

Une contrainte : collecter moins ; notre mission : le faire le mieux possible

O. Haag, P.-A. Pendoli, S. Faivre¹

Résumé

En France, les contraintes budgétaires rendent plus difficile l'embauche d'enquêteurs occasionnels pour prendre en compte des problèmes de collecte. Il devient donc nécessaire de respecter une quotité de travail annuelle préalablement fixée. Pour les enquêtes Insee, réalisées à partir d'un échantillon maître, les difficultés apparaissent lors de l'absence prolongée d'un enquêteur sur l'ensemble de la durée de la collecte d'une enquête. En effet, dans ces conditions une partie du territoire peut devenir non couverte par l'enquête, ce qui génère de fait un biais.

Afin de répondre à cette nouvelle difficulté, deux méthodes ont été mises en œuvre en fonction du moment où le problème est diagnostiqué.

- Si l'« abandon » de zone intervient avant ou juste au début de la collecte, une procédure dite de « sous-allocation » est mise en œuvre. Elle consiste à interroger un minimum de ménages dans chaque zone de collecte au détriment d'autres zones pour lesquelles aucun problème de collecte n'est a priori diagnostiqué. Il s'agit donc de minimiser la dispersion des poids sous contrainte de respect de la charge de collecte.
- Si l'« abandon » de zone intervient en cours de collecte, une priorisation des enquêtes restantes est mise en œuvre. Elle se base sur le R-indicateur (indicateur de Représentativité) qui permet de mesurer le degré de similarité d'un échantillon par rapport à la population de base. L'objectif de cette priorisation en cours de collecte est de s'approcher le plus possible au final d'une équi-probabilité de réponse des répondants. Il est basé sur la dispersion des probabilités de réponse estimées des ménages échantillonnés, et se décline en R-indicateurs partiels mesurant cette représentativité variable par variable. Ces R-indicateurs sont des outils permettant d'analyser la collecte en isolant des groupes de populations sous-représentées. Il est possible d'intensifier les efforts de collecte sur les groupes précédemment identifiés.

Lors de la présentation orale, les deux points avaient été évoqués succinctement. Toutefois, cet article ne traite que du premier point évoqué ci-dessus, à savoir la « sous-allocation ». La priorisation est en cours de mise en œuvre pour la première fois à l'Insee pour l'enquête Patrimoine et elle donnera lieu à un article spécifique qui sera rédigé par A. Rebecq.

Mots Clés : Collecte, échantillonnage, R-indicateurs

1. Rappel sur l'échantillonnage des enquêtes ménages à l'INSEE (Christine et Faivre, 2009)

Il est usuel en statistique d'enquêtes de recourir au tirage d'échantillons à deux degrés dans le cas d'enquêtes en face-à-face. Les données sont collectées par un enquêteur dont une partie du temps de travail est constituée des déplacements entre son domicile et celui du ménage ou de l'individu à interroger. L'intérêt de l'échantillonnage à deux degrés est de concentrer l'échantillon de ménages ou d'individus dans un nombre limité de zones géographiques peu étendues afin :

- de réduire les coûts de collecte (déplacements, temps de repérage) de l'enquête qui dispose d'un budget contraint ;
- d'embaucher un réseau fixe d'enquêteurs qui reçoivent une formation aux techniques d'enquête et acquièrent une expérience dans le domaine, toutes deux mises à profit sur le long terme.

Cette méthode d'échantillonnage des enquêtes nécessite :

- de définir et construire ces zones géographiques (appelées unités primaires) ;
- de sélectionner les zones dans lesquelles concentrer le tirage des ménages de manière à obtenir un échantillon représentatif d'unités primaires (1^{er} degré du tirage) ;

¹ Institut National de la Statistique et des études Économiques, France.

- et de mettre en œuvre le tirage des ménages ou des individus, i.e. les unités secondaires, dans ces zones sélectionnées (2nd degré du tirage).

1.1 Les zones d'action enquêteur (ZAE), unités primaires des enquêtes ménages de l'INSEE

La construction des ZAE intègre la méthodologie du Nouveau Recensement. En particulier, on distingue les petites et les grandes communes.

- Une grande commune (GC) est une commune de 10 000 habitants ou plus. Elle est en partie (8% des logements) interrogée chaque année de Recensement de la Population (RP), pour un cycle de cinq ans du RP. **Une grande commune forme à elle seule une ZAE.**
- Une petite commune (PC) est une commune de moins de 10 000 habitants. Une petite commune est affectée à une année de Recensement donnée, pour un cycle de cinq ans du RP.

Une ZAE PC est un regroupement de petites communes qui satisfait les critères suivants :

1. Comporter au moins une commune de chaque année de Recensement ;
2. Comporter au moins 300 résidences principales de chaque année de Recensement ;
3. Appartenir à une seule région administrative (il y en a 22 en France métropolitaine) ;
4. Avoir une étendue ne dépassant 20 km de rayon.

1.2 Tirage de 1^{er} degré : constitution de l'Échantillon-Maître

Parmi les 3 832 ZAE qui partitionnent le territoire métropolitain, 567 ZAE sont sélectionnées aléatoirement (Guggemos, 2009).

1.3 Tirage de 2nd degré : les logements

Selon le thème de l'enquête, l'unité statistique peut être le ménage ou l'individu. Dans les deux cas, l'unité statistique est atteinte via son logement.

Les grands principes sont les suivants :

- Tirage des logements dans les ZAE de l'Échantillon-Maître. Ces logements tirés sont communément appelés « fiches adresses » pour les enquêteurs ;
- Calcul d'allocations théoriques par ZAE dont l'objectif est de minimiser la dispersion des poids de sondage des logements ;
- Définition éventuelle d'une borne minimum et d'une borne maximum d'enquêtes à réaliser par ZAE afin de prendre en compte la charge des enquêteurs.

2. Les nouvelles conditions d'emploi des enquêteurs (NCEE)

L'entrée en vigueur des NCEE en 2013 impose à l'Institut de faire correspondre le temps de collecte annuel d'un enquêteur à une durée annuelle d'emploi préalablement fixée. Ce temps de collecte comprend :

- Le temps total de déplacement. Celui-ci est fonction :
 1. du nombre de logements à interroger ;
 2. du temps nécessaire pour atteindre la zone d'enquête (y compris le temps de repérage de l'adresse) ;
 3. du nombre moyen de contacts nécessaires avant d'obtenir un rendez-vous.
- Le temps d'enquête total chez les ménages. Celui-ci dépend :
 1. du nombre de ménages à interroger ;
 2. du taux de réponse ;
 3. du temps de passation d'un questionnaire.

Chaque année une planification est donc réalisée afin d'attribuer l'ensemble des échantillons des enquêtes de l'année

suivante aux enquêteurs en respectant leur temps de travail disponible.

Toutefois, par la suite, en cas d'absence prolongée d'un enquêteur sur une période de collecte, il est possible que pour une enquête donnée et une ZAE donnée, très peu voire aucun entretien ne puisse être réalisé. On parle alors de ZAE « orpheline ».

Des marges de manœuvre existent pour prendre en charge ces entretiens dans la ZAE « orpheline ». Il s'agit principalement :

- d'un recours à une réserve, destinée aux remplacements, du temps de travail d'autres enquêteurs du réseau² ;
- d'un recrutement d'un enquêteur occasionnel (sous réserve de disponibilités du plafond d'emploi).

Il peut cependant arriver que, une fois explorées toutes ces possibilités, il ne soit pas possible de dégager de moyens de collecte supplémentaires pour prendre en charge en l'état tout ou partie des entretiens dans la ZAE « orpheline ».

Comme l'abandon complet des enquêtes dans une unité primaire est à proscrire, car il créerait par la suite un biais³ qu'il n'est pas possible de corriger au moyen des traitements post-collecte habituels, il est nécessaire de collecter un minimum d'entretiens dans chaque ZAE de l'Echantillon-Maître.

Dans ce cas, la méthodologie mise en œuvre est :

- de réaliser une partie des entretiens prévus initialement sur la ZAE « orpheline ». Ils doivent être attribués à d'autres enquêteurs du réseau, que l'on appellera « enquêteurs remplaçants » ;
- de décharger ces enquêteurs remplaçants d'une partie des entretiens qu'ils devaient réaliser initialement sur leur ZAE d'origine afin de respecter leur temps de travail affecté à cette enquête⁴.

Se pose alors la question de la méthode à retenir quant à la sélection des logements finalement enquêtés, à la fois dans la ZAE « orpheline » et dans les ZAE d'origine des enquêteurs remplaçants.

C'est pourquoi des travaux ont été mis en œuvre afin de :

- définir une méthodologie qui permette de limiter au maximum la perte de qualité de l'enquête liée à la diminution du nombre d'entretiens réalisés sur le terrain,
- en respectant la contrainte d'interroger un nombre minimum de ménages dans chaque ZAE.

Deux cas de figure peuvent se présenter :

- Un enquêteur est indisponible avant la collecte ou en début de collecte d'une enquête et durant la totalité de la durée de la collecte. La méthode mise en œuvre dans ce cas est celle de la sous-allocation, présentée dans ce document.
- L'enquêteur arrête brusquement sa collecte en cours d'enquête alors qu'elle est déjà largement avancée, ou ne revient pas en cours de collecte alors qu'il était prévu qu'il le fasse. Une méthodologie différente doit être utilisée afin de prioriser l'interview de certains types de ménages.

3. Le recours à la technique de sous-allocation

3.1 Contexte

Avant la collecte ou en début de collecte d'une enquête, on a pris connaissance d'une absence prolongée d'un

² En 2014, la planification initiale ne prenait en compte que 95% de la quotité des enquêteurs. Les 5% restants étaient destinés à la réserve.

³ Des logements de la base de sondage ayant eu une probabilité nulle de figurer dans l'échantillon des répondants.

⁴ Bien entendu, comme en général le temps de déplacement vers la ZAE « orpheline » sera supérieur au temps de déplacement de l'enquêteur vers sa ZAE d'origine, on sera contraint de supprimer plus d'entretiens dans sa ZAE d'origine qu'on ne lui en confiera dans la ZAE « orpheline ».

enquêteur sur la totalité de la durée de la collecte, rendant la ZAE « orpheline ».

3.2 La formalisation du problème à résoudre (Rebecq, 2014)

Critères d'optimisation :

1. **Minimiser la dispersion des poids de sondage** des logements de l'échantillon final (i.e. l'échantillon qui sera finalement collecté) dans la région concernée.
2. **Réduire le moins possible la taille de l'échantillon.**

Contraintes :

Trois contraintes ont été imposées dans le modèle. Elles sont décrites ci-dessous. Alors que la première découle directement du respect des temps de travail, les deux autres ont été ajoutées pour limiter l'impact de la modification de l'échantillon sur les conditions de travail des enquêteurs concernés.

1. Conserver à l'identique le temps de travail des enquêteurs affectés à cette opération.
Hypothèses sur le temps de travail des enquêteurs qui comprend, comme décrit plus haut :
 - *Un temps de déplacement.* On suppose que celui-ci est uniquement fonction de la distance entre le domicile de l'enquêteur et la commune principale de la ZAE de collecte. L'hypothèse implicite est que le nombre moyen de contacts par ménage est le même dans la ZAE d'origine de l'enquêteur remplaçant et dans la ZAE « orpheline », et que les temps de repérage sont identiques, bien que l'enquêteur soit moins habitué à la ZAE « orpheline ». Cette hypothèse est donc un peu optimiste en termes de quotité consommée.
 - *Un temps d'enquête.* On suppose qu'il reste constant. Cette hypothèse est en revanche pessimiste dans la mesure où, comme on l'a vu, l'enquêteur se verra confier in fine moins d'entretiens qu'au départ. Toutes choses égales par ailleurs, il réalisera donc moins d'entretiens et son temps d'enquête sera plus faible qu'à l'origine.
 Sous ces hypothèses, la contrainte porte uniquement sur la conservation du temps total de déplacement.
2. *In fine*, un enquêteur remplaçant n'aura pas à sa charge davantage d'entretiens dans la ZAE « orpheline » que dans sa ZAE d'origine.
3. *In fine*, un enquêteur conservera à sa charge au moins 50 % des entretiens qui lui ont été attribués initialement dans sa ZAE d'origine.

Les informations nécessaires pour mettre en œuvre la méthode :

- Connaître la ZAE « orpheline » avant la collecte ou en début de collecte ;
- connaître les enquêteurs remplaçants potentiellement mobilisables ;
- connaître le temps de travail alloué à l'enquête pour chacun de ces enquêteurs ;
- connaître les temps de déplacement de ces enquêteurs remplaçants entre leur domicile et la commune principale de la ZAE « orpheline » d'une part, et entre leur domicile et la commune principale de leur ZAE habituelle d'autre part.

Formalisation du problème :

Le programme de minimisation région par région

$$(1) \quad \underset{n_{ZAE}}{\text{Min}} \frac{1}{n_{reg}} * \sum_{ZAE \in reg} n_{ZAE} * (Poids_{ZAE}^{log} - Poids_{moy}^{log})^2$$

$$\text{où} \quad n_{reg} = n_{ZAE}^{orpheline} + \sum_{ZAE \in remplaçantes} n_{ZAE} + \sum_{ZAE \in reste \ reg} n_{ZAE}$$

$$Poids_{ZAE}^{log} = \omega_{ZAE} * \frac{N_{ZAE}}{n_{ZAE}}$$

Sous les contraintes par enquêteur

$$(2) \quad n_{ZAE\ rempl}^{initial} * T_{ZAE\ rempl}^{enq} = n_{ZAE\ rempl}^{final} * T_{ZAE\ rempl}^{enq} + n_{ZAE\ orpheline}^{final} * T_{ZAE\ orpheline}^{enq}$$

$$(3) \quad n_{ZAE\ rempl}^{final} \geq \frac{n_{ZAE\ rempl}^{initial}}{2}$$

$$(4) \quad n_{ZAE\ rempl}^{final} \geq n_{ZAE\ orpheline}^{final}$$

3.3 Mise en œuvre pratique :

Les deux ou trois meilleurs scénarios (l'optimum du programme et les scénarios qui s'en approchent le plus) sont proposés aux responsables de la collecte dans la région concernée. Ils choisissent parmi ces scénarios celui qui est le plus adapté aux contraintes de collecte.

Il est à noter que, plus que la durée d'accès à la zone « orpheline », c'est le ratio entre la durée du trajet pour se rendre dans la ZAE d'origine et la durée du trajet pour se rendre dans la ZAE « orpheline » qui joue un rôle important dans le modèle. Car plus ce ratio est grand et plus la perte d'entretiens est importante.

Cela justifie la mise en place d'une communication adaptée vis-à-vis des enquêteurs. Pour chaque scénario proposé, une validation de la faisabilité sur le terrain par les équipes en charge de la collecte est nécessaire.

3.4 Illustration sur un exemple fictif : la ZAE de Vizille pour la collecte de l'enquête Logement 2013

Informations disponibles dans l'échantillon

- La ZAE de Vizille a une allocation de 32 ménages tirés pour l'enquête Logement 2013.
- La région Rhône-Alpes a une allocation totale de 2 591 ménages pour cette enquête.
- La dispersion des poids de sondage dans l'échantillon pour cette région vaut 1 954.
- La dispersion des poids de sondage finaux dans la région, si l'on ne réalisait qu'un entretien dans la ZAE « orpheline », vaudrait 102 039.

Tableau 3.4.1

Informations fournies par la Direction régionale de Rhône-Alpes chargée de superviser la collecte dans la ZAE de Vizille : la liste des enquêteurs mobilisables et leur temps de trajet

Enquêteur en charge de la ZAE de ...	Allocation initiale dans la ZAE	Temps de déplacement de l'enquêteur vers sa ZAE habituelle (en minutes)	Temps de déplacement estimé de l'enquêteur vers la ZAE de Vizille (en minutes)	Ratio des temps de déplacement
La Mure	52	10	25	2.5
Saint-Marcellin	43	29	55	1.9
Grenoble 1	45	27	47	1.7
Grenoble 2	24	16	36	2.25
Grenoble 3	10	10	25	2.5

Lecture : Dans l'échantillon de départ il y avait 52 enquêtes à réaliser dans la ZAE de La Mure. L'enquêteur en charge de la ZAE de La Mure met en moyenne 10 minutes pour atteindre son lieu d'enquête. Il met en revanche 25 minutes en moyenne pour réaliser une enquête dans la commune de Vizille, soit 2,5 fois plus de temps.

Tableau 3.4.2
Résultats obtenus en fonction des enquêteurs mobilisés

Enquêteur(s) remplaçant(s) mobilisé(s) initialement sur...	Taille de l'échantillon en Rhône-Alpes <i>a posteriori</i>	Dispersion des poids de sondage <i>a posteriori</i>
La Mure / Saint-Marcellin	2 544	1 955
La Mure / Saint-Marcellin / Grenoble	2 543	1 992
La Mure / Grenoble	2 540	2 059
La Mure	2 544	2 233
Saint-Marcellin	2 550	2 192
Grenoble	2 546	2 184
Aucun enquêteur remplaçant mobilisé + Hypothèse : 1 FA réalisée dans la ZAE « orpheline »	2 560	102 039

Lecture : En ne mobilisant que les enquêteurs de La Mure et de Saint-Marcellin pour aller réaliser des enquêtes dans la ZAE orpheline de Vizilles, la taille de l'échantillon final collectée sera de 2 544, et la sous-allocation conduit à une dispersion des poids de 1955. Cette dispersion est notre indicateur de qualité : plus il est petit, meilleure est la sous-allocation. Si on ne réalisait qu'une FA dans la zone orpheline, la taille de l'échantillon collecté serait de 2 560 (soit le maximum envisageable), mais la dispersion des poids serait alors la plus forte pour la région, soit 102 039.

Détail du scénario « La Mure / Saint-Marcellin » (c.-à-d. l'optimum)

Dans ce scénario, on mobilise les enquêteurs en charge habituellement des ZAE de la Mure et de Saint-Marcellin pour reprendre une partie des entretiens de la ZAE de Vizille « orpheline ». En contrepartie, ces deux enquêteurs remplaçants se voient retirer une partie des entretiens qui leur avaient été initialement attribués sur leur ZAE d'origine afin de maintenir leur temps de travail constant.

L'algorithme donne les sous-allocations suivantes :

- 15 FA sur 32 sont réalisées dans la ZAE de Vizille
- 35 FA sur 52 sont réalisées dans la ZAE de la Mure
- 30 FA sur 43 sont réalisées dans la ZAE de Saint-Marcellin
- Les enquêteurs grenoblois ne sont pas mobilisés. Il n'y a aucun changement pour les allocations dans les trois ZAE de Grenoble.

Dans chaque ZAE (la ZAE « orpheline » et celles des enquêteurs remplaçants), les ménages qui seront finalement enquêtés sont sélectionnés aléatoirement parmi les ménages tirés initialement dans l'échantillon.

Cette sous-allocation « coûte » donc 15 FA puisqu'on réalise 15 FA sur les 32 dans la zone orpheline, mais que l'on perd 30 FA dans les ZAE voisines.

3.5 Bilan de la mise en œuvre

Cette méthode a été mise en œuvre pour deux enquêtes Ménages de l'Insee.

- Pour l'enquête « Logements 2013 », 6 ZAE étaient orphelines (soit 189 logements). 8 enquêteurs remplaçants ont été mobilisés. Ils ont enquêté 95 logements dans les ZAE « orphelines », tandis que 138 entretiens leur ont été retirés dans leurs ZAE d'origine. Le coût de la méthode a donc été une perte de 43 logements supplémentaires, mais en évitant toute ZAE « orpheline ».
- Pour l'enquête « Conditions de Vie et Sécurité 2014 », 6 ZAE étaient orphelines (soit 212 logements). 10 enquêteurs remplaçants ont été mobilisés. Ils ont enquêté 97 logements dans les ZAE « orphelines », tandis que 176 entretiens leur ont été retirés dans leurs ZAE d'origine. Le coût de la méthode a donc été une perte de 79 logements supplémentaires, mais en évitant toute ZAE « orpheline ».

Elle est actuellement mise en oeuvre pour trois autres enquêtes :

- l'enquête sur les conditions de vie des ménages. Il s'agit d'un panel sur 9 ans. Pour ce dernier, comme les interrogations des vagues 6 à 9 ne sont plus obligatoires, ce sont principalement parmi ces dernières qu'ont été retirées les enquêtes à abandonner dans les "ZAE remplaçantes". En revanche, les enquêtes de la vague 1 des "ZAE orphelines" ont toutes été traitées ;
- l'enquête Patrimoine ;
- une enquête sur la mesure des ressources des jeunes.

4. Conclusion

Cette méthode d'adaptation de la collecte est au cœur de l'épineux débat entre biais et variance. Elle privilégie *a priori* la réduction du biais mais s'accompagne d'une augmentation de la variance des résultats obtenus.

En effet, elle vise principalement à éviter des biais de couverture dans la collecte difficilement corrigeables d'une part, et à favoriser des traitements post collecte de type calage sur marges en essayant de réduire le plus possible la dispersion des poids des individus répondants avant la phase de traitement de la non-réponse totale d'autre part.

Toutefois, elle se traduit *in fine* par une réduction du nombre de répondants, et donc une augmentation de la variance des résultats. Cette réduction provient de la distance supplémentaire à parcourir pour enquêter les logements des ZAE orphelines dans le cas de la sous-allocation. En règle générale, en effet le nombre d'enquêtes réalisées dans la zone orpheline est inférieur au nombre d'enquêtes qui ont été retirées dans la zone habituelle de l'enquêteur remplaçant. En outre, la zone orpheline étant moins bien connue, les temps de repérage sont accrus. Enfin, comme ces zones sont plus éloignées, les efforts de collecte sont parfois moins importants. Des mesures sont en cours afin de comparer les taux de réponse obtenus par un même enquêteur dans sa zone de collecte habituelle et dans la zone orpheline.

Cette méthode n'est donc pas parfaite, mais elle essaye autant que faire ce peut de réduire la perte de qualité des enquêtes liée à une réduction de la taille de l'échantillon en cours de collecte.

Bibliographie

- Christine M., Faivre, S. (2009), « Le projet OCTOPUSSE de nouvel Échantillon-Maître de l'Insee », *Actes des journées de méthodologie statistique*, INSEE, 23- 25 mars 2009.
- Guggemos, F. (2009), « Simulation de tirages de zones d'action enquêteurs pour les enquêtes-ménages de l'Insee », *Actes des journées de méthodologie statistique*, INSEE, 23- 25 mars 2009.
- Rebecq, A. (2014), « *Heuristique Branch-and-bound pour la sous-allocation et la réallocation* », Article présenté au 8e colloque francophone sur les sondage, Dijon, 18-20 novembre 2014.