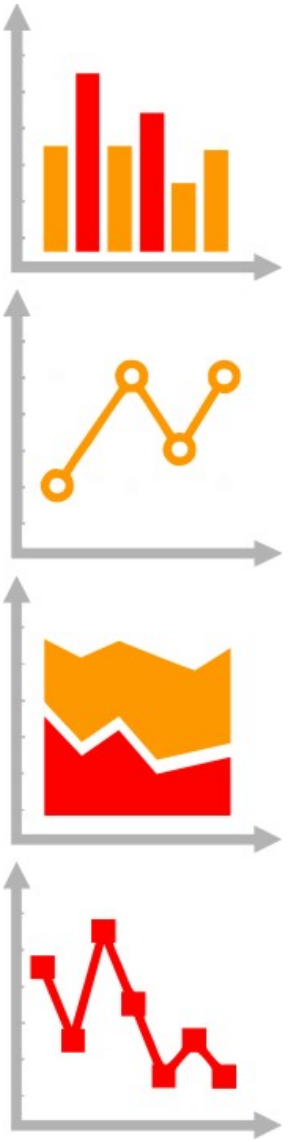
La gestion par partage des poids des changements de contour des entreprises dans l'enquête sectorielle annuelle



JMS 2018

Insee

Mesurer pour comprendre

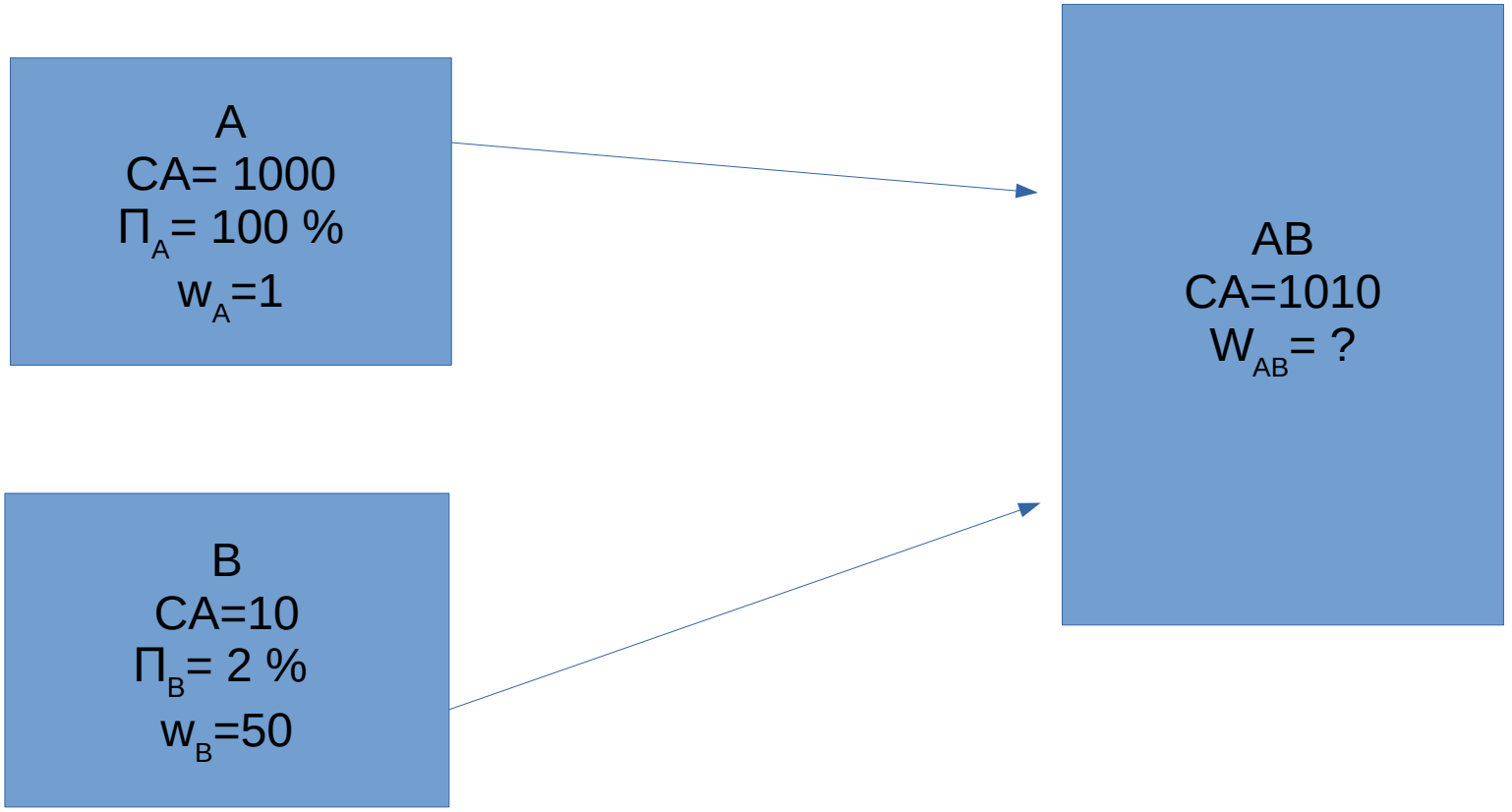
Plan

- Quel est le problème ?
- La méthode généralisée de partage des poids
- L'application de cette méthode dans ESANE.

Quel est le problème ?

- **Depuis 2016 : Nouveau plan de sondage, on tire des entreprises profilées (EP).**
- **L'unité de collecte reste l'unité légale (UL) : lorsqu'une EP est tirée, toutes les UL rattachées sont interrogées.**
- **Le probleme : Au moment du tirage, en novembre, les contours des EP sont provisoires, et c'est plus tard, en mars, que l'information à jour sur les contours peut être utilisée… Mais comment pondérer une EP dont le contour a changé ?**

Le problème : exemple



- **Comment pondérer la nouvelle entreprise AB résultant de la fusion entre A et B ?**

Plusieurs solutions

- **Ne pas mettre à jour les contours…**

- **Passer à 1 le poids de chaque EP concernée par un changement de contour. => C'est la solution utilisée jusqu'ici pour les restructurations d'UL, mais on prédit trop de changements de contours pour que cette solution puisse être adopter niveau entreprises.**

- **Solution "empruntée" aux cas similaires coté ménages : La méthode généralisée de partage des poids (MGPP - voir *Indirect Sampling* de Pierre Lavallée).**

Insee

Méthode généralisée de partage des poids (version liens classiques)

$$w_i = \sum_{k \in i \cap U^A} \tilde{\theta}_{k,i} w_{ik}$$

Avec :

- w_i : le poids final de l'EP *i* ;

- w_{ik} : le poids initial de l'UL *k* rattachée (contours mis à jour) à l'EP *i*, égal à 0 pour les unités légales non échantillonnées ;

- U^A : l'ensemble des UL de la base de sondage (UL du sous-champ 1 et rattachées – contours au moment du tirage - à une EP de la base de sondage) ;

- $\tilde{\theta}_{k,i}$: pondération du lien entre l'EP *i* et l'UL *k* qui lui est rattachée.

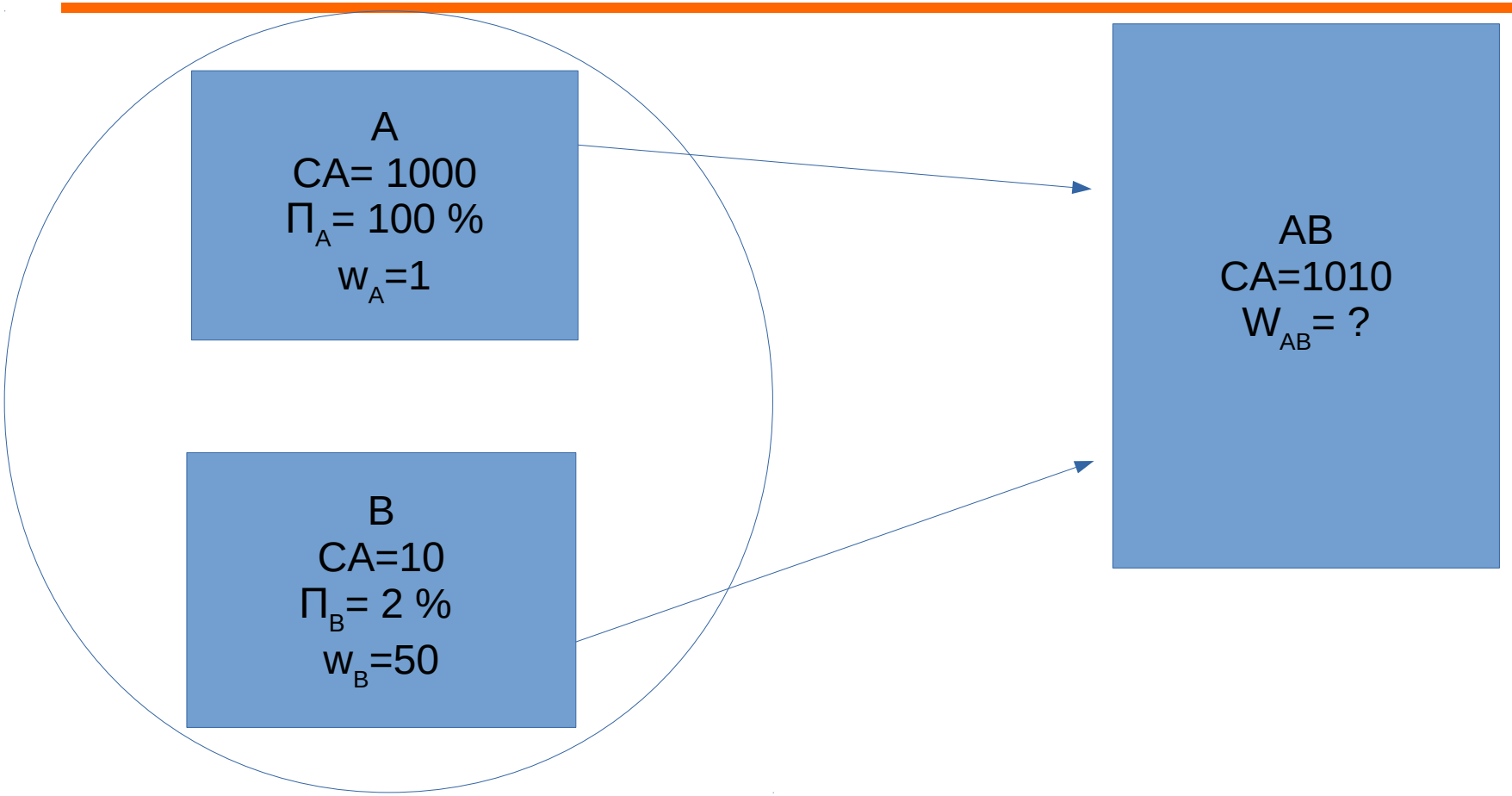
Version liens classiques :

$$\tilde{\theta}_{k,i} = \frac{1}{M_i^{AB}}$$

M_i^{AB} : nombre d'UL rattachées à l'EP i (contours actualisés) et présentes dans la base de sondage

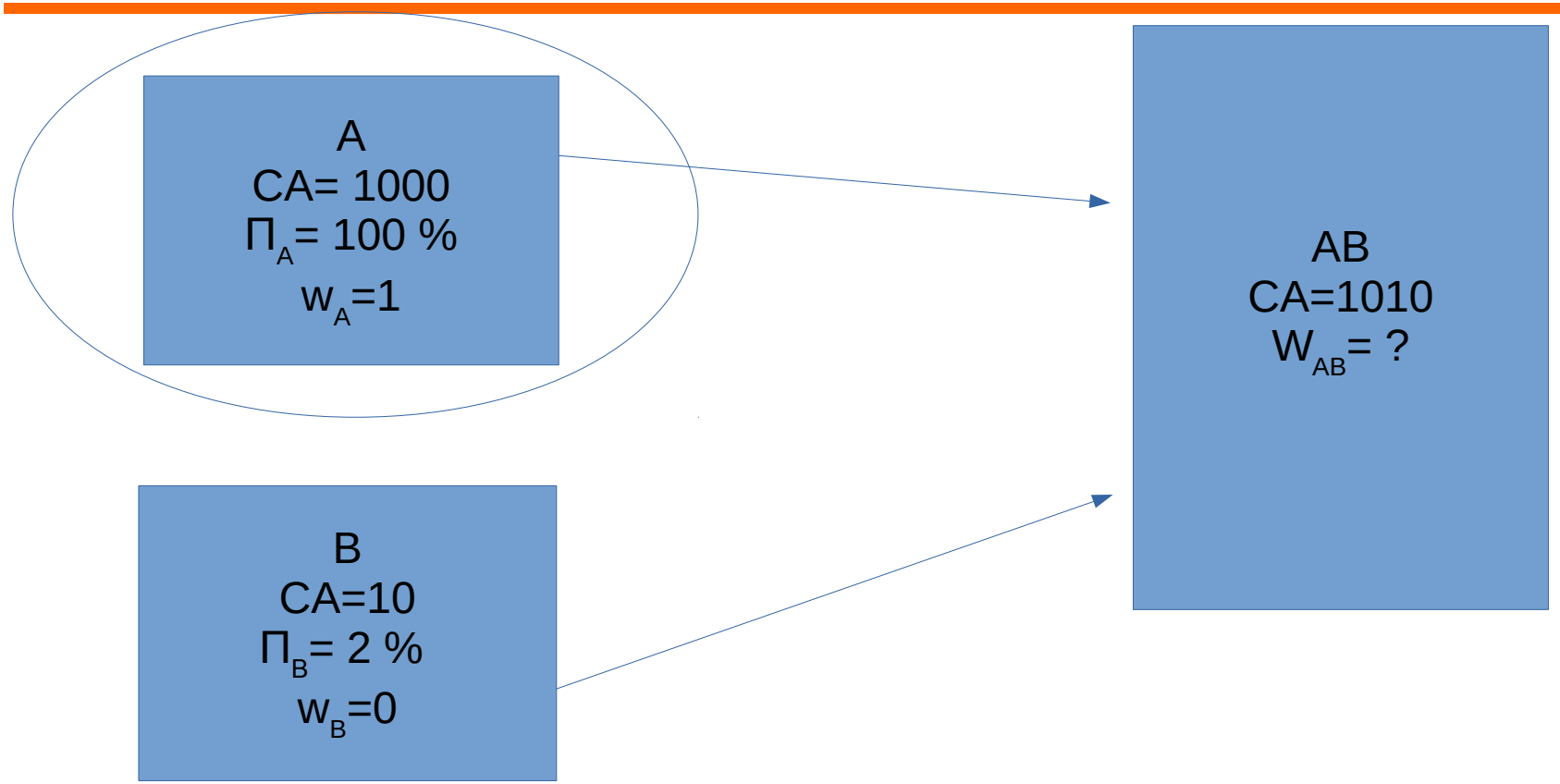
Insee

Méthode généralisée de partage des poids (version liens classiques) Cas 1 : B dans s^A



$$w_{AB} = \frac{w_A + w_B}{2} = \frac{1+50}{2} = 25{,}5$$

Méthode généralisée de partage des poids (version liens classiques) Cas 1 : B pas dans s^A



$$w_{AB} = \frac{w_A + w_B}{2} = \frac{1+0}{2} = 0{,}5$$

Méthode généralisée de partage des poids version liens pondérés par le chiffre d'affaires (CA)

$$w_i = \sum_{k \in i \cap U^A} \tilde{\theta}_{k,i} w_{ik}$$

Avec :

- w_i : le poids final de l'EP *i* ;

- w_{ik} : le poids initial de l'UL *k* rattachée (contours mis à jour) à l'EP *i*, égal à 0 pour les unités légales non échantillonnées ;

- U^A : l'ensemble des UL de la base de sondage (UL du sous-champ 1 et rattachées – contours au moment du tirage - à une EP de la base de sondage) ;

- $\tilde{\theta}_{k,i}$: pondération du lien entre l'EP *i* et l'UL *k* qui lui est rattachée.

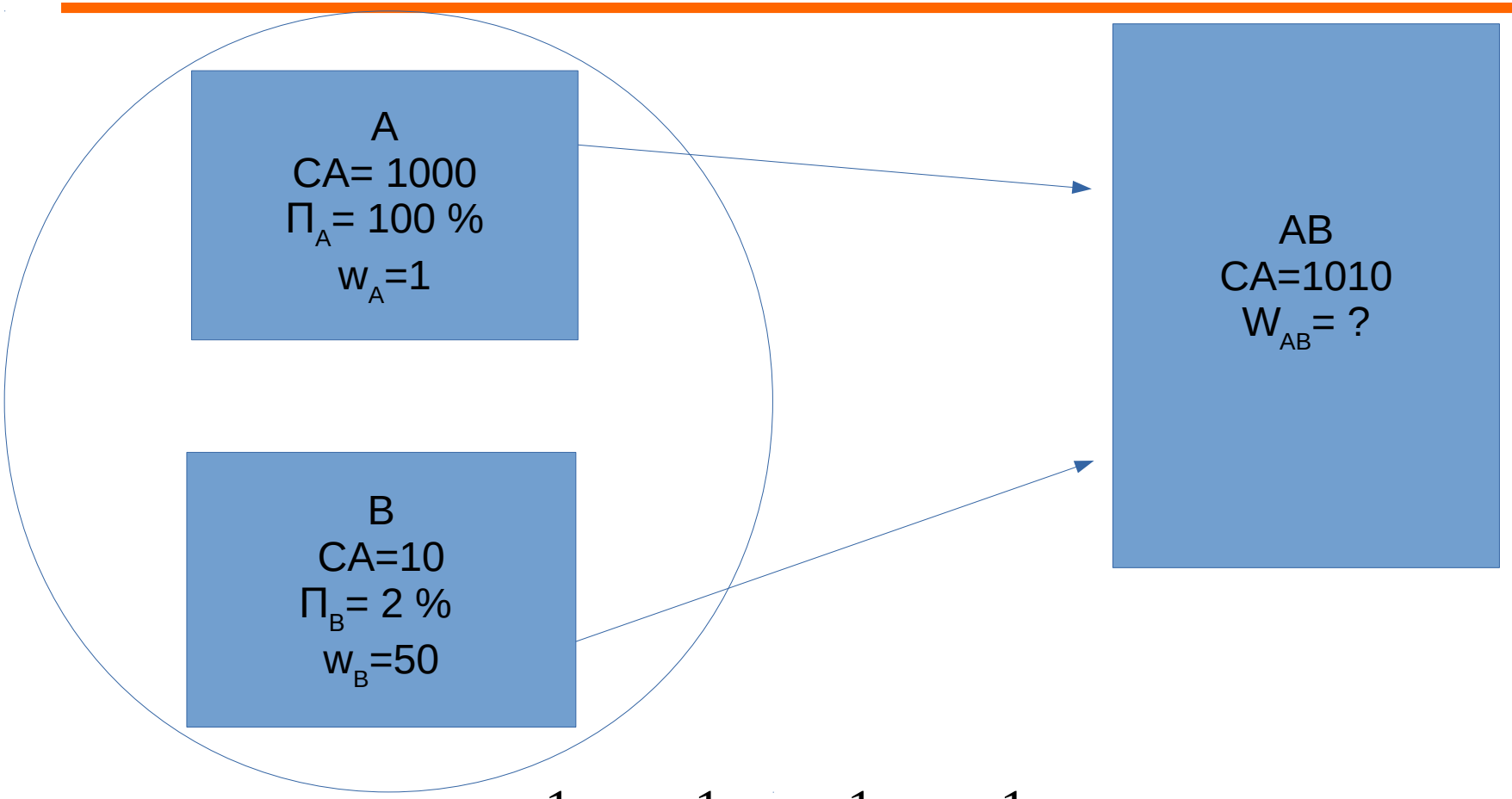
Version liens pondérés par le chiffre d'affaires (CA) :

$$\tilde{\theta}_{k,i} = \frac{CA_k}{\sum_{j \in i \cap U^A} CA_j}$$

Avec CA_k le CA de l'UL k

Insee

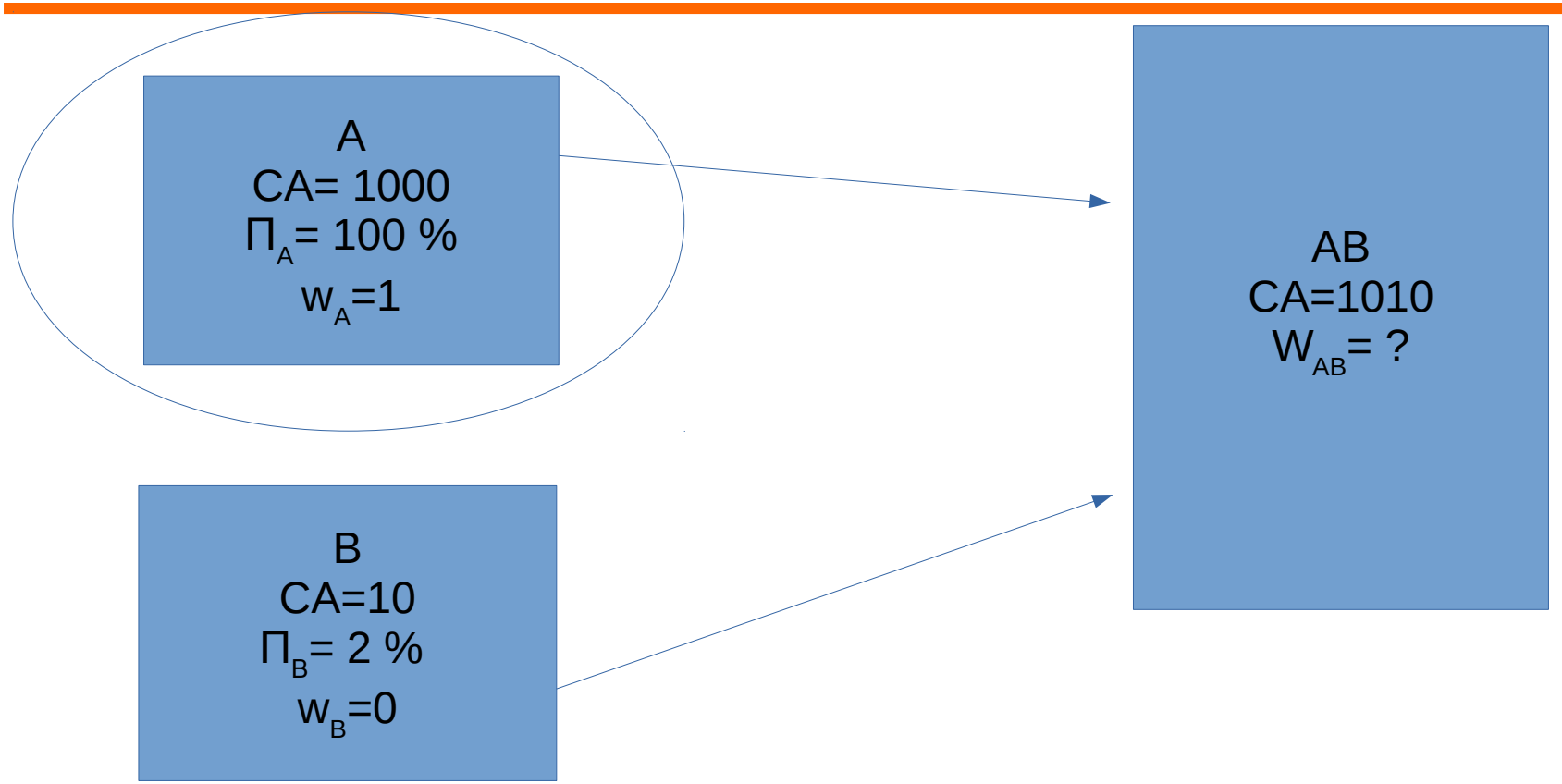
Partage des poids pondéré ; Cas 1 : B dans sA



Classique : $w_{AB} = \frac{1}{2} w_A + \frac{1}{2} w_B = \frac{1}{2} \times 1 + \frac{1}{2} \times 50 = 25{,}5$

Liens pondérés : $w_{AB} = \frac{1000}{1010} w_A + \frac{10}{1010} w_B = \frac{1000}{1010} \times 1 + \frac{10}{1010} \times 50 = 1{,}5$

Partage des poids pondéré ; Cas 2 : B pas dans s^A



Classique : $w_{AB} = \frac{1}{2} w_A + \frac{1}{2} w_B = \frac{1}{2} \times 1 + \frac{1}{2} \times 0 = 0{,}5$

Liens pondérés : $w_{AB} = \frac{1000}{1010} w_A + \frac{10}{1010} w_B = \frac{1000}{1010} \times 1 + \frac{10}{1010} \times 0 = 0{,}99$

Étude par simulations

- 30 000 échantillons tirés selon le nouveau plan de sondage ;
- Comparaison entre la version classique (CL) et la version (CA) où les liens sont pondérés par le CA du partage des poids basée sur l'estimation de la valeur ajoutée totale (2 niveaux d'aggrégation) :
  - « Grands secteurs » (A10 - 8 secteurs) ;
  - NACE 3 positions (195 groupes).
- Cadre simplifié : pas de non-reponse, pas de traitement des valeurs influentes, pas de calage, estimation « simple » (pas d'estimateur composite).

Insee

Indicateurs retenus dans l'étude

- **La comparaison des indicateurs se base sur le coefficient de variation (CV) :**

$$Cv = \frac{\frac{1}{30000}\sqrt{\sum_{r=1}^{30000}\left(\hat{T}_y^m(r) - T_y\right)^2}}{T_y} \qquad RCV = \frac{CV\left(\hat{T}_Y^{CA}\right)}{CV\left(\hat{T}_Y^{CL}\right)}$$

Avec T_y le « vrai » total de valeur ajoutée et $\hat{T}_Y^m(r)$ l'estimateur avec la version m de partage des poids (CL=Classique, CA=liens pondérés par le CA)

- **Biais relatif :**

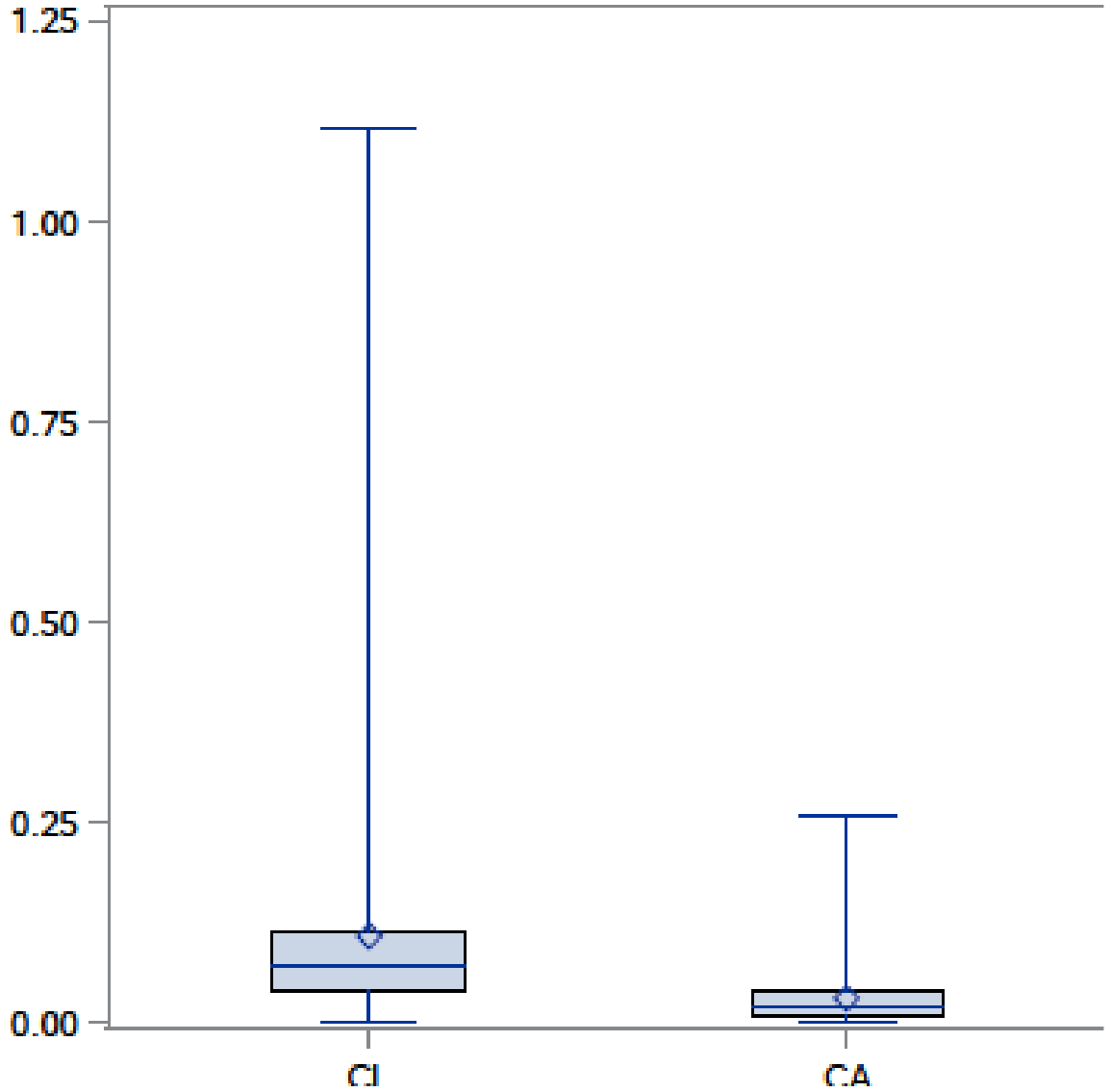
$$Br = \frac{\frac{1}{30000}\sum_{r=1}^{30000}\left(\hat{T}_y^m(r) - T_y\right)}{T_y}$$

Insee

Valeur ajoutée par A10

A10	CV-CL	CV-CA	BR-CL	BR-CA	RCV
AZ Agriculture	7,9%	6,8%	0,0%	0,0%	85,6%
BE Industrie	1,0%	0,2%	0,0%	0,0%	15,4%
FZ Construction	1,5%	0,8%	0,0%	0,0%	53,4%
GI Commerce	2,9%	0,4%	0,0%	0,0%	12,7%
JZ Info,com	3,1%	0,7%	0,0%	0,0%	22,3%
LZ Immobilier	4,5%	2,1%	0,0%	0,0%	46,7%
MN Scien, tec	2,4%	1,1%	0,0%	0,0%	45,0%
RU Services	5,3%	1,9%	0,0%	0,0%	36,0%
Total	**1,1%**	**0,2%**	**0,0%**	**0,0%**	**22,0%**

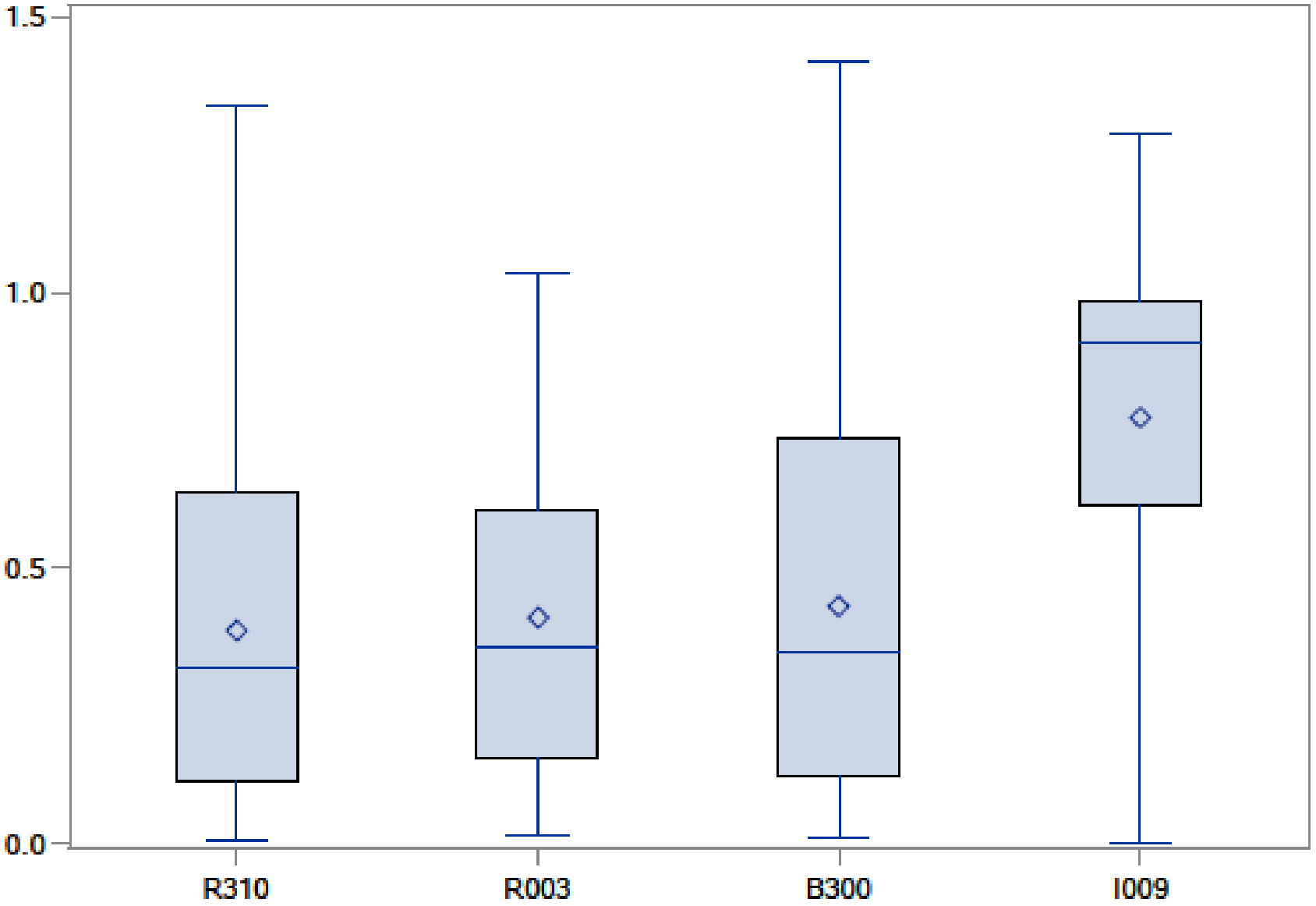
CV de valeur ajoutée par groupe (NACE 3 positions)



Nombre de groupes où le partage des poids avec liens pondérés a le meilleur CV, par variable

Variable	Nombre de groupes (195)
Chiffre d'affaires	164
Valeur ajoutée	173
Passif au bilan	177
Investissement	155

Distribution du RCV par variable



Insee

Résultats de l'étude

- **Les coefficients de variation sont plus faibles avec la version pondérée par le CA pour la grande majorité des activités ;**
- **Pas de biais que les liens soient pondérés ou non ;**
- **Plus la corrélation avec le CA utilisé pour pondérer les liens est forte, meilleurs sont les résultats.**

Conclusion

- Le partage des poids avec liens pondérés apparaît comme la meilleure option pour gérer les changements de contour des entreprises.
- Même si les estimateurs analysés dans cette étude ne sont pas ceux directement utilisés dans Esane, il n'y a pas de raison apparente pour que les conclusions soient différentes lorsque le processus complet d'estimation est pris en compte.
- De plus, les données utilisées sont particulières :
  - il y a une année de décalage entre les contours des EP utilisés pour le tirage et les contours mis à jour ;
  - une nouvelle source a été intégrée au processus permettant de définir les contours.
- En régime courant, le choix de la méthode de partage des poids ne devrait pas avoir autant d'impact que ce qui a été vu dans cette étude.

Perspectives

- Plusieurs pistes d'approfondissement :
  - Tester un cadre plus « général » de passage d'un échantillon d'UL (tiré sans tenir compte de la dimension EP) à un échantillon d'EP.
  - La méthode fonctionnerait-elle pour les restructurations d'unités légales ?
  - Comment adapter la correction de la non-réponse, la winsorisation et le calage sur marges de l'échantillon d'EP mis à jour ?
  - Calculs de précision des estimateurs.

Bibliographie

[1] P. Brion, “Esane, le dispositif rénové de production des statistiques structurelles d’entreprises” Courrier des statistiques n°130, 2011 .

[2] E. Gros, “Esane, ou les malheurs de l’estimation composite : comment gérer les valeurs négatives d’estimateurs par différence”, Actes des Journées de Méthodologie Statistique, 2012

[3] E. Gros, R. Le Gleut “The impact of profiling on sampling”, presentation à l’European Establishment Statistics Workshop, 2017.

[4] P. Lavallée, “Indirect sampling” Springer Series in Statistics, 2007.

La gestion par partage des poids des changements de contour des entreprises dans l'enquête sectorielle annuelle

Merci de votre attention

Arnaud Fizzala
Arnaud.fizzala@insee.fr

Insee
DMCSI – Division Sondages

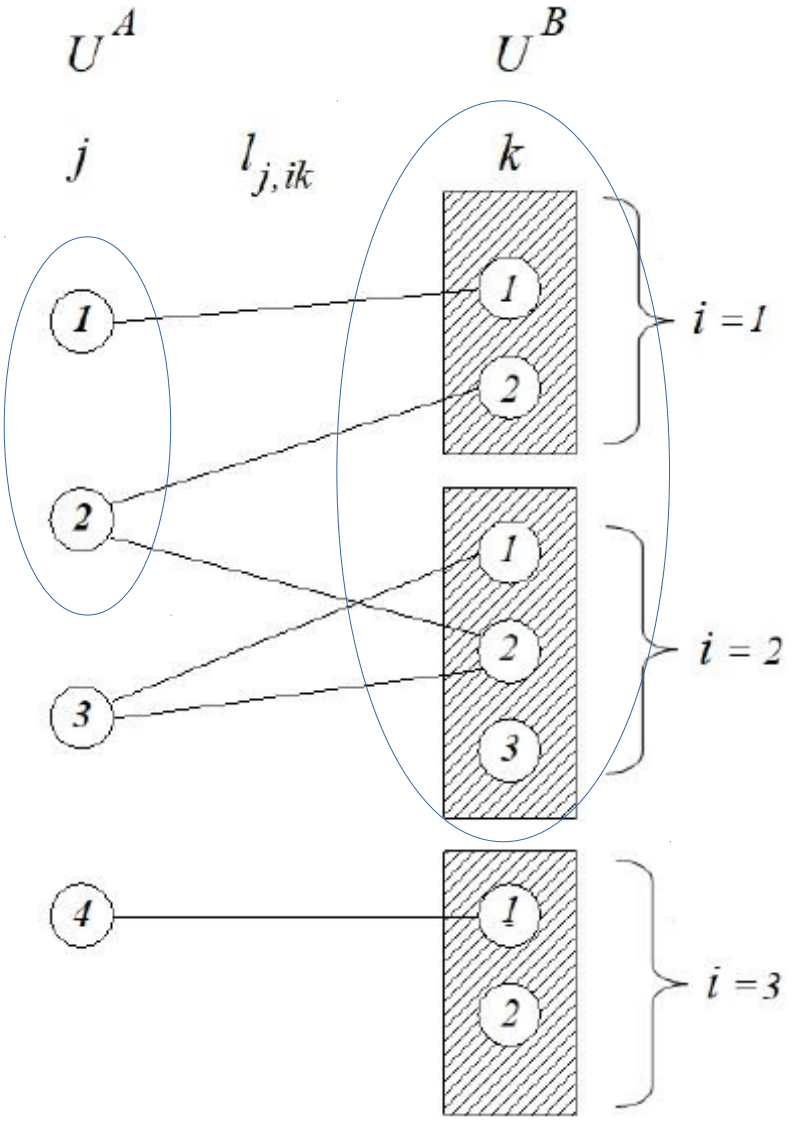
www.insee.fr

@InseeFr

Annexes

GWSM : Illustration by an exemple

- Selection of the units j=1 and j=2 in s^A
- By selecting j=1, we survey the units of the cluster i=1.
- By selecting j=2, we survey the units of the cluster i=1 and the cluster i=2.



Insee

GWSM : Illustration by an exemple

i	k	w'_{ik}	L^B_{ik}	w_i
1	1	$\frac{1}{\pi_1^A}$	1	$\frac{1}{2}\left[\frac{1}{\pi_1^A}+\frac{1}{\pi_2^A}\right]$
1	2	$\frac{1}{\pi_2^A}$	1	
2	1	0 (parce que $t_3 = 0$)	1	$\frac{1}{3}\left[0+\frac{1}{\pi_2^A}+0\right]=\frac{1}{3\pi_2^A}$
2	2	$\frac{1}{\pi_2^A}+0=\frac{1}{\pi_2^A}$	2	
2	3	0 (parce que $l_{j,23}=0$ pour tout j)	0	

$$\hat{Y}^B = \frac{1}{2}\left[\frac{1}{\pi_1^A}+\frac{1}{\pi_2^A}\right]y_{11} + \frac{1}{2}\left[\frac{1}{\pi_1^A}+\frac{1}{\pi_2^A}\right]y_{12} + \frac{y_{21}}{3\pi_2^A} + \frac{y_{22}}{3\pi_2^A} + \frac{y_{23}}{3\pi_2^A}$$

GWSM : Illustration by an exemple

- **Suppose that** $\pi_1^A = 1/3$ and $\pi_2^A = 1$

i	k	w'_{ik}	L_{ik}^B	w_i
1	1	$\frac{1}{\pi_1^A} = 3$	1	$\frac{1}{2}\left[\frac{1}{\pi_1^A} + \frac{1}{\pi_2^A}\right] = 2$
1	2	$\frac{1}{\pi_2^A} = 1$	1	
2	1	0 (parce que $t_3 = 0$)	1	$\frac{1}{3\pi_2^A} = \frac{1}{3}$
2	2	$\frac{1}{\pi_2^A} + 0 = 1$	2	
2	3	0 (parce que $l_{j,23} = 0$ pour tout j)	0	

Insee

Study – Distribution of the weights

- Now we focus on the impact of the GWSM on the biggest units, so we focus on the enterprises in the « take-all stratum ».
- An enterprise is in the « take-all stratum » if at least one legal unit linked to the enterprise has an initial weight of 1.
- As expected, final weights are more concentrated around the value 1 when the links are weighted by turnover. That probably explains in most part that the GWSM with links weighted by turnover estimators are more stable and accurate.

Study – Distribution of the weights of the enterprises from the « take-all » stratum

m	max	P99	P95	P90	Q3	Q2	Q1	P10	P5	P1	min
CL	117	1,00	1,00	1,00	1,00	1,00	1,00	0,94	0,67	0,50	0,08
CA	158	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,97	0,68	0,00

Insee