TRAITEMENT DES VALEURS ATYPIQUES D'UNE ENQUÊTE PAR WINSORIZATION - APPLICATION AUX ENQUÊTES SECTORIELLES ANNUELLES

Thomas DEROYON[1] (*)
(*) Insee, Direction de la méthodologie et de la coordination statistique et internationale

Résumé

Dans une enquête, une unité est atypique si ses réponses diffèrent fortement des réponses des autres unités de l'échantillon ayant le même poids de sondage, par exemple des unités de la même strate de tirage dans le cas d'un sondage stratifié. Les unités atypiques, en augmentant la dispersion des variables d'intérêt dans la population, diminuent la précision des estimateurs pouvant être construits à partir des données d'enquête.

La winsorization est une technique permettant d'identifier et de traiter certaines de ces unités. Dans le cas d'un sondage stratifié, elle consiste à définir, pour une variable d'intérêt, des seuils dans chaque strate de tirage. Si les valeurs déclarées pour une variable d'intérêt dépassent ces seuils, l'unité est considérée comme atypique et sa réponse est « rabotée ». Le total de la variable d'intérêt est alors estimé comme la somme pondérée des valeurs winsorizées.

La winsorization introduit un biais dans les estimateurs, en modifiant les réponses des entreprises, mais diminue la variance des estimations en réduisant la variance des réponses. Le choix des seuils est ce faisant crucial : c'est lui qui détermine si l'arbitrage biais variance qu'introduit la winsorization permet, au final, un gain réel en précision.

Depuis 2008, les enquêtes du dispositif Esane utilisent des techniques de winsorization suivant la méthode proposée par Kokic et Bell [9]. Cette méthode suppose de disposer de données, extérieures à l'enquête, sur la distribution de la variable winsorizée dans les strates de tirage. Pour définir une stratégie d'actualisation annuelle des seuils, nous avons comparé l'utilisation de différentes données possibles. Même si ces stratégies ont des effets proches en termes de précision, nous avons choisi d'utiliser une particularité du dispositif d'Esane, grâce à laquelle nous disposons du chiffre d'affaires fiscal de toutes les entreprises de la base de sondage, pour définir les seuils de winsorization qui seront utilisés, à partir de 2013, pour la winsorization des enquêtes d'Esane.

Abstract

Outliers are sampled units whose responses differ from the responses of units having the same sampling weights, for example belonging to the same stratum in stratified sampling. Outliers cause greater variance for estimators based on survey data.

Winsorization is a robust estimation technique aiming at identifying a certain type of outliers and limiting their effect on estimators variance. In case of stratified sampling, thresholds are defined for each sampling stratum : the responses higher than the thresholds are shrinked. The estimators obtained with this method are biased, but more precise. Kokic and Bell [9] suggested a method to compute threshold guaranteing a gain in precision for estimators. We show how we applied this method to Structural Business Statistics surveys in France.

Mots-clés

[1] thomas.deroyon@insee.fr

Unités influentes, winsorization

Introduction

Les techniques usuelles de sondage permettent d'obtenir des estimateurs non biaisés de paramètres d'intérêt d'une population, comme la moyenne ou le total d'une variable d'intérêt, sans poser d'hypothèses particulières sur la population enquêtée. Ces estimateurs peuvent cependant manquer de précision, aussi des techniques particulières ont-elles été développées pour réduire leur variance. Le calage sur marges en est un exemple.

Les unités atypiques (*outliers* en anglais) sont l'une des causes fréquentes de ce défaut de précision. Une unité atypique est une unité dont les réponses diffèrent fortement des réponses d'autres unités de l'échantillon ayant le même poids de sondage. Dans le cas d'un sondage stratifié, une unité atypique est ainsi une unité ayant des réponses très éloignées des réponses des autres unités de sa strate.

Les unités atypiques induisent une dispersion accrue des grandeurs que nous souhaitons quantifier dans l'enquête, partant une variance plus grande des estimations. Ce problème est particulièrement fréquent dans les enquêtes auprès des entreprises. Les distributions de leurs caractéristiques quantitatives (chiffre d'affaires, masse salariale, investissement…) sont en effet souvent très dissymétriques : pour beaucoup de ces caractéristiques, un faible nombre d'entreprises concentrant une grande part du total de la variable cohabitent avec un nombre important de très petites entreprises dont la contribution aux agrégats est faible.

Deux types d'unités atypiques peuvent être présentes dans un échantillon (voir [5]) :

- Unités atypiques non-représentatives (*non-representative outliers*) : ces situations renvoient à deux cas possibles. Soit les réponses de l'unité sont fausses ; soit, l'unité, bien qu'elle ait été échantillonnée avec un poids de sondage *p* strictement supérieur à 1 et qu'elle ait vocation de ce fait à représenter *p-1* unités de la population en plus d'elle-même, ne peut représenter qu'elle-même. Les entreprises participant à une restructuration (issues de la scission d'une entreprise en plusieurs unités, ou absorbées par une autre entreprise) sont en général considérées comme atypiques et placées, en cours de collecte ou en fin de campagne, dans les strates exhaustives.

- Unités atypiques représentatives (*representative outliers*) : bien que leurs réponses diffèrent fortement de celles des autres entreprises ayant environ le même poids, il n'est pas possible de considérer que les réponses de ces unités sont fausses ou ne représentent qu'elles mêmes.

Les méthodes décrites dans cet article visent à identifier et traiter les unités atypiques représentatives. Les unités atypiques non-représentatives sont, quant à elles, traitées par des procédures de contrôles et redressements (*data editing*) et d'imputation pour les réponses fausses, ou de singularisation (attribution d'un poids de 1) pour les réponses correctes mais non représentatives.

Dans les enquêtes auprès des entreprises à l'Insee et notamment dans les plus importantes d'entre elles, les enquêtes du dispositif Esane, les Enquêtes Sectorielles Annuelles (ESA) et l'Enquête Annuelle de Production (EAP), les unités atypiques représentatives sont identifiées et traitées par winsorization.

Cette méthode repose sur le calcul de seuils sur une variable d'intérêt, les unités atypiques étant identifiées comme celles dont la valeur de la variable dépasse le seuil. La qualité de la méthode repose essentiellement sur la qualité des seuils utilisés. Aussi nous présentons dans un premier temps le principe de la winsorization ainsi que la méthode de calcul des seuils utilisée pour les enquêtes d'Esane. Nous détaillons ensuite comment cette méthode a été adaptée et appliquée aux ESA et à l'EAP à partir de 2008.

.

1. Winsorization : principe et méthode de Kokic et Bell

1.1. Cadre et principe

La winsorization suivant la méthode de Kokic et Bell (voir [9]) s'applique à un sondage stratifié à un degré avec sondage aléatoire simple dans chaque strate.

La population, de taille *N*, est divisée en un ensemble de *H* strates U_h de tailles N_h. Dans chacune de ces strates, un échantillon s_h de taille n_h est sélectionné par sondage aléatoire simple, indépendamment des échantillons tirés dans les autres strates.

Si *X* est une variable d'intérêt de l'enquête, l'estimateur usuel du total de *X*, appelé estimateur d'Horvitz-Thompson, est donné par : $\hat{X} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in s_h} X_i$

Il est sans biais et sa variance est égale à :

$V(\hat{X}) = \sum_{h=1}^{H} N_h^2 (1 - \frac{n_h}{N_h}) \frac{S_h^2(X)}{n_h}$ avec $S_h^2(X) = \frac{\sum_{i \in U_h} (X_i - \overline{X}_h)^2}{N_h - 1}$ et $\overline{X}_h$ la moyenne empirique de *X* dans la strate *h*.

Si une strate contient des unités atypiques représentatives, cela signifie qu'un petit nombre d'unités de la strate ont des valeurs très éloignées de la moyenne de *X* dans la strate, donc que la variance empirique de la variable *X* dans la strate, $S_h^2(X)$, est importante, ce qui nuit directement à la précision de $\hat{X}$.

La winsorization consiste alors à définir un seuil K_h associé à chaque strate et à remplacer *X* par une variable winsorizée X^w égale à *X* si *X* est inférieure au seuil, et inférieure à *X* dans le cas contraire.

Deux types de winsorization sont utilisées :

1. **Winsorization de type 1 :** toutes les valeurs de *X* dépassant le seuil sont tronquées à la valeur du seuil.

2. **Winsorization de type 2 :** seule une part égale au taux de sondage dans la strate des valeurs de *X* au delà du seuil est conservée dans la valeur de la variable winsorizée. Ainsi,

$$X^w = \begin{cases} X, \, si \; X < K_h \\ \frac{n_h}{N_h} X + (1 - \frac{n_h}{N_h}) K_h, \, si \; X > K_h \end{cases}$$

Remarquons que, dans les strates exhaustives, le taux de sondage n_h / N_h est égal à 1 : dans ces strates, une winsorization de type 2 n'a aucun effet.

L'estimateur winsorizé du total de *X*, $\hat{X}^w$, est alors égal à l'estimateur d'Horvitz-Thompson du total de la variable winsorizée X^w.

En tant qu'estimateur du total de *X*, cet estimateur est biaisé. Sa variance dépend par contre de la variance empirique de la variable winsorizée dans chaque strate. Or, la variable winsorizée, obtenue en rabotant les valeurs extrêmes de *X*, est moins dispersée que celle-ci. La winsorization crée donc un arbitrage biais - variance : elle est efficace si le biais qu'elle introduit est plus que compensé par la

baisse de variance qu'elle permet. Dans ce cas, l'erreur quadratique moyenne[2] de l'estimateur winsorizé est plus faible que celle de l'estimateur d'Hovitz-Thompson.

Les deux types de winsorization donnent en général des résultats proches (voir [5]).

La qualité d'une winsorization dépend par contre crucialement du choix des seuils K_h. Kokic et Bell ont proposé, dans le cadre d'une winsorization de type II, une méthode pour calculer des seuils de winsorization optimaux, *i.e.* permettant, sous certaines hypothèses, d'apporter un gain de précision maximal.

1.2. La méthode de calcul des seuils de winsorization de Kokic et Bell

Winsorizer suppose d'identifier des valeurs atypiques, donc de poser des hypothèses sur la distribution de la variable winsorizée, permettant de distinguer les valeurs « normales » des valeurs qui le sont moins. Kokic et Bell supposent ainsi qu'à l'intérieur d'une strate, les valeurs de la variable d'intérêt sur laquelle porte la winsorization sont toutes des réalisations indépendantes d'une même loi.

Sous cette hypothèse, Kokic et Bell proposent de calculer les seuils K_h de manière à ce qu'ils soient indépendants de l'échantillon auquel ils sont appliqués : ils peuvent de ce fait être utilisés quelles que soit les valeurs de la variable winsorizée dans l'échantillon. Les seuils sont ainsi calculés de manière à minimiser l'erreur quadratique de l'estimateur winsorizé, cette erreur étant calculée en tenant compte de l'aléa résultant à la fois du plan de sondage et de la distribution de la variable *X* dans la population.

Ainsi, en moyenne sur l'ensemble des échantillons et sur les valeurs possibles que peut prendre *X* dans ces échantillons, *i.e.* sur l'ensemble des situations auxquelles les données d'enquête peuvent nous confronter, l'estimateur winsorizé a l'erreur quadratique moyenne la plus faible possible.

Kokic et Bell cherchent ainsi à calculer un estimateur winsorizé dont les propriétés prolongent celles de l'estimateur classique d'Horvitz-Thompson. Celui-ci est en effet sans biais dans la mesure où, si tous les échantillons possibles de la population étaient tirés et l'estimateur d'Horvitz-Thompson du total de *X* calculé dans chacun d'entre eux, la moyenne de ces estimateurs pondérés par la probabilité qu'a chaque échantillon d'être sélectionné serait exactement égale au total de *X*. L'estimateur winsorizé calculé avec les seuils de Kokic et Bell n'est plus sans biais, mais il commet l'erreur la plus faible dans l'estimation du total de *X* en moyenne sur tous les échantillons possibles et sur toutes les valeurs possibles de X dans ces échantillons.

Sous ces hypothèses, il est possible d'exprimer le biais et la variance de l'estimateur winsorizé résultant de l'aléa de sondage et de la loi de la variable *X* dans chaque strate, d'écrire le programme de minimisation de l'erreur quadratique moyenne et de le résoudre.

Kokic et Bell montrent qu'à l'optimum le biais de l'estimateur winsorizé est le point où s'annule la fonction *F* définie par :

$$F(B) = -B\left[1 + \sum_h n_h E_h(J_h^*)\right] - \sum_h n_h E_h(X_h^* J_h^*) \quad (1)$$

avec E_h l'espérance selon la loi de *X* dans la strate *h*, $X_h^* = (\frac{N_h}{n_h} - 1)(X_h - \mu_h)$, μ_h l'espérance de *X* dans la strate *h* et J_h^* la variable indicatrice valant 1 si X_h est supérieur à K_h.

et K_h est équivalent asympotiquement, dans chaque strate, à $-\frac{B}{\frac{N_h}{n_h} - 1} + \mu_h$ (2) , quand la taille de la population et de l'échantillon tendent vers l'infini.

[2] somme du carré du biais de l'estimateur et de sa variance.

A l'optimum, les seuils sont donc égaux à l'espérance de *X* dans chaque strate, augmentée d'un terme positif proche de la valeur absolue du biais multipliée par le taux de sondage dans la strate.

Ainsi, quand le taux de sondage est faible, et les poids de sondage élevés dans une strate, le seuil est très proche de l'espérance de *X* dans la strate. Ainsi, si très peu d'unités sont tirées dans une strate, il peut être utile d'en winsorizer beaucoup, car les écarts entre ces valeurs et la moyenne de *X* dans la strate, extrapolés par des poids de sondage élevés, ont une grande incidence sur les estimateurs.

Quand le taux de sondage est proche de 1, à l'inverse, le terme ajouté à μ_h tend vers l'infini : quand le sondage dans une strate s'approche d'une interrogation exhaustive, seules les valeurs très atypiques de X méritent d'être winsorizées, car elles sont seules susceptibles d'avoir un effet suivant qu'elles sont échantillonnées ou pas.

La méthode de calcul des seuils K_h est donc simple : d'abord, il faut estimer le point où la fonction *F* s'annule, *i.e.* le biais optimal *B**. Ensuite, en déduire les seuils K_h par application de la formule (2).

Pour déterminer les zéros de la fonction *F*, il nous faut pouvoir estimer μ_h ainsi que les valeurs de $E_h(J_h^*)$ et $E_h(X_h^* J_h^*)$ pour différentes valeurs du biais[3].

Ces deux espérances sont calculées suivant la loi de la variable *X* dans la strate *h* : pour les estimer, il nous faut soit poser des hypothèses sur la forme de la distribution de *X*, soit disposer d'un ensemble de réalisations de la loi de la variable *X* dans la strate *h*. Kokic et Bell se placent dans la seconde hypothèse : nous disposons d'un ensemble de réalisations $(\tilde{X}_j^h)_{h=1..H,\, j=1..p_h}$ de la loi de la variable *X* dans les strates non exhaustives *h* (*i.e.* de valeurs de *X* dans la population des strates non exhaustives *h*).

Pour une valeur du biais donnée, nous pouvons alors estimer μ_h, $E_h(J_h^*)$ et $E_h(X_h^* J_h^*)$ par :

$$\hat{\mu}_h = \frac{1}{p_h}\sum_{j=1}^{p_h} \tilde{X}_j^h$$

$$\hat{E}_h(J_h^*) = \frac{1}{p_h}\sum_{j=1}^{p_h} I\left[\left(\frac{N_h}{n_h}-1\right)(\tilde{X}_j^h - \hat{\mu}_h) > -B\right]$$

$$\hat{E}_h(X_h^* J_h^*) = \frac{1}{p_h}\sum_{j=1}^{p_h} \left(\frac{N_h}{n_h}-1\right)(\tilde{X}_j^h - \hat{\mu}_h)\, I\left[\left(\frac{N_h}{n_h}-1\right)(\tilde{X}_j^h - \hat{\mu}_h) > -B\right]$$

Il est alors possible d'estimer la valeur de *B* qui minimise la fonction *F*, et d'en déduire les seuils K_h optimaux, suivant une procédure détaillée en annexe A.

Remarquons que, comme les calculs de Kokic et Bell reposent sur l'hypothèse que les seuils K_h sont indépendants de l'échantillon, les réalisations de la loi de *X* dans la strate *h* à partir desquelles nous allons estimer le biais optimal ne peuvent *a priori* pas être celles observées dans l'échantillon de l'enquête que nous cherchons à winsorizer. Kokic et Bell supposent donc que l'utilisateur dispose de valeurs de la variable *X* issues d'une enquête précédente, dont l'échantillon est indépendant de l'échantillon winsorizé.

[3] Ces deux espérances sont en effet des fonctions du biais puisque, asymptotiquement $X_h > K_h \Leftrightarrow X_h^* > -B$ et donc $J_h^* = I(X_h^* > -B)$.

2. Application de la winsorization aux enquêtes d'Esane

2.1. Présentation (rapide) d'Esane

Esane (Elaboration des Statistiques ANnuelles d'Entreprise) est le processus qui estime chaque année les statistiques structurelles d'entreprise, *i.e.* les principales caractéristiques du système productif français. Plus précisément, les statistiques structurelles regroupent principalement deux types d'agrégats :

- les bilans et comptes de résultat des entreprises agrégés par secteur. Ces agrégats estiment par exemple le chiffre d'affaires total des entreprises du commerce de gros, la masse salariale des entreprises du génie civil ou l'investissement des entreprises sidérurgiques ;

- la ventilation du chiffre d'affaires des entreprises d'un secteur suivant leurs différentes activités. Un secteur regroupe en effet les entreprises ayant la même activité principale, mais la plupart d'entre elles réalisent une part de leur production sur d'autres activités. En estimant le chiffre d'affaires réalisé dans chacune de ces activités, Esane permet aux comptables nationaux de construire les comptes de branche, *i.e.* de décrire la fonction de production de chaque bien ou service en France ;

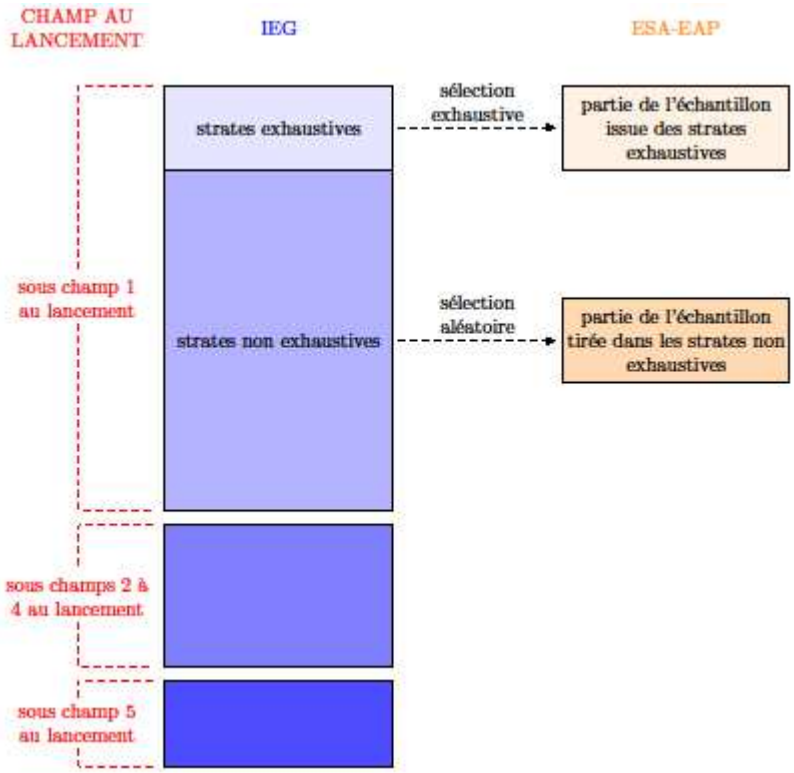
Pour estimer ces différents agrégats, Esane combine données administratives et données d'enquête (voir figure 1) :

- Esane repose sur le répertoire interadministratif Sirene, qui lui fournit chaque année la liste des unités participant *a priori* aux activités productives marchandes dans le champ couvert par la statistique structurelle d'entreprise en France, ainsi que leur secteur en début de campagne ;

- l'Insee récupère chaque année les « liasses fiscales » des entreprises, *i.e.* leur déclaration de revenu à l'administration fiscale permettant le calcul de leur impôt sur les sociétés. Ces fichiers, nommés aussi IEG (pour Information Economique Générale) contiennent, pour toutes les entreprises, l'ensemble des variables du bilan et des comptes de résultats, dans des concepts analogues à ceux utilisés dans les agrégats statistiques, ainsi que le numéro d'identification des entreprises au répertoire Sirene ;

- la source fiscale n'est cependant pas suffisante pour répondre aux besoins d'Esane ; en particulier elle ne contient qu'une ventilation très agrégée du chiffre d'affaires sur trois activités : production de biens, production de services et commerce. Des enquêtes sont donc réalisées sur un échantillon d'entreprises, l'Enquête Annuelle de Production (EAP) pour les entreprises des secteurs industriels (hors agroalimentaire) en métropole et les Enquêtes Sectorielles Annuelles (ESA) pour les entreprises des autres secteurs et des départements d'outre-mer. Ces enquêtes permettent de quantifier la ventilation du chiffre d'affaires en branche. Elles permettent également de réévaluer le secteur des entreprises de l'échantillon.

Figure 1 : Les données utilisées dans le dispositif Esane



Les estimateurs composites d'Esane, décrits dans [1] et [7], tentent de combiner au mieux ces différentes sources d'information. Ils font notamment intervenir des sommes sur l'échantillon des variables déclarées par les entreprises dans leur liasse fiscale, pondérées par les poids des entreprises après traitements post-collecte : ces estimateurs sont donc sensibles aux unités atypiques.

Or, deux facteurs favorisent leur apparition dans les ESA et l'EAP : d'une part, le fait, commun aux enquêtes entreprises et déjà évoqué plus haut, que les distributions de leur chiffre d'affaires par exemple sont très dissymétriques et concentrées ; d'autre part, le fait que les strates utilisées pour tirer l'échantillon des ESA et de l'EAP croisent secteur et tranche d'effectif, celui-ci étant mesuré à l'aide de l'emploi fourni par la source CLAP[4] provisoire. Or, cet effectif n'est pas complètement stabilisé : des entreprises employant effectivement des salariés peuvent notamment avoir un effectif manquant ou nul dans CLAP provisoire, si bien qu'elles sont classées dans une strate de très petites entreprises, et échantillonnées avec un poids de sondage élevé. Elles se distinguent cependant des autres entreprises de leur strate, sans forcément relever pour autant de l'exhaustif : si elles avaient été classées dans la strate de tirage dont elles relèvent vraiment, elles auraient pu avoir un poids de sondage plus faible, mais superieur strictement à 1.

Il n'est par ailleurs pas possible de replacer *ex-post* ces entreprises dans la strate de tirage dont elles auraient dû relever. En effet, cela supposerait que nous sachions *ex-post* où placer chaque entreprise, *i.e.* que l'effectif observé en fin de campagne est correct. Or, si les effectifs définitifs disponibles dans Esane en fin de campagne sont d'une bien meilleure qualité que ceux utilisés pour constituer les strates de tirage, ils contiennent encore des erreurs, susceptibles de générer des unités atypiques. Enfin, déplacer les entreprises dans une autre strate que celle dans laquelle elles ont été tirées, et ce faisant changer leur poids de sondage, méconnaît le fait que la probabilité avec laquelle ces entreprises ont été tirées est bien celle de leur strate d'origine. Changer leur poids de sondage introduit un biais dans l'estimateur pour diminuer sa variance, mais sans que l'arbitrage biais - variance qui en résulte soit maîtrisé. La winsorization par la méthode de Kokic et Bell est ainsi plus

[4] Connaissance Localisée de l'Appareil Productif.

adaptée pour traiter ce type de situations. Nous allons à présent voir comment elle a été appliquée aux ESA et à l'EAP (voir [2], [3] et [8]).

2.2. Adaptation de la méthode de winsorization de Kokic et Bell aux spécificités du plan de sondage d'Esane

Le plan de sondage des ESA et de l'EAP est proche du cadre proposé par Kokic et Bell : les échantillons des ESA et de l'EAP sont en effet tirés par un sondage stratifié à un degré, les strates étant constituées par croisement du secteur des entreprises dans le répertoire au niveau le plus fin (sous-classes de la NAF rev.2), de leur tranche d'effectif en 15 positions et de leur région. Les strates contenant les grandes entreprises sont interrogées exhaustivement, les autres sont échantillonnées avec des taux de sondage variant de 1/3 à environ 1/400.

L'échantillon n'est cependant pas tiré intégralement chaque année : sa partie sélectionnée dans les strates non exhaustives est en effet renouvelée par moitié chaque année, de façon à mesurer avec plus de précision les évolutions des statistiques structurelles d'entreprise.

La stratégie de renouvellement employée dans les ESA est commune aux enquêtes auprès des entreprises à l'Insee (voir [6] pour plus de détails). Chaque entreprise du répertoire reçoit, à sa création, un numéro permanent entier, appelé numéro hexal, aléatoire et donc pair dans la moitié des cas. Ce numéro permet de partager la population aléatoirement en deux parties identiques ; chaque année, seule la partie de l'échantillon sélectionnée dans une des deux moitiés est renouvelée.

Ainsi, par exemple, en 2014, toutes les unités de l'échantillon (hors strates exhaustives) interrogé pour la campagne 2013 ayant un numéro hexal pair restent dans l'échantillon (voir figure 2). Cette partie conservée de l'échantillon est complétée par :

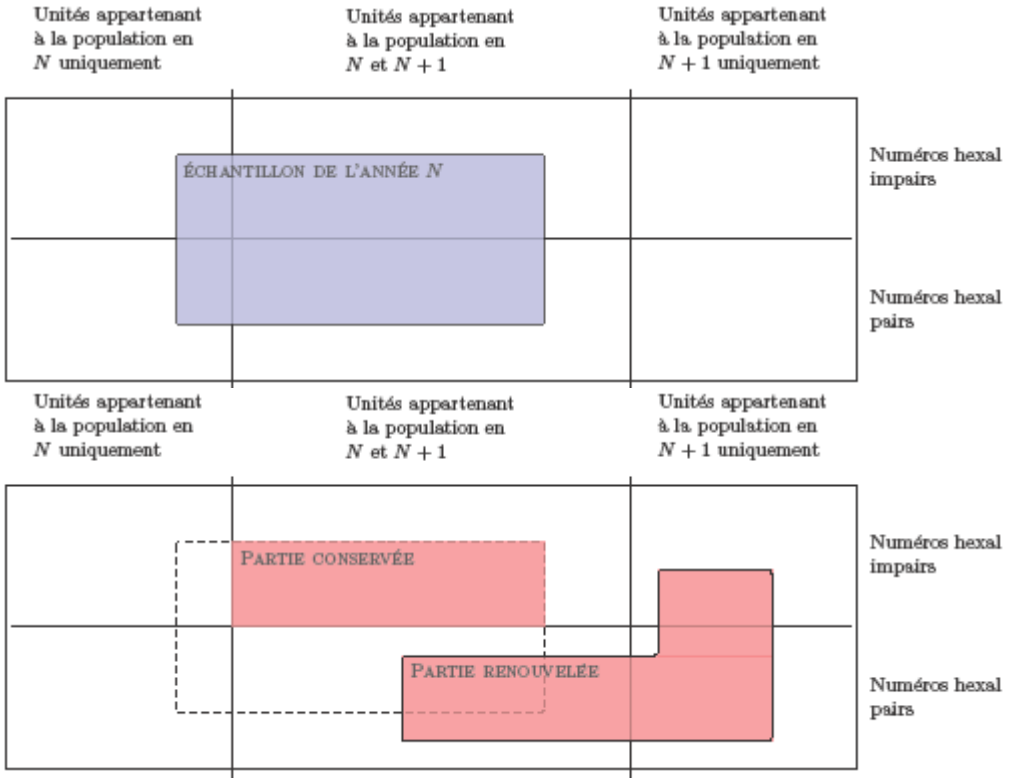
- un nouvel échantillon tiré parmi les entreprises déjà présentes dans la population en 2013 et ayant un numéro hexal impair, selon un plan de sondage stratifié à un degré, les strates étant formées en croisant secteur, tranche d'effectif et région ;

- un échantillon tiré parmi les entreprises créées en 2014, indépendamment de leur numéro hexal

Au final, les entreprises présentes dans une strate de l'échantillon ont été tirées dans deux populations distinctes. Les poids de sondage dans une même strate peuvent de ce fait différer. Dans l'application de la winsorization aux ESA et à l'EAP, cette particularité n'est pas prise en compte: tout se passe comme si toutes les entreprises d'une même strate ont été tirées la même année, dans la même base de sondage, avec le même poids de sondage.

De même, les strates dans lesquelles sont calculés les seuils de winsorization sont constituées en croisant uniquement le secteur et la tranche d'effectif. En effet, tenir compte également de la région conduirait à travailler sur des strates très petites. D'autre part, des taux de sondage identiques sont appliqués, à l'intérieur d'un croisement de secteur et tranche d'effectif, à toutes les régions.

Figure 2 : Renouvellement par moitié de l'échantillon des ESA



L'hypothèse la plus forte adoptée pour appliquer la winsorization aux ESA et à l'EAP est cependant liée à l'ordre dans lequel sont appliqués les traitements *ex-post* des enquêtes.

2.3. Quel ordre pour les traitements ex-post des enquêtes d'Esane ?

Trois traitements sont en effet appliqués aux données des enquêtes après leur collecte, leur contrôle et leur apurement :

- une correction de la non-réponse totale, ayant pour but de limiter le biais introduit par le fait que toutes les entreprises interrogées ne répondent pas. Cette correction est réalisée par imputation dans les strates exhaustives, et par repondération selon la méthode des groupes de réponse homogène (GRH) dans le reste de l'échantillon (voir [4]). Cette méthode revient à supposer que les répondants sont sélectionnés dans l'échantillon par un sondage stratifié à un degré, les strates étant égales aux groupes de réponse homogène. En pratique, les GRH diffèrent des strates de tirage de l'échantillon : ils sont constitués à partir des variables explicatives du comportement de réponse, qui peuvent différer des variables de stratification ;

- un calage sur les marges de nombre d'entreprises et chiffre d'affaires par secteur dans le répertoire, visant à améliorer la précision des estimateurs ;

- la winsorization.

L'ordre initialement envisagé pour ces trois opérations était le suivant : correction de la non-réponse, calage et winsorization.

La winsorization devait ainsi intervenir en toute fin du processus, pour traiter les quelques unités atypiques pouvant être identifiées. Il s'est cependant avéré que celles-ci pouvaient gêner fortement le calage sur marges, voire empêcher qu'il soit mis en œuvre à des niveaux sectoriels suffisamment fins. Aussi, winsorization et calage sur marges ont-ils été finalement inversés.

Dans tous les cas, la winsorization intervient après la correction de la non-réponse. Les seuils sont calculés par strate de tirage croisant secteur et tranche d'effectif et appliqués aux répondants, ce qui revient à supposer que l'échantillon des répondants est sélectionné dans la population initiale par un sondage stratifié à un degré, les strates étant confondues avec les strates de tirage de l'échantillon initial.

Il s'agit d'une hypothèse simplificatrice, qui ne tient pas compte de la manière dont est opérée la correction de la non-réponse par repondération. En effet, la méthode des GRH revient à assimiler la sélection des répondants dans la population à un plan de sondage en deux phases :

- la première phase correspond à la sélection de l'échantillon initial ; c'est un sondage stratifié à un degré, suivant des strates croisant secteur, tranche d'effectif et région ;

- la deuxième phase décrit la sélection des entreprises répondantes parmi les entreprises interrogées ; il s'agit également d'un sondage stratifié à un degré, les strates étant cette fois égales aux groupes de réponse homogène, distincts des strates de tirage de l'échantillon initial.

Remarquons que, comme la non-réponse est traitée par imputation dans les strates exhaustives, et que la winsorization n'a aucun effet sur les unités tirées dans les strates exhaustives, la winsorization n'est appliquée qu'aux strates non exhaustives.

2.4. Comment winsorizer l'ensemble des variables d'une liasse fiscale ?

La liasse fiscale d'une entreprise contient un grand nombre de variables, liées entre elles par de nombreuses relations comptables.

Une entreprise peut n'être atypique que pour certaines de ces variables. Aussi, il serait possible de réaliser une winsorization séparée pour chaque variable de la liasse fiscale. Des seuils seraient calculés pour le chiffre d'affaires, la valeur ajoutée, l'excédent brut d'exploitation, l'investissement,… et les valeurs atypiques identifiées et traitées sur la base de ces seuils.

Cette méthode risque cependant de rompre les relations comptables existant entre les variables d'une même liasse pour les unités winsorizées. Aussi, la winsorization est appliquée à toutes les variables d'une liasse fiscale sur la base du chiffre d'affaires fiscal :

- les unités atypiques sont identifiées à l'aide des seuils définis sur le chiffre d'affaires *X* et la valeur winsorizée X^w calculée pour chaque entreprise des strates non exhaustives ;

- pour toute autre variable *Y* de la liasse fiscale, la valeur winsorizée est calculée comme $Y^w = Y\frac{X^w}{X}$

Cet ajustement est efficace si la variable *Y* est très corrélée au chiffre d'affaires *X*. C'est le cas par exemple de la valeur ajoutée et de la masse salariale. Par contre, cette méthode peut poser problème quand il s'agit de détecter les valeurs atypiques de variables peu corrélées au chiffre d'affaires, comme l'investissement, qui, de manière générale, est une variable complexe à traiter, présentant une grande variance et une faible cohérence temporelle.

3. Comment calculer et actualiser les seuils de winsorization

Pour calculer les seuils de winsorization suivant la méthode de Kokic et Bell, nous devons disposer de données sur le chiffre d'affaires d'un échantillon d'entreprises dans chaque strate de tirage, cet échantillon étant sélectionné indépendamment des entreprises de l'échantillon auquel est appliquée la winsorization.

Lors de la mise en place d'Esane, les seuils ont été calculés à partir des données des Enquêtes Annuelles d'Entreprise (EAE) de 2007. Les EAE sont les prédécesseurs des ESA. Elles couvraient le même champ que les ESA et l'EAP actuelles. Cependant, les EAE dans l'industrie n'interrogeraient (exhaustivement) que les grandes entreprises[5].
Le questionnaire des EAE demandait aux entreprises de détailler leurs comptes de résultat et leur bilan : l'un des objectifs d'Esane a consisté à alléger la charge statistique des entreprises en limitant les enquêtes à la collecte de la ventilation du chiffre d'affaires en branche et en exploitant pour le reste les liasses fiscales.

Ces données présentaient cependant deux limites :

- les EAE ne couvraient pas les strates non exhaustives des secteurs de l'industrie. Aussi aucune donnée n'existait permettant de calculer des seuils de winsorization pour les strates de l'industrie.

- Les données des EAE permettent de calculer des seuils pertinents pour les années proches de 2007 ; mais à mesure que nous nous éloignons de 2007, la distribution du chiffre d'affaires dans chaque strate se déforme et les seuils perdent de leur pertinence.

En 2013, nous avons ainsi mené une étude afin d'actualiser les seuils de winsorization et de définir plus largement une stratégie pour réévaluer chaque année ces seuils. Une stratégie tient en fait au choix d'une base de donnée fournissant des informations sur les chiffres d'affaires des entreprises dans chaque strate de tirage.

Nous avons d'abord testé l'utilisation des données des ESA et de l'EAP. Ces données présentaient cependant deux limites.

D'une part, comme les données utilisées pour calculer les seuils doivent être indépendantes des données auxquels ces seuils sont appliqués, nous ne pouvions utiliser les données de l'enquête de l'année *N*, pas plus que celles de l'année précédente compte tenu de la stratégie de renouvellement par moitié de l'échantillon, pour calculer des seuils de winsorization applicables aux données de l'enquête de l'année *N*. L'enquête la plus proche pouvant être mobilisée est donc l'enquête réalisée en *N-2*. D'autre part, le taille de l'échantillon tiré hors des strates exhaustives a été divisée par moitié environ avec le passage des EAE aux ESA, aussi, nous disposions de peu de données sur lesquelles estimer les caractéristiques de la distribution du chiffre d'affaires par strate. Il en résultait une certaine variation du nombre d'unités winsorizées et de la réduction de chiffre d'affaires induite par la winsorization d'une année sur l'autre, ou une même année en changeant l'enquête utilisée pour calculer les seuils (par exemple en utilisant les enquêtes de 2009 ou de 2010 pour calculer les seuils applicables à 2012).

C'est pourquoi nous avons utilisé une particularité du dispositif Esane : la source fiscale étant exhaustive[6], nous disposons du chiffre d'affaires fiscal pour toutes les entreprises de chaque strate de tirage de l'échantillon. Aussi, il nous est possible d'utiliser cette information, disponible sur la totalité de la base de sondage de l'enquête de l'année *N*, pour calculer les seuils applicables aux unités répondantes à l'enquête de l'année *N*.

Il est peu fréquent de disposer, dans la base de sondage, de la variable que nous souhaitons winsorizer. Cette particularité tient au caractère composite d'Esane, qui mobilise données administratives et données d'enquête, et à la forme des estimateurs qu'il emploie. Dans la plupart des enquêtes, il n'est possible que d'utiliser une édition antérieure de l'enquête pour calculer les seuils ; si l'enquête n'est pas répétée, seules les données de l'enquête sont disponibles.

C'est pourquoi nous avons comparé les effets de la winsorization obtenue avec les seuils calculés sur les données de la base de sondage avec la winsorization due aux seuils calculés sur les données de l'enquête à laquelle ils sont appliqués. Ceci nous donne une indication sur l'importance de l'hypothèse

[5] Employant plus de 20 salariés.
[6] Même si toutes les entreprises ne renvoient pas leur déclaration fiscale, cette non-réponse est traitée par imputation, aussi nous disposons bien d'une liasse fiscale pour toute entreprise du champ d'Esane, que cette liasse ait été effectivement déclarée ou ait été imputée.

d'indépendance des seuils par rapport aux données auxquelles ils sont appliqués, et du risque pris quand elle n'est pas respectée.

Tableau 1 : Nombre d'unités winsorizées par secteur (au niveau section) dans l'échantillon 2012 des ESA et de l'EAP en fonction des données utilisées pour calculer les seuils

	Seuils calculés avec l'enquête 2012	Seuils calculés avec l'enquête 2010	Seuils calculés avec la base de sondage 2012	Seuils calculés avec les EAE 2007
Ensemble de la population	374	469	220	268
A : agriculture	2	2	2	0
B : industries extractives	4	5	1	0
C : industrie manufacturière	139	170	81	0
CA : agroalimentaire	16	23	9	11
D : production et distribution d'électricité	3	0	2	0
E : production et distribution d'eau	10	7	8	0
F : construction	16	22	17	27
G : commerce	39	73	27	55
H : transport et entreposage	19	21	9	12
I : hébergement restauration	22	15	9	21
J : information communication	27	16	6	20
L : activités immobilières	7	7	2	1
M : activités spécialisées, scientifiques et techniques	24	23	12	43
N : activités de services administratifs et de soutien	31	68	32	72
R : arts et spectacles	2	1	1	0
S : autres activités de service	13	16	2	6

Tableau 2 : Diminution du chiffre d'affaires (non pondéré) due à la winsorization par secteur dans l'échantillon 2012 des ESA et de l'EAP en fonction des données utilisées pour calculer les seuils

	Seuils calculés avec l'enquête 2012	Seuils calculés avec l'enquête 2010	Seuils calculés avec la base de sondage 2012	Seuils calculés avec les EAE 2007
Ensemble de la population	-750 903	-1 346 989	-1 120 262	-1 354 927
A : agriculture	-2 075	-4 209	-2 518	0
B : industries extractives	-1 764	-6 856	-2 654	0
C : industrie manufacturière	-105 056	-191 958	-122 739	0
CA : agroalimentaire	-21 483	-40 910	-22 708	-14 717
D : production et distribution d'électricité	-3 826	0	-2 520	0
E : production et distribution d'eau	-20 852	-28 801	-34 265	0
F : construction	-58 287	-119 449	-99 636	-195 191
G : commerce	-340 992	-605 096	-604 051	-845 599
H : transport et entreposage	-12 598	-13 502	-6 138	-8 326
I : hébergement restauration	-10 188	-4 667	-10 754	-30 880
J : information communication	-33 768	-38 454	-24 865	-14 487
L : activités immobilières	-13 579	-22 163	-1 696	-473
M : activités spécialisées, scientifiques et techniques	-74 038	-128 197	-104 819	-82 869
N : activités de services administratifs et de soutien	-49 963	-139 624	-80 034	-161 017
R : arts et spectacles	-1 270	-267	-450	0
S : autres activités de service	-1 164	-2 837	-414	-1 367

Les tableaux 1 et 2 donnent, appliquée aux données des ESA et de l'EAP de 2012 les effets de la winsorization en termes de nombre d'unités winsorizées et de montant de chiffre d'affaires winsorizé, en fonction des données utilisées pour calculer le seuil.

L'utilisation des données de la base de sondage de 2012 se traduit par la winsorization d'un nombre beaucoup plus faible d'unités : 220 contre 374 quand les seuils sont calculés avec les données des ESA et de l'EAP 2012. Il est proche du nombre d'unités winsorizées par les seuils calculés sur les EAE 2007, mais ces derniers ne concernaient que les strates de tirage des ESA et pas l'EAP. Sur le champ des seules ESA[7], les seuils calculés avec les données de la base de sondage amènent à winsorizer 127 entreprises, soit moitié moins que les seuils issus des EAE 2007.

Si les seuils calculés sur la base de sondage de 2012 amènent à winsorizer moins d'unités que les autres seuils, en particulier les seuils issus des enquêtes 2012, le montant de la winsorization qu'ils induisent est élevé, et supérieur à celui induit par les seuils calculés sur les enquêtes. Les seuils de la base de sondage se concentrent ainsi sur un plus faible nombre d'unités, et de strates, mais chacune de ces unités est ajustée plus fortement.

Ainsi, sur les 1 645 strates de tirage non exhaustives formées par le croisement du secteur et de la tranche d'effectif, 171 contiennent au moins une unité winsorizée avec les seuils issus de la base de sondage, contre 283 avec les seuils issus de l'enquête 2012, et 270 avec les seuils issus de l'enquête 2010. Mais les ajustements sur les 171 strates winsorizées sont plus élevés : alors que la moyenne des totaux des ajustements par strate est de -13 000 k€ sur les 171 strates winsorizées avec les seuils issus de la base de sondage, elle n'est que de - 5 300 k€ avec les seuils issus de l'enquête 2012 (- 9 900 k€ avec les seuils issus de l'enquête de 2010) .

Quant aux effets de la winsorization sur la précision des estimateurs :

- les estimateurs winsorizées sont plus précis que les estimateurs non winsorizés, quels que soient les seuils utilisés, comme l'ont montré les travaux de Fabien Guggemos (voir [3] et [8]) ;

- les précisions des estimateurs winsorizés diffèrent peu avec les seuils de winsorization. Aux niveaux sectoriels les plus agrégés, leurs précisions sont quasiment identiques, compte tenu du calage sur marges opéré sur le chiffre d'affaires fiscal par groupe. Des différences peuvent exister, à des niveaux de diffusion très fins, par exemple sur le total du chiffre d'affaires dans certaines sous-classes très fortement winsorizées avec les seuils issus de la base de sondage.

Au final, les seuils obtenus avec la base de sondage nous semblent correspondre le mieux aux objectifs de la winsorization, *i.e.* identifier un nombre restreint d'unités atypiques posant problème et dont le traitement améliore la précision des estimateurs.

Les seuils calculés avec les données de l'enquête à laquelle ils sont appliqués conduisent à winsorizer beaucoup d'entreprises et de strate de tirage, mais avec des ajustements faibles à chaque fois. L'effet final sur la précision des estimateurs est cependant proche de ce qui est obtenu avec les autres seuils respectant les hypothèses de la méthode de Kokic et Bell.

Bibliographie

[1] Brion P., « L'utilisation combinée de données d'enquête et de données administratives pour la production des statistiques structurelles d'entreprise », *Actes des Journées de Méthologie Statistique*, 2009
[2] Brion P., Guggemos F., « Du bon usage de la winsorization… ou comment traiter les entreprises atypiques dans les enquêtes sectorielles annuelles », *Lettre du SSE*, n°65, septembre 2010
[3] Brion P., Gros E., Guggemos F., « La gestion des unités influentes dans l'ESA par winsorization », *Séminaire de Méthodologie Statistique de l'Insee*, Séance du 2 juillet 2013 : Le traitement des unités influentes dans les enquêtes, 2013
[4] Caron N., « La correction de la non-réponse par repondération et par imputation », *Document de travail de l'Insee*, n°M0502, 2005

[7] C'est à dire, hors des sections B, C, D et E, qui correspondent au champ couvert par l'EAP.

[5] Chambers R., « Outlier robust finite population estimation », *Journal of the American Statistical Association*, vol 81, n° 396, pp 1063-1069, december 1986
[6] Demoly E., Gros, E., Fizzala A., « Méthodes et pratiques des enquêtes entreprises à l'Insee », *Journal de la Société Française de Statistiques*, vol 155, n°4 pp 134-159, 2014
[7] Gros E., « Esane, ou les malheurs de l'estimation composite : comment gérer les valeurs négatives d'estimateurs par différence ? », *Actes des Journées de Méthodologie Statistique*, 2009
[8] Guggemos F., « Winsorization dans les enquêtes annuelles auprès des entreprises françaises », *Actes du sixème colloque francophone sur les sondages (Tanger)*, 2010
[9] Kokic P.N., Bell P.A., « Optimal winsorizing cutoffs for a stratified finite population estimator », *Journal of Official Statistics*, vol 10, n° 4, pp 419-435, 1994

Annexe A : Calcul du biais optimal dans la méthode de Kokic et Bell

Soit $Y_j^h = (\frac{N_h}{n_h} - 1)(\tilde{X}_j^h - \mu_h)$. Les $(Y_j^h)_{h=1..H,\, j=1..p_h}$ sont classés par ordre croissant : soit $Y_{(k)}$ la k[ème] valeur et soit *p* le nombre total d'observations[8].

La fonction $\hat{F}(L) = L\left[1 + \sum_h \frac{n_h}{p_h} \sum_{j=1}^{p_h} I(Y_j^h > L)\right] - \sum_h \frac{n_h}{m_h} \sum_{j=1}^{p_h} Y_j^h I(Y_j^h > L)$ est une fonction croissante de *L*. Quand *L* varie entre deux valeurs successives de $Y_{(k)}$, les indicatrices $I(Y_j^h > L)$ restent constantes, égales à 0 ou 1, suivant que le rang de Y_j^h est inférieur ou supérieur à *k* : la fonction $\hat{F}$ est donc affine par morceaux, et continue par morceaux, avec des sauts quand *L* = $Y_{(k)}$. $\hat{F}(Y_{(1)})$ est de plus négatif, et $\hat{F}(Y_{(p)})$ est positif.

Il est donc possible d'encadrer le zéro de $\hat{F}$ en calculant les différentes valeurs de $\hat{F}(Y_{(k)})$ pour *k* variant de *1* à *p*. Le zéro est entre k et k+1 tels que $\hat{F}(Y_{(k)}) < 0$ et $\hat{F}(Y_{(k+1)}) > 0$. Le biais optimal est ensuite estimé pa interpolation linéaire : c'est l'abscisse du point de la droite reliant les points $(Y_{(k)}, \hat{F}(Y_{(k)}))$ et $(Y_{(k+1)}, \hat{F}(Y_{(k+1)}))$ dont l'ordonnée est nulle.

Plus précisément, $\hat{B}^* = \frac{Y_{(k+1)}\hat{F}(Y_{(k)}) - Y_{(k)}\hat{F}(Y_{(k+1)})}{\hat{F}(Y_{(k)}) - \hat{F}(Y_{(k+1)})}$

[8] $p = \sum_{h=1}^{H} p_h$