Le dispositif Esane et les estimations composites

Comment l'utilisation combinée de données administratives et de données d'enquête permet d'améliorer la qualité des données individuelles et des statistiques





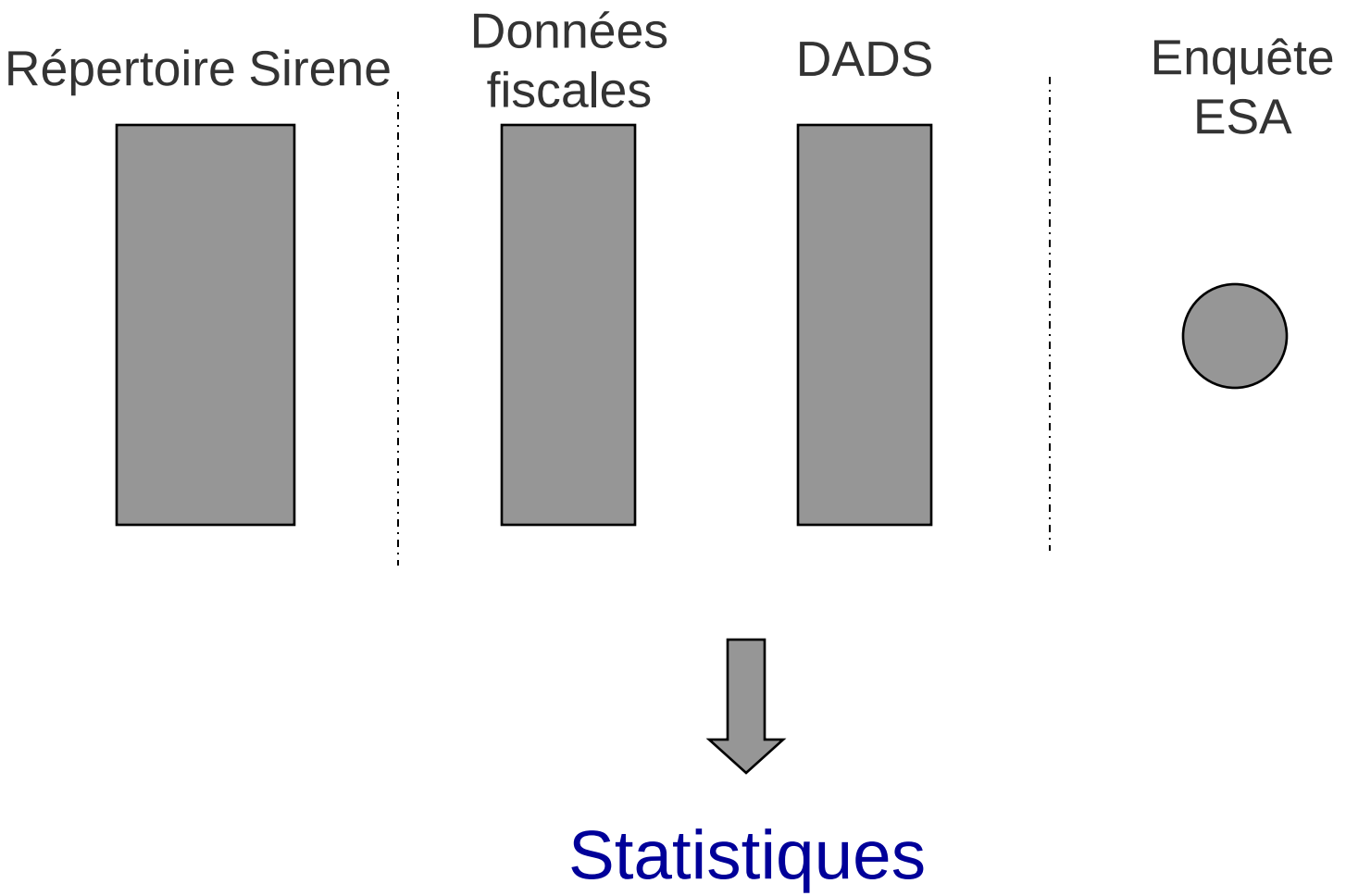
Gros Emmanuel
Insee – DMCSI – Division Sondages

INSEE
INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Le contexte du système Esane

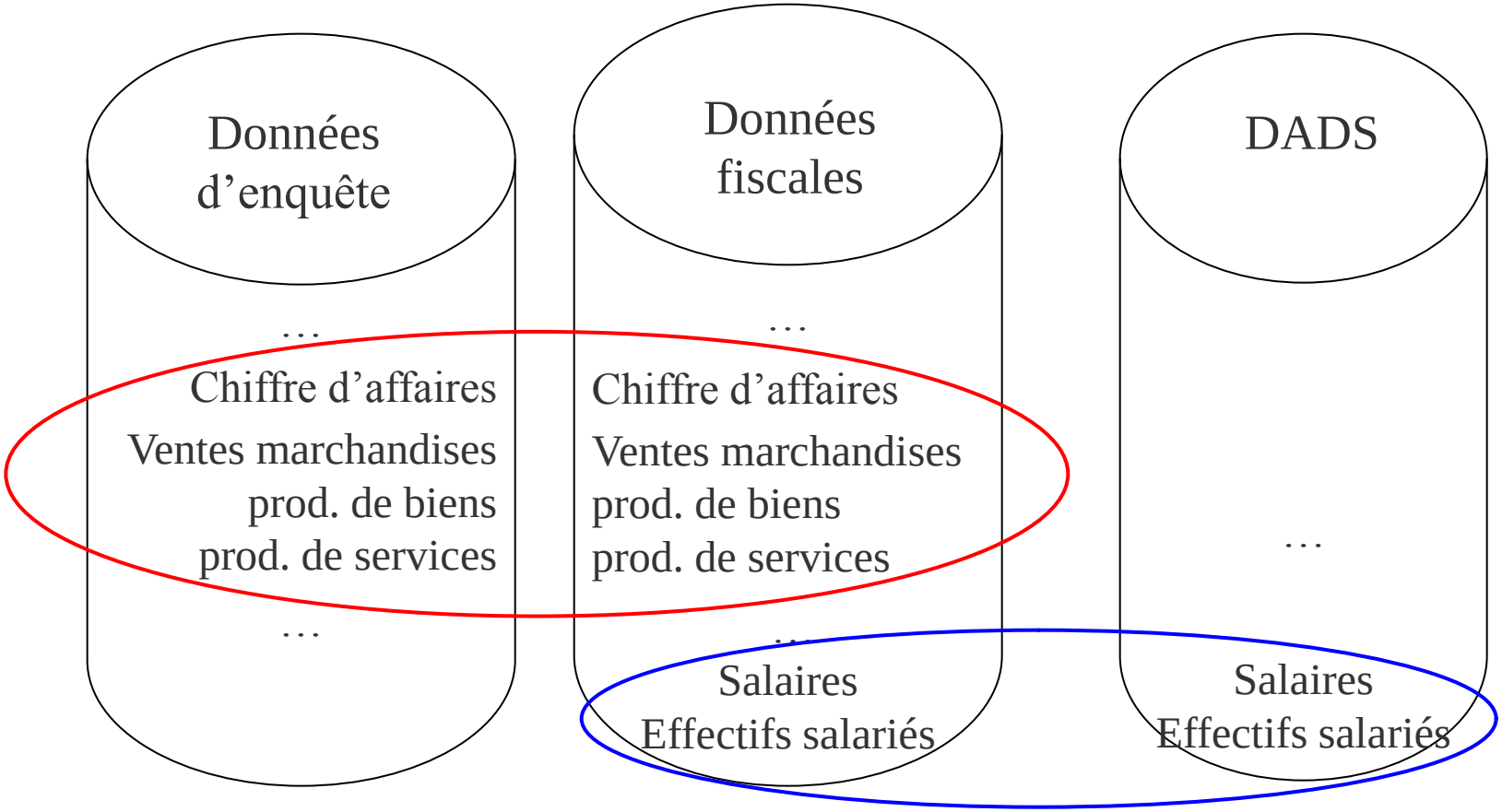
- Résultat d'une refonte complète du processus de production des statistiques structurelles d'entreprises :
  - ✓ Auparavant, deux dispositifs coexistaient en parallèle : enquêtes statistiques (EAE) et exploitation de données fiscales (Suse) ;
  - ✓ Nouveau système fondé sur l'utilisation intensive de 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 $\Rightarrow$ phase de réconciliation des données individuelles ;
  - ✓ en termes d'estimation $\Rightarrow$ utilisation de techniques de calage et d'estimateurs « composites ».

Un dispositif multi-source



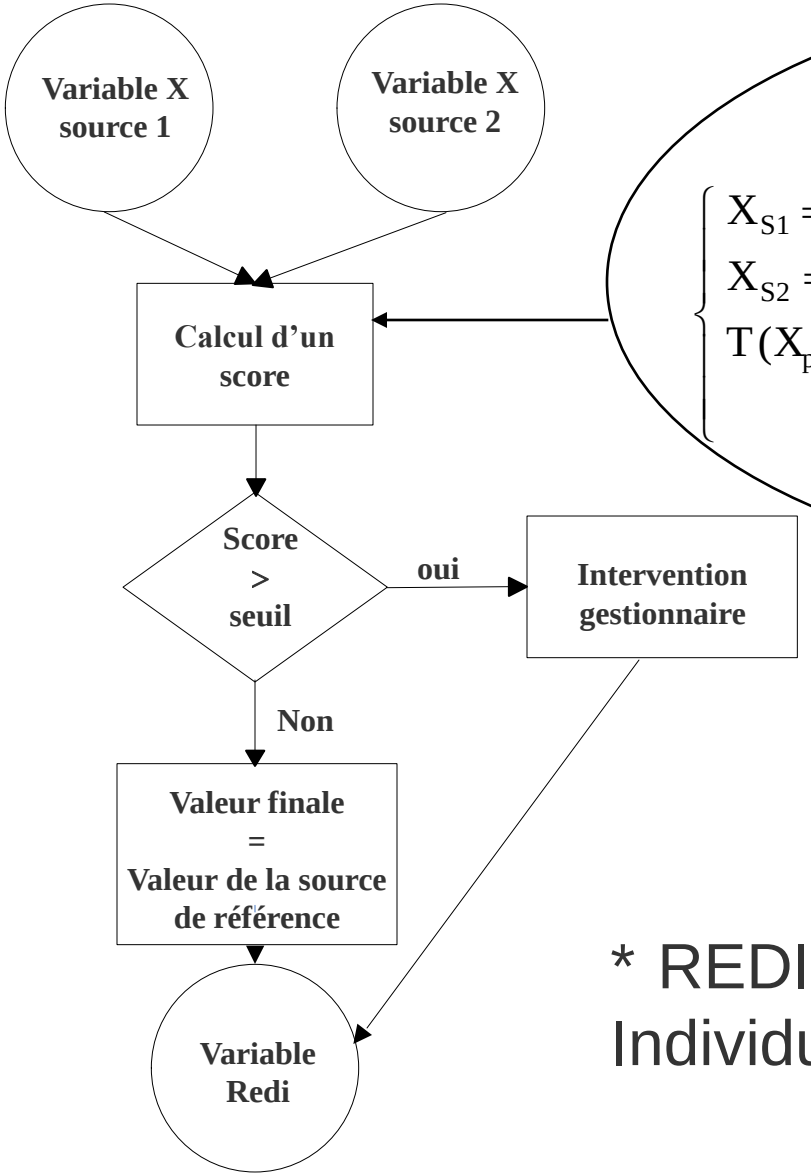
Réconciliation des données individuelles

- 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*



$$\text{score} = \left| \frac{X_{S1} - X_{S2}}{T(X_p)} \right|$$

X_{S1} = valeur de la variable X dans la source 1

X_{S2} = valeur de la variable X dans la source 2

$T(X_p)$ = total de la variable X dans la source de référence, pour le niveau d'agrégation retenu pour le contrôle

* REDI : « REconciliation des Données Individuelles

Impact de Redi sur les données individuelles (1)

- En ce qui concerne le chiffre d'affaires, les données administratives et les données d'enquête sont globalement cohérentes :
  - ✓ 60 % de chiffre d'affaires identiques dans les deux sources, seules 14% des entreprises présentent une différence relative supérieure à 15 % (en valeur absolue) ;
  - ✓ divergence faible – environ 1% – au niveau des agrégats.
- Des divergences plus importantes au niveau de la ventilation du CA.

Variable	Total enquête	Total fiscal	Total final
Chiffre d'affaires	3 338	3 305	3 316
Ventes de marchandises	1 378 41,3%	1 352 40,9%	1 350 40,7%
Production de biens	998 29,9%	841 25,4%	1 079 32,5%
Production de services	962 28,8%	1 112 33,7%	887 26,8%

Impact de Redi sur les données individuelles (2)

- Un choix des sources de référence du processus Redi validé par les contrôles des gestionnaires :

  ⇒ valeur Redi sélectionnée par les gestionnaires après contrôle = valeur de la source de référence dans plus de 95 % des cas pour le CA, dans plus de 80 % des cas pour sa ventilation agrégée.

- Par conséquent, au niveau individuel :

  - ✓ pour le CA, la valeur issue des sources fiscale a été retenue pour 97 % des unités, représentant 98 % du CA total ;
  - ✓ pour la ventilation agrégée, choix de la structure dérivée de l'enquête pour 86 % des unités, représentant 87 % du CA total.

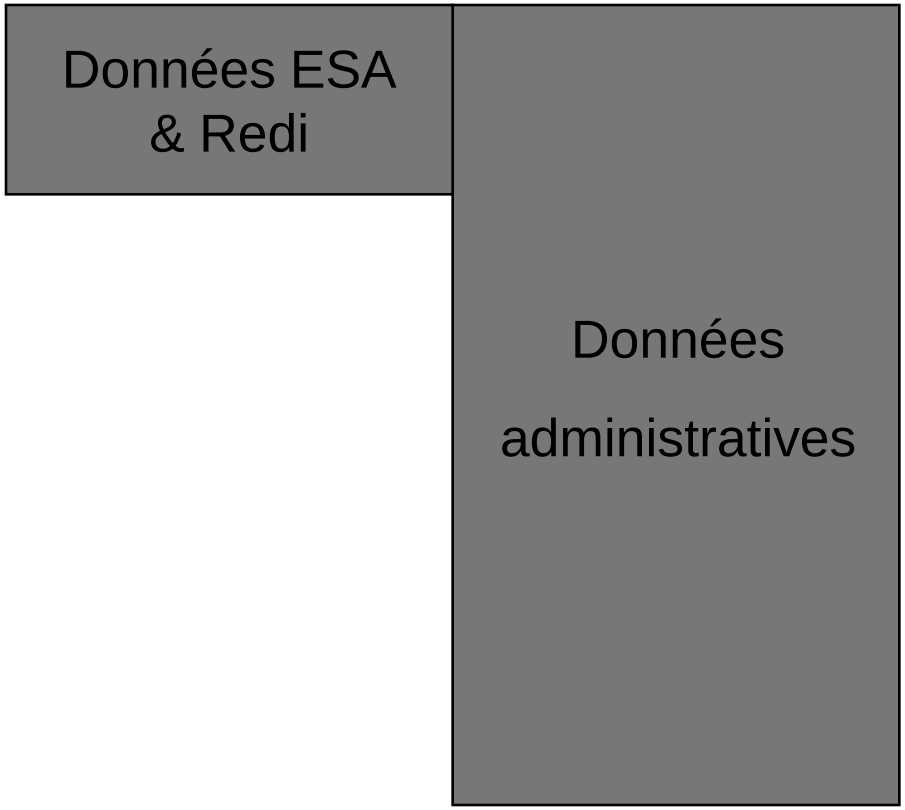
- Efficacité du processus Redi :

  - ✓ moins de 4 % des unités de l'échantillon sont détectées par le processus comme « sérieusement incohérentes »...
  - ✓ ... mais ces unités représentent plus de 40 % de la différence absolue en termes de chiffre d'affaires.

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 essentiellement la ventilation agrégée du chiffre d'affaires ;

  ✓ 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 ;
- ✓ 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)

- Objectif : évaluer l'impact des améliorations méthodologiques – calage et estimateurs spécifiques – mises en œuvre dans le nouveau système.

➔ Simulation d'estimateurs « type EAE » et comparaison de leurs CV avec ceux des nouveaux estimateurs.

- Estimateurs « type EAE » : $\sum_{i \in R} d_i Y_i \, \mathbb{1}_{APE\_enq=X}(i)$

- CV calculés grâce à une macro SAS, qui prend en compte :
  - ✓ pour les estimateurs simulés relatifs à l'ancien système : l'erreur d'échantillonnage de l'enquête, liée au plan de sondage stratifié et à la correction de la non-réponse totale par groupes de réponses homogènes ;
  - ✓ pour les estimateurs Esane : même chose + étape de calage et utilisation d'estimateurs spécifiques.

Impact sur la qualité des estimations (2)

Secteur	CV des estimations relatives à l'ancien système							
	Nombre d'entreprises	Chiffre d'affaires	Ventes de marchandises	Production de biens	Production de services	Salaires	Valeur ajoutée	EBE
IAA	3,9%	0,5%	1,2%	0,5%	0,8%	6,7%	0,8%	0,5%
Construction	0,9%	1,1%	9,1%	1,1%	5,7%	1,5%	1,4%	3,1%
Commerce	1,1%	0,4%	0,4%	2,3%	0,9%	1,1%	0,6%	1,2%
Industrie	1,3%	0,1%	0,3%	0,1%	0,2%	2,5%	0,1%	0,2%
Services	0,5%	0,4%	1,3%	3,5%	0,4%	1,2%	0,5%	1,0%
Transport	2,1%	0,4%	4,3%	1,0%	0,4%	3,6%	0,5%	0,6%
Total	0,40%	0,20%	0,37%	0,28%	0,32%	0,67%	0,26%	0,52%
Sector	CV des estimations relatives au système Esane							
	Nombre d'entreprises	Chiffre d'affaires	Ventes de marchandises	Production de biens	Production de services	Salaires	Valeur ajoutée	EBE
IAA	**3,2%**	**0,3%**	*2,1%*	**0,3%**	*13,3%*	**3,3%**	**0,4%**	**0,3%**
Construction	**0,3%**	**0,6%**	*25,6%*	**0,8%**	*51,4%*	**0,4%**	**1,2%**	*3,3%*
Commerce	**0,5%**	**0,1%**	**0,1%**	*4,1%*	*1,5%*	**0,4%**	**0,4%**	**1,0%**
Industrie	**1,3%**	**0,1%**	*0,3%*	**0,1%**	*1,0%*	**1,4%**	**0,1%**	**0,2%**
Services	**0,3%**	**0,2%**	*4,5%*	*14,7%*	**0,3%**	**0,2%**	**0,3%**	**0,7%**
Transport	**0,6%**	**0,2%**	*5,9%*	*2,3%*	**0,1%**	**1,2%**	**0,1%**	**0,2%**
Total	**0,09%**	**0,05%**	**0,14%**	**0,22%**	**0,25%**	**0,05%**	**0,15%**	**0,37%**

Impact sur la qualité des estimations (3)

Moyennes et quantiles des ratios des CV « estimateur EAE » / « estimateur Esane » (niveau 3 caractères de la NAF)

	Nombre d'entreprises	Chiffre d'affaires	Ventes de marchandises	Production de biens	Production de services	Salaires	Valeur ajoutée	EBE
Moyenne	0,78	0,68	3,64	2,82	9,13	0,65	0,78	0,93
Max	2,20	2,26	130,01	98,94	325,05	3,36	6,34	10,09
Q99	1,89	1,97	35,68	47,93	104,51	3,04	5,13	9,93
Q95	1,02	1,02	11,24	9,01	45,11	1,37	1,80	2,65
Q90	0,99	0,97	6,36	4,43	18,59	1,07	1,12	1,40
Q75	0,93	0,89	2,54	1,40	5,23	0,89	0,92	0,98
Médiane	0,81	0,73	1,34	0,86	1,00	0,68	0,72	0,71
Q25	0,63	0,47	0,78	0,66	0,73	0,41	0,46	0,40
Q10	0,46	0,24	0,40	0,32	0,52	0,20	0,19	0,20
Q5	0,37	0,11	0,30	0,15	0,37	0,13	0,08	0,09
Q1	0,17	0,00	0,00	0,09	0,23	0,03	0,00	0,00
Min	0,16	0,00	0,00	0,07	0,16	0,02	0,00	0,00

➔ Amélioration globale de la qualité, sauf pour la ventilation du chiffre d'affaires.

Impact sur la qualité des estimations (4)

Moyennes et quantiles des ratios des ET « estimateur EAE » / « estimateur Esane » (niveau 5 caractères de la NAF)

	Nombre d'entreprises	Chiffre d'affaires	Ventes de marchandises	Production de biens	Production de services	Salaires	Valeur ajoutée	EBE
Moyenne	0,94	0,91	5,34	4,98	10,63	1,00	1,02	1,18
Max	17,08	2,25	352,95	197,15	918,85	9,75	14,20	17,63
Q99	1,21	1,82	73,30	103,03	140,78	3,48	4,39	6,08
Q95	1,06	1,11	18,12	17,61	34,31	1,50	1,59	2,51
Q90	1,02	1,04	6,24	6,21	15,49	1,16	1,20	1,67
Q75	1,00	1,00	1,98	1,71	3,35	1,04	1,03	1,09
Médiane	0,98	0,97	1,11	1,00	1,20	0,98	0,97	0,96
Q25	0,89	0,84	0,99	0,95	0,98	0,85	0,85	0,84
Q10	0,74	0,61	0,90	0,74	0,78	0,60	0,58	0,59
Q5	0,60	0,39	0,65	0,52	0,62	0,42	0,45	0,41
Q1	0,19	0,08	0,23	0,19	0,32	0,03	0,10	0,08
Min	0,00	0,00	0,00	0,07	0,14	0,00	0,00	0,00

➔ Résultats plus mitigés, même si on observe quand même plutôt un gain de précision, hors ventilation du CA toujours...

Conclusion

- Le dispositif multi-sources Esane présente de nombreux avantages...
  - ✓ amélioration de la cohérence des résultats ;
  - ✓ permet la mise en place de contrôles de cohérence sur quelques variables clefs $\Rightarrow \Delta^{-}$ biais liés aux erreurs de réponse ;
  - ✓ permet l'utilisation de techniques de calage et d'estimateurs spécifiques, ce qui conduit à une amélioration globale de la qualité des estimations sectorielles (à l'exception des variables de ventilation agrégée du chiffre d'affaires).
- ... au prix d'une plus grande complexité du système, qui ne va pas sans poser quelques problèmes
  - ✓ problèmes d'estimation négatives « à tort » lié aux estimations par différence $\Rightarrow$ des estimateurs finaux différents en fonction des niveaux d'agrégation ;
  - ✓ un système plus complexe $\Rightarrow$ plus difficile d'évaluer la contribution d'une unité aux agrégats.

Merci de votre attention !

Contact :

M. Gros Emmanuel
Tél. : 01 41 17 64 91
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