Pour qui sont les bons salaires?

Une estimation débiaisée du rôle des appariements dans les inégalités

Damien Babet, Olivier Godechot, Marco G. Palladino

Journées de méthodologie statistique, 30 mars 2022

INSEE & Sciences Po & Sciences Po

Table of contents

Introduction

Motivation

- *Les inégalités de salaire* ont augmentés dans la plupart des pays développés au cours des dernières décennies

  $\rightarrow$ La France est une **exception**

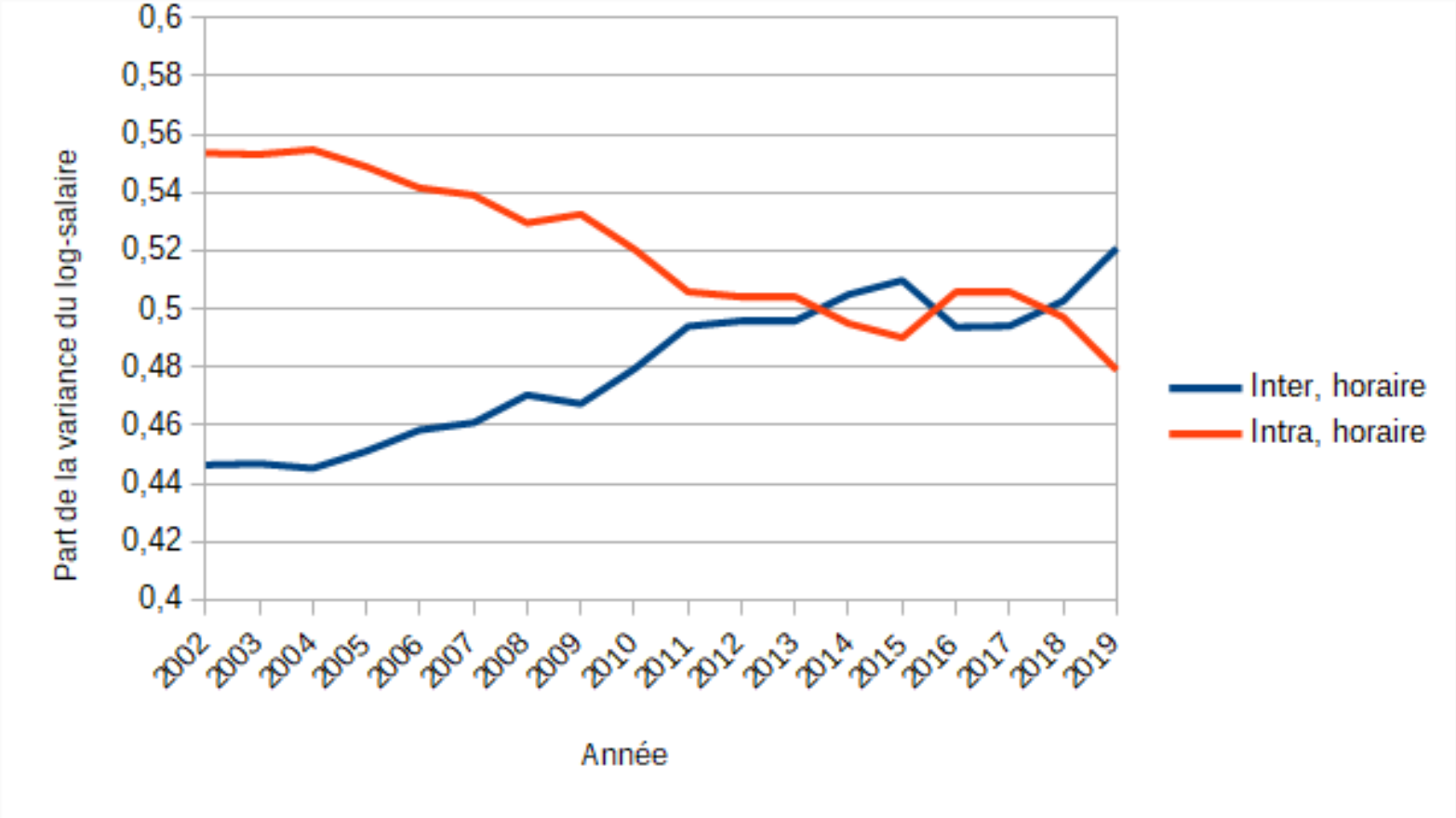
- *Les inégalités de salaire entre firmes* constituent une part croissante des inégalités totales dans la plupart des pays.

  $\rightarrow$ La France **n'est pas une exception**

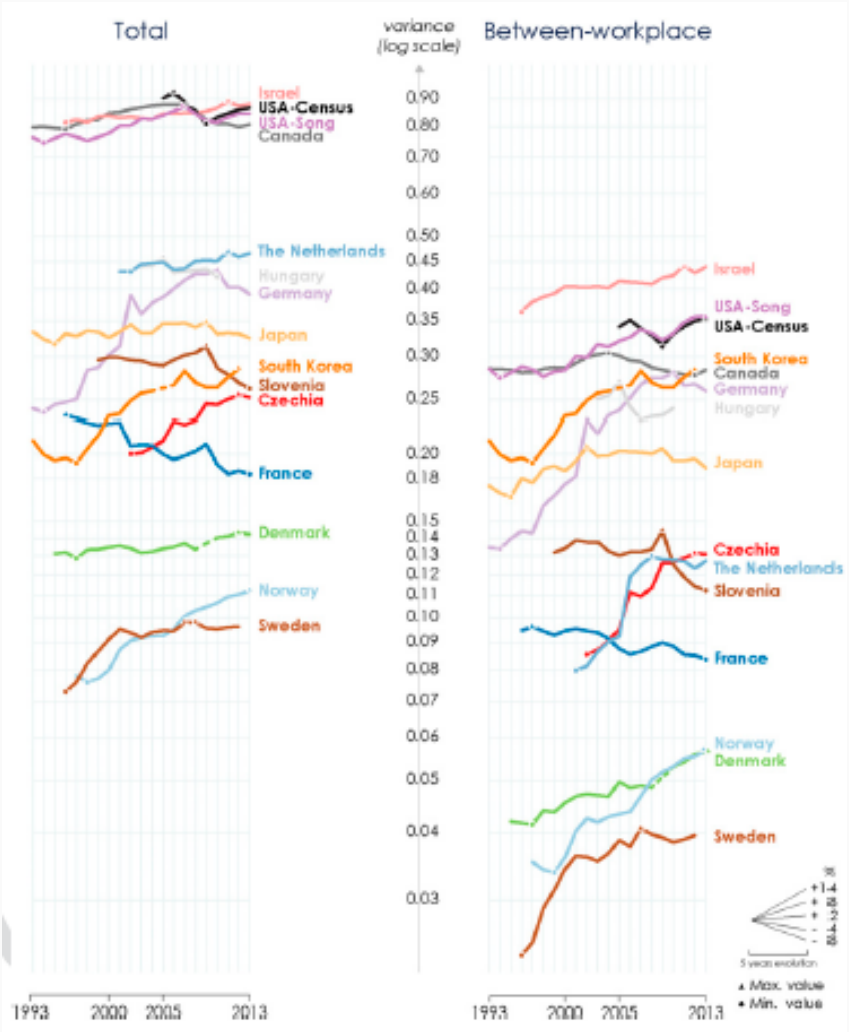
Nos apports

- Nous exploitons un nouveau *panel DADS* ***exhaustif***
- Nous répliquons l'approche de Card et al. (2013) et Song et al. 2019 pour la France et montrons que le *sorting augmente*
- Cette croissance du sorting est robuste à la correction des biais
- Nous explorons les **déterminants de cette dynamique**

Variance du log-salaire, parts intra- et inter-entreprises



Context: rising wage inequalities in rich countries



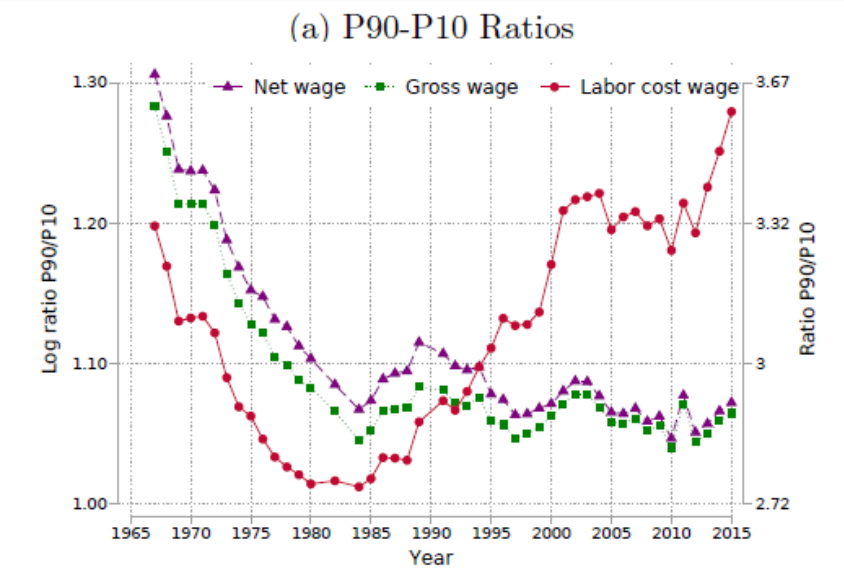
Total variance and between-workplace variance of log-wages

Tomaskovic-Devey et al. 2020

Wage inequality has risen for several decades in most rich countries. France is an exception.

Between-workplace inequalities is a rising part of wage inequalities in most countries. France is no exception

Context: a French exception ?



Wage inequality ratio in France, 1967-2015

Bozio, Breda, and Guillot 2020

The French exception of decreasing wage inequality disappears when wages are measured at the labor cost level.

Les données

DADS: petit et grand panel

Les DADS ne sont pas organisées en panel, nous reconstruisons un panel (quasi) exhaustif

- Il existe un "petit" panel (au 1/24 puis 1/12) utilisé par l'article AKM original, mais la taille compte!
- Les identifiants des DADS exhaustives sont des pseudonymes qui changent chaque année
- Mais les fichiers annuels contiennent des lignes pour tous les postes d'une personne pendant l'année t... et pendant l'année $t-1$
- Il est donc possible d'apparier les postes $t-1$ du fichier de l'année N aux postes t du fichier de l'année $N-1$, puis d'apparier les identifiants individuels d'une année à l'autre.
- L'appariement échoue pour quelques cas très minoritaires, et pour les personnes sans poste salarié pendant une année civile.

panel DADS: la clé d'appariement

```
1  DATA y_1t; SET a;
2       pseudoid=COMPRESS(sexe!!"#"!!siren!!"#"!!nic!!"#"
3           !!nbheur!!"#"!!datdeb!!"#"!!datfin!!"#"!!duree
4           !!"#"!!comr!!"#"!!comt!!"#"!!sonde);
5  RUN;
6
7  DATA yt_1; SET b;
8       pseudoid_b=COMPRESS(sexe!!"#"!!siren!!"#"!!nic!!"#"
9           !!nbheur_1!!"#"!!datdeb_1!!"#"!!datfin_1!!"#"
10          !!duree_1!!"#"!!comr_1!!"#"!!comt_1!!"#!!sonde_1);
11 RUN;
```

DADS exhaustive panel: the SAS code - the merge

```
1  PROC SQL;
2      CREATE TABLE ab AS SELECT * FROM y_1t AS aa
3          FULL JOIN  yt_1 AS bb
4              ON aa.pseudoid=bb.pseudoid_b
5              GROUP BY aa.s_brut, aa.pseudoid
6              HAVING ABS(aa.s_brut-bb.s_brut_1)=
7                  MIN(ABS(aa.s_brut-bb.s_brut_1))
8              AND (0<=bb.age_B-aa.age<2 or age_b=. or age=.)
9              ORDER BY aa.pseudoid,bb.s_brut_1;
10 QUIT;
```

Résultats

Le modèle AKM

Le point de départ est modèle additif du log-salaire de Abowd, Kramarz, and Margolis 1999 (par la suite *AKM*):

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

- y_{it} est le logarithme du salaire horaire du salarié $i = 1, 2, ..., N$ pendant l'année $t = 1, ..., T$, centré par année.
- x_{it} sont les covariables variant dans le temps (âge et âge au carré)
- θ_i est l'effet fixe du salarié i
- ψ_j est l'effet fixe de la firme $j = 1, 2..., J$, qui emploit le salarié i pendant l'année t
- u_{it} est le terme d'erreur.

Hypothèses principales: Pas d'interactions, mobilité exogène, les effets fixes sont relatifs (au sein d'un groupe de salariés et d'entreprises connectés par des mobilités)

Décomposition de la variance

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

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

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

Card, Heining, and Kline 2013 introduisent l'équation 2 et isolent $2Cov(\theta, \psi)$ comme mesure du *sorting*. Song et al. 2019 décomposent davantage entre des composantes inter- et intra-firmes, les premières comprenant les effets fixes des entreprises (ou premiums), le *sorting* et la "ségrégation".

Décomposition de la variance totale

	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		0,206		-0,008
Var (WFE)	0,165	77,1	0,166	78,8	0,158	76,6	-0,007	90,2
Var (FFE)	0,030	14,0	0,029	14,0	0,025	12,0	-0,005	65,2
Var(Xb)	0,024	11,1	0,034	16,0	0,016	7,8	-0,008	97,4
Var(u)	0,009	4,1	0,008	4,0	0,008	4,0	-0,001	6,8
2*Cov(WFE,FFE)	-0,004	-1,8	0,000	-0,2	0,005	2,2	0,008	-106,4
2*Cov(WFE,Xb)	-0,012	-5,7	-0,029	-13,7	-0,007	-3,2	0,006	-69,5
2*Cov(FFE,Xb)	0,002	1,2	0,002	1,1	0,001	0,6	-0,001	16,3
Between-firm	0,091	42,2	0,095	45,2	0,099	47,9	0,008	-103,9
Within-firm	0,124	57,8	0,115	54,8	0,108	52,1	-0,016	203,9
N* (main connected set)	41 703 340		44 733 304		47 038 310			

▸ Detailed between and within decomposition

Décomposition de la variance totale

	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		0,206		-0,008
Var (WFE)	0,165	77,1	0,166	78,8	0,158	76,6	-0,007	90,2
Var (FFE)	0,030	14,0	0,029	14,0	0,025	12,0	-0,005	65,2
Var(Xb)	0,024	11,1	0,034	16,0	0,016	7,8	-0,008	97,4
Var(u)	0,009	4,1	0,008	4,0	0,008	4,0	-0,001	6,8
2*Cov(WFE,FFE)	-0,004	-1,8	0,000	-0,2	0,005	2,2	0,008	-106,4
2*Cov(WFE,Xb)	-0,012	-5,7	-0,029	-13,7	-0,007	-3,2	0,006	-69,5
2*Cov(FFE,Xb)	0,002	1,2	0,002	1,1	0,001	0,6	-0,001	16,3
Between-firm	0,091	42,2	0,095	45,2	0,099	47,9	0,008	-103,9
Within-firm	0,124	57,8	0,115	54,8	0,108	52,1	-0,016	203,9
N* (main connected set)	41 703 340		44 733 304		47 038 310			

▸ Detailed between and within decomposition

Décomposition de la variance totale

	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		0,206		-0,008
Var (WFE)	0,165	77,1	0,166	78,8	0,158	76,6	-0,007	90,2
Var (FFE)	0,030	14,0	0,029	14,0	0,025	12,0	-0,005	65,2
Var(Xb)	0,024	11,1	0,034	16,0	0,016	7,8	-0,008	97,4
Var(u)	0,009	4,1	0,008	4,0	0,008	4,0	-0,001	6,8
2*Cov(WFE,FFE)	**-0,004**	-1,8	0,000	-0,2	0,005	2,2	0,008	-106,4
2*Cov(WFE,Xb)	-0,012	-5,7	-0,029	-13,7	-0,007	-3,2	0,006	-69,5
2*Cov(FFE,Xb)	0,002	1,2	0,002	1,1	0,001	0,6	-0,001	16,3
Between-firm	0,091	42,2	0,095	45,2	0,099	47,9	0,008	-103,9
Within-firm	0,124	57,8	0,115	54,8	0,108	52,1	-0,016	203,9
N* (main connected set)	41 703 340		44 733 304		47 038 310			

▸ Detailed between and within decomposition

Décomposition de la variance totale

	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		0,206		-0,008
Var (WFE)	0,165	77,1	0,166	78,8	0,158	76,6	-0,007	90,2
Var (FFE)	0,030	14,0	0,029	14,0	0,025	12,0	-0,005	65,2
Var(Xb)	0,024	11,1	0,034	16,0	0,016	7,8	-0,008	97,4
Var(u)	0,009	4,1	0,008	4,0	0,008	4,0	-0,001	6,8
2*Cov(WFE,FFE)	**-0,004**	-1,8	**0,000**	-0,2	0,005	2,2	0,008	-106,4
2*Cov(WFE,Xb)	-0,012	-5,7	-0,029	-13,7	-0,007	-3,2	0,006	-69,5
2*Cov(FFE,Xb)	0,002	1,2	0,002	1,1	0,001	0,6	-0,001	16,3
Between-firm	0,091	42,2	0,095	45,2	0,099	47,9	0,008	-103,9
Within-firm	0,124	57,8	0,115	54,8	0,108	52,1	-0,016	203,9
N* (main connected set)	41 703 340		44 733 304		47 038 310			

▸ Detailed between and within decomposition

Décomposition de la variance totale

	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		0,206		-0,008
Var (WFE)	0,165	77,1	0,166	78,8	0,158	76,6	-0,007	90,2
Var (FFE)	0,030	14,0	0,029	14,0	0,025	12,0	-0,005	65,2
Var(Xb)	0,024	11,1	0,034	16,0	0,016	7,8	-0,008	97,4
Var(u)	0,009	4,1	0,008	4,0	0,008	4,0	-0,001	6,8
2*Cov(WFE,FFE)	**-0,004**	-1,8	**0,000**	-0,2	**0,005**	2,2	0,008	-106,4
2*Cov(WFE,Xb)	-0,012	-5,7	-0,029	-13,7	-0,007	-3,2	0,006	-69,5
2*Cov(FFE,Xb)	0,002	1,2	0,002	1,1	0,001	0,6	-0,001	16,3
Between-firm	0,091	42,2	0,095	45,2	0,099	47,9	0,008	-103,9
Within-firm	0,124	57,8	0,115	54,8	0,108	52,1	-0,016	203,9
N* (main connected set)	41 703 340		44 733 304		47 038 310			

▸ Detailed between and within decomposition

Corriger le biais de mobilité limité

Expression du biais

Le *sorting* mesuré à partir du modèle AKM souffre d'un **biais de mobilité limité**, qui peut aussi être décrit comme un biais de paramètre incident, ou du surapprentissage. Nous suivons l'analyse formelle de Kline, Saggio, and Sølvsten 2020. Le modèle AKM est noté de manière concise:

$$y_i = z_i'\alpha + u_i \tag{5}$$

- $\alpha = (\beta, \theta, \psi)$ le vecteur de paramètres de longueur $k = 2 + N + J$ et z_i le vecteur non-aléatoire des régresseurs, incluant les indicatrices de la firme et du salarié.
- $S_{zz} = \sum_{i=1}^{N*} z_i z_i'$ la design matrix (de plein rang lorsqu'on limite la régression à la principale composante connectée et qu'on fixe la moyenne des effets fixes à 0)
- Nos objets d'intérêt: des formes quadratiques $\omega = \alpha' A \alpha$ pour une matrice symétrique $A \in \mathrm{R}^{k \times k}$ que nous choisissons. Cela inclut la variance et la covariance.

Expression du biais

Un estimateur plug-in naïf pour ω donne: $\hat{\omega}^{PI} = \hat{\alpha}' A \hat{\alpha}$ avec $\hat{\alpha}$ l'estimateur OLS des paramètres
$\hat{\alpha} = {S_{zz}}^{-1} \sum_{i=1}^{N*} z_i y_i = \alpha + {S_{zz}}^{-1} \sum_{i=1}^{N*} z_i u_i$.

$$\mathrm{E}[\hat{\omega}^{PI}] - \omega = trace(AV[\hat{\alpha}]) = \sum_{i=1}^{N^*} B_{ii} \sigma_i^2 \tag{6}$$

Avec $B_{ii} = z_i' S_{zz}^{-1} A S_{zz}^{-1} z_i$ représentant l'influence de chaque terme d'erreur sur l'estimateur plug-in.

▸ Limited mobility bias: literature

Correction par split-sampling

Un estimateur plug-in par split-sampling pour une forme quadratique ω devient $\hat{\omega}^{SP} = \hat{\alpha}_0' A \hat{\alpha}_1$ avec $\hat{\alpha}_s$ un estimateur OLS dans l'échantillon $I_s, s = 0, 1$ de taille N_s : $\hat{\alpha}_s = S_{zz,s}^{-1} \sum_{i \in s}^{N_s} y_i z_i = \alpha + S_{zz,s}^{-1} \sum_{i \in s}^{N_s} u_i z_i$ que l'on peut exprimer comme $\hat{\alpha}_s = \alpha + \epsilon_s$.

On montre que :

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

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

Ce terme b_{lm} est nul sous des conditions raisonnables: 1. $\mathrm{E}[u|z] = 0$ et 2. indépendance de $u_i, i \in I_1$ et $u_j, j \in I_0$.

Résultats de la correction par split-sampling

	2002-2006 Variance	2007-2011 Variance	2012-2016 Variance	Change
Uncorrected				
Var (WFE)	0,165	0,166	0,160	-0,005
Var (FFE)	0,030	0,029	0,025	-0,005
2*Cov(WFE,FFE)	-0,004	0,000	0,004	0,007
Corrected				
Cov(WFE_H0,WFE_H1)	0,152	0,156	0,145	-0,007
Cov(FFE_H0,FFE_H1)	0,016	0,016	0,013	-0,003
2*Cov(WFE_H0,FFE_H1)	0,02012	0,02452	0,02459	0,004
2*Cov(WFE_H1,FFE_H0)	0,01992	0,02410	0,02241	0,002
Mean 2*Cov(WFE,FFE)	0,02002	0,02431	0,02350	0,003
N*	29 543 074	32 312 274	32 006 188	

Correction par clustering

Nous suivons Bonhomme, Lamadon, and Manresa 2019: les entreprises sont regroupées en 10 clusters en fonction de la similarité de leurs distributions de salaires. Un algorithme de clustering minimise la distance entre les fonctions de répartition des log-salaires (sur 20 points).

AKM est alors estimé avec des effets fixes par cluster et non plus par entreprise, sous l'hypothèse que les entreprises d'un même cluster ont les mêmes premiums.

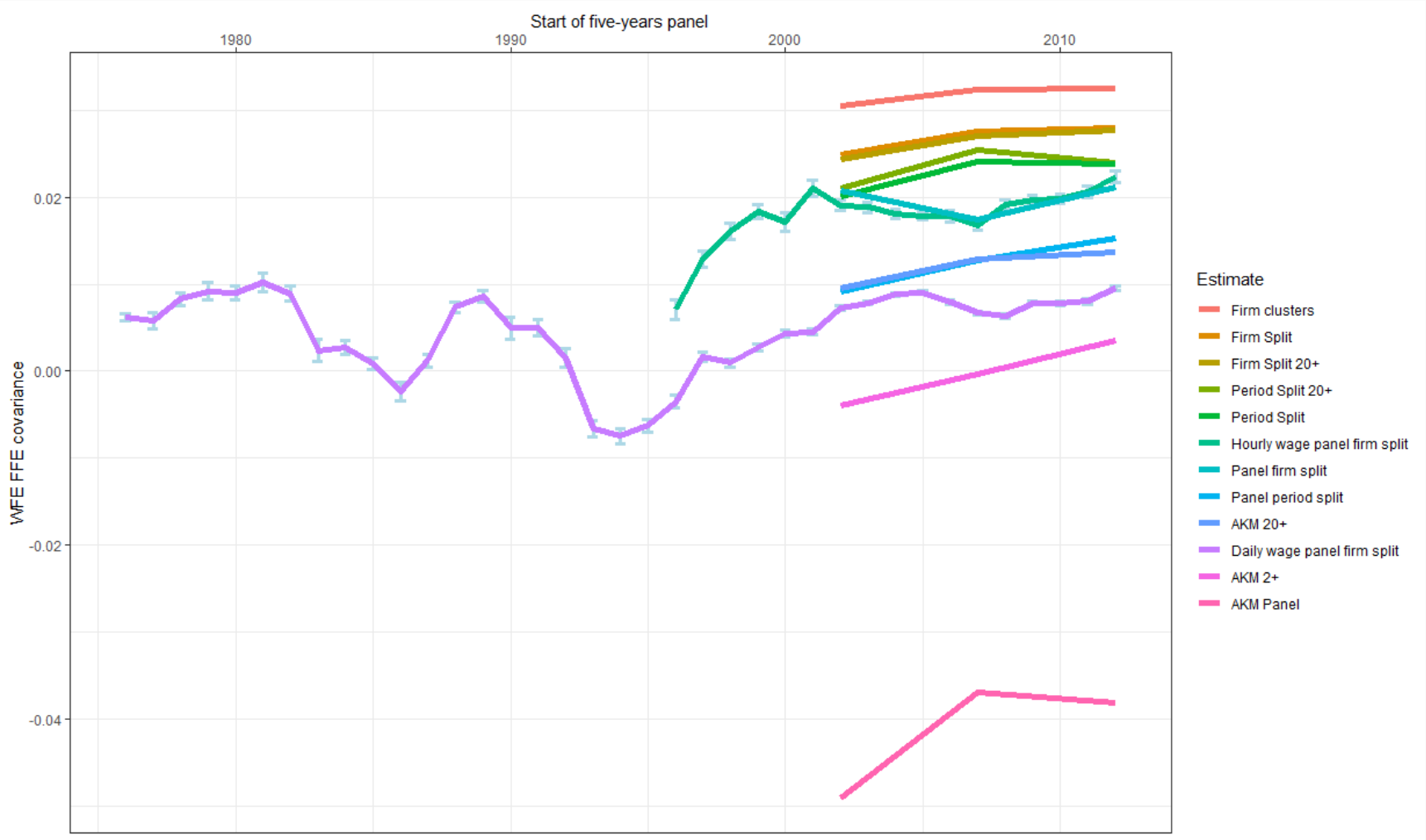
Les clusters sont gros, le biais de mobilité limité devient négligeable.

On trouve une correction du biais plus forte que pour le split-sampling.

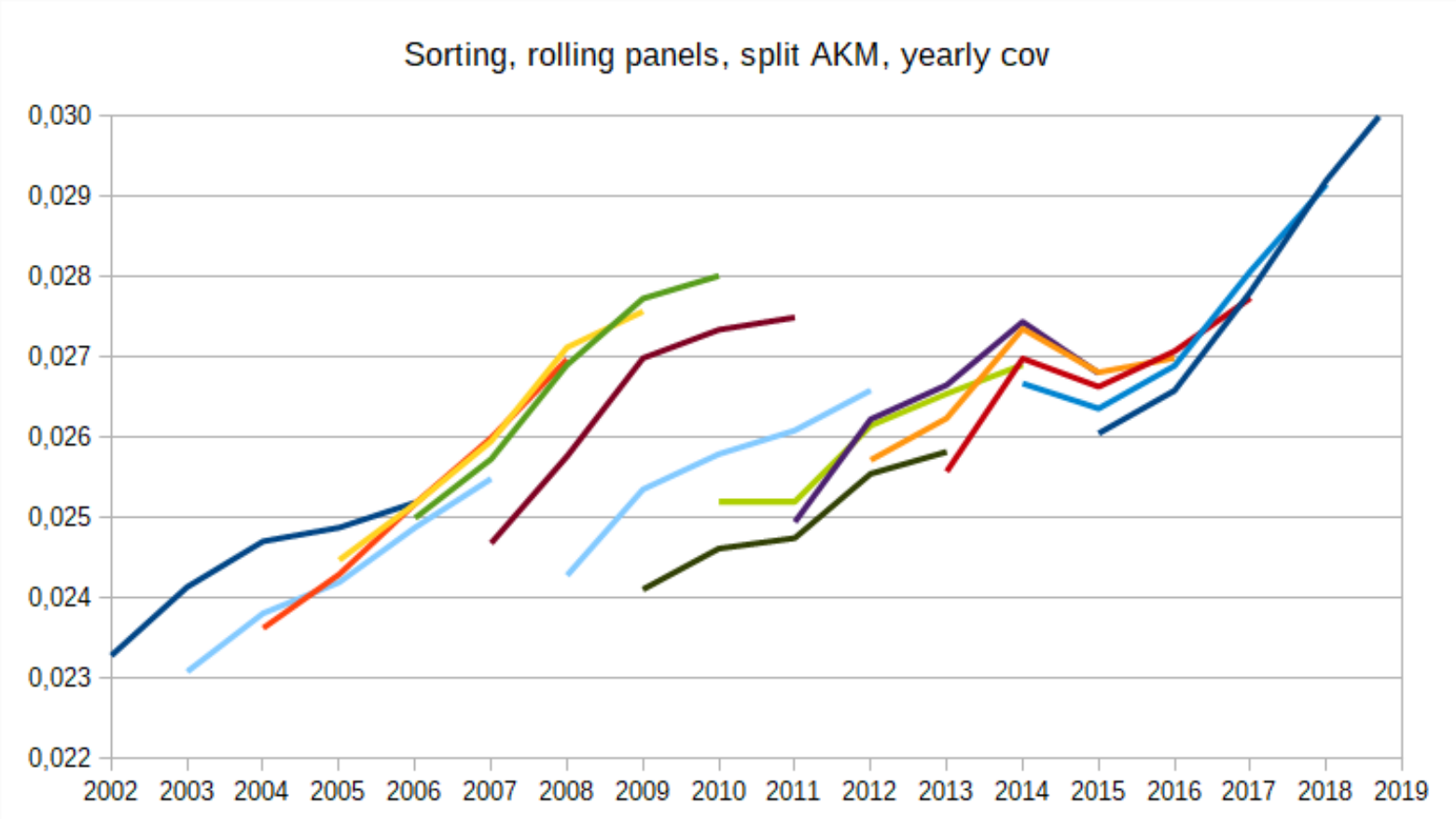
Résultat de la correction par clustering

	2002-2006	2007-2011	2012-2016
Uncorrected			
Total variance	0,214	0,211	0,207
Var (WFE)	0,165	0,166	0,160
Var (FFE)	0,030	0,029	0,025
Var(Xb)	0,024	0,034	0,016
Var(u)	0,009	0,008	0,007
2*Cov(WFE,FFE)	-0,004	0,000	0,004
2*Cov(WFE,Xb)	-0,012	-0,029	-0,007
2*Cov(FFE,Xb)	0,002	0,002	0,001
Corrected			
Total variance	0,214	0,205	0,198
Var (WFE)	0,139	0,142	0,130
Var (CFE)	0,008	0,008	0,007
Var(Xb)	0,032	0,057	0,025
Var(u)	0,016	0,016	0,013
2*Cov(WFE,CFE)	0,031	0,032	0,033
2*Cov(WFE,Xb)	-0,013	-0,051	-0,010
2*Cov(CFE,Xb)	0,002	0,002	0,001

Augmentation du sorting, toutes estimations



Sorting 2002-2019, rolling panel



Comprendre la croissance du sorting

Decomposition de la covariance

Nous prenons comme unité d'observation les firmes i et des covariables X. Nous rassemblons les périodes 1 (2002-2006) et 3 (2012-2016) pour construire un panel, sur lequel nous régressons les effets fixes (salariés et employeurs) estimés en première étape

$$WFE_{it} = \alpha_w + W_i + \beta_w * X_{it} + \epsilon_{it}$$
$$FFE_{it} = \alpha_f + F_i + \beta_f * X_{it} + \omega_{it}$$

où W_i et F_i constituent des *meta-effets fixes* dans chaque régression, respectivement.

D'un point de vue économique, la covariance entre W_i et F_i mesure à quel point le *sorting* est expliqué par des **dynamiques inter-firmes**

Modèle complet

On peut décomposer la covariance pour la période $t \in \{1, 3\}$:

$$
\begin{aligned}
Cov(WFE_t, FFE_t) = Cov(\alpha_w + W_{it} + \beta_w * X_{it} + \epsilon_{it}, \alpha_f + F_i + \beta_f * X_{it} + \omega_{it}) = \\
\boldsymbol{Cov(W_i, F_i)} + Cov(W_i, \beta_f * X_{it}) + Cov(W_i, \omega_{it}) + \\
Cov(\beta_w * X_{it}, F_t i + \boldsymbol{Cov(\beta_w * X_{it}, \beta_f * X_{it})} + Cov(\beta_w * X_{it}, \omega_{it}) + \\
Cov(\epsilon_{it}, F_t i + Cov(\epsilon_{it}, \beta_f * X_{it}) + \boldsymbol{Cov(\epsilon_{it}, \omega_{it})}
\end{aligned}
$$

On s'intéresse principalement aux termes en gras, qui mesurent comment le *sorting* est directement expliqué par les *meta-effets fixes*, les covariables et les residus, respectivement, compte tenu des autres interactions.

Modèle complet - Résultats

Table 1: Décomposition de l'évolution du sorting entre les périodes 1 et 3

		FFE regression					
WFE reg.	Residuals	FE	VA	Size	% Female	Age	Occupation
Residuals	**7.90%**	49.24%	5.12%	-0.21%	-0.57%	1.32%	6.51%
Fixed Effects	-30.76%	**30.91%**	7.31%	6.80%	-1.45%	-0.83%	32.99%
Value Added	-4.48%	5.05%	**1.00%**	-0.01%	0.73%	-0.16%	4.41%
Size	0.13%	0.02%	0.00%	**0.56%**	0.14%	-0.04%	2.70%
% Female	1.64%	3.20%	1.18%	0.35%	**0.26%**	-0.34%	4.19%
Avg Age	2.11%	13.15%	0.87%	0.35%	1.12%	**-0.25%**	1.33%
Occupation	-54.72%	-22.40%	7.48%	-5.60%	4.14%	-0.48%	**18.08%**

Modèles partiels

Deux autres approches possibles:

1. **Addition**:
    - La référence est un modèle sobre, avec uniquement les méta-effets fixes
    - Pour chaque co-variable, on regarde à quel point son **addition** entraîne une diminution de la covariance des résidus
2. **Omission**:
    - La référence est le modèle complet
    - Pour chaque covariable, on regarde à quel point son **omission** modifie la covariance des résidus.

Covariance des résidus, en fonction des covariables ajoutées ou omises

Covariance des résidus (en % de l'évolution du *sorting*)

	"Addition"	*"Omission"*
Meta-FE (Benchmark)	8.29%	
Value-Added	8.21%	7.81%
Occupation	7.67%	8.26%
Incidence of Female	8.30%	7.89%
Avg. Age	8.44%	7.82%
Size	8.25%	7.91%

Effets fixes et nombre d'emplois

Table 2: Nombre d'observation selon la période et le signe des effets fixes

	Entreprises présentes en périodes 1 et 3			
Period	***F>0, W>0***	***F>0, W<0***	***F<0, W<0***	***F<0, W>0***
1	7 622 024	5 590 596	7 805 226	3 668 069
3	9 297 447	6 487 760	7 958 418	3 596 627
3-1	21,98%	16,05%	1,96%	-1,95%
	1 675 423	897 164	153 192	-71 442

	Entreprises présentes seulement une seule période			
Period	***F>0, W>0***	***F>0, W<0***	***F<0, W<0***	***F<0, W>0***
1	2 129 808	2 255 253	3 274 813	1 973 691
3	2 478 767	2 100 389	5 388 803	2 164 239
3-1	16,38%	-6,87%	64,55%	9,65%
	348 959	-154 864	2 113 990	190 548

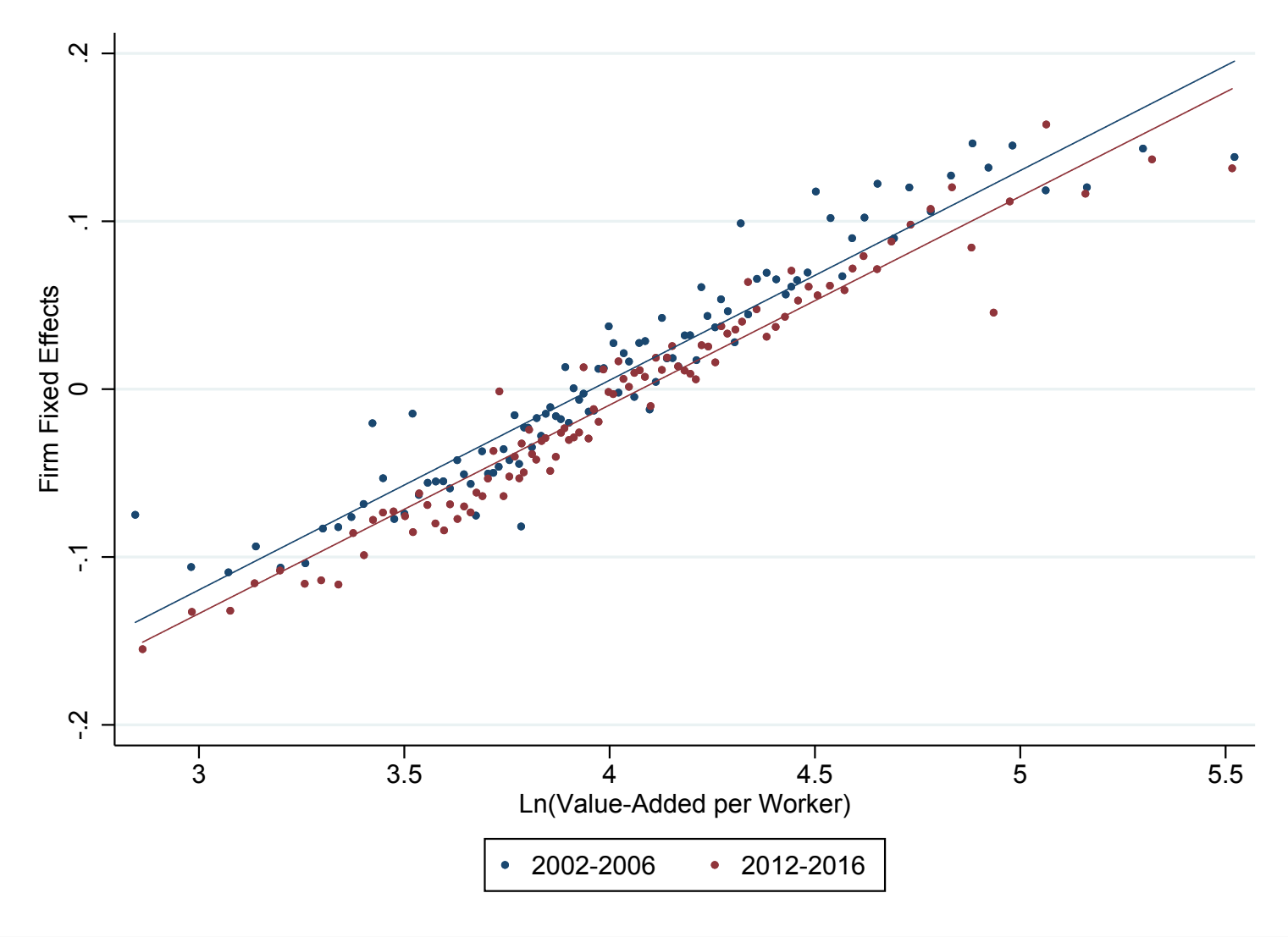
Expliquer le *sorting*

- Nous observons une **augmentation du sorting** ...
- ... qui est **robuste** à de nombreuses méthodes de correction des biais

  $\rightarrow$ Les bons salaires sont, de plus en plus souvent, pour les salariés qui sont déjà les mieux payés
- La dynamique du sorting est notamment liée aux évolutions structurelles de la distribution des professions dans les firmes.
- Nous ne trouvons pas d'effet du **partage de la rente**

  $\rightarrow$ Les firmes les plus productives *n'ont pas* accru plus que d'autres les premiums versés aux salariés

Productivité et partage de la rente, périodes 1 et 3



Appendix : between and within firm decomposition

Results: Between-firm decomposition

		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		0,206		-0,008	
Between-firm variance		0,091	42,2	0,095	45,2	0,099	47,9	0,008	-103,9
Var (m_WFE)		0,058	27,2	0,063	29,7	0,066	32,0	0,008	-97,3
Var (FFE)		0,030	14,0	0,029	14,0	0,025	12,0	-0,005	65,2
Var(m_Xb)		0,004	1,8	0,006	2,6	0,003	1,3	-0,001	16,6
2*Cov(m_WFE,FFE)		-0,004	-1,8	0,000	-0,2	0,005	2,2	0,008	-106,4
2*Cov(m_WFE,m_Xb)		0,000	-0,1	-0,004	-2,0	0,000	-0,2	0,000	1,6
2*Cov(FFE,m_Xb)		0,002	1,2	0,002	1,1	0,001	0,6	-0,001	16,3

Results: Within-firm decomposition

		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		0,206		-0,008	
Within-firm variance		0,124	57,8	0,115	54,8	0,108	52,1	-0,016	203,9
Var (diff_WFE)		0,107	49,9	0,103	49,1	0,092	44,6	-0,015	187,4
Var(diff_Xb)		0,020	9,3	0,028	13,4	0,014	6,5	-0,006	80,8
Var(u)		0,009	4,1	0,008	4,0	0,007	3,3	-0,002	26,8
2*Cov(diff_WFE,diff_Xb)		-0,012	-5,6	-0,025	-11,7	-0,006	-3,0	0,006	-71,1
2*Cov(diff_WFE,u)		0,000	0,0	0,000	0,0	0,000	0,0	0,000	0,0
2*Cov(diff_Xb,u)		0,000	0,0	0,000	0,0	0,000	0,0	0,000	0,0

Appendix: Split sampling proof and robustness

Limited mobility bias: literature

The bias was noticed in Abowd, Kramarz, Lengermann, et al. 2004 and a first correction proposed in Andrews et al. 2008, later shown to be too small because of the effect of the real world graph structure (Jochmans and Weidner 2019).

Random effect models such as Borovičková and Shimer 2017 lead to a higher bias correction and higher estimate of the sorting effect. Bonhomme, Lamadon, and Manresa 2019 add a first step of firm clustering.

Several papers use sample-splitting to correct for the bias, generally without theoretical justification (Chanut 2018, Gerard et al. 2018, Schoefer and Ziv 2021, Drenik et al. 2020). On the contrary, Kline, Saggio, and Sølvsten 2020 uses a "leave-one-out" estimator with proof, but its computation is complex and costly.

Card, Heining, and Kline 2013 and Song et al. 2019 do not correct for the bias, and expect that it remains stable in time and do not change the results on sorting effects dynamics. Bonhomme, Holzheu, et al. 2020 do a systematic comparison and find the bias to be stable in time.

Split-sampling in theory

Split-sampling plug-in estimator for the quadratic form ω 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 : $\hat{\alpha}_s = S_{zz,s}^{-1} \sum_{i \in s}^{N_s} y_i z_i = \alpha + S_{zz,s}^{-1} \sum_{i \in s}^{N_s} u_i z_i$ that we can express as $\hat{\alpha}_s = \alpha + \epsilon_s$.

$$\begin{aligned} \mathsf{E}[\hat{\omega}^{SP}] &= \mathsf{E}[\hat{\alpha}_0' A \hat{\alpha}_1] = \mathsf{E}[trace(\hat{\alpha}_0' A \hat{\alpha}_1)] \\ &= \mathsf{E}[trace((A\hat{\alpha}_0)' \hat{\alpha}_1)] = \mathsf{E}[trace(A\hat{\alpha}_0 \hat{\alpha}_1)'] \\ &= trace(A\mathsf{E}[\hat{\alpha}_1 \hat{\alpha}_0{}']) \end{aligned}$$

We further have:

$$\begin{aligned} \mathsf{E}[\hat{\alpha}_1 \hat{\alpha}_0{}'] &= \mathsf{E}[(\alpha + \epsilon_1)(\alpha + \epsilon_0)'] = \alpha\alpha' + \mathsf{E}[\epsilon_1 \epsilon_0'] \\ trace(A\alpha\alpha') &= \omega \end{aligned}$$

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

Split-sampling in theory

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

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

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

This term has null expectation under mild conditions 1. null conditional expectation $\mathrm{E}[u|z] = 0$ and 2. independence of $u_i, i \in I_1$ and $u_j, j \in I_0$. The second condition might be violated, if for instance u are correlated for different years of the same employer / employee pair, which is likely.

Split-sampling in practice

We randomly split each worker's T_i observations in two equal parts (up to a difference of one) to get two equal split samples. We estimate AKM separetely on samples 0 and 1 and, on the set of firms and workers belonging to the main connected component of both samples, we compute $cov(WFE_0, WFE_1)$, $cov(WFE_0, FFE_1)$, $cov(FFE_0, WFE_1)$ and $cov(FFE_0, FFE_1)$ as corrected estimates of variance and covariance.

Since we have $T = 5$ and only keep years when a worker works for the same firm for more than 360 days, workers can be mobile in both splits only in rare cases when they switch job on the calendar year.

We also check results with a second split strategy, dividing all workers in each firms in two splits, similar to Chanut 2018.

Robustness of split-sampling: sample and split effects

	Period split, common sample	Period split, taylored samples	Firm split, taylored samples
2002-2006			
Cov(WFE_H0,WFE_H1)	0,152	0,152	
Cov(FFE_H0,FFE_H1)	0,016	0,017	0,014
2*Cov(WFE_H0,FFE_H1)	0,020	0,020	0,02536
2*Cov(WFE_H1,FFE_H0)	0,020	0,020	0,02447
Mean 2*Cov(WFE,FFE)	0,02002	0,02016	0,02492
2012-2016			
Cov(WFE_H0,WFE_H1)	0,145	0,145	
Cov(FFE_H0,FFE_H1)	0,013	0,013	0,012
2*Cov(WFE_H0,FFE_H1)	0,025	0,025	0,02797
2*Cov(WFE_H1,FFE_H0)	0,022	0,023	0,02793
Mean 2*Cov(WFE,FFE)	0,02350	0,02381	0,02795

Robustness of split-sampling : simulations

	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

Appendice: régressions des effets fixes

Régression des effets fixes

	FFE		WFE	
log VA/salariés	0,03	***	0,04	***
Log taille	0,00		0,00	***
Âge moyen	0,00	***	-0,01	***
Part de femmes	-0,02	.	-0,11	***
Artisans, commerçants, chefs d'entreprise	0,00	ref	0,00	ref
Professions libérales, enseignants cadres du public, clergé	0,07		-0,20	**
Professions de l'information, des arts et des spectacles	-0,27	***	-0,19	**
Cadres administratifs et commerciaux d'entreprise	-0,05		-0,05	
Ingénieurs et cadres techniques d'entreprise	-0,03		-0,12	*
Professeurs des écoles, instituteurs et assimilés	-0,12	.	-0,28	***
Professions intermédiaires de la santé et du travail social	-0,10		-0,54	***
Professions intermédiaires administratives de la fonction publique	-0,77	**	0,06	
Professions intermédiaires administratives et commerciales des entreprises	-0,12	*	-0,42	***
Techniciens	-0,09	.	-0,52	***
Contremaîtres, agents de maîtrise	-0,18	***	-0,54	***
Employés civils et agents de service de la fonction publique	-0,07		-0,59	***
Policiers et militaires	-0,17	**	-0,54	***
Employés administratifs d'entreprise	-0,17	***	-0,52	***
Employés de commerce	-0,15	**	-0,62	***
Personnels des services directs aux particuliers	-0,18	**	-0,61	***
Ouvriers qualifiés de type industriel	-0,15	**	-0,69	***
Ouvriers qualifiés de type artisanal	-0,17	**	-0,67	***
Chauffeurs	-0,15	**	-0,65	***
Ouvriers qualifiés de la manutention, du magasinage et du transport	-0,14	**	-0,70	***
Ouvriers non qualifiés de type industriel	-0,16	**	-0,72	***
Ouvriers non qualifiés de type artisanal	-0,19	**	-0,64	***
Ouvriers agricoles	0,00		-1,01	***

References i

- Abowd, John M, Francis Kramarz, Paul Lengermann, et al. (2004). "Are good workers employed by good firms? A test of a simple assortative matching model for France and the United States". In: *Unpublished Manuscript*.
- Abowd, John M, Francis Kramarz, and David N Margolis (1999). "High wage workers and high wage firms". In: *Econometrica* 67.2, pp. 251–333.
- Andrews, Martyn J et al. (2008). "High wage workers and low wage firms: negative assortative matching or limited mobility bias?" In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171.3, pp. 673–697.
- Bonhomme, Stéphane, Kerstin Holzheu, et al. (2020). *How Much Should we Trust Estimates of Firm Effects and Worker Sorting?* Tech. rep. National Bureau of Economic Research.

References ii

- Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa (2019). "A distributional framework for matched employer employee data". In: *Econometrica* 87.3, pp. 699–739.
- Borovičková, Katarına and Robert Shimer (2017). *High wage workers work for high wage firms*. Tech. rep. National Bureau of Economic Research.
- Bozio, Antoine, Thomas Breda, and Malka Guillot (2020). "The Contribution of Payroll Taxation to Wage Inequality in France". In.
- Card, David, Jörg Heining, and Patrick Kline (2013). "Workplace heterogeneity and the rise of West German wage inequality". In: *The Quarterly journal of economics* 128.3, pp. 967–1015.
- Chanut, Nicolas (2018). "Distinguishing Between Signal and Noise in the Measurement of the Firm Wage Premium". In: *Available at SSRN 3470571*.

References iii

- Drenik, Andres et al. (2020). *Paying outsourced labor: Direct evidence from linked temp agency-worker-client data*. Tech. rep. National Bureau of Economic Research.
- Gerard, François et al. (2018). *Assortative matching or exclusionary hiring? The impact of firm policies on racial wage differences in Brazil*. Tech. rep. National Bureau of Economic Research.
- Jochmans, Koen and Martin Weidner (2019). "Fixed-Effect Regressions on Network Data". In: *Econometrica* 87.5, pp. 1543–1560.
- Kline, Patrick, Raffaele Saggio, and Mikkel Sølvsten (2020). "Leave-out estimation of variance components". In: *Econometrica* 88.5, pp. 1859–1898.
- Schoefer, Benjamin and Oren Ziv (2021). *Productivity, Place, and Plants: Revisiting the Measurement*. Tech. rep. CEPR Discussion Paper No. DP15676.

References iv

 Song, Jae et al. (2019). "Firming up inequality". In: *The Quarterly journal of economics* 134.1, pp. 1–50.

 Tomaskovic-Devey, Donald et al. (2020). "Rising between-workplace inequalities in high-income countries". In: *Proceedings of the National Academy of Sciences* 117.17, pp. 9277–9283.