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

Damien Babet, Olivier Godechot, Marco G Palladino

In the Land of AKM:

Explaining the Dynamics of Wage Inequality in France[*]

Damien Babet[†] Olivier Godechot[‡] Marco G. Palladino[§]

October 10, 2022

Abstract

We use a newly built and exhaustive matched employer-employee database to study the contribution of firms to the dynamics of wage inequalities in France, where the original Abowd, Kramarz and Margolis (1999) double fixed effects model of log-wages was originally estimated on sampled data. Contrary to other countries, overall wage inequalities decreased in the 2002-2016 period. But France is not entirely an exception: the same polarizing dynamics observed in other countries are operating through an increase in between-firm inequalities. By applying the AKM model of log-wages with workers and firms additive fixed effects, we document increased sorting of high-wage workers to high-wage firms. We correct for bias in estimates of variance and covariance by clustering firms, and by splitting the sample as a simplification of previous methods. The rise in sorting is robust to these correction strategies and linked entirely to firm demographics and worker composition changes over time across firms. Over the same period, bottom earnings percentiles increased more than the rest of the distribution, in line with the rise in the legal minimum wage. As a result, within-firm inequalities decreased, more than offsetting the rising between-firm inequalities.

[*] We thank Kerstin Holzheu, Horng Wong, Matthieu Lequien, Sébastien Roux, Pierre Cahuc, and the seminar participants at the département des études économiques seminar at Insee, and at the Coin meeting. Access to some confidential data, on which this work is based, was made possible within a secure environment provided by CASD – Centre d'accès sécurisé aux données (Ref. 10.34724/CASD). The research and writing of this paper benefited from the monetary support of the following institutions: Agence Nationale de la Recherche (grant ANR-17-CE41-0009-01)

[†] Insee, damien.babet@insee.fr

[‡] Sciences-Po, olivier.godechot@sciencespo.fr

[§] Sciences-Po and INSEAD, marco.palladino@sciencespo.fr

Introduction

The Abowd, Kramarz and Margolis (1999) (hereafter AKM) model of log-wages with additive workers and firm fixed effects was estimated on French data: a panel sample of 1/24th of French wage earners (without 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 was available to researchers: this exhaustivity proved essential to the quality of the estimation for these models. We build such a dataset for France, to bring AKM back to the state of the art in its original land.

Wage inequality is a driving force of economic inequalities. Its rise for several decades in most rich countries is well documented[1].

Firms play a central role in driving these dynamics: in Germany (Card, Heining and Kline, 2013) and in the USA (Song et al., 2019) rising inequality comes in large part from between-firm inequalities. Both papers use AKM model to decompose log-wage variance into three main components: variance in individual workers heterogeneity, variance in firm premium, and covariance between the two, or *sorting*. In both cases, sorting explains a large share of the rise in wage inequalities. High-wage workers tend to work for high-wage firms, and increasingly so.

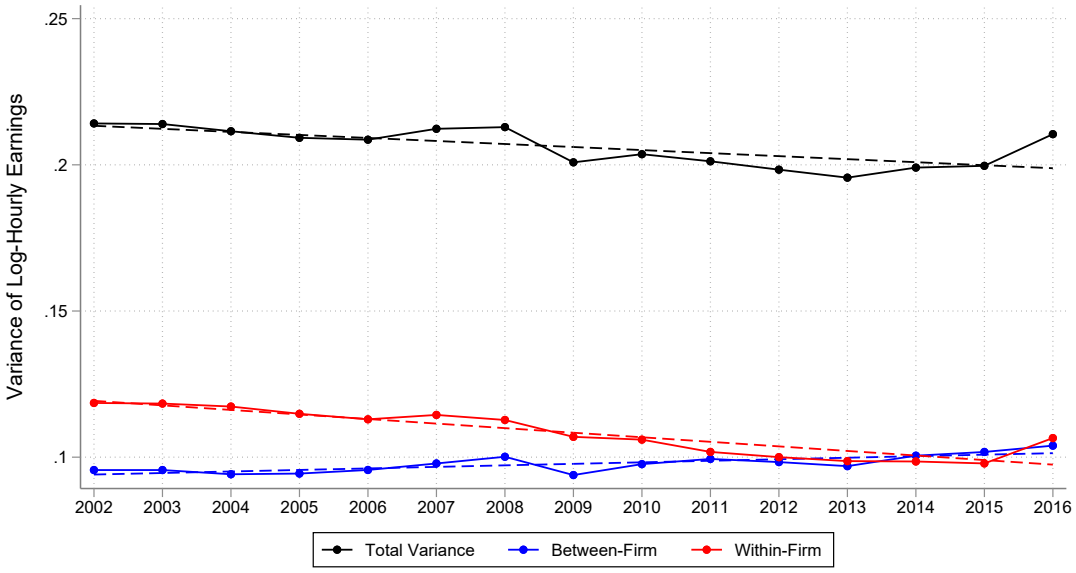
France is an interesting touchstone because it is an exception: wage inequality there has been stable or decreasing in the last decades. The same polarizing dynamics observed in other countries are still operating though, through an increase in between-firm inequalities (Figure 1).

This paper contributes to the literature in three main ways: we use a new dataset to provide first estimates for France, we introduce a new correction strategy for known biases, and we build fine descriptive statistics on firms and wages to analyze the observed dynamics and rule out some potential economic mechanisms.

First, we use a new, quasi-exhaustive matched employers/employees dataset for France first described in Godechot et al. (2020), thus allowing for the first decomposition of

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

Figure 1: Evolution of wage inequality - France



Note: This figure shows the evolution over time of the variance of log-earnings, the between-firm variance of log-earnings and the within-firm variance of log-earnings. 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.

log-wage variance and its evolution for France, between 2002 and 2016[2]. We show that sorting increased as in other countries. During the same period, within-firm inequalities decreased more than offsetting the rising between-firm inequality linked to sorting.

The measure of sorting through AKM models has however a well-known "limited mobility" bias described in Abowd et al. (2004), Andrews et al. (2008), and Bonhomme et al. (2020). This bias stems from the limited number of useful observations for each individual firm and worker parameters. Individual parameters estimations remain consistent, but the variance of the error term is underestimated. Both Card, Heining and Kline (2013) and Song et al. (2019) acknowledge this important bias in the measure of sorting, but expect it to be stable enough in time that it does not impact their dynamic results.

[2] The construction of the dataset is precisely described in annex C

Several correction strategies are available. Andrews et al. (2008) directly correct estimates with a bias correction factor derived from an estimate of the error term variance, with a hypothesis of homoscedasticity that is unrealistic because of 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 a much higher sorting than previous estimates, but while fixed effects models allow for further study of the distribution of this heterogeneity, it is much more difficult with random effects models. Bonhomme, Lamadon and Manresa (2019) cluster firms based on the distance 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, that allows for the estimation of richer models, including interaction and dynamic terms, at the price of the additional hypothesis that clusters are correctly identified. Kline, Saggio and Sølvsten (2020) leave-one-out strategy amounts to the bias correction factor method of Andrews et al. (2008) compatible with heteroscedasticity but is complex and computationally costly on large datasets. We use an analogous but much simpler split-sampling strategy by applying only one split to our data, rather than a leave-one-out method with as many splits as observations. The idea is that fixed effects are estimated independently in each split, so that covariance between firm effects estimated in one sample and workers effects estimated in the other is a debiased measure of sorting. Split-sampling has been used in similar settings by Chanut (2018), Drenik et al. (2020), Goldschmidt and Schmieder (2017), Gerard et al. (2018), Sorkin (2018) and Schoefer and Ziv (2021). We add to these works by generalizing the idea and providing a proof that, under reasonable hypothesis analogous to Kline, Saggio and Sølvsten, split-sampling corrects the limited mobility bias on quadratic terms. We also implement Bonhomme, Lamadon and Manresa (2019) cluster method (without random effects) and find the results coherent with split sampling.

Once the rise in sorting is robustly estimated, it remains to be explained. We show that most of the rise in sorting happened between firms, through changes in the population of firms. We observe a stronger growth in two dimensions: in workforce size for firms

with high premium and high-wage workers (workers with high individual fixed effects), and in the number of firms for firms with low premium and low-wage workers. These trends could reflect a structural evolution in the division of work between firms, arising through outsourcing dynamics - already outlined in other contexts - that expand to the public sector in France. We rule out alternative stories potentially related to the rise in between-firm inequalities. We do not find evidence of increasing return to skills and change in firms' rent-sharing behavior. We leverage further Bonhomme, Lamadon and Manresa (2019): by estimating a model with nonlinear interactions between workers and firms, we investigate the presence and the evolution of production complementarities (Shimer and Smith, 2000; Eeckhout and Kircher, 2011) in France, with the idea that the rise in sorting could reflect a productivity-increasing rise in the quality of matching. We observe that the addition of the idiosyncratic match component of wages does not affect our baseline results: sorting is not explained away when production complementarities are considered.

During the same period, wages at the bottom have grown faster than middle and top percentiles. We provide suggestive evidence that these dynamics are associated to French labor market institutions' nature, reinforcing the findings of recent literature (Bozio, Breda and Guillot, 2020; Kramarz et al., 2021). The mix of a significant increase in the minimum wage and the elimination of all employer-paid payroll taxes around it has proven successful at increasing redistribution and balancing the contemporaneous rise in between-firm inequalities.

Section 1 details how we built our data. Section 2 introduces the methods: variance decomposition of the log-wage stemming from the AKM model and limited bias correction through split-sampling or clustering. We present results in Section 3 and discuss our methods and findings in Section 4.

1 Data

1.1 Building an exhaustive pseudo-panel

We use DADS data, exhaustive yearly files built from tax returns files by firms on their payrolled employees. This is the source for French official statistics on wages evolution. It allows a first decomposition of log-earnings variance in between and within-firm components (Figure 1 above).

We want to use this data in panel form. For workers, the data is pseudonymous with an individual identifying code that changes every year, allowing for cross-section use of the file, but not for long panel use. Panel analysis on French matched employers-employees wage data are traditionally done on the "DADS panel" or "all wage-earners panel", which we dub here the "narrow panel". This panel is built on a sample of 1/24 before 2002 and 1/12 after, sampling the same individuals as a permanent demographic panel, allowing matching. The sampling also allows for additional data quality control and correction work that would be much heavier on 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), Coudin, Maillard and Tô (2018) or Palladino, Roulet and Stabile (2020).

Since 1999, though, AKM models have been better estimated in countries where researchers have had access to exhaustive panel data: USA, Germany, Sweden, Austria, Italy, Norway, Denmark, etc., and for good reason. As in any estimation, the reduction and the sample size raises uncertainty, so that, for instance, firms in the narrow panel need to be roughly 12 times bigger to have their firm fixed effect estimated with the same precision as in the exhaustive data. The decrease in precision translates into an even larger "limited mobility bias" for variance and sorting estimates. AKM identification also relies on mobile workers moving between firms, and is only possible conditional on the group of firms interconnected through such workers. Sampling drastically reduces the proportion of firms belonging to the main connected component, and further reinforces the selection of bigger and more connected firms in the

estimation sample.

Each yearly DADS file for a given year y is also a short panel, with most job variables given both for the current year $t = y$ as well as for the previous year $t - 1$. Direct use of this data as short, two years panel data is possible, but this overlap also allows for matching between yearly files, based on common information (establishment ID, gender, number of hours, job duration in days, start and end dates of the job, municipality of work and residence, earnings and age) between year t of yearfile y–1 and year t–1 of yearfile y. Between 2002 and 2016, matching gives a single match to 98% of the individuals. Matching misses include specific situations where all matching variables are identical for several individuals (such as higher education institutions for civil servants for instance), and rare instances of individual data modifications between one yearly file and the $t - 1$ values of the following year.

By construction, even if matching were perfect, employment spell before and after a career interruption of more than a year cannot be connected to the same individual. Employees who are not matched either due to career interruption or to matching misses still are in the panel, but they appear under multiple ident numbers. We dub the resulting almost exhaustive pseudo panel the "wide panel".

This matching procedure is of no use before 2002. Up to 2001, the various jobs of a single individual were not linked in the files by a corresponding individual ident number. It is not possible to follow one worker through different employers, even in the course of one year. Matching is possible, but only for workers who kept the same job for two years to be matched. Estimation of AKM relies on the different wages a given individual can earn when working for different employers. It cannot be done before 2002[3].

We computed long-term series on sorting from 1976 to 2016, relying on the narrow panel, where information on wages and careers is available since 1976, with some missing years (1981, 1983, 1990) and varying data quality[4].

We mobilize another data source. Exhaustive firm financial data from administrative

[3] the matching procedure is detailed in Section C in annex

[4] Historical series in annex, Figure A1

sources (FICUS/FARE files) matched to our wage files provide value-added per worker and total workforce, irrespective of the sample restriction we use. Value-added and other accounting variables are defined and measured at the legal unit level, not the establishment 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.

1.2 Sample restrictions

Following similar works, we exclude public workers, both for comparability and data quality reasons[6]. We restrict to ordinary jobs, excluding subsidized contracts, interns and apprenticeships. We include both men and women.

We divide the data into three adjacent five-years periods[7]: 2002-2006, 2007-2011, and 2012-2016. 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). For simplicity, we sometimes call such observations a "wage", or a "job". Each worker can appear up to 5 times during each of the sample periods. For individual workers effects, these rather short panels obviously translates into noisy estimations, which increases the need for bias correction.

We have information about the number of hours worked, which is rare in this kind of data. Without it, it is common practice in the literature to set a minimal wage for inclusion and exclude women to reduce the risk of misidentifying part-time workers. We can avoid these exclusions. Our target variable of earnings is the log hourly wage. We restrain the sample to people employed for the full year, so as to limit the impact of annualized payments, and because for people with spells of unemployment or in-

[5] We follow in this other applications of the AKM method on French data. More recently, Insee has started to provide datasets of groups, built from financial links between legal units. There is not enough historical depth yet to measure evolution at this observational unit level

[6] Most notably, public servants are not included before 2009

[7] Card, Heining and Kline (2013) divide their 1985-2009 data into four overlapping seven-years panels, and Song et al. (2019) divide their 1980-2013 data into five adjacent seven-years panels. We chose shorter periods to adapt to our shorter overall panel, but also because the correction for the limited mobility bias diminishes the need for long panels

activity, the total annual number of hours worked (as well as the duration in days) is not entirely reliable and computation of hourly wage lacks precision[8]. We exclude jobs with an hourly wage inferior to 80% of the legal minimum hourly wage for the corresponding year, or above 1000 times the minimum hourly wage, and observations with missing values for sex, age and employer. All restrictions are done after matching, when the wide panel is already constructed. They select specific observations but not individuals, who remain in the wide panel as long as they have been working during the year (or received unemployment benefits).

AKM models are performed on connected sets of firms and workers. Two sets of firms are disconnected if there is not any worker in the data who worked for two different firms, one in each group. A connected set then includes all workers who ever worked in the set's firms. We compute the connected sets and perform all estimations on the main connected set for each period. Table 1 reports descriptive information characterizing both the full population and the largest connected set.

In our historical series computed on the small panel, work hours are not reliable before 1996, but job duration in days and an indicator variable for part-time jobs exist since 1976. In our long-term series we also compute sorting on the daily wage of full-time workers. The level is lower than hourly wage small panel sorting, but the trends are mostly parallel for the post-1996 period where both are known. Other changes in variables in the 40 years period also preclude an exact reconstruction of the selection we choose on the wide panel, notably because the distinction between public and private sector is not consistent.

[8] We found similar results when widening the selection to all individual whose main job during the year lasts more than 90 days. In this larger sample the connectivity is better and the limited mobility bias is reduced

Table 1: Summary statistics for overall sample and individuals and firms in largest connected set

				Log-Hourly Wage	
	Person/yr	Individuals	Firms	Mean	Std.Dev.
	Overall Sample				
2002-2006	47,095,163	17,008,450	884,179	2.66	0.46
2007-2011	51,415,168	17,479,889	1,075,512	2.78	0.46
2012-2016	54,478,051	17,333,027	1,130,560	2.86	0.45
	Largest Connected Set				
2002-2006	41,703,340	14,904,618	367,257	2.68	0.46
2007-2011	44,733,304	14,957,510	412,817	2.81	0.46
2012-2016	47,038,310	14,894,154	394,969	2.89	0.45

Note: 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.

2 Methodology

2.1 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 $i = 1, 2, ..., N$ during year $t = 1, ..., T$, demeaned by the average log-hourly wage for all workers during year t so that $\bar{y}_t = 0$[9]. Time-varying covariates x_{it} are limited to age and age squared. θ_i is the fixed effect of individual worker i, and ψ_j is the fixed effect of individual firm $j = 1, 2..., J$, both supposed constant in time for the duration of the panel, firm $j(i,t)$

[9] We also checked with a model inspired by Card et al. (2018), where log-wages are not centered, years are right-hand explanatory variables and age is included as a cubic polynomial constrained to be flat at 40 to avoid colinearity. Results in table A6 are generally robust to this change in specification but show a lower rise in sorting. Figures A8 and A10 suggests this specification is more sensitive to firm fixed effect changes during the great recession

being the employer of worker i during year t. u_{it} is the idiosyncratic error term. We further note $\mathbf{F} = (\mathbf{1}_{j=j(i,t)})$ the $N^* \times J$ matrix of the bipartite graph of workers/firms connections through time, with $N^* = NT$.

This model rests on two notable hypotheses:

- No interaction effect between firm type and worker type: the fixed effects are additive (in the log-wage). We suppose the firm specific wage premium will be the same for all workers, men and women, young and old, skilled or not.
- Exogenous mobility: the residual term u_{it} has null expectation conditional on the variables x_{it}, i, t and j, as is classical, but also conditional on the matrix $\mathbf{F}$. This means in particular that wages before or after a job change are on average the same as if there had been no job change.

Although both hypotheses can appear unrealistic and have been subjected to scrutiny, they seem to provide reasonable approximations. Bonhomme, Lamadon and Manresa (2019) build a model that allows for interaction and find only slight departures to the additive linear model. We replicate their model on our data and reach the same conclusion. Di Addario et al. (2021) develop an extension of the two-way fixed effects model à la AKM with two firm fixed effects. They add to the fixed effect for the "destination" firm hiring the worker (the classic firm fixed effect) a fixed effect for the "origin" firm, reflecting the wage level necessary to "poach" a worker from a given firm. They found that destination effects are more than 13 times as variable as origin effects across firms, implying that a more dynamic specification does not increase much the model explanatory power. Another potential specification error results from the possible evolution of "fixed" effects. Firm premium might change with time, and workers age profile are heterogeneous, as shown in the French case in Magnac and Roux (2021) on DADS data. Formally, this simply contradicts the hypothesis of a null conditional expectation for the residual term. Regarding year to year variations in firm premium at least, Engbom, Moser and Sauermann (2022) and Lachowska et al. (2020) provide some reassurance. We analyze this source of heterogeneity in annex E.3[10]. Still,

[10] More generally, de Chaisemartin and d'Haultfoeuille (2020) and de Chaisemartin and D'Haultfoeuille

we find signs that changes in firm premium with time and age composition evolution of the population of workers might affect our results.

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 five-years periods: 2002-2006, 2007-2011 and 2012-2016. Like them, we also assume the AKM model to correctly describe wages and use it as the basis for variance decomposition. Ignoring for simplicity of exposition the time-varying workers variables x_{it}, we can describe for each period a decomposition of $V(y)$ 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 distinguishes within-firms and between-firms components of wage variance, and extend the law of total variance $V(y) = E[V(y|j)] + V[E(y|j)]$ 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. Our interest lies first with the evolution of the sorting component of this decomposition, $2Cov(\theta, \phi)$, which is by construction entirely contained in the between-firm component of wage variance. We further construct a measure of segregation following Song et al. (2019).

(2022) raised an alarm about the consequences of heterogeneity in panel treatment effects or, in our case, fixed effects. The problem does apply here if the AKM model is ill specified, and we are not aware of any existing solution in this setting.

$$\text{Segregation Index} : \frac{Var(\bar{\theta}_j)}{Var(\theta_i)} \tag{5}$$

Segregation captures the extent to which high-wage workers tend to work with one another, and low-wage workers with one another. Like other dimensions of segregation (residential, school, etc.), this has a strong impact on social mixing, but no direct impact on overall wages inequality, since the increased between-firm variance stems from a decrease in within-firm variance, leaving the overall distribution unchanged.

2.3 Limited mobility bias

We follow Kline, Saggio and Sølvsten (2020) for the description of this bias. They provide a simple framework that neatly generalizes on any quadratic form of the estimated parameters. We start with a simplified notation of our linear model with only one parameter vector and one observation index i (replacing the (i,t) couple in the detailed model):

$$y_i = z_i'\alpha + u_i \tag{6}$$

With $\alpha = (\beta, \theta, \psi)$ our parameter vector of length $k = 2 + N + J$ and z_i the non-random regressors vector of the (worker * year) i's observation characteristics, including the indicator vector for worker and firm. We note $S_{zz} = \sum_{i=1}^{N^*} z_i z_i'$ the design matrix (with full rank when we limit the sample to the main connected set). Our objects of interest are (weighted) variances and covariances of parts of the α vector and can be described as quadratic forms $\omega = \alpha' A \alpha$ for a chosen symetric matrix $A \in \mathbf{R}^{k \times k}$. We can choose A so as to compute the quantities studied here, weighted by the number of worker-year observations[11].

Our naive plug-in estimator for ω is thus $\hat{\omega}^{PI} = \hat{\alpha}' A \hat{\alpha}$ with $\hat{\alpha}$ an OLS estimate $\hat{\alpha} = {S_{zz}}^{-1} \sum_{i=1}^{N^*} z_i y_i = \alpha + {S_{zz}}^{-1} \sum_{i=1}^{N^*} z_i u_i$. $\hat{\alpha}$ contains $\hat{\theta}$ and $\hat{\psi}$ which we later name *WFE*

[11] For instance, var(WFE) is computed with a matrix A filled with 0 except for the $N \times N$ square corresponding to the N WFE estimates of the parameter vector, which we fill with a generic term $-1/N$ and a diagonal term $1 - 1/N$. Covariances objects are built with analogous matrices, with for instance a $1 - m_j$ term for the $(2 + i, 2 + N + j)$ and $(2 + N + j, 2 + i)$ positions in the matrix if the worker i is in firm j and a generic $-m_j$ term if the worker i is not working in firm j

(for workers fixed effects) and *FFE* (for firms fixed effects). The estimation error in $\hat{\alpha}$ will result in a systematic bias in $\hat{\omega}^{PI}$ equal to a linear combination of the unknown and possibly heteroscedastic variances σ_i^2 of the error terms u_i. From classic results on quadratic forms, Kline, Saggio and Sølvsten deduce :

$$\mathbf{E}[\hat{\omega}^{PI}] - \omega = trace(A\mathbf{V}[\hat{\alpha}]) = \sum_{i=1}^{N^*} B_{ii}\sigma_i^2 \tag{7}$$

With $B_{ii} = z_i' S_{zz}^{-1} A S_{zz}^{-1} z_i$ representing the influence of each (squared) error term on the plug-in estimator. This bias exists for all linear models, but usually for a small parameter dimension k the S^{-2} term insures relatively fast convergence. Here however k is large, and so is, potentially, B_{ii}. Moreover the complex structure of the design matrix, reflecting the complex network of worker / firm connections, is present both in matrices S and A when computing $cov(\theta, \psi)$, leaving way to even stronger bias.

This expression for the bias leads Kline, Saggio and Sølvsten to an obvious correction strategy: estimating the σ_i^2 error terms, and thus the bias itself. This can be done with a leave-one-out strategy that is computationally costly, adding a factor of the order $N^* = NT$ to the computation. Kline, Saggio and Sølvsten provide a more tractable estimation method through a high number of random projections (in the hundreds). Bonhomme et al. (2020) still finds the method demanding and further approximate it, though they worry the succession of approximate estimations (with those usual in AKM models) might have consequences that are not well understood.

For purely computational reasons, we favour a split-sampling strategy that only demands two estimations on two half-samples, at worst doubling computing time.

2.3.1 Split-sampling bias correction

With $\hat{\omega}^{SP}$ the split-sampling estimate of any quadratic form ω of the parameters, we show in annex D that the bias is

$$\mathbf{E}[\hat{\omega}^{SP}] - \omega = trace(AS_{zz,1}^{-1}\mathbf{E}(B)(S_{zz,0}^{-1})') \tag{8}$$

With the indices 1 and 0 indexing the two split samples I_1 and I_0 and B the matrix with the generic term:

$$b_{lm} = \sum_{i \in I_1}^{N_1} u_i z_{l,i} \sum_{j \in I_0}^{N_0} u_j z_{m,j}$$

This term has null expectation under mild conditions: 1. null conditional expectation $\mathbf{E}[u|z] = 0$ and 2. independence of $u_i, i \in I_1$ and $u_j, j \in I_0$. If the variance-covariance matrix of u is diagonal, the bias disappears whatever the matrices A and S_i might be. The second condition might be violated, if for instance u_i are correlated for different years of the same employer / employee pair, which is likely. We do not derive here the complete conditions for consistency, nor the convergence rate when the number of parameters rises with the number of observations. We follow in this most of the AKM decomposition literature, which implicitly rely on the big size of the data used. However consistency and convergence depend on the variance of error terms but also on relations between the quadratic form matrix A and the design matrix S, hence on the mobility network. We refer to the proofs and discussion in Kline, Saggio and Sølvsten in the different but related leave-one-out context.

Of course there is an additional cost of split-sampling in increased uncertainty coming from the reduced effective sample size ($S_{zz,s}$ has half the observations of S_{zz}). With our data, we observe that this uncertainty is small compared to the size of the bias reduction effect. One can reduce this uncertainty by repeatedly estimating the quadratic form through split-sampling and averaging the results. The procedure reaches arbitrary precision, only limited by the computational cost. We also report Monte Carlo experiment results and standard deviations computed on multiple random splits in annex D, that confirms the stability of the procedure.

The main limitation of split sampling is the impact of the split on the bipartite graph and its main connected set, and the sample-splitting strategy has to be considered in this regard. In each split sample, the main connected set is smaller than in the original sample and both are distinct, so that the common sample of workers and

firms belonging to the main connected set in both split samples is reduced, and so is the corresponding parameter vector of individual effects.

The most simple split strategy is a direct random split of observations in two equally sized samples. By balancing the sampling by worker, splitting for each worker the periods of observation, one increases the odds that each worker is present in both samples' main connected set. We dub this method "period splitting". On the contrary, by splitting individuals rather than observations, one increases the connectivity in each set (because individual careers are kept intact), but each worker's fixed effect is estimated only once: one looses the capacity to correct the $var(\theta)$ and $var(u)$ quadratic forms through split sampling. If this splitting of individuals is balanced by firm, it increases the odds that each firm's fixed effect is estimated in each sample. We dub this method "firm splitting"[12]. With 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.

Period splitting might not completely correct the limited mobility bias for the reason mentioned above: it is likely that u_i are correlated for several observations of the same employer / employee pair. This problem is attenuated by the specifics of our setting. Because we keep only observations with full year jobs, movers are generally observed only four years or less among the five in the panel. It is unlikely that after the random split of these observations, an individual worker would remain a mover in both samples. Because the estimation relies exclusively on movers, a given individual residuals would generally not be correlated to errors on both sides of the split. Still, by keeping all observations of one individual on one side of the split, the firm splitting method avoids entirely this drawback. Consequently, we favor firm splitting to compute a debiased sorting effect, but use period splitting to compute the complete variance decomposition.

[12] Chanut (2018) provides a description of this split sampling strategy to correct the limited mobility bias. He describes a way to compute such a split, uses this method on the French narrow panel and shows, on a toy example, that it succeeds in correcting the bias.

2.3.2 Firm-clustering

We also implemented Bonhomme, Lamadon and Manresa (2019) strategy. We ran a firm-clustering algorithm with 5000 clusters (around 1% of the number of firms in 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 mobility network between clusters is very dense and each cluster's fixed effect estimate has very low variance, thus correcting the limited mobility bias. 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[13]. Even so, it remains plausible that the segregation of workers could bias the clustering, with firms being clustered based on some combination of their own fixed effects and their average workers' fixed effects. An AKM estimation following this procedure would then show higher sorting, and lower cluster effect variance, than is really the case. Bonhomme, Lamadon and Manresa (2019) aknowledge the risk, mention job-market models that satisfy the conditions for cluster identification[14], and provide in-depth robustness analysis that suggests it is of limited impact in practice.

3 Results

3.1 A robust rise in sorting

In Table 2, the two variance decompositions 2 and 4 are applied to our three periods. We use the classic AKM estimates of θ and ψ fixed effects as our baseline estimates, before applying any of the corrections discussed above[15]. Overall wage inequalities in the private sector, measured as log-hourly wage variance, went down slightly be-

[13] 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

[14] "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)

[15] We present this initial result first for simplicity and because bias correction gives qualitatively similar results

tween 2002 and 2016 in France. The decomposition of the variance shows that while the within-firm log-hourly wage variance was also diminishing, between-firm wage inequalities rose during the same period. They accounted for 42% of total log-hourly wage variance in 2002-2006, and 48% in 2012-2016. Log-hourly wage variance was lower by 3.7% in the third period compared to the first: within-firm inequalities accounted for 204% of this evolution, and between-firm inequalities for -104%. France's diminishing wages inequalities during the period are atypical among developed countries, but the rise of between-firm wage inequalities matches the results of Song et al.. for the US from 1978 to 2013.

This rise is robust to the various checks we conducted, most notably to the correction of the limited mobility bias, which appears very important, even when restricting to firms with more than 20 observations per year. Figure 2 reports the results coming from the principal corrections discussed in Sections 2.3.1 and 2.3.2[16]. All corrections push upward the estimates of sorting without changing qualitatively the trends over time.

The intensity of the rise is lower in corrected estimates, suggesting the limited mobility bias decreased overall on the period, possibly due to increased connectivity of the firm-workers network. This evolution shows that a perfect stability of the limited mobility bias cannot be assumed, and that the dynamics of sorting are best measured based on corrected estimates.

Split sampling correction behaves as expected, with period-splitting showing signs of an incomplete correction of the bias compared to firm-splitting. The firm-clustering method gives results reasonably close to the split sampling, reproducing results from Bonhomme et al. (2020) when comparing their clustering and random effect model to the Kline, Saggio and Sølvsten (2020) leave one out method.

Long-term series are imperfect, as is clear when compared to exhaustive data estimates on recent years, but they might provide some indication of past trends[17]. They suggest

[16] see tables A1, A3, A4 and A5 for the full decomposition with the different corrections.

[17] Historical series are presented together with corrected and uncorrected sorting results in Figure A1. The lowest curve shows uncorrected sorting as measured on the narrow panel, the very data used in the original Abowd, Kramarz and Margolis (1999) paper, but for subsequent years.

Table 2: Decomposition of wage variance and its evolution - uncorrected AKM1+

		2002-2006		2007-2011		2012-2016		Change from 2002-2006 to 2012-2016	
		Comp.	Share	Comp.	Share	Comp.	Share	Comp.	Share
Total variance	Var(y)	0.214		0.211		0.207		-0.008	
	Var (θ)	0.165	77.1	0.166	78.8	0.160	77.6	-0.005	63.8
	Var (ψ)	0.030	14.0	0.029	14.0	0.025	12.3	-0.005	60.8
	Var(Xb)	0.024	11.1	0.034	16.0	0.016	7.9	-0.008	99.5
	Var(u)	0.009	4.1	0.008	4.0	0.007	3.3	-0.002	28.2
	2*Cov($\theta.\psi$)	-0.004	-1.8	0.000	-0.2	0.004	1.7	0.007	-98.9
	2*Cov($\theta.Xb$)	-0.012	-5.7	-0.029	-13.7	-0.007	-3.3	0.005	-69.7
	2*Cov($\psi.Xb$)	0.002	1.2	0.002	1.1	0.001	0.6	-0.001	16.2
Between-firm variance	Var($\bar{y}$)	0.091	42.2	0.095	45.2	0.099	47.9	0.009	-113.2
	Var ($\bar{\theta}$)	0.058	27.2	0.063	29.7	0.067	32.2	0.008	-110.5
	Var (ψ)	0.030	14.0	0.029	14.0	0.025	12.3	-0.005	60.8
	Var($\bar{X}B$)	0.004	1.8	0.006	2.6	0.003	1.3	-0.001	16.9
	2*Cov($\bar{\theta}.\psi$)	-0.004	-1.8	0.000	-0.2	0.004	1.7	0.007	-98.9
	2*Cov($\bar{\theta}.\bar{X}B$)	0.000	-0.1	-0.004	-2.0	0.000	-0.2	0.000	2.3
	2*Cov($\psi.\bar{X}B$)	0.002	1.2	0.002	1.1	0.001	0.6	-0.001	16.2
Within-firm variance	Var($y - \bar{y}$)	0.124	57.8	0.115	54.8	0.108	52.1	-0.016	213.2
	Var ($\theta - \bar{\theta}$)	0.107	49.9	0.103	49.1	0.094	45.3	-0.013	174.3
	Var($Xb - \bar{X}b$)	0.020	9.3	0.028	13.4	0.014	6.6	-0.006	82.7
	Var(u)	0.009	4.1	0.008	4.0	0.007	3.3	-0.002	28.2
	2*Cov($\theta - \bar{\theta}.Xb - \bar{X}b$)	-0.012	-5.6	-0.025	-11.7	-0.006	-3.1	0.005	-72.0
	2*Cov($\theta - \bar{\theta}$, u)	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0
	2*Cov($Xb - \bar{X}b$, u)	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0
Segregation	$\frac{Var(\bar{\theta}_j)}{Var(\theta_i)}$	0.351		0.379		0.419		0.068	
N* (largest connected set)		41,703,340		44,733,304		47,038,310			

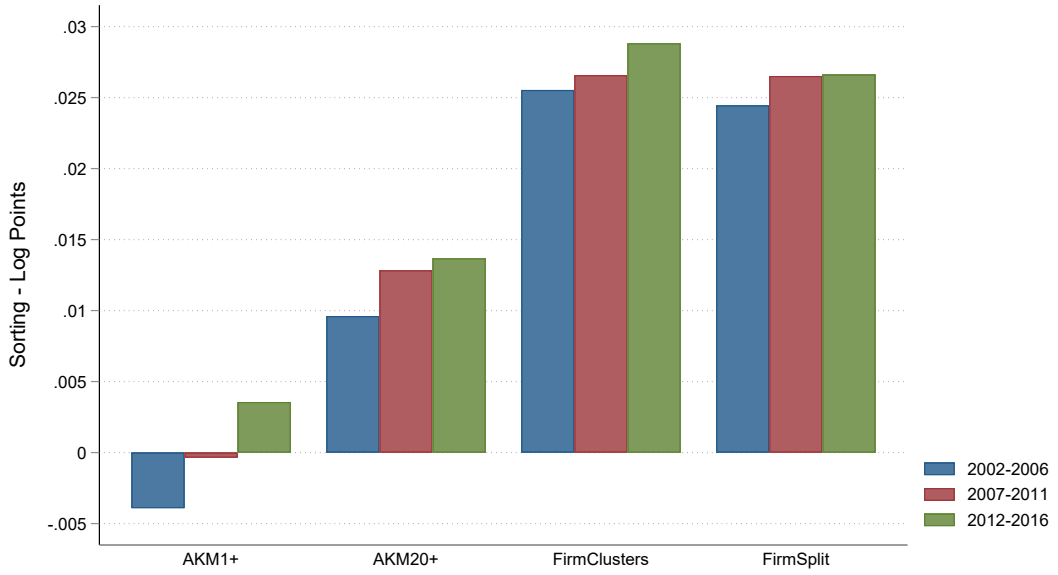
Note: Only individuals employed for at least 360 days by the same firm during the year are included for a given year. Individuals and firms in public administration are not included. Decomposition based on Equations 2, 3 and 4, weighted by worker-year observations. The largest connected set entails the group of firms connected by worker mobility.

that most of the rise in sorting actually predates our main period of study, and show fluctuations that are evocative of pro-cyclicality.

3.2 Explaining the rise in between-firms inequalities

In this section, we test which factors are behind the empirical trends we find, particularly the rise in sorting. Like variance or mean, sorting is fundamentally a distributional statistic. It has no meaning at the individual firm level, and cannot be studied

Figure 2: Sorting (2*Cov(θ.ψ)) over time - Baseline and selected correction strategies



Note: 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. This figure reports the estimates of sorting - by period - coming from the different strategies described in Sections 2.3.1 and 2.3.2.

directly as a characteristic regressed on explanatory variables. We also move beyond variance and covariance and study other, non-quadratic distributional statistics of fixed effects[18]. For this reason, we use uncorrected estimates in this section[19]. We examine several empirical arguments to sort through potential economic mechanisms. We find that the rise in sorting is mostly accounted for by composition effects in the population of firms: firm destruction, creation and growth favored sorting.

3.2.1 Firm demography and composition effects

We get a first look at composition effects with a simple partition of all firms in the first and third periods into four categories. Fixed effects estimates are centered on

[18] We also experimented with panel regressions of firm-level workers and firms fixed effects on firm-level variables, followed by period specific covariance decomposition. The method is complex and generates numerous, hard to interpret interaction terms. Even so, most of the rise in sorting was accounted for by changes in the size and population of firms with given fixed effects, rather than in firm specific changes in fixed effects. We focus here on simpler results conveying the same message

[19] Other correction strategy might be available in this context, such as Jochmans and Weidner (2021). The split-sampling method could also possibly be adapted. We leave these corrections for future research.

0 for each period, and we partition firms depending on the sign of their estimated fixed effect (averaged over the two periods for firms present in both periods) and the sign of the mean workers' fixed effect of their employees (similarly averaged). The partition distinguishes four "quadrants": covariance increasing quadrants (high wage firms with high wage workers and low wage firms with low wage workers) and covariance decreasing quadrants (low wage firms with high wage workers and high wage firms with low wage workers).

Table 3 additionally differentiates between firms that are present in both periods and firms present either in 2002-2006 or in 2012-2016. The first column captures the number of workers where both average worker (W) and firm fixed effects (F) are above zero (F>0, W>0)[20]. Other columns map the rest of the possible combinations for F and W values. Stayers firms grew in size, remarkably so in the area where both FEs are positive. 1-period firms are much more concentrated in period 3 in positive sorting areas, particularly where both fixed effects are negative. There were in period 1 almost 4 millions jobs in firms from this quadrant that did not survive to period 3, but there were 6.3 millions jobs there in new firms in period 3, the highest growth in person-year terms in table 3. In other words, even if no worker changed job between the two periods (except from firm destruction and creation), and even if fixed effects were the same in both periods, sorting would have increased simply because the number and size of firms increased more in the sorting groups than in the non-sorting groups[21].

Goldschmidt and Schmieder (2017) show for Germany that the rise in sorting is explained by outsourcing of low productivity activities. Our composition results are compatible with such a mechanism, and more generally with any dynamics of firm specialization taking place through firm creation, destruction and growth, rather than internal changes. Bilal and Lhuillier (2021) estimates on French data a structural model where the gains of outsourcing come from savings on rent-sharing that high-productivity firms realize on outsourced workers. They also estimate that outsourcing

[20] Worker and firm fixed effects are normalized so they are comparable across periods

[21] This is not due to a general change in size distribution. A significant shift in the size distribution over time could explain the rise in between-firm inequalities. Figure A2 plots the fraction of firms below a given size, by period (2002-2006 vs 2012-2016). The two distributions overlap almost perfectly.

Table 3: Composition effects : firm demographics and mean fixed effects by sorting quadrant

	Person-Year Observations				Mean Worker Fixed Effects				Mean Firm Fixed Effects			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	F>0, W>0	***F>0, W<0***	***F<0, W<0***	***F<0, W>0***	***F>0, W>0***	***F>0, W<0***	***F<0, W<0***	***F<0, W>0***	***F>0, W>0***	***F>0, W<0***	***F<0, W<0***	***F<0, W>0***
Firms present in both periods (2-periods)												
2002-2006	9,676,361	6,925,595	8,459,452	3,850,883	0.2184	-0.1398	-0.1643	0.1690	0.1181	0.0827	-0.0909	-0.1116
2012-2016	11,100,000	7,216,232	9,785,778	4,416,734	0.2302	-0.1327	-0.1617	0.1735	0.1196	0.0798	-0.0936	-0.1128
Diff	1,423,639	290,637	1,326,326	565,851	0.0118	0.0071	0.0026	0.0044	0.0015	-0.0029	-0.0027	-0.0012
Firms present only in one period (1-period)												
	F>0, W>0	***F>0, W<0***	***F<0, W<0***	***F<0, W>0***	***F>0, W>0***	***F>0, W<0***	***F<0, W<0***	***F<0, W>0***	***F>0, W>0***	***F>0, W<0***	***F<0, W<0***	***F<0, W>0***
2002-2006	2,571,145	3,341,323	3,979,486	2,898,952	0.2402	-0.2174	-0.1745	0.2162	0.1280	0.1294	-0.1191	-0.2021
2012-2016	2,712,241	2,685,078	6,320,317	2,996,942	0.2488	-0.2375	-0.1872	0.2120	0.1033	0.1277	-0.1134	-0.1957
Diff	141,096	-656,245	2,340,831	97,990	0.0086	-0.0201	-0.0127	-0.0042	-0.0247	-0.0016	0.0057	0.0064

Note: 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 firms present in both periods and present either in 2002-2006 or in 2012-2016 are analyzed separately. Firms are further divided into four quadrants according to the value of the estimated worker (W) and firm fixed effects (F) (see Equation 1), averaged over both periods for staying firms. Worker and firm fixed effects are normalized so they are comparable across periods. Columns 1, 5, and 9 refer to the number of workers where both worker (W) and firm fixed effects (F) are above zero (F>0, W>0). The other columns map the rest of the possible combinations for F and W values. For firms present in both periods, the (not employment weighted) 2-periods average of firm and worker fixed effects are considered in the allocation to the four quadrants. Entries in columns 5-8 and 9-12 are the average worker and firm fixed effects, respectively, by quadrant.

spending rose in France from 6% of the aggregate wage bill in 1996 to 11% in 2007 and 19% in 2015. Both papers define contractor firms and outsourcing events from detailed firm's industry, occupations and joint mobility of clusters of workers. Bilal and Lhuillier (2021) additionally use a firm survey (EAE) on intermediate inputs to measure expenditure on external workers.

We also use the four quadrants to describe the evolution of fixed effects themselves for a given group of firms between the periods. The change in the mean workers fixed effects for a firm between 2002 and 2016 might come from workers' turnover, or from a change in the WFE estimation of a given worker in two different panels, like the change in FFE for a given firm. Columns 5 to 12 in table 3 does not show any strong pattern of increasing sorting intensity within quadrants.

3.2.2 The central role of occupations

Table 4 presents simple descriptive data that help to investigate the role of occupations. We use the fixed effect quadrants already mentioned and map the evolution of the occupational structure for the three groups with the most important evolution in size: surviving firms with either high premiums and high wage workers or low premium and low wage workers, and 1-period firms with low premium, low-wage workers. The group of high wage, stayer firms have more managers and engineers in period 3 than in period 1, and fewer blue collar workers and associate professionals. Low wage, stayer firms have a very stable occupational structure, but the little evolution that exists mirrors, in reverse, the previous group: the correlation between the two evolution columns ("Diff") is -0.8. Low wage firms observed only once are very different in the first and the last period, with a strong decrease in the share of industrial occupations and associate professionals, and a strong increase in the share of health and social workers, personal services employees and civil servant. We consider only the private sector. In the private sector, these last occupational categories include mostly previously public activities, such as postal services or private employers of mostly public

Table 4: Change in the occupational structure by type of firm

		2-periods, F>0, W>0			2-periods, F<0, W<0			1-period, F<0, W<0		
Occupation	Code	2002-2006	2012-2016	Diff	2002-2006	2012-2016	Diff	2002-2006	2012-2016	Diff
Entrepreneurs	20	0.0057	0.0063	0.0006	0.0065	0.0070	0.0004	0.0087	0.0063	-0.0024
Professionals	32	0.0049	0.0043	-0.0006	0.0028	0.0023	-0.0006	0.0025	0.0245	0.0219
Artists and media professionals	35	0.0107	0.0104	-0.0002	0.0005	0.0008	0.0003	0.0014	0.0023	0.0010
Managers	37	0.1601	0.1999	0.0399	0.0415	0.0335	-0.0080	0.0258	0.0249	-0.0009
Engineers	38	0.1533	0.1998	0.0465	0.0187	0.0160	-0.0026	0.0149	0.0071	-0.0078
Primary school teachers	42	0.0030	0.0029	-0.0001	0.0026	0.0021	-0.0005	0.0021	0.0206	0.0184
Health and social workers	43	0.0040	0.0046	0.0006	0.0212	0.0174	-0.0038	0.0100	0.1336	0.1236
Public administration intermediates	45	0.0001	0.0003	0.0002	0.0002	0.0000	-0.0001	0.0016	0.0164	0.0148
Business administration intermediates	46	0.1758	0.1114	-0.0644	0.0648	0.0894	0.0246	0.0851	0.0342	-0.0509
Technicians	47	0.0997	0.0995	-0.0002	0.0273	0.0233	-0.0041	0.0233	0.0157	-0.0076
Intermediate supervisors	48	0.0362	0.0315	-0.0047	0.0242	0.0285	0.0042	0.0304	0.0096	-0.0209
Public administration clerks	52	0.0015	0.0019	0.0004	0.0344	0.0265	-0.0079	0.0243	0.1248	0.1006
Security agents	53	0.0026	0.0035	0.0009	0.0258	0.0192	-0.0066	0.0249	0.0154	-0.0094
Business administration clerks	54	0.0988	0.1112	0.0124	0.0905	0.0854	-0.0051	0.0963	0.0821	-0.0143
Retail salespersons	55	0.0073	0.0263	0.0191	0.1617	0.1514	-0.0103	0.1088	0.0705	-0.0383
Personal service employees	56	0.0034	0.0034	0.0000	0.0570	0.0510	-0.0060	0.0532	0.1597	0.1066
Skilled manufacturing workers	62	0.1330	0.0959	-0.0371	0.0791	0.0949	0.0158	0.1126	0.0282	-0.0843
Skilled artisans	63	0.0199	0.0230	0.0031	0.0608	0.0625	0.0017	0.0849	0.0512	-0.0337
Drivers	64	0.0085	0.0062	-0.0023	0.1044	0.0986	-0.0059	0.0917	0.0756	-0.0162
Handling, transport skilled workers	65	0.0226	0.0209	-0.0016	0.0348	0.0353	0.0005	0.0295	0.0102	-0.0193
Unskilled manufacturing workers	67	0.0437	0.0306	-0.0132	0.0611	0.0781	0.0169	0.0877	0.0337	-0.0541
Unskilled artisans	68	0.0052	0.0061	0.0009	0.0796	0.0765	-0.0031	0.0801	0.0506	-0.0295
Farm workers	69	0.0001	0.0001	0.0000	0.0002	0.0002	0.0000	0.0003	0.0029	0.0026

Note: All firms and individuals in firms with at least 2 employees 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. Firms are divided into types as described in Table 3. We consider only the types that experienced a remarkable change in the person-year observations across time (see Table 3). We report the incidence of a certain 2-digits occupation in 2002-2006 vs 2012-2016, and the difference (in percentage points), by type.

occupations, such as caregivers and service workers in health and education).

3.2.3 Potential alternative channels

The economic literature has dealt extensively with skill-biased technological change (Acemoglu and Autor, 2011). A rise in skill premium could explain the increase in sorting and segregation we document. However, worker fixed effects dispersion does not diverge significantly in our period (tables 2, A5). An increase in the skill premium would likely impact the link between occupations and worker fixed effects[22]. The association between occupations and worker fixed effects is very stable between the periods and does not show evidence of an increase in occupational effects variance[23]. Since it is also unlikely that the distribution of worker skills changed dramatically in our relatively short time span, we rule out this channel.

Firm pay premiums have stayed substantially stable everywhere, and its variance has been slightly decreasing[24]. We further elaborate on this evidence by investigating in Figure A3 the employment-weighted relationships between value added per worker and the estimated firm fixed effects at the firm level, by period. The period fitting lines' slopes are basically identical. We conclude that firm pay premium dispersion and levels have been stable and that the distribution of firm rents has not become more skewed towards high-wage workers.

Finally, a vast theoretical literature has linked sorting to complementarities in production (Shimer and Smith, 2000; Eeckhout and Kircher, 2011). If worker and firm attributes affect mobility when interacted, there would be complementary patterns in earnings for different types of workers. The AKM additive specification does not control for this idiosyncratic match component of wages. In order to test whether this creates bias in our approach, we leverage further Bonhomme, Lamadon and Manresa

[22] to be distinguished from the evolution of the distribution of occupations in firms, discussed in the previous section

[23] Table A10 in annex gives the results.

[24] This decline is more pronounced without bias correction, which corroborates the idea that the bias decreased with time. The decline of premium variance is compatible with an "eclipse of rent-sharing" as recently documented by Acemoglu, He and le Maire (2022) for Denmark and the US

(2019) by estimating a correlated random effect model that allows for nonlinear interactions between workers and firms (see Section G for more details about the methodology). We perform a similar log-wage decomposition as in Equation 4 by working with a linear projection of log-wage on worker and firm types. Sorting is still increasing over time, and the rise is quantitatively close to our other estimates.

3.3 Explaining the decline in within-firm inequalities

Why did within-firm inequalities decrease? The rise in segregation through increasing firm specialization is only part of the story. Indeed, as explained in Section 2.2, changes in the segregation index as defined in Equation 5 mechanically do not impact overall wages inequality dynamics. So it is not sufficient to explain France's diminishing inequalities.

A simple statistic to understand within-firm dynamics over time is to plot the average change in earnings for employees ranging from the top1%-paid employee down to the employees in the first percentile (Figure A4). Contrary to evidence in the US, particularly in mega-firms (Song et al., 2019), wages at the bottom have not stagnated in France. On the contrary, they have grown at a higher rate than the rest of the distribution.

Figure A5 shows how this pace of growth is strongly correlated with hourly minimum wage growth, which experienced a steep increase in the first year of our panel following the need of harmonization after the 35-hours a week Reform. This suggestive evidence points towards a potential role for labor market institutions in the French context. The link between P1 and hourly minimum wage is partly mechanical though for two main reasons. First, it is mandatory to pay at least the SMIC in most sectors and for most employees. Second, we define the population based on the hourly minimum wage as explained in Section 1.2.

On a similar note, Bozio, Breda and Guillot (2020) shows that since the 1970's the redistributive effects of payroll taxation have regularly increased. Considered before tax, labour cost inequalities have increased in France at a comparable rate as in other

countries.

Kramarz et al. (2021) concludes as well that the early 2000s' significant increase in the minimum wage—followed by the workweek's reduction to 35 hours and the elimination of all employer-paid payroll taxes around the minimum wage—translates into a significant rise in the bottom percentiles of the earnings distribution in the 2000s. They observe that this mix of policies was particularly beneficial for women, young people and workers in rural and remote municipalities.

France experienced the same mechanisms raising inequalities elsewhere, only to be balanced by increasing redistribution.

4 Discussion

The AKM model has proven a robust description of wages. But the limited mobility bias is a serious limitation that led to an important underestimation of sorting, which can appear null or even negative. Our results confirm that, once corrected for this bias, sorting accounts for more than 10% of overall wage inequalities, measured as the variance of the log-wage. Although less seriously, the bias also impacts the measure of the evolution of sorting, likely because mobility intensity and patterns do evolve in time. We found however that sorting did increase in France, as it did in the USA and Germany, even though log-wage variance in France remained stable throughout the period.

To investigate the causes of this rise in sorting, we mobilize various descriptive methods. Like measures of inequalities, sorting is a distributional statistic, not a characteristic of individuals or firms. It is not directly amenable to classical econometric analysis. We find that firm demographics account for a large part of the rise in sorting: high-premiums, high wages firms have grown more than others, and newly created firms tend to be more often low-wage, low-premium than the one they replace. Both phenomena would point toward a structural evolution in the division of work between firms, such as an increased externalisation of low-value added tasks. Other statistical sources might better inform these phenomena.

Other important open questions are both methodological and substantive. Methodologically, there are additional limitations that are not yet well understood. One is that fixed effects are not fixed. The complete consequences of this specification error are difficult to grasp for the moment, but they might impact the measure of sorting, as well as the other component of the decomposition, especially when short-term economic fluctuations are large. A related question is the measure of the age and experience components of wages, when it matters to disentangle yearly effects, age and individual effects. We found fluctuations in the interactions between these terms that are suggestive of some estimation artefact, but still resistant to alternative specifications.
On the substance, there is more to explore about the interplay of French institutional features and sorting. It appears that wage inequalities in France have been controlled, for most of the period, by an increase in the redistributive power of payroll taxes and by irregular increments in the minimum wage. Both phenomena necessarily impact the shape of the distribution of wages and likely interact with firm pay policies, the job market, and sorting. We focused here on wages, not labor costs, because it better reflects workers incentives, and with the goal to better describe observed inequalities. Still, a replication of our analysis on labour costs in addition to wages could further our understanding of the French exception.

References

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.

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*.

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." National Bureau of Economic Research.

Acemoglu, Daron, and David Autor. 2011. "Skills, Tasks and Technologies: Implications for Employment and Earnings." In *Handbook of Labor Economics*. Vol. 4 of *Handbook of Labor Economics*, , ed. O. Ashenfelter and D. Card, Chapter 12, 1043–1171. Elsevier.

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.

Bergé, Laurent. 2018. "Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm." *CREA Discussion Papers*, , (13).

Bilal, Adrien, and Hugo Lhuillier. 2021. "Outsourcing, inequality and aggregate output." National Bureau of Economic Research.

Bonhomme, Stéphane, Kerstin Holzheu, Thibaut Lamadon, Elena Manresa, Magne Mogstad, and Bradley Setzler. 2020. "How Much Should we Trust Estimates of Firm Effects and Worker Sorting?" National Bureau of Economic Research.

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

Borovičková, Katarína, and Robert Shimer. 2017. "High wage workers work for high wage firms." National Bureau of Economic Research.

Bozio, Antoine, Thomas Breda, and Malka Guillot. 2020. "The Contribution of Payroll Taxation to Wage Inequality in France."

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." *Available at SSRN 3470571*.

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.

de Chaisemartin, Clément, and Xavier d'Haultfoeuille. 2020. "Two-way fixed effects estimators with heterogeneous treatment effects." *American Economic Review*, 110(9): 2964–96.

de Chaisemartin, Clément, and Xavier D'Haultfoeuille. 2022. "Two-way Fixed Effects and Differences-in-Differences Estimators with Several Treatments." National Bureau of Economic Research.

Di Addario, Sabrina L, Patrick M Kline, Raffaele Saggio, and Mikkel Sølvsten. 2021. "It Ain't Where You're From, It's Where You're At: Hiring Origins, Firm Heterogeneity, and Wages." National Bureau of Economic Research Working Paper 28917.

Drenik, Andres, Simon Jäger, Miguel Pascuel Plotkin, and Benjamin Schoefer. 2020. "Paying outsourced labor: Direct evidence from linked temp agency-worker-client data." National Bureau of Economic Research.

Eeckhout, JAN, and PHILIPP Kircher. 2011. "Identifying Sorting—In Theory." *The Review of Economic Studies*, 78(3): 872–906.

Engbom, Niklas, Christian Moser, and Jan Sauermann. 2022. "Firm pay dynamics." National Bureau of Economic Research.

Gaure, Simen. 2013. "lfe: Linear group fixed effects." *The R Journal*, 5(2): 104–117. User documentation of the 'lfe' package.

Gerard, François, Lorenzo Lagos, Edson Severnini, and David Card. 2018. "Assortative matching or exclusionary hiring? The impact of firm policies on racial wage differences in Brazil." National Bureau of Economic Research.

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.

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.

Jochmans, Koen, and Martin Weidner. 2021. "Inference on a distribution from noisy draws." *arXiv preprint arXiv:1803.04991*.

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

Kramarz, Francis, Elio Nimier-David, Thomas Delemotte, et al. 2021. "Inequality and earnings dynamics in France: National policies and local consequences." Center for Research in Economics and Statistics.

Lachowska, Marta, Alexandre Mas, Raffaele D Saggio, and Stephen A Woodbury. 2020. "Do firm effects drift? Evidence from Washington administrative data." National Bureau of Economic Research.

Magnac, Thierry, and Sébastien Roux. 2021. "Heterogeneity and wage inequalities over the life cycle." *European Economic Review*, 103715.

OCDE. 2021. *The Role of Firms in Wage Inequality.*

Palladino, Marco G, Alexandra Roulet, and Mark Stabile. 2020. "Revisiting the Contribution of Firm Pay Policies to the Gender Wage Gap."

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. 2021. "Productivity, Place, and Plants: Revisiting the Measurement." CEPR Discussion Paper No. DP15676.

Shimer, Robert, and Lones Smith. 2000. "Assortative Matching and Search." *Econometrica*, 68(2): 343–370.

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.

ANNEXES

A Appendix Tables

Table A1: Decomposition of wage variance and its evolution - Firms with 20+ employees

		2002-2006		2007-2011		2012-2016		Change from 2002-2006 to 2012-2016	
		Comp.	Share	Comp.	Share	Comp.	Share	Comp.	Share
Total variance	Var(y)	0.212		0.210		0.207		-0.005	
	Var (θ)	0.160	75.3	0.163	77.5	0.156	75.5	-0.004	67.1
	Var (ψ)	0.021	9.8	0.021	9.9	0.018	8.8	-0.003	50.4
	Var(Xb)	0.022	10.2	0.034	16.2	0.016	7.6	-0.006	113.4
	Var(u)	0.008	4.0	0.008	3.9	0.008	4.0	0.000	5.0
	2*Cov(θ, ψ)	0.010	4.5	0.013	6.1	0.015	7.1	0.005	-94.5
	2*Cov(θ, Xb)	-0.010	-4.9	-0.030	-14.5	-0.007	-3.2	0.004	-69.4
	2*Cov(ψ, Xb)	0.002	1.0	0.002	0.8	0.001	0.3	-0.002	27.9
Between-firm variance	Var($\bar{y}$)	0.088	41.3	0.092	44.0	0.097	46.7	0.009	-164.6
	Var ($\bar{\theta}$)	0.051	24.3	0.056	26.6	0.061	29.7	0.010	-182.6
	Var (ψ)	0.021	9.8	0.021	9.9	0.018	8.8	-0.003	50.4
	Var($\bar{X}b$)	0.003	1.5	0.005	2.3	0.002	1.0	-0.001	19.2
	2*Cov($\bar{\theta}, \psi$)	0.010	4.5	0.013	6.1	0.015	7.1	0.005	-94.5
	2*Cov($\bar{\theta}, \bar{X}b$)	0.001	0.2	-0.004	-1.7	0.000	-0.1	-0.001	15.0
	2*Cov($\psi, \bar{X}b$)	0.002	1.0	0.002	0.8	0.001	0.3	-0.002	27.9
Within-firm variance	Var($y - \bar{y}$)	0.124	58.7	0.118	56.0	0.110	53.3	-0.014	264.6
	Var ($\theta - \bar{\theta}$)	0.108	51.0	0.107	50.9	0.095	45.9	-0.013	249.7
	Var($Xb - \bar{X}b$)	0.019	8.7	0.029	14.0	0.013	6.5	-0.005	94.3
	Var(u)	0.008	4.0	0.000	0.0	0.000	0.0	-0.008	156.5
	2*Cov($\theta - \bar{\theta}$,$Xb - \bar{X}b$)	-0.011	-5.1	-0.027	-12.7	-0.006	-3.0	0.005	-84.4
	2*Cov($\theta - \bar{\theta}$,u)	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0
	2*Cov($Xb - \bar{X}b$,u)	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0
Segregation	$\frac{Var(\bar{\theta}_j)}{Var(\theta_i)}$	0.322		0.343		0.393		0.071	
N* (largest connected set)		34,319,605		36,782,639		39,472,450			

Note: In each period, all firms and individuals in firms with at least 20 different employees over the period, and in the main connected component, 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. Decomposition based on Equations 2, 3 and 4.

Table A2: Decomposition of wage variance and its evolution - Establishments with 20+ employees

		2002-2006		2012-2016		Change from 2002-2006 to 2012-2016	
		Comp.	Share	Comp.	Share	Comp.	Share
Total variance	Var(y)	0.217		0.215		-0.002	
	Var (θ)	0.161	74.3	0.161	75.0	0.000	-4.1
	Var (ψ)	0.028	12.9	0.025	11.4	-0.003	176.5
	Var(Xb)	0.020	9.1	0.015	6.9	-0.005	250.1
	Var(u)	0.008	3.8	0.008	3.7	0.000	9.8
	2*Cov(θ, ψ)	0.007	3.1	0.012	5.7	0.006	-281.3
	2*Cov(θ, Xb)	-0.009	-4.0	-0.006	-2.9	0.003	-127.4
	2*Cov(ψ, Xb)	0.002	0.9	0.000	0.2	-0.002	76.6
Between-establishment variance	Var($\bar{y}$)	0.101	46.8	0.113	52.6	0.012	-588.5
	Var ($\bar{\theta}$)	0.062	28.4	0.074	34.6	0.013	-642.4
	Var (ψ)	0.028	12.9	0.025	11.4	-0.003	176.5
	Var($\bar{X}b$)	0.003	1.5	0.002	1.1	-0.001	46.2
	2*Cov($\bar{\theta}, \psi$)	0.007	3.1	0.012	5.7	0.006	-281.3
	2*Cov($\bar{\theta}, \bar{X}b$)	0.000	0.0	-0.001	-0.3	-0.001	35.9
	2*Cov(FFE.m_Xb)	0.002	0.9	0.000	0.2	-0.002	76.6
Within-establishment variance	Var($y - \bar{y}$)	0.115	53.2	0.102	47.4	-0.014	688.5
	Var ($\theta - \bar{\theta}$)	0.099	45.8	0.087	40.4	-0.013	638.3
	Var($Xb - \bar{X}b$)	0.016	7.6	0.012	5.8	-0.004	204.0
	Var(u)	0.008	3.9	0.000	0.0	-0.008	429.0
	2*Cov($\theta - \bar{\theta}$,$Xb - \bar{X}b$)	-0.009	-4.0	-0.006	-2.6	0.003	-163.5
	2*Cov($\theta - \bar{\theta}$,u)	0.000	0.0	0.000	0.0	0.000	0.0
	2*Cov($Xb - \bar{X}b$,u)	0.000	0.0	0.000	0.0	0.000	0.0
Segregation	$\frac{Var(\bar{\theta}_j)}{Var(\theta_i)}$	0.383		0.461		0.078	
N* (largest connected set)		30,282,946		33,911,928			

Note: All establishments and individuals in establishments with at least 20 employees are included. Only individuals employed for at least 360 days by the same establishment during the year are included for a given year. Individuals and establishments in public administration are not included. Decomposition based on Equations 2, 3 and 4.

Table A3: Decomposition of wage variance and its evolution - Split-sampling correction with firm split

		2002-2006		2007-2011		2012-2016		Change from 2002-2006 to 2012-2016	
		Comp.	Share	Comp.	Share	Comp.	Share	Comp.	Share
Total variance	Var(y)	0.211		0.206		0.201		-0.011	
	**Var (θ)								
	Var (ψ)	0.014	6.5	0.014	7.0	0.013	6.5	-0.001	6.6
	Var(Xb)	0.022	10.6	0.036	17.6	0.018	8.8	-0.005	44.0
	**Var(u)								
	2Cov(θ,ψ)	0.024	11.6	0.027	12.9	0.027	13.3	0.002	-20.0
	2Cov(θ,Xb)	-0.010	-4.7	-0.031	-15.0	-0.008	-3.9	0.002	-18.1
	2Cov(ψ,Xb)	0.002	1.1	0.002	1.0	0.001	0.4	-0.002	14.5
Between-firm variance	Var($\bar{y}$)	0.091	43.0	0.094	45.8	0.097	48.6	0.006	-59.7
	Var ($\bar{\theta}$)	0.043	20.5	0.047	23.0	0.054	26.9	0.011	-97.5
	Var (ψ)	0.014	6.5	0.014	7.0	0.013	6.5	-0.001	6.6
	Var($\bar{X}B$)	0.004	1.9	0.007	3.2	0.003	1.6	-0.001	6.7
	2Cov($\bar{\theta}$,ψ)	0.024	11.6	0.027	12.9	0.027	13.3	0.002	-19.9
	2Cov($\bar{\theta}$,$\bar{X}B$)	0.000	0.2	-0.004	-1.9	-0.001	-0.4	-0.001	10.5
	2Cov(ψ,$\bar{X}B$)	0.002	1.1	0.002	1.0	0.001	0.4	-0.002	14.5
Within-firm variance	Var($y - \bar{y}$)	0.120	57.0	0.111	54.2	0.103	51.4	-0.017	159.7
	**Var ($\theta - \bar{\theta}$)								
	Var($Xb - \bar{X}b$)	0.018	8.7	0.030	14.4	0.014	7.2	-0.004	37.4
	**Var(u)								
	2Cov($\theta - \bar{\theta}$,$Xb - \bar{X}b$)	-0.010	-4.8	-0.026	-12.9	-0.007	-3.5	0.003	-28.6
	**2Cov($\theta - \bar{\theta}$,u)								
	2Cov($Xb - \bar{X}b$,u)	0.000	0.0	0.000	0.0	0.000	0.0	0.000	-0.1
****Segregation**	$\frac{Var(\bar{\theta}_j)}{Var(\theta_i)}$	0.255		0.280		0.333		0.078	
N*		36,113,649		39,853,353		41,362,051			

Note: 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. Estimation on firms present in both main connected component in each split sample. Decomposition based on Equations 2, 3 and 4. Split-sampling method described in Section 2.3.1.

** : These parameters' estimates are not corrected by firm-split. Segregation index computed with uncorrected $Var(\theta)$ estimates

Table A4: Decomposition of wage variance and its evolution - Split-sampling correction with period split

		2002-2006		2007-2011		2012-2016		Change from 2002-2006 to 2012-2016	
		Comp.	Share	Comp.	Share	Comp.	Share	Comp.	Share
Total variance	Var(y)	0.213		0.211		0.205		-0.008	
	Var(θ)	0.150	70.4	0.154	73.1	0.145	70.7	-0.005	63.4
	Var(ψ)	0.017	7.8	0.016	7.7	0.014	7.0	-0.002	27.4
	Var(Xb)	0.020	9.2	0.034	16.3	0.016	7.8	-0.004	45.4
	**Var(u)								
	2*Cov(θ, ψ)	0.019	9.1	0.024	11.1	0.023	11.1	0.003	-42.2
	2*Cov(θ, Xb)	-0.008	-3.6	-0.030	-14.1	-0.007	-3.4	0.001	-6.5
	2*Cov(ψ, Xb)	0.002	0.9	0.002	0.9	0.001	0.3	-0.001	17.7
Between-firm variance	Var($\bar{y}$)	0.089	41.7	0.093	44.3	0.096	46.8	0.007	-90.1
	Var ($\bar{\theta}$)	0.048	22.3	0.051	24.1	0.057	27.7	0.009	-117.0
	Var (ψ)	0.017	7.8	0.016	7.7	0.014	7.0	-0.002	27.4
	Var($\bar{X}b$)	0.003	1.4	0.005	2.4	0.002	1.2	-0.001	7.5
	2*Cov($\bar{\theta}, \psi$)	0.019	8.9	0.023	10.9	0.023	11.0	0.004	-44.9
	2*Cov($\bar{\theta}, \bar{X}b$)	0.001	0.2	-0.004	-1.8	-0.001	-0.4	-0.001	16.1
	2*Cov($\psi, \bar{X}b$)	0.002	0.9	0.002	0.9	0.001	0.3	-0.001	17.7
Within-firm variance	Var ($y - \bar{y}$)	0.125	58.3	0.118	55.7	0.109	53.2	-0.015	190.1
	Var($\theta - \bar{\theta}$)	0.103	48.1	0.103	48.6	0.089	43.2	-0.014	174.0
	Var($Xb - \bar{X}b$)	0.017	7.8	0.029	13.9	0.014	6.7	-0.003	38.0
	**Var(u)								
	2*Cov($\theta - \bar{\theta}$,$Xb - \bar{X}b$)	-0.008	-3.7	-0.026	-12.2	-0.006	-2.9	0.002	-23.1
	2*Cov($\theta - \bar{\theta}$,u)	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.1
	2*Cov($Xb - \bar{X}b$,u)	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0
Segregation	$\frac{Var(\bar{\theta}_j)}{Var(\theta_i)}$	0.316		0.330		0.391		0.076	
N*		34,649,503		37,882,676		40,420,363			

Note: 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. Estimation on firms and, when necessary, individuals present in both main connected component in each split sample. Decomposition based on Equations 2, 3 and 4. Split-sampling method described in Section 2.3.1.
** : These parameters' estimates are not corrected by period-split

Table A5: Decomposition of wage variance and its evolution - Firm Clustering

		2002-2006		2007-2011		2012-2016		Change from 2002-2006 to 2012-2016	
		Comp.	Share	Comp.	Share	Comp.	Share	Comp.	Share
Total variance	Var(y)	0.213		0.208		0.203		-0.009	
	Var (θ)	0.158	74.2	0.160	76.8	0.149	73.4	-0.009	91.0
	Var (ψ)	0.007	3.2	0.007	3.2	0.006	3.0	-0.001	5.7
	Var(Xb)	0.026	12.2	0.036	17.2	0.017	8.3	-0.009	97.8
	Var(u)	0.010	4.5	0.009	4.3	0.009	4.3	-0.001	7.9
	2*Cov(θ,ψ)	0.026	12.0	0.027	12.8	0.029	14.2	0.003	-35.0
	2*Cov(θ,Xb)	-0.015	-7.1	-0.032	-15.2	-0.008	-3.9	0.007	-77.2
	2*Cov(ψ,Xb)	0.002	1.1	0.002	1.0	0.001	0.7	-0.001	9.8
Between-cluster variance	Var($\bar{y}$)	0.082	38.6	0.086	41.2	0.092	45.2	0.010	-106.3
	Var ($\bar{\theta}$)	0.042	19.6	0.045	21.6	0.052	25.4	0.010	-106.9
	Var (ψ)	0.007	3.2	0.007	3.2	0.006	3.0	-0.001	5.7
	Var($\bar{X}B$)	0.001	0.5	0.001	0.6	0.001	0.3	0.000	4.7
	2*Cov($\bar{\theta}$,ψ)	0.026	12.0	0.027	12.8	0.029	14.2	0.003	-35.0
	2*Cov($\bar{\theta}$,$\bar{X}B$)	0.005	2.2	0.004	2.0	0.003	1.6	-0.001	15.4
	2*Cov(ψ,$\bar{X}B$)	0.002	1.1	0.002	1.0	0.001	0.7	-0.001	9.8
Within-cluster variance	Var($y - \bar{y}$)	0.131	61.4	0.122	58.8	0.111	54.8	-0.019	206.3
	Var ($\theta - \bar{\theta}$)	0.116	54.6	0.115	55.1	0.098	48.0	-0.019	197.8
	Var($Xb - \bar{X}b$)	0.025	11.7	0.035	16.6	0.016	8.0	-0.009	93.1
	Var(u)	0.010	4.5	0.009	4.3	0.009	4.3	-0.001	7.9
	2*Cov($\theta - \bar{\theta}$,$Xb - \bar{X}b$)	-0.020	-9.4	-0.036	-17.2	-0.011	-5.5	0.009	-92.6
	2*Cov($\theta - \bar{\theta}$,u)	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0
	2*Cov($Xb - \bar{X}b$,u)	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0
Segregation	$\frac{Var(\bar{\theta}_j)}{Var(\theta_i)}$	0.264		0.282		0.346		0.082	
N*		44,618,999		48,953,550		52,369,828			

Note: 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. Decomposition based on Equations 2, 3 and 4. Firm clustering method described in Section 2.3.2.

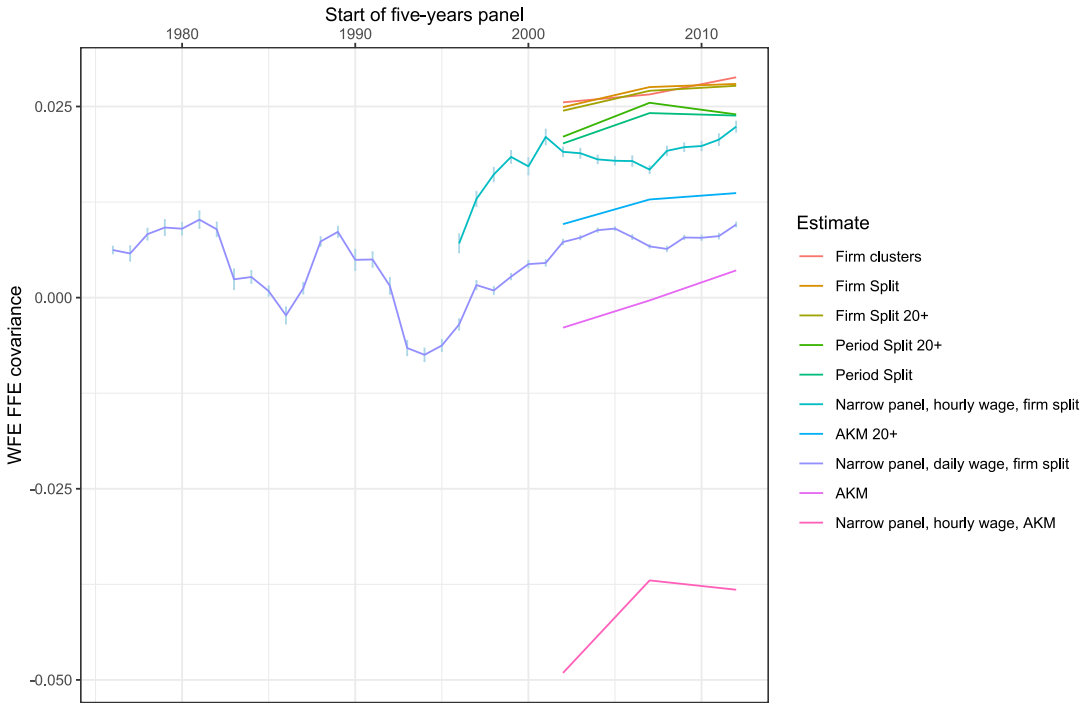
Table A6: Decomposition of wage variance and its evolution - Year fixed effects - Firm-split correction

		2002-2006		2007-2011		2012-2016		Change from 2002-2006 to 2012-2016	
		Comp.	Share	Comp.	Share	Comp.	Share	Comp.	Share
Total variance	Var (y)	0.212		0.206		0.2		-0.012	
	***Var (θ)*								
	Var (ψ)	0.014	6.5	0.014	7	0.013	6.4	-0.001	8
	Var (Xb)	0.003	1.5	0.003	1.3	0.002	0.9	-0.001	10.9
	***Var (u)*								
	2Cov(θ,ψ)	0.026	12.1	0.028	13.5	0.027	13.6	0.001	-11.6
	2Cov(θ,Xb)	0	0.1	0.001	0.4	0.001	0.7	0.001	-9
	2Cov(ψ,Xb)	0.001	0.3	0.001	0.3	0.001	0.3	0	0.6
Between-firm variance	Var my	0.091	42.7	0.094	45.6	0.097	48.6	0.006	-51.8
	Var ($\bar{\theta}$)	0.046	21.7	0.048	23.3	0.054	26.9	0.008	-61.4
	Var (ψ)	0.014	6.5	0.014	7	0.013	6.4	-0.001	8
	Var ($\bar{X}B$)	0	0.1	0	0.1	0	0.1	0	0.9
	2Cov($\bar{\theta}$,ψ)	0.026	12.2	0.028	13.5	0.027	13.6	0.001	-11.6
	2Cov($\bar{\theta}$,$\bar{X}B$)	0.001	0.5	0.001	0.6	0.001	0.7	0	-2.3
	2Cov(ψ,$\bar{X}B$)	0.001	0.3	0.001	0.3	0.001	0.3	0	0.6
Within-firm variance	Var($y - \bar{y}$)	0.122	57.3	0.112	54.4	0.103	51.4	-0.019	151.8
	***Var ($\theta - \bar{\theta}$)*								
	Var($Xb - \bar{X}b$)	0.003	1.3	0.002	1.1	0.002	0.8	-0.001	10
	***Var(u)*								
	2Cov($\theta - \bar{\theta}$,$Xb - \bar{X}b$)	-0.001	-0.4	0	-0.2	0	0	0.001	-6.5
	***2*Cov($\theta - \bar{\theta}$,u)*								
	2Cov($Xb - \bar{X}b$,u)	0	0	0	0	0	0	0	0
****Segregation**	$\frac{Var(\bar{\theta}_i)}{Var(\theta_i)}$	0.258		0.28		0.319		0.061	
N*		35,892,335		39,625,364		40,155,758			

Note: AKM estimation with fixed effects for years and a cubic function of age flat at 40. No selection for age. Bias corrected with firm split.
** : These parameters' estimates are not corrected by firm-split
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. Estimation on firms present in both main connected component in each split sample. Decomposition based on Equations 2, 3 and 4. Split-sampling method described in Section 2.3.1.

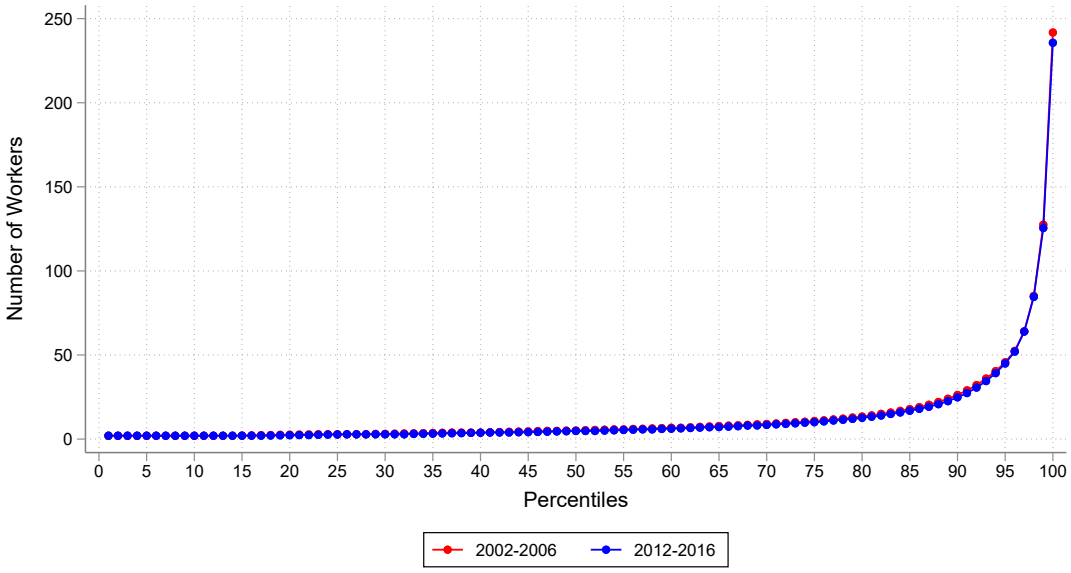
B Appendix Figures

Figure A1: Sorting, historical series



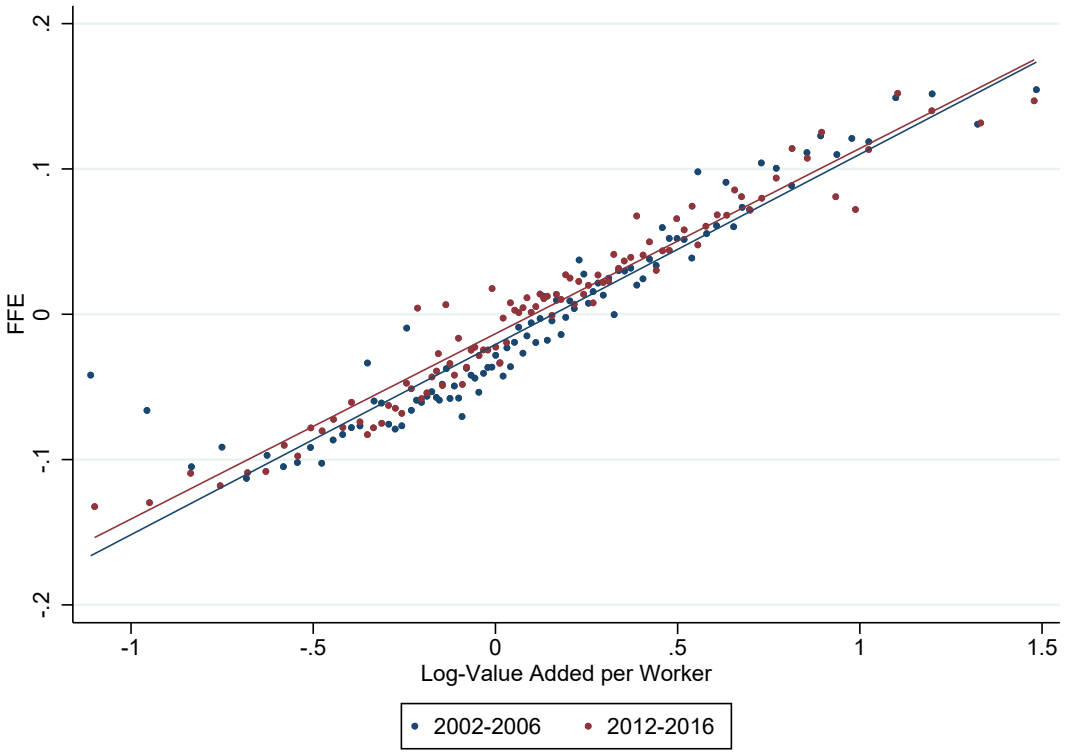
Note: All firms and individuals in firms with at least 1 employee are included. Only individuals employed for at least 360 days by the same firm during the year are included for a given year. Individuals and firms in public administration are not included. Sorting estimates by five years periods. Long term series are computed on the narrow panel, on rolling 5 years period, corrected by firm-splitting, with mean estimate and confidence intervals computed on repeated (split) sampling with 20 repetitions, reflecting only the noise stemming from the randomness of the split. Estimates on the narrow panel are particularly affected by sampling error and selection of bigger and more connected firms. Years 1981, 1983, 1990 are missing. There have been several changes of scope and variable definition since 1976.

Figure A2: Cumulative firm size distribution



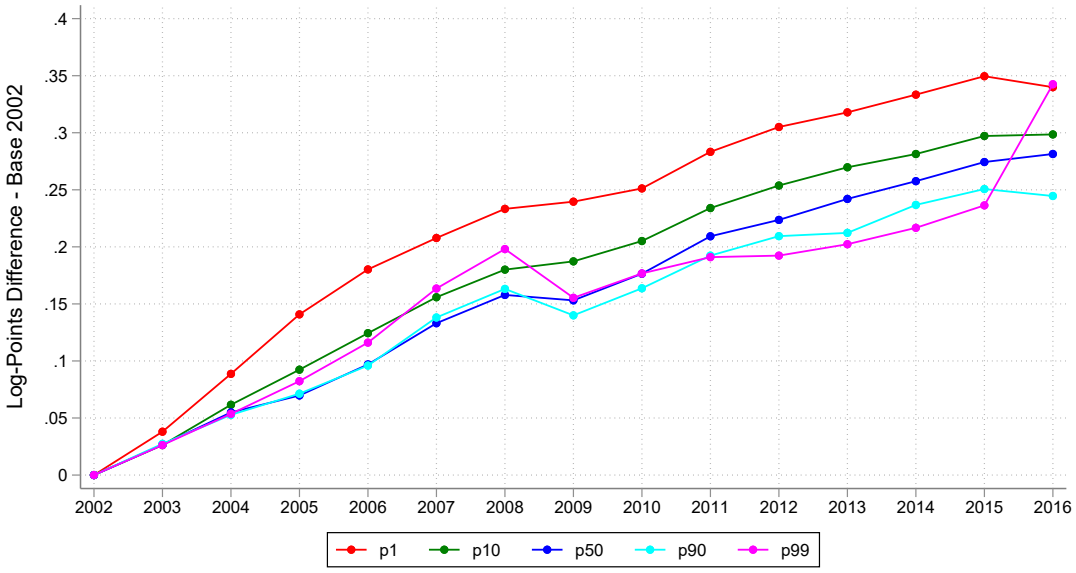
Note: All firms and individuals in firms with at least 1 employee are included. Only individuals employed for at least 360 days by the same firm during the year are included for a given year. Individuals and firms in public administration are not included. This figure shows the fraction of firms below a given size, by period.

Figure A3: Rent-Sharing - Firm Fixed Effects vs Log Value Added/Worker



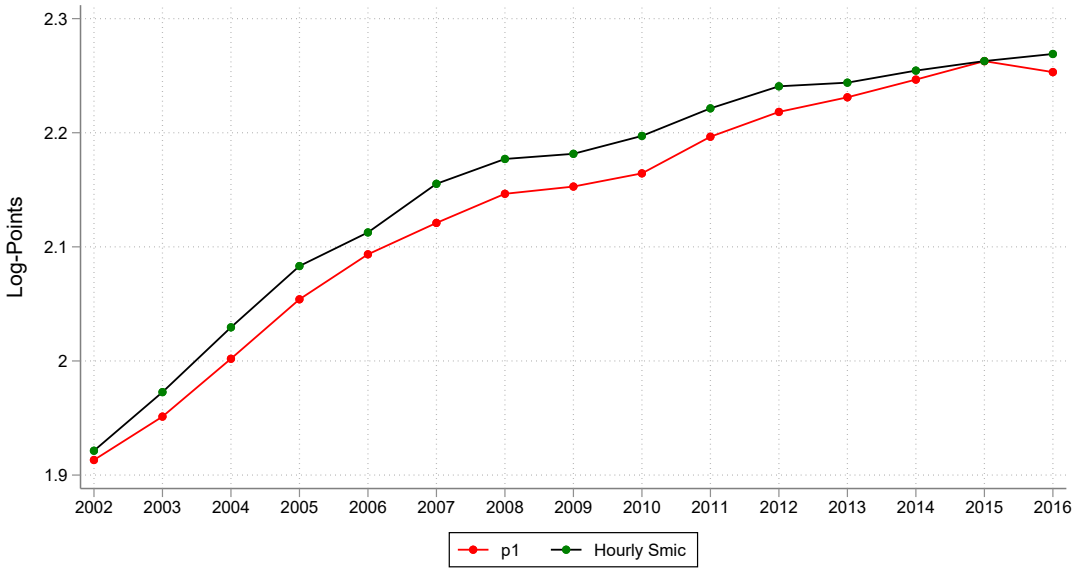
Note: All firms and individuals in firms with at least 1 employee are included. Only individuals employed for at least 360 days by the same firm during the year are included for a given year. Individuals and firms in public administration are not included. Points shown represent mean estimated firm fixed effects from AKM models, averaged across firms in 100 percentile bins of mean log value added per worker. Period best-fitting lines coming from employment-weighted OLS are reported.

Figure A4: Change in percentiles of real earnings relative to 2002 - Individuals



Note: 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. This figure plots the average change in log-earnings for employees ranging from the top1%-paid employee down to the employees in the first percentile relative to 2002.

Figure A5: Evolution of P1 of real earnings and of real hourly minimum wage



Note: 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 figure plots the evolution over time of the first percentile of real earnings and of real hourly minimum wage (in log points).

C Constructing a DADS full panel

C.1 Chaining the yearfiles

The French DADS is not a proper panel dataset as there is no individual IDs before 2002 and after 2002 the individual IDs are specific to each yearfile. However, each yearfile y contains information both on the current year t and the preceding 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 Insee's authorization to chain the DADS yearly files in order to create a full panel of the wage earning population between 1994 and 2020. Between 1994 and 2001, as there is no individual IDs enabling to track mobilities within the yearfiles, we can only match the "stayers", employees as long as they stay in the same establishment: for instance an individual who moves from workplace j to workplace k will be given two different IDs. After 2002, individual IDs in the original yearfiles, even if yearfile specific, enables us to match both stayers and 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 DADS regional files[25]. Within the regional file, we keep the job for which a worker i has the highest pay.

We create the following keys in for the year t of yearfile $y-1$:

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

```
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);
```

However, as there are some discrepancies in the ages and the wages reported for the same year in yearfile $y-1$ and y, we do not use them directly in the matching key. We use the having property of the SQL procedure, in order to select the match with the minimal difference between the two wages and an absolute age difference below 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 was adapted to account for 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 at the end

the regional code. For instance: the ID for the 10th observation of Paris region (code: 11) will be 1011.

- In 2013 ($y = 2013$ and $y - 1 = 2012$), the `SONDE` variable leads to some 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 yearfile $y - 1$ and y. We thus excluded the number of hours from the matching key and we added the minimal difference in the number of hours in the having clause.

We count the number of matches based on the procedure and we attribute 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 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 identification, we opted for a conservative procedure in order to identify two individuals as the same person, by using the maximal available overlapping information. When the procedure leads to multiple matches, we do not impute any identification. However these duplicates remain rare, around 0.4% of the observations. Most matching failures are due to observations for which we don't find any match.

Figure A6 gives a first proxy of the quality of the matching. Generally, we find a single match for 98% of the observations of the overlapping years of two yearfiles. The quality of the matching declines between 2016 and 2018, dropping to 91-93% and

resumes back to 97% in 2019, probably as a result of the switch from DADS to the DSN[26]. With the existing procedure, the match is poor for yearfile 2002 (and similarly in 1995), as the consequence of the major transformation of the DADS between the 1994-2001 series and the 2002-2020 serie[27].

Despite a high level of matches, the matching has some limitations. We must bear in mind, that a false positive is still possible (but unlikely). Second, in order to be identified as the same person, an employee needs to be present each year as a wage earner in the DADS. This means that we cannot link the initial ID of an employee who was either further unemployed, self-employed or state civil servant (before 2009) for more than a calendar year to her subsequent employment periods. Overall, the quality of the match seems sufficiently good to run AKM panel regressions.

C.3 How to use the ID files

Hence, 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[28]:

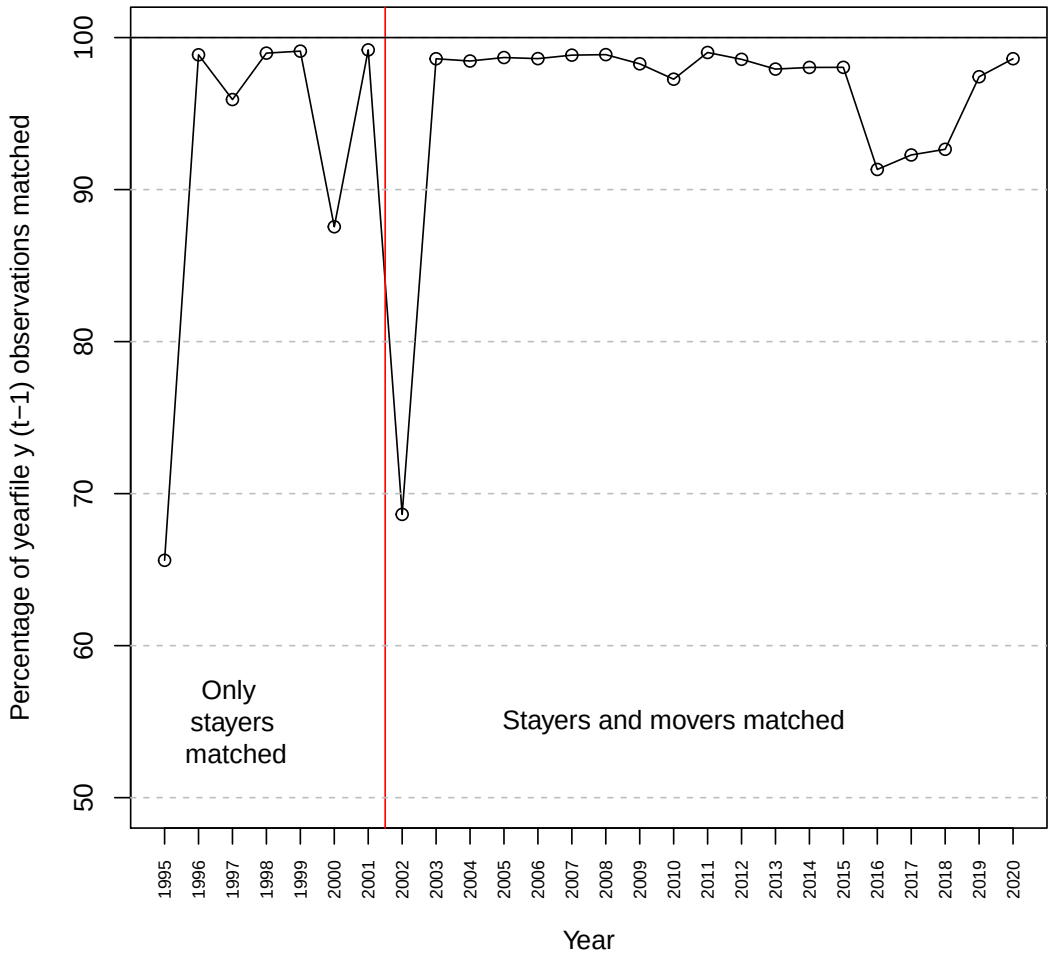
```
PROC SQL;
 CREATE TABLE b2010b
 AS SELECT * FROM b2010 AS aa
 LEFT JOIN psid.psid_2010 AS bb
 ON aa.ident_s=bb.ident_s;
QUIT;


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

[26] The "déclaration sociale nominative" is a new monthly administrative source replacing the "déclarations annuelles de données sociales". INSEE produces from the DSN a yearly datafile on the DADS format for series continuity

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

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

Figure A6: Quality of the match



Note: The figure present the proportion of observations from the year $t-1$ of yearfile y for which we found a single match in the year t of yearfile $y-1$. 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.

Before 2002, in order to get permanent IDs necessary for the match with the `PSID_yyyy` files, one needs to create an ID in the 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;
```

D Proof of split-sampling bias correction

We suppose that observations are randomly split in two half samples, and that each sample retains the same connectivity as the full sample. In practice, we reduce the data to firms and individuals (for period-split, or only firms for firm-split) that belong to both main connected components in each split sample. This condition is analogous to the Kline, Saggio and Sølvsten (2020) leave-one-out condition that the main connected sample is not disconnected when any one observation is removed. We start again with a simplified notation of the AKM model in equation 6:

$$y_i = z_i'\alpha + u_i \tag{9}$$

With $\alpha = (\beta, \theta, \psi)$ our parameter vector of length $k = 2 + N + J$ and z_i the non-random regressors vector of the (worker * year) i's observation characteristics, including the indicator vector for worker and firm. For a given symetric matrix A corresponding to a given quadratic form ω of interest, our split-sampling plug-in estimator becomes $\hat{\omega}^{SP} = \hat{\alpha}_0' A \hat{\alpha_1}$ with $\hat{\alpha}_s$ an OLS estimate in the sample $I_s, s = 0,1$ of size N_s : $\hat{\alpha}_s = S_{zz,s}^{-1} \sum_{i \in s}^{N_s} z_i y_i = \alpha + S_{zz,s}^{-1} \sum_{i \in s}^{N_s} z_i u_i$, with $S_{zz,s} = \sum_{i \in s}^{N_s} z_i z_i'$ the split sample design matrix (with full rank when we limit the sample to the split sample main connected set). We can express $\hat{\alpha}_s$ as $\hat{\alpha}_s = \alpha + \epsilon_s$. We calculate the bias by expressing a scalar as the trace of a (1,1) matrix, as in the classic demonstration on the expectation of quadratic forms

$$\begin{aligned}
\mathbf{E}[\hat{\omega}^{SP}] &= \mathbf{E}[\hat{\alpha}_0' A \hat{\alpha_1}] = \mathbf{E}[trace(\hat{\alpha}_0' A \hat{\alpha_1})] \\
&= \mathbf{E}[trace(\hat{\alpha_1} \hat{\alpha}_0' A)] && \text{by propriety of trace()} \\
&= \mathbf{E}[trace(A \hat{\alpha_1} \hat{\alpha}_0')] && \text{idem} \\
&= trace(A \mathbf{E}[\hat{\alpha}_1 \hat{\alpha_0}']) && \text{A non-random, trace() is linear}
\end{aligned}$$

We further have:

$$\begin{aligned}\mathbf{E}[\hat{\alpha}_1\hat{\alpha_0}'] &= \mathbf{E}[(\alpha+\epsilon_1)(\alpha+\epsilon_0)'] \\ &= \alpha\alpha' + \mathbf{E}[\epsilon_1\epsilon_0']\end{aligned}$$

And: $trace(A\alpha\alpha') = \omega$

The bias is thus equal to:

$$\mathbf{E}[\hat{\omega}^{SP}] - \omega = trace(A\mathbf{E}[(S_{zz,1}^{-1}\sum_{i\in I_1}^{N_1} z_i u_i)(S_{zz,0}^{-1}\sum_{j\in I_0}^{N_0} z_j u_j)'])$$

$$\mathbf{E}[\hat{\omega}^{SP}] - \omega = trace(A\mathbf{E}[(S_{zz,1}^{-1}\sum_{i\in I_1}^{N_1} u_i z_i)(S_{zz,0}^{-1}\sum_{j\in I_0}^{N_0} u_j z_j)'])$$

$$= trace(AS_{zz,1}^{-1}\underbrace{\mathbf{E}[(\sum_{i\in I_1}^{N_1} u_i z_i)(\sum_{j\in I_0}^{N_0} u_j z_j)']}_{\text{matrix } (b_{lm})}(S_{zz,0}^{-1})')$$

with generic term:

$$b_{lm} = \sum_{i\in I_1}^{N_1} u_i z_{l,i} \sum_{j\in I_0}^{N_0} u_j z_{m,j}$$

The variance of the split sampling estimator of a quadratic form stems from the random errors u_i, but also from the randomness of the split. We can abstract from this last source by considering only a given split, and the corresponding sample of firms and individuals connected in both split samples. The size of the variance then depends on the quadratic form matrix A and the design matrices $S_{zz,s}$ and their relation, so that additional hypothesis are needed for this variance to be of finite value. Kline, Saggio and Sølvsten (2020) discuss these conditions in the context of leave-one-out. When introducing the randomness of the split however, A and $S_{zz,s}$ become random matrices. Finally, to study the consistency and convergence of our estimator, we need to consider the series of these random matrices when some index n of the number of observations grows. We leave the study of the precise conditions for consistency and convergence to further research. Instead, we checked that the estimator showed reasonable stability over multiple random splits, and recovered known values in Monte Carlo experiments.

D.1 Simulations

We generated simulated workers and firms fixed effects with sorting and noise, calibrated on measured distributions, to get simulated wages on the observed match. It is important to keep the real mobility network, on which the bias depends. On table A7, split sample correction recovers the true value on the sample of firms belonging in connected component in both splits, that differs from the full sample true values because firms are bigger and there is less firm variance in fixed effects and mean workers effects. This kind of simulation cannot however reproduce potential selection effects in the sample reduction.

Table A7: Simulated wage: true fixed effects and estimations

	true	AKM	true, split sample	corrected
Var WFE	0,134	0,165	0,135	0,135
Var FFE	0,014	0,030	0,009	0,009
2*Cov(WFE,FFE)	0,022	-0,004	0,025	
2*Cov(WFE_H1,FFE_H0)				0,025
2*Cov(WFE_H0,FFE_H1)				0,024
N*	41 703 340	41 703 340	29 543 074	29 543 074

Simulation on 2002-2006 data, corrected estimates with period split method.
First column: true simulated values. Second column: classic AKM estimation results. Third column: true simulated values on the restricted sample used to compute the split-sampling correction. Fourth column: slip-sampling corrected estimation results.

D.2 Multiple random splits

We checked the stability of the (firm split) split sample 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. We plot the means of 20 split sample estimations and a confidence interval on this mean in figure A1.

On our main estimates, we limited the multiple random split experiment to the firm-split on the first and third periods for computational reasons. In table A8 we report the mean and standard deviation of 20 estimations.

The standard deviations are very small relatively to the estimates, the sizes of the bias correction and the evolution between period. Our split sampling corrected results do

Table A8: Decomposition of wage variance - Mean and standard deviation over 20 firm split estimations

		2002-2006		2012-2016	
		Mean	SD	Mean	SD
Total variance	Var(y)	0.213	0.00003	0.204	0.00003
	Var (ψ)	0.014	0.0001	0.013	0.0001
	Var(Xb)	0.021	0.00003	0.017	0.00002
	2Cov(θ,ψ)	0.024	0.0001	0.027	0.0001
	2Cov(θ,Xb)	−0.009	0.00004	−0.008	0.00002
	2Cov(ψ,Xb)	0.002	0.00002	0.001	0.00001
	2Cov(ψ,u)	−0.000	0.00001	−0.000	0.00000
	2Cov(Xb,u)	0.000	0.00000	0.000	0.00000
Between-firm variance	Var($\bar{y}$)	0.089	0.00003	0.096	0.00002
	Var ($\bar{\theta}$)	0.044	0.0001	0.054	0.0001
	Var($\bar{X}B$)	0.003	0.00000	0.003	0.00000
	2Cov($\bar{\theta}$,ψ)	0.024	0.0001	0.027	0.0001
	2Cov($\bar{\theta}$,$\bar{X}B$)	0.000	0.00002	−0.001	0.00001
	2Cov(ψ,$\bar{X}B$)	0.002	0.00002	0.001	0.00001
Within-firm variance	Var($y - \bar{y}$)	0.124	0.00002	0.108	0.00002
	Var($Xb - \bar{X}b$)	0.018	0.00003	0.014	0.00002
	2Cov($\theta - \bar{\theta}$,$Xb - \bar{X}b$)	−0.009	0.00003	−0.007	0.00002
	2Cov($Xb - \bar{X}b$,u)	−0.000	0.00000	0.000	0.00000
N. of firms		138,437	317	141,444	272
N* of obs		35,946,471	18,879	41,112,856	15,894

Note: Mean and standard deviations computed on 20 estimations similar to table A3, on firms belonging to both main connected components

not stem from random split noise. This exercise can also be interpreted as a bootstrap, indicative more generally of a high stability of AKM decomposition statistics in our large dataset.

E Log-wage variance decomposition from alternate specifications

We conduct multiple robustness experiments with alternate specifications. For practical reasons, most of these experiments were computed on DADS files available within INSEE (the producer of the files), a slightly different version from the files available to researchers through CASD. The results might differ slightly from the baselines results in the main text for this reason. We leveraged the lower computation time of the R package *fixest*[29], up to ten times faster the the R package *lfe*[30] or the Stata package *reghdfe*[31], although we used all three methods depending on the setting.

E.1 Split-sample estimates

Split sample results are presented in tables A3 and A4. . Table A9 shows that almost all the reduction in bias comes from the increase in estimated sorting in firms with less than 200 employees.

Table A9: Sorting decomposition by firm size group

		2002-2006				2012-2016			
	Firm size category	< 20	$20-200$	$200-1000$	> 1000	< 20	$20-200$	$200-1000$	> 1000
Split-sampling	Overall sorting			0.0244				0.0266	
	Between			0.0016				0.0021	
	Within	0.0285	0.0245	0.0226	0.0204	0.0257	0.0223	0.0258	0.0253
AKM	Overall sorting			-0.0040				0.0041	
	Between			0.0019				0.0026	
	Within	-0.0842	-0.0042	0.0190	0.0200	-0.0695	-0.0007	0.0232	0.0254

Note: 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. Split sampling with the firm split method. Firm size is computed on the observed workers employed by the firm in the panel. Sorting is computed as elsewhere as $2Cov(\theta, \psi)$, while between sorting is $2Cov(\bar{\theta}, \bar{\psi})$ with the average computed for each size group, and within is sorting computed separately for each size group.

[29] Bergé (2018)
[30] Gaure (2013)
[31] Correia (2016)

E.2 Model with year fixed effects and a cubic function of age

Card et al. (2018)'s AKM model specification differs from ours. While we demean wages per year, they use year fixed effects. To avoir colinearity between year effects, workers effects and age, they suppress the linear term in age and keep only a quadratic and a cubic term in (age-40). We implement this specification and present results in table A6, with split sampling bias correction with the firm-split method (variance of workers effects and residual are not corrected).

With this specification, covariance between workers effects and age functions are close to zero. The rise in sorting is still visible but less important.

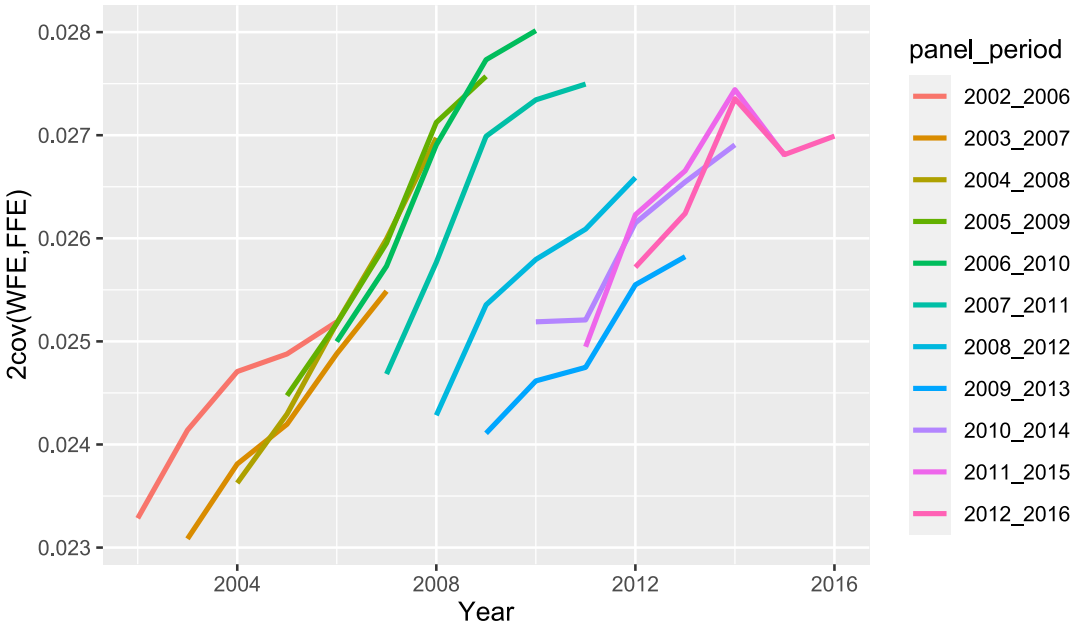
E.3 Rolling panels and yearly log-wage variance decomposition

Are fixed effects really fixed? To better understand their stability or lack thereof, we implement the rolling AKM (R-AKM) method of Lachowska et al. (2020) by estimating firm-split corrected AKM on five-years panels for every possible starting year in our data. We then apply log-wage variance decomposition separately for each year of each panel estimation. Within each five-years panel, year-to-year variations in sorting reflect composition effects (year to year changes in the populations of firms and workers) and pure sorting (changes in worker-to-firm matching). If fixed effects were fixed, and the AKM estimation correct, the yearly results would be the same whatever the panel used for estimation. On the contrary, we observe systematic shifts between different panels' estimates, for a given year (Figure A7). Sampling differences between panels due to the main connected component selection might explain some of the shift, but it is most likely due to changes in the fixed effects of a given firm or a given worker, estimated in different, overlapping panels. This, in turn, could be a mechanical consequence of an evolution of individuals fixed effects themselves, from year to year.

We conclude that the rise in sorting stemming from pure matching and composition effects is probably even stronger than the one we measure. The variance of firm effects, measured for the same years on subsequent, overlapping panels diminishes strongly around 2008, and again in 2012, likely a reflection of changes in firms pay policies

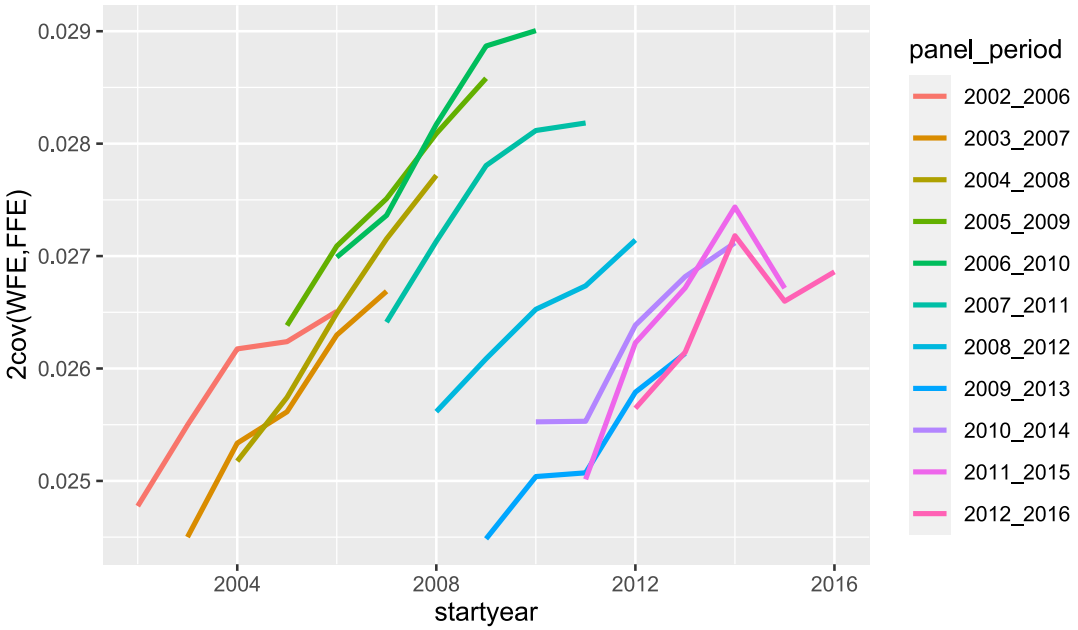
during and after the financial crisis and the Eurozone crisis (Figure A9). The decrease of premium variance could explain the decline in estimated sorting, for a given year, when measured in subsequent, overlapping panels around the time of the financial crisis.

Figure A7: Sorting, rolling panels



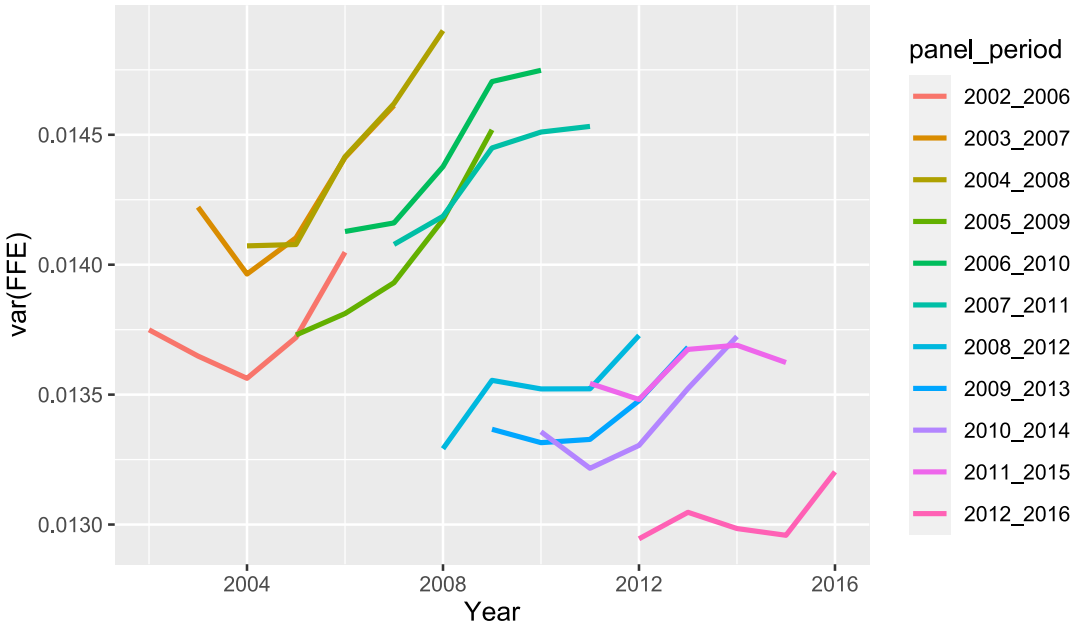
Note: All firms and individuals in firms with at least 1 employee are included. Only individuals employed for at least 360 days by the same firm during the year are included for a given year. Individuals and firms in public administration are not included. Fixed effects estimates by panel of five years periods, split sample bias correction with firm splitting. For each panel, log wage variance decomposition is computed separately for each year.

Figure A8: Sorting estimated with year fixed effects, rolling panels



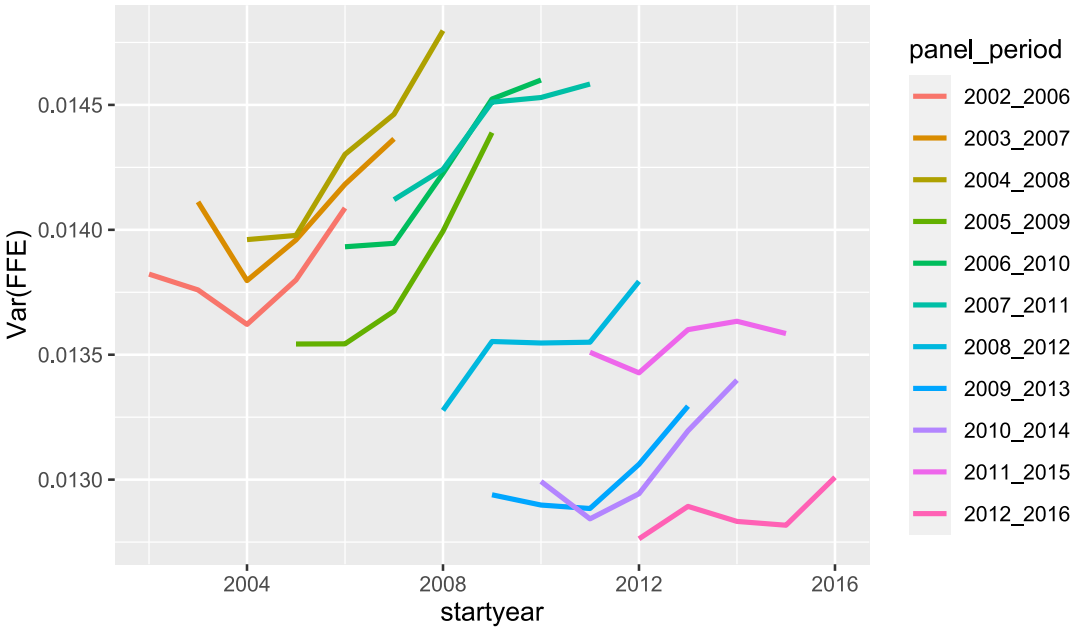
Note: All firms and individuals in firms with at least 1 employee are included. Only individuals employed for at least 360 days by the same firm during the year are included for a given year. Individuals and firms in public administration are not included. Fixed effects estimates via AKM with year fixed effects and cubic function of (age-40), by panel of five years periods, split sample bias correction with firm splitting. For each panel, log wage variance decomposition is computed separately for each year.

Figure A9: Var(FFE), rolling panels



Note: 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. Fixed effects estimates by panel of five years periods, split sample bias correction with firm splitting. For each panel, log wage variance decomposition is computed separately for each year.

Figure A10: Var(FFE) estimated with year fixed effects, rolling panels



Note: 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. Fixed effects estimates via AKM with year fixed effects and cubic function of (age-40), by panel of five years periods, split sample bias correction with firm splitting. For each panel, log wage variance decomposition is computed separately for each year.

F Fixed effects regressions

We regressed at the observation-level the estimated firms and workers fixed effects (without correction) on occupation, sex and industry (table A10). This is similar to a firm level regression of FFE on industry, occupational shares and proportion of women, weighted by worker-year observations, or a individual level regression of WFE on occupation, sex, and industry shares (in year), weighted by worker-year observations. Estimates are relatively stables from period 1 to period 3. We do not observe strong patterns where a change in the association between an occupation, industry or sex and each fixed effect would imply an increase in sorting, though these results are mostly descriptive. They do not account for interactions between terms (if, for instance, women's shares in industries changed over the period) and compositional effects (for instance, because sex is associated with sorting, an increase of the share of the minority would increase sorting).

Table A10: Fixed effects regressions on occupation, sex and industry

		2002-2006					2012-2016				
		WFE	se		FFE	se	WFE	se		FFE	se
Intercept		-0.138	0.002		-0.089	0.002	-0.181	0.002		-0.050	0.002
Occupation											
20	Entrepreneurs	1.293	0.004		0.033	0.003	1.230	0.007		0.026	0.003
32	Professionals	0.724	0.028		0.041	0.013	* 0.766	0.009		0.049	0.005
35	Artists and media professionals	0.567	0.007		0.142	0.007	0.585	0.011		0.110	0.009
37	Managers	0.777	0.004		0.116	0.003	0.717	0.003		0.093	0.003
38	Engineers	0.665	0.003		0.112	0.003	0.627	0.005		0.085	0.003
42	Primary school teachers	0.407	0.012		0.032	0.008	0.305	0.009		0.008	0.006 ns
43	Health and social workers	0.360	0.005		0.072	0.005	0.336	0.004		0.025	0.003
45	Public admin. intermediates	0.254	0.009		0.029	0.008	0.097	0.01		0.003	0.004 ns
46	Business admin. intermediates	0.273	0.003		0.082	0.003	0.279	0.003		0.060	0.003
47	Technicians	0.195	0.003		0.086	0.003	0.208	0.003		0.055	0.003
48	Intermediate supervisors	0.226	0.003		0.062	0.003	0.227	0.003		0.042	0.003
52	Public admin. clerks	0.008	0.006	ns	0.037	0.005	0.019	0.004		0.035	0.004
53	Security agents	-0.045	0.005		-0.017	0.006	* 0.002	0.006	ns	-0.029	0.005
54	Business admin. clerks	0.094	0.002		0.057	0.003	0.140	0.005		0.037	0.003
55	Retail salespersons	0	ref		0	ref	0	ref		0	ref
56	Personal service employees	-0.016	0.005	*	0.023	0.005	-0.052	0.004		-0.016	0.004
62	Skilled manufacturing workers	-0.023	0.003		0.049	0.003	0.010	0.003	*	0.035	0.003
63	Skilled artisans	-0.033	0.002		0.022	0.003	-0.002	0.003	ns	0.013	0.003
64	Drivers	-0.081	0.003		-0.021	0.004	-0.094	0.01		-0.045	0.004
65	Handling, transport skilled workers	-0.092	0.003		0.057	0.003	-0.046	0.005		0.045	0.004
67	Unskilled manufacturing workers	-0.115	0.003		0.032	0.003	-0.083	0.003		0.018	0.003
68	Unkilled artisans	-0.159	0.003		-0.061	0.003	-0.101	0.004		-0.058	0.003
69	Farm workers	-0.059	0.04	ns	0.053	0.024	. -0.066	0.018		0.014	0.018 ns
Industry											
AC	Farming and industry	0.023	0.003		0.052	0.003	0.044	0.003		0.042	0.003
DE	Utilities	0.025	0.005		0.079	0.005	0.033	0.005		0.078	0.004
F	Construction	0.019	0.003		0.058	0.003	0.021	0.003		0.097	0.003
G	Commerce	0	ref		0	ref	0	ref		0	ref
H	Transport	0.018	0.003		0.015	0.004	0.029	0.014	*	0.025	0.005
I	Hotels, tourism, catering	-0.018	0.005		-0.030	0.007	0.011	0.004	*	-0.004	0.004 ns
J	Media	-0.009	0.004	.	0.079	0.004	-0.014	0.005	.	0.020	0.004
K	Financial services	0.011	0.005	.	0.132	0.005	0.062	0.006		0.086	0.004
LM	Real estate, professional services	-0.009	0.005	.	0.015	0.005	* 0.072	0.005		0.021	0.004
N	Administrative services	0.009	0.003	*	0.005	0.003	ns -0.040	0.004		-0.014	0.004
OPQ	Health, educ. and public admin.	0.010	0.005	.	-0.029	0.004	-0.036	0.004		-0.069	0.004
R	Arts and recreation	0.034	0.014	.	0.001	0.013	ns 0.052	0.014		-0.012	0.011 ns
STU	Other	0.030	0.009		-0.034	0.009	-0.034	0.007		-0.047	0.006
Sex											
	Women	-0.143	0.001		-0.005	0.001	-0.115	0.001		-0.007	0.001

Note : all parameter estimates significant at the 0.001 level, except if indicated : *$p < 0.01$; .$p < 0.05$; ns: non significant

All firms and individuals in firms with at least 2 employees 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.

G Correlated random effects model

Following Bonhomme, Lamadon and Manresa (2019), we perform their same conservative sample selection (in addition to the criteria already outlined in Section 1) in order to capture individual job changes between existing firms as precisely as possible. In practice, for the first period in our analysis, we drop observations in years 2003 to 2005. We further keep firms that have at least one worker in both 2002 and 2006 and workers present in years 2002 and 2006. We do the same for the other two periods (2007-2011 and 2012-2016).

We estimate a non-linear static earnings model of the type:

$$Y_{i,t} = a(k(i,t)) + b(k(i,t)) \times \alpha_i + \epsilon_{i,t} \tag{10}$$

Where i denotes the individual, t the year, Y is the demeaned log hourly wage, and $k(i,t)$ the cluster to which is assigned the firm in which individual i is employed in year t. We follow the estimation procedure of Bonhomme, Lamadon and Manresa (2019) such that the α_i are random effects and the number of clusters is equal to 10. Identification still come from mobility across clusters.

Once estimated the model parameters, we perform a similar log-wage decomposition as in Equation 4 by working with a linear projection of log-wage on worker and firm types. The results of the variance decomposition is reported in Table A11.

Table A11: Variance decomposition (x100) - non-linear static model

	$\frac{Var(\alpha)}{Var(Y)}$	$\frac{Var(\Psi)}{Var(Y)}$	$\frac{2Cov(\alpha,\Psi)}{Var(Y)}$	$\frac{Var(\epsilon)}{Var(Y)}$	$Corr(\alpha,\Psi)$
2002-2006	66.49	2.23	14.52	16.76	53.37
2007-2011	64.96	2.14	15.18	17.72	58.66
2012-2016	63.87	2.02	16.04	18.07	62.40

Period samples are selected as explained in Section G. α denotes the worker effect, and Ψ denotes the cluster effect. Estimates are expressed in percentage of overall variance of log-wage (Y).