ÉTUDES ET SIMULATIONS RELATIVES À LA NOUVELLE MÉTHODE DE COORDINATION DES ÉCHANTILLONS D'ENQUÊTES ENTREPRISES ET ÉTABLISSEMENTS DÉVELOPPÉE A L'INSEE

Emmanuel GROS[1] (*)

(*) Insee, Direction de la méthodologie et de la coordination statistique et internationale

Résumé

Le système statistique public réalise chaque année un nombre important d'enquêtes auprès des entreprises et des établissements. L'objectif de la coordination négative d'échantillons est de favoriser, lors du tirage d'un échantillon, la sélection d'entreprises n'ayant pas déjà été sélectionnées lors d'enquêtes récentes, tout en conservant le caractère sans biais des échantillons. Elle s'inscrit donc dans une démarche de réduction de la charge statistique imposée aux petites entreprises – les grandes entreprises, à partir d'un certain seuil, étant systématiquement enquêtées dans la plupart des enquêtes. La coordination positive vise quant à elle à maximiser le recouvrement entre les échantillons coordonnés, soit dans une optique de panélisation, soit une fois encore dans l'objectif d'une réduction de la charge statistique par le biais cette fois-ci d'une réduction de la taille des questionnaires.

Depuis fin 2013, une nouvelle méthode de coordination des échantillons des enquêtes auprès des entreprises est utilisée en production à l'Insee. Cette méthode – qui fait partie de la famille des procédures de coordination s'appuyant sur l'utilisation de numéros aléatoires permanents – repose sur la notion de fonction de coordination. Ces fonctions, définies pour chaque unité et chaque nouveau tirage en fonction des charges de réponse aux enquêtes passées, transforment les nombres aléatoires permanents des unités de façon à assurer la sélection d'un échantillon qui satisfasse à l'objectif de coordination, en général négative, recherché tout en respectant le plan de sondage voulu.

Après avoir rappelé les fondements théoriques de cette méthode de coordination dans le cas du tirage aléatoire simple stratifié, le présent article aborde différentes problématiques spécifiques à la coordination d'échantillon – biais de rétroaction, incompatibilité de la procédure avec l'algorithme de tirage systématique, coordination entre échantillons de niveaux différents – et présente les résultats de nombreuses simulations effectuées sur données réelles pour évaluer les performances de cette procédure de coordination.

Abstract

The public statistical system carries out each year a significant number of businesses and establishments surveys. The objective of the negative coordination of samples is to foster, when selecting a sample, the selection of businesses that have not already been selected in recent surveys, while preserving the unbiasedness of the samples. This coordination contributes to reduce the statistical burden of small businesses – large businesses, from a certain threshold, are systematically surveyed in most surveys.

Since late 2013, a new sampling coordination method is used at Insee. This method, using Permanent Random Numbers (PRN) assigned to each unit, is based on the notion of *coordination function*, defined for each unit and each new drawing, which transforms permanent random numbers.

[1] emmanuel.gros@insee.fr

This paper presents the main principles of this method in the case of stratified simple random sampling, broaches some specific issues related to the coordination methods – feedback bias, incompatibility with systematic sampling, sample coordination between surveys based on different kind of units – and reports the results of many simulations assessing the properties of this coordination method.

Mots-clés

Coordination d'échantillons, numéros aléatoires permanents, fonction de coordination, charge de réponse, échantillons stratifiés.

1. La méthode de coordination d'échantillons d'enquêtes auprès des entreprises de l'Insee

On présente ici les fondements théoriques de la méthode en se limitant au cas du sondage aléatoire simple stratifié, qui est l'échantillonnage le plus souvent utilisé à l'Insee pour les enquêtes-entreprises. Cette méthode, proposée par Christian Hesse dans [1] et étudiée par Pascal Ardilly dans [2], est exposée plus en détail dans [3] par Fabien Guggemos et Olivier Sautory.

Remarque : cette partie théorique est reprise *in extenso* – modulo quelques amendements mineurs – de l'article d'Olivier Sautory [4].

1.1. Fonction de coordination – Sélection des échantillons

Le concept de fonction de coordination joue un rôle essentiel dans la méthode :

> Une fonction de coordination g est une application mesurable de [0,1[dans [0,1[qui conserve la loi uniforme ; elle a donc pour propriété de conserver la longueur des intervalles – et des réunions d'intervalles – par image réciproque.

On attribue à chaque unité k de la base de sondage un nombre aléatoire permanent ω_k, tiré dans la loi de probabilité uniforme sur [0,1[. Les tirages des ω_k sont indépendants les uns des autres.

On considère une succession d'enquêtes t = 1, 2,… (t désigne à la fois la date et le numéro de l'enquête). On suppose que l'on a défini pour chaque unité k une fonction de coordination $g_{k,t}$, « judicieusement choisie » (voir § 1.2) qui change à chaque enquête t.

L'échantillon S_t correspondant à l'enquête t, obtenu par un sondage aléatoire simple stratifié, est obtenu de la façon suivante : dans chaque strate (h,t) de taille $N_{(h,t)}$, on sélectionne les $n_{(h,t)}$ unités correspondant aux $n_{(h,t)}$ plus petites valeurs $g_{k,t}(\omega_k)$, k = 1… $N_{(h,t)}$.

Démonstration

> Les $N_{(h,t)}$ nombres aléatoires (ω_k) associés aux $N_{(h,t)}$ unités de la strate ayant été tirés indépendamment dans la loi de probabilité uniforme sur [0,1[, notée P, les $N_{(h,t)}$ nombres $g_{k,t}(\omega_k)$ sont eux-mêmes tirés indépendamment dans la loi P, en raison de la propriété des fonctions de coordination, et les n plus petites valeurs $g_{k,t}(\omega_k)$ donnent bien un échantillon aléatoire simple de taille $n_{(h,t)}$ dans la strate.

1.2. Construction d'une fonction de coordination à partir d'une fonction de charge

1.2.1. Charge de réponse cumulée et fonction de coordination

On note $\mathbf{\Omega} = (\Omega_1,\ldots, \Omega_N)$ le vecteur aléatoire dont la réalisation est le vecteur $\boldsymbol{\omega} = (\omega_1,\ldots, \omega_N)$ composé des N nombres aléatoires ω_k associés aux unités k de la population.

On note $I_{k,t}(\mathbf{\Omega})$ l'indicatrice d'appartenance de l'unité k à l'échantillon S_t, égale à 1 si les valeurs de $\boldsymbol{\omega}$ conduisent à sélectionner l'unité k, et 0 sinon : il s'agit d'une variable aléatoire, dépendant du vecteur $\mathbf{\Omega}$.

On note $\gamma_{k,t}$ la charge de réponse « potentielle »[2] d'une unité k pour l'enquête t[3]. La charge de réponse effective est donc une variable aléatoire $\gamma_{k,t}(\mathbf{\Omega}) = \gamma_{k,t} I_{k,t}(\mathbf{\Omega})$, et la charge de réponse cumulée sur toutes les enquêtes de 1 à t est une fonction de $\mathbf{\Omega}$ égale à :

$$\Gamma_{k,t}(\mathbf{\Omega}) = \sum_{u \le t} \gamma_{k,u}.I_{k,u}(\mathbf{\Omega}) \quad (1)$$

On souhaite définir, pour chaque unité k, une fonction de coordination $g_{k,t}$ fondée sur $\Gamma_{k,t-1}$, la charge cumulée de l'unité k jusqu'à l'enquête t-1. Pour répondre à l'objectif de coordination négative (***sélectionner en priorité, pour un tirage donné, les unités qui ont eu la plus faible charge de réponse dans le passé***), et compte tenu du mode de sélection des unités choisi (***la probabilité qu'une unité soit sélectionnée est d'autant plus élevée que $g_{k,t}(\omega_k)$ est petit***), une propriété souhaitée pour les fonctions de coordination est la suivante :

$$\Gamma_{k,t-1}(\boldsymbol{\omega}^{(1)}) < \Gamma_{k,t-1}(\boldsymbol{\omega}^{(2)}) \Rightarrow g_{k,t}(\omega_k^{(1)}) \le g_{k,t}(\omega_k^{(2)})$$

où $\boldsymbol{\omega}^{(1)}$ et $\boldsymbol{\omega}^{(2)}$ désignent deux réalisations du vecteur $\mathbf{\Omega}$, et $\omega_k^{(i)}$ (i=1,2) la $k^{ème}$ composante du vecteur $\boldsymbol{\omega}^{(i)}$.

Cette condition n'est pas facile à manipuler, car la charge cumulée $\Gamma_{k,t}(\mathbf{\Omega})$ est une fonction du vecteur $\mathbf{\Omega}$, i.e. non seulement du nombre aléatoire Ω_k associé à l'unité k, mais de tous les autres nombres aléatoires. Nous verrons plus loin comment on peut la remplacer par une fonction $\Gamma'_{k,t}(\Omega_k)$ qui dépende uniquement du nombre aléatoire Ω_k. La propriété attendue pour une fonction de coordination $g_{k,t}$ s'écrit alors :

$$\Gamma'_{k,t-1}(\omega_k^{(1)}) < \Gamma'_{k,t-1}(\omega_k^{(2)}) \Rightarrow g_{k,t}(\omega_k^{(1)}) \le g_{k,t}(\omega_k^{(2)}) \quad (2)$$

où $\omega_k^{(1)}$ et $\omega_k^{(2)}$ désignent deux réalisations du nombre aléatoire Ω_k.

1.2.2. Construction d'une fonction de coordination

On omet les indices k et t, pour simplifier les notations. Ainsi ω désigne un réel compris entre 0 et 1. On note C la fonction de charge, supposée mesurable bornée : $\omega \in [0,1[\rightarrow C(\omega) \in \mathbb{R}$. On veut lui associer une fonction de coordination g telle que :

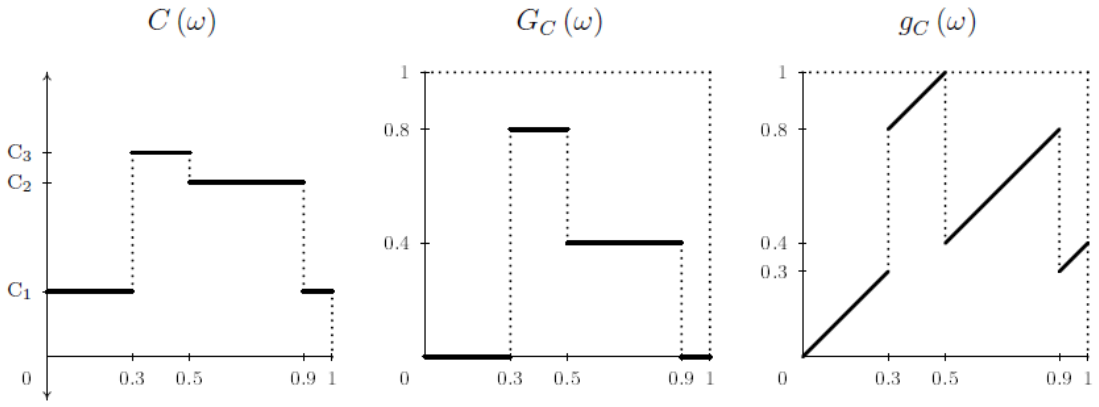
$$C(\omega^{(1)}) < C(\omega^{(2)}) \Rightarrow g(\omega^{(1)}) \le g(\omega^{(2)}) \quad (2')$$

On définit la fonction $G_C = F_C(C)$, où F_C désigne la fonction de répartition de C :

$$\forall \omega \in [0,1] \quad G_C(\omega) = P\left(u \middle| C(u) < C(\omega)\right)$$

[2] i.e. si elle est interrogée (et si elle répond…)
[3] que l'on supposera souvent identique pour toutes les unités pour une enquête donnée

La fonction G_C, qui a une image incluse dans [0,1[, vérifie la propriété (2'), mais n'est pas une fonction de coordination, dès que la fonction C possède des « paliers », i.e. des sous-intervalles de [0,1[où C est constante (G_C possède alors les mêmes paliers).
On peut alors construire une *fonction de coordination* g_C, égale à G_C en dehors des paliers, et constituée de morceaux de fonctions affines de pente 1 sur les paliers de G_C, comme l'illustre la figure suivante, où la fonction C est une fonction étagée ayant 4 paliers :



1.3. Mise en œuvre dans le cas d'un tirage aléatoire simple stratifié

Avec ce mode de tirage, on sélectionne une unité k dans l'échantillon S_t si le nombre aléatoire $g_{k,t}(\omega_k)$ figure parmi les n plus petits nombres $g_{i,t}(\omega_i)$ associés à toutes les unités i de la base de sondage[4]. L'inclusion de k dans S_t dépend donc de l'ensemble des nombres aléatoires ω_i de toutes les unités i de la base de sondage, et la fonction indicatrice $I_{k,t}$, de même que la charge cumulée $\Gamma_{k,t}$, sont des fonctions du vecteur $\boldsymbol{\Omega}$. Il est donc nécessaire de remplacer l'indicatrice $I_{k,t}$ par une indicatrice approchée $I^a_{k,t}$, qui lui sera proche tout en ne dépendant que de ω_k.

1.3.1. L'indicatrice approchée – La fonction de charge espérée

La meilleure approximation possible de la fonction indicatrice $I_{k,t}(\boldsymbol{\Omega})$ qui ne dépende que de Ω_k, au sens de la norme L_2, est son espérance conditionnelle par rapport à Ω_k :

$$I^a_{k,t}(\omega) = E\left(I_{k,t}(\boldsymbol{\Omega}) \,\middle|\, \Omega_k = \omega\right) = P\left(k \in S_t \,\middle|\, \Omega_k = \omega\right)$$

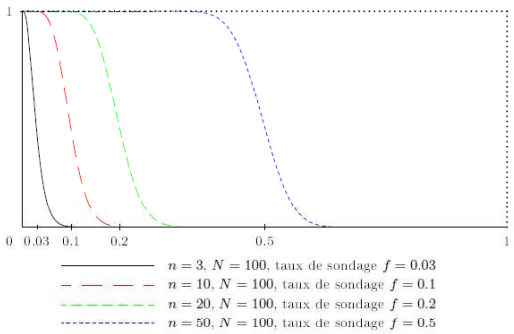
En supposant les fonctions de coordination bijectives[5], on montre que :

$$I^a_{k,t}(\omega) = P\left(k \in S_t \,\middle|\, g_{k,t}(\Omega_k) = g_{k,t}(\omega)\right) = b_{k,t}\left(g_{k,t}(\omega)\right)$$

où $1 - b_{k,t}(x)$ est la fonction de répartition de la loi beta de paramètres N et N – n + 1.

On peut voir dans les graphiques ci-dessous la forme de la fonction b(x) pour quelques valeurs de n et N.

[4] en omettant l'indice de strate.
[5] ce qui est vérifié dans la méthode présentée ici, mais qui n'est pas une propriété intrinsèque d'une fonction de coordination.



Une fonction b(x) a l'allure suivante : une première partie "presque horizontale" proche de 1 (sélection "presque certaine" de l'unité dans l'échantillon), une troisième partie "presque horizontale" proche de 0 (non-sélection "presque certaine" de l'unité). Entre les deux, une partie décroissante "à forte pente", correspondant à un intervalle sur l'axe des abscisses plus ou moins long, à peu près centré sur la valeur n/N, égale aux taux de sondage : c'est autour de cette valeur qu'il y a la plus grande incertitude sur la sélection ou non de l'unité dans l'échantillon.

Le remplacement de la fonction indicatrice par une indicatrice approchée fait que la fonction de charge cumulée est elle-même remplacée, dans l'expression (1) du §1.2.1, par une charge cumulée espérée $\Gamma^{e}_{k,t}$, conditionnellement à Ω_k : $\Gamma^{e}_{k,t}(\omega) = \sum_{u=1}^{t} \gamma_{k,u}\, I^{a}_{k,u}(\omega)$

Pour que la méthode mise en œuvre conduise à des échantillons sans biais, il est nécessaire d'utiliser cette charge espérée, et non la charge réelle, qui est fondée sur les inclusions observées de l'unité k dans les différents échantillons : $\Gamma_{k,t} = \sum_{u=1}^{t} \gamma_{k,u}\, \mathbb{1}(k \in S_u)$

1.3.2. Approximation par des fonctions étagées – Construction de la fonction de coordination

Les fonctions indicatrices approchées $I^{a}_{k,t}(\omega)$ et les fonctions de charge cumulée espérées $\Gamma^{e}_{k,t}$ ne sont pas des fonctions étagées, ni même des fonctions que l'on peut « calculer » facilement. On va simplifier la forme de la fonction indicatrice approchée $I^{a}_{k,t} = b_{k,t}$ de la façon suivante :

❶ On divise l'intervalle [0,1[en L[6] intervalles de longueurs égales $I_\ell = \left[\frac{\ell-1}{L}; \frac{\ell}{L}\right[$ $\ell = 1 \ldots L$.

❷ On remplace la fonction indicatrice approchée par une fonction affine par morceaux $\tilde{b}_{k,t}$ prenant les mêmes valeurs que $b_{k,t}$ aux extrémités des intervalles I_ℓ.

❸ On calcule la valeur moyenne $\beta_{k,t}(\ell)$ de $\tilde{b}_{k,t}$ sur chaque intervalle I_ℓ.

❹ On définit la fonction $\beta_{k,t}$ par : $\forall \omega \in I_\ell \quad \beta_{k,t}(\omega) = \beta_{k,t}(\ell)$.

$\beta_{k,t}$ est donc une approximation de la fonction indicatrice approchée $I^{a}_{k,t}$, sous la forme d'une fonction constante sur chaque intervalle I_ℓ. On en déduit la fonction de charge cumulée espérée « approchée » :

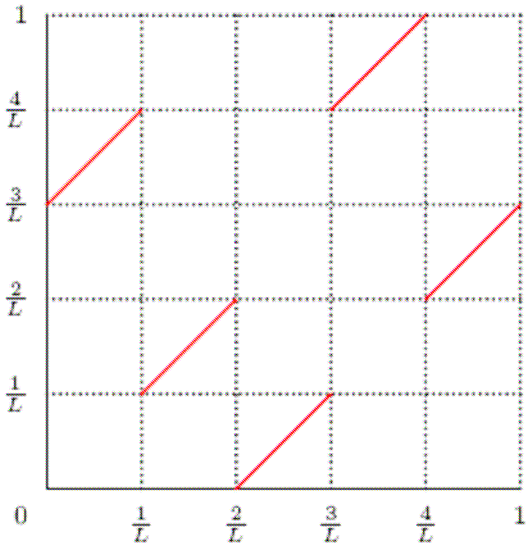
$$\Gamma^{ea}_{k,t}(\omega) = \sum_{u=1}^{t} \gamma_{k,u}\, \beta_{k,u}\left(g_{k,u}(\omega)\right)$$

$\Gamma^{ea}_{k,t}$, tout comme les fonctions $\beta_{k,u}$, est une fonction étagée, constante sur chaque intervalle I_ℓ.

1.3.3. Construction de la fonction de coordination

[6] L étant un nombre entier « assez élevé » (au moins supérieur à 50).

On est donc dans le même cas de figure que celui présenté dans l'exemple du §1.2.2. On déduit de la fonction $\Gamma^{ea}_{k,t}$ une fonction « G », également constante sur chaque intervalle I_ℓ, puis une fonction de coordination g, dont un exemple est présenté ci-contre, avec L = 5.



Elle est complètement définie par une permutation σ sur {1, 2, 3…L}.

Son expression est la suivante :

$$\forall \omega \in \left[\frac{\ell-1}{L};\frac{\ell}{L}\right[\quad g_\sigma(\omega) = \frac{\sigma(\ell)-1}{L} + (\omega - \frac{\ell-1}{L})$$

Il ne reste plus qu'à déterminer la permutation σ. Pour ce faire, il faut alors revenir à la propriété fondamentale (2') de la fonction de coordination : sa valeur est d'autant plus petite que le critère (la charge cumulée) est petit. Or, $g_\sigma(\omega)$ est d'autant plus petit sur $\left[\frac{l-1}{L};\frac{l}{L}\right[$ que $\sigma(\ell)$ est petit. On va donc classer les valeurs $\sigma(\ell)$ exactement comme on classe les valeurs de charge cumulée $\Gamma^{ea}_{k,t}(\ell)$. En notant ℓ_i l'identifiant de l'intervalle standardisé de rang i lorsqu'on trie les charges cumulées dans l'ordre croissant :

$$\Gamma^{ea}_{k,t}(\ell_1) \le \Gamma^{ea}_{k,t}(\ell_2) \le \ldots \le \Gamma^{ea}_{k,t}(\ell_L) \Leftrightarrow \sigma(\ell_1) \le \sigma(\ell_2) \le \ldots \le \sigma(\ell_L)$$

Comme σ doit être une permutation, donc bijective, on impose la règle supplémentaire suivante : si $\Gamma^{ea}_{k,t}(p) = \Gamma^{ea}_{k,t}(q)$ et $p < q$, alors $\sigma(p) < \sigma(q)$. *In fine*, ceci revient à imposer des inégalités strictes dans le classement des $\sigma(\ell)$, ce qui conduit à $\sigma(\ell_i) = i$ et définit complètement la permutation σ.

1.3.4. La procédure de sélection des échantillons – Cas séquentiel

À chaque étape, on se place au sein d'une strate donnée, dont on omet l'indice.

Sélection de l'échantillon S_1

On initialise la charge à 0 : $\forall$ k $\Gamma_{k,0}(\omega) = 0$ pour tout $\omega \in$ [0,1[.

Il n'y a aucune coordination à réaliser : on sélectionne les n unités correspondant aux n plus petites valeurs ω_i, i =1…N, ce qui revient à prendre comme fonction de coordination pour toute unité k l'identité sur [0,1[: $\forall$ k $g_{k,1}(\omega) = \omega$ pour tout $\omega \in$ [0,1[.

Pour toute unité k, la charge réelle vaut $\Gamma_{k,1} = \gamma_{k,1} \mathbb{1}(k \in S_1)$, en notant $\gamma_{k,1}$ sa charge de réponse pour l'enquête 1. Mais sa *fonction de charge espérée approchée* utilise la *fonction indicatrice approchée* $\beta_{k,1}$:

$$\Gamma^{ea}_{k,1}(\omega) = \gamma_{k,1}\beta_{k,1}(\omega)$$

Comme conséquence de la forme de la fonction $\beta_{k,1}$, commentée plus haut, les fonctions de charge réelle et espérée coïncideront en général sur l'intervalle [0,1[, sauf sur un voisinage de la valeur n/N ; ce

sont les unités de nombre aléatoire proche de n/N pour lesquelles l'appartenance à l'échantillon est *a priori* la plus incertaine.

Sélection de l'échantillon S_2

Pour chaque unité k, on utilise la fonction de charge espérée $\Gamma_{k,1}^{ea}$ comme « charge C » (au sens du §1.2.2) pour construire sa *fonction de coordination* $g_{k,2}$ pour le tirage du 2ème échantillon S_2, comme indiqué au §1.3.3.

On sélectionne les n unités correspondant aux n plus petites valeurs $g_{i,2}(\omega_i)$.

Pour toute unité k, sa *fonction de charge cumulée espérée approchée* après ce tirage vaut :

$$\Gamma_{k,2}^{ea}(\omega) = \Gamma_{k,1}^{ea}(\omega) + \gamma_{k,2}\ \beta_{k,2}\left(g_{k,2}(\omega)\right)$$

en notant $\gamma_{k,2}$ sa charge de réponse pour l'enquête 2, et $\beta_{k,2}$ la *fonction indicatrice approchée* correspondant à ce tirage.

Sélection de l'échantillon S_t

Plus généralement, pour la sélection de l'échantillon S_t, on construit pour chaque unité k sa *fonction de coordination* $g_{k,t}$, à partir de la *fonction de charge cumulée espérée approchée* $\Gamma_{k,t-1}^{ea}$.

On sélectionne les n unités correspondant aux n plus petites valeurs $g_{i,t}(\omega_i)$.

Pour toute unité k, sa *fonction de charge cumulée espérée approchée* après ce tirage vaut :

$$\Gamma_{k,t}^{ea}(\omega) = \Gamma_{k,t-1}^{ea}(\omega) + \gamma_{k,t}\ \beta_{k,t}\left(g_{k,t}(\omega)\right)$$

en notant $\gamma_{k,t}$ sa charge de réponse pour l'enquête t, et $\beta_{k,t}$ la *fonction indicatrice approchée* correspondant à ce tirage.

1.3.5. La procédure de sélection des échantillons – Cas non séquentiel

La méthode permet, pour une enquête donnée E_t, de coordonner le tirage de son échantillon S_t avec un ensemble de p enquêtes du passé, notées dans l'ordre chronologique (E^1, E^2, …, E^p), en ne prenant pas nécessairement en compte toutes les enquêtes réalisées entre la date t_1 de l'enquête E^1 et la date courante t.

❶ Pour chaque unité k, il faut calculer sa *fonction de charge cumulée espérée approchée* préalable à la construction de sa fonction de coordination :

$$\Gamma_{k}^{ea}(\omega) = \sum_{e=1}^{p} \gamma_{k,e}\,\beta_{k,e}\left(g_{k,e}(\omega)\right)$$

Il faut donc connaître, pour chaque enquête e :

- la charge de réponse $\gamma_{k,e}$ de l'unité k pour l'enquête e. Cette charge vaut 0 si l'unité k n'appartient pas au champ de l'enquête e (en particulier si elle a été créée après le tirage de l'enquête) ;
- l'indicatrice approchée $\beta_{k,e}$, qui est fonction de la taille de la strate à laquelle appartient l'unité k lors du tirage de l'enquête e, et de la taille de l'échantillon dans cette strate ;
- la fonction de coordination $g_{k,e}$, qui a été calculée au moment du tirage de l'enquête e.

Remarque : pour mettre en œuvre une telle coordination, il est donc nécessaire de conserver pour chaque enquête sa base de sondage, contenant en particulier l'identifiant h de strate, les informations relatives au plan de sondage (pour chaque strate sa taille N_h et l'allocation de l'échantillon n_h), et pour chaque unité appartenant au champ de l'enquête sa fonction de coordination, i.e. la permutation σ sur $\{1, 2, 3 \ldots L\}$ associée à cette fonction.

❷ On construit pour chaque unité k sa *fonction de coordination* $g_{k,t}$, à partir de la *fonction de charge cumulée espérée approchée* Γ_k^{ea}.

❸ On sélectionne les n unités correspondant aux n plus petites valeurs $g_{i,t}(\omega_i)$.

2. Évaluation de la procédure de coordination en situation de production et coordination entre échantillons de niveaux différents

Une première évaluation empirique de la méthode sur données simulées a été menée en 2012. Les résultats de ces simulations, exposés dans [3], se sont révélés très satisfaisants : la méthode de coordination s'est ainsi révélée à la fois très efficace – en conduisant à des gains considérables[7] en termes de répartition de la charge de réponse sur les différentes unités de la population – et remarquablement robuste vis à vis des paramètres des différents plans de sondage – taux de sondage, stratification, taux de recouvrement entre le champ des différentes enquêtes, charges associées aux différentes enquêtes, etc.

Cette première étude a été doublement complétée par les travaux de Kevin Rosamont-Prombo présentés dans [5] :

- d'une part, les tests sur données simulées ont été étendus et enrichis, et les résultats de ces nouvelles simulations sont venus confirmer ceux obtenus précédemment ;
- d'autre part, on a procédé à une première application de la méthode sur données réelles, à partir des bases de sondages relatives aux enquêtes TIC 2008 à 2012, dont les résultats se sont révélés aussi satisfaisants que ceux effectués sur données simulées. Ces premiers tests sur données réelles ont également permis de montrer que cette nouvelle méthode de coordination pouvait être utilisée de conserve avec la technique dite du « numéro hexal » utilisée pour la gestion des échantillons rotatifs dans les enquêtes entreprises (cf. [6] pour plus de détail sur cette procédure), et de comprendre et détailler les interactions entre ces deux méthodes et la façon dont elles s'imbriquent en cas d'utilisation simultanée.

Les simulations et résultats présentés dans la suite de cet article viennent enrichir et compléter les travaux précédemment menés selon différents axes :

- application de la procédure sur données réelles en grandeur nature afin de tester la faisabilité opérationnelle de la méthode et ses performances en situation de production ;
- test d'une procédure de coordination « multi-niveaux » permettant de coordonner simultanément des échantillons d'unités de niveaux différents (unités légales et établissements par exemple) ;
- simulations relatives à l'impact du paramètre de charge associé à chaque enquête sur la qualité de la coordination.

2.1. Test de la procédure en situation de production : tirages coordonnés de vingt enquêtes unités légales successives

[7] Par rapport à des tirages indépendants.

Afin de juger de la faisabilité opérationnelle de la méthode de coordination proposée, ainsi que de ses propriétés en situation de production en termes de répartition de la charge de réponse, nous avons procédé à une simulation sur données réelles en grandeur nature. La simulation a consisté :

- à partir de l'enquête sectorielle annuelle (ESA) de 2008, qui constitue ainsi l'enquête initiant la séquence de tirages coordonnés dans nos simulations ;
- à enchaîner ensuite, par ordre chronologique, le tirage de 19 autres enquêtes unités légales[8] :
  - en respectant systématiquement au mieux les plans de sondage mis en œuvre lors des tirages effectifs de ces enquêtes : critères de stratification et allocations, renouvellement par moitié, tiers ou quart de certains échantillons, coordination positive[9] d'une partie de l'échantillon de l'enquête « Points de vente » avec l'échantillon de l'ESA 2009, etc.
  - en **coordonnant** systématiquement le tirage de chaque échantillon **négativement avec l'ensemble des enquêtes passées**[10].

Une séquence de 20 tirages **indépendants** a également été réalisée, afin de pouvoir juger de la qualité de la procédure de coordination en termes de répartition de la charge d'enquête.

Du point de vue opérationnel, la méthode ne pose strictement aucun problème :

- les temps de calculs restent raisonnables : environ 8 heures pour la séquence complète de tirage des 20 enquêtes ;
- les besoins de stockage également : l'ensemble des tables stockant, pour chacune des 20 enquêtes, les permutations permettant de définir les fonctions de coordination nécessaires à la procédure occupe un espace d'environ 6 Go, et la plus grosse de ces tables, relative à l'ESA 2008, pèse environ 900 Mo ;
- la procédure de coordination se marie parfaitement en situation de production avec la technique du « numéro hexal » utilisée pour la gestion des échantillons rotatifs.

Du point de vue de la qualité statistique de la procédure, on observe, comme attendu, une bien meilleure répartition de la charge d'enquête entre les différentes unités de la population lorsque les tirages sont coordonnés. Le tableau 1 présente la distribution de la variable « charge d'enquête » – ici, le nombre d'échantillons auxquels une unité appartient – selon les deux scénarios de tirage indépendants et coordonnés. La coordination étant, à juste titre, sans effet sur les strates exhaustives[11], les parties exhaustives des échantillons ont été exclues des calculs de charge afin de pouvoir juger de la qualité de la procédure sur son champ d'action réel.

Charge d'enquête, hors exhaustifs	Fréquence selon le scénario de tirage retenu		Écarts entre les scénarios de tirages
	Tirages indépendants	Tirages coordonnés	
0	3 981 423	3 952 718	-28 705
1	257 692	290 783	33 091
2	126 430	136 787	10 357
3	34 542	27 012	-7 530
4	6 012	475	-5 537
5	1 500	38	-1 462
6	180	6	-174
7	39	0	-39
8	1	0	-1

Tableau 1 : distribution de la charge d'enquête, hors parties exhaustives, selon le scénario de tirage retenu.

[8] Il s'agit des enquêtes TIC 2010 à 2012, IPEA 2010 et 2011, Acemo-TPE 2010 à 2012, ESA 2009 à 2011, Points de vente 2010, SINE 2010, CVTS4, CIS 2010, ENDD 2011, Qualité énergétique mise en œuvre par les entreprises dans les bâtiments 2012, CAM 2012 et TIC-TPE 2012.
[9] Voir plus loin comment est mise en œuvre la coordination positive.
[10] Dans ces simulations, la charge attribuée à chaque enquête est constante et égale à 1.
[11] Strates contenant les unités incluses d'office dans l'échantillon.

Comme attendu, et dans la droite ligne des résultats obtenus lors des tests sur données simulées et sur les données réelles relatives aux seules enquêtes TIC, on observe un resserrement de la distribution autour de 1, i.e. un étalement de la charge d'enquête : le nombre d'unités interrogées plus de deux fois diminue dans des proportions importantes, de même que le nombre d'unité non échantillonnées, au profit d'une augmentation très nette du nombre d'unités sélectionnées dans une seule enquête, et dans une moindre mesure du nombre d'unités présentes dans deux échantillons. Ce dernier point, déjà observé dans les simulations sur données réelles relatives aux seules enquêtes TIC, découle de l'existence de parties conservées d'un millésime à l'autre pour les enquêtes à échantillons rotatifs – ESA, TIC, IPEA et Acemo-TPE. Si l'on exclut du calcul de la charge d'enquête les parties conservées de ces différents échantillons, les résultats sont encore plus parlants, comme le montre le tableau 2.

Charge d'enquête, hors exhaustifs et parties conservées	Fréquence selon le scénario de tirage retenu		Écarts entre les scénarios de tirages
	Tirages indépendants	Tirages coordonnés UL seules	
0	3 981 423	3 952 718	-28 705
1	391 840	445 402	53 562
2	30 494	9 084	-21 410
3	3 670	606	-3 064
4	374	9	-365
5	18	0	-18

Tableau 2 : distribution de la charge d'enquête, hors parties exhaustives et parties conservées, selon le scénario de tirage retenu

Notons par ailleurs que cette procédure de coordination permet également de procéder à une **coordination positive** entre deux enquêtes : il suffit pour cela d'affecter une **charge négative** à l'enquête que l'on souhaite coordonner positivement avec l'enquête que l'on tire. On a ainsi attribué une charge négative à l'ESA 2009 lors du tirage d'un sous-échantillon de l'enquête « Points de vente ». Les résultats obtenus en termes de recouvrement entre les deux échantillons sont très satisfaisants, légèrement supérieurs à ceux observés avec la méthode de coordination précédemment utilisée à l'Insee, fondée sur une autre technique.

Enfin, dans le tableau 2, le fait qu'un certain nombre d'unités restent sélectionnées dans plus d'un échantillon s'explique principalement par :

- la coordination positive d'un sous-échantillon de l'enquête « Points de vente » avec l'échantillon de l'ESA 2009 ;
- l'existence de strates avec des taux de sondage élevés dans certaines enquêtes.

Ainsi, sur les 9 084 unités présentent dans deux échantillons dans le cas de la procédure de tirages coordonnés, 2 909 le sont du fait de la coordination positive mentionnée précédemment. Pour les 6 175 unités restantes, 50 % d'entre elles appartiennent, dans un des deux échantillons dans lesquels elles sont sélectionnées, à une strate présentant un taux de sondage supérieur à 50 %, et 45 % appartiennent à une strate présentant un taux de sondage compris entre 20 % et 50 %.

2.2. Coordination entre échantillons de niveaux différents

La coordination entre échantillons de niveaux différents constitue un problème délicat, qui n'avait jusqu'à présent ni été traité en détail d'un point de vue théorique dans le document de travail de Christian Hesse, ni *a fortiori* abordé dans les différents tests.

Pour gérer cette articulation entre sondages d'unités relatives à des niveaux différents – disons unités légales *versus* établissements pour fixer les idées, mais la méthode vaut également, *mutatis mutandis*, pour d'autres niveaux (entreprises, groupes, etc.) –, nous reprenons la démarche évoquée rapidement au paragraphe 8.5 du document de travail de Christian Hesse, ce qui conduit à la procédure suivante :

- ❶ on génère un jeu de numéros aléatoires permanents selon une loi uniforme sur [0;1] pour les établissements, et on attribue aux unités légales le numéro aléatoire de leur établissement principal[12]. On dispose ainsi, pour chaque niveau, d'un jeu de numéros aléatoires permanents issu d'une loi uniforme sur [0;1], avec un lien univoque [UL ↔ établissement principal] entre ces deux jeux ;

- ❷ chaque univers – unités légales d'un coté, établissements de l'autre – est ensuite soumis à son propre système d'enquêtes coordonnés – ce qui implique en particulier la gestion de fonctions de coordination propres à chaque niveau –, la coordination entre échantillons d'unités légales et échantillons d'établissements s'effectuant exclusivement par l'intermédiaire du lien [UL ↔ établissement principal] selon le schéma suivant :

  - → lors du tirage d'un échantillon d'unités légales, la coordination avec les échantillons relatifs aux enquêtes établissements s'effectue en faisant « remonter » au niveau unité légales, pour chaque unité légale les fonctions de charges de son établissement principal dans les différentes enquêtes établissements ;

  - → à l'inverse, lors du tirage d'un échantillon d'établissements, la coordination avec les échantillons relatifs aux enquêtes unités légales s'effectue en faisant « descendre » au niveau établissement, pour chaque établissement principal, les fonctions de charges de son unité légale dans les différentes enquêtes unités légales ;

Ainsi, cette méthode assure une coordination entre enquêtes unités légales et établissements via la prise en compte, pour un tirage relatif à un niveau donné, des charges d'enquêtes relatives à l'autre niveau grâce au seul lien [UL ↔ établissement principal] : pour un sondage d'unités légales, les charges d'enquête des seuls établissements principaux sont intégrées au calcul de la charge pesant sur les unités légales, et réciproquement, pour un sondage d'établissement, les charges d'enquêtes des unités légales sont prises en compte uniquement dans la charge des établissements principaux.

Un inconvénient possible de cette procédure est qu'elle ne permet ni la prise en compte pour un tirage d'unités légales des charges pesant sur leurs établissements secondaires, ni l'intégration dans le calcul de la charge des établissements secondaires des charges d'enquêtes relatives à leur unité légale. Il n'y a donc aucun lien avec cette procédure entre charges d'enquêtes des unités légales et charges d'enquête des établissements secondaires. Ceci semble toutefois consubstantiel à cette technique de coordination. En effet, cette méthode repose sur l'utilisation de numéros aléatoires permanents et de fonctions de coordination, selon le schéma suivant :

- ❶ lors du tirage d'une enquête donnée, pour chaque unité de la population, les fonctions de coordinations relatives à cette unité dans les enquêtes passées avec lesquelles on souhaite coordonner servent de base au calcul d'une fonction de charge cumulée ;

- ❷ cette fonction de charge cumulée permet ensuite de déterminer la fonction de coordination de l'unité pour le tirage en cours ;

- ❸ enfin, l'application de cette fonction de coordination au numéro aléatoire permanent de l'unité permet de définir le numéro aléatoire transformé de l'unité qui sert de base au tirage de l'échantillon.

Or, la validité de cette procédure de coordination est subordonnée au fait que le numéro aléatoire d'une unité k, utilisé à chaque tirage d'enquête, est **permanent,** identique d'une enquête à l'autre. Dès lors, une coordination entre une enquête « unités légales » et une enquête « établissements », qui aurait pour objectif de réduire le « cumul » de charges sur une unité légale et sur ses établissements, ne peut se faire via cette procédure qu'à la condition que l'unité légale et les établissements qui lui sont associés dans le système de coordination aient le même numéro aléatoire.

Par ailleurs, pour chaque niveau, ces numéros doivent avoir été générés aléatoirement selon une loi uniforme sur [0;1], afin que la sélection des unités présentant les plus petits numéros aléatoire transformés corresponde bien à un sondage aléatoire simple. La réunion de ces deux conditions conduit

[12] Qui sera souvent le siège social lors de la création de l'unité légale.

alors nécessairement à ne pouvoir « mettre en correspondance » dans le système intégré une unité légale qu'avec un seul de ses établissements.

Nous avons testé cette procédure de coordination multi-niveaux, en intégrant dans nos simulations 8 enquêtes établissements[13] en sus des vingt enquêtes unités légales déjà mentionnées précédemment. Nous avons distingué trois scénarios :

- tirages indépendants des 28 enquêtes, en respectant les différents plans de sondage mis en œuvre lors des tirages effectifs ;
- tirages coordonnés[14] des vingt enquêtes unités légales d'une part et des huit enquêtes établissements d'autre part, sans coordination entre les deux niveaux de tirage ;
- tirages coordonnés[15] des 28 enquêtes unités légales et établissements via la procédure de coordination multi-niveaux présentée ci-dessus.

Nous avons ensuite comparé les performances de ces différentes stratégies de tirage en termes de répartition de la charge d'enquête[16] – toujours hors parties exhaustives et parties conservées des échantillons. Contrairement aux simulations précédentes, qui n'impliquaient que des unités de même niveau et pour lesquelles le concept de charge d'enquête allait par conséquent de soi, plusieurs notions différentes de charge d'enquête peuvent être définies ici, selon le niveau auquel on se place – charge d'enquête pesant sur les unités légales ou sur les établissements – et le lien [UL ↔ établissement] que l'on retient pour le calcul de la charge à un niveau donné.

Le tableau 3 présente la distribution de la charge d'enquête de niveau unité légale avec prise en compte, au titre des tirages établissements, des charges d'enquête des seuls établissements principaux[17].

Charge d'enquête de niveau unité légale, hors exhaustifs et parties conservées, établissements principaux uniquement	Fréquence selon le scénario de tirage retenu			Écarts entre les scénarios de tirages :		
	Tirages indépendants	Tirages coordonnés séparés	Tirages coordonnés multi-niveaux	indépendants & coordonnés séparés	coordonnés séparés & coordonnés multi-niveaux	indépendants & coordonnés multi-niveaux
0	4 670 676	4 651 954	4 634 250	-18 722	-17 704	-36 426
1	410 016	439 355	474 286	29 339	34 931	64 270
2	40 095	34 824	18 230	-5 271	-16 594	-21 865
3	8 072	4 679	4 125	-3 393	-554	-3 947
4	2 142	813	737	-1 329	-76	-1 405
5	578	93	92	-485	-1	-486
6	121	5	2	-116	-3	-119
7	20	0	1	-20	1	-19
8	3	0	0	-3	0	-3

Tableau 3 : distribution de la charge d'enquête de niveau unité légale avec prise en compte des charges d'enquête des seuls établissements principaux, hors parties exhaustives et parties conservées, selon le scénario de tirage retenu

Les résultats obtenus sont sans surprise et parfaitement cohérents : par rapport à des tirages indépendants, la stratégie de coordination séparée conduit à une meilleure répartition de la charge, et ce phénomène est encore largement accentué lorsque l'on procède à une coordination multi-niveaux.

Point de vue dual, le tableau 4 présente la distribution de la charge d'enquête de niveau établissement, avec affectation des charges d'enquête des unités légales aux seuls établissements principaux[18], et conduit aux mêmes conclusions.

[13] Il s'agit des enquêtes Ecmoss 2010 à 2013, Reponse 2011, ECET 2012, Conditions de travail 2012 et Déchets 2013.
[14] Coordination négative avec l'ensemble des enquêtes passées, charge unitaire pour chaque enquête.
[15] Là encore, coordination négative avec l'ensemble des enquêtes passées, charge unitaire pour chaque enquête.
[16] Ici encore, il s'agit du nombre d'échantillons auxquels une unité appartient.
[17] La charge d'enquête de niveau unité légale ainsi définie correspond à celle utilisée par la procédure de coordination multi-niveaux pour les tirages d'échantillons d'unités légales.
[18] La charge d'enquête de niveau établissement ainsi définie correspond cette fois-ci à celle utilisée par la procédure de coordination multi-niveaux pour les tirages d'échantillons d'établissements. Notons qu'en termes de comptages, la différence avec le tableau 3 résulte exclusivement de la prise en compte dans ce tableau 4 des établissements secondaires.

Charge d'enquête de niveau établissement, hors exhaustifs et parties conservées, charges UL affectées aux seuls EP	Fréquence selon le scénario de tirage retenu			Écarts entre les scénarios de tirages :		
	Tirages indépendants	Tirages coordonnés séparés	Tirages coordonnés multi-niveaux	indépendants & coordonnés séparés	coordonnés séparés & coordonnés multi-niveaux	indépendants & coordonnés multi-niveaux
0	5 511 266	5 488 093	5 470 203	-23 173	-17 890	-41 063
1	445 399	482 381	517 459	36 982	35 078	72 060
2	45 445	38 004	21 436	-7 441	-16 568	-24 009
3	9 234	5 044	4 502	-4 190	-542	-4 732
4	2 354	828	753	-1 526	-75	-1 601
5	606	94	93	-512	-1	-513
6	122	5	2	-117	-3	-120
7	20	0	1	-20	1	-19
8	3	0	0	-3	0	-3

Tableau 4 : distribution de la charge d'enquête de niveau établissement avec affectation des charges d'enquête des unités légales aux seuls établissements principaux, hors parties exhaustives et parties conservées, selon le scénario de tirage

Enfin, si l'on s'intéresse à nouveau à la charge d'enquête de niveau unité légale, mais en prenant en compte cette fois-ci dans le calcul de la charge, au titre des tirages établissements, les charges d'enquête de tous les établissements – chaque établissement sélectionné dans un échantillon, qu'il soit principal ou secondaire, est ainsi comptabilisé dans la charge de son unité légale ; l'idée sous-jacente étant que la charge d'enquête pesant sur un établissement secondaire d'une unité légale donnée peut également être « ressentie » au niveau cette l'unité légale –, on obtient les résultats du tableau 5.

Ces résultats viennent confirmer ceux du tableau 3 : la procédure de tirages coordonnés multi-niveaux permet un resserrement important de la distribution autour de un, et donc une meilleure répartition de la charge d'enquête entre les différentes unités. L'existence d'unité légales possédant de très nombreux établissements[19], associée au fait que les charges d'enquêtes des établissements secondaires ne soit pas prises en comptes lors des tirages d'unités légales par la procédure, explique la persistance de charge d'enquête élevées.

Charge d'enquête de niveau unité légale, hors exhaustifs et parties conservées, ensemble des établissements	Fréquence selon le scénario de tirage retenu			Écarts entre les scénarios de tirages :		
	Tirages indépendants	Tirages coordonnés séparés	Tirages coordonnés multi-niveaux	indépendants & coordonnés séparés	coordonnés séparés & coordonnés multi-niveaux	indépendants & coordonnés multi-niveaux
0	4 662 214	4 644 324	4 627 001	-17 890	-17 323	-35 213
1	411 407	438 741	472 463	27 334	33 722	61 056
2	41 640	37 096	21 573	-4 544	-15 523	-20 067
3	9 494	6 528	5 725	-2 966	-803	-3 769
4	3 255	2 070	2 017	-1 185	-53	-1 238
5	1 363	951	977	-412	26	-386
6	719	539	489	-180	-50	-230
7	395	323	321	-72	-2	-74
8	252	205	213	-47	8	-39
9	165	165	159	0	-6	-6
10	133	105	101	-28	-4	-32
11 à 20	421	433	441	12	8	20
21 à 30	111	94	93	-17	-1	-18
30 à 50	77	66	69	-11	3	-8
Plus de 50	77	83	81	6	-2	4

Tableau 5 : distribution de la charge d'enquête de niveau unité légale avec prise en compte des charges d'enquête de l'ensemble des établissements, hors parties exhaustives et parties conservées, selon le scénario de tirage retenu

2.3. Coordination entre échantillons avec charges différentiées selon les enquêtes

Jusqu'à présent, les différentes simulations réalisées dans le cadre de cette étude se plaçaient dans le contexte d'enquêtes présentant des charges identiques et unitaires : la charge d'enquête pesant sur une entreprise était ainsi égale au nombre d'échantillons auxquels cette unité appartenait. Toutefois, les enquêtes menées par le système statistique public ne présentent pas toutes le même degré de complexité : certaines enquête sont relativement légères et n'imposent que peu de contraintes aux unités enquêtées, tandis que d'autres sont particulièrement complexes avec des questionnaires difficiles et long à remplir. Ainsi, la fréquence de sélection ne constitue pas nécessairement le meilleur critère pour mesurer et par conséquent « optimiser » la charge d'enquête pesant sur les entreprises. La

[19] À titre d'exemple, la SNCF compte plus de 3 000 établissements, et Orange près de 2 500…

méthode de coordination proposée autorisant la différenciation des charges associées aux différentes enquêtes, nous avons relancé la séquence de tirages coordonnés des 28 enquêtes unités légales et établissements présentées au point précédent, en distinguant cette fois-ci les enquêtes « lourdes[20] » – temps de réponse moyen ou médian déclaré dans le dossier de passage devant le comité du label supérieur à 30 minutes –, qui se voyaient attribuer une charge égale à 2, des enquêtes « légères » qui conservaient une charge unitaire. Le tableau 6 compare les performances de cette séquence de tirages coordonnés avec charges différentiées et de la séquence de tirages coordonnés avec charges unitaires, tant en termes de répartition de la fréquence de sélection des unités qu'en termes de répartition de la charge d'enquête réelle.

Fréquence de tirage de niveau unité légale, hors exhaustifs et parties conservées, coordination multi-niveaux	Fréquence selon le scénario de tirage retenu			Ecarts entre les scénarios de tirages :		
	Tirages indépendants	Tirages coordonnés, charges =	Tirages coordonnés, charges ≠	indépendants & coordonnés charges =	coordonnés « charges = » versus « charges ≠ »	indépendants & coordonnés charges ≠
0	4 670 676	4 634 250	4 634 612	-36 426	362	-36 064
1	410 016	474 286	473 826	64 270	-460	63 810
2	40 095	18 230	18 057	-21 865	-173	-22 038
3	8 072	4 125	4 345	-3 947	220	-3 727
4	2 142	737	802	-1 405	65	-1 340
5	578	92	79	-486	-13	-499
6	121	2	2	-119	0	-119
7	20	1	0	-19	-1	-20
8	3	0	0	-3	0	-3
Charge réelle de niveau unité légale, hors exhaustifs et parties conservées, coordination multi-niveaux	**Tirages indépendants**	**Tirages coordonnés, charges =**	**Tirages coordonnés, charges ≠**	**indépendants & coordonnés charges =**	**coordonnés « charges = » versus « charges ≠ »**	**indépendants & coordonnés charges ≠**
0	4 670 676	4 634 250	4 634 612	-36 426	362	-36 064
1	241 297	275 147	272 716	33 850	-2 431	31 419
2	180 664	204 014	206 347	23 350	2 333	25 683
3	18 828	9 696	9 995	-9 132	299	-8 833
4	13 954	6 047	5 807	-7 907	-240	-8 147
5	3 331	1 454	1 417	-1 877	-37	-1 914
6	1 931	850	697	-1 081	-153	-1 234
7	657	185	91	-472	-94	-566
8	269	73	37	-196	-36	-232
9	74	5	4	-69	-1	-70
10	28	2	0	-26	-2	-28
11	12	0	0	-12	0	-12
12	2	0	0	-2	0	-2

Tableau 6 : distribution de la fréquence de sélection et de la charge d'enquête de niveau unité légale, hors parties exhaustives et parties conservées, selon le scénario de tirage retenu – différenciation des charges « binaire »

La prise en compte des charges ainsi différentiées au sein de la procédure de coordination ne dégrade que très marginalement la répartition de la fréquence de sélection des unités : on n'observe qu'une légère augmentation du nombre d'unités sélectionnées dans plusieurs échantillons. En contrepartie, on constate une diminution du nombre d'unités concernées par des charges d'enquête réelle supérieures à trois, qui s'opère au prix d'une augmentation du nombre d'unités affectées par une charge d'enquête réelle de deux, et dans une moindre mesure de trois.

Ce phénomène résulte vraisemblablement[21] du processus suivant : dans le cas d'un tirage avec charges identiques entre les enquêtes, l'algorithme cherche à chaque tirage à minimiser le recouvrement de l'échantillon en cours de tirage avec l'ensemble des échantillons déjà tirés ; lorsque les charges sont différentiées, lors du tirage d'un échantillon donné, l'algorithme cherche à minimiser en priorité le recouvrement de l'échantillon en cours de tirage avec les échantillons relatifs aux enquêtes lourdes déjà tirées, ce qui peut le conduire à augmenter légèrement le recouvrement avec certains échantillons d'enquêtes légères. D'où une augmentation conjointe du nombre d'unités sélectionnées dans deux enquêtes légères ou plus et du nombre d'unités sélectionnées dans seulement une enquête lourde, qui conduit à cet accroissement du nombre d'unités affectées par une charge de deux.

Il est également possible de différencier de manière plus fine les charges associées aux différentes enquêtes, en fonction du temps de réponse moyen ou médian déclaré dans le dossier de passage

[20] Il s'agit des enquêtes ESA 2008 à 2011, Ecmoss 2010 à 2013, Reponse 2011 et Qualité énergétique mise en œuvre par les entreprises dans les bâtiments 2012.

[21] Cette interprétation est d'ailleurs confortée par le fait que, si l'on refait la simulation avec une charge associée aux enquêtes lourdes égale à 3, on observe un phénomène similaire : diminution des charges supérieures à trois, augmentation du nombre d'unités affectées par une charge d'enquête réelle égale à trois.

devant le comité du label par exemple. En outre, on peut également envisager de faire décroître la charge associée à une enquête en fonction de son antériorité dans le processus de coordination : en effet, lors du tirage d'une enquête donnée, il semble opportun de privilégier la coordination avec les enquêtes les plus récentes au détriment d'enquêtes plus anciennes.

Nous avons donc considéré le plan de coordination plus complexe suivant :

- chaque enquête est coordonnée négativement avec les 19 enquêtes précédentes (la totalité des enquêtes précédentes pour les vingt premières enquêtes)…
- … avec des charges s'étalant de 0,25 pour les enquêtes IPEA – très légères – à 6 pour les enquêtes Ecmoss – particulièrement lourdes…
- …et diminuant de 10 % par « rang d'antériorité » dans le processus de coordination à partir du onzième rang.

Le tableau 7 présente les résultats de ce plan de coordination complexe à l'issue du tirage de la 28ème et dernière enquête par rapport au plan de coordination simple – coordination de la 28ème enquête avec les 20 enquêtes précédentes affectées de charges unitaires – en termes de répartition de la fréquence de sélection des unités et en termes de répartition de la charge d'enquête réelle.

Fréquence de tirage de niveau unité légale, hors exhaustifs et parties conservées, coordination multi-niveaux	Fréquence selon le scénario de tirage retenu			Écarts entre les scénarios de tirages :		
	Tirages indépendants	Tirages coordonnés, charges =	Tirages coordonnés, charges ≠	indépendants & coordonnés charges =	coordonnés « charges = » versus « charges ≠ »	indépendants & coordonnés charges ≠
0	4 326 056	4 310 146	4 309 931	-15 910	-215	-16 125
1	189 686	217 115	217 690	27 429	575	28 004
2	18 057	9 629	9 232	-8 428	-397	-8 825
3	4 110	2 082	2 104	-2 028	22	-2 006
4	1 026	222	232	-804	10	-794
5	234	9	16	-225	7	-218
6	32	2	0	-30	-2	-32
7	4	0	0	-4	0	-4
Charge réelle de niveau unité légale, hors exhaustifs et parties conservées, coordination multi-niveaux	**Tirages indépendants**	**Tirages coordonnés, charges =**	**Tirages coordonnés, charges ≠**	**indépendants & coordonnés charges =**	**coordonnés « charges = » versus « charges ≠ »**	**indépendants & coordonnés charges ≠**
0	4 337 654	4 326 179	4 325 320	-11 475	-859	-12 334
1	123 785	137 192	136 920	13 407	-272	13 135
2	14 752	13 109	13 865	-1 643	756	-887
3	44 045	45 346	45 510	1 301	164	1 465
4	2 725	910	902	-1 815	-8	-1 823
5	352	84	52	-268	-32	-300
6	9 639	12 436	13 796	2 797	1 360	4 157
7	2 357	1 482	1 387	-875	-95	-970
8	1681	1340	730	-341	-610	-951
9	561	166	82	-395	-84	-479
10	206	67	13	-139	-54	-193
11	61	2	2	-59	0	-59
12	901	736	578	-165	-158	-323
13	196	93	27	-103	-66	-169
14	228	58	20	-170	-38	-208
15	41	4	1	-37	-3	-40
16	19	1	0	-18	-1	-19
17	2	0	0	-2	0	-2

Tableau 7 : distribution de la fréquence de sélection et de la charge d'enquête de niveau unité légale, hors parties exhaustives et parties conservées, selon le scénario de tirage retenu – différenciation des charges complexe

Les résultats sont en ligne avec ceux obtenus précédemment : l'impact en termes de répartition de la fréquence de sélection des unités s'avère toujours extrêmement marginal, tandis qu'en termes de répartition de la charge d'enquête réelle, on constate à nouveau une diminution du nombre d'unités concernées par des charges supérieures à 6 – valeur qui constitue dans cette simulation la charge maximale associée à une enquête – et une concentration plus importantes sur les charges de 3 – charge associée à plusieurs enquêtes relativement lourdes lors du processus de coordination – et 6 – charge associée aux enquêtes Ecmoss lors du processus de coordination.

Enfin, nous avons testé l'impact d'une variation de charge portant sur une enquête particulière, lorsque l'objectif prioritaire de la coordination est cette fois-ci d'assurer un recouvrement minimal entre l'enquête en cours de tirage et une autre enquête spécifique, et non plus de répartir au mieux la charge entre l'ensemble des enquêtes. Ce cas de figure est susceptible de se présenter lorsque l'on renouvelle une

partie d'un échantillon rotatif : il est alors fréquent de chercher à disjoindre autant que faire ce peut la partie renouvelée de l'échantillon en cours de tirage de l'échantillon relatif au millésime précédent de l'enquête, afin de minimiser le nombre d'unités qui seront présentes dans le panel pendant plus d'un cycle de rotation.

Nous avons donc comparé, pour le tirage de la partie renouvelée de l'ESA 2010, qui constitue la douzième enquête de notre simulation et dont l'échantillon est renouvelé par moitié, différentes stratégies de tirages coordonnés avec les onze enquêtes précédentes, en faisant varier la charge associée à l'ESA 2009[22]: charge unitaire, charge égale à 2 ou charge égale à 14. Nous nous sommes intéressés à l'impact de cette variation de charge tant du point de vue de la qualité de la répartition globale de la charge d'enquête sur ces douze enquêtes que du recouvrement entre les échantillons hors exhaustifs relatifs à la partie renouvelée de l'ESA 2010 et à l'ESA 2009. Le tableau 8 présente les résultats de ces simulations.

Fréquence de tirage de niveau unité légale, hors exhaustifs et parties conservées	Fréquence selon le scénario de tirage retenu				Écarts entre les scénarios de tirages :		
	Tirages indépendants	Tirages coordonnés charge ESA 2009 = 1	Tirages coordonnés charge ESA 2009 = 2	Tirages coordonnés charge ESA 2009 = 14	indépendants & coordonnés charges =	coordonnés « charges = » versus « charge ESA09 =2 »	coordonnés « charges = » versus « charge ESA09 =14 »
0	4 097 491	4 085 252	4 085 273	4 085 386	-12 239	21	134
1	305 167	328 524	328 487	328 269	23 357	-37	-255
2	16 012	5 994	6 005	6 102	-10 018	11	108
3	1 089	65	70	78	-1 024	5	13
4	72	1	1	1	-71	0	0
5	5	0	0	0	-5	0	0
Recouvrement entre échantillon ESA 2010 renouvelé et ESA 2009	2 468	258	215	186	-2 210	-43	-72

Tableau 8 : distribution de la charge d'enquête de niveau unité légale, hors parties exhaustives et parties conservées, et recouvrement entre les échantillons « ESA 2010 renouvelé » et « ESA 2009 », selon la charge affectée à l'ESA 2009 dans le plan de coordination

On peut tout d'abord noter que la méthode de coordination avec charges unitaires pour toutes les enquêtes conduit déjà à d'excellents résultats en terme de recouvrement entre les deux échantillons évoqués ci-dessus, puisqu'on passe de 2 468 unités présentes dans les deux échantillons hors exhaustifs dans le cas de tirage indépendants à seulement 258 dans le cas de tirages coordonnés. Par ailleurs, en ce qui concerne l'impact de l'augmentation de la charge associée à l'ESA 2009 dans le plan de coordination relatif au tirage de la partie renouvelée de l'ESA 2010, les résultats obtenus sont conformes aux attentes : cette augmentation se traduit bien par une diminution du recouvrement entre les échantillons de ces deux enquêtes. Ce gain en termes de recouvrement s'effectue au prix d'une légère dégradation de la qualité de la coordination globale, puisqu'on observe une augmentation – qui reste cependant mesurée, même dans le scénario extrême où la charge associée à l'ESA 2009 est fortement augmentée – du nombre d'unités sélectionnées dans plus d'un échantillon, augmentation elle-même croissante avec la charge associée à l'ESA 2009. Les gains en termes de recouvrement et pertes en termes de qualité de la coordination globale sont du même ordre de grandeur.

3. Étude de deux questions d'ordre méthodologique

Pour compléter cette étude sur les propriétés de la méthode de coordination, nous nous sommes intéressés à deux questions d'ordre méthodologique : d'une part le problème du biais de rétroaction, et d'autre part la question du tirage systématique sur données triées.

3.1. Le problème du biais de rétroaction

Le biais de rétroaction est un problème relatif à la coordination d'échantillon qui apparaît lorsque le mécanisme d'échantillonnage n'est pas indépendant des procédures de réintroduction d'information dans la base de sondage. Plus précisément, si les résultats d'une enquête A servent à mettre à jour les bases de sondage d'enquêtes postérieures à cette dernière, et si les échantillons de ces enquêtes postérieures sont tirés de façon coordonnée avec l'enquête A, alors les échantillons ainsi sélectionnés conduiront à des estimations biaisées[23].

[22] La charge associée aux onze autres enquêtes est ici systématiquement unitaire.
[23] Pour une démonstration mathématique de ce résultat, on peut se reporter à l'annexe 1 du document de travail [7].

Ce phénomène se révèle particulièrement problématique dans l'optique d'une coordination globale des enquêtes entreprises voulue par l'Insee. En effet, la majeure partie des enquêtes auprès des entreprises menées par le service statistique public voient leurs bases de sondage constituées à partir du répertoire Sirus, et ce dernier est régulièrement mis à jour à partir des résultats des différentes enquêtes. Ainsi, les cessations détectées au cours de certaines enquêtes sont signalées au répertoire Sirus et prises en compte par ce dernier[24]. Autre exemple, les code APE de classement sectoriel des unités du répertoire Sirus – codes constituant la base de la stratification sectorielle mise en œuvre dans la quasi-totalité des enquêtes entreprises – sont mis à jour chaque année en fonction des résultats des enquêtes ESA et EAP[25]. Dès lors, la mise en place d'une procédure de coordination globale des enquêtes entreprises implique :

- soit de ne plus mettre à jour les bases de sondage des différentes enquêtes à partir des résultats d'autres enquêtes, en particulier les résultats en termes de classement sectoriel issus des enquêtes ESA et EAP. Ceci semble difficilement envisageable, puisque cela conduirait à retenir dans les échantillons un nombre de plus en plus important d'unités cessées, hors champ ou mal classées, ce qui ne manquerait pas d'être rapidement préjudiciable à la qualité des enquêtes en termes de précision et de robustesse des résultats ;

- soit d'exclure de la procédure de coordination les enquêtes dont les résultats affectent le plus le répertoire Sirus, à savoir les enquêtes ESA et EAP. C'est l'option retenue par l'office fédéral de la statistique suisse, au sein duquel les enquêtes spécifiques de mise à jour sont tirées indépendamment des enquêtes coordonnées de production statistique. Toutefois, dans la mesure où l'ESA constitue la plus importante – en termes de taille d'échantillon – des enquêtes menées auprès des entreprises par le service statistique public et s'avère relativement « lourde », l'exclure du processus de coordination ne serait pas complètement satisfaisant ;

- soit de passer outre le problème du biais de rétroaction, en considérant que ce dernier est suffisamment faible pour être négligeable par rapport aux désagréments liés aux deux autres alternatives précédemment énoncées.

Afin de pouvoir trancher entre les deux dernières options, il était nécessaire de quantifier l'ampleur du biais de rétroaction. Pour ce faire, nous avons procédé par simulation à partir des données relatives aux campagnes Esane 2008 à 2011. Plus précisément, nous avons procédé[26] pour les ESA « Commerce de gros » 2008 à 2011, d'une part 5 000 tirages indépendants et d'autre part à 5 000 tirages coordonnés[27]. Puis, à partir de variables fiscales disponibles pour l'ensemble des unités, nous avons estimé, pour chaque scénario, le biais moyen relatif pour les estimations sectorielles de niveau sous-classe. Le tableau 9 présente la distribution de ces biais relatifs pour les principales variables fiscales du dispositif Esane en 2011 – on trouvera en annexe les mêmes tableaux pour 2009 et 2010, qui présentent des résultats similaires à ceux de 2011.

Le secteur du commerce de gros se caractérise par un taux de changement d'APE[28] dans les bases de sondages de l'ordre de 3 % par an[29] sur la période 2008-2011, contre environ 1,5 % par an sur l'ensemble du champ ESA. Et malgré cette « démographie sectorielle » relativement importante[30], le fait de procéder à des tirages coordonnés ne semble pas induire de biais significatif et systématique dans les estimations par rapport à une stratégie de tirages indépendants.

---

[24] Ce point n'est pas nécessairement le plus problématique, dans la mesure où ces cessations sont également susceptibles d'être détectées, certes de manière plus tardive, par des sources administratives indépendantes des enquêtes.

[25] Il s'agit du point qui pose le plus de difficultés, puisque, contrairement au critère d'activité des unités, ces enquêtes sont les seules sources d'informations sur le classement sectoriel des unités.

[26] Selon un plan de sondage proche de celui effectivement mis en œuvre en termes de stratification et d'allocations, la principale différence consistant en un abandon du principe de renouvellement par moitié de l'échantillon.

[27] Plus précisément, chaque millésime est coordonné négativement avec l'ensemble des millésimes passés, l'ESA 2008 constituant le point de départ de ces simulations.

[28] Hors créations et cessations – majoritairement connues via des sources administratives indépendantes des ESA – et entrées/sorties de champ Esane – en nombre négligeable du fait de l'étendue de ce dernier.

[29] Environ 1 % d'entrées dans le secteur du commerce de gros, 1,3 % de sorties du secteur du commerce de gros et 0,7 % de changements d'APE au sein de ce secteur.

[30] Des premières simulations, menées sur le secteur des transports, conduisaient à des résultats similaires, mais moins « significatifs » du fait du plus faible nombre de changements d'APE observés dans ce secteur, de l'ordre de 1,5 %.

Tirages indépendants									
Variable	**Nombre d'entreprises**	**Chiffre d'affaires**	**Total des achats**	**Salaires**	**Valeur ajoutée**	**EBE**	**Résultat comptable**	**Total de l'actif**	**Total du passif**
Moyenne	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	-0,1%	0,0%	0,0%
maximum	0,3%	0,1%	0,1%	0,2%	0,2%	1,0%	7,6%	0,9%	1,0%
P99	0,3%	0,1%	0,1%	0,2%	0,2%	1,0%	7,6%	0,9%	1,0%
P95	0,1%	0,1%	0,1%	0,1%	0,1%	0,4%	1,0%	0,1%	0,1%
P90	0,1%	0,1%	0,1%	0,1%	0,1%	0,2%	0,4%	0,1%	0,1%
P75	0,0%	0,0%	0,0%	0,0%	0,0%	0,1%	0,1%	0,0%	0,0%
mediane	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
P25	0,0%	0,0%	0,0%	0,0%	0,0%	-0,1%	-0,2%	0,0%	0,0%
P10	-0,1%	-0,1%	-0,1%	-0,1%	-0,1%	-0,1%	-0,4%	-0,1%	-0,1%
P5	-0,1%	-0,1%	-0,2%	-0,1%	-0,1%	-0,2%	-1,1%	-0,3%	-0,2%
P1	-0,2%	-0,6%	-0,6%	-0,4%	-0,4%	-0,4%	-14,5%	-1,0%	-1,1%
minimum	-0,2%	-0,6%	-0,6%	-0,4%	-0,4%	-0,4%	-14,5%	-1,0%	-1,1%

Tirages coordonnés									
Variable	**Nombre d'entreprises**	**Chiffre d'affaires**	**Total des achats**	**Salaires**	**Valeur ajoutée**	**EBE**	**Résultat comptable**	**Total de l'actif**	**Total du passif**
Moyenne	0,0%	0,0%	0,0%	0,0%	0,0%	-0,6%	3,2%	0,0%	0,0%
maximum	0,1%	0,2%	0,2%	0,3%	0,3%	0,7%	197,5%	0,4%	0,8%
P99	0,1%	0,2%	0,2%	0,3%	0,3%	0,7%	197,5%	0,4%	0,8%
P95	0,1%	0,2%	0,2%	0,2%	0,2%	0,4%	0,7%	0,2%	0,2%
P90	0,1%	0,1%	0,1%	0,1%	0,1%	0,3%	0,3%	0,1%	0,2%
P75	0,0%	0,1%	0,1%	0,0%	0,0%	0,1%	0,1%	0,0%	0,1%
mediane	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
P25	0,0%	0,0%	0,0%	0,0%	0,0%	-0,1%	-0,1%	0,0%	0,0%
P10	-0,1%	-0,1%	-0,1%	0,0%	-0,1%	-0,2%	-0,5%	-0,1%	-0,1%
P5	-0,2%	-0,1%	-0,1%	-0,1%	-0,1%	-0,3%	-1,2%	-0,1%	-0,1%
P1	-0,2%	-0,1%	-0,2%	-0,1%	-0,1%	-38,7%	-5,5%	-0,5%	-0,6%
minimum	-0,2%	-0,1%	-0,2%	-0,1%	-0,1%	-38,7%	-5,5%	-0,5%	-0,6%

Tableau 9 : moyenne et distribution des biais relatifs pour les estimations de niveau sous-classe des principales variables fiscales du dispositif Esane en 2011, selon les scénarios de tirage

Seul un secteur – le 4618Z « Intermédiaires spécialisés dans le commerce d'autres produits spécifiques » – présente des biais relatifs importants pour les variables « excédent brut d'exploitation » (-38,7 %) et « résultat comptable : bénéfice ou perte » (+197,5 %) lorsque les tirages sont coordonnés. Ces deux variables se caractérisent toutefois par une dispersion importante, avec beaucoup d'unités déclarant des montants faibles[31] et quelques unités déclarant des montants très élevés, en positif comme en négatif ; en particulier, une unité du 4618Z présente un EBE égal au résultat net de -389 millions d'euros, alors qu'elle appartient à une strate à faible taux de sondage (1/76). Aussi, les biais relatifs importants observés pour ces deux variables dans le 4618Z traduisent-ils moins un biais de rétroaction lié aux tirages coordonnés que le caractère influent de cette unité pour ces variables et donc l'imprécision des biais relatifs mesurés dans ce secteur avec seulement 5 000 échantillons simulés. Pour preuve, si l'on procède à 5 000 tirages indépendants supplémentaires et que l'on calcule les biais relatifs dans le 4618Z à partir des 10 000 tirages indépendants ainsi disponibles, on obtient un biais relatif pour l'EBE de 25 % et pour le résultat net de -142 %, alors même que la stratégie de tirages indépendants est par construction sans biais…

Des analyses similaires menées sur les biais relatifs pour les estimations par tranche d'effectifs des entreprises du secteur du commerce de gros, dont les résultats sont donnés dans les tableaux en annexe, conduisent aux mêmes conclusions.

Ces résultats nous amènent à penser que le biais de rétroaction potentiel, lié à la coordination avec les enquêtes ESA et EAP qui servent à mettre à jour les bases de sondages, est suffisamment faible pour pouvoir être négligé[32]. En conséquence, les enquêtes ESA et EAP sont inclues dans la procédure de coordination globale des enquêtes entreprises mise en place par l'Insee.

3.2. Test d'une procédure de « sur-stratification » visant à remplacer le tirage systématique au sein des strates

[31] Plus de 98 % des unités présentent un EBE ou un résultat net inférieur, en valeur absolue, à 2 millions d'euros.

[32] Et ce d'autant plus que l'on ne coordonnera pas avec l'ensemble des enquêtes ESA et EAP passées, mais selon toute vraisemblance avec seulement le dernier ou les deux derniers millésimes…

Les échantillons des enquêtes auprès des entreprises sont quasi-systématiquement tirés selon des plans de sondages stratifiés, avec tirage à probabilités égales au sein de chaque strate. En outre, il n'est pas rare que le tirage des unités au sein de chaque strate soit effectué par tirage systématique après tri des unités au sein de chaque strate selon un certain critère. Cette procédure de tirage – qui permet d'obtenir, au sein de chaque strate, une répartition des unités de l'échantillon proche de celle observée dans la base de sondage pour le critère de tri – s'avère malheureusement totalement incompatible avec la méthode de coordination proposée. Cependant, le tirage systématique s'apparentant à une stratification implicite à allocations proportionnelles, il est possible de prendre en compte le critère jadis « contrôlé » par le tirage systématique dans la méthode de tirage coordonné en procédant comme suit :

❶ ajout d'un niveau de stratification supplémentaire définit par ledit critère (préalablement discrétisé le cas échéant, par exemple s'il s'agit d'un chiffre d'affaires) ;

❷ passage des allocations relatives à la stratification initiale aux allocations relatives à la stratification finale, plus fine, par « allocations proportionnelles ». Plus précisément, chaque strate finale i étant incluse dans une strate initiale h, on calcule le nombre d'unités à tirer n_i dans la strate finale i en y appliquant le taux de sondage t_h de la strate initiale h correspondante, puis en arrondissant les quantités ainsi obtenues via une méthode d'arrondis contrôlés ;

❸ regroupement automatique des strates fines afin d'obtenir des allocations non nulles pour les strates de tirage finales post-regroupements.

Cette procédure, dite de « sur-stratification », permet par construction d'obtenir un échantillon équivalent, en termes de « représentativité » vis-à-vis du critère considéré, à celui issu d'un tirage systématique stratifié avec la stratification initiale. Elle induit en revanche une augmentation du nombre de strates, ce qui pourrait influer sur la qualité de la coordination entre les enquêtes sans qu'il soit toutefois possible de prévoir *a priori* dans quel sens et dans quelle mesure.

Afin de quantifier l'impact de cette procédure de sur-stratification sur la qualité de la coordination en termes de répartition de la charge d'enquête, nous avons procédé à des simulations portant sur cinq enquêtes actuellement tirées par tirage systématique[33], en distinguant trois scénarios :

- séquence de tirages systématiques et indépendants en respectant les différents plans de sondage mis en œuvre lors des tirages effectifs de ces enquêtes ;
- séquence de tirages coordonnés de ces cinq enquêtes – CIS constituant le point de départ de la séquence, et chaque enquête étant coordonnée négativement avec l'ensemble des enquêtes passées – en respectant les différents plans de sondage mis en œuvre lors des tirages effectifs de ces enquêtes et en abandonnant purement et simplement le critère de tirage systématique ;
- séquence de tirages coordonnés de ces cinq enquêtes – même schéma de coordination que dans le scénario précédent – avec cette fois-ci sur-stratification pour prendre en compte le critère de tirage systématique. On passe ainsi, sur l'ensemble des 5 enquêtes, de 1 988 strates initiales à 6 378 strates fines après sur-stratification.

Pour chaque scénario, nous avons effectué 100 séquences de tirages différentes, pour lesquelles nous avons calculé les charges d'enquêtes hors exhaustifs. Comme on peut le constater dans le tableau 10, qui présente les distributions moyennes de ces charges d'enquête hors exhaustifs selon les trois scénarios, la procédure de sur-stratification envisagée ne joue que très marginalement sur la qualité de la coordination en termes de répartition de la charge d'enquête.

La procédure de sur-stratification présentée ci-dessus permet donc de prendre en compte avec la méthode de coordination les critères jadis « contrôlés » via les tirages systématiques, sans que l'augmentation du nombre de strates induite par la sur-stratification n'affecte la qualité de la coordination.

[33] À savoir les enquêtes CIS 2010 – tirage systématique par région –, ENDD 2011 – tirage systématique par région –, CVTS4 – tirage systématique par chiffre d'affaires –, Qualité énergétique mise en œuvre par les entreprises dans les bâtiments 2012 – tirage systématique par chiffre d'affaires – et CAM 2012 – tirage systématique par type d'unité (indépendante, filiale ou tête de groupe, d'un groupe français ou multinational, etc.).

Charge d'enquête, hors exhaustifs	Fréquence moyenne selon le scénario de tirage retenu			Écarts entre tirages indépendants et coordonnés « simples »	Écarts entre tirages coordonnés « simples » et coordonnés avec sur-stratification
	Tirages systématiques indépendants	Tirages coordonnés « simples »	Tirages coordonnés avec sur-stratification		
0	630 452	627 016	626 896	-3 436	-120
1	37 029	43 703	43 784	6 674	81
2	3 258	213	251	-3 045	38
3	188	1	2	-187	1
4	6	0	0	-6	0

Tableau 10 : distribution de la charge d'enquête, hors parties exhaustives, selon le scénario de tirage retenu

Conclusion

L'ensemble des simulations grandeur nature menées sur données réelles sont donc venues confirmer l'efficacité de la méthode de coordination en termes de répartition de la charge de réponse sur les différentes unités de la population, ainsi que sa remarquable robustesse vis à vis des différents paramètres – taux de sondage, stratifications des différentes enquêtes, nombre de strates, taux de recouvrement entre le champ des différentes enquêtes, charges associées aux différentes enquêtes, etc. Elles ont en outre permis d'en valider la faisabilité opérationnelle. Par ailleurs, une procédure de coordination entre échantillons de niveaux différents a été développée et testée, avec là encore des résultats pleinement satisfaisants. Enfin, des simulations ont permis d'évaluer l'ampleur du biais de rétroaction associé à la coordination avec les enquêtes ESA et EAP qui servent à mettre à jour les bases de sondages, et de conclure que ce dernier était suffisamment faible pour être négligé.

Le passage en production de cette méthode de coordination à l'Insee fin 2013, à l'occasion du tirage de l'échantillon de l'ESA 2013, a conduit à une adaptation du répertoire statistique Sirus – Système d'immatriculation au répertoire des unités statistiques, utilisé pour constituer la très grande majorité des bases de sondages des enquêtes auprès des entreprises menées par l'Insee et les différents SSM – et impose quelques contraintes supplémentaires aux concepteurs d'enquête. Ces différents aspects pratiques, ainsi qu'un premier bilan après 18 mois d'utilisation, sont présentés dans la contribution associée d'Anaïs Levieil-Guillon « Mise en œuvre et résultats de la nouvelle procédure de coordination des échantillons des enquêtes auprès des entreprises et/ou des établissements » [8].

Bibliographie

[1] Christian Hesse, « Généralisation des tirages aléatoires à numéros aléatoires permanents, ou la méthode JALES+ », *Document de travail Insee E0101*, 2001.

[2] Pascal Ardilly, « Présentation de la méthode JALES+ conçue par Christian Hesse », *Document de travail interne Insee*, 2009.

[3] Fabien Guggemos et Olivier Sautory, « Sampling Coordination of Business Surveys Conducted by Insee », *Proceedings of the Fourth International Conference of Establishment Surveys*, Montréal, Canada, 11-14 juin 2012.

[4] Olivier Sautory, « La coordination des échantillons d'enquêtes entreprises et établissements : une nouvelle méthode développée a l'Insee. », *Actes du 8ème colloque francophone sur les sondages*, Dijon, 18-20 novembre 2014.

[5] Kevin Rosamont-Prombo, « La coordination des échantillons d'entreprises », *Rapport de stage effectué à l'Insee*, 2012.

[6] Elvire Demoly, Arnaud Fizzala et Emmanuel Gros, « 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] Christian Hesse, « Sampling coordination : a review by country », *Document de travail Insee E9908*, 1999.

[8] Anaïs Levieil-Guillon, « Mise en œuvre et résultats de la nouvelle procédure de coordination des échantillons des enquêtes auprès des entreprises et/ou des établissements », *Actes des 12èmes Journées de Méthodologie Statistique*, 31 mars – 02 avril 2015, Insee.

Annexe

Tirages indépendants

Variable	Nombre d'entreprises	Chiffre d'affaires	Total des achats	Salaires	Valeur ajoutée	EBE	Résultat comptable	Total de l'actif	Total du passif
Moyenne	0,0%	-0,1%	-0,1%	-0,1%	-0,1%	0,0%	-0,1%	-0,1%	-0,1%
maximum	0,1%	0,4%	0,3%	0,1%	1,6%	6,8%	4,1%	1,8%	1,4%
P99	0,1%	0,4%	0,3%	0,1%	1,6%	6,8%	4,1%	1,8%	1,4%
P95	0,1%	0,2%	0,2%	0,1%	0,1%	1,0%	2,4%	0,2%	0,2%
P90	0,1%	0,1%	0,1%	0,0%	0,0%	0,2%	0,2%	0,1%	0,1%
P75	0,0%	0,0%	0,0%	0,0%	0,0%	0,1%	0,1%	0,0%	0,0%
mediane	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
P25	0,0%	0,0%	0,0%	0,0%	0,0%	-0,1%	-0,1%	0,0%	0,0%
P10	-0,1%	-0,1%	-0,1%	-0,1%	-0,1%	-0,2%	-0,6%	-0,2%	-0,2%
P5	-0,1%	-0,2%	-0,2%	-0,2%	-0,3%	-1,1%	-1,9%	-0,8%	-0,8%
P1	-0,2%	-3,5%	-3,0%	-4,7%	-6,3%	-11,8%	-6,7%	-3,5%	-3,1%
minimum	-0,2%	-3,5%	-3,0%	-4,7%	-6,3%	-11,8%	-6,7%	-3,5%	-3,1%

Tirages coordonnés

Variable	Nombre d'entreprises	Chiffre d'affaires	Total des achats	Salaires	Valeur ajoutée	EBE	Résultat comptable	Total de l'actif	Total du passif
Moyenne	0,0%	0,0%	0,0%	0,0%	0,0%	0,1%	0,0%	0,0%	0,0%
maximum	0,2%	0,5%	0,5%	0,3%	0,5%	2,3%	2,7%	0,6%	0,6%
P99	0,2%	0,5%	0,5%	0,3%	0,5%	2,3%	2,7%	0,6%	0,6%
P95	0,1%	0,2%	0,2%	0,1%	0,4%	1,2%	1,2%	0,2%	0,2%
P90	0,1%	0,1%	0,1%	0,0%	0,1%	0,3%	0,6%	0,1%	0,1%
P75	0,0%	0,0%	0,0%	0,0%	0,0%	0,1%	0,1%	0,0%	0,0%
mediane	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
P25	0,0%	-0,1%	-0,1%	0,0%	0,0%	-0,1%	-0,1%	0,0%	0,0%
P10	-0,1%	-0,1%	-0,1%	-0,1%	-0,1%	-0,2%	-0,3%	-0,1%	-0,1%
P5	-0,1%	-0,2%	-0,1%	-0,1%	-0,1%	-0,4%	-0,7%	-0,2%	-0,3%
P1	-0,2%	-0,2%	-0,2%	-0,2%	-0,3%	-2,4%	-3,0%	-0,5%	-0,6%
minimum	-0,2%	-0,2%	-0,2%	-0,2%	-0,3%	-2,4%	-3,0%	-0,5%	-0,6%

Moyenne et distribution des biais relatifs pour les estimations de niveau sous-classe des principales variables fiscales du dispositif Esane en 2009, selon les scénarios de tirage

Tirages indépendants

Variable	Nombre d'entreprises	Chiffre d'affaires	Total des achats	Salaires	Valeur ajoutée	EBE	Résultat comptable	Total de l'actif	Total du passif
Moyenne	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	11,7%	0,0%	0,0%
maximum	0,4%	0,6%	0,8%	0,4%	0,4%	1,1%	682,3%	0,5%	0,7%
P99	0,4%	0,6%	0,8%	0,4%	0,4%	1,1%	682,3%	0,5%	0,7%
P95	0,2%	0,2%	0,3%	0,1%	0,1%	0,7%	4,6%	0,2%	0,3%
P90	0,2%	0,2%	0,2%	0,1%	0,1%	0,4%	0,6%	0,2%	0,2%
P75	0,0%	0,0%	0,0%	0,0%	0,0%	0,1%	0,2%	0,0%	0,0%
mediane	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
P25	0,0%	0,0%	0,0%	0,0%	0,0%	-0,1%	-0,2%	-0,1%	-0,1%
P10	-0,1%	-0,1%	-0,1%	-0,1%	-0,2%	-0,4%	-0,6%	-0,1%	-0,1%
P5	-0,2%	-0,2%	-0,2%	-0,1%	-0,2%	-0,8%	-1,7%	-0,3%	-0,3%
P1	-0,4%	-1,9%	-1,9%	-0,3%	-0,5%	-1,9%	-7,0%	-0,6%	-0,4%
minimum	-0,4%	-1,9%	-1,9%	-0,3%	-0,5%	-1,9%	-7,0%	-0,6%	-0,4%

Tirages coordonnés

Variable	Nombre d'entreprises	Chiffre d'affaires	Total des achats	Salaires	Valeur ajoutée	EBE	Résultat comptable	Total de l'actif	Total du passif
Moyenne	0,0%	0,0%	0,0%	0,0%	0,0%	-0,1%	8,2%	0,0%	0,0%
maximum	0,2%	0,4%	0,4%	0,1%	0,2%	4,5%	491,5%	0,3%	0,3%
P99	0,2%	0,4%	0,4%	0,1%	0,2%	4,5%	491,5%	0,3%	0,3%
P95	0,1%	0,2%	0,2%	0,1%	0,1%	0,4%	0,9%	0,2%	0,2%
P90	0,1%	0,1%	0,1%	0,1%	0,1%	0,2%	0,3%	0,1%	0,1%
P75	0,0%	0,0%	0,0%	0,0%	0,0%	0,1%	0,1%	0,0%	0,0%
mediane	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
P25	0,0%	0,0%	0,0%	0,0%	0,0%	-0,1%	-0,2%	0,0%	0,0%
P10	-0,1%	-0,1%	-0,1%	-0,1%	-0,1%	-0,3%	-0,6%	-0,2%	-0,2%
P5	-0,1%	-0,2%	-0,2%	-0,1%	-0,1%	-1,5%	-2,0%	-0,3%	-0,3%
P1	-0,1%	-0,3%	-0,3%	-0,1%	-0,4%	-8,7%	-4,7%	-0,5%	-0,6%
minimum	-0,1%	-0,3%	-0,3%	-0,1%	-0,4%	-8,7%	-4,7%	-0,5%	-0,6%

Moyenne et distribution des biais relatifs pour les estimations de niveau sous-classe des principales variables fiscales du dispositif Esane en 2010, selon les scénarios de tirage

Tirages indépendants

Variable	Nombre d'entreprises	Chiffre d'affaires	Total des achats	Salaires	Valeur ajoutée	EBE	Résultat comptable	Total de l'actif	Total du passif
0 ou non-renseigné	0,0%	0,1%	0,1%	0,3%	0,0%	0,0%	-0,1%	0,0%	0,0%
1 à 5 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,2%	0,0%	0,0%	0,0%
6 à 9 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
10 à 19 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,1%	-0,1%	-0,1%	0,0%
20 à 29 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	-0,1%	0,0%	0,0%
30 à 49 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
50 à 99 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
100 à 199 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
200 salariés ou plus	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%

Tirages coordonnés

Variable	Nombre d'entreprises	Chiffre d'affaires	Total des achats	Salaires	Valeur ajoutée	EBE	Résultat comptable	Total de l'actif	Total du passif
0 ou non-renseigné	0,0%	0,1%	0,1%	-0,2%	-0,1%	0,0%	0,1%	-0,1%	0,0%
1 à 5 salariés	0,0%	0,1%	0,1%	0,0%	0,0%	0,1%	-0,2%	0,0%	0,0%
6 à 9 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
10 à 19 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	-0,1%	-0,2%	0,0%	-0,1%
20 à 29 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
30 à 49 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
50 à 99 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
100 à 199 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
200 salariés ou plus	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%

Biais relatifs pour les estimations par tranche d'effectifs des principales variables fiscales du dispositif Esane en 2009, sur le secteur du commerce de gros, selon les scénarios de tirage

Tirages indépendants

Variable	Nombre d'entreprises	Chiffre d'affaires	Total des achats	Salaires	Valeur ajoutée	EBE	Résultat comptable	Total de l'actif	Total du passif
0 ou non-renseigné	0,0%	0,0%	0,0%	-0,5%	-0,1%	0,3%	-0,1%	0,1%	0,1%
1 à 5 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,1%	0,1%	0,0%	0,0%
6 à 9 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,1%	0,0%	0,0%	0,0%
10 à 19 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,1%	0,2%	0,0%	0,0%
20 à 29 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,1%	0,0%	0,0%	0,0%
30 à 49 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
50 à 99 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
100 à 199 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
200 salariés ou plus	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%

Tirages coordonnés

Variable	Nombre d'entreprises	Chiffre d'affaires	Total des achats	Salaires	Valeur ajoutée	EBE	Résultat comptable	Total de l'actif	Total du passif
0 ou non-renseigné	0,0%	-0,1%	-0,1%	-0,1%	0,0%	0,1%	0,2%	0,0%	0,0%
1 à 5 salariés	0,0%	-0,1%	-0,1%	0,0%	-0,1%	-0,2%	0,0%	-0,1%	-0,1%
6 à 9 salariés	0,0%	0,1%	0,1%	0,0%	0,1%	0,3%	0,2%	0,0%	0,0%
10 à 19 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,1%	0,0%	0,0%
20 à 29 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
30 à 49 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,1%	0,0%	0,0%	0,0%
50 à 99 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
100 à 199 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
200 salariés ou plus	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%

Biais relatifs pour les estimations par tranche d'effectifs des principales variables fiscales du dispositif Esane en 2010, sur le secteur du commerce de gros, selon les scénarios de tirage

Tirages indépendants

Variable	Nombre d'entreprises	Chiffre d'affaires	Total des achats	Salaires	Valeur ajoutée	EBE	Résultat comptable	Total de l'actif	Total du passif
0 ou non-renseigné	0,0%	-0,5%	-0,5%	-0,7%	-0,3%	-1,6%	-1,3%	-0,2%	-0,2%
1 à 5 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,1%	0,1%	0,0%	0,0%
6 à 9 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	-0,1%	0,0%	0,0%	0,0%
10 à 19 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	-0,1%	-0,1%	0,0%	0,0%
20 à 29 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
30 à 49 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
50 à 99 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
100 à 199 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
200 salariés ou plus	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%

Tirages coordonnés

Variable	Nombre d'entreprises	Chiffre d'affaires	Total des achats	Salaires	Valeur ajoutée	EBE	Résultat comptable	Total de l'actif	Total du passif
0 ou non-renseigné	0,0%	0,3%	0,3%	0,2%	0,0%	2,4%	2,6%	0,1%	0,2%
1 à 5 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,2%	0,0%	0,0%	0,0%
6 à 9 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
10 à 19 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,1%	0,1%	0,0%	0,0%
20 à 29 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
30 à 49 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
50 à 99 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
100 à 199 salariés	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
200 salariés ou plus	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%

Biais relatifs pour les estimations par tranche d'effectifs des principales variables fiscales du dispositif Esane en 2011, sur le secteur du commerce de gros, selon les scénarios de tirage