La gestion des unités influentes dans l'ESA par winsorisation





Philippe Brion
Emmanuel Gros
Fabien Guggemos

Insee
Mesurer pour comprendre

02/07/2013

L'enquête sectorielle annuelle (ESA)

- Enquête annuelle
- Variables recueillies auprès des entreprises :
  - ✓ en premier lieu le chiffre d'affaires et sa répartition par activité ;
  - ✓ des variables sectorielles spécifiques.
- Exploitation conjointe avec les sources fiscales → statistiques sur données comptables.
- Outputs : statistiques
  - ✓ par secteur d'activité ;
  - ✓ et par branches d'activité.

Le problème des « representative outliers »

- Plan de sondage de l'ESA :
  - ✓ échantillon renouvelé par moitié
  - ✓ sondage aléatoire simple stratifié à un seul degré

strates ≈ secteur d'activité ⊗ tranche d'effectifs

➔ Apparition de points atypiques non aberrants (representative outliers, cf. Chambers (1986)). Il s'agit d'unités :

  - ✓ dont les réponses sont anormalement élevées par rapport aux autres unités de leur strate d'appartenance...
  - ✓ ...sans qu'il s'agisse pour autant d'erreurs de mesure.

Le problème des « representative outliers » (2)

- Pourquoi ces points atypiques non aberrants ?

  ↳ À cause du différentiel entre données de la base de sondage utilisées pour construire les strates (avant le début de l'enquête) & données recueillies lors de l'enquête.

Changement de secteur d'activité Evolution à la hausse des effectifs	⇒	Entreprise classée dans une mauvaise strate ↳ En général, poids de sondage trop élevé

- Ces points augmentent fortement la variance des estimations → nécessité de les « traiter ».

Le problème des « representative outliers » (3)

- Objectif : réduire la variance des estimations

  ↳ Par modification de certaines des caractéristiques des unités atypiques non aberrantes : réduction de la valeur extrême déclarée ou alternativement diminution du poids de sondage de l'unité.

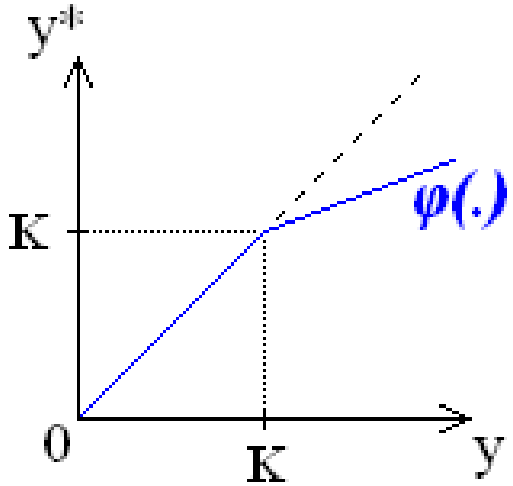
  → Prix à payer : introduction d'un léger biais à la baisse

⇒ Procédure de winsorisation, à évaluer en termes d'erreur quadratique moyenne (EQM).

Winsorisation dans l'ESA : principe

Réduire la dispersion des valeurs observées en révisant à la baisse les valeurs extrêmes

$$y_{hi}^{*} = \varphi_h(y_{hi}) = \begin{cases} \frac{n_h}{N_h} y_{hi} + \left(1 - \frac{n_h}{N_h}\right) K_h & \text{si} \quad y_{hi} \geq K_h \\ y_{hi} \quad \text{si} \quad y_{hi} < K_h \end{cases}$$



Pas de correction dans les strates exhaustives (poids égal à 1)

$$\hat{Y}^{\text{Winsor}} = \sum_{i \in s} w_i y_i^{*} = \sum_{h=1}^{H} \left(\frac{N_h}{n_h}\right) \sum_{i \in s_h} y_{hi}^{*}$$

Winsorisation dans l'ESA : principe (2)

- Comment déterminer les seuils par strates K_h ?
  - ✓ Hypothèse : y_{hi} variables aléatoires i.i.d, espérance μ_h
  - ✓ Minimisation de l'EQM de l'estimateur $\hat{Y}^{Winsor}$ du total de Y.

→ Kokic & Bell (1994) : asymptotiquement, quand EQM minimisée

$$\forall h,\ (N_h/n_h-1)(K_h-\mu_h) \sim \text{Biais de } \hat{Y}^{Winsor}(K_1,\ldots K_H)$$

↳ Système de H équations à H inconnues qui se ramène à 1 équation sur le biais, $F(B)=0$, puis $K_h = -(N_h/n_h-1)^{-1}B + \mu_h$

Définition des seuils par strate → problème de dimension 1

Winsorisation sur le poids / sur les variables ?

- Déterminer des seuils de winsorisation pour chaque variable d'intérêt ?

  ↳ Possible, mais détruit la cohérence entre le traitement des variables (ex: 1 entreprise atypique pour une variable, mais pas pour une autre variable…)

➔ Choix retenu pour l'ESA :

❶ Seuils K_h pour le chiffre d'affaires CA

❷ Pour toute autre variable z, $z^*_{hi} = z_{hi} \dfrac{CA^*_{hi}}{CA_{hi}}$

↳ <u>hypothèse :</u> bonne corrélation entre CA et z

Winsorisation sur le poids / sur les variables ? (2)

- Permet de « transférer » la winsorisation de la variable vers le poids :

$$w_{hi} z^{*}_{hi} = w_{hi} \frac{CA^{*}_{hi}}{CA_{hi}} z_{hi} = w^{*}_{hi} z_{hi}, \quad \text{avec} \quad w^{*}_{hi} = w_{hi} \frac{CA^{*}_{hi}}{CA_{hi}}$$

↳ poids « winsorisé » unique pour chaque entreprise :

$$w^{*}_{hi} = 1 + \frac{K_h \left(w_{hi} - 1 \right)}{CA_i}$$

→ de sorte que $\forall z, \quad \hat{Z}^{\text{Winsor}} = \sum_s w^{*}_{hi} z_{hi} \quad \left(= \sum_s w_{hi} z^{*}_{hi} \right)$

Simulations et déterminations des seuils K_h

- Seuils déterminés à partir de données passées (EAE 2007)
- Procédure de winsorisation appliquée pour chaque domaine de diffusion (secteurs agrégés)
- Au final :
  - ↳ seulement 343 unités winsorisées sur près de 150 000…
  - ↳ mais des gains élevés en EQM et de faibles biais pour les variables d'intérêt étudiées

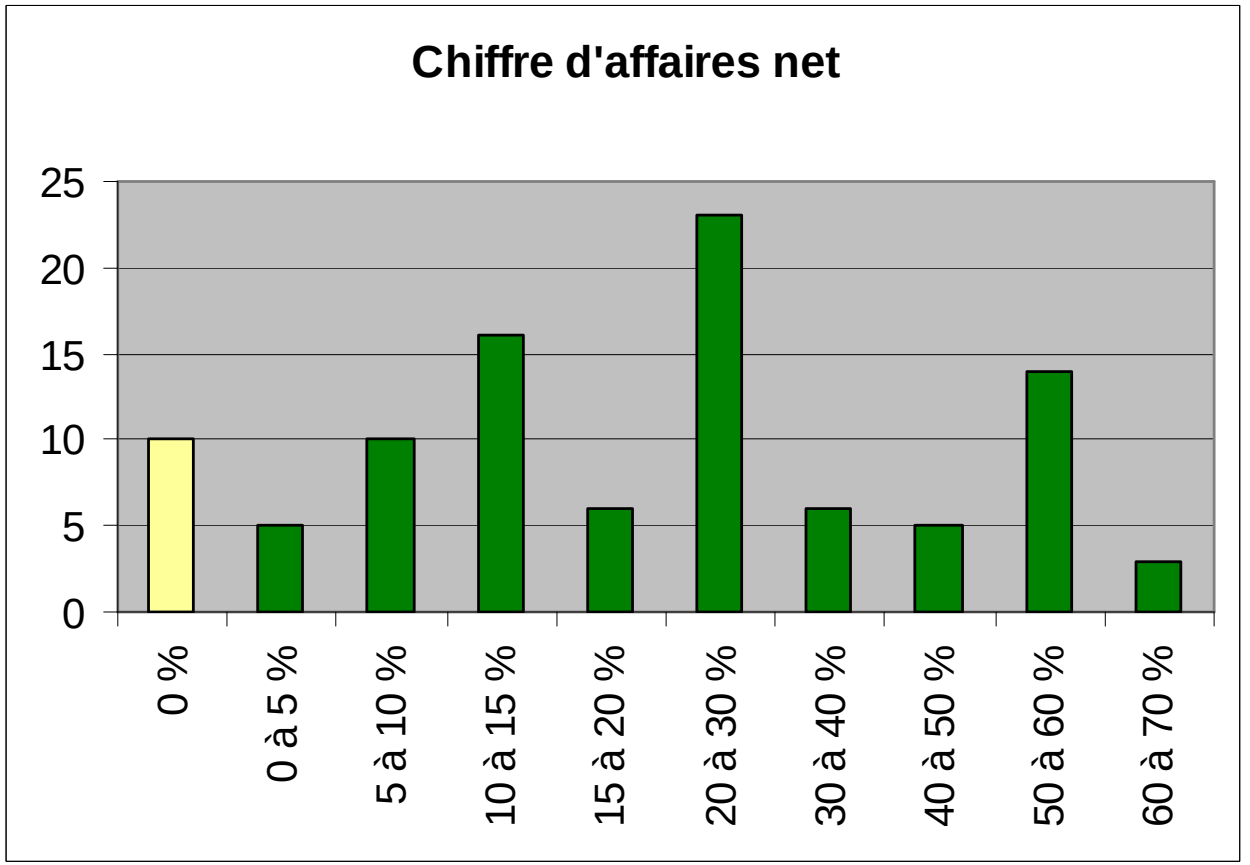
Résultats des simulations

Exemple du commerce de détail (poste 47 de la nomenclature d'activité française), variable d'intérêt = CA

Domaine de diffusion	Biais relatif (en %)	Gain en EQM (en %)
47.1	0.166	30.588
47.2	0.360	3.392
47.3	1.222	14.097
47.4	0.807	12.307
47.5	1.731	50.829
47.6	0.475	12.706
47.7	0.452	28.416
47.8	0.441	3.301
47.9	0.627	20.502

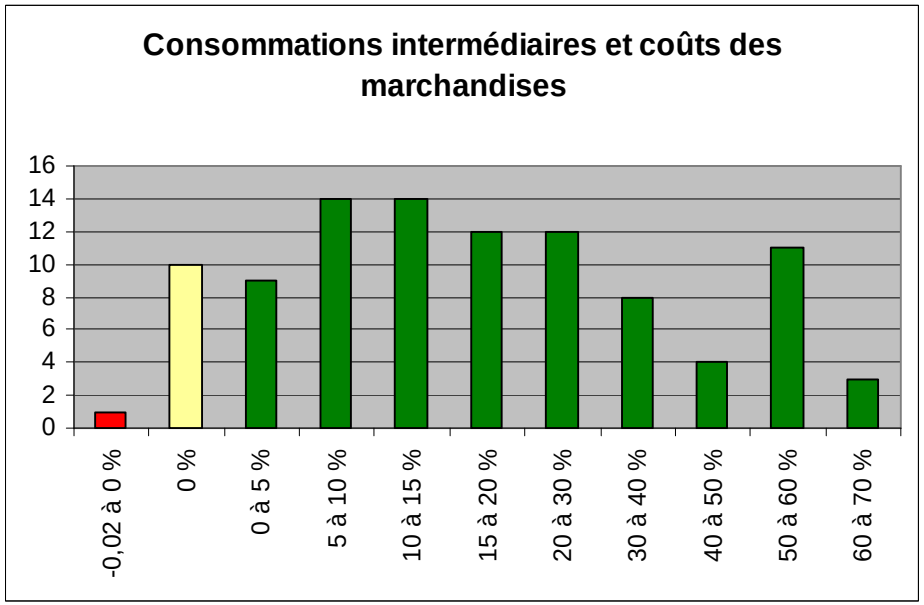
Résultats des simulations (2)

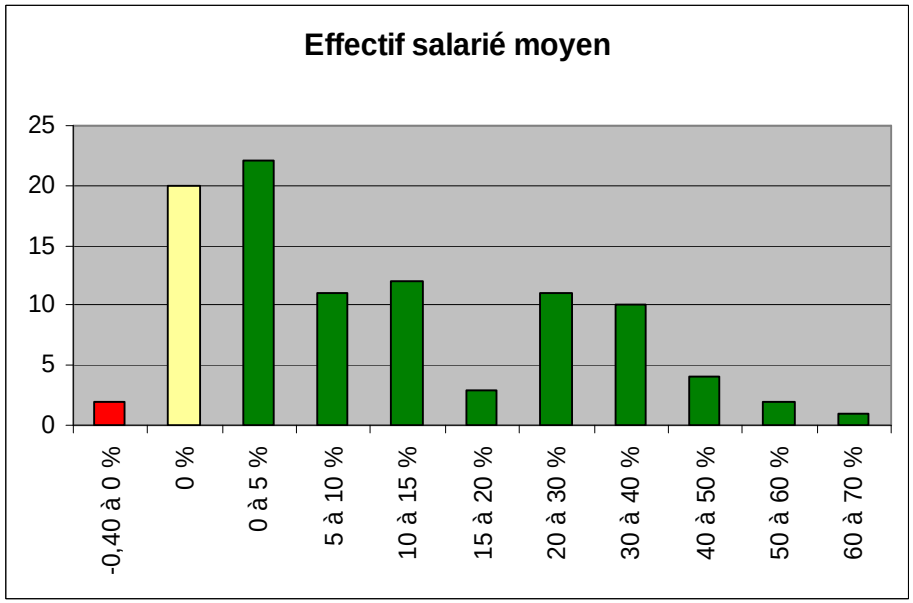
Nombre de domaines de diffusion par tranches de gains en erreur quadratique moyenne



Résultats des simulations (3)

Et sur des variables autres que le CA…





Mise en œuvre dans l'ESA 2008

- Winsorisation effectuée après CNR et calage.
- Deux difficultés rencontrées :

  ❶ pas de seuils calculés pour les secteurs de l'industrie et des IAA, faute de données EAE disponibles.

  ↳ calcul de seuils « moyens » par tranche de taille pour effectuer une winsorisation *a minima*

  ❷ winsorisation appliquée à estimateur calé → les poids utilisés ne sont pas les poids de sondage initiaux mais des poids issus d'une CNR et d'un calage.

  ↳ adaptation des seuils en conséquence : $K^{*}_{hi} = K_h \frac{d_i}{w_i}$

- 245 unités winsorisées, pour un impact à la baisse de 16 Md€ sur l'estimateur du CA total, soit 0,4 % → cohérent avec les résultats des simulations sur EAE 2007.

Mise en œuvre dans l'ESA 2009

- Problème lié au renouvellement par moitié de l'échantillon de l'ESA : on conserve la moitié de l'échantillon avec les caractéristiques – strates de tirage, poids de sondage – relatives au tirage N-1.

➔ favorise l'apparition de « stratum jumpers » importants, qui viennent perturber la procédure de calage !

- Un exemple d'unité winsorisée:
  - ✓ sélectionnée en 2008 avec un poids de 140 (pour cause de CA et d'effectif au lancement non renseignés), conservée en 2009 ;
  - ✓ poids de 166 après CNR ⊕ CA 2009 de 180 M€ : contribution de 30 Md€ à l'agrégat sectoriel → calage sur le CA par raking-ratio ramène à zéro le poids de cette unité, et calage borné impossible...

↳ passage à une winsorisation effectuée avant le calage.

Mise en œuvre dans l'ESA 2009 (2)

- La winsorisation avant calage permet de traiter les unités atypiques qui perturbent la procédure de calage

  ↳ intérêt « pratique » de la winsorisation...

- Retour sur l'exemple précédent :

  ✓ la winsorisation ramène le poids après CNR de 166 à 2,4 ;

  ↳ calage avec méthode logit et bornes à [0,5 – 2] passe sans problème...

- Au final, 223 unités winsorisées, pour un impact à la baisse de 51 Md€ sur l'estimateur avant calage du CA total → Δ^+ impact de la winsorisation, cohérent avec le fait que la stratégie de renouvellement par moitié « favorise » l'apparition d'unités « fortement » atypiques.

Mise en œuvre dans les ESA 2010 & 2011

- Aucun changement de méthodologie par rapport à la campagne 2009.

- En 2010, 227 unités winsorisées, pour un impact à la baisse de 36 Md€ sur l'estimateur avant calage du CA total.

- En 2011, 265 unités winsorisées, pour un impact à la baisse de 35 Md€ sur l'estimateur avant calage du CA total.

↳Processus « stabilisé » en régime de croisière...

Analyse de la winsorisation sur Esane 2010

- Objectif : évaluer l'impact, en termes d'EQM, de la procédure de winsorisation sur les estimations d'une campagne Esane.

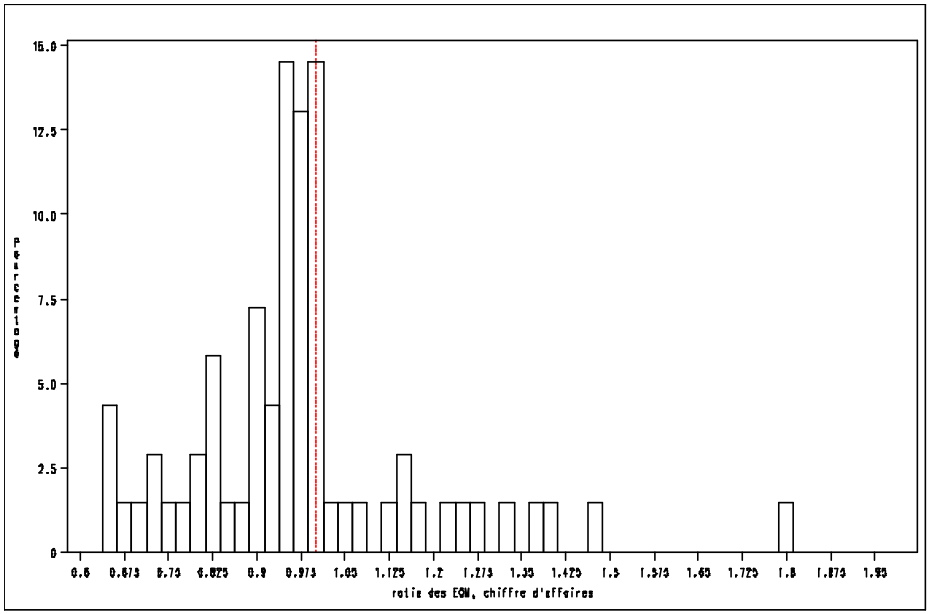
  ↳ Évaluation rigoureuse « impossible » : nécessiterait de réaliser un calage sans winsorisation, ce que la présence de point atypiques rend très difficile...

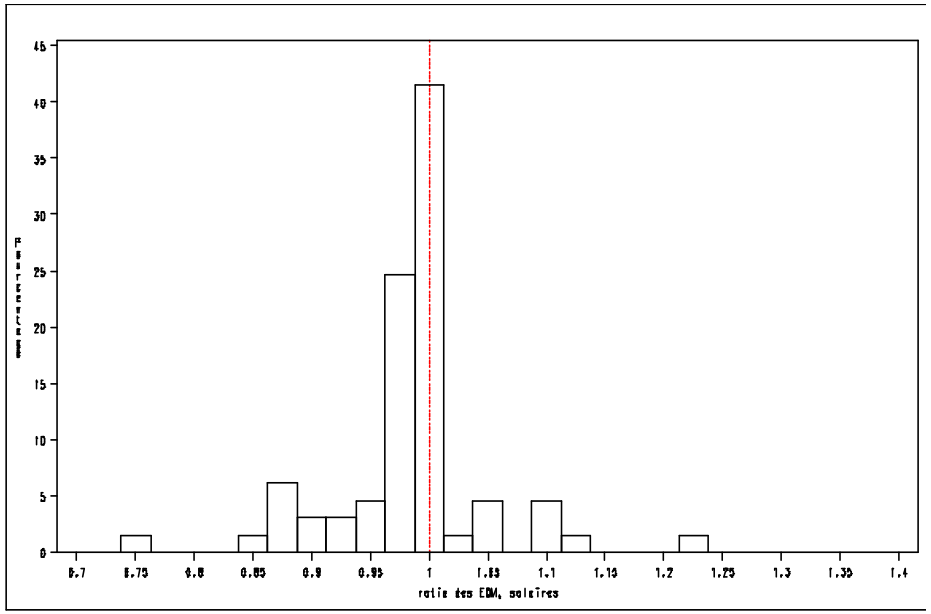
- Méthode retenue : raisonner sur l'estimateur HT après correction de la non-réponse mais avant calage

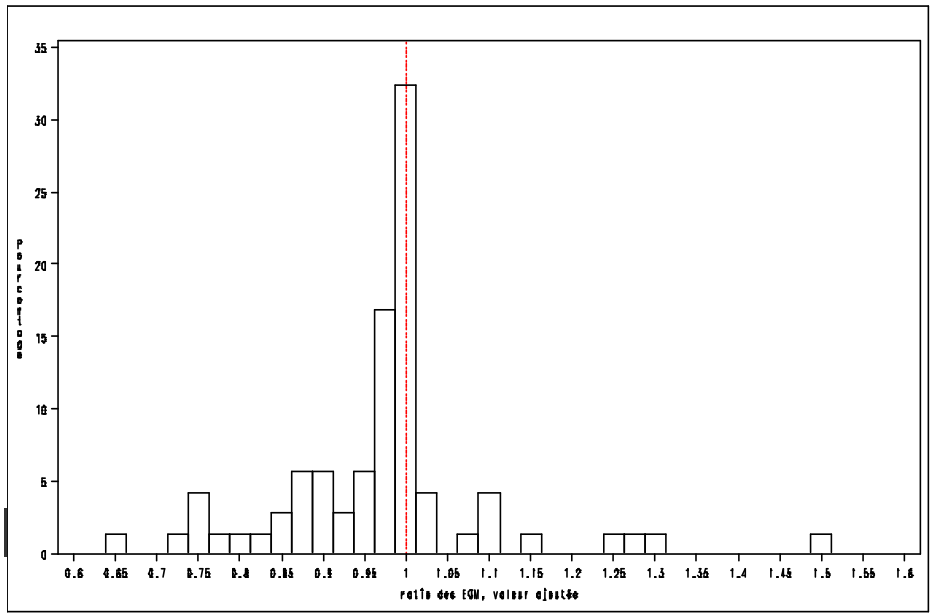
  ↳ calcul de la précision des estimations par groupe sans winsorisation et avec winsorisation

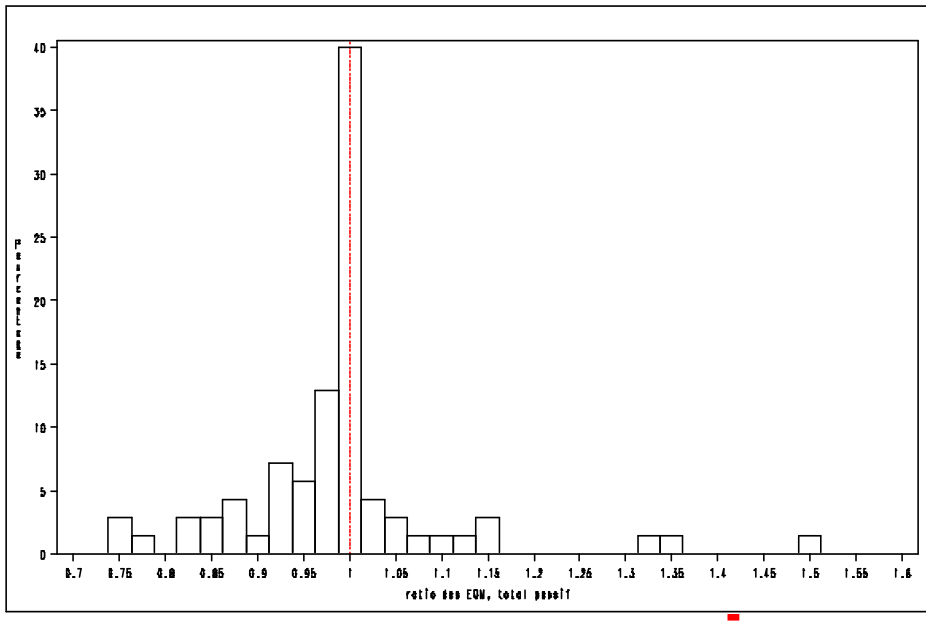
  ↳ comparaison des EQM estimés entre les deux scénarios

Analyse de la winsorisation sur Esane 2010 (2)









Perspectives

- Analyser plus en détail les interactions entre winsorisation et calage.

- À faire : actualiser les seuils K_h de la procédure

  ↳ Actualisation possible chaque année, en appliquant la procédure de Kokic & Bell à l'échantillon de l'ESA de la campagne en cours ;

  ↳ Détermination des K_h par l'approche biais conditionnel et estimateur « min-max » ?

  ↳ Actuellement, seuils déterminés pour estimation du CA par groupe → quid de la qualité de la winsorisation sur l'estimation globale ?

- Procédures de winsorisation sur d'autres variables, au moins comme aide à la vérification des données...

Merci de votre attention !

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