Champ et hors champ : impact de la démographie des entreprises. Cas de l'enquête Acemo-TPE

Ludovic VINCENT

Dares, Sous-direction salaires, travail et relation professionnelle

Résumé

Pour mener une enquête statistique, il est primordial de définir précisément la thématique et le champ de son enquête. Malheureusement, les sources disponibles pour constituer la base de sondage ne permettent pas toujours de se limiter exactement à la population d'intérêt.
Certaines unités sélectionnables se retrouvent alors dans l'échantillon, et pourtant hors du champ de l'enquête. Ce problème existe particulièrement pour les enquêtes dont le champ est défini par des seuils, comme l'enquête Acemo-TPE, administrée auprès des entreprises de 1 à 9 salariés.
Lors de la collecte, des unités interrogées se signalent hors champ, permettant de déceler facilement, parmi les répondants, ceux concernés par l'enquête des autres. Il est en revanche plus compliqué de statuer sur les unités qui n'ont pas retourné le questionnaire.
L'utilisation de sources externes peut apporter de nouvelles informations, inconnues au moment de la constitution de la base de sondage, et ainsi permettre de distinguer parmi ces non retours, les unités du champ des unités hors champ. Certaines demeureront cependant encore sans statut. On envisagera alors de faire appel à des méthodes statistiques de repondération par exemple pour que les unités dont on connait le statut représentent ces unités restées indécises.
Si le traitement du hors champ n'a pas toujours d'impact sur les résultats produits, son ampleur doit obliger le statisticien à réfléchir à son traitement.

Mots-clés

hors-champ, échantillonnage, biais, sondage

Introduction

Lors de la conception d'une enquête statistique, deux éléments doivent être précisément définis dès le début : la problématique étudiée et la population cible. L'objectif est alors de capter au mieux cette population cible afin de collecter l'information voulue. Or, bien souvent, la constitution de la base de sondage se révèle plus complexe que prévu, et ne correspond pas exactement au champ désiré. En effet, S'il parait aisé de définir les critères d'appartenance au champ de l'enquête, il est bien plus difficile de les traduire en terme pratique pour construire la base de sondage dans laquelle sera tiré l'échantillon de l'enquête : certaines unités sont ainsi exclues à tort du champ de l'enquête, d'autres se retrouvent, également à tort, dans la base de sondage.

Cet article se propose d'établir l'impact théorique sur les estimateurs de la sélection d'unités hors du champ de l'enquête dans l'échantillon, puis de décrire une méthode pour limiter cet impact.

L'enquête Activité et Condition d'Emploi de la Main-d'Oeuvre dans les très petites entreprises (Acemo-TPE) porte sur les entreprises (ou unités légales) de 1 à 9 salariés. L'existence de seuil et la forte démographie de ces entreprises ainsi que la variabilité de leur effectif salarié font que la constitution d'une base de sondage « parfaite » est impossible. Le problème d'unité dite « hors champ » dans l'échantillon y est ainsi très prégnant. Cette enquête apparaît ainsi comme un terrain pratique pour expérimenter les méthodes proposées pour le traitement du Hors champ, et voir leur apport et leurs limites.

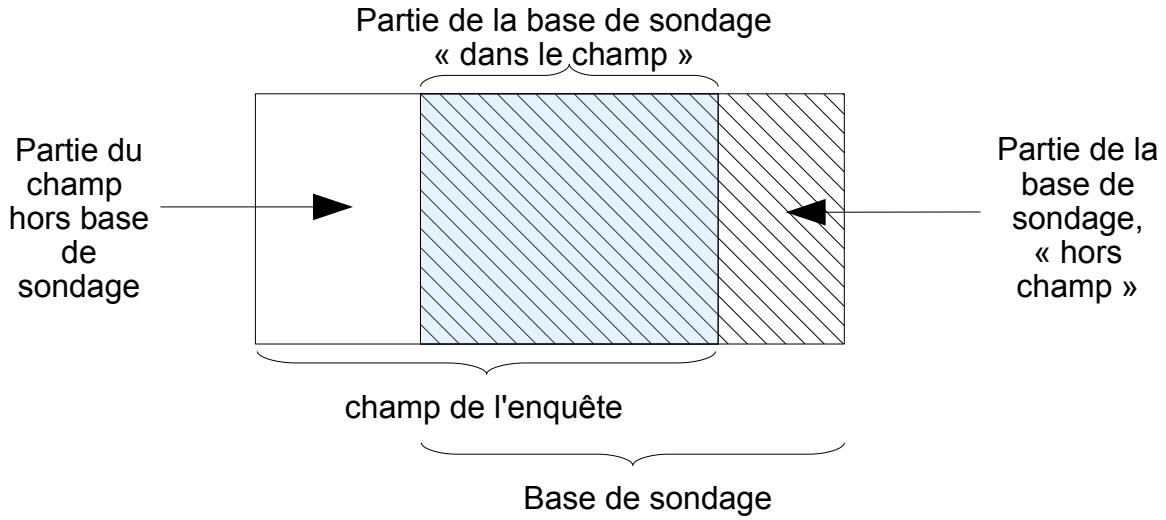
1. Champ et Hors champ – définition et origine du problème

Par définition, est considérée comme unité hors du champ d'une enquête toute unité statistique dont les caractéristiques ne relèvent pas de celles définissant le champ considéré.
Ainsi, l'enquête patrimoine va-t-elle s'intéresser aux ménages résidant dans un logement ordinaire en France, et aux individus de ce ménage.
L'enquête Acemo-TPE, quant à elle, va s'intéresser aux entreprises de France métropolitaine, du secteur marchand non agricole ayant 1 à 9 salariés au 31 décembre de l'année précédent l'enquête.

Pour mener à bien l'enquête, il va donc s'agir de déterminer précisément la population répondant à ces caractéristiques. Or, lors de la constitution de la base de sondage, les inexactitudes des informations des différentes unités génèrent trois populations :

- la population d'unités appartenant au champ désiré et présente dans la base de sondage,
- une population d'unités appartenant au champ de l'enquête, mais non captée lors de la construction de la base,
- une population d'unités considérée à tort dans le champ de l'enquête, dite hors champ.

Figure 1 : Représentation du champ d'une enquête et de sa base de sondage



Ainsi, l'enquête patrimoine 2010 a-t-elle recensé, après collecte, 11,6% des logements sélectionnés comme hors champ, car ne correspondant pas, finalement, à des résidences principales.

Plusieurs causes sont à l'origine de cet écart entre base de sondage et population d'intérêt. En premier lieu, c'est l'imperfection des bases de données elles-mêmes à disposition du statisticien qui génère ce problème. Le champ peut être défini par des variables dont les valeurs sont très volatiles dans le temps. Il peut être alors difficile pour les unités elle-même d'en donner une valeur exacte à une date donnée. De plus, la mise à jour des bases de données peut être lente et ainsi ne pas permettre de connaître la véritable valeur au moment de la constitution de la base de sondage (mais n'avoir qu'une information plus ancienne).

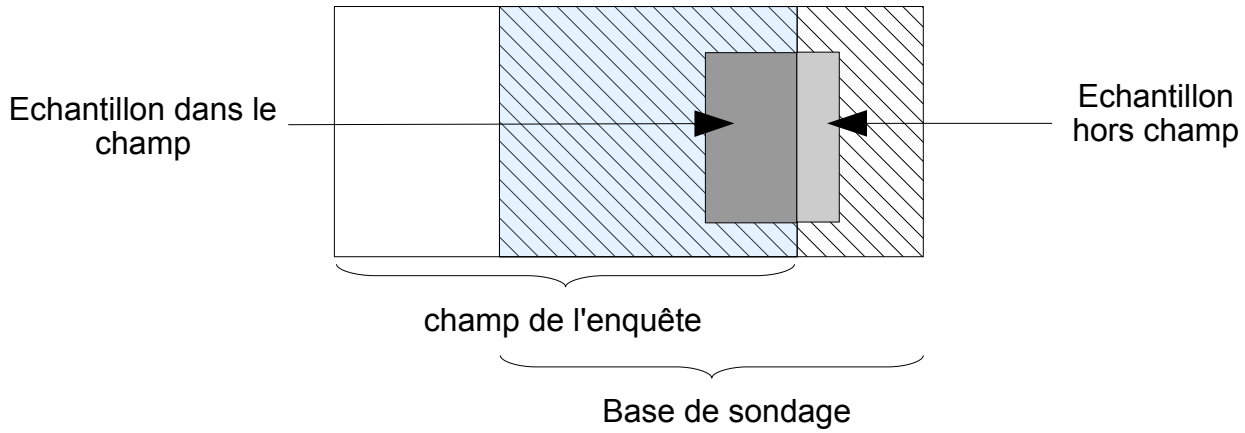
En second lieu, des problèmes de date peuvent engendrer un écart entre population cible et population d'intérêt. Ainsi, les caractéristiques définissant le champ d'une enquête font référence à des dates, et les bases de sondage sont bien souvent construites antérieurement à cette date de référence ; on ne peut donc connaître parfaitement la réalité à la date de référence. Par exemple, pour l'enquête Acemo trimestrielle, enquête de conjoncture, les questionnaires sont envoyés aux unités enquêtées quelques jours avant la date de référence (fin du trimestre). La base de sondage doit ainsi être figée en amont de la date de référence.
L'écart temporel entre collecte et date de référence (ou date de constitution de la base de sondage) engendre également des erreurs de champ (impossibilité de capter des entreprises du champs, erreur de déclaration...),

Pour éviter tout hors champ, il faudrait ainsi connaître exactement la situation de tous les individus de la population à la date de référence, et être en mesure d'interroger toutes les unités de l'échantillon. Ce qui n'est pas réaliste.

2. Biais sur les estimateurs

Si la base de sondage contient des unités hors champ, il est vraisemblable que, lors du tirage d'échantillon, certaines d'entre elles se retrouver interrogées.

Figure 2 : Représentation d'un échantillon par rapport au champ d'une enquête et à sa base de sondage



Faute de pouvoir détecter les unités du champ de celle hors champ, dans l'échantillon, l'indicateur produit ne portera pas sur le champ de l'enquête, mais sur la base de sondage. Pour peu que la variable d'intérêt prenne des valeurs très différentes de part et d'autre de la frontière du champ, l'indicateur se trouvera fortement biaisé.

De même, la partie du champ de l'enquête, absente de la base de sondage ne pourra être représentée par l'échantillon. Un second biais est ainsi généré, bien plus délicat à matérialiser, puisqu'il ne peut être estimé par l'échantillon tiré. Par la suite, nous supposerons que toutes les unités appartenant au champ de l'enquête sont contenues dans la base de sondage.

Biais généré par la partie « hors champ » de la base de sondage

Notations utilisées :

Espaces et variable
- U, la population de la base de sondage
- s, un échantillon tiré dans U
- I_k, l'indicatrice de présence de l'individu k dans l'échantillon s
- Y, une variable d'intérêt (un total) sur la population U
- y_i , la valeur de la variable d'intérêt pour chaque unité i de la population U.

Par la suite, on indicera les notations :
- par C pour les données relatives à la population du champ,
- par HC pour les données relatives à la population hors champ,
- par R pour les répondants (ou retournant),
- par NR pour les non retour,
- par indécis pour les indécis.

Poids et probabilités

- N le nombre d'unités de U (N_C, celui de U_C)
- n le nombre d'unités de s (n_C, celui de s_C)
- n_R le nombre d'unités répondantes de s (n_{R_C}, celui de s_C)
- π_i, la probabilité d'inclusion de l'individu i dans l'échantillon s

Notations mathématiques :

- $\hat{Y}$ est un estimateur de Y ,
- E(Y) est l'espérance de Y,
- Biais($\hat{Y}$) est le biais de l'estimateur $\hat{Y}$,
- V($\hat{Y}$) est la variance de $\hat{Y}$.

Nous considérerons, par simplification, que le plan de sondage est un tirage aléatoire simple.

La valeur de la variable Y sur la population U s'écrit :

$$Y = \sum_{i \in U} y_i \qquad (1)$$

Un estimateur de ce total est l'estimateur de Horvitz Thompson (que l'on notera HT) :

$$\hat{Y}_{HT} = \sum_{i \in s} \frac{1}{\pi_i} * y_i \qquad (2)$$

Il est sans biais, c'est à dire que la moyenne du total estimé sur l'ensemble des échantillons de taille n[1] est égale à la valeur de Y :

$$E(\hat{Y}_{HT}) = E\left(\sum_{k \in U} \frac{1}{\pi_k} * I_k * y_k\right) = \sum_{k \in U} \frac{1}{\pi_k} * E(I_k) * y_k = \sum_{k \in U} y_k = Y \qquad (3)$$

En considérant qu'une partie de la population est hors champ, la valeur de Y s'écrit :

$$Y = \sum_{i \in U} y_i = \sum_{i' \in U_C} y_{i'} + \sum_{i'' \in U_{HC}} y_{i''} = Y_C + Y_{HC} \qquad (4)$$

L'estimateur de HT peut se décomposer en :

$$\hat{Y}_{HT} = \sum_{i \in s} \frac{1}{\pi_i} * y_i = \sum_{i' \in s_C} \frac{1}{\pi_{i'}} y_{i'} + \sum_{i'' \in s_{HC}} \frac{1}{\pi_{i''}} y_{i''} = \hat{Y}_C + \hat{Y}_{CH} \qquad (5)$$

La valeur de Y sur la population totale est différente de la valeur de Y sur la population d'intérêt (Y_C), et l'estimateur HT ne permet pas de donner une valeur « sans biais[2] » de Yc.

Le « biais » peut ainsi s'écrire :

$$Biais(\hat{Y}_{HT}) = E(\hat{Y}_{HT}) - Y_C = \sum_{k' \in U_C} \frac{1}{\pi_{k'}} y_{k'} * E(I_{k'}) + \sum_{k'' \in U_{HC}} \frac{1}{\pi_{k''}} y_{k''} * E(I_{k''}) - Y_C$$

$$Biais(\hat{Y}_{HT}) = \sum_{k' \in U_C} y_{k'} + \sum_{k'' \in U_{HC}} y_{k''} - Y_C = Y_C + Y_{HC} - Y_C$$

$$Biais(\hat{Y}_{HT}) = Y_{HC} \qquad (6)$$

Le biais de l'estimateur correspond donc à la valeur de la variable d'intérêt dans la population hors champ.

Comment régler le problème ?

Pour que le résultat produit ne comporte pas ce biais, on souhaite que les unités hors champ de l'échantillon

1 Il s'agit de l'espérance de l'estimateur. En réalité, nous le verrons par la suite, l'estimateur de Horvitz-Thompson est sans biais si tous les individus ont une probabilité de sélection non nulle.

2 C'est-à-dire, dans ce contexte, tel que le total sur le champ est bien estimé.

n'influe pas dans la variable estimée.
Par exemple, une solution est d'arriver à ne pas sélectionner ces unités dans l'échantillon. En d'autre terme, pour avoir un estimateur « sans biais » à partir de la population initiale, la probabilité d'inclusion doit être telle que :

- les individus hors champ ne peuvent être dans l'échantillon,
- les individus du champ ont tous la même chance d'être dans l'échantillon[3].

Soit[4],

- $\pi'_i = 0 \quad \text{si } i \notin U_{HC}$
- $\pi'_i = \alpha \quad \text{si } i \in U_{HC} \quad \text{avec } \alpha = \frac{n_C}{N_c}$

D'après (5), l'estimateur de HT s'écrit alors[5] :

$$\hat{Y}_{HT} = \sum_{i \in s} \frac{1}{\pi'_i} * y_i = \sum_{i' \in s_C} \frac{1}{\pi'_{i'}} y_{i'} \quad \text{car la probabilité d'inclusion dans l'échantillon des unités hors champ est nulle}$$

$$\hat{Y}^{bis}_{HT} = \sum_{i \in s_C} \frac{1}{\pi'_i} * y_i = \frac{N_C}{n_C} \sum_{i' \in s_C} y_{i'} = \hat{Y}_C \qquad (7)$$

L'estimateur de HT sur l'ensemble de la population (base de sondage) se trouve biaisé du fait de la probabilité nulle de tirer certains individus[6]. Mais il estime sans biais la population du champ, c'est-à-dire la population dont on exclut les individus de probabilité de sélection nulle.

En d'autre terme, s'il était possible de ne sélectionner dans l'échantillon que des individus du champ, l'estimateur estimerait **sans biais** la variable d'intérêt **sur la population cible**.

Cependant, n'échantillonner que les unités du champ parmi la population revient à être en capacité de construire une base de sondage sans hors champ.
Une autre solution pour que les unités sélectionnées n'interviennent pas dans le calcul est de considérer nulle la variable d'intérêt pour chaque unité hors champ. Ainsi, Y_{HC} est nulle, et l'estimateur d'Horvitz Thompson se retrouve « sans biais » :

$$Biais(\hat{Y}_{HT}) = \sum_{k' \in U_C} y_{k'} + \sum_{k'' \in U_{HC}} y_{k''} - Y_C = \sum_{k' \in U_C} y_{k'} + 0 - Y_C = 0 \qquad (8)$$

On écrira donc notre estimateur ainsi[7] :

$$\hat{Y}^{*}_{HT} = \sum_{i \in s} \frac{1}{\pi_i} y_i = \sum_{i' \in s_C} \frac{1}{\pi_{i'}} y_{i'} + \sum_{i'' \in s_{HC}} \frac{1}{\pi_{i''}} y_{i''} = \frac{N}{n} \sum_{i' \in s_C} y_{i'} + \frac{N}{n} \sum_{i'' \in s_{HC}} 0 = \hat{Y}_C \qquad (9)$$

Ce traitement du problème sous-entend toutefois une hypothèse : il est possible, au moment du calcul de l'estimateur, de séparer exhaustivement dans l'échantillon les unités du champ des unités hors du champ de l'enquête. Cela ne pouvant être réalisé au moment de la constitution de la base de sondage, une information supplémentaire est nécessaire.
Si les réponses à l'enquête peuvent permettre de lever le doute, l'exploitation s'expose à un nouveau problème, celui de la non réponse. En effet, parmi les non-répondants, certains n'appartiennent pas au champ de l'enquête.
On ne connaît pas l'appartenance au champ pour tous les individus de l'échantillon (valeur spécifique de y

3 Bien prendre en compte le hors champ serait repondérer l'échantillon « dans le champ » de telle sorte qu'il représente exactement la population du champ. A population de taille égale, la précision en serait affectée (échantillon plus petit pour représenter la même population), mais l'indicateur ne serait pas biaisé sur la population d'intérêt.

4 Il s'agit à ce stade d'estimer la probabilité de réponse conditionnellement à l'appartenance au champ.

5 On retrouve ici une tirage stratifié, en considérant comme strate les « hors champ » et les « dans le champ ».

6 L'estimateur proposé est justifié par l'hypothèse que le comportement de réponse à la variable y diffère sensiblement entre échantillon dans le champ et hors champ. A défaut, l'estimateur classiquement corrigé de la non-réponse serait aussi bon sinon meilleur (car l'estimation de la probabilité de réponse serait effectuée sur un échantillon plus grand)

7 En écrivant cela, on suppose également que notre échantillon est représentatif dans les mêmes proportions du champ et du hors champ. On se ramène alors à un sondage stratifié avec taux de sondage égale dans chaque strate.

mise à nulle pour les hors champ) ; on ne peut clairement établir la part de l'échantillon appartenant au champ, mais seulement celle des répondants.

En prenant en compte tous ces paramètres, la valeur du total peut s'écrire :

$$Y = \sum_{k' \in U_C} y_{k'} + \sum_{k'' \in U_{HC}} y_{k''} + \sum_{l' \in U_{NR\_C}} y_{l'} + \sum_{l'' \in U_{NR\_HC}} y_{l''} \qquad (10)$$

Un estimateur de ce total sera :

$$\hat{Y} = \sum_{i \in s} \frac{1}{\pi_i} y_i = \sum_{i' \in s_{R_C}} \frac{1}{\pi_{i'}} y_{i'} + \sum_{i'' \in s_{HC}} \frac{1}{\pi_{i''}} y_{i''} + \sum_{j' \in s_{NR\_C}} \frac{1}{\pi_{j'}} y_{j'} \qquad (11)$$

3. Proposition pour le traitement du hors champ

3.1. Exemple de détermination du hors champ dans des enquêtes

Pratiquement toutes les enquêtes sont confrontées au problème du hors-champ. Si certaines le négligent complètement (et ainsi prennent le risque de biaiser les résultats), la plupart dénombrent les unités hors du champ, et les excluent de l'exploitation. Cette pratique revient à considérer comme nulle la valeur de la variable d'intérêt parmi le hors champ ; seuls les éléments du champ participent au calcul du total estimé.

Par exemple, dans les enquêtes ESA, les unités hors champ sont ainsi définies : Sont présumées cessées – ou hors champ – les unités non affiliées aux régimes d'imposition micro, sans liasse fiscale pendant trois années successives et n'ayant effectué aucune déclaration de TVA l'année d'intérêt de l'enquête.

Pour beaucoup d'enquêtes ménages, la collecte en face-à-face permet de distinguer aisément champ et hors champ. Les enquêtes entreprises, plus souvent auto déclarée, utilisent des procédés variés.

Nombre de celles-ci utilisent des sources externes comme Sirus (répertoire d'entreprises) ou les DADS (déclaration annuelle des données sociales), non disponibles ou non consolidées au moment de la constitution de la base de sondage.

C'est le cas de l'enquête TIC qui combine l'utilisation du fichier Sirus et des déclarations de TVA pour déterminer, lors de l'exploitation, les unités hors champ.

Les enquêtes annuelles d'entreprises (EAE) avaient recours a une enquête plus légère (les enquêtes d'amélioration du répertoire – EAR) pour statuer notamment sur le caractère cessé ou non des unités interrogées.

3.2. Démarche pour le traitement du hors champ

Nous allons ici proposer une démarche permettant un traitement efficient du hors champ.

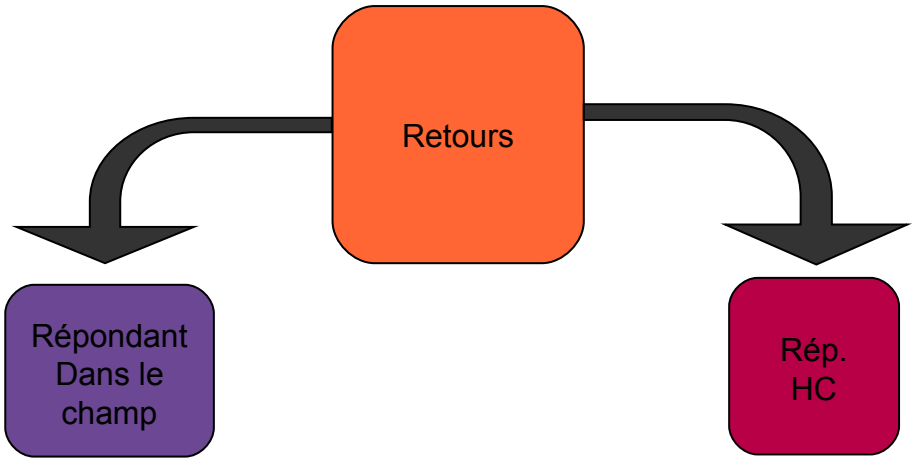
Le hors champ prenant sa source dans l'écart entre la base de sondage et la population d'intérêt, la première étape consiste à tout mettre en œuvre pour limiter cet écart.
Il est ainsi important de définir les contours du champ de l'enquête, grâce à des conditions précises sur des variables fiables et mesurables.

Par exemple, définir le champ de l'enquête TPE comme l'ensemble des entreprises en France métropolitaine, appartenant au SMNA et dont l'effectif total est de 1 à 9 salarié ne suffit pas : il manque une date de référence.
A l'inverse, l'enquête sur le commerce électronique ne peut être menée en réduisant le champ aux seules entreprises effectuant ce genre de commerce, car il sera impossible d'établir une base de sondage à partir

d'une variable suffisamment fiable. Il faudra donc l'élargir en gardant en tête qu'un traitement futur du hors champ s'avérera nécessaire.
Enfin, il faut tenter de construire une base de sondage la plus à jour possible au regard des dates de l'enquête.
Cette problématique est assez forte, par exemple, pour le recensement de la population qui voit certains logements disparaître, et d'autre apparaître entre le tirage de l'échantillon et la date de référence du 1er janvier suivant.

En cours de collecte, et après retour des questionnaires, le traitement du hors champ passe par l'apurement des répondants.
L'objectif ici est de déterminer, parmi les répondants, les unités appartenant au champ des unités hors champ.

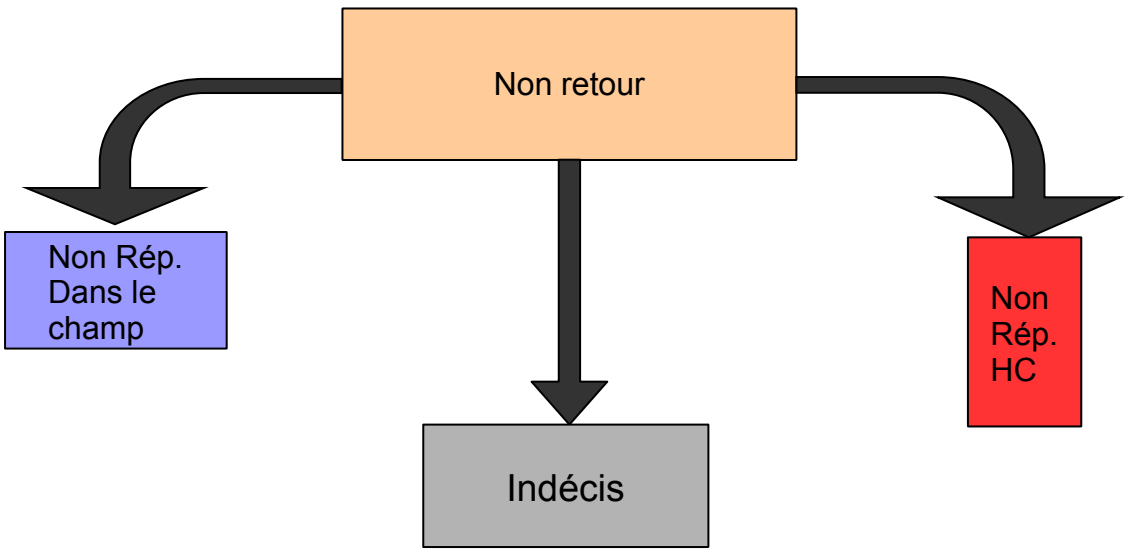
Figure 3 : Séparation des questionnaires retournés en deux catégories : champ et hors champ.



Pour statuer sur les questionnaires retournés, il est nécessaire d'utiliser une information inconnue au moment du tirage. L'utilisation d'information supplémentaire auxiliaire ou les réponses internes au questionnaires peuvent permettre d'effectuer cette distinction.

Pour le traitement des non répondants, l'idée est au départ la même que pour les répondants : distinguer les unités appartenant au champ des unités hors champ.
Or si l'utilisation d'une nouvelle source externe peut permettre de faire cela, on ne peut utiliser les questionnaires. Il est alors impossible de statuer sur certaines unités de l'échantillon, nommés « indécis » par la suite.

Figure 4 : Séparation des questionnaires non retournés en trois catégories : champ, hors champ et indécis



Après traitement de ces répondants et non-répondants à l'aide de diverses sources externes, l'enquête Innovation 2001/2002 fait part de 512 cas d'indécision parmi les non-répondants, c'est-à-dire d'unités dont on ne peut déterminer l'appartenance au champ.

L'approche présentée ici propose de déterminer statistiquement le caractère « hors champ » ou dans le champ de ces unités en prédisant une variable dichotomique « champ de l'enquête ».

Quatre méthodes sont proposées.

Tout d'abord, les méthodes que l'on peut dire radicales :
- tous les indécis sont hors champ (Méthode i)
- tous les indécis appartiennent au champ (Méthode ii).

Les deux options peuvent trouver leur justification.
Dans le premier cas, le manque de connaissance des unités interrogées et leur non réponse peut conduire à penser qu'elles ne répondent pas au critère de sélection de l'échantillon. Pour une entreprise, il peut ainsi s'agir d'une cessation d'activité.
Dans le second cas, l'appartenance au champ peut se justifier par la présence de l'unité dans l'échantillon : si elle y est, ce n'est pas par hasard. La probabilité est donc non nulle qu'elle appartienne au champ de l'enquête.

Les deux autres méthodes proposées sont intermédiaires.
La première consiste à repondérer les unités au statut connu de telle sorte qu'elles représentent les unités indécises (Méthode iii).
Dans la seconde méthode, il s'agit, à l'aide d'un modèle économétrique, de statuer sur le caractère hors champ ou non de l'unité, en construisant ce modèle à partir des unités non indécises (Méthode iiii).

Le traitement du hors champ se révèle donc très proche du traitement de la non réponse à la question « appartenance au champ ».

Reprenons l'écriture (10) de la variable d'intérêt en y ajoutant une population indécise.
En prenant en compte tous les paramètres, la valeur du total peut s'écrire :

$$Y = \sum_{k' \in U_C} y_{k'} + \sum_{k'' \in U_{HC}} y_{k''} + \sum_{l' \in U_{NR\_C}} y_{l'} + \sum_{l'' \in U_{NR\_HC}} y_{l''} + \sum_{l''' \in U_{indec}} y_{l'''} \qquad (11')$$

Un estimateur de ce total sera :

$$\hat{Y} = \sum_{i \in s} \frac{1}{\pi_i} y_i = \sum_{i' \in s_{R\_C}} \frac{1}{\pi_{i'}} y_{i'} + \sum_{i'' \in s_{R\_HC}} \frac{1}{\pi_{i''}} y_{i''} + \sum_{j' \in s_{NR\_C}} \frac{1}{\pi_{j'}} y_{j'} + \sum_{j'' \in s_{NR\_HC}} \frac{1}{\pi_{j''}} y_{j''} + \sum_{j''' \in s_{indec}} \frac{1}{\pi_{j'''}} y_{j'''} \qquad (12)$$

Dans les faits, on peut considérer un indécis comme un non répondant pour la variable « champ ». Vu de cette manière, les différentes méthodes de traitement de la non réponse peuvent permettre de corriger la représentativité des unités de l'échantillon afin de prendre en compte les indécis.
Nous décidons ici de présenter une méthode de repondération de l'échantillon pour le traitement de cette « non réponse » à la variable « champ ». Un estimateur du total Y pourra ainsi être noté :

$$\hat{Y}_{indec} = \sum_{i \in s} \frac{1}{\pi_i} y_i = \sum_{i' \in s / s_{indécis}} \frac{1}{\pi_i} y_i + \sum_{i'' \in s_{indécis}} \frac{1}{\pi_{i''}} y_{i''} = \sum_{i' \in s / s_{indécis}} \frac{N}{n} * \frac{n}{n - n_{indécis}} y_i = \sum_{i' \in s / s_{indécis}} \frac{N}{n - n_{indécis}} y_i \qquad (13)$$

Cette méthode revient à faire l'hypothèse que la répartition champ/hors champ est la même parmi les indécis que parmi ceux dont le statut est connu.
L'estimateur utilisé est alors considéré comme étant sans biais.

En appliquant cette formule à la formule (12) qui prenait en compte la non réponse, on obtient l'estimateur suivant pour notre total sur la population :

$$\hat{Y}_{final} = \sum_{i \in s} \frac{1}{\pi_i} y_i = \sum_{i' \in s_{R\_C}} \frac{1}{\pi_{i'}} y_{i'} + \sum_{i'' \in s_{R\_HC}} \frac{1}{\pi_{i''}} y_{i''} + \sum_{j' \in s_{NR\_C}} \frac{1}{\pi_{j'}} y_{j'} + \sum_{j'' \in s_{NR\_HC}} \frac{1}{\pi_{j''}} y_{j''} + \sum_{j''' \in s_{indec}} \frac{1}{\pi_{j'''}} y_{j'''} \qquad (12)$$

$$\hat{Y}_{final}=\sum_{i\in s}\frac{1}{\pi_i}y_i=\sum_{i'=1}^{n_{R\_C}}\frac{N}{n}y_{i'}+\sum_{j'=1}^{n_{NR\_C}}\frac{N}{n}y_{j'}+\sum_{j''=1}^{n_{HC}}\frac{N}{n}y_{j''}+\sum_{j'''=1}^{n_{indec}}\frac{N}{n}y_{j'''}$$ *hors champ répondant et non répondant sont regroupés* (14)

$$\hat{Y}_{final}=\sum_{i'=1}^{r_C}\frac{N}{n-n_{indécis}}y_{i'}+\sum_{j'=1}^{n_{NR\_C}}\frac{N}{n-n_{indécis}}y_{j'}+\sum_{j''=1}^{n_{HC}}\frac{N}{n-n_{indécis}}y_{j''}$$ *traitement des indécis* (15)

$$\hat{Y}_{final}=\sum_{i'=1}^{n_{R\_C}}\frac{N}{n-n_{indécis}}*\frac{n_C}{n_{R\_C}}y_{i'}+\sum_{j''=1}^{n_{HC}}\frac{N}{n-n_{indécis}}*1*y_{j''}$$ *traitement des non répondants* (16)

$$\hat{Y}_{final}=\sum_{i'=1}^{n_{R\_C}}\frac{N}{n-n_{indécis}}*\frac{n_C}{n_{R\_C}}y_{i'}+\sum_{j''=1}^{n_{HC}}\frac{N}{n-n_{indécis}}*0$$ *traitement des hors champ* (17)

$$\hat{Y}_{final}=\sum_{i'=1}^{n_{R\_C}}\frac{N}{n-n_{indécis}}*\frac{n_C}{n_{R\_C}}y_{i'}$$ (18)

Au final, la formule montre que pour traiter les indécis, non répondants et hors champ :

- on repondère l'ensemble des individus dont on connaît le statut pour prendre en compte les indécis,
- puis on repondère les répondants du champ pour prendre en compte les non répondants du champ,
- enfin, on met à « 0 » la valeur de la variable d'intérêt pour les individus hors champ.

Par construction, cet estimateur peut être considéré « sans biais ».
En effet, il peut s'écrire :

$$\hat{Y}_{final}=\sum_{i'=1}^{n_{R\_C}}\frac{N}{n'}*\frac{n_C}{n_{R\_C}}y_{i'}+\sum_{j''=1}^{n_{HC}}\frac{N}{n'}*1*y_{i'}\quad\text{avec}\quad n'=n_{R\_C}+n_{NR\_C}+n_{HC}=n-n_{indécis}$$ [8]

On retrouve alors un estimateur sans biais sur la population totale, et considéré sans biais sur la population d'intérêt en émettant l'hypothèse de y_i nulle pour tous les individus hors champ.

La variance pourrait s'écrire de manière simplifiée :

$$Var\left(\hat{Y}_{final}\right)=N^2\frac{1-\frac{n'}{N}}{n'}S^2+N^2\frac{1-\frac{n_{R\_C}}{n_C}}{n_{R\_C}}S^2$$

où S² est la variance empirique modifiée de Y sur la population.

Ainsi, plus le nombre d'indécis est élevé, plus la variance est forte ; plus le nombre de répondants du champ est faible par rapport au nombre d'individu du champ, plus la variance est forte.
Le fait de traiter les hors champ permet cependant de rapporter le nombre de répondants au nombre de hors champ, ce qui fait diminuer la variance[9].

Il est intéressant de mettre cet estimateur en regard d'un estimateur classique n'incluant pas le traitement du hors champ :

$$\hat{Y}^*_{final}=\sum_{i'=1}^{r}\frac{N}{r}y_{i'}$$ (19)

Si aucune unité est hors champ et aucune indécise, alors :

$n_{indécis}=0$

$n_{R\_C}=r$ et $n_C=n$

Ainsi :

$$\hat{Y}_{final}=\sum_{i'=1}^{n_{R\_C}}\frac{N}{n-n_{indécis}}*\frac{n_C}{n_{R\_C}}y_{i'}=\sum_{i'=1}^{n_R}\frac{N}{n}*\frac{n}{r}y_{i'}=\sum_{i'=1}^{n_R}\frac{N}{r}y_{i'}=\hat{Y}^*_{final}$$ (20)

8 Une fois les indécis traités, tous les hors champ sont considérés comme répondants, puisque leur valeur de y est fixée nulle.

9 Par contre, l'estimateur de la variance empirique pourrait être dégradé si le nombre de répondants du champ est trop faible (même s'il correspond à l'ensemble des entreprises interrogées).

4. Application à l'enquête TPE

4.1. Description de l'enquête

L'enquête Acemo TPE a pour objectif de compléter les autres enquêtes Acemo sur le champ des entreprises de 1 à 9 salariés. La partie principale du questionnaire repose sur un tableau permettant de détailler, pour chaque salarié de l'entreprise, la durée du contrat, le type de contrat (apprentissage, alternance...), la durée hebdomadaire et s'il est bénéficiaire des revalorisations du SMIC.
Depuis 2013, l'enquête contient un module quadriennal tournant permettant d'obtenir de l'information sur diverses thématiques et ainsi de compléter la connaissance de la statistique publique dans les petites structures de l'économie française. Les modules concernent le dialogue social en entreprise, la formation professionnelle ou encore l'épargne salariale.

L'enquête permet également de compléter l'information obtenue par l'enquête Acemo trimestrielle sur les emplois vacants et l'application de convention collective.

L'enquête interroge près de 55 000 entreprises[10], parmi les quelques 1,2 millions d'entreprises de 1 à 9 salariés du secteur marchand non agricole en France Métropolitaine.
Elle couvre trois millions de salariés parmi les 23 millions de salariés de l'économie en France métropolitaine.

L'échantillon est tiré par l'Insee à partir du nouveau répertoire statistique Sirus (répertoire servant de base de sondage à de nombreuses enquêtes auprès des entreprises). L'enquête est panélisée, et chaque année, un quart de l'échantillon de l'enquête Acemo TPE est renouvelé.
Le plan de sondage est stratifié selon un croisement de la taille d'entreprise en quatre tranches[11] et du secteur d'activité en 88 postes.
Le nombre d'unités tirées est déterminé selon l'allocation de Neyman améliorée, afin d'obtenir une précision maximale dans chaque strate de l'estimation du nombre de salariés, à partir de l'échantillon.

La collecte débute généralement fin mars par voie postale[12]. Il est demandé de répondre dans les six semaines. Un rappel est envoyé deux mois plus tard, avec un délai de réponse d'un mois.
La collecte est fermée à la fin de l'été, pour une exploitation pendant l'automne et une diffusion des résultats en fin d'année.

4.2. Définition du champ de l'enquête et constitution de la base de sondage

Les enquêtes Acemo couvrent pratiquement l'ensemble du secteur marchand non agricole en France Métropolitaine. Leur champ est défini à partir d'une restriction sur certaines activités principales exercées (APE) et certaines catégories juridiques des unités légales.

Ainsi, les enquêtes Acemo couvrent l'ensemble des salariés de France Métropolitaine des entreprises du secteur privé de l'économie française. Sont exclues les entreprises appartenant aux secteurs de l'agriculture, de la pêche et de la sylviculture, à l'administration publique, à l'action sociale relevant de la loi de 1901, aux activités des ménages et les activités extraterritoriales.

Pour l'enquête TPE, on restreint en plus ce champ aux unités légales de 1 à 9 salariés. Les intérimaires et stagiaires ne sont pas comptés parmi les effectifs salariés, contrairement aux apprentis et emplois aidés.

Le champ est ainsi clairement défini et repose sur des variables plutôt fiables que sont les secteurs d'activités des entreprises et l'effectif salariés au 31/12.

Cependant, la base de sondage est issu de Sirus, qui ne permet pas, au moment de la constitution de la base (février), d'avoir précisément les effectifs au 31/12. De plus, la collecte a lieu 3 à 4 mois après la date de référence de l'enquête, écart assez important, ce qui peut générer des erreurs de champ pour les unités interrogées.

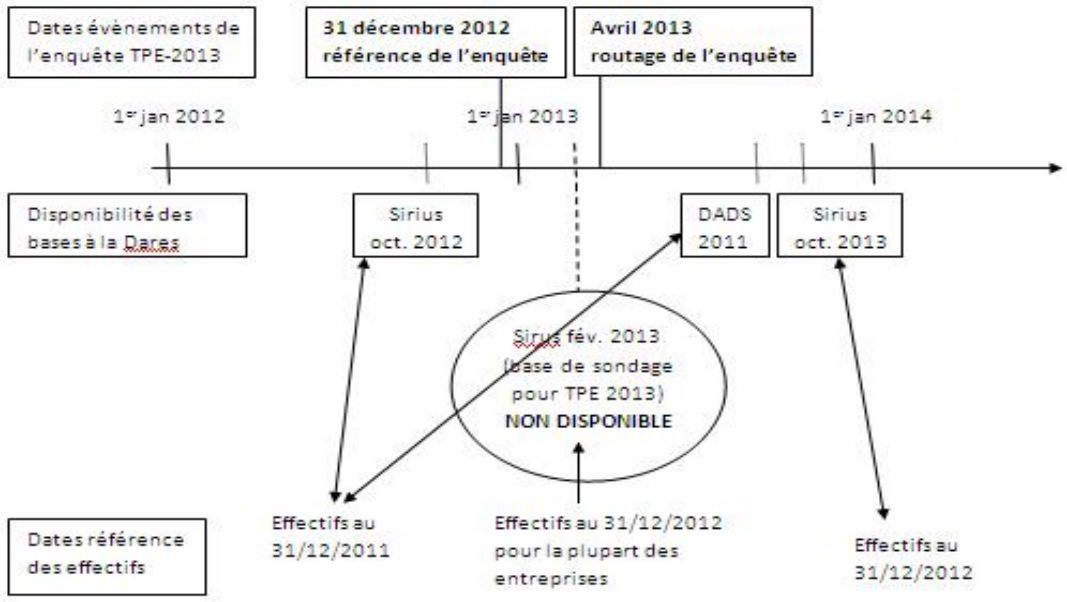
10 On entend par entreprise « Unité légale »
11 Les quatre tranches de taille sont : 1 salarié, 2 salariés, 3 à 5 salariés et 6 à 9 salariés.
12 La collecte par internet débutera en 2017.

4.3. Apurement des données

L'exploitation de l'enquête a lieu 6 mois après le début de la collecte. L'apurement des données peut se faire grâce au questionnaire lui même qui permet de détecter l'effectif salarié à la date de référence, variable à la base de la quasi totalité du hors champ dans l'enquête.
D'autre part, le délai entre collecte et exploitation permet d'obtenir des informations supplémentaires qui permettront de statuer sur nombre de questionnaires non retournés.

Le schéma suivant montre les bases de données à disposition de la Dares entre le tirage d'échantillon et l'exploitation de l'enquête.

Figure 5 : Bases de données à disposition de la Dares entre tirage d'échantillon et exploitation de l'enquête

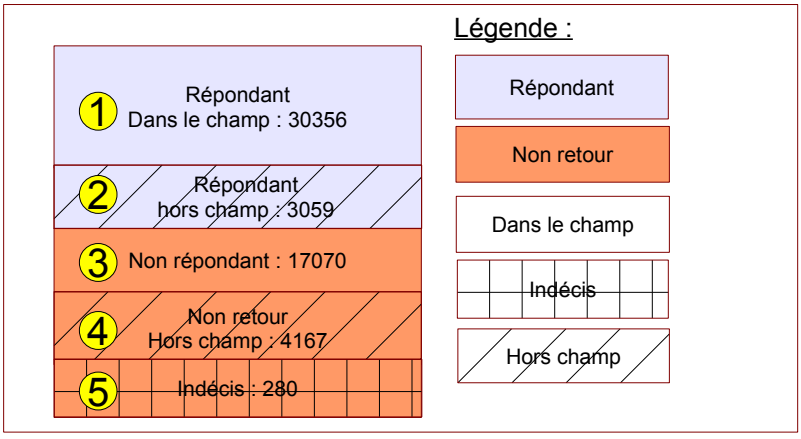


Nous avons ainsi choisi de considérer hors champ :

- les unités légales présentent dans Sirus en 2012 et pas en 2013 ;
- les unités légales ayant un effectif nul ou > 9 dans Sirus en 2013.

Au final, l'échantillon peut se décomposer en cinq catégories :

Figure 6 : Répartition de l'échantillon TPE2013 en cinq catégories : répondants du champ, répondants hors champ, non répondants, non retour hors champ et indécis (après détection des indécis)



Source : Acemo-TPE 2013

On peut ainsi constater que les unités considérées hors champ dans l'échantillon sont non négligeables puisqu'elles représentent presque 10 % des répondants, et 20% des non retour.
A l'inverse, les critères pris pour définir non retour du champ et non retour hors champ limite à un très faible pourcentage les indécis.

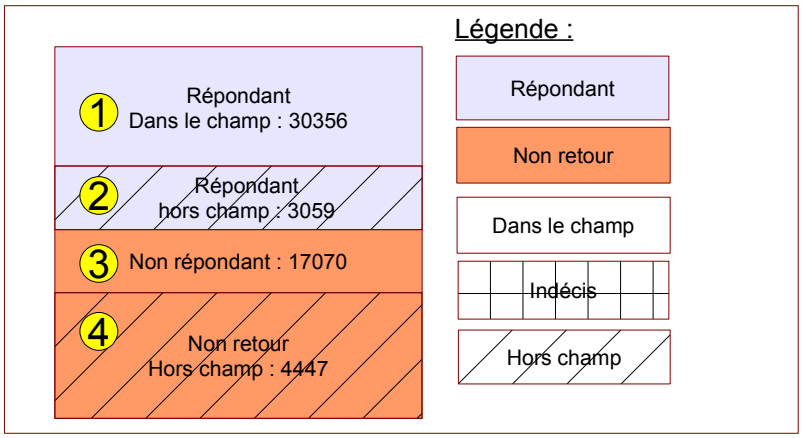
4.4. Application de méthodes pour le traitement des indécis

Du fait du faible taux d'indécis, les quatre méthodes proposées n'auront qu'un impact limité sur les statistiques produites. Nous nous proposons d'en détailler deux :

- tous les indécis sont hors champ (Méthode i)
- repondération de l'ensemble des non indécis pour représenter les indécis (Méthode iii).

Méthode i : On considère l'ensemble des indécis comme des hors champ. Cette possibilité peut se justifier par l'absence d'informations possédées. La répartition de l'échantillon devient alors :

Figure 7 : Répartition de l'échantillon TPE2013 entre répondants du champ, répondants hors champ, non répondants, non retour hors champ et indécis, après traitement des indécis – Méthode i



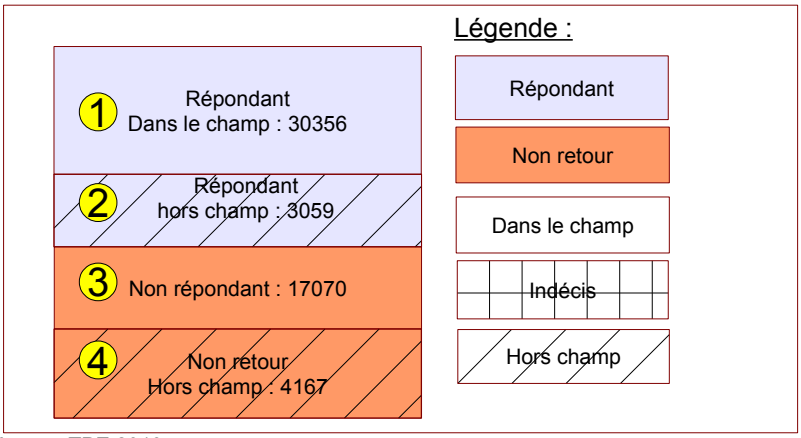
Source : Acemo-TPE 2013

On repondère alors les 30 356 répondants pour qu'ils représentent les 17 070 non répondants du champ. C'est avec cette option, que la part de hors champ est la plus importante (n'appartient au champ que ce qu'on a pu « prouver » y être).

Méthode iii : On traite les indécis, par repondération des quatre premières catégories. En augmentant le poids des unités au statut connu, on représente celles au statut inconnu.

Nous repondérons dans chaque strate les entreprises dans le champ et hors champ en affectant un coefficient constant par strate. Ce coefficient correspond classiquement au nombre d'entreprises de la strate rapporté au nombre d'entreprises au statut connu.
L'échantillon se trouve alors diminué des 280 indécis.

Figure 8 : Répartition de l'échantillon TPE2013 entre répondants du champ, répondants hors champ, non répondants, non retour hors champ et indécis, après traitement des indécis – Méthode iii



Source : Acemo-TPE 2013

Le poids de chaque catégorie en est légèrement modifié.

Tableau 1 : Modification des poids par traitement des indécis par la méthode iii

	Somme des poids hors indécis, avant correction	Somme des poids hors indécis, après correction
Hors champ	135151	136383
Champ	886338	891336

Source : Acemo-TPE 2013

La somme des poids des indécis, avant correction est de 6 230.

Cette repondération, par construction, modifie très peu la répartition pondérée champ/hors champ ; dans le cas de TPE et avec les options prises, c'est la distinction « Non répondant – non retour hors champ » qui est la plus impactante.

Le tableau suivant montre l'impact de chacune de ces deux méthodes par rapport aux données publiées :

Tableau 2 : Estimation du taux de temps partiel et du taux de bénéficiaires de la revalorisation du Smic au 1er janvier 2013 par taille d'entreprises en fonction de la méthode de repondération choisie, sur les TPE

		Résultats publiés	Résultat avec la pondération i pour le hors champ	Résultat avec la pondération iii pour le hors champ
taux de temps partiel	1 salarié	41,2%	41,1%	41,1%
	2 salariés	36,8%	36,9%	36,9%
	3 à 5 salariés	30,2%	30,3%	30,3%
	6 à 9 salariés	23,9%	22,9%	22,9%
	Ensemble	30,26%	30,09%	30,10%
taux de bénéficiaires de la revalorisation du Smic au 1er janvier 2013	1 salarié	36,0%	35,9%	35,9%
	2 salariés	33,3%	33,1%	33,1%
	3 à 5 salariés	27,8%	27,8%	27,8%
	6 à 9 salariés	22,2%	21,7%	21,7%
	Ensemble	27,61%	27,53%	27,52%

source : Acemo-TPE2013

Les deux estimations proposées font état de 30,1% des salariés des entreprises de 1 à 9 salariés qui travaillent à temps partiel. Quelque soit le secteur (en 17 postes) et la taille de l'entreprise (en quatre tranches), les deux pondérations ne se distinguent jamais l'une de l'autre de plus de cinq centièmes. Même dans une répartition plus précise de la nomenclature d'activité (38 postes), les différences les plus fortes sont de l'ordre du dixième. À l'inverse, à ce niveau plus détaillé, traiter le hors champ modifie les taux publiés

(sans correction du hors champ). On constate d'ailleurs que les plus gros écarts avec les résultats publiés se situent à la frontière des 9 salariés, atteignant jusqu'à 1% de différence.

Le même constat peut être fait sur la statistique donnant la proportion de bénéficiaires de la revalorisation du Smic. Si l'indicateur est très proche entre les deux pondérations corrigeant le hors champ, des écarts sont à noter par rapport aux indicateurs publiés. Et une nouvelle fois, la différence la plus importante se situe au niveau des entreprises de 6 à 9 salariés.

Prendre en compte le hors champ dans la pondération des questionnaires exploités ne change pas le message porté par la publication faite à partir de l'enquête TPE-2013 mais modifie légèrement les chiffres publiés, particulièrement à la frontière du champ.

Conclusion

Toute enquête statistique est confrontée au hors champ : il est en effet utopique d'espérer une base de sondage reflétant exactement la réalité.
Dans certaines enquêtes, ce phénomène est d'ailleurs conséquent, et il ne peut être ignoré, au risque de biaiser fortement les résultats produits.
L'objectif du procédé décrit ici n'est donc pas d'annuler le hors champ, mais de le restreindre et de le prendre en compte de la constitution de la base de sondage à l'exploitation de l'enquête, afin de diminuer son effet. La méthode propose de modéliser le hors champ afin de prédire l'état des unités échantillonnées pour lesquelles on ne peut aisément statuer.
Afin de limiter plus encore ce cas, la prédiction du hors champ dès la base de sondage permettrait de limiter l'interrogation à tort d'unités dans l'échantillon.

Bibliographie

1 - BUISSON B., *« Champ et hors-champ : le cas des enquêtes thématiques entreprises »,* N°67 - série La lettre du SSE, 2011.
2 - BRION P., CARON N., PIETRI-BESSY P., *« Redresser la non-réponse totale dans les enquêtes auprès des entreprises : les pièges à éviter - Illustration avec l'enquête innovation »* Journées Mondiales de la statistique, 2005.
3 - ARDILLY P., *Les techniques de sondage,* 2e édition Paris : Éditions Technip, 2006.
4 - BRILHAULT G. et CARON N., *« Correction de la non-réponse totale : par imputation ou par repondération ? »*, Document de travail de la Direction des Statistiques d'Entreprises de l'INSEE N°E2004/01, 2004.
5 - CARON N., *« Les principales techniques de correction de la non-réponse et les modèles associés »*, Document n°9604 - série "Méthodologie Statistique" de l'INSEE, 1996.
6 - HAZIZA D., *« Traitement de la non-réponse dans les enquêtes",* Polycopié de la formation continue diplômante des attachés, 2008.
7 - KOUBI M., MATHERN S., *« La nouvelle méthode d'échantillonnage de l'enquête trimestrielle ACEMO depuis 2006 - Amélioration de l'allocation de Neyman»*, Document d'étude n°146 (2009).