LA COORDINATION DES ÉCHANTILLONS D'ENQUÊTES ENTREPRISES ET ÉTABLISSEMENTS : UNE NOUVELLE MÉTHODE DÉVELOPPÉE A L'INSEE.

Olivier Sautory[1], Emmanuel Gros[2] & Fabien Guggemos[3]

[1]*Insee, DMCSI, 18 bd A. Pinard, 75675 Paris Cedex 14, olivier.sautory@insee.fr*
[2]*Insee, DMCSI, 18 bd A. Pinard, 75675 Paris Cedex 14, emmanuel.gros@insee.fr*
[3]*Insee, DSDS, 18 bd A. Pinard, 75675 Paris Cedex 14, fabien.guggemos@insee.fr*

Résumé.

Le papier présente la nouvelle méthode de coordination des échantillons des enquêtes auprès des entreprises développée à l'Insee. En particulier, la coordination négative a pour objectif 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 : cela permet de réduire la charge statistique imposée aux *petites* entreprises – les *grandes* entreprises étant systématiquement enquêtées dans la plupart des enquêtes. Cette méthode, utilisant des nombres aléatoires permanents attribués aux unités, est fondée sur la notion de *fonction de coordination*, définie pour chaque unité et chaque nouveau tirage, qui transforme ces nombres aléatoires permanents. Elle est utilisée de façon opérationnelle depuis la fin 2013.

Mots-clés.

Coordination d'échantillons, nombres aléatoires, charge de réponse, échantillons stratifiés.

1 La méthode de coordination d'échantillons d'enquêtes-entreprises de l'Insee

On indique ici les grands principes de la méthode, présentée en détail dans Guggemos et Sautory (2012), en se limitant au cas du tirage aléatoire simple stratifié, qui est l'échantillonnage le plus souvent utilisé à l'Insee pour les enquêtes-entreprises. Cette méthode a été proposée par Ch. Hesse (2001) et étudiée par P. Ardilly (2009).

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 déterministe $g_{k,t}$, « judicieusement choisie » (voir § 1.2) et 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… $N_{(h,t)}$.

Justification : 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 »[1] d'une unité k pour l'enquête t[2]. 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 quelconques 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 quelconques du nombre aléatoire Ω_k.

[1] i.e. si elle est interrogée (et si elle répond…)
[2] cette charge potentielle sera souvent supposée identique pour toutes les unités d'une enquête donnée

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 de [0,1[. On note C la fonction de charge cumulée, supposée mesurable bornée : $\omega \in [0,1[\rightarrow C(\omega) \in \mathbb{R}$. On veut lui associer une fonction de coordination g telle que :

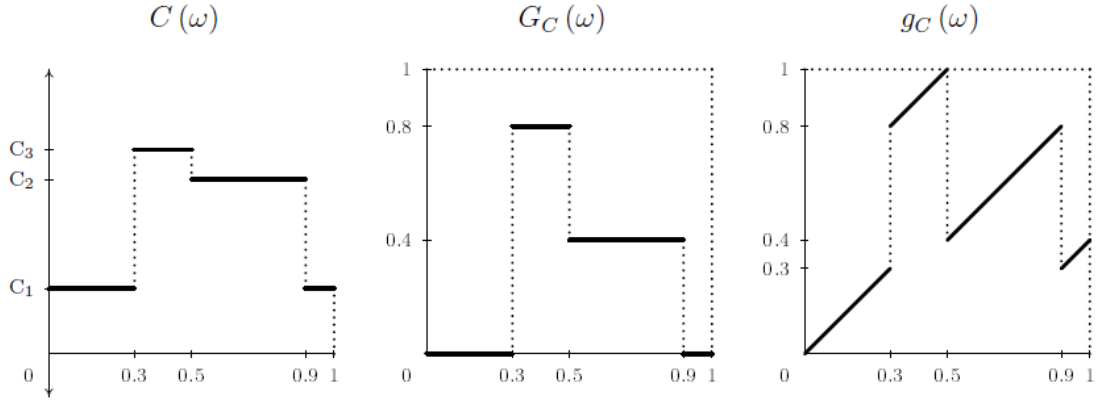
$$C(\omega^{(1)}) < C(\omega^{(2)}) \Rightarrow g(\omega^{(1)}) \leq g(\omega^{(2)}) \quad (2')$$

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

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

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[3]. 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[4], on montre que :

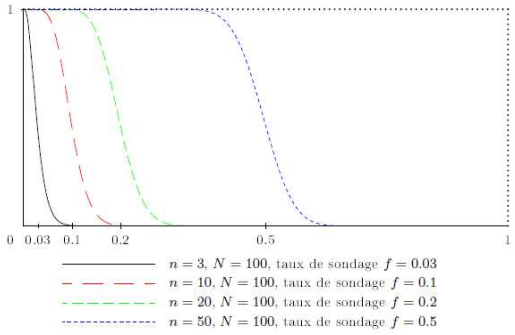
$$I^a_{k,t}(\omega) = P\left(k \in S_t \middle| g_{k,t}(\Omega_k) = g_{k,t}(\omega)\right) = b_{k,t}\left(g_{k,t}(\omega)\right)$$

[3] en omettant l'indice de strate.

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

où $1 - b_{k,t}$ 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.



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 implique que la fonction de charge cumulée soit 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 : $\Gamma_{k,t} = \sum_{u=1}^{t} \gamma_{k,u}\, \mathbb{1}(k \in S_u)$. C'est ce point qui garantit en effet que les fonctions de coordination sont construites de façon déterministe (elles ne dépendent pas de la réalisation observée **ω** du vecteur aléatoire **Ω**) ; de ce fait, les probabilités de sélection des unités dans l'échantillon d'une enquête ne dépendent pas de l'*appartenance ou non* de ces unités aux échantillons des enquêtes précédentes, mais des *probabilités* pour qu'elles y aient été sélectionnées.

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 :

1. On divise l'intervalle [0,1[en L[5] intervalles de longueurs égales $I_\ell = \left[\frac{\ell-1}{L}; \frac{\ell}{L}\right[$ $\ell = 1 \ldots L$.
2. 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.
3. On calcule la valeur moyenne $\beta_{k,t}(\ell)$ de $\tilde{b}_{k,t}$ sur chaque intervalle I_ℓ.
4. On définit la fonction $\beta_{k,t}$ par : $\forall \omega \in I_\ell \quad \beta_{k,t}(\omega) = \beta_{k,t}(\ell)$.

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

$\beta_{k,t}$ 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.

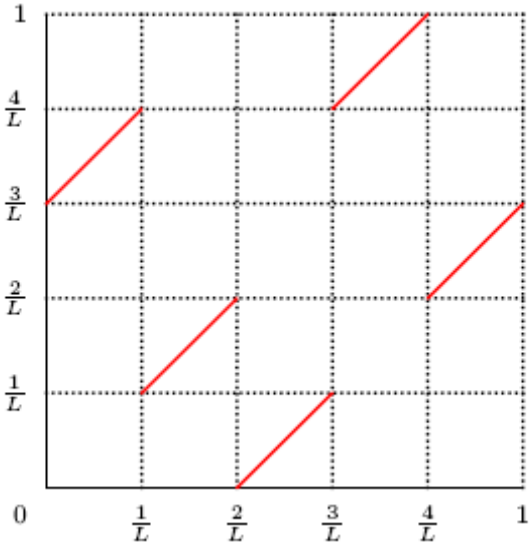
1.3.3 Construction de la fonction de coordination

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 :

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



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.

(a) 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…L} associée à cette fonction.

(b) 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}.

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

2 Simulations et tests grandeur nature

Des premiers tests empiriques obtenus par simulations sont présentés dans [3], ils se sont révélés très satisfaisants : la nouvelle méthode de coordination négative s'est ainsi révélée à la fois très efficace – en conduisant à des gains considérables, par rapport à des tirages indépendants, en termes de répartition de la charge réelle d'enquête 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.

Les résultats suivants sont extraits d'une note d'Emmanuel Gros (2014).

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

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 sur les différentes unités de la population, on a 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 sur des unités légales[6] :
  - en respectant les plans de sondage mis en œuvre lors des tirages effectifs de ces enquêtes : critères de stratification et allocations, renouvellement par moitié, par tiers ou quart de certains échantillons, coordination positive[7] 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 avec l'ensemble des enquêtes passées[8].

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 pas de problème :

- les temps de calculs restent raisonnables : environ 8 heures pour la séquence complète de tirage des 20 enquêtes ;

[6] Il s'agit des enquêtes TIC 2010, IPEA 2010, Acemo-TPE 2010, ESA 2009, Points de vente 2010, SINE 2010, TIC 2011, ESA 2010, IPEA 2011, CVTS4, Acemo-TPE 2011, CIS 2010, ENDD 2011, TIC 2012, Qualité énergétique mise en œuvre par les entreprises dans les bâtiments 2012, ESA 2011, Acemo-TPE12, CAM 2012 et TIC-TPE 2012.

[7] Voir plus loin comment est mise en œuvre la coordination positive.

[8] Dans ces simulations, la charge attribuée à chaque enquête est constante et égale à 1.

– 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.

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 sans effet sur les strates exhaustives[9], 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.

On observe un resserrement de la distribution autour de 1, i.e. un étalement de la charge totale d'enquête sur l'ensemble des unités : 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 découle de l'existence de parties conservées d'un millésime à l'autre pour les enquêtes à échantillons rotatifs. 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

[9] Strates contenant les unités incluses d'office dans l'échantillon, en général les unités ayant un effectif salarié au-delà d'un certain seuil.

Notons par ailleurs que cette méthode 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 présentes 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

Comment la méthode proposée peut-elle être utilisée pour coordonner des échantillons d'enquêtes portant sur différents types d'unités « emboîtés » – établissements, unités légales, entreprises, groupes – dans un système de coordination « intégré » ? On se place par la suite dans le cas le plus classique, à savoir la coordination d'enquêtes unités légales et d'enquêtes établissements, sachant que la méthodologie présentée ci-dessous peut s'appliquer à toute coordination sur des unités « emboîtées ».

La validité de la procédure de coordination présentée ici est subordonnée au fait que le numéro aléatoire d'une unité k, utilisé à chaque tirage d'enquête lors de la construction de la fonction de coordination, 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 sont générés aléatoirement selon une loi uniforme sur [0;1]. La réunion de ces deux conditions conduit alors nécessairement à ne pouvoir « mettre en correspondance » dans le système intégré une unité légale qu'avec un seul de ses établissements.

La procédure adoptée pour la coordination d'enquêtes unités légales (UL) et établissements est donc la suivante :

- on génère les nombres aléatoires permanents pour les établissements, tirés indépendamment selon une loi uniforme sur [0;1], et **on attribue à chaque unité légale le numéro aléatoire de son établissement principal**[10]. 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 $\leftrightarrow$ établissement principal] entre ces deux jeux ;

[10] qui sera souvent le siège social lors de la création de l'unité légale

– la coordination entre échantillons d'unités légales et échantillons d'établissements s'effectue selon le schéma suivant :

(a) Tirage d'un échantillon d'unités légales

Pour chaque UL, on fait intervenir, dans le calcul de sa charge cumulée[11], les fonctions de charges de son établissement principal dans les différentes enquêtes établissements avec lesquelles on souhaite coordonner l'enquête unité légale en cours.

Conséquence : les charges d'enquête des établissements non principaux ne sont pas prises en compte dans le système de coordination.

(b) Tirage d'un échantillon d'établissements

Pour l'établissement principal d'une UL, on fait intervenir, dans le calcul de sa charge cumulée, les fonctions de charges de l'unité légale dans les différentes enquêtes unité légales avec lesquelles on souhaite coordonner l'enquête établissements en cours.

Conséquence : les charges d'enquêtes des unités légales sont prises en compte uniquement dans la charge des établissements principaux.

Des simulations similaires à celles présentées au §2.2, en ajoutant 8 enquêtes établissements aux 20 enquêtes unités légales, montrent l'efficacité, en termes de répartition de la charge d'enquête, du système de coordination « intégré », ou « multi-niveaux », présenté ci-dessus, par rapport à des systèmes de coordination « séparés » unités légales et établissements.

2.3 Étude de deux questions d'ordre méthodologique

2.3.1 Le problème du biais de rétroaction

On peut montrer que 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[12]. Ce phénomène, appelé « biais de rétroaction », se révèle *a priori* particulièrement problématique car 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. En particulier, 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 structurelles ESA et EAP (enquête annuelle de production).

Des simulations sur données réelles ont montré que le biais de rétroaction potentiel, lié à la coordination avec les enquêtes ESA et EAP, est suffisamment faible pour pouvoir être négligé[13]. Ce résultat permet d'inclure les enquêtes ESA et EAP dans la procédure de coordination globale des enquêtes entreprises.

[11] plus précisément la « fonction de charge cumulée espérée approchée »…

[12] Pour une démonstration mathématique de ce résultat, on peut se reporter à l'annexe 1 du document de travail E9908 : *Sampling coordination : a review by country*, Hesse Christian (Insee, 1999).

[13] 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…

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

Il n'est pas rare que le tirage des unités au sein de chaque strate ne soit pas un sondage aléatoire simple *stricto sensu*, mais un 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, est 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 dans la méthode de tirage coordonné le critère jadis « contrôlé » par le tirage systématique :

- on ajoute un niveau de stratification supplémentaire défini par ledit critère (préalablement discrétisé le cas échéant, par exemple s'il s'agit d'un chiffre d'affaires) ;
- on passe des allocations relatives à la stratification initiale aux allocations relatives à la stratification finale, plus fine, par « allocations proportionnelles », au prix si nécessaire d'un regroupement des nouvelles strates « trop fines », afin d'obtenir des allocations non nulles pour les strates de tirage finales.

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. Des simulations ont montré que cette augmentation n'affecte pas la qualité de la coordination.

Bibliographie

[1] Hesse, Ch. (2001), Généralisation des tirages aléatoires à numéros aléatoires permanents, ou la méthode JALES, *document de travail Insee E0101*.

[2] Ardilly, P. (2009), Présentation de la méthode JALES+ conçue par Christian Hesse, *document de travail interne Insee*.

[3] Guggemos, F. & Sautory, O. (2012), La coordination d'échantillons d'enquêtes auprès des entreprises mise en place à l'Insee, *11^{e} Journées de méthodologie statistique de l'Insee*.

[4] Gros, E. (2014), Résultats des simulations et tests grandeur nature de la nouvelle méthode de coordination pour les échantillons des enquêtes Entreprises, *note interne Insee*.