Esane ou les malheurs de l'estimation composite

Comment gérer les valeurs négatives d'estimateurs par différence ?





Gros Emmanuel
Insee – UMS-E

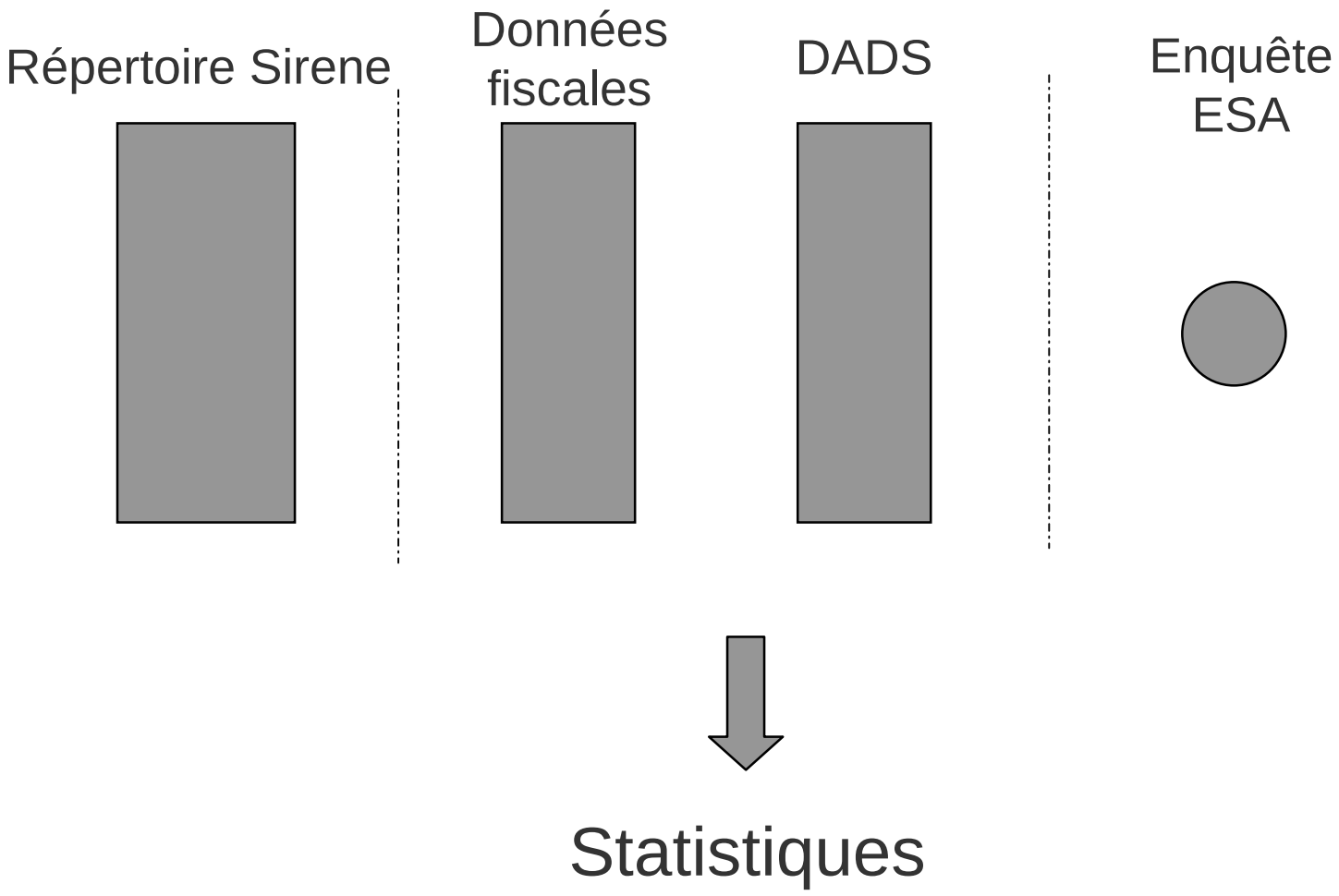
INSEE
INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Le contexte du système Esane

- Résultat d'une refonte complète du processus de production des statistiques structurelles d'entreprises :
  - ✓ Auparavant, deux dispositifs coexistaient en parallèle : enquêtes statistiques (EAE) et exploitation de données fiscales (Suse) ;
  - ✓ Nouveau système fondé sur l'utilisation intensive de sources administratives, complétées par des données d'enquête.
- Un nouveau dispositif plus complexe, qui « unifie » les deux sources et ouvre de nouvelles possibilités :
  - ✓ en ce qui concerne la cohérence des données ⇒ phase de réconciliation des données individuelles ;
  - ✓ en termes d'estimation ⇒ utilisation de techniques de calage et estimation par différence.

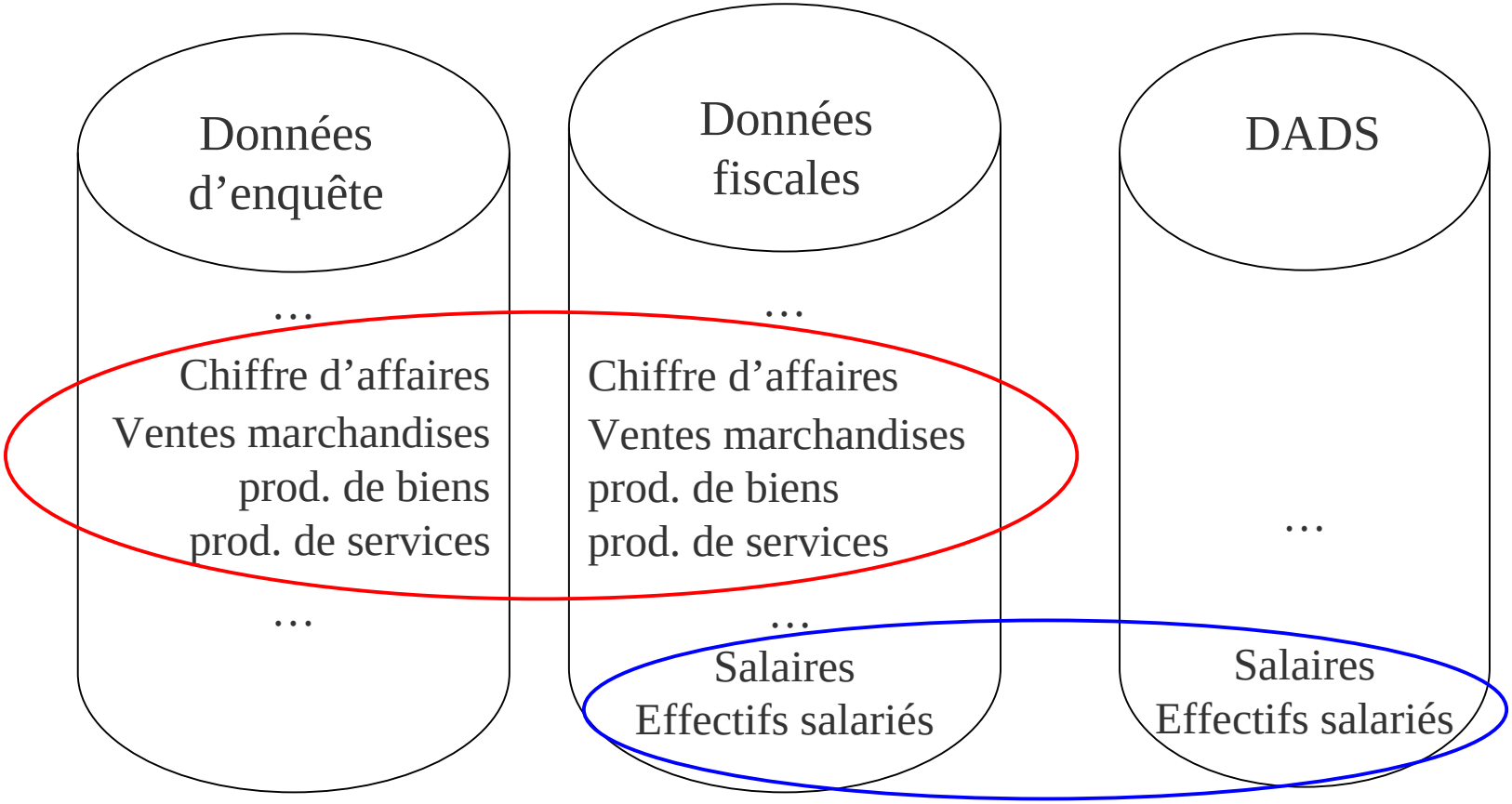
➔ Estimations négatives problématiques à gérer !

Un dispositif multi-source



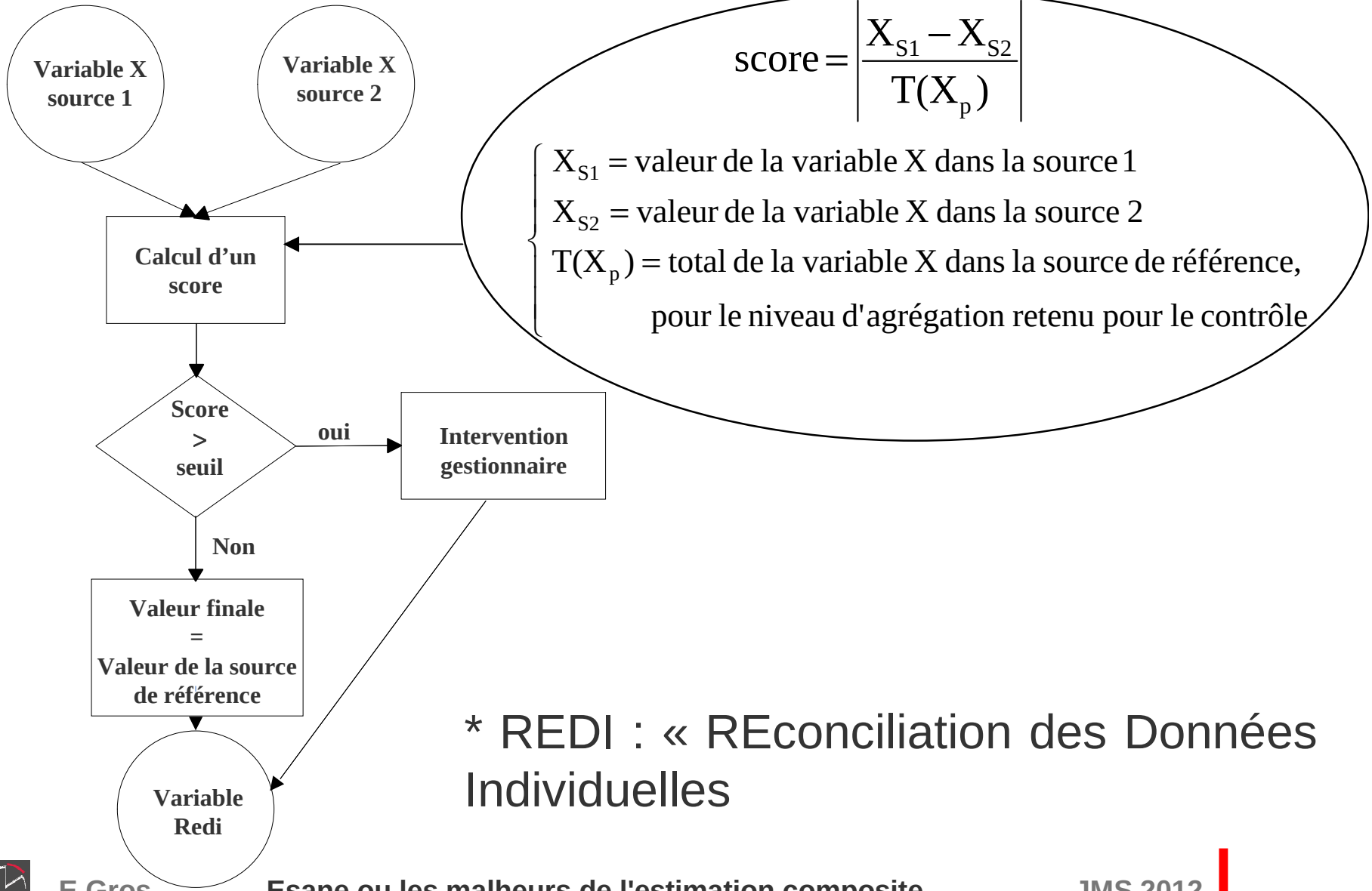
Réconciliation des données individuelles

- Redondance d'informations entre les différentes sources :



➔ Permet la mise en place d'une procédure de contrôle de cohérence et de réconciliation des données individuelles.

Le processus REDI*

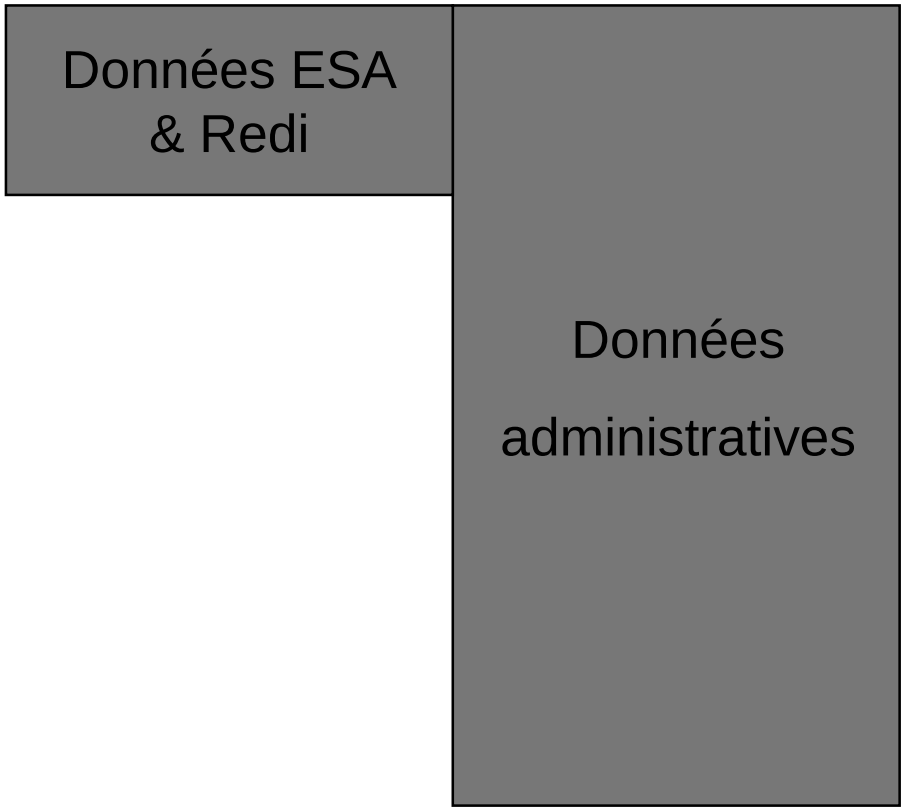


* REDI : « REconciliation des Données Individuelles

Une procédure d'estimation spécifique (1)

- Problématique méthodologique : produire des statistiques exploitant conjointement des données administratives et des données d'enquête.

Contexte



Procédures statistiques

⇒ Mise en œuvre de techniques de calage

⇒ Utilisation d'estimateurs spécifiques : estimation par différence

Une procédure d'estimation spécifique (2)

- Point de départ : l'estimateur usuel $\sum_{i \in R} d_i Y_i$
- Première étape : calage sur données administratives

$\Rightarrow$ modification des poids w_i des unités selon les équations de calage suivantes :

$$\begin{cases} \sum_{i \in R} w_i CA^{fiscal}(i)\, \mathbb{1}_{APE\_rep=X}(i) = \sum_{i \in U\text{-exhaustif}} CA^{fiscal}(i)\, \mathbb{1}_{APE\_rep=X}(i) \\ \sum_{i \in R} w_i\, \mathbb{1}_{APE\_rep=X}(i) = \sum_{i \in U\text{-exhaustif}} \mathbb{1}_{APE\_rep=X}(i) \end{cases}$$

où APE_rep est le code APE issu du répertoire Sirene et $CA^{fiscal}(i)$ le chiffre d'affaires de l'entreprise i issu des données fiscales.

$\Rightarrow$ niveau sectoriel retenu pour le calage = groupe (trois 1ers caractères du code APE) pour limiter la dispersion des poids.

Une procédure d'estimation spécifique (3)

- seconde étape : pour les statistiques sectorielles, l'existence de deux codes APE – celui ex ante du répertoire APE_rep, et celui issu de l'enquête statistique APE_enq –, conduit à proposer d'utiliser l'estimateur par différence $\hat{Y}^{X}_{diff}$ suivant :

$$\sum_{i \in exh \oplus R} w_i Y_i^{Redi} \mathbb{1}_{APE\_enq=X}(i) + \sum_{i \in U} Y_i^{fiscal} \mathbb{1}_{APE\_rep=X}(i) - \sum_{i \in exh \oplus R} w_i Y_i^{fiscal} \mathbb{1}_{APE\_rep=X}(i)$$

$\Rightarrow$ Les variables $Y_i^{Redi} \mathbb{1}_{APE\_enq=X}(i)$ et $Y_i^{fiscal} \mathbb{1}_{APE\_rep=X}(i)$ étant en général fortement corrélées, et même très souvent identiques, estimateur globalement plus efficace que l'estimateur usuel.

$\Rightarrow$ Mais pas de garantie sur le signe des agrégats obtenus !

Cas d'apparition d'estimations négatives « à tort »

- Les estimations problématiques relèvent de deux cas de figure distincts :

  ✓ variables fortement affectées par le processus Redi, et pour lesquelles $Y_i^{Redi} \neq Y_i^{fiscal} \Rightarrow$ concerne ventilation agrégée du CA et les achats & variations de stocks correspondants ;

  ✓ estimation de statistiques portant sur des petits domaines $\Rightarrow$ concerne les estimations de quantités à un niveau fin, ainsi que les estimations portant sur des variables comptables à occurrences rares.

- Problèmes rares – et relevant majoritairement du 1er cas – tant qu'on raisonne au niveau groupe ou supérieur, beaucoup plus fréquents – du fait du 2nd cas cette fois – pour les estimations de niveau fin.

Nouvelle procédure d'estimation – principes

- Système Esane riche et complexe ⇒ modification en profondeur de la procédure d'estimation.
- Abandon du principe d'estimation directe systématique :
  - ✓ Estimation directe pour les variables « élémentaires »
  - ✓ Estimation indirecte via les relations comptables pour les variables « soldes ».
- Différenciation de la méthode d'estimation selon le niveau de détail des statistiques produites :
  - ✓ niveau groupe : estimateur par différence avec gestion des estimations problématiques au niveau des agrégats (les estimations de niveaux supérieurs s'en déduisant) ;
  - ✓ niveau infra-groupe : ventilation des estimations de niveau groupe selon des structures observées dans l'enquête.

Procédure d'estimation niveau groupe & supra

❶ Calcul, pour les variables élémentaires, de l'estimateur par différence $\hat{Y}^{G}_{diff}$ au niveau groupe.

❷ Gestion des estimations problématiques sur les agrégats de niveau groupe de variables élémentaires ainsi calculés :

- ✓ pour la ventilation agrégée du CA : mise à zéro de la variable négative, report du montant négatif ainsi traité sur la variable correspondant à la branche principale de l'entreprise et ajustement en conséquence des achats et des variations de stock correspondants ;
- ✓ Pour les autres variables : estimations considérées comme non significatives et donc non diffusées.

❸ Calcul des agrégats « soldes » résultant d'une équation comptable.

❹ Calcul des agrégats de niveau supra-groupe.

Procédure d'estimation niveau infra-groupe

❶ Pour les variables élémentaires, ventilation de l'agrégat niveau groupe $\hat{Y}^{G}_{diff}$ selon la « structure Horvitz-Thompson » propre à chaque variable élémentaire :

$$\hat{Y}^{D}_{ratio} = \hat{Y}^{G}_{diff} \frac{\hat{Y}^{D}_{HT}}{\hat{Y}^{G}_{HT}} = \hat{Y}^{G}_{diff} \frac{\sum_{i \in exh \oplus R} w_i Y_i^{Redi} \mathbb{1}_{Domaine\_enq=D}(i)}{\sum_{i \in exh \oplus R} w_i Y_i^{Redi} \mathbb{1}_{Groupe\_enq=G}(i)}$$

❷ Calcul des agrégats « soldes » résultant d'une équation comptable.

Impact sur la qualité des estimations (1)

- Aux niveaux groupe et supérieurs :
  - ✓ introduction de biais dans les estimations de ventilation chiffre d'affaires ; les estimations de chiffre d'affaires et de valeur ajoutée restent sans biais ;
  - ✓ légère perte d'information – agrégats non diffusés – pour les autres variables.
- Au niveau infra-groupe :
  - ✓ nouvel estimateur asymptotiquement sans biais ;
  - ✓ le nouvel estimateur mobilise de manière moins intensive les informations administratives de niveau fin $\Rightarrow$ théoriquement, perte de précision des estimations.
  - ✓ Calculs de précision effectués sur estimations de niveau sous-classe (code APE sur 5 caractères) pour quantifier l'ampleur de la diminution de précision.

Impact sur la qualité des estimations (2)

Moyennes et quantiles des ratios des écarts-types « estimateur par différence » / « estimateur par ventilation » (niveau 5 caractères de la NAF)

	Chiffre d'affaires	Valeur ajoutée	Excédent brut d'exploitation	Nombre d'entreprises	Effectif salarié	Frais de personnel	Total de l'actif	Total du passif
Moyenne	0,90	0,89	0,84	0,89	0,85	0,89	0,86	0,86
Max	11,53	9,58	4,01	1,53	7,79	10,04	6,04	6,20
Q99	2,85	2,96	3,09	1,42	2,64	3,05	3,33	3,49
Q95	1,53	1,46	1,54	1,15	1,25	1,33	1,82	1,87
Q90	1,21	1,18	1,19	1,07	1,08	1,13	1,20	1,22
Q75	1,00	1,00	1,01	1,00	0,99	1,00	1,00	1,00
Médiane	0,90	0,89	0,86	0,95	0,88	0,88	0,86	0,85
Q25	0,68	0,68	0,52	0,81	0,69	0,68	0,56	0,56
Q10	0,37	0,39	0,22	0,63	0,40	0,37	0,30	0,27
Q5	0,18	0,21	0,13	0,47	0,25	0,26	0,16	0,13
Q1	0,07	0,06	0,02	0,08	0,00	0,05	0,05	0,03
Min	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

➔ Perte de précision mesurée

Conclusion

- La gestion des estimations problématiques du système Esane a nécessité des changements importants par rapport à la méthode proposée initialement :
  - ✓ abandon du principe d'estimation directe pour toutes les variables ;
  - ✓ méthodes d'estimations différenciées selon le niveau de détail des statistiques produites, avec en particulier l'abandon de l'estimateur par différence au niveau infra-groupe au profit d'un estimateur ventilant les agrégats groupes selon des structures observées dans l'enquête.
- La nouvelle mécanique d'estimation garantit un signe valide aux agrégats obtenus et assure par construction une cohérence parfaite des estimations, au prix :
  - ✓ d'une légère perte d'information au niveau groupe ;
  - ✓ d'une perte de précision mesurée au niveau infra-groupe.

Merci de votre attention !

Contact :

M. Gros Emmanuel
Tél. : 01 41 17 56 87
Courriel : emmanuel.gros@insee.fr

Insee
18 bd Adolphe-Pinard
75675 Paris Cedex 14
www.insee.fr

Informations statistiques :
www.insee.fr / Contacter l'Insee
09 72 72 4000
(coût d'un appel local)
du lundi au vendredi de 9h00 à 17h00

INSEE