heining, and Kline for Germany, and more generally with the global trend established for 14 high-wage countries (Tomaskovic-Devey et al. 2020). The increase in between-firm variance can be decomposed into several components (Equation 4): the rise in $Var(\bar{\theta})$ reflecting segregation, the rise in $2Cov(\theta, \psi)$ capturing sorting, the decrease in firm premium variance $Var(\psi)$, and age effects[20].

It can be concluded that the growing gap between firms is mainly driven by the increase in segregation, and to a lesser extent by the increase in sorting. In 2002-2007, segregation accounted for 20.4% of total log-hourly wage variance, increasing to 26.2% in 2014-2019. Thus, the increase in segregation, accounts for 93% of the increase in the between component of the wage variance. The contribution of sorting remains more modest. Sorting represented 12.4% of inequalities in the first period and 13.1% in the last. The increase in sorting therefore accounts for 13% of the increase in the between-component of the wage variance. Finally, the decrease in the variance of the firm fixed effects contributes negatively to the increase in the between wage variance,

[19] it is important to note a qualitative discrepancy between the overall trend and the specific annual log-hourly wage variance as depicted in Figure 1. This discrepancy can be attributed to small methodological breaks in the original data, which were corrected in the graphical evidence but could not be accounted for in the AKM estimations.

[20] In Table 2 and all tables referring to the decompositions 2 and 4, we abstract from the presentation of the variance of the year effects and the associated covariances for simplicity and tractability. Thus, the term *Xb* refers only to age effects. The estimates for the year effects, which always account for less than 1% of the total period-specific variance of log wages, are available upon request.

by a magnitude of -7%.

TABLE 2. Decomposition of wage variance and its evolution
Split-sampling with firm splitting

		2002-2007		2008-2013		2014-2019		Change from 2002-2007 to 2014-2019
		Comp.	Share	Comp.	Share	Comp.	Share	Diff.
Total variance	Var(y)	0.213		0.208		0.220		0.006
	Var (θ)	**	**	**	**	**	**	**
	Var (ψ)	0.014	6.5	0.014	6.6	0.013	5.8	-0.001
	Var(Xb)	0.003	1.4	0.002	1.1	0.003	1.1	-0.001
	Var(u)	**	**	**	**	**	**	**
	2*Cov(θ,ψ)	0.026	12.4	0.027	12.9	0.029	13.1	0.002
	2*Cov(θ,Xb)	0.000	0.0	0.000	0.2	-0.002	-0.7	-0.002
	2*Cov(ψ,Xb)	0.001	0.3	0.001	0.3	0.001	0.3	0.000
Between-firm variance	Var($\bar{y}$)	0.088	41.1	0.094	45.0	0.103	46.8	0.015
	Var ($\bar{\theta}$)	0.043	20.4	0.049	23.8	0.057	26.2	0.014
	Var (ψ)	0.014	6.5	0.014	6.6	0.013	5.8	-0.001
	Var($\overline{Xb}$)	0.000	0.1	0.000	0.1	0.000	0.1	0.000
	2*Cov($\bar{\theta}$,ψ)	0.026	12.4	0.027	12.8	0.029	13.1	0.002
	2*Cov($\bar{\theta}$,$\overline{Xb}$)	0.001	0.6	0.001	0.6	0.001	0.7	0.000
	2*Cov(ψ,$\overline{Xb}$)	0.001	0.3	0.001	0.3	0.001	0.3	0.000
Within-firm variance	Var($y - \bar{y}$)	0.126	58.9	0.114	55.0	0.117	53.2	-0.009
	Var ($\theta - \bar{\theta}$)	**	**	**	**	**	**	**
	Var($Xb - \overline{Xb}$)	0.003	1.3	0.002	1.0	0.002	1.1	0.000
	Var(u)	**	**	**	**	**	**	**
	2*Cov($\theta - \bar{\theta}$,$Xb - \overline{Xb}$)	-0.001	-0.5	-0.001	-0.4	-0.003	-1.4	-0.002
	2*Cov($\theta - \bar{\theta}$, u)	**	**	**	**	**	**	**
	2*Cov($Xb - \overline{Xb}$, u)	0.000	0.0	0.000	0.0	0.000	0.0	0.000
N		52,154,249		54,628,124		53,141,843		

Note: This table presents the decomposition of wage variance and its evolution over three periods using the split-sampling method with firm splitting. "Comp." denotes the component of variance, while "Share"indicates the percentage of total *Var*(y). The last column show the change in levels from 2002-2007 to 2014-2019. The decomposition is based on Equations 2, 3, and 4. The estimation is performed on the sample of firms present in both main connected components in each split sample. The split-sampling method with firm splitting is described in Section 2.3.1.
** : These parameters' estimates are not directly corrected by the firm splitting method

The significant increase in segregation is consistent with recent research by Godechot et al. (2024), which shows a consistent increase in workplace segregation across twelve advanced capitalist economies. Their study shows that top earners, particularly those in the top 10% and 1% brackets, are becoming increasingly isolated from the rest of the wage distribution. In particular, among the countries studied, France has seen the most pronounced increase. However, Godechot et al.'s approach focuses on wage segregation using exposure measures based on gross earnings, without decomposing the respective roles of person and firm fixed effects. This paper extends this analysis by highlighting that the increase in segregation in France is mainly due to the fact that workers are increasingly clustered in firms based on their permanent person-specific components of wages – or, in other words, their average labor market value.

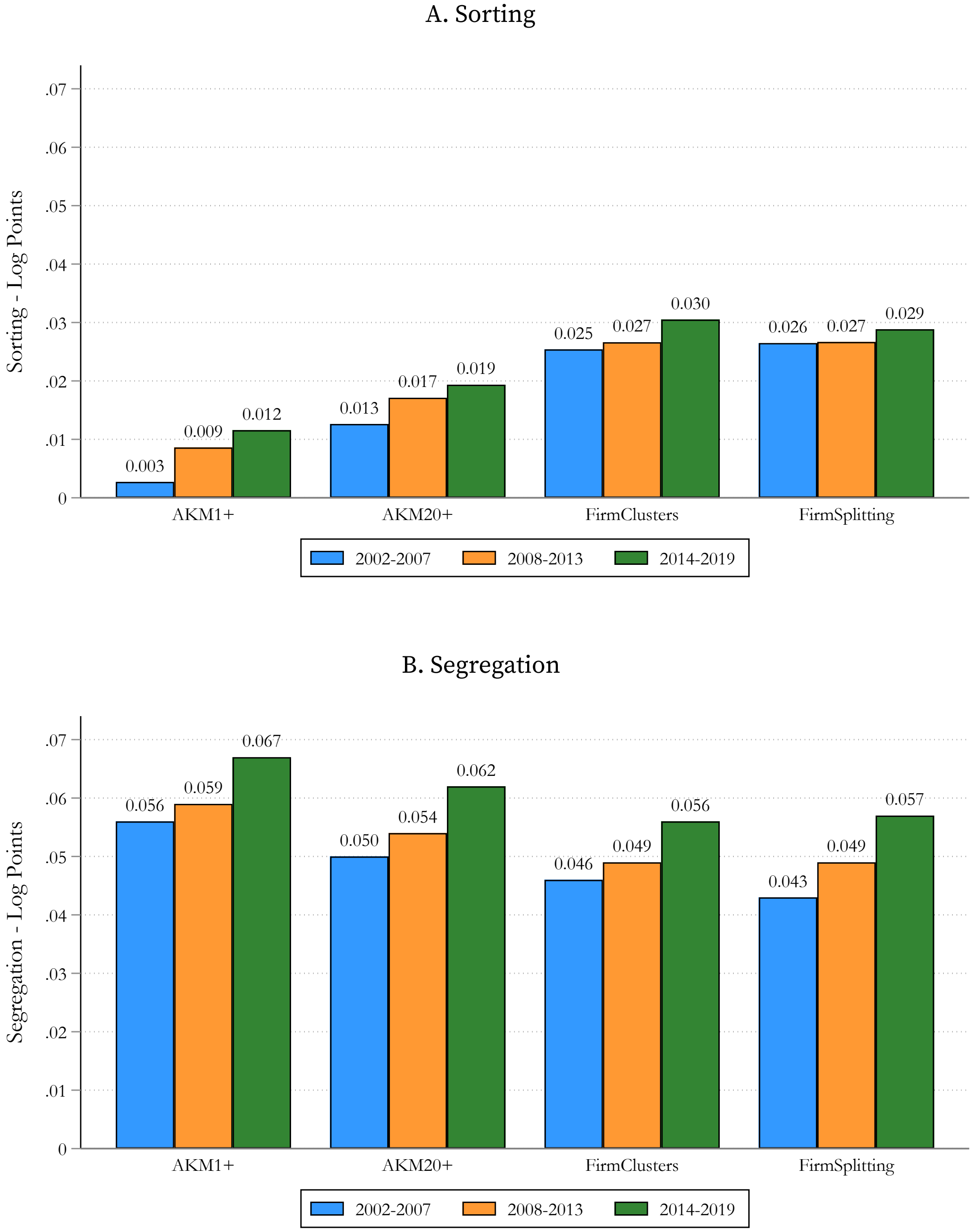
3.1. The importance of Correcting for the Limited Mobility Bias

Previous studies focusing on the evolution of sorting and segregation, such as Song et al. (2019), have attempted to mitigate the mobility bias by imposing a firm size cutoff (n>20) and assumed that above this threshold, the bias, if not entirely eliminated, would remain constant over time. Card, Heining, and Kline (2013) make the same assumption without imposing a size cutoff. In this subsection, we will show that this hypothesis is questionable, at least for France.

Figure 2 shows the levels and evolution of sorting and segregation with four specifications obtained from the main corrections discussed in Sections 2.3.1 and 2.3.2[21]. AKM 1+ shows the estimates before correcting the results for the limited mobility bias. AKM 20+ shows the results estimated when restricting the analysis to firms with more than 20 observations per year in the spirit of Song et al. (2019). In addition to the firm splitting technique, our preferred estimate already presented above, we also show results from cluster-AKM, which proposes an alternative correction for the limited mobility bias.

[21] see Tables A2, A3, 2, and A7 for the full decomposition uncorrected and with the different corrections.

FIGURE 2. Sorting and Segregation over time
Baseline and selected correction strategies


A. Sorting

B. Segregation

Notes: This figure show the estimates of Sorting – $2 * Cov(\theta, \psi)$ – and Segregation – $Var(\bar{\theta})$ by period - coming from a standard AKM estimate (AKM1+), an AKM estimate limited to firms with at least 20 employees (AKM20+), and the two bias correction strategies described in Sections 2.3.1 and 2.3.2. Tables A2, A3, 2, and A7 report for the full wage variance decompositions for the four different methods.

Panel A of Figure 2 shows that the limited mobility bias leads to a substantial underestimation of the magnitude of sorting. While the AKM estimate attributes 1% to

5% of overall wage inequality to sorting, the baseline corrected estimates attribute 12% to 13%. The firm-clustering method produces results close to those of the split-sample method, confirming the magnitude of the bias.

Importantly, the limited mobility bias is not constant over time, as assumed in previous research. In France, the intensity of the increase in sorting is much higher in the uncorrected estimates, or in the partially uncorrected estimates such as AKM20+. For example, in the latter estimates, the increase in sorting accounts for 40% of the increase in between firm inequality - a figure that is similar to Song et al.'s findings for the US. Properly correcting for the limited mobility bias with firm clusters, or even more so with our firm splitting technique, leads to a much smaller increase. This suggests an overall reduction in the limited mobility bias over the period, mainly concentrated in the smallest firms (see Table 3), both those with less than 20 employees and those with between 20 and 200 employees. The reduction in the bias is not driven by an increase in the size of small firms. In fact, within the AKM largest connected set, the average size of small firms decreases from 7.48 to 6.92 between 2002-2007 and 2014-2019. We also examine whether the reduction in bias is due to an increase in the connectedness of these small firms, as measured by the average degree centrality, i.e., the average number of firms to which a firm is connected. However, this measure is remarkably stable over time for very small firms, while it increases for firms with between 20 and 200 employees. This finding is consistent with Jochmans and Weidner (2019), who show that simple connectivity measures may not fully capture the full network structure relevant for statistical inference.

Panel B of Figure 2 shows that the limited mobility bias also leads to an overestimation of segregation measures. While the AKM estimate attributes 26% to 31% of overall wage inequality to segregation, the baseline corrected estimates attribute 20% to 26%. Again, the firm-clustering method produces results close to those of the split-firm method, confirming the magnitude of the bias. However, while correcting for limited mobility leads to similar corrections in the levels of sorting and segregation, it does not have a similar effect on evolution. The overestimation of the segregation coefficient is higher in period 1, where the limited mobility bias is stronger. Its correction thus reveals a stronger contribution of segregation to the increase in the between-component of the wage variance (93% instead of 73% in the uncorrected estimates). These findings highlight the importance of using corrected estimates to accurately capture the dynamics of sorting and segregation. In the French case, this leads to a reassessment of the importance of segregation and a reduction in the role of sorting.

TABLE 3. Sorting decomposition by firm size group and over time

		2002-2007				2014-2019			
		< 20	20 – 200	200 – 1000	> 1000	< 20	20 – 200	200 – 1000	> 1000
Firm Splitting	Overall	0.0264				0.0287			
	Between	0.0028				0.0031			
	Within	0.0259	0.0228	0.0238	0.0250	0.0251	0.0214	0.0249	0.0301
AKM	Overall	0.0027				0.0116			
	Between	0.0031				0.0037			
	Within	-0.0910	-0.0071	0.0202	0.0251	-0.0660	0.0014	0.0228	0.0303
Avg. Size		7.48	52.30	401.20	3,444.23	6.92	53.14	397.09	3,339.53
Avg. Degree Centrality		2.11	8.64	47.20	244.99	2.10	9.49	50.13	265.90

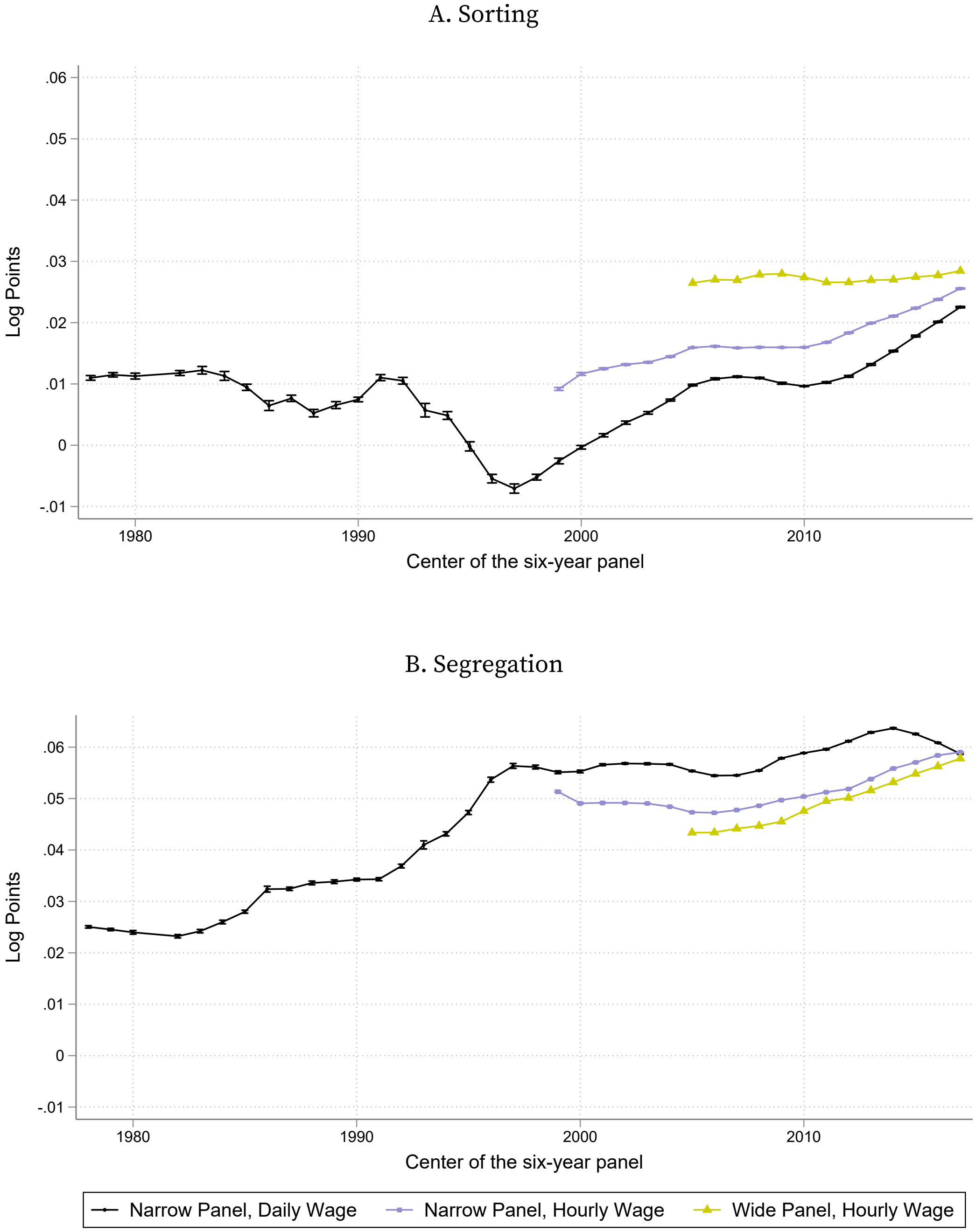
Notes: This table presents sorting decomposition by firm size group for two periods (2002-2007 and 2014-2019) and two methods (standard AKM vs firm splitting). The split-sampling method with firm splitting is described in Section 2.3.1. Estimations performed on the largest connected set for AKM and on firms present in both main connected components in each split sample for firm splitting. Firm size groups are based on the number of observed workers employed by the firm in the panel. The decomposition includes overall sorting (total covariance between worker and firm effects), between sorting ($2Cov(\bar{\theta}, \bar{\psi})$, where averages are computed for each size group), and within sorting (computed separately for each size group). Average firm size and average degree centrality are computed within the AKM largest connected set. Degree centrality represents the number of firms to which a firm is connected through worker mobility.

3.2. Historical Trends

To provide further insight, we document the chronology of the dynamics of sorting and segregation from 1976 onwards using long-term series (see Figure 3). Although long-term series are imperfect, they provide a clear indication of past trends. Sorting started to increase in the mid-1990s, before the period under study. The narrow panel estimates are particularly affected by the selection of larger and more connected firms, which is why we see a much more pronounced increase between 2002 and 2019 compared to the full dataset. Indeed, Table 3 documents a more pronounced increase in sorting in larger firms. The dynamics of sorting show signs of procyclicality in the most recent period, which is more evident in the narrow panel and thus likely to be more pronounced for larger, more connected firms. The trend of increased segregation can be observed from the early 1980s, with a flattening from the mid-1990s to the mid-2000s, before resuming its upward trajectory during our main period of interest, with no interruption during the financial crisis.

In Figure A1 we report the exact same series, using uncorrected estimates for sorting and segregation. The upward dynamics of segregation is much less pronounced, and sorting is always negative for the narrow panel, suggesting a strong bias in the estimation over the entire period. The original study by Abowd et al. (2004), where they applied the Abowd, Kramarz, and Margolis (1999) framework to US and French data, reports a covariance of person and firm effects of -0.0562 using the narrow panel between 1976 and 1996, which is very close to our average results for this period using the same data.

FIGURE 3. Sorting and Segregation over time
Historical series


A. Sorting

B. Segregation

Note: This figure presents estimates of sorting ($2 * Cov(\theta, \psi)$) and segregation ($Var(\bar{\theta})$) using rolling six-year periods. We include only individuals employed by the same firm for at least 360 days in the wide panel, and 90 days in the narrow panel, for a given year. The wider selection criteria for the narrow panel aims to enhance connectivity. Public administration employees and firms are excluded from the analysis. All estimates are corrected by the split-sampling method with firm splitting. For long-term series computed on the narrow panel, we present mean estimates with confidence intervals derived from 20 repetitions of split-sampling, which reflect only the noise stemming from the randomness of the split. Estimates from the narrow panel are particularly affected by the selection of bigger and more connected firms. Data for the years 1981, 1983, and 1990 are missing. There have been several changes in scope and variable definition since 1976.

3.3. Decomposing the Decline in Within-Firm Inequalities

France experienced a decrease in its within-firm variance component, falling from 0.126 in period 1 to 0.117 in period 3 (Table 2). This evolution contrasts with the US case, where inequality increases both between and within firms (Song et al. 2019). While this decline is not unique to France, it is more pronounced there than in other OECD countries such as Denmark or the Netherlands (Tomaskovic-Devey et al. 2020).

Our split-sampling method with firm splitting does not allow a full decomposition of the evolution of the within-firm variance because it does not allow estimating the cross-covariance of the worker fixed effects that are exclusively assigned to one of the two splits. However, it is possible to gain insight into its composition through several approaches: an approximation to the variance of the residuals, which allows the completion of missing estimates (Appendix D.2 and Table A5), the period splitting method (Table A6), the cluster AKM (Table A7), and even the uncorrected 1+ and 20+ AKMs (Tables A2 and A3), all of which give similar results. As shown in Table A5, the decrease in the within-firm variance component is primarily due to the decrease in the within-firm variance of the worker fixed effects ($\Delta(V(\theta - \bar{\theta}))$). This -0.005 decrease in the log wage variance accounts for 60% of the decrease in the within-firm variance. It results mechanically and additively from the fact that the increase in the variance of the worker fixed effects ($\Delta V(\theta) = +0.008$) is smaller than the increase in the firm average variance of the worker fixed effects ($\Delta V(\bar{\theta}) = +0.014$). These results confirm our assessment of increasing worker segregation in France. In contrast to the U.S. case, where worker fixed effects increasingly differ both between firms and within firms (Song et al. 2019), France clearly shows a pattern where, in addition to divergence between firms, workers increasingly work within firms alongside others with similar labor market values.

4. Explaining the Rise in Between-Firm Inequalities

In this section, we test which factors are behind the between-firm empirical trends we find by examining the role of occupations, the impact of changing skill premiums, and the potential influence of firm size distributions and wage premiums dynamics.

4.1. The Role of Occupation

An important feature of our data is the availability of information on occupations[22]. To what extent can the changes in inequality between firms be explained by differences in the occupational mix across firms? To examine how occupations drive between-firm inequalities, we start by defining ω_o as the occupation-specific component of the worker fixed effect (θ_i) for occupation o and $\varepsilon_{i,t}$ as the individual-specific component[23].

$$\varepsilon_{i,t} = \theta_i - \omega_{o(i,t)} \tag{5}$$

The occupation-specific component ω_o captures the average wage effect associated with a particular occupation, reflecting the general skill level and market value of that occupation. For example, this component would capture the typically higher worker fixed effects associated with being an engineer or a manager. The individual-specific component $\varepsilon_{i,t}$, on the other hand, represents the part of a worker's permanent wage that cannot be explained by his or her current occupation alone[24]. This could reflect individual characteristics such as innate ability, quality of education, or other unobserved skills that make a worker more or less valuable relative to others in the same occupation.

To relate these findings to the increase in between-firm inequalities documented in Section 3, we abstract from the role of covariates and express the variance of firm-average wages as:

$$\begin{aligned} V(\bar{y}) &\approx V(\bar{\omega} + \bar{\varepsilon}) + 2 * Cov(\omega_{+}\varepsilon, \psi) = \\ &V(\bar{\omega}) + V(\bar{\varepsilon}) + 2 * Cov(\bar{\omega}, \bar{\varepsilon}) + \\ &2 * Cov(\omega, \psi) + 2 * Cov(\varepsilon, \psi) \end{aligned} \tag{6}$$

We analyze the changes in each component between the first and last periods. $\Delta(\bar{\omega})$ captures changes in the variance of the firm-average occupation-specific components, which could result from changes in the occupational composition across firms or from changes in the occupation-specific components themselves. $\Delta V(\bar{\varepsilon})$ measures changes

[22] We use the two-digit level of *catégories socioprofessionnelles*, a French statistical nomenclature that can be further explored in more detail here.

[23] We use estimates corrected by firm splitting. For more details on how the parameters are recovered, see Section D.2.

[24] Within a six-year period, we first retrieve ω_o as the average worker fixed effect in occupation o over the years. We then calculate $\varepsilon_{i,t}$ by subtracting θ_i and ω_o each year. Thus, $\varepsilon_{i,t}$ varies within a worker if he or she changes occupation at some point within the six-year period.

in the extent to which workers with similar individual-specific components cluster together. $\Delta Cov(\bar{\omega}, \bar{\varepsilon})$ captures changes in the relationship between the occupational composition of a firm and the individual-specific components of its workers. The last two terms, $\Delta Cov(\omega, \psi)$ and $\Delta Cov(\varepsilon, \psi)$, represent the change in the sorting of workers into firms based on the occupation-specific and individual-specific components of the worker fixed effects, respectively. Table 4 presents this decomposition.

TABLE 4. Decomposition of changes in between-firm wage variance

Period	$V(\bar{\omega})$	$V(\bar{\varepsilon})$	$2Cov(\bar{\omega}, \bar{\varepsilon})$	$2Cov(\omega, \psi)$	$2Cov(\varepsilon, \psi)$
2002-2007	0.0304	0.0091	0.0048	0.0211	0.0055
2008-2013	0.0313	0.0108	0.0081	0.0184	0.0085
2014-2019	0.0358	0.0131	0.0092	0.0186	0.0101
Diff (2014-19 vs 2002-07)	0.0054	0.0039	0.0044	-0.0024	0.0046

Notes: This table decomposes the changes in between-firm wage variance based on Equation 6. ω represents the occupation-specific component of the worker fixed effect, while ε is the individual-specific component. $\bar{\omega}$ and $\bar{\varepsilon}$ are firm-level averages of these two components. ψ represents the firm fixed effect. All estimates are corrected by the split-sampling method with firm splitting.

The first component, $V(\bar{\omega})$, increased by 0.0054 between 2002-2007 and 2014-2019, indicating an increase in the variance of the firm-average occupation-specific components. It's important to note, however, that this increase alone does not necessarily imply greater occupational specialization. The increase in $V(\bar{\omega})$ could simply reflect an increase in the overall dispersion of occupation-specific wage components $V(\omega)$ in the labor market, without significant changes in the occupational composition within and across firms. We will return to this point later.

We also observe an increase in $V(\bar{\varepsilon})$, indicating increasing segregation based on individual-specific components. This means that workers with similar individual-specific components within their occupation are increasingly likely to work together. The increase in 2Cov$(\bar{\omega}, \bar{\varepsilon})$ suggests a strengthening relationship between a firm's composition of occupation-specific components and its workers' individual-specific components. In other words, firms with a higher proportion of occupations associated with high occupation-specific components are increasingly likely to employ individuals with higher individual-specific components within these occupations. Regarding the sorting channel, we observe different trends for occupation-specific and individual-specific components. The covariance between firm fixed effects and occupation-specific components, 2Cov(ω, ψ), decreased slightly by 0.0024. In contrast, the covariance between the firm fixed effects and the individual-specific components, $2\text{Cov}(\varepsilon, \psi)$, increased by 0.0046. This suggests that while the sorting based on occupations has slightly weakened, the sorting based on individual-specific components has become more pronounced.

Unpacking Occupational Dynamics. While our analysis of between-firm wage variance components provides valuable insights into segregation and sorting patterns, it raises important questions about the nature of changes in the occupation-specific component itself. Specifically, we observed an increase in $V(\bar{\omega})$. However, this increase alone does not necessarily imply increased occupational specialization within and across firms without knowing how occupation-specific average labor market values have evolved. To this end, we now turn our attention to a more detailed analysis of ω. We first examine how much of the worker fixed effect is explained by occupation and how this has changed over time. We then decompose changes in $V(\omega)$ to distinguish between the effects of shifts in occupational composition and changes in occupation-specific average market values.

First, we observe that $V(\omega)$ decreased by 0.0066 between 2002-2007 and 2014-2019 (Table A9), with the proportion of the worker fixed effect explained by occupation (measured by $\frac{V(\omega)}{V(\theta)}$) consequently decreasing from 62% to 55%. The simultaneous decrease in $V(\omega)$ and increase in $V(\bar{\omega})$ implies that $V(\omega - \bar{\omega})$ (the within-firm component of occupational wage effects) has decreased significantly (-0.0120). This suggests that occupational specialization within firms has indeed increased over time. In other words, while the overall dispersion of occupation-specific market values has narrowed, the differences between firms in terms of their occupational composition have become more pronounced. Firms have become more specialized in terms of the occupation-specific components they employ, tending toward greater homogeneity within firms and heterogeneity between firms.

To better understand the factors that shape the occupation-specific component of the worker fixed effect and to gain more insight into trends for specific occupations, we construct an occupation-level dataset. After normalizing ω to have a period-specific population mean of zero, we can express its variance as:

$$V(\omega_{o,t}) = \sum_{o} p_{o,t} \cdot \omega_{o,t}^2 \tag{7}$$

where $p_{o,t}$ is the employment share in occupation o for the six-year period t. Changes in $V(\omega)$ across periods can result from shifts in worker distribution across occupations (hereafter, $\Delta p(\omega)$) or changes in occupation-specific ω ($\Delta\omega(p)$). We analyze these changes using a Kitagawa-Oaxaca-Blinder decomposition (Kitagawa 1955; Oaxaca 1973; Blinder 1973):

$$\Delta V(\omega_o) = \sum_o \underbrace{(p_{o,t+1} - p_{o,t}) \cdot \omega^2_{o,t+1}}_{\Delta p(\omega)} + \sum_o \underbrace{p_{o,t} \cdot (\omega^2_{o,t+1} - \omega^2_{o,t})}_{\Delta\omega(p)} \tag{8}$$

We report the average of the two possible formulations of this decomposition[25], along with occupation-period-specific ω and p values in Table 5.

[25] In the alternative formulation, we use p at time $t+1$ to hold p constant and ω^2 at time t to hold ω^2 constant.

TABLE 5. Occupational structure, average worker fixed effect, and decomposition of changes

		2002-2007		2014-2019		Decomposition	
Occupation	CS	ω	p	ω	p	$\Delta\, p(\omega)$	$\Delta\omega(p)$
CEOs	23	1.11	0.65%	0.94	0.63%	-0.0002	***-0.0022***
Professionals	31	0.52	0.16%	0.37	0.24%	0.0002	-0.0003
Public administration managers	33	0.58	0.08%	0.38	0.14%	0.0002	-0.0002
Scientific professions	34	0.71	0.69%	0.47	1.32%	***0.0023***	***-0.0028***
Artists and media professionals	35	0.34	0.62%	0.32	0.43%	-0.0002	-0.0001
Managers	37	0.61	8.18%	0.50	10.97%	**0.0088**	**-0.0114**
Engineers	38	0.53	7.04%	0.44	9.61%	**0.0061**	**-0.0067**
School teachers	42	0.09	0.88%	0.03	1.35%	0.0000	-0.0001
Health and social workers	43	0.09	3.16%	0.00	3.93%	0.0000	-0.0003
Public administration intermediates	45	0.04	0.09%	0.07	0.05%	0.0000	0.0000
Business administration intermediates	46	0.08	12.33%	0.04	8.47%	-0.0001	-0.0004
Technicians	47	0.05	5.81%	0.02	6.25%	0.0000	-0.0001
Intermediate supervisors	48	0.12	3.38%	0.07	2.84%	-0.0001	-0.0003
Public administration clerks	52	-0.24	2.76%	-0.28	3.33%	0.0004	*0.0006*
Security agents	53	-0.23	0.86%	-0.24	1.00%	0.0001	0.0001
Business administration clerks	54	-0.15	10.88%	-0.13	10.60%	-0.0001	*-0.0007*
Retail salespersons	55	-0.29	5.92%	-0.31	6.83%	*0.0008*	*0.0007*
Personal service employees	56	-0.31	3.56%	-0.32	4.49%	*0.0009*	0.0003
Skilled manufacturing workers	62	-0.15	11.47%	-0.15	8.83%	*-0.0006*	0.0000
Skilled artisans	63	-0.17	5.08%	-0.20	4.63%	-0.0002	0.0005
Drivers	64	-0.20	4.26%	-0.22	4.13%	-0.0001	0.0004
Handling, transport skilled workers	65	-0.23	2.74%	-0.22	2.65%	0.0000	0.0000
Unskilled manufacturing workers	67	-0.30	6.17%	-0.29	3.97%	***-0.0019***	-0.0002
Unskilled artisans	68	-0.34	2.88%	-0.35	2.76%	-0.0001	0.0002
Farm workers	69	-0.28	0.25%	-0.25	0.38%	0.0001	-0.0001
Total						0.0162	-0.0229

Notes: ω represents the average worker fixed effect for each occupation, and p represents the proportion of workers in each occupation. ω is normalized so that its period-specific population average is zero. CS represents the two-digit code for *catégories socioprofessionnelles*, which can be further explored here. $\Delta\, p(\omega)$ represents the change in variance due to shifts in the distribution of workers across occupations, while $\Delta\omega(p)$ represents the change due to variations in occupation-specific worker fixed effects (see Equation 8). Values in the decomposition columns are formatted with different levels of emphasis to highlight their magnitude. The total row shows the sum of each component of the decomposition across all occupations. All estimates are corrected by the split-sampling method with firm splitting

Changes due to shifts in the distribution of workers ($\Delta p(\omega)$) show an overall positive effect (0.0162), indicating that changes in occupational shares have generally increased variance. This is mainly due to the relative expansion of occupations with high occupation-specific components, such as managers (0.0088) and engineers (0.0061). However, this increase is more than offset by changes in the occupation-specific components ($\Delta\omega(p)$), which show a larger negative effect (-0.0229). This negative effect is largely due to decreases in the occupation-specific components of the same occupations: managers (-0.0114) and engineers (-0.0067). The net result is a decrease in the overall variance of ω, suggesting a complex dynamic. While the share of occupations typically associated with high-skilled workers is increasing, there's a simultaneous decrease in the occupation-specific components within those occupations. Importantly, the variance of ε (the individual-specific component) within these high-skilled occupations has remained stable over time (Table A10). This suggests that the observed decline in occupation-specific components is not due to increased polarization of individual-specific components within occupations. Rather, it suggests a broader shift in the composition of workers in these occupations, possibly indicating a form of "skill dilution" or changes in occupation-specific tasks and responsibilities. We prefer the first explanation because it is consistent with recent research. In their study of the German labor market from 1985 to 2010, Böhm, von Gaudecker, and Schran (2024) found that growing occupations tend to attract relatively less-skilled workers with lower levels of occupation-specific human capital and therefore lower market value. They document that this phenomenon is particularly evident in high-skilled occupations such as managers and professionals.

Taking Stock. In sum, we observe increased segregation based on both occupation-specific and individual-specific components of worker fixed effects, as well as a stronger association between these components at the firm level. The increase in between-firm wage variance is primarily driven by increased occupational segregation, stronger clustering of workers with similar individual-specific components, and a tighter relationship between firms' occupational composition and their workers' idiosyncratic characteristics. These findings suggest a nuanced evolution in the labor market, where firm-level occupational specialization is increasing, and within-occupation worker sorting across firms is becoming more pronounced based on unobserved, individual-specific components. These trends are accompanied by complex dynamics in the occupation-specific components themselves. While the share of occupations typically associated with high-skilled workers is increasing, we observe a simultaneous decline in their average market value.

4.2. Potential Alternative Channels

The economic literature has dealt extensively with skill-biased technological change (Acemoglu and Autor 2011). An increase in the skill premium could explain the increase in sorting and segregation we document. Indeed, the worker fixed effects dispersion Var(θ) diverges significantly over the study period (Appendix A). This may be due to either increasing returns to skill, changes in the distribution of worker skills, or both. Our analysis shows a decline in the average market value of high-skilled occupations despite their growing share of employment. At first glance, this might suggest stable or even declining returns to skill. However, Böhm, von Gaudecker, and Schran (2024) argue that this could be due to a "marginal selection effect", where growing occupations attract relatively less-skilled workers, potentially masking rising skill prices. By developing a model that disentangles changes in skill prices from changes in the composition of workers within occupations, they show rising skill prices in high-wage, growing occupations. Thus, we cannot rule out rising returns to skill in our context, despite the decline in average occupation-specific wage components of high-skill job types. To assess the impact of changing skill prices on segregation and sorting, we follow the approach of Song et al. (2019). Assuming a fixed distribution of worker skills in the economy over our relatively short panel, we calculate the change in returns to skills r between period 1 (2002-2007) and period 3 (2014-2019) as:

$$\frac{r^3}{r^1} = \sqrt{\frac{\mathrm{V}(\theta_i^3)}{\mathrm{V}(\theta_i^1)}} = \sqrt{\frac{0.161}{0.152}} \approx 1.03, \tag{9}$$

where the values for the variances are taken from Table A5. Consequently, the increases in segregation and sorting due solely to changes in skill prices can be expressed as:

$$\frac{\mathrm{V}(\bar{\theta}^3)}{\mathrm{V}(\bar{\theta}^1)} = \left(\frac{r^3}{r^1}\right)^2 \approx 1.06 \tag{10}$$

$$\frac{2\mathrm{Cov}(\bar{\theta}^3, \psi^3)}{2\mathrm{Cov}(\bar{\theta}^1, \psi^1)} = \frac{r^3}{r^1} \approx 1.03 \tag{11}$$

Segregation and sorting increased by 33 and 8%, respectively. The mechanical effect of changing returns to skill can thus account for about 18% of the increase in segregation and 38% of the increase in sorting, or about one-fifth of the total increase in between-firm inequality, suggesting an overall modest role.

The variance of the firm wage premiums has declined slightly[26]. This decline in premium variance is consistent with an "eclipse of rent sharing" as recently documented by Acemoglu, He, and le Maire (2022) for Denmark and the US. To investigate this trend further, we analyze the firm-level, employment-weighted relationships between value added per worker and firm fixed effects across different periods, as illustrated in Figure A2. Our results show that for a given level of period-demeaned productivity, the distribution of shared rents in the most recent period shows a slight increase at the bottom relative to the previous period, while converging at the top. This suggests that the overall decrease in the variance of firm effects is due to an increase in rents at the lower end of the firm rent distribution. However, the slopes of the period fitting lines are visually very similar and statistically indistinguishable. We conclude that the decrease in the variance of firm effects and the relatively small increase in rents at the bottom of the distribution do not contribute significantly to explaining the observed increase in between-firm inequalities. On the contrary, their influence may be in the opposite direction.

Changes in the distribution of firm size could potentially have a significant impact on the results for between-firm wage inequality and occupational segregation. The theoretical effect of firm size is ambiguous. On the one hand, if there had been a substantial shift towards larger firms over time, we might expect to see less occupational segregation, as larger firms often have more diverse occupational structures. On the other hand, larger firms might also contribute to increased between-firm wage inequality through higher wage premiums, thus playing a more important role in worker sorting processes. However, when we examine the cumulative firm size distribution shown in Figure A3, we find that there has been remarkably little change in the firm size distribution between the 2002-2007 and 2014-2019 periods. Even if we look at the very top of the firm size distribution, we find that the average firm size of firms with more than 1000 employees has remained stable, with only a slight decline (Table 3). This stability suggests that the changes observed in our main results are not driven by shifts in the firm size landscape.

[26] This decline is more pronounced without bias correction, supporting the idea that the bias has declined over time.

5. Discussion

This paper introduces a simple adaptation of the AKM model to analyze French wage inequalities, addressing the limited mobility bias through split-sampling correction. Our results show that this bias affects not only the measurement of sorting but also its evolution over time, emphasizing the importance of corrected estimates for accurately capturing sorting dynamics. Despite stable overall wage inequality in France from 2002 to 2019, we observe increases in both sorting and segregation, mirroring trends in the US and Germany. However, our corrected estimates show that segregation plays a more prominent role than sorting in driving these changes. Moreover, French segregation manifests itself in both increasing divergence of worker fixed effects between firms and increasing homogeneity within firms.

Our estimation method has two main limitations. First, it assumes independence in the mobility structure between splits, and that the reduction in the size of the main connected set after the split doesn't significantly affect the results. Second, like most AKM models, we rely on the assumption of exogenous mobility. However, we believe that these problems are likely to be minor. The robustness of our results to alternative methods supports the validity of our splitting assumptions, while previous studies have validated the exogenous mobility assumption in similar contexts.

The French case exemplifies generalized assortative matching, with high-wage workers increasingly segregating from low-wage workers, working with similar others, and concentrating, but at a slower pace, in high-wage firms. This evolution is primarily driven by the reconfiguration of occupational combinations within and between firms, rather than by changes in the returns to skill or the rent sharing behavior of firms. Rather than a radical change in education, formal and informal skills, our findings point to an evolution in the complex responses of firms to multiple pressures and changes in the business environment. Technological advancements, particularly digitalization and improved information technologies, have facilitated remote work coordination and reduced the need for traditional hierarchical structures within firms (Bilal and Lhuillier 2021). At the same time, financial pressures, especially from analysts and shareholders, have pushed companies to focus on core activities and simplify their structures for easier monitoring (Zuckerman 2004). The rise of a new generation of managers, educated in the shareholder value paradigm, has further accelerated these changes (Acemoglu, He, and le Maire 2022; Jung and Shin 2019). These factors have collectively contributed to various forms of "fissuring" in the workplace, including outsourcing, subcontracting, franchising, and subsidiarization. Such practices have been shown to promote occupational and earnings segregation (Weil 2014; Godechot et al. 2024). The resulting reconfiguration of work organization has led to increased separation of different occupational groups across firms, while at the same time increasing homogeneity within firms. As firms are also sites of socialization and identity formation, this increasing segregation may have broader implications for social cohesion (Chetty et al. 2022).

References

Abowd, John M, Francis Kramarz, Paul Lengermann, and Sébastien Pérez-Duarte. 2004. "Are good workers employed by good firms? A test of a simple assortative matching model for France and the United States." Unpublished Manuscript.

Abowd, John M, Francis Kramarz, and David N Margolis. 1999. "High wage workers and high wage firms." *Econometrica* 67 (2): 251–333.

Abowd, John M, Francis Kramarz, and Sebastien Roux. 2006. "Wages, mobility and firm performance: Advantages and insights from using matched worker–firm data." *The Economic Journal* 116 (512): F245–F285.

Acemoglu, Daron, and David Autor. 2011. "Skills, Tasks and Technologies: Implications for Employment and Earnings." *Handbook of Labor Economics* 4: 1043–1171.

Acemoglu, Daron, Alex He, and Daniel le Maire. 2022. "Eclipse of Rent-Sharing: The Effects of Managers' Business Education on Wages and the Labor Share in the US and Denmark." NBER Working Paper No. 29874.

Andrews, Martyn J, Len Gill, Thorsten Schank, and Richard Upward. 2008. "High wage workers and low wage firms: negative assortative matching or limited mobility bias?" *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171 (3): 673–697.

Barth, Erling, Alex Bryson, James C Davis, and Richard Freeman. 2016. "It's where you work: Increases in the dispersion of earnings across establishments and individuals in the United States." *Journal of Labor Economics* 34 (S2): S67–S97.

Bergé, Laurent. 2018. "Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm." Crea Discussion Paper No. 13.

Bergeaud, Antonin, Clément Malgouyres, Clément Mazet-Sonilhac, and Sara Signorelli. 2021. "Technological Change and Domestic Outsourcing." IZA Discussion Paper No. 14603.

Bilal, Adrien, and Hugo Lhuillier. 2021. "Outsourcing, inequality and aggregate output." NBER Working Paper No. 29348.

Blinder, Alan S. 1973. "Wage discrimination: reduced form and structural estimates." *Journal of Human resources*: 436–455.

Böhm, Michael J., Hans-Martin von Gaudecker, and Felix Schran. 2024. "Occupation Growth, Skill Prices, and Wage Inequality." *Journal of Labor Economics* 42 (1): 201–243.

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa. 2019. "A distributional framework for matched employer employee data." *Econometrica* 87 (3): 699–739.

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa. 2022. "Discretizing unobserved heterogeneity." *Econometrica* 90 (2): 625–643.

Bonhomme, Stéphane, Kerstin Holzheu, Thibaut Lamadon, Elena Manresa, Magne Mogstad, and Bradley Setzler. 2023. "How Much Should We Trust Estimates of Firm Effects and Worker Sorting?." *Journal of Labor Economics* 41 (2): 291–322.

Borovičková, Katarína, and Robert Shimer. 2017. "High wage workers work for high wage firms." NBER Working Paper No. 24074.

Card, David, Ana Rute Cardoso, Joerg Heining, and Patrick Kline. 2018. "Firms and labor market inequality: Evidence and some theory." *Journal of Labor Economics* 36 (S1): S13–S70.

Card, David, Jörg Heining, and Patrick Kline. 2013. "Workplace heterogeneity and the rise of West German wage inequality." *The Quarterly journal of economics* 128 (3): 967–1015.

Chanut, Nicolas. 2018. "Distinguishing Between Signal and Noise in the Measurement of the Firm Wage Premium." Working Paper.

Chetty, Raj, Matthew O. Jackson, Theresa Kuchler, Johannes Stroebel, Nathaniel Hendren, Robert B. Fluegge, Sara Gong, Augustine Sun, and William Wood. 2022. "Social Capital I: Measurement and Associations with Economic Mobility." *Nature* 608 (7921): 108–121.

Correia, Sergio. 2016. "Linear Models with High-Dimensional Fixed Effects: An Efficient and Feasible Estimator." Working Paper.

Coudin, Elise, Sophie Maillard, and Maxime Tô. 2018. "Family, firms and the gender wage gap in France." IFS Working Papers W18/01.

Davis, Steve J., and John Haltiwanger. 1991. "Wage Dispersion between and within U.S. Manufacturing Plants, 1963-86." *Brookings Papers on Economic Activity* 22 (1991 Micr): 115–200.

Di Addario, Sabrina L, Patrick M Kline, Raffaele Saggio, and Mikkel Sølvsten. 2023. "It Ain't Where You're From, It's Where You're At: Hiring Origins, Firm Heterogeneity, and Wages.." *Journal of Econometrics* 233 (2): 340–374.

Dorn, David, Johannes Schmieder, and James Spletzer. 2018. "Domestic Outsourcing in the United States." Working Paper.

Drenik, Andres, Simon Jäger, Miguel Pascuel Plotkin, and Benjamin Schoefer. 2023. "Paying Outsourced Labor: Direct Evidence from Linked Temp Agency-Worker-Client Data." *The Review of Economics and Statistics* 105 (1): 206–216.

Engbom, Niklas, Christian Moser, and Jan Sauermann. 2023. "Firm pay dynamics." *Journal of Econometrics* 233 (2): 396–423.

Frederiksen, Anders, Lisa B Kahn, and Fabian Lange. 2020. "Supervisors and performance management systems." *Journal of Political Economy* 128 (6): 2123–2187.

Gaure, Simen. 2013. "lfe: Linear group fixed effects." *The R Journal* 5 (2): 104–117.

Gerard, François, Lorenzo Lagos, Edson Severnini, and David Card. 2021. "Assortative Matching or Exclusionary Hiring? The Impact of Employment and Pay Policies on Racial Wage Differences in Brazil." *American Economic Review* 111 (10): 3418–57.

Godechot, Olivier, Paula Apascaritei, István Boza, Lasse Folke Henriksen, Are Skeie Hermansen, Feng Hou, Naomi Kodama, Alena Křížková, Jiwook Jung, Marta M Elvira et al. 2020. "The great separation: Top earner segregation at work in high-income countries." MaxPo Discussion Paper.

Godechot, Olivier, Mirna Safi, and Matthew Soener. 2021. "The Intersection of Organizational Inequalities: How Gender, Migrant Status, and Class Inequality Relate to Each Other in French Workplaces." SciencesPo OSC Papers.

Godechot, Olivier, Donald Tomaskovic-Devey, István Boza, Lasse Henriksen, Are Skeie Hermansen, Feng Hou, Jiwook Jung, Naomi Kodama, Alena Křížková, Zoltán Lippényi, Silvia Maja Melzer, Eunmi Mun, Halil Sabanci, Max Thaning et al. 2024. "The great separation: Top Earner Segregation at Work in Advanced Capitalist Economies." *American Journal of Sociology* 130 (2): 439–495.

Goldschmidt, Deborah, and Johannes F Schmieder. 2017. "The rise of domestic outsourcing and the evolution of the German wage structure." *The Quarterly Journal of Economics* 132 (3): 1165–1217.

Jochmans, Koen, and Martin Weidner. 2019. "Fixed-Effect Regressions on Network Data." *Econometrica* 87 (5): 1543–1560.

Jung, Jiwook, and Taekjin Shin. 2019. "Learning not to diversify: The transformation of graduate business education and the decline of diversifying acquisitions." *Administrative Science Quarterly* 64 (2): 337–369.

Kitagawa, Evelyn M. 1955. "Components of a difference between two rates." *Journal of the american statistical association* 50 (272): 1168–1194.

Kline, Patrick, Raffaele Saggio, and Mikkel Sølvsten. 2020. "Leave-out estimation of variance components." *Econometrica* 88 (5): 1859–1898.

Lachowska, Marta, Alexandre Mas, Raffaele Saggio, and Stephen A. Woodbury. 2023. "Do firm effects drift? Evidence from Washington administrative data." *Journal of Econometrics* 233 (2): 375–395.

Oaxaca, Ronald. 1973. "Male-female wage differentials in urban labor markets." *International economic review*: 693–709.

OECD. 2021. *The Role of Firms in Wage Inequality: Policy Lessons from a Large Scale Cross-Country Study*. Éditions OCDE, Paris.

Postel-Vinay, Fabien, and Jean-Marc Robin. 2002. "Equilibrium wage dispersion with worker and employer heterogeneity." *Econometrica* 70 (6): 2295–2350.

Schoefer, Benjamin, and Oren Ziv. 2022. "Productivity, Place, and Plants." *The Review of Economics and Statistics*: 1–46.

Song, Jae, David J Price, Fatih Guvenen, Nicholas Bloom, and Till Von Wachter. 2019. "Firming up inequality." *The Quarterly journal of economics* 134 (1): 1–50.

Sorkin, Isaac. 2018. "Ranking Firms Using Revealed Preference." *The Quarterly Journal of Economics* 133 (3): 1331–1393.

Tomaskovic-Devey, Donald, Anthony Rainey, Dustin Avent-Holt, Nina Bandelj, István Boza, David Cort, Olivier Godechot, Gergely Hajdu, Martin Hällsten, Lasse Folke Henriksen et al. 2020. "Rising between-workplace inequalities in high-income countries." *Proceedings of the National Academy of Sciences* 117 (17): 9277–9283.

Weil, David. 2014. *The Fissured Workplace: Why work became so bad for so many and what can be done to improve it*. Harvard University Press, Cambridge MA.

Zuckerman, Ezra W. 2004. "Structural incoherence and stock market activity." *American Sociological Review* 69 (3): 405–432.

Appendix A. Appendix Tables

TABLE A1. Summary statistics - detailed

		2002-2007			2008-2013			2014-2019		
		Universe	Sample 1	Sample 2	Universe	Sample 1	Sample 2	Universe	Sample 1	Sample 2
	Firm size group									
	<20	19.09%	12.04%	5.06%	20.62%	13.48%	5.99%	20.05%	12.43%	5.29%
	20-200	30.47%	32.72%	32.65%	30.20%	33.07%	33.85%	28.30%	31.23%	31.72%
	200-1000	20.23%	22.09%	24.74%	19.74%	21.83%	24.46%	20.24%	22.64%	25.43%
	>1000	30.41%	33.36%	37.38%	29.29%	31.81%	35.69%	31.63%	33.62%	37.80%
	Occupation									
23	CEOs	0.70%	0.69%	0.61%	0.89%	0.79%	0.67%	0.84%	0.69%	0.56%
31	Professionals	0.15%	0.15%	0.16%	0.19%	0.20%	0.20%	0.23%	0.24%	0.24%
35	Artists and media professionals	0.69%	0.65%	0.60%	0.58%	0.55%	0.51%	0.47%	0.45%	0.41%
37	Managers	7.70%	8.01%	8.33%	8.79%	9.15%	9.50%	10.12%	10.71%	11.20%
38	Engineers	6.35%	6.78%	7.29%	6.91%	7.47%	8.06%	8.44%	9.24%	9.96%
42	Primary school teachers	0.95%	0.92%	0.84%	1.34%	1.35%	1.27%	1.40%	1.39%	1.31%
43	Health and social workers	3.08%	3.16%	3.16%	3.48%	3.55%	3.53%	3.88%	3.97%	3.93%
46	Business administration intermediates	12.04%	12.27%	12.37%	8.42%	8.82%	9.11%	7.78%	8.28%	8.64%
47	Technicians	5.36%	5.65%	5.96%	5.66%	6.01%	6.32%	5.65%	6.08%	6.40%
48	Intermediate supervisors	3.22%	3.33%	3.44%	2.55%	2.70%	2.82%	2.57%	2.76%	2.91%
52	Public administration clerks	2.66%	2.75%	2.78%	2.98%	3.01%	3.12%	3.34%	3.28%	3.39%
53	Security agents	0.77%	0.82%	0.89%	1.00%	1.01%	1.09%	0.87%	0.96%	1.04%
54	Business administration clerks	11.37%	11.05%	10.73%	12.04%	11.78%	11.38%	10.95%	10.78%	10.40%
55	Retail salespersons	6.11%	5.96%	5.87%	6.79%	6.76%	6.66%	6.83%	6.87%	6.79%
56	Personal service employees	4.33%	3.79%	3.35%	5.09%	4.59%	4.04%	5.53%	4.86%	4.18%
62	Skilled manufacturing workers	10.62%	11.19%	11.74%	8.61%	9.11%	9.49%	8.03%	8.63%	9.03%
63	Skilled artisans	6.42%	5.63%	4.58%	6.59%	5.89%	4.85%	5.97%	5.21%	4.12%
64	Drivers	4.14%	4.23%	4.27%	4.26%	4.33%	4.39%	4.19%	4.07%	4.18%
65	Handling, transport skilled workers	2.55%	2.66%	2.81%	2.44%	2.59%	2.75%	2.37%	2.56%	2.73%
67	Unskilled manufacturing workers	5.80%	6.06%	6.26%	5.02%	5.24%	5.34%	3.70%	3.93%	4.02%
68	Unskilled artisans	3.40%	2.94%	2.80%	3.52%	3.19%	3.08%	3.04%	2.80%	2.72%
69	Farm workers	0.56%	0.31%	0.20%	0.72%	0.37%	0.21%	0.94%	0.51%	0.28%
	Industry									
AC	Farming and industry	24.86%	25.66%	26.63%	20.28%	21.10%	21.70%	18.96%	19.96%	20.78%
DE	Utilities	2.09%	2.16%	2.46%	2.12%	2.30%	2.53%	2.18%	2.41%	2.65%
F	Construction	7.41%	6.83%	5.85%	8.13%	7.57%	6.61%	7.47%	6.79%	5.90%
G	Commerce	18.25%	17.70%	17.07%	17.62%	17.36%	16.96%	17.56%	17.48%	17.18%
H	Transport	5.33%	5.67%	6.00%	6.13%	5.92%	6.29%	7.14%	5.67%	6.40%
I	Hotels, tourism, catering	3.58%	3.23%	2.80%	3.79%	3.46%	2.93%	3.99%	3.55%	2.92%
J	Media	4.47%	4.70%	5.05%	4.38%	4.69%	5.00%	4.54%	4.93%	5.22%
K	Financial services	6.38%	6.81%	7.33%	5.28%	5.60%	5.99%	5.39%	5.71%	6.10%
LM	Real estate, professional services	7.28%	6.73%	6.16%	7.36%	6.95%	6.45%	7.87%	7.55%	7.21%
N	Administrative services	4.51%	4.00%	4.15%	6.00%	5.62%	5.67%	5.88%	5.63%	5.67%
OPQ	Health, education	11.23%	11.85%	12.18%	12.54%	13.55%	14.12%	13.12%	14.24%	14.86%
R	Arts and recreation	0.59%	0.53%	0.48%	1.09%	0.94%	0.80%	1.15%	0.98%	0.83%
STU	Other	2.24%	1.85%	1.41%	2.41%	2.03%	1.62%	2.22%	1.77%	1.38%

Note: This table presents the workforce composition across different categories for the three time periods. 'Universe' represents the starting sample after applying the restrictions outlined in Section 1.2, while 'Sample 1' and 'Sample 2' refer to firms in the largest connected set and firms in both connected sets, respectively. The largest connected set entails the group of firms connected by worker mobility. Firms in both connected sets refer to firms present in both main connected components in each split sample for the "firm splitting" method (Section 2.3.1). We use the two-digit level of *catégories socioprofessionnelles,* a French statistical nomenclature for classifying occupations which can be explored in more detail here.

TABLE A2. Decomposition of wage variance and its evolution

Firms with 1+ employees (uncorrected AKM)

		2002-2007		2008-2013		2014-2019		Change from 2002-2007 to 2014-2019
		Comp.	Share	Comp.	Share	Comp.	Share	Diff.
Total variance	$\text{Var}(y)$	0.212		0.205		0.217		0.005
	$\text{Var}\ (\theta)$	0.166	78.2	0.159	77.2	0.170	78.5	0.004
	$\text{Var}\ (\psi)$	0.028	13.2	0.024	11.8	0.023	10.5	-0.005
	$\text{Var}(Xb)$	0.003	1.5	0.002	1.2	0.003	1.2	0.000
	$\text{Var}(u)$	0.010	4.6	0.010	4.7	0.010	4.7	0.000
	$2^*\text{Cov}(\theta,\psi)$	0.003	1.3	0.009	4.2	0.012	5.3	0.009
	$2^*\text{Cov}(\theta,Xb)$	0.000	0.0	0.000	0.1	-0.002	-0.8	-0.002
	$2^*\text{Cov}(\psi,Xb)$	0.001	0.3	0.001	0.4	0.001	0.3	0.000
Between-firm variance	$\text{Var}(\bar{y})$	0.089	41.8	0.094	45.9	0.104	47.8	0.015
	$\text{Var}\ (\bar{\theta})$	0.056	26.4	0.059	28.8	0.067	30.8	0.011
	$\text{Var}\ (\psi)$	0.028	13.2	0.024	11.8	0.023	10.5	-0.005
	$\text{Var}(\bar{Xb})$	0.000	0.1	0.000	0.1	0.000	0.1	0.000
	$2^*\text{Cov}(\bar{\theta},\psi)$	0.003	1.3	0.009	4.2	0.012	5.3	0.009
	$2^*\text{Cov}(\bar{\theta},\bar{Xb})$	0.001	0.5	0.001	0.6	0.001	0.7	0.000
	$2^*\text{Cov}(\psi,\bar{Xb})$	0.001	0.3	0.001	0.4	0.001	0.3	0.000
Within-firm variance	$\text{Var}(y-\bar{y})$	0.123	58.2	0.111	54.1	0.113	52.2	-0.010
	$\text{Var}\ (\theta-\bar{\theta})$	0.110	51.8	0.103	48.5	0.103	47.7	-0.007
	$\text{Var}(Xb-\bar{Xb})$	0.003	1.3	0.002	1.1	0.002	1.1	0.000
	$\text{Var}(u)$	0.010	4.6	0.010	4.7	0.010	4.7	0.000
	$2^*\text{Cov}(\theta-\bar{\theta},Xb-\bar{Xb})$	-0.001	-0.5	-0.001	-0.5	-0.003	-1.5	-0.002
	$2^*\text{Cov}(\theta-\bar{\theta},\ u)$	0.000	0.0	0.000	0.0	0.000	0.0	0.000
	$2^*\text{Cov}(Xb-\bar{Xb},\ u)$	0.000	0.0	0.000	0.0	0.000	0.0	0.000
Segregation Index	$\frac{Var(\bar{\theta}_j)}{Var(\theta_i)}$	0.338		0.373		0.393		0.055
N		58,666,316		61,413,372		59,550,288		

Notes: This table presents the decomposition of wage variance and its evolution over three periods using the standard AKM model. "Comp." denotes the component of variance, while "Share"indicates the percentage of total $Var(y)$. The last column show the change in levels from 2002-2007 to 2014-2019. The decomposition is based on Equations 2, 3, and 4. The number of observations refer to the largest connected set, i.e. the group of firms connected by worker mobility.

TABLE A3. Decomposition of wage variance and its evolution

Firms with 20+ employees (uncorrected AKM)

		2002-2007		2008-2013		2014-2019		Change from 2002-2007 to 2014-2019
		Comp.	Share	Comp.	Share	Comp.	Share	Diff.
Total variance	Var(y)	0.211		0.206		0.218		0.007
	Var (θ)	0.163	77.2	0.157	76.1	0.169	77.6	0.006
	Var (ψ)	0.021	9.9	0.018	8.9	0.017	7.8	-0.004
	Var(Xb)	0.003	1.4	0.002	1.1	0.002	1.1	-0.001
	Var(u)	0.010	4.5	0.009	4.6	0.010	4.6	0.001
	2*Cov(θ,ψ)	0.013	6.0	0.017	8.3	0.019	8.9	0.007
	2*Cov(θ,Xb)	0.000	-0.1	0.000	0.2	-0.001	-0.7	-0.001
	2*Cov(ψ,Xb)	0.001	0.3	0.001	0.3	0.001	0.3	0.000
Between-firm variance	Var($\bar{y}$)	0.086	40.6	0.092	44.5	0.101	46.3	0.015
	Var ($\bar{\theta}$)	0.050	23.9	0.054	26.3	0.062	28.5	0.012
	Var (ψ)	0.021	9.9	0.018	8.9	0.017	7.8	-0.004
	Var($\bar{X}b$)	0.000	0.1	0.000	0.1	0.000	0.1	0.000
	2*Cov($\bar{\theta}$,ψ)	0.013	6.0	0.017	8.3	0.019	8.9	0.007
	2*Cov($\bar{\theta}$,$\bar{X}b$)	0.001	0.5	0.001	0.6	0.001	0.7	0.000
	2*Cov(ψ,$\bar{X}b$)	0.001	0.3	0.001	0.3	0.001	0.3	0.000
Within-firm variance	Var($y - \bar{y}$)	0.125	59.4	0.114	55.5	0.117	53.7	-0.008
	Var ($\theta - \bar{\theta}$)	0.113	53.3	0.107	49.8	0.107	49.1	-0.005
	Var($Xb - \bar{X}b$)	0.003	1.3	0.002	1.0	0.002	1.0	-0.001
	Var(u)	0.010	4.5	0.009	4.6	0.010	4.6	0.001
	2*Cov($\theta - \bar{\theta}$,$Xb - \bar{X}b$)	-0.001	-0.6	-0.001	-0.4	-0.003	-1.3	-0.002
	2*Cov($\theta - \bar{\theta}$, u)	0.000	0.0	0.000	0.0	0.000	0.0	0.000
	2*Cov($Xb - \bar{X}b$, u)	0.000	0.0	0.000	0.0	0.000	0.0	0.000
Segregation Index	$\frac{Var(\bar{\theta}_j)}{Var(\theta_i)}$	0.309		0.346		0.368		0.058
N		51,624,477		53,214,824		52,140,050		

Note: This table presents the decomposition of wage variance and its evolution over three periods using the standard AKM model. The analysis is restricted to firms with more than 20 workers per year. "Comp." denotes the component of variance, while "Share"indicates the percentage of total $Var(y)$. The last column show the change in levels from 2002-2007 to 2014-2019. The decomposition is based on Equations 2, 3, and 4. The number of observations refer to the largest connected set, i.e. the group of firms with more than 20 workers per year connected by worker mobility.

TABLE A4. Decomposition of wage variance and its evolution

Establishments with 20+ employees (uncorrected AKM)

		2002-2007		2008-2013		2014-2019		Change from 2002-2007 to 2014-2019
		Comp.	Share	Comp.	Share	Comp.	Share	Diff.
Total variance	$\text{Var}(y)$	0.214		0.210		0.221		0.007
	$\text{Var}\ (\theta)$	0.161	75.4	0.157	74.7	0.170	76.6	0.008
	$\text{Var}\ (\psi)$	0.027	12.4	0.023	10.9	0.021	9.6	-0.005
	$\text{Var}(Xb)$	0.003	1.3	0.002	1.1	0.002	1.0	-0.001
	$\text{Var}(u)$	0.009	4.3	0.009	4.4	0.010	4.5	0.001
	$2^*\text{Cov}(\theta,\psi)$	0.012	5.4	0.017	8.1	0.018	8.2	0.007
	$2^*\text{Cov}(\theta,Xb)$	0.000	0.0	0.000	0.1	-0.001	-0.6	-0.001
	$2^*\text{Cov}(\psi,Xb)$	0.001	0.3	0.001	0.3	0.001	0.3	0.000
Between-firm variance	$\text{Var}(\bar{y})$	0.099	46.1	0.105	50.1	0.114	51.3	0.015
	$\text{Var}\ (\bar{\theta})$	0.058	27.3	0.063	30.1	0.072	32.3	0.013
	$\text{Var}\ (\psi)$	0.027	12.4	0.023	10.9	0.021	9.6	-0.005
	$\text{Var}(\overline{Xb})$	0.000	0.1	0.000	0.1	0.000	0.1	0.000
	$2^*\text{Cov}(\bar{\theta},\psi)$	0.012	5.4	0.017	8.1	0.018	8.2	0.007
	$2^*\text{Cov}(\bar{\theta},\overline{Xb})$	0.001	0.5	0.001	0.6	0.001	0.6	0.000
	$2^*\text{Cov}(\psi,\overline{Xb})$	0.001	0.3	0.001	0.3	0.001	0.3	0.000
Within-firm variance	$\text{Var}(y-\bar{y})$	0.115	53.9	0.105	49.9	0.108	48.7	-0.007
	$\text{Var}\ (\theta-\bar{\theta})$	0.103	48.1	0.098	44.6	0.098	44.3	-0.005
	$\text{Var}(Xb-\overline{Xb})$	0.003	1.2	0.002	1.0	0.002	0.9	0.000
	$\text{Var}(u)$	0.009	4.3	0.009	4.4	0.010	4.5	0.001
	$2^*\text{Cov}(\theta-\bar{\theta},Xb-\overline{Xb})$	-0.001	-0.5	-0.001	-0.5	-0.003	-1.3	-0.002
	$2^*\text{Cov}(\theta-\bar{\theta},\ u)$	0.000	0.0	0.000	0.0	0.000	0.0	0.000
	$2^*\text{Cov}(Xb-\overline{Xb},\ u)$	0.000	0.0	0.000	0.0	0.000	0.0	0.000
Segregation Index	$\frac{Var(\bar{\theta}_j)}{Var(\theta_i)}$	0.362		0.403		0.422		0.060
N		47,376,218		48,510,939		48,242,569		

Note: This table presents the decomposition of wage variance and its evolution over three periods using the standard AKM model. The analysis is restricted to establishments with more than 20 workers per year. "Comp." denotes the component of variance, while "Share"indicates the percentage of total $Var(y)$. The last column show the change in levels from 2002-2007 to 2014-2019. The decomposition is based on Equations 2, 3, and 4. The number of observations refer to the largest connected set, i.e. the group of establishments with more than 20 workers per year connected by worker mobility.

Table A5. Decomposition of wage variance and its evolution

Split-sampling with firm splitting, with an approximation for missing estimates

		2002-2007		2008-2013		2014-2019		Change from 2002-2007 to 2014-2019
		Comp.	Share	Comp.	Share	Comp.	Share	Diff.
Total variance	Var(y)	0.213		0.208		0.220		0.006
	Var (θ)**	0.152	71.5	0.149	71.9	0.161	73.2	0.008
	Var (ψ)	0.014	6.5	0.014	6.6	0.013	5.8	-0.001
	Var(Xb)	0.003	1.4	0.002	1.1	0.003	1.1	-0.001
	Var(u)**	0.015	7.0	0.014	6.6	0.015	7.0	0.001
	2*Cov(θ,ψ)	0.026	12.4	0.027	12.9	0.029	13.1	0.002
	2*Cov(θ,Xb)	0.000	0.0	0.000	0.2	-0.002	-0.7	-0.002
	2*Cov(ψ,Xb)	0.001	0.3	0.001	0.3	0.001	0.3	0.000
Between-firm variance	Var($\bar{y}$)	0.088	41.1	0.094	45.0	0.103	46.8	0.015
	Var ($\bar{\theta}$)	0.043	20.4	0.049	23.8	0.057	26.2	0.014
	Var (ψ)	0.014	6.5	0.014	6.6	0.013	5.8	-0.001
	Var($\overline{Xb}$)	0.000	0.1	0.000	0.1	0.000	0.1	0.000
	2*Cov($\bar{\theta}$,ψ)	0.026	12.4	0.027	12.9	0.029	13.1	0.002
	2*Cov($\bar{\theta}$,$\overline{Xb}$)	0.001	0.6	0.001	0.6	0.001	0.7	0.000
	2*Cov(ψ,$\overline{Xb}$)	0.001	0.3	0.001	0.3	0.001	0.3	0.000
Within-firm variance	Var($y - \bar{y}$)	0.126	58.9	0.114	55.0	0.117	53.2	-0.009
	Var ($\theta - \bar{\theta}$)**	0.107	50.3	0.098	47.3	0.102	46.3	-0.005
	Var($Xb - \overline{Xb}$)	0.003	1.3	0.002	1.0	0.002	1.1	0.000
	Var(u)**	0.015	7.0	0.014	6.6	0.015	7.0	0.001
	2*Cov($\theta - \bar{\theta}$,$Xb - \overline{Xb}$)	-0.001	-0.5	-0.001	-0.4	-0.003	-1.4	-0.002
	2*Cov($\theta - \bar{\theta}$, u)**	0.000	0.0	0.000	0.0	0.000	0.0	0.000
	2*Cov($Xb - \overline{Xb}$, u)	0.000	0.0	0.000	0.0	0.000	0.0	0.000
Segregation Index	$\frac{Var(\bar{\theta}_j)}{Var(\theta_i)}$	0.285		0.331		0.358		0.075
N		52,154,249		54,628,124		53,141,843		

Note: This table presents the decomposition of wage variance and its evolution over three periods using the split-sampling method with firm splitting. "Comp." denotes the component of variance, while "Share"indicates the percentage of total *Var*(y). The last column show the change in levels from 2002-2007 to 2014-2019. The decomposition is based on Equations 2, 3, and 4. The estimation is performed on the sample of firms present in both main connected components in each split sample. The split-sampling method with firm splitting is described in Section 2.3.1.

** : These parameters' estimates are not directly corrected by firm split. The procedure for recovering them is described in Section D.2.

TABLE A6. Decomposition of wage variance and its evolution

Split-sampling with period splitting

		2002-2007		2008-2013		2014-2019		Change from 2002-2007 to 2014-2019
		Comp.	Share	Comp.	Share	Comp.	Share	Diff.
	Var (θ)	0.153	71.5	0.151	71.9	0.162	73.1	0.009
	Var (ψ)	0.014	6.6	0.014	6.6	0.011	5.1	-0.003
	Var(Xb)	0.003	1.4	0.002	1.1	0.002	1.1	-0.001
	Var(u)**	0.016	7.4	0.019	7.1	0.019	8.7	0.003
	2*Cov(θ,ψ)	0.026	12.2	0.026	12.5	0.027	12.2	0.001
	2*Cov(θ,Xb)	0.000	-0.2	0.000	0.0	-0.002	-1.0	-0.002
	2*Cov(ψ,Xb)	0.001	0.3	0.001	0.3	0.001	0.3	0.000
Between-firm variance	Var$(\bar{y})$	0.088	40.9	0.094	44.7	0.102	46.2	0.015
	Var $(\bar{\theta})$	0.045	21.2	0.052	24.6	0.061	27.5	0.016
	Var (ψ)	0.014	6.6	0.014	6.6	0.011	5.1	-0.003
	Var$(\overline{Xb})$	0.000	0.1	0.000	0.1	0.000	0.1	0.000
	2*Cov$(\bar{\theta},\psi)$	0.026	12.3	0.026	12.5	0.027	12.3	0.001
	2*Cov$(\bar{\theta},\overline{Xb})$	0.001	0.5	0.001	0.6	0.001	0.6	0.000
	2*Cov$(\psi,\overline{Xb})$	0.001	0.3	0.001	0.3	0.001	0.3	0.000
Within-firm variance	Var$(y-\bar{y})$	0.127	59.1	0.116	55.3	0.119	53.8	-0.007
	Var $(\theta-\bar{\theta})$	0.108	50.4	0.100	47.4	0.102	46.0	-0.006
	Var$(Xb-\overline{Xb})$	0.003	1.3	0.002	1.0	0.002	1.0	-0.001
	Var(u)**	0.016	7.4	0.015	7.1	0.019	8.7	0.003
	2*Cov$(\theta-\bar{\theta},Xb-\overline{Xb})$	-0.001	-0.7	-0.001	-0.5	-0.003	-1.5	-0.002
	2*Cov$(\theta-\bar{\theta}, u)$	0.000	0.0	0.000	0.0	0.000	0.0	0.000
	2*Cov$(Xb-\overline{Xb}, u)$	0.000	0.0	0.000	0.0	0.000	0.0	0.000
Segregation Index	$\frac{Var(\bar{\theta}_j)}{Var(\theta_i)}$	0.296		0.342		0.376		0.080
N		46,525,005		48,960,858		47,512,825		

Note: This table presents the decomposition of wage variance and its evolution over three periods using the split-sampling method with period splitting. "Comp." denotes the component of variance, while "Share"indicates the percentage of total $Var(y)$. The last column show the change in levels from 2002-2007 to 2014-2019. The decomposition is based on Equations 2, 3, and 4. The estimation is performed on the sample of firms and, when necessary, individuals present in both main connected components in each split sample. The split-sampling method with period splitting is described in Section 2.3.1.

** Var(u) is not directly corrected by period splitting. However, since all other components of the variance decomposition are known, $var(u)$ can be inferred by subtraction.

Table A7. Decomposition of wage variance and its evolution

Firm Clustering

		2002-2007		2008-2013		2014-2019		Change from 2002-2007 to 2014-2019
		Comp.	Share	Comp.	Share	Comp.	Share	Diff.
Total variance	Var(y)	0.211		0.203		0.212		0.002
	Var (θ)	0.164	77.8	0.157	77.6	0.165	77.5	0.001
	Var (ψ)	0.005	2.3	0.005	2.4	0.005	2.4	0.000
	Var(Xb)	0.003	1.5	0.002	1.2	0.003	1.3	0.000
	Var(u)	0.011	5.1	0.010	5.1	0.011	5.1	0.000
	2*Cov(θ,ψ)	0.025	12.1	0.027	13.1	0.030	14.4	0.005
	2*Cov(θ,Xb)	0.000	0.2	0.000	0.0	-0.003	-1.2	-0.003
	2*Cov(ψ,Xb)	0.001	0.2	0.001	0.3	0.001	0.3	0.000
Between-firm variance	Var($\bar{y}$)	0.078	37.1	0.082	40.7	0.094	44.0	0.015
	Var ($\bar{\theta}$)	0.046	21.9	0.049	24.2	0.056	26.2	0.009
	Var (ψ)	0.005	2.3	0.005	2.4	0.005	2.4	0.000
	Var($\overline{Xb}$)	0.000	0.0	0.000	0.0	0.000	0.0	0.000
	2*Cov($\bar{\theta}$,ψ)	0.025	12.1	0.027	13.1	0.030	14.4	0.005
	2*Cov($\bar{\theta}$,$\overline{Xb}$)	0.001	0.7	0.001	0.7	0.002	0.8	0.000
	2*Cov(ψ,$\overline{Xb}$)	0.001	0.2	0.001	0.3	0.001	0.3	0.000
Within-firm variance	Var($y - \bar{y}$)	0.132	62.9	0.120	59.3	0.119	56.0	-0.013
	Var ($\theta - \bar{\theta}$)	0.118	55.9	0.108	53.4	0.109	51.3	-0.009
	Var($Xb - \overline{Xb}$)	0.003	1.4	0.002	1.2	0.003	1.3	0.000
	Var(u)	0.011	5.1	0.010	5.1	0.011	5.1	0.000
	2*Cov($\theta - \bar{\theta}$,$Xb - \overline{Xb}$)	-0.001	-0.5	-0.001	-0.7	-0.004	-2.0	-0.003
	2*Cov($\theta - \bar{\theta}$, u)	0.000	0.0	0.000	0.0	0.000	0.0	0.000
	2*Cov($Xb - \overline{Xb}$, u)	0.000	0.0	0.000	0.0	0.000	0.0	0.000
Segregation Index	$\frac{\text{Var}(\bar{\theta})}{\text{Var}(\theta)}$	0.282		0.312		0.338		0.056
N		61,925,099		66,706,199		65,410,886		

Note: This table presents the decomposition of wage variance and its evolution over three periods using the firm clustering method. "Comp." denotes the component of variance, while "Share"indicates the percentage of total *Var*(y). The last column show the change in levels from 2002-2007 to 2014-2019. The decomposition is based on Equations 2, 3, and 4. Firm clustering method is described in Section 2.3.2.

TABLE A8. Decomposition of wage variance and its evolution

Split-sampling correction with firm splitting - 90+ days worked in a year

		2002-2007		2008-2013		2014-2019		Change from 2002-2007 to 2014-2019
		Comp.	Share	Comp.	Share	Comp.	Share	Diff.
Total variance	Var(y)	0.212		0.200		0.211		-0.001
	Var (θ)	**	**	**	**	**	**	**
	Var (ψ)	0.015	6.9	0.014	7.1	0.012	5.8	-0.003
	Var(Xb)	0.005	2.2	0.003	1.7	0.004	1.7	-0.001
	Var(u)	**	**	**	**	**	**	**
	2*Cov(θ,ψ)	0.026	12.3	0.027	13.5	0.029	13.7	0.003
	2*Cov(θ,Xb)	0.007	3.2	0.006	2.9	0.006	2.7	-0.001
	2*Cov(ψ,Xb)	0.001	0.7	0.001	0.7	0.001	0.6	0.000
Between-firm variance	Var($\bar{y}$)	0.089	42.1	0.092	45.7	0.101	47.8	0.012
	Var ($\bar{\theta}$)	0.040	19.1	0.043	21.6	0.053	24.9	0.012
	Var (ψ)	0.015	6.9	0.014	7.1	0.012	5.8	-0.003
	Var($\overline{Xb}$)	0.001	0.4	0.001	0.3	0.001	0.3	0.000
	2*Cov($\bar{\theta}$,ψ)	0.026	12.4	0.027	13.5	0.029	13.7	0.003
	2*Cov($\bar{\theta}$,$\overline{Xb}$)	0.004	1.9	0.004	1.8	0.004	2.1	0.000
	2*Cov(ψ,$\overline{Xb}$)	0.001	0.7	0.001	0.7	0.001	0.6	0.000
Within-firm variance	Var($y - \bar{y}$)	0.123	57.9	0.109	54.3	0.110	52.2	-0.013
	Var ($\theta - \bar{\theta}$)	**	**	**	**	**	**	**
	Var($Xb - \overline{Xb}$)	0.004	1.9	0.003	1.4	0.003	1.4	-0.001
	Var(u)	**	**	**	**	**	**	**
	2*Cov($\theta - \bar{\theta}$,$Xb - \overline{Xb}$)	0.003	1.3	0.002	1.1	0.001	0.5	-0.002
	2*Cov($\theta - \bar{\theta}$, u)	**	**	**	**	**	**	**
	2*Cov($Xb - \overline{Xb}$, u)	0.000	0.0	0.000	0.0	0.000	0.0	0.000
N		84,924,827		89,176,024		91,436,161		

Note: This table presents the decomposition of wage variance and its evolution over three periods using the split-sampling method with firm splitting, extending the sample selection to individuals employed for at least 90 days by the same firm during the year. "Comp." denotes the component of variance, while "Share"indicates the percentage of total *Var*(y). The last column show the change in levels from 2002-2007 to 2014-2019. The decomposition is based on Equations 2, 3, and 4. The estimation is performed on the sample of firms present in both main connected components in each split sample. The split-sampling method with firm splitting is described in Section 2.3.1.

** : These parameters' estimates are not corrected by firm-split

Table A9. Decomposition of changes in worker fixed effects variance

Period	$V(\theta)$	$V(\omega)$	$V(\varepsilon)$
2002-2007	0.1522	0.0950	0.0572
2008-2013	0.1492	0.0873	0.0619
2014-2019	0.1605	0.0884	0.0721
Diff (2014-19 vs 2002-07)	0.0084	-0.0066	0.0149

Notes: This table shows the decomposition of changes in worker fixed effects variance over the three time periods. $V(\theta)$ represents the variance of worker fixed effects, $V(\omega)$ the variance of the occupation-specific component, and $V(\varepsilon)$ the variance of the individual-specific component. All estimates are corrected by the split-sampling method with firm splitting.

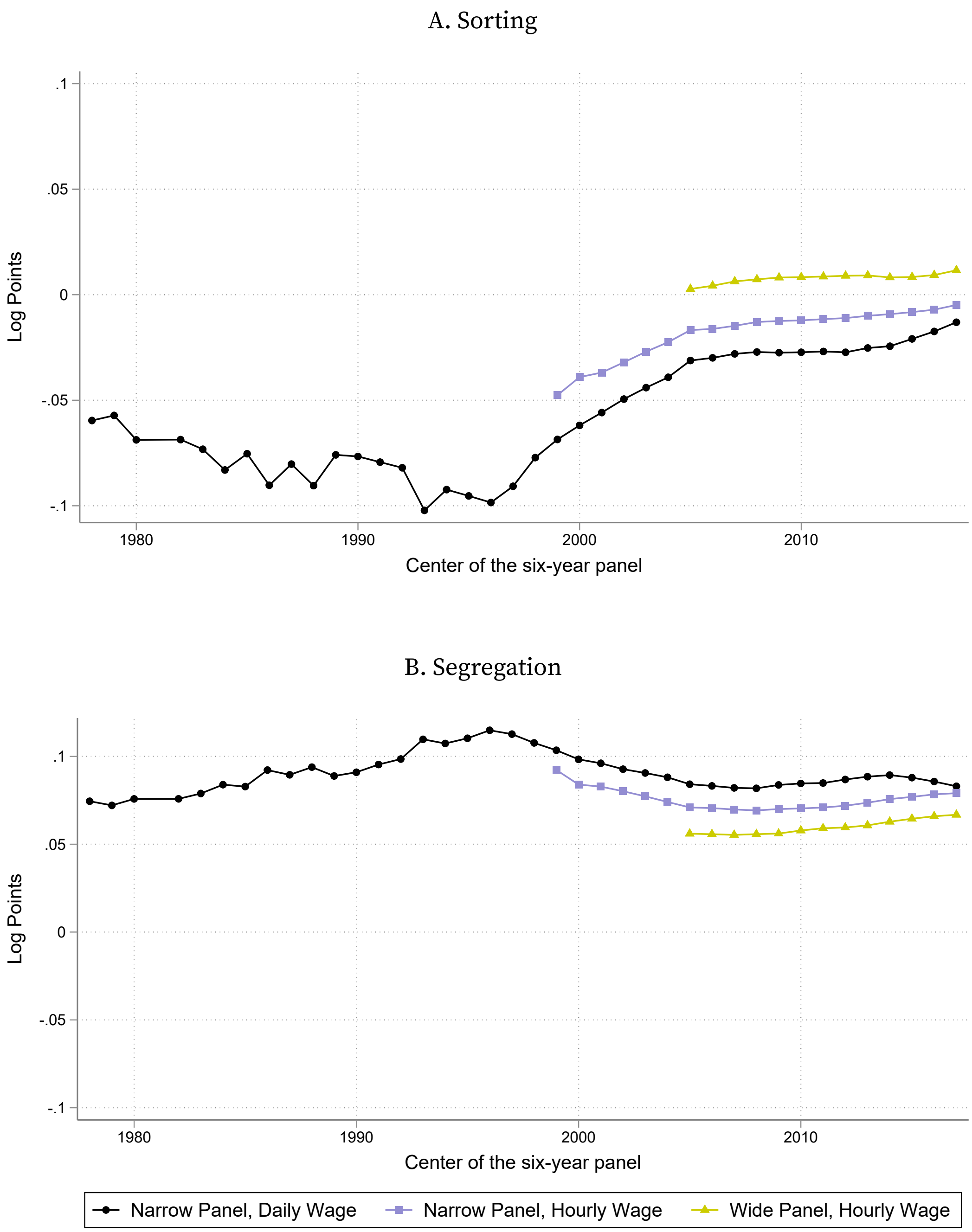
TABLE A10. Within-occupation variance of individual-specific component

		$Var_o(\varepsilon_{i,t})$		
	CS	2002-2007	2008-2013	2014-2019
CEOs	23	0.29	0.33	0.43
Professionals	31	0.17	0.19	0.20
Public administration managers	33	0.16	0.15	0.11
Scientific professions	34	0.21	0.18	0.19
Artists and media professionals	35	0.18	0.17	0.19
Managers	37	0.18	0.17	0.19
Engineers	38	0.12	0.11	0.13
School teachers	42	0.17	0.13	0.14
Health and social workers	43	0.08	0.08	0.08
Public administration intermediates	45	0.09	0.08	0.12
Business administration intermediates	46	0.09	0.08	0.08
Technicians	47	0.06	0.06	0.07
Intermediate supervisors	48	0.07	0.07	0.07
Public administration clerks	52	0.05	0.04	0.04
Security agents	53	0.04	0.06	0.07
Business administration clerks	54	0.06	0.07	0.08
Retail salespersons	55	0.04	0.04	0.05
Personal service employees	56	0.05	0.04	0.05
Skilled manufacturing workers	62	0.05	0.05	0.05
Skilled artisans	63	0.06	0.06	0.06
Drivers	64	0.04	0.04	0.03
Handling, transport skilled workers	65	0.04	0.05	0.05
Unskilled manufacturing workers	67	0.04	0.05	0.05
Unskilled artisans	68	0.04	0.05	0.05
Farm workers	69	0.09	0.09	0.10

Notes: $Var_o(\varepsilon_{i,t})$ represents the within-occupation variance of the individual-specific component of the worker fixed effect. CS represents the two-digit code for *catégories socioprofessionnelles*, which can be further explored here. This data complements the information on average worker fixed effects and occupational shares presented in Table 5. All estimates are corrected by the split-sampling method with firm splitting

Appendix B. Appendix Figures

FIGURE A1. Historical Series - Uncorrected Estimates

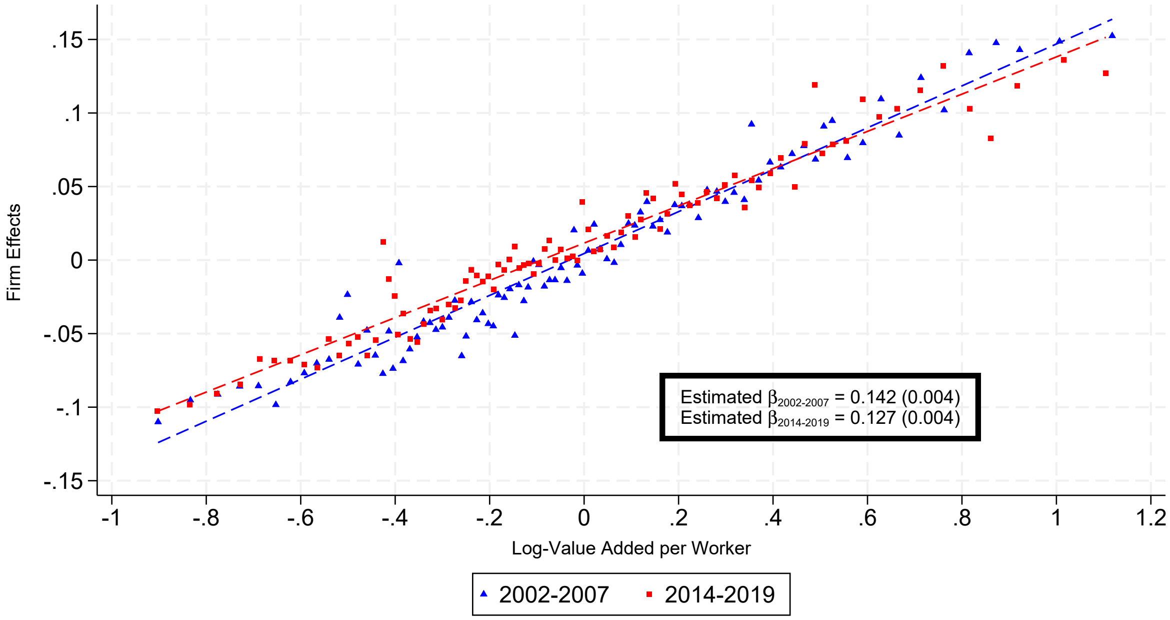


A. Sorting

B. Segregation

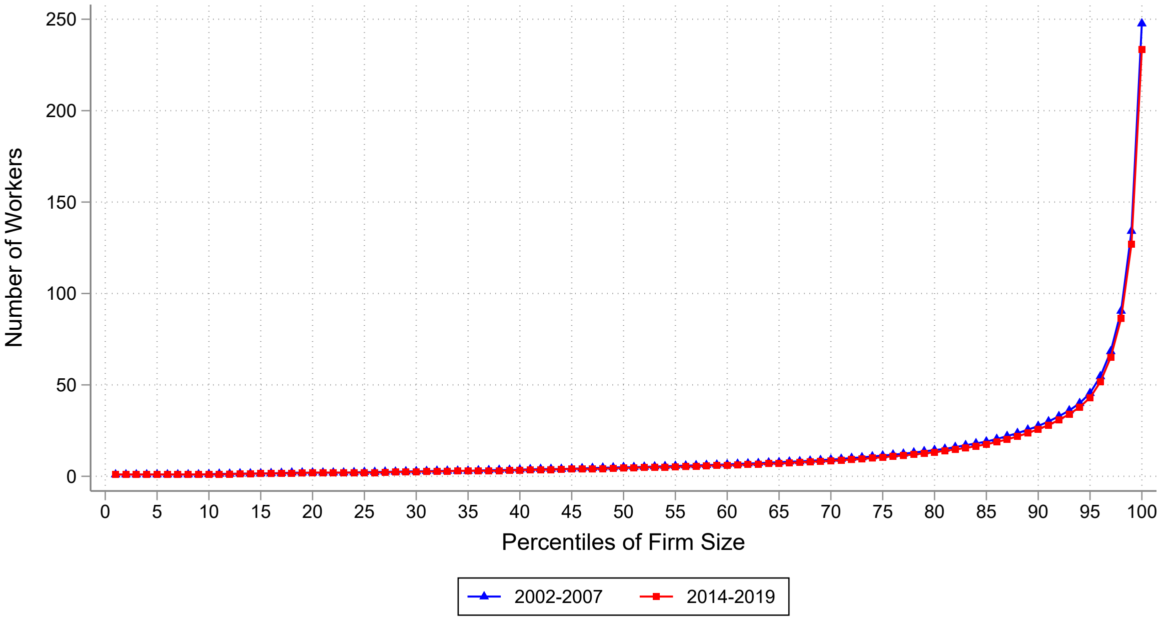
Note: *Note*: This figure presents estimates of sorting ($2 * Cov(\theta, \psi)$) and segregation ($Var(\bar{\theta})$) using rolling six-year periods. We include only individuals employed by the same firm for at least 360 days in the wide panel, and 90 days in the narrow panel, for a given year. The wider selection criteria for the narrow panel aims to enhance connectivity. Public administration employees and firms are excluded from the analysis. All estimates come from a standard AKM model. Estimates from the narrow panel are particularly affected by the selection of bigger and more connected firms. Data for the years 1981, 1983, and 1990 are missing. There have been several changes in scope and variable definition since 1976.

FIGURE A2. Rent-Sharing - Firm Fixed Effects vs Log Value Added/Worker



Note: This figure shows the firm-level, employment-weighted relationships between value added per worker and firm fixed effects across periods. Only firms in the largest connected set with available information on value added are included. The points shown represent mean estimated firm fixed effects from the AKM models, averaged across firms in 100 percentile bins of period-demeaned log value added per worker. Period best-fitting lines from employment-weighted OLS are reported.

FIGURE A3. Cumulative firm size distribution



Note: This figure shows the share of firms below a certain size, by period. Only firms in the largest connected set, i.e. the group of firms connected by worker mobility, are included. Size is defined as the average annual number of workers over the years within a period.

Appendix C. Constructing a BTS full panel

C.1. Chaining the yearfiles

The French BTS is not a proper panel dataset of workers because before 2002 there are no individual IDs after 2002 the individual IDs are specific to each annual file (herefter "yearfile"). However, each yearfile y contains information on both the current year t and the previous year $t-1$ (variables for the year $t-1$ end with "_1"). We therefore take advantage of this overlap to build a pseudo-panel based on common information between year t of yearfile $y-1$ and year $t-1$ of yearfile y. We obtained permission from Insee to chain the BTS annual files in order to create a full panel of the wage-earning population between 1994 and 2020. From 1994 to 2001, the annual files lack unique worker IDs, making it possible to track only workers who stayed in the same firm (the "stayers"). During this period, if a worker changed jobs, he would appear as two different people in the same yearfile. After 2002, the introduction of yearfile-specific individual IDs allowed us to match both stayers and those workers who change jobs within the yearfile (the "movers").

In order to conduct the match, we used the following variables: sex (`SEXE`), firm ID (`SIREN`), establishment ID (`NIC`), number of hours (`NBHEUR` or `NBHEUR_1`), starting day of the job during the year (`DATDEB` or `DATDEB_1`), ending day of the job during the year (`DATFIN` or `DATFIN_1`), number of days between starting and ending day (`DUREE` or `DUREE_1`), municipality of residence (`COMR` or `COMR_1`), municipality of work (`COMT` or `COMT_1`), being part of the sample used for the DADS panel (`SONDE` or `SONDE_1`), and gross wage (`S_BRUT` or `S_BRUT_1`) and age (`AGE`). We run the match with a SAS script at the regional level, using the BTS regional files[27]. Within the regional file, we keep the job for which a worker i has the highest pay. We create the following keys for the year t of yearfile $y-1$:

```
pseudoid=COMPRESS(SEXE!!"#"!!SIREN!!"#"!!NIC!!"#"!!ROUND(NBHEUR,1)!!"#"!!
DATDEB!!"#"!!DATFIN!!"#"!!DUREE !! "#" !!COMR!!"#"!!COMT !! "#" !! SONDE);
```

and the following for the year $t-1$ of yearfile y:

```
pseudoid_b=COMPRESS(SEXE!!"#"!!SIREN!!"#" !!NIC!!"#"!!ROUND(NBHEUR_1,1)!!"#"!!
DATDEB_1!!"#"!!DATFIN_1!!"#"!!DUREE_1!! "#"!! COMR_1!!"#"!!COMT_1!!"#" !! SONDE_1);
```

[27] In the current project, we restricted the match to mainland France and excluded overseas departments (DOM).

However, since there are some discrepancies in the ages and wages reported for the same year in the yearfile $y - 1$ and y, we do not use them directly in the matching key. We use the HAVING property of the SQL procedure to select the match with the smallest difference between the two wages and an absolute age difference of less than two years.

```
PROC SQL;
    CREATE TABLE ab (DROP=pseudoid pseudoid_b S_BRUT S_BRUT_1 AGE)
        AS SELECT * FROM a1 (KEEP=pseudoid s_brut IDENT_S ID2 REGT
        AGE NBHEUR) AS aa
    FULL JOIN b1 (keep=pseudoid_b s_brut_1 IDENT_S ID2_B AGE
        DEP_NAISS NBHEUR_1 rename=(IDENT_S=IDENT_S_B AGE=AGE_B))
        AS bb
    ON aa.pseudoid=bb.pseudoid_B
    GROUP BY aa.S_BRUT,aa.PSEUDOID
    HAVING ABS(aa.s_brut-bb.s_brut_1)=MIN(ABS(aa.s_brut-bb.s_brut_1))
    AND (0<=bb.AGE_B-aa.AGE<2 or AGE_B=. or AGE=.)
    ORDER BY aa.PSEUDOID, bb.s_brut_1;
QUIT;
```

This code has been adjusted to account for the fileyear specificity.

- For years before 2002 ($y < 2002$ and $y - 1 < 2001$), we create an individual ID based on the initial row numbers in each regional file, to which we add the regional code at the end. For example, the ID for the 10th observation of the Paris region (code: 11) will be 1011.

- In 2013 ($y = 2013$ and $y - 1 = 2012$), the variable `SONDE` causes a mismatch and is excluded from the pseudoid key.

- After 2013 ($y > 2013$ and $y - 1 > 2012$) we found that the number of hours for the same year differed between the yearfile $y - 1$ and y. So we excluded the number of hours from the matching key and added the minimal difference in number of hours to the having clause.

We count the number of matches based on the procedure, and we assign the same ID only to workers with a single match. Finally, we chain the different IDs starting from the first year of the DADS (1994). The ID files (*PSID*_1994 to *PSID*_2020) contain the ID of the year (`IDENT_S`) and a permanent ID (`IDENT_ALL`), which is based on the initial ID of

an employee when she first appears in the DADS, to which we add the year of the first appearance on two digits. The full SAS script pseudo_id.sas is available at the following address:
http://olivier.godechot.free.fr/hopfichiers/pseudo_id.zip. It comes with three additional SAS scripts for creating DADS files with the identifier IDENT_ALL included, for creating and adding seniority variables, and for correcting information on workers' location of birth and citizenship.

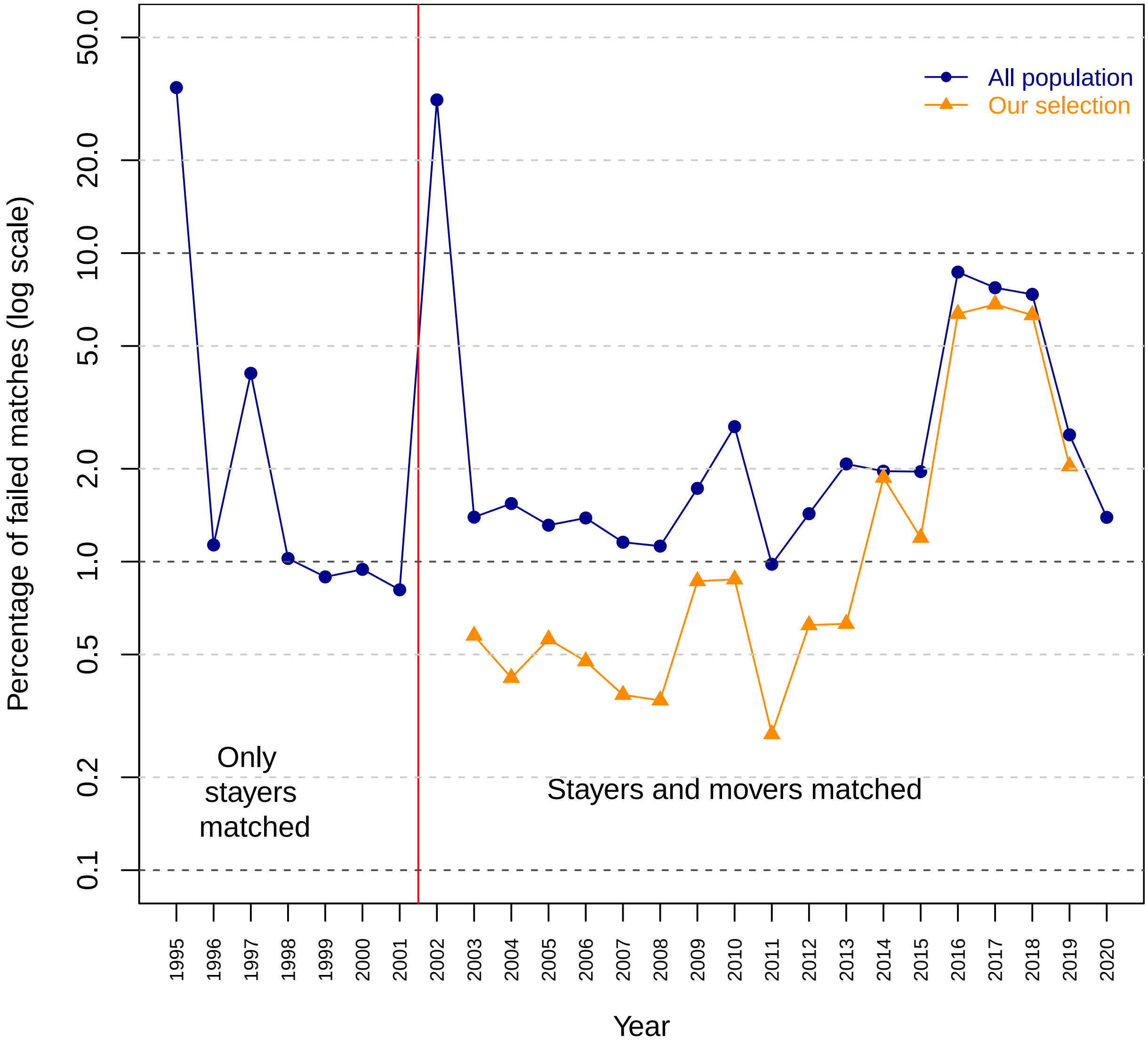
C.2. Quality of the identification

To avoid false identifications, we have chosen a conservative procedure to identify two individuals as the same person by using the maximum amount of overlapping information available. If the procedure results in multiple matches, we do not impute an identification. However, these duplicates remain rare, about 0.4% of the observations. Most failed matches are due to observations for which we do not find a match. Figure A4 gives a first indication of the quality of the match. In general, we find that the match fails for 1 to 2% of the overlapping years of two yearfiles. The quality of the matching decreases between 2016 and 2018, and the failed matches increase to 7 to 9%, and return to 3% in 2019, probably as a result of the switch from DADS to the DSN[28]. With the existing procedure, the match is poor for the 2002 yearfile (and similarly for 1995), as a consequence of the major transformation of the BTS between the 1994-2001 series and the 2002-2020 series[29]. Restricting the selection to the one used for this analysis, an even lower failure rate is observed. The failed match rate is generally lower than 1% for the majority of the years. As in the total population, it increases sharply between 2016 and 2018, reaching 6 to 7%.

[28] The "déclaration sociale nominative" is a new monthly administrative source that replaces the "déclarations annuelles de données sociales." INSEE produces from the DSN an annual file in the BTS format for the continuity of the series.

[29] One could probably improve the match for these years by eliminating variables in the matching key that are incomplete or coded differently. We have already eliminated the number of hours for 2002, and this increased the match rate from 60 to 68%.

FIGURE A4. Failed matches



Note: The figure reports the percentage of failed matches from the year $t-1$ of yearfile y. Before 2002, the lack of individual ID in the initial dataset makes it impossible to follow the movers. In 2002 and after, we can match both stayers and movers. The graph further separates the full population from the sample selection used in the analysis and described in Section 1.2.

Table A11 presents a detailed examination of the factors correlated with failed matches. The results show that failed matches are slightly more common among older workers, and significantly more common among agricultural workers (a very small category). During the 2016-2018 period, the rate of failed matches increased significantly more for engineers and in the utilities sector than for other categories. Logistic regressions, available upon request, generally confirm these findings.

Despite the high number of matches, the matching process is not without limitations. It is important to note that a false positive is still possible, although unlikely. Second, in

TABLE A11. Comprehensive Match Failure Analysis

		All years	2016-2018	Other years
	Sex			
	Male	1.90%	6.90%	0.80%
	Female	1.70%	6.00%	0.80%
	Age			
	25 and less	1.40%	4.80%	0.70%
	26 to 35	1.70%	6.20%	0.80%
	36 to 50	1.80%	6.70%	0.90%
	51 to 60	1.90%	7.00%	1.00%
	61 and more	2.10%	7.90%	1.10%
	Occupation			
20	CEOs	2.30%	9.30%	0.90%
32	Professionals	2.80%	8.60%	1.60%
35	Artists and media professionals	2.60%	9.00%	1.20%
37	Managers	2.40%	9.60%	0.90%
38	Engineers	3.20%	13.80%	0.90%
42	Primary school teachers	2.20%	7.20%	1.20%
43	Health and social workers	1.00%	3.50%	0.50%
45	Public administration intermediates	2.60%	2.50%	2.60%
46	Business administration intermediates	1.70%	4.50%	1.00%
47	Technicians	1.60%	5.80%	0.70%
48	Intermediate supervisors	1.80%	7.30%	0.60%
52	Public administration clerks	1.20%	4.30%	0.60%
53	Security agents	1.30%	4.00%	0.70%
54	Business administration clerks	1.30%	4.10%	0.70%
55	Retail salespersons	1.40%	4.80%	0.60%
56	Personal service employees	1.80%	5.40%	1.00%
62	Skilled manufacturing workers	1.50%	5.50%	0.70%
63	Skilled artisans	1.40%	5.60%	0.50%
64	Drivers	1.10%	3.60%	0.60%
65	Handling, transport skilled workers	1.30%	4.20%	0.70%
67	Unskilled manufacturing workers	1.70%	5.20%	1.00%
68	Unskilled artisans	1.80%	6.80%	0.70%
69	Farm workers	18.20%	24.80%	16.70%
	Industry			
AC	Farming and industry	2.30%	8.60%	0.90%
DE	Utilities	3.50%	14.60%	1.10%
F	Construction	1.80%	8.60%	0.40%
G	Commerce	1.40%	4.70%	0.70%
H	Transport	1.20%	3.70%	0.70%
I	Hotels, tourism, catering	1.20%	4.30%	0.60%
J	Media	2.20%	7.30%	1.10%
K	Financial services	2.40%	7.00%	1.40%
LM	Real estate, professional services	1.60%	5.60%	0.80%
N	Administrative services	1.70%	5.80%	0.80%
OPQ	Health, education	1.70%	5.50%	0.80%
R	Arts and recreation	2.10%	6.50%	1.10%
STU	Other	1.80%	6.60%	0.80%

Note: This table represents a detailed breakdown of match failure analysis across various categories for different years. The ’All years’ column provides the overall percentage, ’2016-2018’ shows the percentage for that specific time period, and ’Other Years’ shows the percentage for years outside of 2016-2018. The table refers to the sample selection described in Section 1.2.

order for an employee to be identified as the same person, he or she must be present in the BTS as a wage earner each year. This means that we cannot link the initial identification of a worker who was either unemployed, self-employed or a civil servant (before 2009) for more than one year with subsequent employment periods. However, the quality of the match seems sufficient to run AKM panel regressions.

C.3. How to use the ID files

In order to add the permanent ID to a given datafile (for instance a file `b2010` for the year 2010), the procedure is as follows:[30]

```
PROC SQL;
 CREATE TABLE b2010b
 AS SELECT * FROM b2010 AS aa
 LEFT JOIN psid.psid_2010 AS bb
 ON aa.ident_s=bb.ident_s;
QUIT;

data b2010c; set b2010b;
 if Missing(ident_all) then ident_all=ident_s*100+substr(AN,3,4);
run;
```

Before 2002, in order to get permanent IDs necessary for the match with the `PSID_yyyy` files, one needs to create an ID in each regional file (prior to any selection) as follows (for instance for Paris Region in 1997):

```
DATA b1197; SET po1997.post1197;
 ident_s=_N_*100+REG;
RUN;
```

Appendix D. Split-sampling bias correction

In the following sections, we show that the split-sample strategy yields unbiased estimates of the quadratic forms of the parameters under reasonable assumptions. In particular, this method effectively corrects for the limited mobility bias inherent in standard estimation techniques (Appendix D.1). Our implementation of split sampling

[30] The script pseudo_id_use.sas also provides a macro program to run these steps automatically.

involves certain complexities aimed at maximizing the connectivity within the data. To ensure clarity and reproducibility, we provide a detailed description of the procedure and the calculation of the corrected estimates in Appendix D.2. Similar to the classical AKM model, we lack theoretical results on the uncertainty associated with our estimates, mainly because this uncertainty is intrinsically linked to the structure of the mobility network and how it evolves as the number of observations increases. Nevertheless, we empirically verify the precision and stability of our estimates through multiple random splits (Appendix D.3) and simulation exercises (Appendix D.4).

D.1. Proof of unbiasedness

In this subsection, we provide a proof that the split-sample estimator is unbiased for the quadratic forms of interest in the context of the AKM model. Following Kline, Saggio, and Sølvsten (2020), we start with a simplified notation of the AKM model:

$$y_i = z_i'\alpha + u_i \tag{A1}$$

With $\alpha = (\beta, \theta, \psi)$ our parameter vector of length $k = 2 + N + J$ and z_i the non-random vector of regressors for the person-year observation i. We are interested in estimating a quadratic form of the parameters:

$$\mu = \alpha' A \alpha, \tag{A2}$$

where A is a known symmetric matrix corresponding to the variance components we aim to estimate (e.g., variance of firm effects, covariance between worker and firm effects, etc.). We randomly split the sample into two disjoint subsamples I_0 and I_1, each containing about half of the observations. Each split retains the same connectivity as the full sample, analogous to the leave-one-out condition in Kline, Saggio, and Sølvsten (2020), where the main connected sample remains connected even if a single observation is removed. For each split $s \in \{0, 1\}$, we estimate the parameter vector $\hat{\alpha}_s$:

$$\hat{\alpha}_s = S_{zz,s}^{-1} \sum_{i \in I_s} z_i y_i, \tag{A3}$$

where $S_{zz,s} = \sum_{i \in I_s} z_i z_i'$ is the design matrix for split s, assumed to be of full rank due to the connectivity condition. We can express $\hat{\alpha}_s$ in terms of the true parameter α and an estimation error ϵ_s:

$$\hat{\alpha}_s = \alpha + \epsilon_s, \tag{A4}$$

where

$$\epsilon_s = S_{zz,s}^{-1} \sum_{i \in I_s} z_i u_i. \tag{A5}$$

Our split-sample plug-in estimator for the quadratic form μ is then

$$\hat{\mu}^{SP} = \hat{\alpha}_0' A \hat{\alpha}_1. \tag{A6}$$

PROPOSITION A1. *Under the following assumptions:*

(a) ***Independence of Errors Within a Split****: The error terms have an expectation of zero ($E[u_i] = 0$) and are independent of z_i.*

(b) ***Independence of Errors Across Splits****: The error terms u_i, $i \in I_1$, are independent of u_j, $j \in I_0$.*

(c) ***Non-random Regressors****: The regressors z_i are non-random and fixed in repeated samples.*

(d) ***Full Rank Design Matrices****: The design matrices $S_{zz,s}$ are of full rank in each split s.*

Then, the split-sample estimator $\hat{\mu}^{SP} = \hat{\alpha}_0' A \hat{\alpha}_1$ is an unbiased estimator of the quadratic form $\mu = \alpha' A \alpha$, i.e.,

$$E[\hat{\mu}^{SP}] = \mu.$$

PROOF. We start by taking the expected value of $\hat{\mu}^{SP}$:

$$\begin{aligned}
E[\hat{\mu}^{SP}] &= E[\hat{\alpha}_0' A \hat{\alpha}_1] \\
&= E[\text{trace}(\hat{\alpha}_0' A \hat{\alpha}_1)] \qquad \text{(A7)} \\
&= E[\text{trace}(\hat{\alpha}_1 \hat{\alpha}_0' A)] \qquad \text{(A8)} \\
&= E[\text{trace}(A \hat{\alpha}_1 \hat{\alpha}_0')] \qquad \text{(A9)} \\
&= \text{trace}\left(A E[\hat{\alpha}_1 \hat{\alpha}_0']\right), \qquad \text{(A10)}
\end{aligned}$$

where we have used the properties of the trace operator and the fact that A is non-random. Substituting $\hat{\alpha}_s = \alpha + \epsilon_s$ from Equation (A4), we have

$$\begin{aligned} E[\hat{\alpha}_1\hat{\alpha}_0'] &= E\left[(\alpha + \epsilon_1)(\alpha + \epsilon_0)'\right] \\ &= \alpha\alpha' + \alpha E[\epsilon_0'] + E[\epsilon_1]\alpha' + E[\epsilon_1\epsilon_0']. \end{aligned} \tag{A11}$$

Since $E[\epsilon_s] = 0$ (because $E[u_i] = 0$ and z_i are non-random), the middle terms vanish:

$$E[\epsilon_s] = S_{zz,s}^{-1} \sum_{i \in I_s} z_i E[u_i] = 0. \tag{A12}$$

Thus, Equation (A11) simplifies to

$$E[\hat{\alpha}_1\hat{\alpha}_0'] = \alpha\alpha' + E[\epsilon_1\epsilon_0']. \tag{A13}$$

Substituting Equation (A13) back into Equation (A10), we have

$$\begin{aligned} E[\hat{\mu}^{SP}] &= \text{trace}\left(A\left(\alpha\alpha' + E[\epsilon_1\epsilon_0']\right)\right) \\ &= \text{trace}(A\alpha\alpha') + \text{trace}\left(AE[\epsilon_1\epsilon_0']\right). \end{aligned} \tag{A14}$$

Note that

$$\mu = \alpha' A \alpha = \text{trace}(A\alpha\alpha'). \tag{A15}$$

Therefore, the bias of the estimator is

$$\text{Bias} = E[\hat{\mu}^{SP}] - \mu = \text{trace}\left(AE[\epsilon_1\epsilon_0']\right). \tag{A16}$$

We compute $E[\epsilon_1\epsilon_0']$ explicitly:

$$E[\epsilon_1\epsilon_0'] = S_{zz,1}^{-1} E\left[\left(\sum_{i \in I_1} z_i u_i\right)\left(\sum_{j \in I_0} z_j u_j\right)'\right](S_{zz,0}^{-1})'. \tag{A17}$$

Since u_i and u_j are independent for $i \in I_1$ and $j \in I_0$, the expected cross-product is zero:

$$E\left[\left(\sum_{i \in I_1} z_i u_i\right)\left(\sum_{j \in I_0} z_j u_j\right)'\right] = \sum_{i \in I_1}\sum_{j \in I_0} z_i E[u_i u_j] z_j' = 0. \tag{A18}$$

Therefore,

$$E[\epsilon_1 \epsilon_0'] = 0. \tag{A19}$$

Substituting Equation (A19) back into Equation (A16), we find

$$\text{Bias} = \text{trace}\,(A \times 0) = 0 \tag{A20}$$

Thus,

$$E[\hat{\mu}^{SP}] = \mu. \tag{A21}$$

While we have established the unbiasedness of the split-sample estimator, analyzing its variance and consistency requires further considerations. The variance of $\hat{\mu}^{SP}$ arises from two sources: the randomness of the error terms u_i, and the randomness introduced by splitting the sample. To simplify the analysis, we may consider a fixed split and focus on the variance due to the error terms. The variance depends on the properties of the matrices A, $S_{zz,s}$, and their interactions. However, a comprehensive analysis of the variance requires additional assumptions about the distribution of u_i and the structure of the design matrices $S_{zz,s}$. Kline, Saggio, and Sølvsten (2020) discusses these conditions in the context of leave-one-out. Studying the consistency and convergence of $\hat{\mu}^{SP}$ involves considering how the design matrices and the mobility network evolve as the sample size increases. This is a complex issue because the connectedness of the data plays a crucial role in the estimation of fixed effects. This would imply additional hypothesis about how the mobility network changes when we add more years, more firms or more workers. We leave the detailed study of these properties to future research. Instead, we have empirically verified the stability and accuracy of our estimator through multiple random splits (see Section D.3) and simulation studies (see Section D.4). These empirical validations provide evidence that our estimator performs well in practice.

D.2. Practical details of split algorithms and calculation of corrected estimates

In this subsection, we describe in more detail the split sampling procedure implemented in our study. Inspired by Chanut (2018), the firm splitting algorithm aims to create two balanced samples while maintaining connectivity. It starts by randomly selecting a year from the 6-year panel. For each firm in that year, workers are split into two groups of equal size[31]. To improve connectivity, the algorithm splits stayers and movers separately, ensuring that movers are present in both halves if there are more than two in the firm for the given year. The process then continues by randomly selecting another year and repeating the process for “new workers” not yet assigned to a split. This continues for all six years of the panel.

The period splitting algorithm is designed to ensure that each worker with at least two years of observations has at least one observation in each split, thereby maximizing the number of identifiable parameters. For each eligible worker, the algorithm randomly selects a pivot year (except for the last year). All observations up to the end of the pivot year are assigned to a randomly chosen split, while all observations after the pivot year are assigned to the other split. This method maintains temporal consistency within each split for each worker.

After applying the AKM model to each split sample, we can associate two estimates for each parameter: $\theta_{i,s}$, $\psi_{i,s}$, $X_{i,s}$, where ’s’ denotes the split. The variance and covariance terms are computed using observations with estimates from both splits. For instance, the variance of firm effects is estimated as $\widehat{var}(\psi) = cov(\psi_{i,0}, \psi_{i,1})$ using the sample of all observations in firms belonging to both main connected components in each split sample.

Some estimates can not be computed directly by this method. This is always the case for $var(u)$: the residual is not a parameter and is specific to each observation. If all other components of the variance decomposition are known, $var(u)$ can be estimated by subtraction, as in Table A6. In the case of firm splitting, when there is no corrected value of $var(\theta)$, we have to proceed differently. To estimate $var(u)$, we define the Individual Residual (IR) as:

$$IR_{i,t,s} = y_{i,t,s} - \widehat{\psi_{i,t,1-s}} - \widehat{Xb_{i,t,1-s}}$$

where $s \in \{0, 1\}$ denotes the split, $y_{i,t,s}$ is the observed wage for individual i at time t in split s , $\widehat{\psi_{1-s}}$ and $\widehat{Xb_{1-s}}$ are the firm fixed effect and estimated effect of observable characteristics from the complementary split $1 - s$.

Proposition A2. *Under the assumptions of the AKM model and independence between splits,*

$$\frac{1}{N} \sum_{i,t,s} \left(cov(y_{i,t,s}, IR_{i,t,s} | i) \right) = var(u)$$

[31] If a firm has an odd number of workers, one of the two samples will have one more worker.

PROOF. We start by expressing the conditional covariance:

$$\begin{aligned} cov(y_{i,t,s}, IR_{i,t,s}|i) = cov[&\theta_i + \psi_{J(i,t)} + X_{i,t,s}b + u_{i,t,s}, \\ &\theta_i + (\psi_{J(i,t)} - \hat{\psi}_{J(i,t),1-s}) + (X_{i,t,s}b - X_{i,t,1-s}\hat{b}_{1-s}) + u_{i,t,s}|i] \end{aligned}$$

Note that $var(\theta_i|i) = 0$ because it is constant for a given individual. Under the assumptions of independence between splits and the properties of the AKM model, all covariance terms involving estimates from the complementary split have an expectation of zero. Therefore, the only non-zero term left in the conditional covariance is $var(u_{i,t,s}|i)$, which equals $var(u)$ under the assumption of homoskedasticity. Taking the average across all individuals, time periods, and splits[32] completes the proof:

$$\frac{1}{N}\sum_{i,t,s} cov(y_{i,t,s}, IR_{i,t,s}|i) = var(u)$$

We finally retrieve $var(\theta)$ by subtraction.

Calculation of the occupation-specific component of worker fixed effects. In Section 4.1 we examine the role of occupations in between-firm wage inequality. We compute the occupation-specific ω and individual-specific ε components of the worker fixed effects in Equation 5 separately in each split. For the firm splitting, each observation gets two estimates of the occupation-specific component, since all occupations are present in both main connected sets of each split. In Table 5, we report the average of these two estimates. The individual-specific component is only estimated in one split. The variance of the individual-specific effect in Table A9 is thus computed as the difference between the corrected estimates of $var(\theta)$ and $var(\omega)$.

D.3. Multiple random splits

We tested the stability of the firm splitting estimators with multiple random splits on two different data sets. First, on the long-term historical series, which are computed on the smaller and less connected "narrow panel" and show more variability due to splitting. In Figure 3, we plot the mean of 20 split sample estimates and a confidence interval around this mean. In our main estimates, for computational reasons, we limited

[32] In practice, $cov(y, IR|i)$ is estimated with the sample size correction factor $n/(n-1)$ because there are typically few observations per worker.

the multiple random split experiments to the first and third periods. In Table A12 we report the mean and standard deviation of 20 estimates. The standard deviations are very small relative to the estimates, the size of the bias correction, and the evolution between periods. Our split sampling corrected results are not due to random split noise. This exercise can also be interpreted as a bootstrap, which more generally indicates the high stability of the AKM decomposition statistics in our large dataset.

TABLE A12. Decomposition of wage variance and its evolution
Mean and standard deviation over 20 firm split estimations

		2002-2007		2014-2019	
		Mean	SD	Mean	SD
Total variance	Var(y)	0.213	0.00004	0.220	0.00002
	Var(ψ)	0.014	0.00004	0.013	0.00004
	Var(Xb)	0.003	0.00000	0.003	0.00000
	2Cov(θ,ψ)	0.026	0.00007	0.029	0.00005
	2Cov(θ,Xb)	0.000	0.00001	-0.002	0.00001
	2Cov(ψ,Xb)	0.001	0.00000	0.001	0.00000
Between-firm variance	Var($\bar{y}$)	0.088	0.00003	0.103	0.00002
	Var($\bar{\theta}$)	0.043	0.00007	0.057	0.00007
	Var($\bar{XB}$)	0.000	0.00000	0.000	0.00000
	2Cov($\bar{\theta}$,ψ)	0.027	0.00007	0.029	0.00005
	2Cov($\bar{\theta}$,$\bar{XB}$)	0.001	0.00000	0.001	0.00000
	2Cov(ψ,$\bar{XB}$)	0.001	0.00000	0.001	0.00000
Within-firm variance	Var($y - \bar{y}$)	0.126	0.00002	0.117	0.00001
	Var($Xb - \bar{Xb}$)	0.003	0.00000	0.002	0.00000
	2Cov($\theta - \bar{\theta}$,$Xb - \bar{Xb}$)	-0.001	0.00001	-0.003	0.00001
N of obs		52,152,944	8,709.74	53,142,355	9,360.74

Note: Mean and standard deviations computed on 20 estimations similar to table 2, on firms belonging to both main connected components

D.4. Simulations

To validate our split-sampling method, we conducted a simulation study using the 2002-2007 data framework. We generated simulated workers' and firms' fixed effects, incorporating sorting and noise, calibrated to match as much as possible the observed distributions. Crucially, we maintained the real mobility network from the actual data, as the bias in the estimation depends heavily on this network structure. This approach allows us to create a controlled environment where we know the true values of the variance components we aim to estimate. Table A13 presents the results of this simulation exercise. In the first column, we establish the ground truth for the full AKM estimation sample. These numbers represent the actual values of Var(θ), Var(ψ), and 2Cov(θ, ψ) computed from our simulated fixed effects. This column serves as our baseline, showing what an ideal estimation method should recover. The second column shows the limitations of standard AKM estimation when applied to our simulated data. By comparing these estimates to the ground truth in the first column, we can clearly see the bias inherent in traditional AKM methods – specifically, the overestimation of variances and underestimation of covariance. The third column presents the ground truth for the specific subsample used in split-sampling estimation. This subsample consists of firms that belong to the connected component in both splits, which results in a slightly different set of true values. Finally, the fourth column shows the results of our split-sampling estimation. By comparing these estimates to the subsample ground truth in the third column, we can assess how well our method recovers the true values within its operational sample. It's worth noting that while our simulation captures many key aspects of the estimation challenge, it cannot fully replicate potential selection effects that might occur in real-world sample reductions. Despite this limitation, the results clearly show that our split-sampling method successfully recovers the true variance components and addresses the biases found in standard AKM estimation.

TABLE A13. Simulated wage: true fixed effects and estimations

	(1)	(2)	(3)	(4)
Var(θ)	0.1499	0.1630	0.1499	
Var(ψ)	0.0140	0.0165	0.0137	0.0138
2Cov(θ,ψ)	0.0251	0.0205	0.0246	0.0251
N of obs	58,666,317	58,666,317	52,146,451	52,146,451

Simulation on 2002-2007 data, corrected estimates with period split method. First column: ground truth on AKM estimation sample (true quadratic terms computed on simulated fixed effects). Second column: AKM estimates (on simulated fixed effects). Third column: ground truth on the split-sampling estimation sample. Fourth column: slip-sampling estimates.