La coordination des enquêtes entreprises et établissement à l'Insee

Études et simulations relatives à la nouvelle procédure de coordination développée à l'Insee





Gros Emmanuel
Insee – DMCSI – Division Sondages

Insee Mesurer pour comprendre

La coordination d'échantillon

- L'objectif de la coordination d'échantillon est de prendre en compte les échantillons d'enquêtes précédentes lors de la sélection d'un nouvel échantillon…

- … dans une optique de réduction de la charge statistique imposée aux petites entreprises (les grandes étant en général interrogées de manière exhaustive)…

- … tout en conservant le caractère sans biais des estimations.

- Deux types de coordination :

  - ✓ coordination négative (la plus fréquemment utilisée à l'Insee) : favoriser la sélection d'entreprises non interrogées lors des enquêtes récentes → minimise le recouvrement entre échantillons ;

  - ✓ coordination positive : favoriser la sélection d'entreprises déjà interrogées lors d'une enquête précédente → maximise le recouvrement entre échantillons.

La nouvelle procédure de coordination à l'Insee

- Une méthode basée sur des numéros aléatoires permanents…

  ↳ Chaque unité de la population se voit attribuer un numéro aléatoire permanent ω_k, tiré dans la loi de probabilité uniforme sur [0,1[, indépendamment d'une unité à l'autre.

- … qui repose sur le concept de fonction de coordination:

  ↳ fonction de coordination g = fonction mesurable de [0;1] dans [0;1], qui conserve la loi uniforme :

  si P est la loi uniforme sur [0;1], alors on a $P^g=P$.

  → si les $(\omega_k)_{k\in U}$ sont des numéros aléatoires tirés indépendamment dans la loi P, alors les numéros aléatoires transformés $(g(\omega_k))_{k\in U}$ sont eux-mêmes tirés indépendamment dans la loi P.

↳ Sélection d'un échantillon par sondage aléatoire simple stratifié: à l'aide d'une fonction de coordination g_k « judicieusement choisie », sélectionner dans chaque strate h de taille N_h les n_h plus petits numéros aléatoire transformés $g_k(\omega_k)$.

Les fonctions de coordination

- Pour le tirage d'un échantillon donné, la fonction de coordination g_k va prendre en compte la charge de réponse cumulée de l'unité k pour répondre à l'objectif de coordination négative ou positive :

  ↳ pour une coordination négative (resp. positive), g_k est définie de telle sorte que plus l'unité k présente une charge de réponse cumulée élevée, plus le numéro $g_k(\omega_k)$ sera élevé (resp. faible).

- Les fonctions de coordination des enquêtes passées servent de base pour calculer une charge de réponse cumulée…

- … permet ensuite de déterminer la fonction de coordination de chaque unité pour le tirage en cours.

$$\left.\begin{array}{ll} g_{k,1}(\omega) & \gamma_{k,1}\left(g_{k,1}(\omega)\right) \\ \vdots & \Rightarrow \vdots \\ g_{k,t-1}(\omega) & \gamma_{k,t-1}\left(g_{k,t-1}(\omega)\right) \end{array}\right\} \Rightarrow \Gamma_{k,t-1}(\omega) = \sum_{u \le t-1} \gamma_{k,u}\left(g_{k,u}(\omega)\right) \Rightarrow g_{k,t}(\omega)$$

Une méthode complète et efficace

- Une méthode très complète :
  - ✓ permet de coordonner aussi bien positivement que négativement ;
  - ✓ permet de coordonner plusieurs enquêtes entre elles…
  - ✓ … tout en différentiant les charges de réponse assignées à chaque enquête ;
  - ✓ permet de coordonner des échantillons de niveaux différents (unités légales et établissements par exemple).
- De premières évaluations de la méthode menées sur données simulées ont prouvé son efficacité en termes de répartition de la charge de réponse sur les différentes unités de la population…
- … ainsi que sa robustesse vis à vis des paramètres des différents plans de sondage : taux de sondage, différence de stratification entre enquêtes, taux de recouvrement entre le champ des différentes enquêtes, charges associées aux différentes enquêtes, etc.

Test de la procédure en situation de production (1)

- Tirages **coordonnés** de vingt enquêtes successives sur des unités légales à partir de l'enquête sectorielle annuelle (ESA) de 2008 :

  - ✓ 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** 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 (charge associée à chaque enquête fixée à 1).

➔ A permis de valider la **faisabilité opérationnelle** de la méthode.

- 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.

Test de la procédure en situation de production (2)

→ Meilleure répartition de la charge d'enquête entre les différentes unités de la population lorsque les tirages sont coordonnés

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

Distribution de la charge d'enquête, hors parties exhaustives, selon le scénario de tirage retenu.

Test de la procédure en situation de production (3)

→ Encore plus net en excluant du calcul de la charge d'enquête les parties conservées des différents échantillons.

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

Distribution de la charge d'enquête, hors parties exhaustives et parties conservées, selon le scénario de tirage retenu

Coordination d'échantillons de niveaux différents

- Cette méthode permet de coordonner des échantillons de niveaux différents, par exemple unités légales et établissements, selon la procédure suivante :

  ❶ pour chaque unité légale, définition d'un **lien permanent** entre l'unité légale et l'un de ses établissements ;

  ❷ génération de numéros aléatoires pour les établissements, et **attribution à chaque unité légale du numéro aléatoire de son établissement principal**.

➔ Coordination « multi-niveaux » :

- ✓ Pour le tirage d'un échantillon d'unités légales : pour chaque unité légale, prise en compte dans le calcul de sa charge cumulée des fonctions de charge de son établissement principal ;

- ✓ Réciproquement, pour le tirage d'un échantillon d'établissements : pour chaque établissement principal, prise en compte dans le calcul de sa charge cumulée des fonctions de charge de son unité légale.

Test de la coordination « multi-niveaux »

- Reprise et enrichissement de la simulation précédente :
  - ✓ ajout de 8 enquêtes établissements aux 20 enquêtes unités légales de la séquence de tirage ;
  - ✓ trois plans de coordination différents : tirages indépendants, tirages coordonnés « séparés » et tirages coordonnés « multi-niveaux ».

→ Résultats cohérents qui confirment les analyses précédentes.

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

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 conservées, selon le scénario de tirage retenu

Impact du paramètre de charge

- Reprise de la simulation précédente, en distinguant enquêtes « légères » (charge = 1) et enquêtes lourdes (charge = 2)

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	4 670 676	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 et plus	385	80	41	-305	-39	-344

Le problème du biais de rétroaction

- En théorie, si les résultats d'une enquête A servent à mettre à jour la base de sondage d'une enquête B postérieure à A et coordonnée avec l'enquête A, alors l'échantillon de B conduit à des estimations biaisées.

➔ **Problème :** les enquêtes du SSP sont en général tirées dans Sirus, qui est mis à jour à partir des résultats de ces enquêtes. En particulier, codes APE actualisés à partir des ESA et EAP…

↳ Simulations sur données réelles à partir des données d'Esane :

- ✓ Séquences de tirage « ESA08 CG → ESA09 CG → ESA10 CG → ESA11 CG » : 5000 tirages indépendants et 5000 avec coordination négative ;
- ✓ Calcul de biais relatifs sur des estimations sectorielles et par tranche de taille à partir de données fiscales disponibles pour toutes les unités.

➔ Biais de rétroaction suffisamment faible pour pouvoir être négligé

Tirage systématique & « sur-stratification »

➢ **Tirage systématique** sur données triées fréquemment utilisé à l'Insee mais **incompatible** avec la procédure de coordination.

↳ Prise en compte du critère auparavant « contrôlé » par tirage systématique via une « sur-stratification » :

✓ ajout d'un niveau de stratification supplémentaire défini par ledit critère ;

✓ passage des allocations relatives à la stratification initiale à celles relatives à la stratification finale par « allocations proportionnelles ».

➔ Faible impact sur la qualité de la coordination.

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 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

Conclusions

- Une méthode très complète :
  - ✓ coordination négative et positive ;
  - ✓ différentiation des charges de réponse assignées à chaque enquête ;
  - ✓ coordination d'échantillons de niveaux différents.
- Les études menées aussi bien sur des données simulées que sur des données réelles en situation de production ont prouvé à la fois **efficacité** et la **robustesse** de cette méthode (ainsi que sa faisabilité opérationnelle…).
- Utilisée en production à l'Insee depuis fin 2013 (tirage ESA 2013).

➔ Pour un premier bilan pratique après 16 mois d'utilisation, cf. la contribution associée d'Anaïs Levieil-Guillon.

Merci de votre attention !

Contact :

Emmanuel Gros
Tél. : 01 41 17 64 91
Courriel : emmanuel.gros@insee.fr

Insee
18 bd Adolphe-Pinard
75675 Paris Cedex 14
www.insee.fr
Informations statistiques :
www.insee.fr / Contacter l'Insee
09 72 72 4000
(coût d'un appel local)
du lundi au vendredi de 9h00 à 17h00

Insee