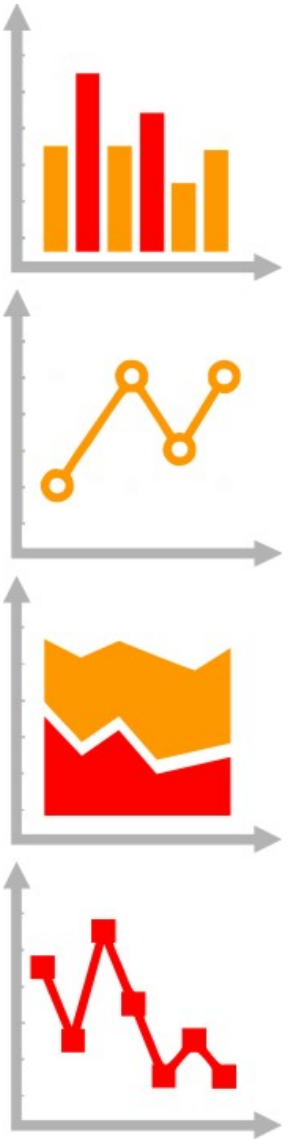
Comment redresser un échantillon d'unités légales tirées via leurs entreprises ?



L'adaptation de la correction de la non-réponse, du traitement des valeurs influentes et du calage de l'Enquête Sectorielle Annuelle et de l'Enquête Annuelle de Production à leur nouveau plan de sondage

JMS 2018

Insee
Mesurer pour comprendre

Les traitements post-collecte

Partie exhaustive (120 000 UL environ):

- Les traitements post-collecte se limitent à de l'imputation pour corriger la non-réponse (partielle ou totale). => Non concernée par la présentation.

Partie non-exhaustive (60 000 UL environ):

- Correction de la non-réponse
- Traitement des valeurs influentes
- Calage

L'essentiel de la présentation portera sur le traitement des valeurs influentes car la méthode utilisée jusqu'ici (Winsorisation avec seuils de Kokic et Bell) a été conçue pour s'appliquer dans le strict cadre d'échantillons aléatoires simples stratifiés.

Correction de la non-réponse totale

- **Méthode : repondération au sein de groupes de réponse homogène (variables mobilisées : APE, effectif, chiffre d'affaires, présence d'une liasse fiscale, ancienneté de l'unité, zone géographique...).**

- **Pas d'étude possible avant les premieres réponses...**

- **Pour les traitements 2016 (délais serrés) : Adaptation minimale : ajout de l'indicatrice d'appartenance à une EP comme variable supplémentaire et potentiellement mobilisable (mais peu mobilisée) pour constituer les groupes de réponse homogène.**

CNRT : analyse après traitement

Etre rattachée à une EP ne semble pas jouer "directement" sur la probabilité de réponse d'une UL.

- La variable n'est pas significative avec une régression logistique cherchant à prédire le fait de répondre lorsque les autres variables habituellement utilisées (activité, taille...) pour les GRH sont présentes.

Les UL appartenant à une même EP semblent avoir des comportements de réponse corrélés.

- On trouve davantage de cas où toutes les UL répondent (ou aucune UL ne répond) qu'on ne le devrait en se basant sur les comportements de réponse des UL indépendantes… Le problème ne concerne que peu d'UL. Continuer à utiliser les méthodes traditionnelles, ne devrait donc pas poser de gros problèmes.

Traitement des valeurs influentes

Les valeurs influentes étudiées ici sont supposées ne pas être des erreurs, c'est-à-dire que c'est la combinaison d'un poids souvent élevé et d'une valeur importante (mais "vraie") d'une variable qui fait que cette dernière est influente.

- **Strata jumper par exemple.**

Les unités de la partie exhaustive, du fait que leur poids vaut 1, ne peuvent pas être influentes au sens que nous donnons à ce terme dans notre étude.

Winsorisation – Cadre et principe

Cadre : Sondage aléatoire simple stratifié.

Principe :

- **choix d'une variable d'intérêt Y**
- **definition de seuils K_h dans chaque strate**
- **Calcul de la variable winsorisée Y_w obtenue en "rabottant" les Y des unités dépassant le seuil relatif à leur strate.**

$$\begin{cases} y_i^w = y_i \text{ si } y_i < K_h \\ y_i^w = \frac{n_h}{N_h} y_i + \left(1 - \frac{n_h}{N_h}\right) K_h \text{ si } y_i \geqslant K_h \end{cases}$$

Comment redresser un échantillon d'UL tirées via leurs EP ?

Insee

Winsorisation dans Esane

Variable winsorisée : Chiffre d'affaires fiscal

Winsorisation appliquée par groupe (NACE, 3 positions)

K_h obtenus par la méthode de Kokic et Bell

- **Minimisation de l'erreur quadratique moyenne de l'estimateur du total de Y**

Changements dûs au nouveau plan de sondage

Depuis l'édition 2016, on tire des EP mais on se base sur les réponses des UL pour élaborer les résultats (notamment les résultats en UL).

- **Le plan de sondage n'est plus un sondage aléatoire simple stratifié d'UL, mais un tirage aléatoire simple stratifié de grappes d'UL (les EP).**

La méthode de Kokic et Bell a été développée dans le cadre des tirages aléatoires simple stratifiés… Il n'y a pas de garantie que la méthode fonctionne avec le nouveau plan de sondage.

Une étude pour vérifier que la méthode fonctionne encore...

Une étude se basant sur 1000 replications du nouveau plan de sondage a été conduite.

On y compare :

- **L'estimateur d'Horvitz-Thompson ;**
- **L'estimateur « Kokic and Bell » appliqué comme si les UL étaient tirées selon un tirage aléatoire simple stratifié d'UL ;**
- **Deux versions d'estimateurs robustes basés sur les biais conditionnels et qui tiennent compte du vrai plan de sondage.**

Les estimateurs robustes

Il s'agit d'une alternative possible à la Winsorisation.

Biais conditionnel : $B_i = E_p\left(\hat{t}_y / I_i = 1\right) - t_y$

- ***C'est une mesure de l'influence de i.***

Estimateur robuste : $\hat{t}_{yR} = \hat{t}_y - \frac{1}{2}\left(B_{min} + B_{max}\right)$

2 versions testées dans l'étude :

V1 – Tirage de Poisson des entreprises (facile à implémenter);

V2 – Tirage stratifié d'entreprises (vrai plan de sondage).

Simulations (1/2)

Sélection de 1000 échantillons avec le nouveau plan de sondage;

Estimation pour les 207 activités du chiffre d'affaires total avec les différents estimateurs présentés precedemment.

Comme le chiffre d'affaires est connu pour l'ensemble des UL de la BdS, le "vrai total" est connu...

On calcule pour chaque activité et chaque estimateur l'erreur quadratique moyenne associée (MSE pour Mean Square Erreur).

Formule pour un estimateur X: $$MSE = \frac{1}{1000}\sum_{k=1}^{1000}\left(\hat{t}_{yX} - t_y\right)^2$$

Simulations (2/2)

On la divise par la MSE de l'estimateur d'Horvitz-Thompson correspondant, et on obtient ce que nous appelons le MSER (Mean Square Error Ratio) :

$$MSER = \frac{\left(\frac{1}{1000}\sum_{k=1}^{1000}\left(\hat{t}_{yX} - t_y\right)^2\right)}{\left(\frac{1}{1000}\sum_{k=1}^{1000}\left(\hat{t}_{yHT} - t_y\right)^2\right)}$$

MSER de l'estimateur X représente la proportion de MSE restante lorsque l'on utilise X plutôt que l'estimateur d'Horvitz-Thompson.

Le meilleur estimateur est celui avec le plus petit MSER.

Resultats - Distribution du MSER par activité

Quantile	Kokic and Bell	Robust V1	Robust V2
100 %	100 %	131 %	141 %
99 %	100 %	108 %	100 %
95 %	88 %	101 %	95 %
90 %	84 %	98 %	92 %
75 %	77 %	93 %	87 %
50 %	67 %	83 %	78 %
25 %	43 %	61 %	59 %
10 %	16 %	39 %	39 %
5 %	10 %	31 %	29 %
1 %	1 %	24 %	22 %
0 %	1 %	22 %	19 %

Insee

Comment redresser un échantillon d'UL tirées via leurs EP ?

La winsorisation en pratique pour 2016

On a appliqué une winsorisation avec les seuils de Kokic et Bell calculés comme si les UL étaient tirées selon un sondage aléatoire simple stratifié.

Résultats proches des résultats des années précédentes : environ 270 unités ont été winsorisées pour un montant de chiffre d'affaires « rogné » d'environ 30 millions de k€.

Adaptation du calage

Jusqu'à l'édition 2015 : calage sur le chiffre d'affaires fiscal et sur le nombre d'unités légales par activité (3 positions de la NACE).

Nouveau plan de sondage : ce calage reste a priori toujours possible mais les poids avant calage, dans un même domaine de calage, sont a priori plus dispersés qu'auparavant.

Test de la procédure de calage qui était utilisée jusqu'à l'édition 2015 sur 100 itérations de tirage d'échantillon selon le nouveau plan de sondage.

Le calage a convergé dans la plupart des secteurs lors de ces simulations.

Pour les résultats 2016, il y a eu un peu plus de secteurs où le calage n'a pas convergé "tout seul", sans poser pour autant de véritable problème en pratique...

Conclusion

Correction de la non-réponse totale : ajout de l'indicatrice d'appartenance à une EP comme variable mobilisable pour constituer les GRH ;

Traitement des valeurs influentes : On continue à utiliser la Winsorisation comme si les UL étaient tirées via un plan de sondage aléatoire simple stratifié ;

Calage : Même procédure qu'auparavant, pas de problème particulier de convergence.

=> Adaptations légères en vérifiant leur efficacité en amont via des simulations.

Bibliographie

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

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

[3] T. Deroyon “Traitement des valeurs atypiques d’une enquête par winsorization - application aux enquêtes sectorielles annuelles”. Acte des Journées de Méthodologie Statistique, 2015.

[4] C. Favre Martinoz, D. Haziza, J-F. Beaumont “A method of determining the winsorization threshold, with an application to domain estimation” Survey Methodology, vol. 41, n°1 (June): 57-77 , 2015.

[5] P.N. Kokic, P.A. Bell “Optimal winsorizing cut-offs for a stratified finite population estimator”, Journal of Official Statistics, vol. 10, n° 4: 419-435, 1994.

[6] C. Favre Martinoz, D. Haziza, J-F. Beaumont, 2016 “Robust Inference in Two-phase Sampling Designs with Application to Unit Non-response” Scandinavian journal of statistics vol. 43:1019-1034 ;

[7] P. Brion, E. Gros, 2015 “Statistical estimators using jointly administrative and survey data to produce french structural business statistics” Journal of Official Statistics, 31(4): 589–609.

Comment redresser un échantillon d'unités légales tirées via leurs entreprises ?

Merci de votre attention

Arnaud Fizzala
Arnaud.fizzala@insee.fr

Insee
DMCSI – Division Sondages

www.insee.fr

@InseeFr

Annexes

Resultats – Rapport entre le MSE des estimateurs robustes et le MSE de l'estimateur de Kokic et Bell par activité

Quantile	Robust V1 / Kokic et Bell	Robust V2 / Kokic et Bell
100 %	27,6	27,5
99 %	23,0	22,8
95 %	3,6	3,5
90 %	2,3	2,2
75 %	1,5	1,4
50 %	1,3	1,2
25 %	1,2	1,1
10 %	1,1	1,0
5 %	1,0	1,0
1 %	0,5	0,6
0 %	0,3	0,4

Insee

Comment redresser un échantillon d'UL tirées via leurs EP ?

Resultats - Distribution du MSER des estimateurs du total d'autres variables avec les poids winsorisés, par activité

Quantile	Turnover	Value added	Investments	Number of legal units
100 %	100 %	100 %	100 %	124 %
99 %	100 %	100 %	100 %	120 %
95 %	88 %	99 %	100 %	108 %
90 %	84 %	97 %	100 %	105 %
75 %	77 %	90 %	99 %	102 %
50 %	67 %	81 %	92 %	100 %
25 %	43 %	64 %	68 %	99 %
10 %	16 %	31 %	28 %	96 %
5 %	10 %	20 %	12 %	93 %
1 %	1 %	3 %	4 %	90 %
0 %	1 %	0 %	0 %	82 %

Insee

Comment redresser un échantillon d'UL tirées via leurs EP ?

Etre rattachée à une EP ne semble pas jouer “directement” sur la probabilité de réponse d’une UL

Le taux de réponse global des UL appartenant à une EP est un peu plus important que celui des UL indépendantes, mais l’effet disparait lorsque l’on compare par taille d’UL en effectif salarié. La variable n’est pas significative avec une régression logistique cherchant à prédire le fait de répondre lorsque les autres variables habituellement utilisées pour les GRH sont présentes.

Tranche d’effectif	Nombre d’UL	Taux de réponse non pondéré	
		UL indépendantes	UL non indépendantes
0 salarié	24 703	44,0%	43,8%
1 salarié	8 423	54,7%	54,6%
2 à 5 salariés	12 141	59,6%	64,6%
6 à 10 salariés	5 477	64,6%	67,9%
>10 salariés	7 052	70,8%	73,2%
Total	**57 796**	**53,5%**	**63,1%**

Les UL appartenant à une même EP semblent avoir des comportements de réponse corrélés

On trouve davantage de cas où toutes les UL répondent (ou aucune UL ne répond) qu'on ne le devrait en se basant sur les comportements de réponse des UL indépendantes…

Méthode :

- Régression logistique cherchant à prédire le fait de répondre en se limitant aux UL indépendantes ;

- Imputation d'une probabilité de répondre aux UL rattachées à une EP en se basant sur le modèle obtenu avec une base limitée aux UL indépendantes ;

- Comparaison des répartitions de taux de réponse par taille d'EP en nombre d'UL.

Les UL appartenant à une même EP semblent avoir des comportements de réponse corrélés

Observé

Nombre d'UL dans l'EP	Nombre d'UL répondantes					
	0	1	2	3	4	5
1	32%	68%	0%	0%	0%	0%
2	25%	26%	49%	0%	0%	0%
3	22%	14%	22%	42%	0%	0%
4	23%	9%	10%	11%	47%	0%
5	18%	13%	0%	15%	25%	30%

Simulations

Nombre d'UL dans l'EP	Nombre d'UL répondantes					
	0	1	2	3	4	5
1	34%	66%	0%	0%	0%	0%
2	14%	45%	41%	0%	0%	0%
3	7%	27%	41%	24%	0%	0%
4	4%	15%	31%	34%	16%	0%
5	1%	7%	19%	32%	29%	12%

Comment redresser un échantillon d'UL tirées via leurs EP ?

Insee

Les UL appartenant à une même EP semblent avoir des comportements de réponse corrélés

Résultat à confirmer et à étudier plus en détails car plusieurs autres éléments peuvent intervenir :

- La qualité du modèle permettant d'estimer la probabilité de réponse ;

- La « particularité » de l'échantillon d'EP dont on dispose ;

- La variabilité des répartitions selon le nombre d'UL répondantes (avec une approche de type quelle probabilité que la répartition que l'on observe soit « si éloignée » de la répartition théorique ?)

- Le fait d'appartenir à une EP modifie peut être le système de relance ?

=> Le problème ne concerne que peu d'UL. Continuer à utiliser les méthodes traditionnelles, ne devrait donc pas poser de gros problèmes.