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

Arnaud Fizzala

Insee, Direction de la méthodologie et de la coordination statistique et internationale

arnaud.fizzala@insee.fr

Mots-clés : Redressements, Non-réponse, calage, winsorisation, entreprises, profilage.

Résumé

Depuis le millésime 2016, le tirage des échantillons de l'enquête sectorielle annuelle (ESA) et de l'enquête annuelle de production (EAP), qui font partie du système d'élaboration des statistiques annuelles d'entreprises (Esane) est réalisé au niveau des entreprises profilées. Lorsqu'une entreprise profilée est tirée, toutes[1] les unités légales relevant du champ de l'enquête (en tant qu'unité légale) qui lui sont rattachées sont sélectionnées dans l'échantillon d'unités légales correspondant. On envoie alors un questionnaire aux unités légales de cet échantillon, et les réponses des entreprises profilées sont ensuite « reconstituées » à partir des retours de questionnaires des unités légales.

Les premiers résultats concernant l'année 2016 ont été diffusés début 2018, au niveau unités légales seulement, le règlement européen sur les statistiques structurelles annuelles n'exigeant l'élaboration de statistiques au niveau des entreprises qu'à partir de l'année 2017. À partir du millésime 2017, les résultats d'ESANE seront élaborés au niveau des entreprises afin de répondre à ce règlement européen, mais aussi au niveau des unités légales car le règlement européen encadrant les comptes nationaux n'est lui pas passé au concept d'entreprises profilées.

Cette perspective de devoir continuer, pour au moins quelques années, à produire des résultats au niveau des unités légales malgré le passage du plan de sondage à un niveau entreprises a conduit à adapter différents aspects du processus d'élaboration des résultats d'ESANE.

L'article présente l'adaptation au nouveau plan de sondage des traitements post-collecte (traitement de la non-réponse, traitement des valeurs influentes, calage) d'ESANE en se limitant[2] à l'élaboration des résultats au niveau unités légales.

[1] En pratique, toutes les unités légales ne sont pas forcément interrogées et on procède alors par imputation pour celles contribuant le moins au chiffre d'affaires de l'entreprise profilée.

[2] L'adaptation des traitements post-collecte pour l'élaboration des résultats au niveau des entreprises est encore en cours d'étude et fait l'objet d'une autre proposition d'article aux JMS 2018.

C'est notamment le traitement des valeurs influentes qui a du être adapté car la méthode utilisée jusqu'ici - winsorisation avec des seuils constitués selon la méthode de Kokic et Bell - a été conçue pour s'appliquer dans le strict cadre d'échantillons aléatoires simples stratifiés[3]. En effet, d'un « point de vue unités légales », le tirage aléatoire simple stratifié d'entreprises profilées correspondant au nouveau plan de sondage s'apparente à un tirage d'unités légales en grappes, les grappes correspondant aux entreprises profilées.

Plusieurs stratégies ont été envisagées afin d'adapter le traitement des valeurs influentes au nouveau plan de sondage en entreprises :
- Appliquer la winsorisation avec les seuils de Kokic et Bell « en faisant comme si » l'échantillon d'unités légales était tiré selon un plan de sondage aléatoire simple stratifié d'unités légales ;
- Appliquer des méthodes alternatives basées sur le biais conditionnel et s'adaptant facilement au nouveau plan de sondage.

Une comparaison de ces stratégies a pu être réalisée en amont de la production des premiers résultats 2016 en s'appuyant sur des variables fiscales disponibles pour l'ensemble des unités légales[4], et sur des simulations de tirage d'échantillons d'entreprises selon le nouveau plan de sondage de l'ESA et de l'EAP. Cette comparaison, dont la présentation détaillée constituera l'essentiel de l'article, a mené à retenir la winsorisation avec les seuils de Kokic et Bell « en faisant comme si » l'échantillon d'unités légales était tiré selon un plan de sondage aléatoire simple stratifié d'unités légales. Et c'est effectivement la méthode qui a été appliquée pour l'élaboration des premiers résultats 2016.

Les autres traitements post-collecte nécessitent moins d'ajustements pour s'adapter au nouveau plan de sondage et seront donc évoqués de façon plus succincte dans l'article. Quelques tests de convergence du calage ont tout de même été réalisés et laissent penser que le calage devrait pouvoir continuer à être pratiqué en utilisant les mêmes variables et les mêmes marges que « sous » l'ancien plan de sondage[5]. Le traitement de la non-réponse ne pouvait pas être étudié en amont de la réception des premières données et n'a été modifié qu'à la marge pour l'élaboration des premiers résultats : l'indicatrice d'appartenance à une entreprise profilée a été ajoutée comme variable mobilisable pour constituer les groupes de réponses homogènes.

Abstract

The sample design of the French Structural Business Statistics survey has changed for the 2016 edition. Now we no longer sample « legal units » but « enterprises ». Data is still collected on legal units, with the following rule : when we select an enterprise, then all legal units within this enterprise will be surveyed. This paper develops the adaptation of non-response treatments, influential values treatments and calibration to this new context. This is mainly the influential values treatments which had to be adapted because the method used until now (winsorization) has been built in a stratified random sampling framework. The main matter of the paper is so on adaptation of winsorisation.

We show that winsorization with the Kokic and Bell thresholds, applied as if the sampling were a stratified sampling of legal units, seems to be the best option to deal with influential values. We also test alternative methods based on conditional bias, but it leads, in our context, to poorer results with some problems to solve to be operational.

[3] Cadre qui est en effet rencontré de façon systématique ou presque dans les enquêtes alimentant la statistique d'entreprises française.
[4] Il n'était donc pas nécessaire d'attendre les résultats de l'enquête pour commencer à étudier les méthodes.
[5] En pratique, le calage s'est effectivement déroulé sans difficulté particulière pour l'élaboration des premiers résultats au niveau unités légales portant sur 2016.

1. Introduction

Depuis le millésime 2016, le tirage des échantillons de l'enquête sectorielle annuelle (ESA) et de l'enquête annuelle de production (EAP), qui font partie du système d'élaboration des statistiques annuelles d'entreprises (Esane) [1] est réalisé au niveau des entreprises profilées [2]. Lorsqu'une entreprise profilée est tirée, toutes[6] les unités légales relevant du champ de l'enquête (en tant qu'unité légale) qui lui sont rattachées sont sélectionnées dans l'échantillon d'unités légales correspondant. On envoie alors un questionnaire aux unités légales de cet échantillon, et les réponses des entreprises profilées sont ensuite « reconstituées » à partir des retours de questionnaires des unités légales.

Les premiers résultats concernant l'année 2016 ont été diffusés début 2018, au niveau unités légales seulement, le règlement européen sur les statistiques structurelles annuelles n'exigeant l'élaboration de statistiques au niveau des entreprises qu'à partir de l'année 2017. À partir du millésime 2017, les résultats d'ESANE seront élaborés au niveau des entreprises afin de répondre à ce règlement européen, mais aussi au niveau des unités légales car le règlement européen encadrant les comptes nationaux n'est lui pas passé au concept d'entreprises profilées.

Cette perspective de devoir continuer, pour au moins quelques années, à produire des résultats au niveau des unités légales malgré le passage du plan de sondage au niveau entreprises a conduit à adapter différents aspects du processus d'élaboration des résultats d'ESANE.

L'article présente l'adaptation au nouveau plan de sondage des traitements post-collecte d'ESANE en se limitant à l'élaboration des résultats au niveau unités légales. Il aborde brièvement le traitement de la non-réponse totale pour lequel peu d'adaptations ont été réalisées, puis s'attarde plus longuement sur l'adaptation du traitement des valeurs influentes qui constitue le « coeur » de l'article, et se termine par des tests de convergence du calage en appliquant les mêmes marges et bornes de rapport de poids qu'avec l'ancien plan de sondage.

2. Adaptation du traitement de la non-réponse

Lorsqu'une entreprise profilée est tirée, on envoie un questionnaire aux unités légales (UL) qui lui sont rattachées. Pour la diffusion en UL, on peut craindre que le fait d'appartenir à une EP influence la probabilité de réponse de l'UL qui reçoit le questionnaire. Ce possible comportement de réponse au niveau de l'EP existait peut-être déjà précédemment, mais n'était pas pris en compte jusqu'ici. Le changement du plan de sondage, qui fait qu'à présent, et par construction, nous disposons de beaucoup d'UL appartenant aux mêmes EP, fait que si ce comportement existe, il devient plus important de le prendre en compte. Cette partie de l'article tente de défricher deux sujets liés au traitement de la non-reponse des UL à l'ESA ou l'EAP pour la diffusion en UL :
- l'influence du fait d'être rattachée à une entreprise profilée sur la probabilité de répondre d'une unité légale ;
- l'indépendance des comportements de réponse des unités légales rattachées à une même entreprise profilée.

[6] En pratique, toutes les unités légales ne sont pas forcément interrogées et on procède alors par imputation pour les celles contribuant le moins au chiffre d'affaires de l'entreprise profilée.

2.1. Données

Pour cette partie de l'étude, on se base sur l'échantillon d'UL utilisé pour les traitements post-collecte réalisés fin 2017 par la division Sondages pour l'élaboration des résultats semi-définitifs d'Esane 2016. L'analyse porte donc sur la partie non exhaustive de l'échantillon uniquement et les unités légales hors-champ ont été exclues. Le fichier comporte 57 796 unités légales dont 4 855 sont rattachées à une EP.

2.2. Le fait d'appartenir à une EP influence-t-il la probabilité de réponse d'une UL ?

Le taux de réponse non pondéré des UL indépendantes est de 54 % contre 63 % pour les UL rattachées à une EP (Tableau 1). Cependant cet écart ne semble pas vraiment dû au fait d'être rattaché à une EP ou non, mais plutôt à d'autres critères comme par exemple la taille des UL. En effet, lorsque l'on compare les taux de réponse par tranche d'effectif, ces derniers deviennent très proches :

Tableau 1 : taux de réponse non pondérés par tranche d'effectif des UL

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%**

De façon plus rigoureuse, on peut vérifier, à partir d'une régression logistique expliquant le fait de répondre que le coefficient associé au fait d'être rattaché à une EP n'est pas significatif, dès lors que les principales variables explicatives (secteur d'activité, tranche d'effectif et réception d'une liasse fiscale) sont introduites dans la régression[7] :

[7]Le modèle de régression logistique présenté dans ce tableau est estimé sous l'hypothèse que les comportements de réponse entre unités légales sont indépendants, y compris quand elles appartiennent à la même entreprise. Cette hypothèse est forte et les éléments présentés dans la deuxième partie de la note tendent à la rejeter.

Tableau 2 : Régression logistique expliquant le fait de répondre

Paramètre		Coefficient	p-value
Constante		0,7714	<,0001
secteur	AZ	0,1083	0,3331
secteur	BE	0,0343	0,2621
secteur	FZ	-0,1478	<,0001
secteur	GI	-0,0426	0,0575
secteur	JZ	-0,0957	0,007
secteur	LZ	-0,1744	<,0001
secteur	MN	0,0277	0,2825
Indicatrice liasse fiscale imputée		-1,2065	<,0001
Tranche effectif	0 salarié	-0,4385	<,0001
Tranche effectif	1 salarié	-0,1953	<,0001
Tranche effectif	2 à 5 salariés	-0,00048	0,9788
Tranche effectif	6 à 10 salariés	0,1931	<,0001
Indicatrice d'appartenance à une EP		**0,00425**	**0,8989**

2.3. Les Comportements de réponse des UL au sein d'une EP sont-ils corrélés ?

Dans cette partie, on essaie de savoir si les UL d'une même EP ont tendance à adopter le même comportement de réponse ou non. La démarche adoptée dans la suite s'appuie sur la répartition de la proportion d'UL répondantes par taille d'EP (en nombre d'UL) : s'il y a un comportement de réponse au niveau de l'EP, les proportions d'UL répondantes devraient se concentrer aux extrémités (toutes les UL répondent ou aucune UL ne répond).

Une première difficulté rencontrée est que les EP de la partie non exhaustive sont souvent des « petites » EP au sens où elles ne comportent qu'une ou deux UL dans plus de 80 % des cas (tableau 3).

Tableau 3 : répartition des EP par taille (en nombre d'UL)

Taille de l'EP en nombre d'UL	nombre d'EP	%
1	1043	42
2	1022	41
3	273	11
4	79	3
5	40	2
6	24	1
7	7	0
8	7	0
9	5	0
10	3	0
11	1	0
12	3	0
13	1	0
16	2	0
17	1	0
Total	**2511**	**100**

Dans la suite de l'analyse, on se limite aux EP comportant 5 UL ou moins. Les autres cas étant trop peu fréquents dans nos données pour pouvoir être étudiés. Les proportions d'UL répondantes sont alors les suivantes :

Tableau 4 : Nombre d'UL répondantes par taille d'EP (en nombre d'UL)

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%

La deuxième difficulté consiste à savoir à quoi comparer la distribution des proportions de réponses obtenues. En effet, la répartition de la proportion d'UL répondantes dépend des probabilités de réponse des UL elles-mêmes : par exemple si celles-ci sont élevées, on devrait avoir plus d'EP avec une proportion « importante » d'UL répondantes… Différencier l'effet venant des caractéristiques des UL et l'effet venant d'un éventuel comportement de réponse au niveau de l'EP semble difficile.

Nous avons procédé de la façon suivante afin d'effectuer cette comparaison :

a) Élaboration d'un modèle logistique[8] expliquant le fait de répondre à partir des données limitées aux UL indépendantes ;

b) Application de ce modèle sur les UL rattachées à une EP pour déterminer une probabilité de répondre « si l'unité était indépendante » ;

c) Simulations (10 000 tirages de Poisson selon les probabilités de réponse affectées à l'étape b) pour obtenir la répartition moyenne des proportions d'UL répondantes « si les UL étaient indépendantes ».

Tableau 5 : Nombre d'UL répondantes par taille d'EP, si les UL étaient indépendantes (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%

La répartition obtenue par simulations (tableau 5) est assez différente de la répartition observée (tableau 4), avec notamment des valeurs plus faibles au niveau des répartitions extrêmes, c'est-à-dire correspondant à « aucune UL ne répond » ou « toutes les UL répondent ». Cela va donc dans le sens d'un comportement de réponse au niveau EP, bien que plusieurs autres éléments peuvent intervenir, notamment :

- 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 ?

2.4. Conclusion sur l'adaptation du traitement de non-réponse

D'après les premiers résultats obtenus sur Esane 2016, le fait d'appartenir à une EP ne semble pas en tant que tel, jouer sur la probabilité de réponse d'une UL. En effet, à caractéristiques communes (secteur d'activité, taille, réception ou non d'une liasse fiscale) les taux de réponses des UL indépendantes sont équivalents aux taux de réponse des autres UL.

En revanche, il semble y avoir un lien entre les comportements de réponse des UL au sein d'une EP. D'après notre analyse, assez frustre, on observe davantage d'EP avec des comportements « extrêmes », c'est-à-dire avec aucune UL répondante ou à l'inverse toutes ses UL répondantes, que ce qu'on devrait observer si les UL étaient indépendantes. Cette observation est en contradiction

[8] Les variables explicatives sont celles du tableau 2, sauf l'indicatrice d'appartenance à une EP.

avec l'hypothèse de comportements de réponse indépendants des UL formulée lors du traitement de la non-réponse, mais elle semble concerner pour le moment un nombre assez restreint d'UL. Continuer à utiliser les méthodes traditionnelles, ne devrait donc pas poser de gros problèmes, mais il serait intéressant d'avoir des retours coté terrain sur ce comportement de réponse niveau EP ainsi que de se documenter du côté méthodologie sur des méthodes pouvant s'appliquer lorsque les comportements de réponse des unités sont corrélés. L'étude a été réalisée sur la partie non-exhaustive de l'échantillon, car c'est sur cette partie uniquement que la non-réponse totale est traitée par repondération. Les résultats de l'étude auraient probablement été différents sur la partie exhaustive. Cependant le fait qu'il y ait des comportements de réponse au niveau EP ne devrait pas avoir d'impact visible sur la partie exhaustive car la non-réponse totale y est principalement traitée en imputant par la réponse de l'unité légale une année précédente.

Pour l'élaboration des résultats semi-définitifs 2016, le traitement de la non-réponse s'est déroulé via des procédures « automatiques »[9], proches de ce qui était fait les années précédentes, avec cependant le fait d'appartenir à une entreprise profilée comme variable supplémentaire et potentiellement mobilisable pour constituer les groupes de réponse homogène.

3. Adaptation du traitement des valeurs influentes

Nous nous intéressons dans cette partie à l'adaptation du traitement des valeurs influentes au nouveau plan de sondage des ESA pour la diffusion en unités légales. 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. Il s'agit typiquement d'une « strata jumper », c'est-à-dire d'une unité que l'on pensait être « petite » au moment du tirage, qui a été tirée avec un taux de sondage faible impliquant un poids élevé, mais qui s'avère être une « grande » unité au moment de la collecte des données. 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.
La méthode utilisée jusqu'ici - winsorisation avec des seuils constitués selon la méthode de Kokic et Bell [3] - a été conçue pour s'appliquer dans le strict cadre d'échantillons aléatoires simples stratifiés[10]. En effet, d'un « point de vue unités légales », le tirage aléatoire simple stratifié d'entreprises profilées correspondant au nouveau plan de sondage s'apparente à un tirage d'unités légales en grappes, les grappes correspondant aux entreprises profilées.

Plusieurs stratégies ont été envisagées afin d'adapter le traitement des valeurs influentes au nouveau plan de sondage en entreprises :
- Appliquer la winsorisation avec les seuils de Kokic et Bell « en faisant comme si » l'échantillon d'unités légales était tiré selon un plan de sondage aléatoire simple stratifié d'unités légales ;
- Appliquer des méthodes alternatives se basant sur le biais conditionnel [4] et s'adaptant facilement au nouveau plan de sondage.

Une comparaison de ces stratégies a pu être réalisée en amont de la production des premiers résultats 2016 en s'appuyant sur des variables fiscales disponibles pour l'ensemble des unités légales[11], et sur des simulations de tirage d'échantillons d'entreprises selon le nouveau plan de sondage de l'ESA et de l'EAP.

3.1. Données utilisées

[9] Via la macro sas treedisc mettant en œuvre l'algorithme CHAID.
[10] Cadre qui est en effet rencontré de façon systématique ou presque dans les enquêtes alimentant la statistique d'entreprises française.
[11] Il n'était donc pas nécessaire d'attendre les résultats de l'enquête pour commencer à étudier les méthodes.

Les données utilisées pour cette partie de l'étude se basent sur la base de sondage (BdS) qui avait été constituée par la division sondage début 2016 sur des données 2015 pour tester le nouveau plan de sondage des enquêtes Esane, restreinte à la partie non exhaustive. Notre population d'intérêt correspond aux UL dans le champ des enquêtes Esane (dit sous-champ 1) au lancement et rattachées à une EP du sous-champ 1 au lancement, encore actives ou présumées actives.

Notre base d'UL correspondant à notre population d'intérêt comporte[12] au final 2 276 634 UL.

3.2. Le traitement des valeurs influentes par winsorisation

Depuis l'édition 2008 d'Esane, les valeurs influentes sont traitées par winsorisation. Cette méthode se base sur la détermination d'un seuil dans chaque strate de tirage au-delà duquel les valeurs sont réduites. Plus précisément, les valeurs Y^w d'une variable d'intérêt Y positive après winsorisation sont définies par :

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

Avec :

- K_h : le seuil de la strate h ;
- n_h : le nombre d'unités sélectionnées dans l'échantillon dans la strate h ;
- N_h : le nombre d'unités dans la base de sondage et dans la strate h.

L'estimateur winsorisé correspond alors à l'estimateur d'Horvitz-Thompson de la variable winsorisée :

$$\hat{t_{y^w}} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_i^w$$

Contrairement à l'estimateur d'Horvitz-Thompson qui est sans biais, l'estimateur winsorisé sous-estime systématiquement le total t_Y de la variable d'interêt Y. En revanche, sa variance dépend de la dispersion de la variable Y^w dans chaque strate, qui est par construction plus petite que celle de la variable originale Y. La winsorisation correspond donc à un compromis entre biais et variance : elle sera efficace si le biais introduit est plus que compensé par le gain de variance obtenu, ou autrement dit, si l'erreur quadratique moyenne (EQM) de l'estimateur winsorisé est inférieure à celle de l'estimateur d'Horvitz-Thompson.

Bien entendu, le choix des seuils K_h est crucial pour la qualité de la winsorisation. Le choix des seuils a été étudié par Kokic et Bell en 1994 [5], qui ont proposé une méthode permettant, sous certaines hypothèses, d'optimiser les seuils, c'est-à-dire de déterminer les seuils qui minimisent l'erreur quadratique moyenne de l'estimateur winsorisé.

[12]En pratique, nous les identifions comme l'intersection entre les 2 504 391 UL sus-mentionnées et les UL qui appartiennent au fichier dit « Calculpoids », fourni pour calculer les poids finaux d'ESANE 2015, et qui est restreint aux unités légales qui, en fin de campagne, sont considérées comme ayant une activité économique réelle.

Jusqu'à l'édition 2015, les valeurs influentes étaient traitées par winsorisation de la variable R310 (chiffre d'affaires des liasses fiscales, disponible pour toutes les UL du champ) pour l'estimateur du R310 total au niveau du secteur d'activité de l'UL sur 3 positions[13] [3]. Les seuils de winsorisation étaient déterminés par la méthode de Kokic et Bell, permettant donc de déterminer les seuils qui minimisent l'EQM pour l'estimation du R310 total. Le poids des unités winsorisées était ensuite modifié selon la formule $w_i^w = w_i \frac{y_i^w}{y_i}$ afin que l'effet de la winsorisation sur le R310 soit « transmis » aux autres variables[14] que le R310.

Le cadre théorique de la méthode de Kokic et Bell exige que le plan de sondage soit stratifié avec tirage aléatoire simple dans chaque strate. Le passage à un tirage au niveau EP, fait que l'on sort un peu de ce cadre théorique pour le calcul d'estimateurs au niveau UL et il n'est donc pas garanti que l'application des seuils de Kokic et Bell, comme si le tirage se faisait au niveau UL, continue à donner de bons résultats.

Une première façon d'appréhender l'impact du nouveau plan de sondage sur l'efficacité de la méthode actuelle de winsorisation est de regarder à quel point les poids des UL d'une même APE et tranche d'effectif (strates de tirage que l'on considère pour calculer les seuils de winsorisation selon la méthode de Kokic et Bell) varient.

Pour cela, nous avons calculé des statistiques dans la base de sondage par strate (Coefficient de variation des poids, écart entre le poids maximal et le poids minimal (étendue), proportion d'UL dont le poids correspond au poids modal de la strate) en considérant[15] que le poids était l'inverse du taux de sondage appliqué à l'EP à laquelle est rattachée l'UL. Le tableau 6 correspond à la distribution de ces statistiques par strate[16] (une observation=une strate).

Tableau 6 : Indicateurs de distribution des poids par strate UL

Quantile	CV	Étendue	% mode
100%	401,7	797	100%
99%	118,1	265	100%
95%	59,5	153	100%
90%	38,7	96	100%[17]
75%	20,7	50	99%
50%	11,9	20	97%
25%	6,1	9	91%
10%	2,0	1	84%
5%	0,0	0	80%
1%	0,0	0	68%
0%	0,0	0	47%

[13]Correspondant au groupe de la nomenclature d'activité économique, qui ne doit pas être confondu avec la notion de groupe d'entreprises

[14]L'estimateur winsorisé pour une variable Y correspond alors à l'estimateur par expansion de la variable Y mais utilisant les poids modifiés w_i^w.

[15]On ne donne « habituellement » cette valeur qu'aux seules unités sélectionnées dans l'échantillon.

[16]Afin que les statistiques soient relativement « solides » nous nous sommes limités aux strates comportant au moins 30 UL. Cela représente 1 632 strates.

[17]La valeur présentée ici correspond à l'arrondi de 99,88 %. Tous les groupes ayant 100 % des UL avec le même poids ont un CV et une étendue nuls.

On voit que le plus souvent, peu d'unités ont un poids différent des autres dans une strate donnée, mais que l'écart de poids peut être important. Du fait de la concentration importante des poids au mode de la strate, on s'attend à ce que la méthode de winsorisation actuelle donne plutôt de bons résultats avec le nouveau plan de sondage. Néanmoins, comme le calcul des seuils se fait en considérant que toutes les unités d'une même strate d'UL ont le même poids de sondage, égal au poids moyen de la strate, la méthode ne peut tenir compte des écarts de poids de sondage à l'intérieur des strates d'unités légales et risque de ce fait de rater certaines unités influentes.

Pour vérifier cela, nous avons réalisé une étude basée sur 1 000 simulations de tirages au niveau EP, en comparant l'estimateur winsorisé à partir de seuils de winsorisation calculés selon la méthode de Kokic et Bell en faisant comme si les UL avaient été tirées directement dans des strates d'UL[18] à l'estimateur « classique » d'Horvitz-Thompson d'une part mais aussi à des estimateurs dits « robustes » traitant les valeurs influentes selon d'autres méthodes basées sur le concept de biais conditionnel[19].

Avant de présenter les simulations plus en détails, nous présentons quelques éléments sur les estimateurs robustes.

3.3. Les estimateurs robustes

Le cadre de l'estimation robuste est décrit dans les articles [4] et [6] cités en bibliographie. Nous ne rappelons ici que les éléments indispensables à la compréhension de l'étude.

L'estimateur robuste pour une variable y s'écrit :

$$\hat{t_{yR}} = \hat{t_y} - \frac{1}{2}(B_{min} + B_{max})$$

où $\hat{t_y}$ correspond à l'estimateur classique d'Horvitz-Thompson du total de la variable y et B_{min} et B_{max} correspondent[20] aux biais conditionnels maximaux et minimaux des unités appartenant à l'ensemble des répondants. Cet estimateur est constitué de manière à minimiser le biais conditionnel de l'unité ayant le biais conditionnel maximal.

Le biais conditionnel est une mesure de l'influence d'une observation. Il correspond, pour un plan de sondage donné, au biais observé en moyenne sur l'ensemble des échantillons comprenant l'observation.

$$B_i = E_p(\hat{t_y} / I_i = 1) - t_y$$

B_{min} et B_{max} dépendent donc du plan de sondage qui a été appliqué pour obtenir l'échantillon et de la façon dont on modélise le comportement de réponse des unités. L'estimateur robuste est donc construit de manière à minimiser l'influence de l'unité la plus influente.

Pour notre étude, nous avons assimilé l'ensemble du processus aboutissant au fichier de répondants (tirage d'un échantillon d'UL et réponse d'une partie de ces dernières) à un tirage en deux phases : une première phase correspondant au tirage des UL à qui on envoie un questionnaire et une deuxième phase correspondant au « tirage » des UL qui répondront.

[18]Cela revient en particulier à considérer que toutes les UL d'une même APE et tranche d'effectif ont le même poids, ce qui n'est pas forcément vrai.
[19]Le biais conditionnel est une mesure de l'influence d'une observation. Il correspond, pour un plan de sondage donné, au biais observé en moyenne sur l'ensemble des échantillons comprenant l'observation.
[20]Bmin et Bmax dépendent de la variable y bien que ce ne soit pas explicite dans les notations.

La deuxième phase a été considérée comme un tirage poissonien d'UL. Cela revient à supposer que le fait qu'une unité légale réponde n'influence pas la probabilité qu'une autre unité légale de l'échantillon réponde. Il s'agit là d'une modélisation classique[21] du comportement de réponse dans les études de plan de sondage.

La première phase, qui correspond au tirage des unités légales, a été considérée selon deux versions, aboutissant à deux estimateurs différents :
- 1 :Tirage poissonien des EP ;
- 2 : Tirage stratifié des EP (considérées comme des grappes d'UL) ;

C'est la version 2 qui correspond au « vrai » plan de sondage, mais la version 1 présente des avantages pratiques (voir annexe). Aussi, même si on s'attend à ce que les résultats issus de cette version 1 soient moins bons en termes de précision des estimateurs obtenus que ceux issus de la version 2, il est intéressant de l'étudier car elle est très simple à mettre en production.

Dans le cas d'un plan de sondage en deux phases, avec une deuxième phase poissonienne, le biais conditionnel d'une unité i s'écrit[22] (formule 6 de l'article [6]) :

$$B_i = \sum_{j \in U} \left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}} - 1 \right) y_j + \frac{1}{\pi_{1i}} \left(\frac{1}{\pi_{2i}} - 1 \right) y_i$$

Pour évaluer dans la version 1 le biais conditionnel, nous distinguons 3 sous-cas (en notant E l'EP contenant l'UL i) pour évaluer les termes de la somme :

a) $j = i$ alors $\pi_{1ij} = \pi_{1i} = \pi_{1E} = \frac{m_h}{M_h}$;

b) $j \neq i \text{ et } j \in E$ alors $\pi_{1ij} = \pi_{1i} = \pi_{1E} = \frac{m_h}{M_h}$;

c) $j \neq i \text{ et } j \notin E$ alors $\pi_{1ij} = \pi_{1i}\pi_{1j}$ et le terme de la somme est nul

Nous pouvons en déduire :

$$B_i^1 = \left(\frac{M_h}{(m_h \cdot r_i)} - 1 \right) y_i + \left(\frac{M_h}{m_h} - 1 \right) (y_E - y_i)$$

Remarque : dans le cas où E recouvre des UL appartenant à différents secteurs d'activité à 3 positions, y_E représente la somme des y <u>des UL de même secteur d'activité à 3 positions que l'UL i</u>.

Dans la version 2, il faut distinguer 4 sous-cas pour évaluer les termes de la somme :

a) $j = i$ alors ; $\pi_{1ij} = \pi_{1i} = \pi_{1E} = \frac{m_h}{M_h}$

[21]Notons tout de même que l'on suppose ici que les comportements de réponse des unités légales d'une même EP ne sont pas liés entre eux, ce qui semble inexact d'après la partie 2 de cet article. Nous négligeons cependant cette dépendance dans la mise en œuvre de la méthode de robustification.

[22]Les probabilités indicées par 1 correspondent aux probabilités d'inclusion simples et doubles définies par la première phase du plan de sondage et celles indicées par 2 aux probabilités d'inclusion définies par la deuxième phase.

b) $j \neq i \; et \; j \in E$ alors ; $\pi_{1ij} = \pi_{1i} = \pi_{1E} = \frac{m_h}{M_h}$

c) $j \neq i \; et \; j \notin E_i et \; E_j \in h$ alors $\pi_{1ij} = \frac{m_h}{M_h} \frac{(m_h - 1)}{(M_h - 1)}$

c) $j \neq i \; et \; j \notin E_i et \; E_j \notin h$ alors $\pi_{1ij} = \pi_{1i} \pi_{1j}$ et le terme de la somme est nul

Nous pouvons en déduire :

$$B_i^2 = (\frac{M_h}{(m_h \cdot r_i)} - 1) y_i + (\frac{M_h}{m_h} - 1)(y_E - y_i) + (\frac{M_h}{m_h} \frac{(m_h - 1)}{(M_h - 1)} - 1)(t_{yh} - y_E)$$

Remarquons que : $B_i^2 = B_i^1 + (\frac{M_h}{m_h} \frac{(m_h - 1)}{(M_h - 1)} - 1)(t_{yh} - y_E)$

Dans la version 2, on tient compte du niveau du R310 de E par rapport à sa strate d'appartenance, car (contrairement à un tirage poissonien) tirer E va réduire les chances de tirer d'autres EP dans la strate.

En pratique, pour calculer les estimateurs robustes correspondant aux versions 1 et 2 sus-mentionnées, on procède en 3 temps :

1 – Calcul de B_i pour chaque UL de la BdS.
2 – Recherche parmi les UL répondantes de B_{min} et B_{max} par secteur d'activité (nous considérons que chaque UL ne peut être influente que pour l'estimateur de son secteur d'activité).
3 – Calcul de l'estimateur d'Horvitz-Thompson du total de la variable y.
4 – Calcul de l'estimateur robuste du total de la variable y.

3.4. Simulations

Pour évaluer la qualité des différents estimateurs, nous avons réalisé 1 000 tirages d'échantillons selon le nouveau plan de sondage. Puis nous avons simulé de la non-réponse à partir des taux de réponse observés dans les données Esane 2015. Concrètement, nous avons appliqué, par groupe de réponse homogène utilisé pour la correction de la non-réponse totale Esane 2015, un tirage de Poisson avec pour probabilité d'inclusion le taux de réponse observé dans ce GRH dans les données Esane 2015. Nous avons ensuite corrigé la non-réponse par repondération dans ces GRH. Enfin, nous avons calculé pour chacun des 207 secteurs d'activité présents dans notre base d'UL les différentes versions des estimateurs ainsi que le « vrai total » de R310 (puisqu'il s'agit d'une variable fiscale, disponible pour l'ensemble des unités du champ). Cela permet d'estimer l'EQM pour un estimateur X du total de R310 d'un secteur d'activité :

$$EQM = \frac{1}{1000} \sum_{k=1}^{1000} (\hat{t_{yX}} - t_y)^2$$

Afin de rendre davantage lisible les valeurs d'EQM obtenues, on étalonne par l'EQM de l'estimateur d'Horvitz-Thompson, qui sert ainsi de référence, et on analyse en fait le rapport de l'EQM de l'estimateur X et de l'EQM de l'estimateur d'Horvitz-Thompson :

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

Les résultats présentés ci-dessous correspondent à la distribution des REQM obtenus (une observation=un secteur d'activité).

Tableau 7 : Distribution des EQM des différents estimateurs du R310 total rapportés aux EQM de l'estimateur d'Horvitz-Thompson

Quantile	Kokic et Bell	Robuste V1	Robuste 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 %

On voit que l'estimateur winsorisé avec des seuils déterminés selon la méthode de Kokic et Bell est celui aboutissant aux meilleurs résultats en termes d'EQM de l'estimation du R310 total, même si les hypothèses d'application de ne sont pas tout à fait respectées. Les estimateurs robustes, notamment le V2, donnent tout de même de très bons résultats : meilleurs que l'estimateur d'Horvitz-Thompson dans plus de 95 % des cas.

Ceci est sans doute lié au fait que, contrairement à la méthode de Kokic et Bell dont l'objectif est de minimiser l'erreur quadratique moyenne de l'estimateur winsorisé, les estimateurs robustes définis grâce au biais conditionnel sont construits de manière à obtenir l'estimateur dont le biais conditionnel maximal est minimal, i.e. pour lesquels la valeur la plus influente a l'influence la plus faible. Ce faisant, ils n'ont pas forcément l'EQM la plus faible, mais ils auraient pu tout de même obtenir ici une EQM plus faible que les estimateurs winsorisés, les hypothèses sous lesquelles ces derniers sont efficaces n' étant pas tout à fait respectées

Tableau 8 : Distribution des Biais relatifs des différents estimateurs du R310 total

Quantile	Kokic et Bell	Robuste V1	Robuste V2	Horvitz-Thompson
100 %	0 %	0 %	201 %[23]	15 %
99 %	0 %	-1 %	0 %	5 %
95 %	-1 %	-1 %	-1 %	2 %
90 %	-1 %	-2 %	-1 %	2 %
75 %	-3 %	-3 %	-3 %	1 %
50 %	-5 %	-6 %	-4 %	0 %
25 %	-7 %	-12 %	-8 %	0 %
10 %	-13 %	-23 %	-14 %	-1 %
5 %	-24 %	-30 %	-21 %	-1 %
1 %	-52 %	-61 %	-33 %	-5 %
0 %	-69 %	-77 %	-48 %	-7 %

Pour essayer de comprendre si l'estimateur « Kokic et Bell » est meilleur que les estimateurs robustes de façon systématique ou non, on peut simplement faire le rapport entre les EQM de l'estimateur Kokic et Bell et les EQM des estimateurs robustes.

Tableau 9 : Distribution du rapport entre les EQM des différents estimateurs robustes du R310 total et l'EQM de l'estimateur winsorisé avec les seuils « Kokic et Bell »

Quantile	Robuste V1 / Kokic et Bell	Robuste 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

On voit ainsi que l'estimateur winsorisé selon la méthode de Kokic et Bell est « meilleur » que l'estimateur Robuste (V1 ou V2) dans plus de 90 % des secteurs.

[23]Le biais relatif de 201 % est observé pour le groupe 303, il contient une unité avec un R310 de 34 067 kú représentant plus de la moitié du R310 total du groupe. Nous pensons que la valeur atypique du biais provient de la présence de cette unité.

L'estimateur winsorisé selon la méthode de Kokic et Bell apparaît comme la meilleure option au vu de ces résultats et de problèmes de mise en pratique des estimateurs robustes (voir annexe)[24]. À noter de plus qu'avec la méthode de Kokic et Bell[25], 216 UL sont winsorisées en moyenne (variant en fonction de la réplication d'échantillon entre 172 et 265), ce qui correspond à l'ordre de grandeur du nombre d'UL winsorisées actuellement. Le montant de R310 « rogné » est de 21,9 millions de k€ en moyenne (variant en fonction de la réplication d'échantillon entre 5 et 238[26] millions de k€), ce qui est légèrement inférieur au montant « rogné » pour les estimations définitives 2015 (34 millions de k€).

Afin de mesurer l'impact sur d'autres variables que le R310 des poids winsorisés selon la méthode de Kokic et Bell, nous avons calculé l'EQM pour des estimateurs de totaux d'autres variables en utilisant les poids qui ont été winsorisés sur la variable R310 :
- La valeur ajoutée ;
- L'investissement ;
- Le nombre d'UL (total de la variable constante égale à 1).

Tableau 10 : Impact des poids winsorisés sur l'EQM d'estimateurs d'autres variables

Quantile	R310	R003	INV	NB_UL
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 %

On voit que même sur l'investissement, qui est une variable a priori peu corrélée au chiffre d'affaires, le traitement des valeurs influentes améliore l'EQM. Sur le nombre d'unités légales, le traitement des valeurs influentes semble avoir un effet assez neutre : amélioration dans la moitié des cas et dégradation pour l'autre moitié des cas, de l'ordre de 20 % au maximum.

[24]Dans la suite de la note, on n'a pas calculé les indicateurs selon les estimateurs robustes à cause de problèmes pratiques pour intégrer l'effet du traitement dans les poids.
[25]Les problèmes pratiques de calcul de poids empêchent de fournir cette statistique avec les autres méthodes.
[26]95 % des réplications aboutissent à un montant « rogné » inférieur à 44 millions de kú.

Tableau 11: Impact des poids winsorisés sur le BR d'estimateurs d'autres variables

Quantile	R310	R003	INV	NB_UL
100 %	0 %	13 %	527 %[27]	3 %
99 %	0 %	2 %	9 %	2 %
95 %	-1 %	0 %	6 %	0 %
90 %	-1 %	0 %	1 %	0 %
75 %	-3 %	-1 %	-1 %	0 %
50 %	-5 %	-3 %	-3 %	0 %
25 %	-7 %	-5 %	-7 %	-1 %
10 %	-13 %	-11 %	-19 %	-2 %
5 %	-24 %	-23 %	-37 %	-4 %
1 %	-52 %	-69 %	-61 %	-9 %
0 %	-69 %	-187 %[28]	-70 %	-13 %

3.5. Conclusion sur l'adaptation du traitement des valeurs influentes

Au vu des résultats obtenus dans cette étude, la winsorisation avec les seuils de Kokic et Bell, appliquée comme si les UL étaient tirées selon un plan de sondage aléatoire simple stratifié paraît être la meilleure option pour traiter les valeurs influentes avec le nouveau plan de sondage. Les résultats obtenus avec cette méthode sont meilleurs pour le chiffre d'affaires comme pour d'autres variables plus ou moins corrélées au chiffre d'affaires (valeur ajoutée, investissement).

Cette winsorisation a d'ailleurs été mise en pratique lors de la production des résultats semi-définitifs d'Esane 2016, aboutissant à des résultats proches de ceux obtenus lors de l'étude présentée dans cet article : 267 unités ont été winsorisées pour un montant de R310 « rogné » de 32 millions de k€.

Les résultats obtenus dans cette étude sont évidemment très liés aux indicateurs choisis pour comparer les estimateurs. En effet, les seuils de Kokic et Bell minimisent l'EQM de l'estimateur du total de la variable winsorisée, ce qui correspond justement à notre indicateur. Ce choix est cohérent avec les objectifs d'Esane qui sont d'estimer des totaux de variables pour la plupart très liées au chiffre d'affaires. Si les objectifs d'Esane et les indicateurs choisis pour réaliser l'étude n'avaient pas été ceux-ci, les résultats auraient pu être très différents. Dans l'étude, les totaux sont estimés par un estimateur par expansion. Ce type d'estimateur est bien connu et simple à étudier mais ne correspond pas aux estimateurs composites utilisés pour produire les résultats d'Esane. De plus, les estimateurs finaux d'Esane sont calés sur le R310 et sur le nombre d'unités légales par activité (3 positions de la NACE) et sont « combinés » aux données fiscales, qui sont exhaustives [7]. Si ce travail est repris dans l'avenir, il sera intéressant d'essayer de prendre en compte le processus d'estimation complet.

[27]Le biais relatif de 527 % pour le groupe 501 vient de la présence d'une unité particulièrement influente avec une valeur d'investissement de -14 680 kú pour un total pour le groupe 501 de 1 616 kú. Lorsqu'elle est échantillonnée, cette unité est systématiquement winsorisée avec un poids divisé par 3 environ.

[28]Le biais de -187 % pour le groupe 511 vient d'une unité avec une valeur de R003 particulièrement atypique : -15 924 kú, qui est systématiquement winsorisée lorsqu'elle est sélectionnée dans l'échantillon. Le total pour le groupe vaut 526 kú.

4. Adaptation du calage

Jusqu'à l'édition 2015, les poids finaux d'Esane étaient calés sur le R310 et sur le nombre d'unités légales par activité (3 positions de la NACE). Avec le 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. Afin de vérifier si cette dispersion pose ou non des problèmes de convergence du calage, nous avons testé la procédure de calage qui était utilisée jusqu'à l'édition 2015 sur 100[29] itérations de tirage d'échantillon selon le nouveau plan de sondage[30]. Pour chaque itération, nous avons, dans chacun des 66 secteurs de calage, essayé un calage sur le nombre d'UL et sur le total de R310[31] avec la méthode logit. Différentes bornes de rapports de poids maximales et minimales, qui sont des paramètres en entrée du calage avec une méthode logit, ont été testées : [0,5-2] ; [0,3 – 3] ; [0,2 – 5]. Nous avons ensuite simplement comptabilisé, pour chaque itération et chaque borne, le nombre de secteurs où le calage ne converge pas.

L'ensemble a été lancé deux fois : une fois avec en entrée du calage les poids corrigés de la non-réponse mais non winsorisés et une seconde fois avec en entrée du calage les poids corrigés de la non-réponse et winsorisés.

Tableau 12 : Nombre de secteurs où le calage ne converge pas

	Bornes	Moyenne	Ecart-type	Minimum	Maximum
Poids non winsorisés	[0,5 - 2]	4,8	1,9	1	10
	[0,3 - 3]	1,3	1	0	6
	[0,2 - 5]	0,7	0,7	0	2
Poids winsorisés	[0,5 - 2]	3,7	1,7	0	9
	[0,3 - 3]	1	0,9	0	5
	[0,2 - 5]	0,4	0,6	0	2

Les résultats obtenus correspondent à des ordres de grandeurs que nous avons rencontrés dans les dernières éditions des ESA. À moins qu'un des futurs échantillons soit réellement atypique, risque qui se présentait de toute façon chaque année avec l'ancien plan de sondage, son calage ne devrait pas poser plus de problèmes que d'habitude : dans une majorité de secteurs, le calage devrait converger avec des bornes « raisonnables » et quelques secteurs (a priori 10 au maximum) devront être examinés de plus près et éventuellement modifiés à la marge (ajustement des bornes de calage ou regroupement des secteurs).

On peut noter aussi que ces simulations incitent plutôt[32] à réaliser le calage après traitement des valeurs influentes, comme cela est fait actuellement.

Lors du calage réalisé pour la production des résultats semi-définitifs d'Esane 2016, 8 secteurs ne convergeaient pas avec les bornes [0,5 – 2], ce qui est un peu supérieur à ce qui est observé en moyenne dans nos simulations, sans poser pour autant de véritable problème en pratique.

[29]Nous avons pris 100 itérations et non 1000 pour des raisons de temps de traitement. Il faut en effet, pour chaque itération, tester le calage sur chacun des 66 secteurs de calage en essayant plusieurs bornes.

[30]Nous avons aussi simulé de la non-réponse et calculé des poids corrigés de la non-réponse, comme pour les simulations concernant le traitement des valeurs influentes (nous avons en pratique utilisé les 100 premières itérations des simulations précédentes).

[31]Le calage du total de R310 est en général décliné à un niveau plus fin que le calage sur le nombre d'unités légales, comme cela est fait actuellement.

[32]L'indicateur n'a pas été calculé mais il est possible que le nombre d'unités avec un poids inférieur à 1 à l'issue du calage soit plus élevé quand on utilise les poids winsorisés, ces derniers étant plus proches de 1.

5. Conclusion

Les travaux d'adaptation des traitements post-collecte d'Esane au nouveau plan de sondage pour la diffusion des résultats au niveau unités légales ont été menées dans l'année 2017. Pour le traitement de la non-réponse totale, l'étude s'est limitée à vérifier qu'il n'y avait pas trop de corrélations entre les probabilités de répondre des différentes unités légales d'une même entreprise. Il semble qu'il y en ait au moins un peu, mais finalement les tailles d'EP dans la partie non exhaustive étant limitées, négliger les corrélations entre les probabilités de réponse des UL d'une même EP ne devrait pas poser de « gros » problème au moment des estimations. Le traitement des valeurs influentes représentait clairement l'enjeu le plus important de ces études. Deux solutions ont été testées : la première consistait à essayer d'appliquer la méthode actuelle de winsorisation dans un cadre qui n'est pas tout à fait celui dans lequel elle est optimale, la seconde à essayer une méthode alternative basée sur les biais conditionnels. Cette seconde méthode est plus « souple » mais la finalité n'est pas la minimisation de l'erreur quadratique moyenne des estimateurs. D'après nos simulations, la méthode de winsorisation aboutit à de meilleurs résultats en termes d'erreur quadratique moyenne, même si le cadre dans lequel on l'applique n'est pas le cadre dans lequel elle a été conçue. Enfin, la dispersion des poids des UL d'une même activité a priori plus importante qu'auparavant avec le nouveau plan de sondage pouvait faire craindre une convergence moins aisée des procédures de calage. Les simulations ont permis de vérifier que le calage ne posait pas plus de problèmes qu'auparavant, ce qui s'est d'ailleurs confirmé lors de l'élaboration des résultats semi-définitifs 2016.

6. Annexes : Problèmes de mise en pratique rencontrés avec les estimateurs robustes

La forme des estimateurs robustes $\hat{t}_R = \hat{t} - \frac{1}{2}(B_{min} + B_{max})$: estimateur d'Horvitz-Thompson avec un terme additionnel peut poser des problèmes pratiques pour entrer dans les chaînes de programme existantes. En effet, pour s'appliquer à n'importe quelle autre variable d'intérêt, la méthode de traitement des unités influentes doit pouvoir se traduire par l'attribution à chaque unité répondante de l'échantillon d'un poids d'estimation traité des unités influentes. La méthode classique pour obtenir un tel poids est de traduire l'estimateur robuste sous une forme du type :

$$\hat{t}_R = \sum_{i \in S} w_i \, y_{iR}$$

L'enjeu est alors de trouver le moyen de calculer y_{iR}. Le poids traité des unités influentes est alors obtenu de manière analogue à celle appliquée pour obtenir des poids winsorisés dans la méthode de Kokic et Bell.

L'article [6] propose une méthode pour passer de la formule de l'estimateur robuste proposée par Beaumont, Haziza et Ruiz-Gazen à la somme pondérée des valeurs y_{iR} que nous avons essayée de mettre en œuvre sur les données. Elle repose sur une constante c, telle que :

$$y_{iR} = y_i - \frac{B_i - \psi_c(B_i)}{w_i} \quad \text{avec} \quad \psi_c(B_i) = sign(B_i) \times min(|B_i|, c)$$

Un algorithme, décrit dans l'article, permet de calculer les y_{iR} pour l'ensemble de nos données. Avec la forme de ψ_c , on voit que seules les unités ayant les biais conditionnels les plus importants (supérieurs en valeur absolue à c) auront une valeur y_{iR} différente de y_i.

Dans Esane, actuellement, on travaille ensuite avec des poids winsorisés, cela permet de transférer à d'autres variables l'effet du traitement des valeurs influentes pour le R310 (avec l'idée qu'une observation influente pour le R310 a de fortes chances d'être influente pour d'autres variables, notamment lorsqu'elles sont liées au R310) en respectant les relations comptables existant entre variables (si par exemple le chiffre d'affaires est winsorisé mais que la valeur ajoutée et les consommations intermédiaires ne sont pas modifiées, la relation comptable liant ces trois variables n'est plus respectée).

Pour passer des y_{iR} au w_{iR}, il « suffit » d'utiliser la formule suivante :

$$w_{iR} = w_i \frac{y_{iR}}{y_i}$$

En pratique cette formule pose problème dès lors que y_i est nul. Nous rencontrons ce problème, au moins une fois parmi les 1 000 réplications d'échantillons, dans un tiers des secteurs[33] avec nos données.

Il arrive en effet que des unités dont le chiffre d'affaires est nul soient influentes. Cela ne se produit jamais quand le plan de sondage des unités légales est considéré comme poissonien, mais peut se produire quand le plan de sondage est décrit de manière adéquate (enchaînement d'un sondage par grappes stratifié puis d'un sondage poissonien).

Des pistes de solutions pourraient être :

- d'utiliser l'estimateur d'Horvitz-Thompson pour les secteurs concernés ;

- travailler sur une forme de passage du y_{iR} au w_{iR} différente (qui n'impose pas $y_{iR}>0$).

Dans l'article [4] une autre méthode est décrite, plus proche de la méthode actuelle de winsorisation, pour modifier les valeurs de y de façon à intégrer le traitement des valeurs influentes. On recherche ainsi un seuil K tel que :

$$\begin{array}{ll} y_{iR} = y_i & w_i y_i \leq K \\ y_{iR} = \dfrac{K}{w_i} & w_i y_i > K \end{array} \quad \text{et} \quad \hat{t}_R = \sum_{i \in S} w_i y_{iR}$$

L'avantage de cette méthode est qu'elle modifie les unités avec le $w_i y_i$ le plus grand plutôt que celles avec le biais conditionnel le plus grand. Cela permet de contourner la difficulté rencontrée avec la méthode précédente (les $y_i=0$ ne seront pas dans les $w_i y_i$ les plus grands).

[33] Ce sont plutôt des secteurs avec peu d'unités (ressortant dans beaucoup de réplications), mais pas uniquement, par exemple le secteur 472 pose problème dans 32 réplications alors qu'il y a 1 122 répondants en moyenne.

En revanche on voit un autre problème apparaître concernant tous les cas où $\hat{t}_R > \hat{t}$. En effet on voit que de la façon dont K est construit, on a forcément $y_{iR} \leq y_i$ et donc $\hat{t}_R \leq \hat{t}$

Dès lors que $-\frac{1}{2}(B_{min}+B_{max})>0$ le calcul de K est impossible. Cela se produit :

- Jamais avec la version V1 de l'estimateur robuste (les biais conditionnels avec un tirage de Poisson ne sont jamais négatifs si y_i>=0) ;

- Au moins une fois parmi les 1 000 réplications d'échantillons, pour la moitié des secteurs avec la version V2.

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.