Les traitements post-collecte de l'ESA-EAP en entreprises

Séminaire de Méthodologie Statistique
20/11/2019

Insee
Mesurer pour comprendre

01 Mise à jour des entreprises

Pourquoi ?

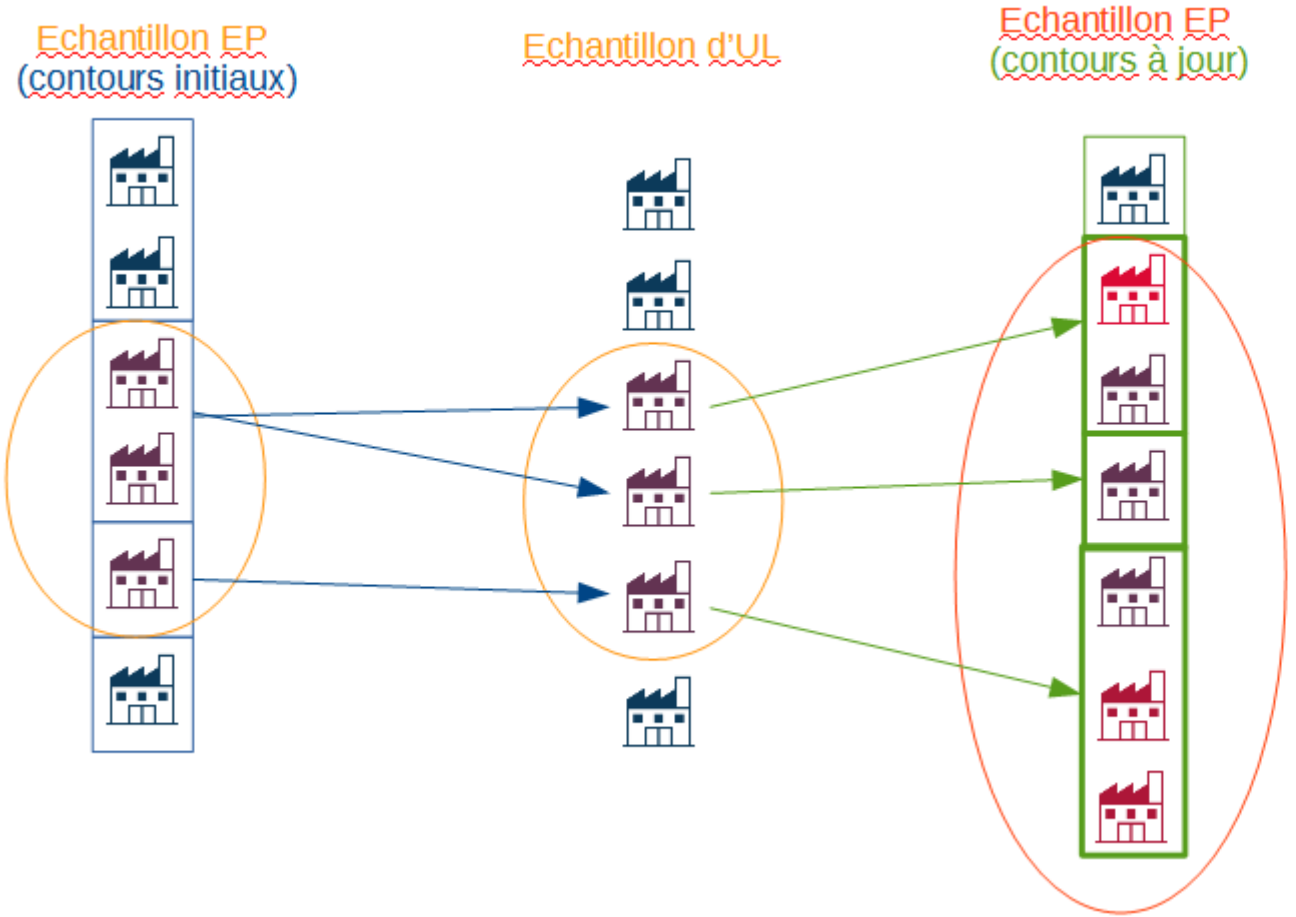
Depuis 2016 : Nouveau plan de sondage, on tire des entreprises profilées (EP).

L'unité de collecte reste l'unité légale (UL) : lorsqu'une EP est tirée, toutes les UL rattachées sont interrogées.

Le problème : 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.

On met à jour l'échantillon en considérant qu'une EP (contours à jour) est dans l'échantillon si au moins une de ses UL est dans l'échantillon initial d'UL.

Mise à jour du contour des entreprises profilées : impact sur l'échantillon



Questionnaire

Imputations

Mise à jour du contour des entreprises profilées : impact sur l'échantillon ESA 2017

Composition de l'échantillon avant/après mise à jour

EP Tirage	UL	EP mises à jour
105 000	180 000 (155 000 envois de questionnaires)	105 000 (200 000 UL)

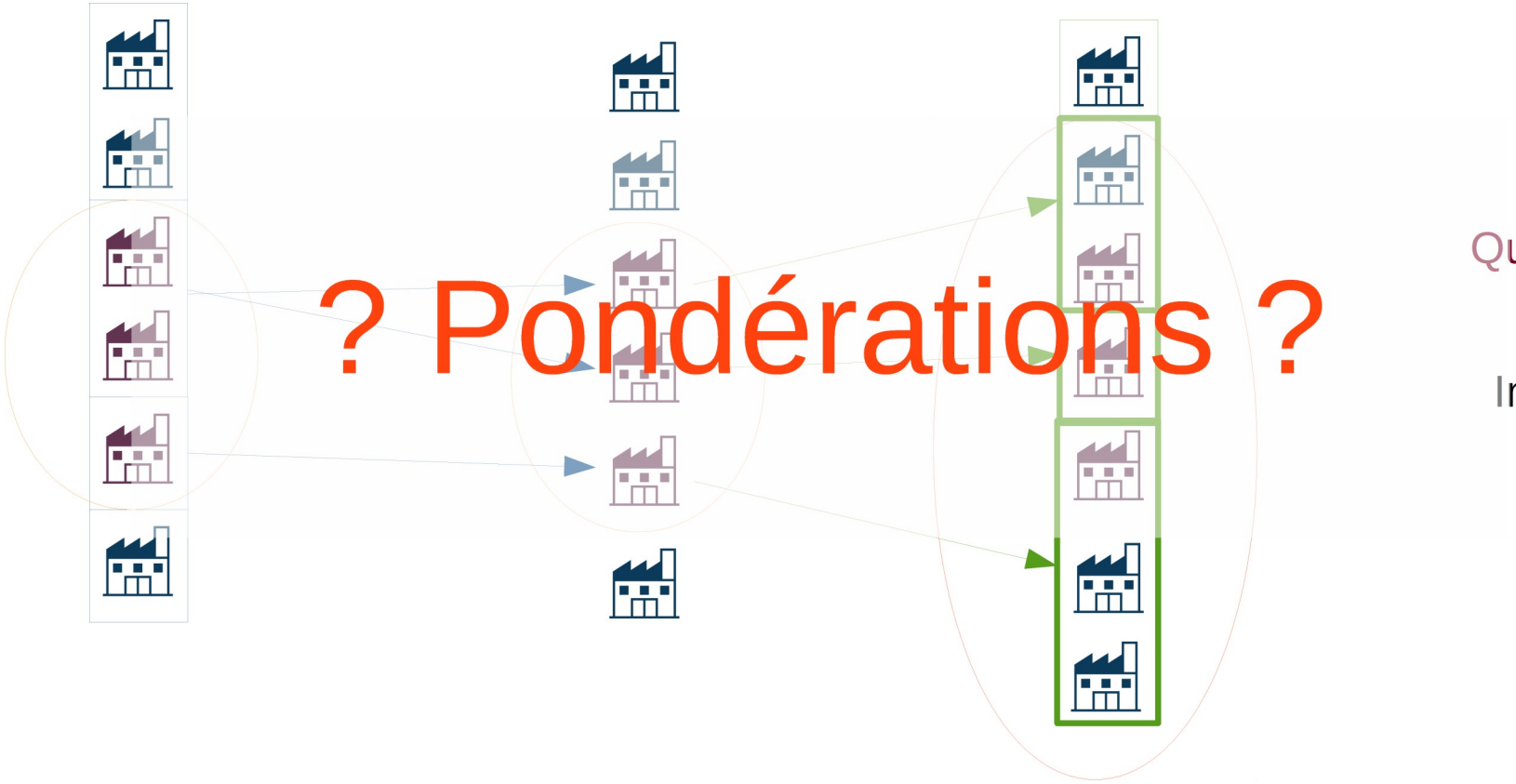
=> 45 000 imputations *(données N-1 en priorité et moyennes de classes à défaut)* à réaliser pour calculer les données consolidées

Mise à jour du contour des entreprises profilées : impact sur l'échantillon

Echantillon EP
(contours initiaux)

Echantillon d'UL

Echantillon EP
(contours à jour)



? Pondérations ?

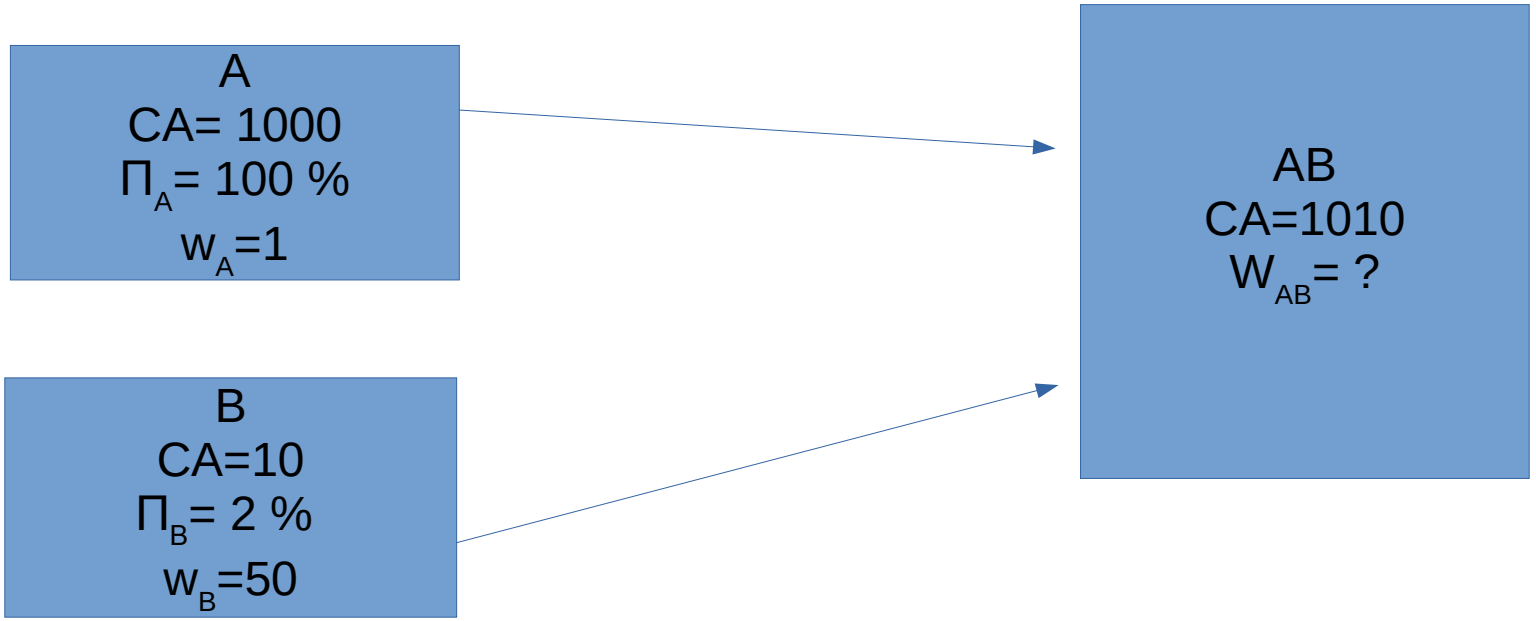
Questionnaire

Imputations

Mise à jour du contour des entreprises profilées : Exemple

Tirage (Novembre N)

Mise à jour (Mars N+1)



Comment pondérer la « nouvelle » EP *AB* ?

Mise à jour du contour des entreprises profilées : Solutions

Calculer la probabilité de sélection des EP une fois le contours mis à jour. => Solution abandonnée car le calcul ne parait pas faisable dès que la situation est complexe.

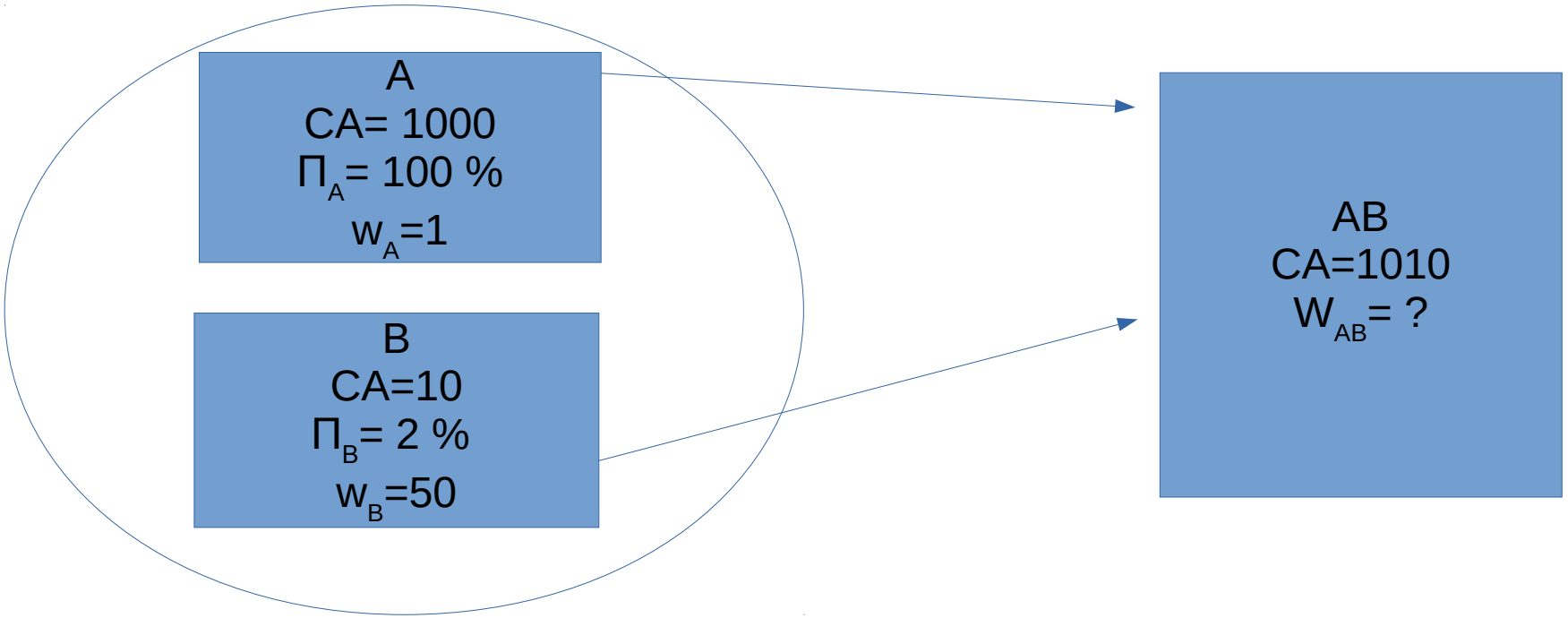
Passer à 1 le poids de chaque EP concernée par un changement de contour. => C'est la solution utilisée jusqu'ici pour les restructurations d'UL, mais on prédit trop de changements de contours pour que cette solution puisse être adoptée niveau entreprises.

Partage des poids (courant côté ménages). Cette méthode est « facile » à appliquer ici car on considère qu'une UL est rattachée à une seule EP.

- Version « classique » : moyenne des poids des UL rattachées et présentes dans la base de sondage.
- Version « entreprise » : moyenne, pondérée par le chiffre d'affaires, des poids des UL rattachées et présentes dans la base de sondage.

Avec w_k=0 si l'UL k n'est pas dans l'échantillon initial d'UL.

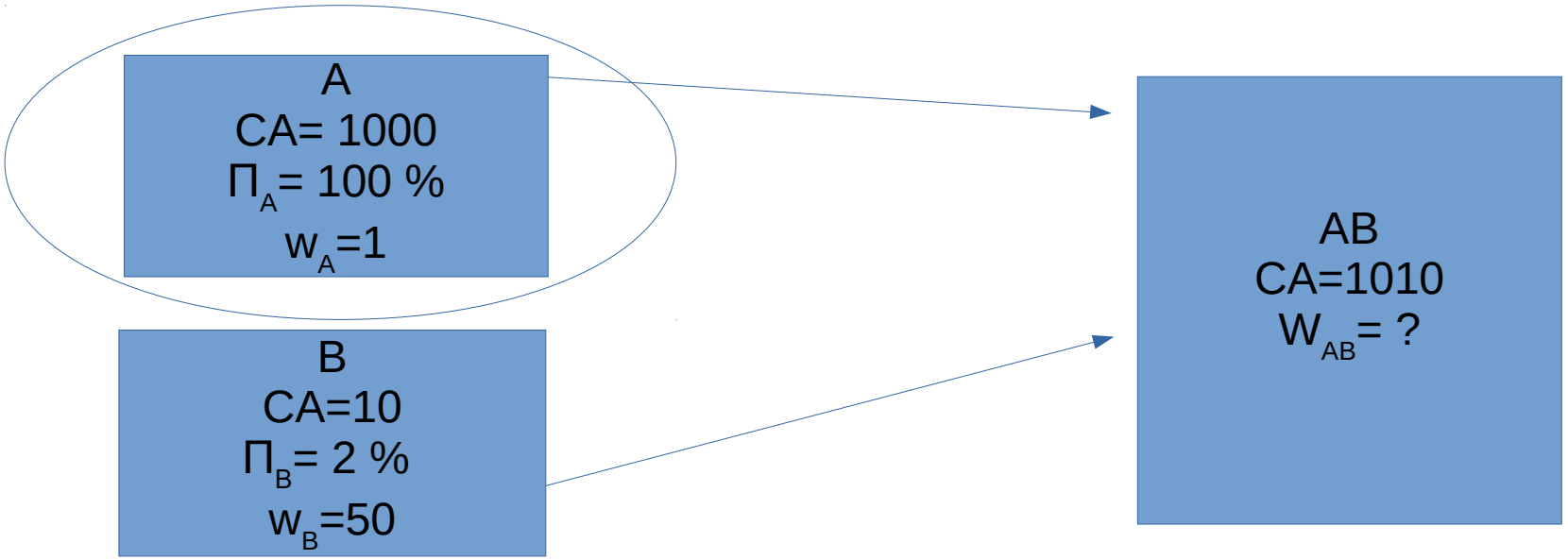
Exemple : Cas 1 : A et B sont dans l'échantillon initial



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

Exemple : Cas 2 : B n'est pas dans l'échantillon initial

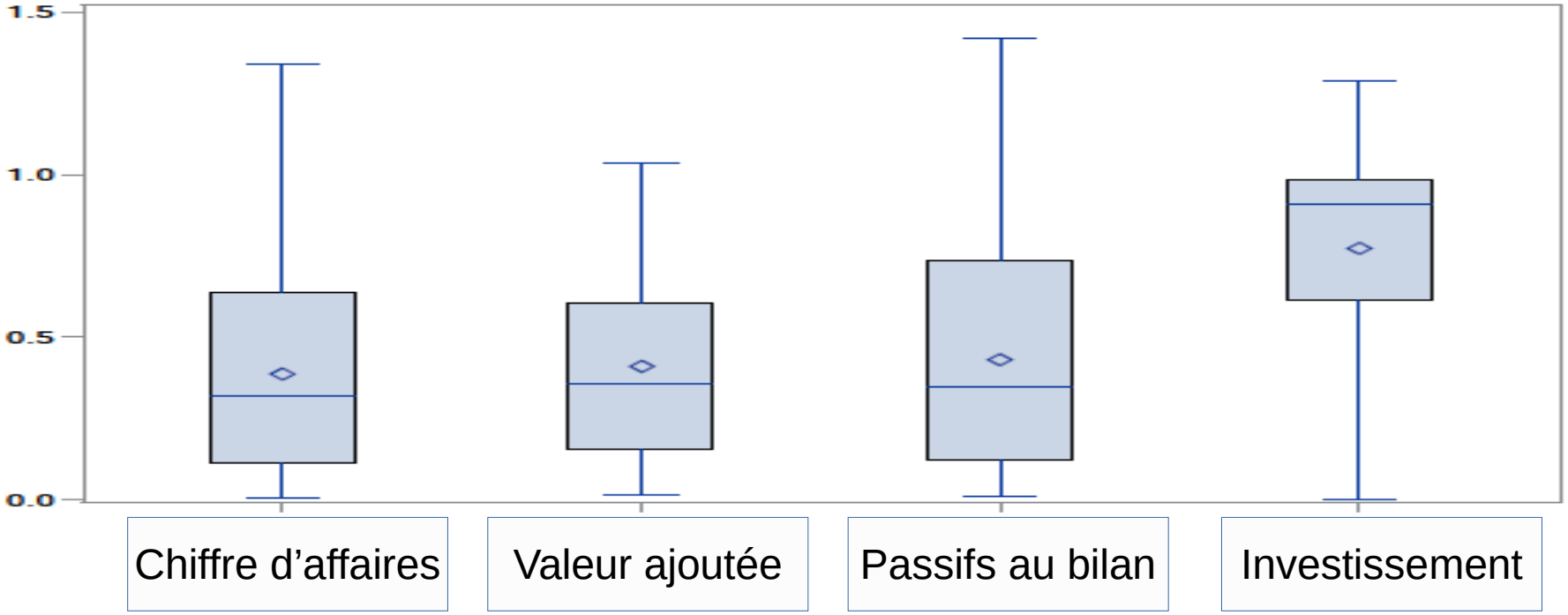


Classique : $$w_{AB}=\frac{1}{2}w_A+\frac{1}{2}w_B=\frac{1}{2}\times 1+\frac{1}{2}\times 0=0{,}5$$

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

Pondérer les liens par le chiffre d'affaires des UL conduit à des estimateurs plus précis

Simulations : 30 000 échantillons, Cv(Partage des poids « entreprises ») / Cv(Partage des poids « classique »), par activité (NACE 3 positions).



Mise à jour du contour des entreprises profilées : L'exhaustif

La probabilité de tirage d'une EP (contours à jour) correspond à la probabilité qu'au moins une de ses UL appartienne à l'échantillon initial.

En général, cette probabilité est complexe à calculer… Sauf si une des UL appartient à la partie exhaustive de l'échantillon d'UL => Dans ce cas on est sûr (probabilité de 1) de tirer l'UL en question et donc sûr (probabilité de 1) que l'EP en question appartienne à l'échantillon d'EP.

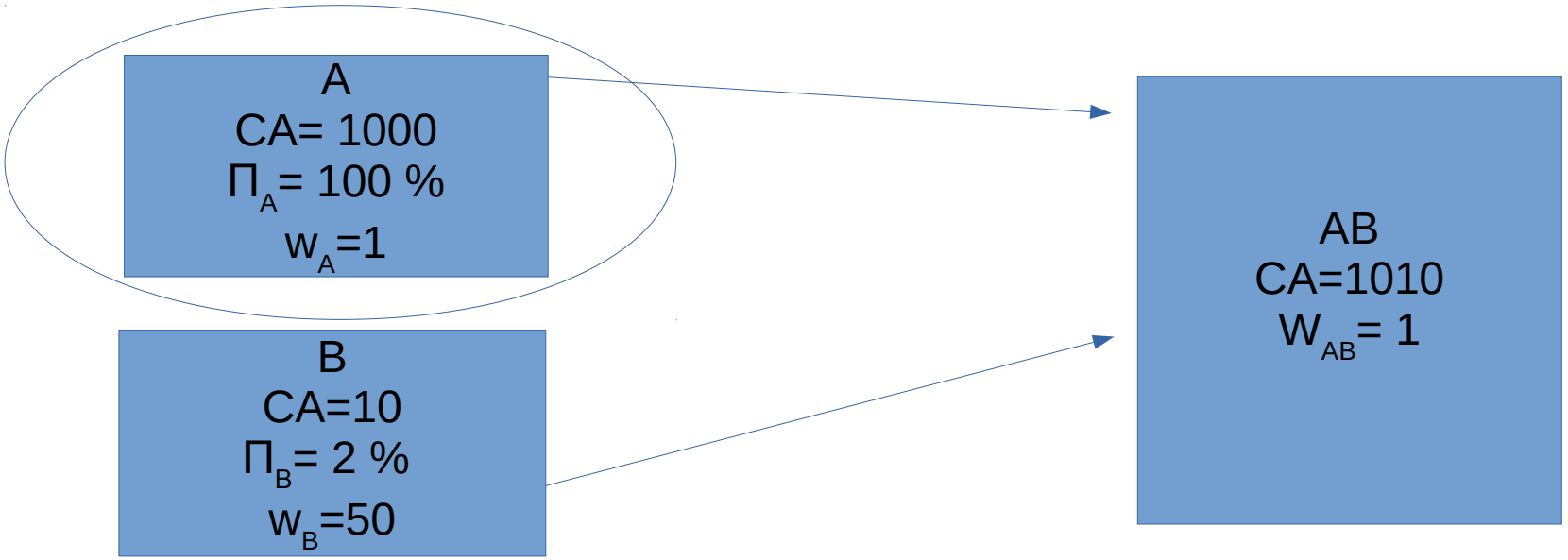
Finalement, on applique la règle suivante pour pondérer les EP (contours à jour) :

Si au moins une UL dans l'exhaustif alors poids=1 (EP « exhaustives »)

Si aucune UL dans l'exhaustif alors Partage des poids pondéré par le chiffre d'affaires.

! Echantillon ESA 2017 : 27 500 EP, 80% dans l'exhaustif, 150 EP avec un poids <1.

Exemple : Finalement que B soit dans l'echantillon ou non...



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

W_{AB}=1 (car A est dans l'exhaustif)

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

02 Correction de la non-réponse

Correction de la non-réponse

Les réponses ne sont pas obtenues directement auprès des EP mais constituées à partir des réponses des UL

UL imputées (pas dans l'échantillon initial d'UL)

UL imputées (dans l'échantillon initial d'UL mais non répondantes)

UL répondantes

EP répondante si les UL répondantes représentent au moins 70% du CA cumulé de l'EP

Correction de la non-réponse totale

La non-réponse totale est traitée différemment entre :

- Les EP exhaustives (au moins une UL dans la partie exhaustive) : Imputations se basant sur les réponses (imputées ou vraies réponses) au niveau des UL rattachées.

- Les EP non-exhaustives (aucune UL dans la partie exhaustive) : Repondération des EP non-exhaustive répondantes dans des groupes de réponse homogène.

Dès lors que le caractère répondant des EP est défini (règle des 70%), la correction de la non-réponse totale au niveau des EP non-exhaustive ne pose pas de problème particulier.

En pratique la correction de la non réponse est opérée séparément entre EP et UL indépendantes car les variables mobilisables ne sont pas les mêmes.

Correction de la non-réponse totale

Variables mobilisées pour constituer les groupes de réponse homogène

UL Indépendantes	EP
- Effectif - Chiffre d'affaires - indicatrice liasse imputée - Indicatrice entrepreneur individuel - groupe (APE 3 positions) - durée d'existence - zeat	- Effectif - Chiffre d'affaires des UL cumulé - part de chiffre d'affaires des UL cumulé venant d'une liasse imputée - groupe (APE 3 positions) - durée d'existence de l'UL avec le plus grand CA - zeat de l'UL avec le plus grand CA

Correction de la non-réponse totale

Taux de réponse en 2017

	UL Indépendantes	EP
Partie exhaustive	72%	75%
Partie non exhaustive	55%	61%

03 Valeurs influentes

Unités influentes

Unité influente : Unité dont la présence dans l'échantillon fait beaucoup varier l'estimateur. En général : combinaison d'un poids élevé et d'une valeur importante (mais "vraie", il ne s'agit pas ici de la détection de données erronées).

Les unités de la partie exhaustive, du fait qu'elles appartiennent systématiquement à l'échantillon, ne peuvent pas être influentes au sens que nous donnons à ce terme ici.

Les unités influentes sont parfois traitées en ramenant leur poids de sondage a 1 (i.e les passer dans l'exhaustif) et en repondérant les autres unités de la strate de tirage.

Facile à mettre en place, cette solution pose question :

1) *Le caractère subjectif de la détection d'une unité : à partir de quel moment une unité est-elle considérée comme influente ?*
2) *le passage systématique du poids à 1 : pourquoi pas une autre valeur ?*

*La Winsorisation avec les seuils de **Kokic et Bell** apporte une réponse "scientifique" à ces questions.*

Winsorisation

Winsorisation : Méthode se basant sur des seuils K_h par strate. Lorsque le chiffre d'affaires de l'unité dépasse le seuil correspondant à sa strate, on réduit (pas forcément à 1) le poids de l'unité. Cela va entraîner :

- *Stabilisation des estimateurs (réduction de la variance due à l'échantillonnage)*
- *Biais de sous-estimation*

Le choix des seuils est crucial pour que l'arbitrage biais/variance soit bon.

Les seuils de Kokic et Bell minimisent l'erreur quadratique moyenne de l'estimateur winsorisé.

Problème : La méthode est prouvée pour le sondage aléatoire simple stratifié, ce qui n'est plus notre cadre après la mise à jour des contours !

Winsorisation adaptée

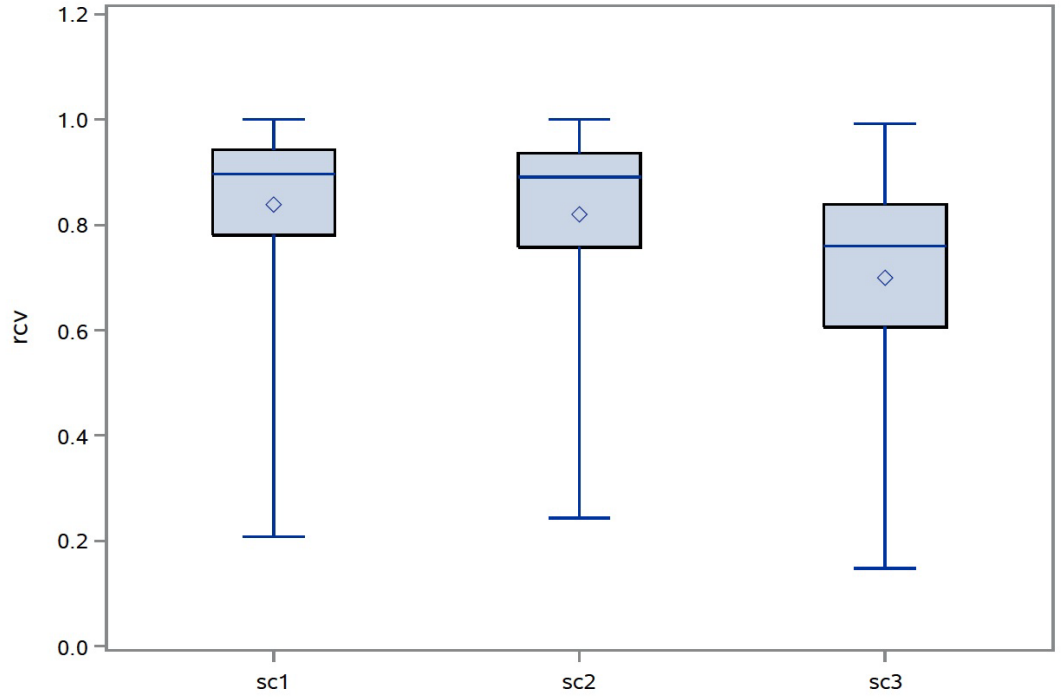
Solutions envisagées :

1) *Winsoriser l'échantillon original et effectuer ensuite le partage des poids a partir des poids winsorisés*
   - *Cadre théorique respecté (car échantillon original est un SAS stratifié)*
   - *Non détection des unités devenant influentes du fait du partage des poids*

2) *Faire comme si l'échantillon après partage des poids etait issu d'un SAS stratifié et appliquer "directement" la winsorisation*
   - *Facile à mettre en place*
   - *Conditions de calcul des seuils de Kokic et Bell ne sont pas vérifiées…*

3) *Winsoriser l'échantillon original avec une variable transformée (variables Z) permettant d'obtenir, avec l'échantillon original, le total de la variable visée après partage des poids*
   - *Est-ce que winsoriser la variable Z permet de détecter les unités influentes pour la variable finale ?*

Simulations : 50 000 tirages, CV(avec winsorisation)/CV(sans winsorisation) par NACE 3 positions.



04 Calage

Calage

Calage : Opération visant à retrouver un total connu en modifiant le "moins possible" les poids initiaux.

Jusqu'ici :

- *Totaux connus via la Base de Sondage (Chiffre d'affaires par NACE 3 positions et Nombre d'unités par NACE 2 positions)*
- *Calage de la partie non-exhaustive uniquement*

Impact du passage en EP :

- *Du fait du partage des poids, l'estimateur n'est plus naturellement calé sur les strates de tirage => Potentiellement plus de difficultés pour atteindre les marges de calage.*
- *On cale séparément les UL indépendantes et les EP car les données ne doivent pas tenir compte des résultats de l'enquête :*
  - *UL : caractéristiques au lancement actualisées avec les informations ne venant pas de l'enquête (Sirus et nouvelles liasses fiscales).*
  - *EP : APE après partage des poids (car à ce moment les réponses à l'enquête ne sont pas disponibles et la consolidation n'en tient donc pas compte), et CA venant des nouvelles liasses fiscales des UL cumulées (car il n'est pas possible de consolider sans tenir compte des résultats de l'enquête).*

05 Estimation

Estimateurs (Variables fiscales)

Jusqu'ici: Estimateurs par différence

$$\hat{Y} = \sum_{i \in U} Y_i + \sum_{i \in s} w_i [y_i - Y_i]$$

Avec : U la population, s l'échantillon, Y_i la variable au lancement, y_i la variable collectée.

Impact du passage en EP : Du fait de l'absence de données « au lancement » au niveau EP et de l'impossibilité de ne pas tenir compte des résultats de l'enquête pour consolider les données, on sépare l'estimateur entre :

- *UL indépendantes : on continue l'estimateur par différence*
- *EP : on utilise un estimateur par expansion « classique »*

$$\hat{Y} = \sum_{i \in U_{ind}} Y_i + \sum_{i \in s_{ind}} w_i [y_i - Y_i] + \sum_{i \in s_{EP}} w_i y_i$$

Références

Partage des poids

P. Lavallée, Indirect sampling Springer Series in Statistics, 2007.

P. Lavallée and S. Labelle-Blanchet. Indirect sampling applied to skewed populations. Survey Methodology, Volume 39, Number 1, June 2013.

Adaptation des traitements post-collecte en EP

A. Fizzala. "Adaptation of Winsorization caused by weight share method", poster présenté à NTTS 2019 (New Techniques and Technologies for Statistics).

A. Fizzala. "La gestion par partage des poids des changements de contour des entreprises dans l'Enquête Sectorielle Annuelle", JMS 2018.

***La méthodologie en général** : Fiches méthodologiques diffusées sur le site internet de l'Insee (Winsorisation, calage, traitement de la non-réponse totale, traitement de la non-réponse partielle….).*

Retrouvez-nous sur :

insee.fr

Arnaud Fizzala
DMCSI – Division Sondages – Section Entreprises

Insee
Mesurer pour comprendre