STRATÉGIE DE SÉLECTION ET DE COORDINATION D'ÉCHANTILLONS POUR LES ENQUÊTES À UN ET DEUX DEGRÉS

Jean-Louis TAMBAY(*)

(*) Statistique Canada, Division des méthodes d'enquêtes auprès des ménages

Introduction

De Keyfitz[1] en 1951 à Rivière[2] en 2001, diverses méthodes ont été conçues pour contrôler le chevauchement entre des échantillons à un degré. Cependant, la sélection et la coordination d'échantillons à un et deux degrés tirés d'une même base deviennent difficiles si l'on tient à gérer la croissance de la population et à contrôler le chevauchement des échantillons. Une stratégie d'échantillonnage est proposée. Celle-ci vise à offrir un maximum de flexibilité tout en tenant compte d'un nombre élevé de besoins. La stratégie a été développée dans le cadre de l'initiative d'Architecture opérationnelle du Bureau de Statistique Canada. Cette initiative vise l'efficience organisationnelle par des mécanismes tels que la centralisation de services et l'utilisation de méthodes et d'outils communs. Un des projets proposés sous cette initiative est le développement d'une nouvelle fonction pour la sélection et la coordination d'échantillons pour les enquêtes auprès des ménages.

La stratégie d'échantillonnage proposée est assez simple. Il s'agit surtout d'un processus mécanique de sélection qui repose sur l'utilisation de nombres aléatoires permanents alloués aux unités (p. ex., le logement) ou à des groupes d'unités. Pour situer le contexte qui a mené à son développement, on présente d'abord le plan de sondage de l'Enquête sur la population active (EPA) dans la section 1. Ceci nous conduit, dans la section 2, à identifier des caractéristiques souhaitables pour un système pour la sélection d'échantillons à un et à deux degrés. La section 3 décrit la stratégie de base proposée pour les échantillons à deux degrés lorsqu'on dispose d'une base de sondage qui est mise à jour régulièrement. Dans la section 4, on établit que la stratégie de base équivaut à un échantillon aléatoire simple (ÉAS) sans remise et présente quelques résultats sur son fonctionnement. Les sections 5 à 8 portent sur des aspects particuliers de la stratégie : son utilisation pour aider la coordination de la collecte pour les enquêtes à deux degrés, son adaptation pour reproduire le plan de sondage actuel de l'EPA, une variante pour mieux contrôler les tailles d'échantillon, et une façon de s'en servir pour contrôler le chevauchement des échantillons des enquêtes à un et à deux degrés. Dans la section 9, on compare la stratégie à celle de l'EPA et à un ÉAS vis-à-vis des caractéristiques souhaitables présentées dans la section 2. Dans la section 10, on introduit un nouvel aspect, l'estimation des tendances, qui nous a obligés à reconsidérer l'adoption de l'approche de base pour l'EPA. Ceci nous mène à proposer une variante de la méthode pour ménager les besoins de l'EPA et ceux d'autres enquêtes. L'article conclut avec la section 11.

1. Plan de sondage de l'Enquête sur la population active (EPA)

1.1. Plan de sondage général

L'Enquête sur la population active est la source officielle pour les estimations mensuelles touchant l'emploi et le chômage au Canada[3]. L'enquête repose sur un échantillon d'environ 60 000 ménages tiré d'une base aréolaire. Les ménages sélectionnés sont suivis pendant six mois consécutifs et un sixième d'entre eux est renouvelé chaque mois. Dans les dix provinces, l'EPA se sert principalement d'un échantillon stratifié à deux degrés pour sélectionner les ménages. Les unités primaires

d'échantillon sont des grappes dont la taille est d'environ 250 logements. Dans chaque grappe sélectionnée, on obtient ou génère une liste de logements et on divise la liste en « départs » qui sont en fait des échantillons systématiques distincts de même taille (à une unité près). Selon la strate, le plan vise généralement à avoir entre 8 et 12 logements par départ. Ces départs constituent les unités secondaires d'échantillonnage. La base de sondage de l'EPA contient environ 60 000 grappes qui sont réparties en un peu plus de 1 000 strates. Environ 6 000 grappes sont échantillonnées.

Dans presque toutes les strates, l'échantillon de grappes est tiré avec probabilités proportionnelles à la taille (PPT) selon la méthode Rao-Hartley-Cochran (RHC). Cette méthode permet de maximiser le chevauchement des grappes sélectionnées après que leurs probabilités de sélection aient été mises à jour, par exemple, en réponse à une croissance extrême de la taille d'une grappe. Par le passé, chaque rotation d'une nouvelle grappe dans l'échantillon entraînait des coûts de listage et, parfois, une redistribution des charges de travail du personnel de terrain.

Selon la méthode RHC, on effectue d'abord une répartition aléatoire mais équilibrée des grappes dans chaque strate à ses groupes de renouvellement (il y en a généralement six). Dans un groupe de renouvellement donné, le nombre total de départs est fixé à la fraction de sondage inverse (FSI) pour sa strate et le nombre de départs attribué à chaque grappe est proportionnel à sa taille en termes de logements.

Dans chaque groupe de renouvellement, l'EPA sélectionne un départ dans une grappe suivant un procédé qui est illustré dans la figure 1. Supposons qu'un groupe de renouvellement comprend trente-trois départs (donc FSI = 33) alloués à quatre grappes. On trie les grappes aléatoirement dans le groupe de renouvellement et, dans chaque grappe, on suit la séquence des départs à partir d'un départ choisi au hasard (les premiers départs d'une grappe ont en général une unité de plus que les derniers). Pour démarrer le processus de sélection, on choisit un nombre au hasard entre 1 et 33. Ce nombre identifie la grappe et le départ sélectionnés initialement. Si on avait choisi le nombre 12 on aurait échantillonné le N^{o} de départ 6 dans la grappe N^{o} 082.

On effectue une rotation de tous les départs sélectionnés dans le premier groupe de renouvellement en janvier et en juillet, à ceux du deuxième groupe en février et en août, … , et ceux du sixième groupe en juin et en décembre. La rotation de l'échantillon se fait en se déplaçant le long de la liste, reprenant du début lorsqu'on arrive à la fin. Si le déplacement nous mène dans une nouvelle grappe, on effectue une rotation de grappe. Lors qu'une grappe est introduite à l'échantillon, on doit dresser la liste des logements et générer les départs, c'est-à-dire allouer les logements aux départs.

N° de grappe	219 (9 départs)									082 (7 départs)							447 (8 départs)								153 (9 départs)								
N° de départ	6	7	8	9	1	2	3	4	5	4	5	6	7	1	2	3	1	2	3	4	5	6	7	8	8	9	1	2	3	4	5	6	7

Figure 1 : Ordonnancement aléatoire de 4 grappes et de leurs 33 départs dans un groupe de renouvellement

D'autres enquêtes à deux degrés peuvent également se servir de la base de sondage aréolaire de l'EPA et par le fait même de la liste des départs pour choisir leurs échantillons. Après avoir déterminé leurs tailles d'échantillon par strate de l'EPA, il leur suffit de sélectionner, dans chaque strate, les prochains départs dans un ou plusieurs des groupes de renouvellement. Il est à noter, qu'aux fins de sélection, les groupes de renouvellement d'une strate sont identiques et donc interchangeables.

1.2. Caractéristiques du plan de sondage

Le plan de sondage de l'EPA offre des caractéristiques attrayantes. Il minimise la rotation des grappes et permet aux enquêtes de partager les grappes de l'échantillon, ce qui entraîne des économies des coûts de listage et de déplacements.

Le plan de sondage permet aussi de traiter la croissance dans la population. Dans les grappes échantillonnées, les nouveaux logements (« naissances ») sont identifiés périodiquement par les intervieweurs et répartis aux départs présents. Une grappe à forte croissance, ou dont la taille réelle

est supérieure à sa taille selon la base, peut être sous-échantillonnée ou, dans des cas extrêmes et rares, la grappe ou la région entière peuvent être restratifiées. Lors d'une restratification, les nombres de départs sont recalculés et le chevauchement avec l'ancien échantillon est maximisé selon la méthode proposée par Keyfitz[1]. Pour les cas de croissance moins élevée, un processus appelé « stabilisation » sert à maintenir la taille de l'échantillon. Il consiste à retirer un sous-échantillon systématique de logements dans chaque groupement de strates appelé « zone de stabilisation ». Pour l'EPA le processus de stabilisation est appliqué séparément pour chaque groupe de renouvellement.

Le plan de sondage de l'EPA a ses limites. La composition des strates et des grappes, et l'allocation de l'échantillon, ne sont pas optimales pour les autres enquêtes qui dépendent du même plan. Par exemple, il n'y a pas suffisamment de grappes échantillonnées pour satisfaire aux besoins de l'Enquête sur la santé dans les collectivités canadiennes (ESCC). Cette enquête a des besoins d'échantillons pour des régions sociosanitaires qui ne correspondent pas aux strates de l'EPA. L'ESCC doit utiliser un procédé à deux phases inefficace et lourd pour répondre à ses besoins (lequel comprend la sélection de grappes « à venir » selon le plan de sondage de l'EPA).

L'utilisation des départs par d'autres enquêtes peut également entraîner du gaspillage d'échantillons. Par exemple, si une enquête a besoin d'une taille d'échantillon équivalente à 2,2 départs dans une strate, elle devra sélectionner trois départs et se servir de la stabilisation pour éliminer l'excédent d'échantillons. Pour permettre l'estimation de la variance même les petites enquêtes sélectionnent un minimum de deux départs pas strate. Ce gaspillage accélère le processus de rotation, ce qui pourrait être indésirable dans les régions où les coûts de listage de grappes sont élevés et dans les strates où les fractions de sondage élevées peuvent mener à l'épuisement des départs en quelques années.

Le plan de l'EPA n'a pas non plus été conçu pour traiter des naissances identifiées sur l'ensemble de la population, ce qui deviendra possible à la suite de la décision de remplacer les activités de listage par l'utilisation d'un registre d'adresses pour la grande majorité de la population (les exceptions sont les régions rurales et éloignées). Finalement, la croissance élevée et inégale dans les grappes peut augmenter la variance échantillonnale et, dans les cas extrêmes, entraîner une restratification locale coûteuse.

2. Caractéristiques souhaitables d'un système pour la sélection d'échantillons à un et à deux degrés

Dans ce qui suit, on présente des caractéristiques souhaitables d'un système de sélection d'échantillons pour les enquêtes à un et à deux degrés (dont l'EPA). Il est supposé que l'on veuille, pour certaines enquêtes, choisir des échantillons à deux degrés pour réduire les coûts de collecte et, pour d'autres, choisir des échantillons à un degré qui tiennent compte de l'information disponible sur la base de sondage pour effectuer une stratification plus poussée de la population.

On se penche particulièrement sur les aspects opérationnels d'un tel système d'échantillonnage. On suppose que l'on possède une base de sondage détaillée qui permet de diviser la population en grappes et qui fournisse de l'information auxiliaire utile au niveau des unités (par exemple, logements).

Caractéristiques souhaitables :

a) Le système devrait être simple, mais permettre une plus grande flexibilité dans les plans de sondage (notamment à la stratification et la répartition).

b) Il devrait permettre des modifications aux tailles d'échantillon, par exemple à la suite d'achats d'échantillons supplémentaires, et éviter le gaspillage d'échantillon par l'utilisation poussée de la stabilisation ou du sous-échantillonnage en réponse à une croissance non désirable de la taille d'échantillon.

c) Il devrait pouvoir s'inscrire dans le cadre de produits généralisés, et utiliser des éléments et des méthodes communs.

d) Il devrait permettre l'utilisation efficace d'une base de sondage qui serait mise à jour régulièrement, principalement dans les centres urbains.

e) Dans certains cas, il devrait permettre de reproduire l'ancien plan de sondage de l'EPA (méthode RHC) – par exemple, là où l'on voudrait prolonger la durée de vie des grappes dans l'échantillon.

f) Il devrait pouvoir produire des échantillons à un et à deux degrés des mêmes unités (c.-à-d. de logements). Dans les échantillons à un degré, les unités seraient stratifiées individuellement à partir de données auxiliaires disponibles sur la base.

g) Il devrait produire des échantillons coordonnés (distincts) afin de réduire le fardeau de réponse, mais le faire sans causer de biais de sélection.

h) Il devrait permettre la coordination de la collecte pour les enquêtes à deux degrés.

i) Il devrait faciliter la transition d'échantillons à la suite d'un remaniement, par exemple, en simplifiant l'évitement du chevauchement avec l'ancien échantillon.

La caractéristique h) mérite plus d'explications. Actuellement, l'EPA, l'ESCC et l'Enquête sur les dépenses des ménages (EDM) coordonnent leurs échantillons non seulement pour être dans les mêmes grappes, mais aussi pour visiter ces grappes les mêmes jours autant que possible. Ceci aide à réduire les coûts de collecte pour les interviews en personne. Cette coordination repose sur le calendrier des opérations de collecte de l'EPA. La période de collecte de l'EPA a typiquement lieu la quatrième semaine du mois. L'utilisation d'entrevues personnelles est plus élevée lors du premier mois de collecte puisque la collecte de numéros de téléphone des répondants lors de la première entrevue sur le terrain permet de réaliser les cinq entrevues subséquentes via les centres d'appels téléphoniques. L'EDM, qui a une période de collecte d'un mois, et a généralement besoin de moins de six départs par strate, se coordonne avec l'EPA en deux étapes. Premièrement, avant même de sélectionner les départs, elle attribue un mois de collecte à chacun de ses départs. Deuxièmement, elle s'arrange pour sélectionner les départs dans les groupes de renouvellement dont l'échantillon est renouvelé par l'EPA le mois voulu. Par exemple, supposons que l'EDM a deux départs dans une strate qu'elle aimerait visiter en mars et juillet. Les départs utilisés par l'enquête seront donc ceux des groupes de renouvellement 3 (pour mars) et 1 (pour juillet). Les choses sont plus compliquées pour l'ESCC. Cette enquête utilise des périodes de collecte de deux mois et nécessite un échantillon de grappes supplémentaires. Mais le même principe est utilisé pour accroître le chevauchement avec le premier mois dans l'EPA et/ou avec le mois de collecte de l'EDM.

3. Stratégie générale pour les échantillons à deux degrés

3.1. Stratégie de base

On présente ici la stratégie de base, des raffinements seront ajoutés ultérieurement. On suppose que dans une grande partie de la population, par exemple les centres urbains, il sera possible de suivre l'évolution de l'ensemble de la population et d'allouer chaque nouvelle unité (ou logement) à une grappe et à un départ. La taille des grappes et des départs n'est pas importante, mais pour des besoins de flexibilité et d'efficience les départs devraient avoir à peu près la même taille dans une région.

Quoique la stratégie vise à générer des ÉAS stratifiés de départs, sa caractéristique principale est l'utilisation des grappes comme unités de base pour la sélection et la coordination d'échantillons. Ceci

veut dire que les enquêtes à deux degrés seront obligées d'utiliser le même ensemble de grappes et de départs, quoiqu'elles pourront stratifier les grappes comme elles veulent.

Un processus de sélection à deux degrés est utilisé dans chaque strate. Ce processus comporte quatre étapes. L'étape d'initialisation n'est nécessaire que lorsque les départs sont créés. L'identification de nouvelles unités (naissances ou nouveaux logements) et leur répartition aux départs a lieu régulièrement, par exemple à chaque trimestre.

Initialisation : On génère deux nombres aléatoires permanents pour chaque départ, disons S1 et S2.

Sélection des grappes : Après avoir stratifié les grappes, on utilise la séquence des valeurs de S1 pour déterminer l'ordre de sélection de grappes pour toutes les enquêtes qui utilisent cette même stratification. L'échantillon obtenu est presque un échantillon PPT avec remise de grappes (les grappes sont remises dans l'échantillon, mais pas leurs valeurs de S1).

Sélection des départs : Pour chaque grappe sélectionnée, on prend le nombre voulu de départs, k ($k \geq 1$), en suivant la séquence des valeurs de S2 à l'intérieur de la grappe. La valeur de k peut dépendre de l'enquête ou de la strate. Afin d'éviter les chevauchements d'échantillons, toute enquête à deux degrés qui tire un échantillon dans une grappe doit suivre la même séquence de valeurs de S2 et choisir la prochaine valeur disponible, en reprenant la séquence du début lorsqu'on touche à la fin.

Sélection des naissances : Les nouvelles unités (naissances) dans une grappe sont allouées à un départ suivant un ordre établi. Si une grappe a un nombre élevé de naissances, disons équivalents à au moins trois départs, on place ces naissances dans de nouveaux départs. Ce procédé est illustré plus loin (figure 4). Des valeurs S1 et S2 sont générées pour les nouveaux départs et sont insérées dans les séquences de valeurs S1 et S2 présentes pour leur strate et leur grappe.

L'étape de sélection des grappes est illustrée par les figures 2 et 3. Prenons une strate de quatre grappes ayant 7, 9, 8 et 6 départs, soit 30 départs en tout. La figure 2 montre les valeurs S1 qui auraient été générées pour ces départs. On sélectionne les grappes suivant l'ordre des valeurs de S1, tel qu'illustré par la figure 3. Les quatre premières sélections respectives donnent les grappes 3, 2, 1 et 2.

Grappe 1 (7 départs) :	887	192	659	738	130	427	773	
Grappe 2 (9 départs) :	814	223	518	144	739	051	405	667	944
Grappe 3 (8 départs) :	046	409	714	275	151	382	760	482	
Grappe 4 (6 départs) :	631	309	878	218	919	535			

Figure 2: Valeurs de S1 pour les 30 départs des 4 grappes d'une strate

046_3	051_2	130_1	144_2	151_3	192_1	218_4	223_2	275_3	309_4	382_3	405_2	409_3	427_1	482_3
518_2	535_4	631_4	659_1	667_2	714_3	738_1	739_2	760_3	773_1	814_2	878_4	887_1	919_4	944_2

Figure 3 : Séquence des valeurs de S1 indiquant l'ordre de sélection de grappes pour l'étape 1 (les indices donnent les grappes)

À l'étape suivante, la séquence des valeurs de S2 détermine l'ordre de sélection des départs dans chaque grappe. Supposons que les valeurs respectives de S2 pour les sept départs de la grappe 1 étaient 286, 506, 732, 178, 564, 355 et 941. Le quatrième départ sera choisi en premier (S2=178), puis le premier (S2=286), puis le sixième, et ainsi de suite. Cette séquence servira à toutes les enquêtes à deux degrés qui sélectionneront des départs dans cette grappe.

Pour S1 autant que pour S2 lorsqu'on arrive en fin de liste, on reprend la séquence du début. Pour la séquence de valeurs de S2, cela veut dire que des départs seraient à nouveau sélectionnés, et leurs

unités seraient enquêtées à nouveau (pas forcément par la même enquête). Ce phénomène devrait se produire rarement dans un intervalle de quelques années seulement.

Le contrôle du chevauchement pour les enquêtes à deux degrés est donc assez bien géré par ce mécanisme de sélection. Le chevauchement avec les enquêtes à un degré est également contrôlé au niveau de la grappe. Dans chaque grappe, une proportion fixe, mais mobile, des départs comprenant les départs qui ont été sélectionnés récemment par des enquêtes à deux degrés, et ceux qui le seront prochainement, est hors limite pour les enquêtes à un degré. On reviendra sur ce sujet dans la section 8.

3.2. Gestion de la croissance

La figure 4 illustre le procédé pour une attribution initiale de 32 unités, puis deux ensembles de naissances, dans la grappe 1. À gauche, on montre l'attribution des 32 unités et de cinq naissances numérotées de 33 à 37. À droite, on ajoute 16 naissances, ce qui occasionne la création de trois nouveaux départs pour la grappe. La création de nouveaux départs sert à préserver les tailles des départs et à éviter la formation de départs de taille énorme en réponse à une croissance excessive. Des départs peuvent toujours devenir énormes, mais pour se faire ils devront avoir grandi progressivement.

Départ	1	2	3	4	5	6	7
S2	286	506	732	178	564	355	941
Unité	1	2	3	4	5	6	7
	8	9	10	11	12	13	14
	15	16	17	18	19	20	21
	22	23	24	25	26	27	28
	29	30	31	32	*33*	*34*	*35*
	36	*37*					

Départ	1	2	3	4	5	6	7	*8*	*9*	*10*
S2	286	506	732	178	564	355	941	*680*	*79*	*351*
Unité	1	2	3	4	5	6	7	*38*	*39*	*40*
	8	9	10	11	12	13	14	*41*	*42*	*43*
	15	16	17	18	19	20	21	*44*	*45*	*46*
	22	23	24	25	26	27	28	*47*	*48*	*49*
	29	30	31	32	33	34	35	*50*	*51*	*52*
	36	37	*53*							

Figure 4 : Attribution de 32 unités, puis deux ensembles de 5 et 16 naissances, à la grappe 1

Les nouveaux départs perturbent l'ordre de sélection des grappes et des départs. Suite à l'ajout de trois départs, la plus petite valeur de S2 dans la grappe 1 est le 79, ce qui correspond au neuvième départ, c.-à-d. un nouveau départ. Si les départs 4, 1 et 6 de la grappe avaient déjà été sélectionnés, le prochain départ sélectionné aurait été le deuxième (S2=506). Si seulement les départs 4 et 1 avaient été choisis, le prochain départ aurait été le dixième (S2=351). De la même manière, la sélection des grappes sera perturbée par les valeurs de S1 générées pour les nouveaux départs dans la strate.

Cette manière de procéder en présence de nouveaux départs ne donne pas exactement les probabilités de sélection voulues, et en conséquence, il est nécessaire d'apporter un ajustement au procédé. Le problème vient du fait que lorsqu'on choisit les nombres aléatoires (S1 ou S2) selon la séquence établie, on s'arrête forcément sur une valeur actuelle, celle de la dernière unité sélectionnée. Quand on ajoute de nouveaux nombres aléatoires, ceux-ci ont un léger avantage pour la sélection parce que le dernier nombre aléatoire sélectionné, qui correspond à une ancienne unité, ne peut pas être choisi à moins d'avoir un recensement dans la strate. Plus précisément, si on a X anciens nombres aléatoires et B nouveaux, et qu'on arrête la séquence des S1 ou S2 sur un des X anciens nombres, un échantillon de n unités sur $N = X + B$ donnerait x et b anciennes et nouvelles unités, où x suivrait la loi hypergéométrique avec les paramètres $N–1$, $X–1$ et n et où b suivrait la loi hypergéométrique avec les paramètres $N–1$, B et n. Il faudrait plutôt que x suive la loi hypergéométrique avec les paramètres N, X et n et que b suive la loi hypergéométrique avec les paramètres N, B et n.

Pour rectifier le problème on propose l'approche suivante pour obtenir un échantillon de taille n sur $N = X + B$. Après avoir ajouté les B nouveaux nombres aléatoires aux X anciens, on sélectionne le

prochain nombre aléatoire en suivant la liste à partir du dernier nombre utilisé (un ancien). Si ce prochain nombre appartient à une nouvelle unité, on la retient avec probabilité π et on choisit les n–1 unités suivantes. Appelons ceci le cas 1. Si l'unité est rejetée on choisit les n prochaines unités (cas 2). Si le prochain nombre aléatoire est celui d'une ancienne unité on retient cette unité et prend les n–1 unités suivantes (cas 3).

On fixe π pour donner des tailles d'échantillon espérées de E(x) = nX/N et E(b) = nB/N. Soit :

$$\mathrm{E}(x) = \Pr(\text{cas 1})\,\mathrm{E}(x|\text{cas 1}) + \Pr(\text{cas 2})\,\mathrm{E}(x|\text{cas 2}) + \Pr(\text{cas 3})\,\mathrm{E}(x|\text{cas 3})$$

$$\frac{nX}{N} = \frac{\pi B}{N-1}\,\frac{(n-1)(X-1)}{N-2} + \frac{(1-\pi)B}{N-1}\,\frac{n(X-1)}{N-2} + \frac{X-1}{N-1}\left(1 + \frac{(n-1)(X-2)}{(N-2)}\right)$$

Ce qui donne

$$\pi = 1 - \frac{(N-2)n}{(X-1)N}.$$

Dans de rares occasions, il peut y avoir tellement de nouveaux nombres aléatoires que la valeur de π en devienne négative. Cela se produit si (X–1)/(N–2) est inférieur à la fraction de sondage n/N. Dans ces situations, on rejette le premier nombre aléatoire s'il correspond à une nouvelle unité et rejette avec une certaine probabilité 1–ρ le deuxième nombre aléatoire s'il correspond lui aussi à une nouvelle unité, et on choisit les prochaines unités pour en avoir n en tout. On continue de la sorte si ρ est également négatif; quoique dans ces situations extrêmes il sera probablement préférable d'effectuer une restratification.

4. Quelques résultats

Pour une strate quelconque, soient :

N_i le nombre de départs dans la grappe i de la strate, soit la « taille » de la grappe i
N le nombre total de départs dans la strate ($N = \Sigma_i N_i$)
n le nombre de grappes sélectionnées dans la strate suivant la séquence des valeurs S1
n_i le nombre de fois que la grappe i a été sélectionnée ($\Sigma_i n_i = n$)
k le nombre de départs choisi à chaque sélection de grappes (l'échantillon comprend nk départs)

L'utilisation des valeurs S1 donne à peu près un échantillon avec PPT sans remise de grappes, sauf que les probabilités de sélection des grappes ne sont pas fixes. La probabilité que la grappe i soit sélectionnée lors du s-ième tirage est égale à N_i/N. La probabilité conditionnelle que la grappe i soit sélectionnée lors du t-ième tirage si la grappe j avait été sélectionnée lors du s-ième tirage est égale à $N_i/(N-1)$. La probabilité que la grappe i soit sélectionnée lors des s-ième et t-ième tirages est égale à $N_i(N_i-1)/[N(N-1)]$.

Le nombre de fois qu'une grappe i est sélectionnée, soit n_i , suit la loi hypergéométrique avec les paramètres N, N_i et n. Si $k = 1$ et $n \le \min\{N_i\}$ les départs sélectionnés dans la strate constitueront l'équivalent d'un ÉAS sans remise de n départs sur N. Le poids de sondage de chaque départ sélectionné sera égal à N/n. Si $k > 1$ et $nk \le \min\{N_i\}$, pour l'estimation de la variance, on traitera les n sélections de k départs comme n répétitions indépendantes (on procèdera sans doute pareillement même lorsque k sera égal à 1). Si la fraction de sondage nk/N est faible, la surestimation de la variance sera négligeable. Chaque ensemble de k départs sélectionné aura un poids d'échantillonnage de N/nk.

On démontre que l'échantillon de n départs est semblable à un ÉAS sans remise quand $k = 1$ et $n \le \min\{N_i\}$. Pour la loi hypergéométrique $E(n_i) = nN_i/N$, $V(n_i) = nN_i(N-n)(N-N_i)/[N^2(N-1)]$ et, pour les grappes i et j, $Cov(n_i, n_j) = -nN_iN_j(N-n)/[N^2(N-1)]$.

Pr(le départ s dans la grappe i est sélectionné)

$= \sum_{n_i} \Pr(n_i|N, N_i, n)$ Pr (le départ s est sélectionné$|n_i)$

$= \sum_{n_i} \Pr(n_i|N, N_i, n)\ (n_i/N_i)$

$= E(n_i)/N_i = n/N.$

Pr (les départs s et t dans la grappe i sont sélectionnés)

$= \sum_{n_i} \Pr(n_i|N, N_i, n)$ Pr (les départs s et t sont sélectionnés$|n_i)$

$= \sum_{n_i} \Pr(n_i|N, N_i, n)\ [n_i(n_i - 1)]/[N_i(N_i - 1)]$

$= E(n_i(n_i - 1))/[N_i(N_i - 1)]$

$= \{V(n_i) + [E(n_i)]^2 - E(n_i)\}\ /\ [N_i(N_i - 1)]$

$= n(n - 1)/[N(N - 1)].$

Pr(les départs s de la grappe i et t de la grappe j sont sélectionnés) $=$

$\sum_{n_i} \sum_{n_j} \Pr\left(n_i, n_j\middle|N, N_i, N_j, n\right)$ Pr(le départ s est sélectionné$|n_i)$ Pr$\left(\text{le départ } t \text{ est sélectionné}\middle|n_j\right)$

$= \sum_{n_i} \sum_{n_j} \Pr\left(n_i, n_j\middle|N, N_i, N_j, n\right)\ (n_i/N_i)(n_j/N_j)$

$= E\left(n_i n_j\right)/(N_i N_j)$

$= \left\{Cov\left(n_i, n_j\right) + E(n_i)E(n_j)\right\}/(N_i N_j)$

$= n(n - 1)/[N(N - 1)].$

Ce qui correspond aux probabilités de sélection d'un ÉAS sans remise de *n* unités sur *N*.

La probabilité de sélection d'un départ lorsque $k > 1$ et $nk \leq \min\{N_i\}$ est donnée. Pour une grappe *i*, on définit l'ensemble *s* comme la suite de *k* départs sélectionnés en suivant la séquence des valeurs de S2 à partir du départ *s* (on boucle la séquence à la fin). Compte tenu de la séquence de S2 :

Pr (l'ensemble s de la grappe i est sélectionné)

$= \sum_{n_i} \Pr(n_i|N, N_i, n)$ Pr (l'ensemble s est sélectionné$|n_i)$

$= \sum_{n_i} \Pr(n_i|N, N_i, n)\ (n_i/N_i)$

$= E(n_i)/N_i = n/N.$

Pr (le départ s de la grappe i est sélectionné)

$= \sum_{n_i} \Pr(n_i|N, N_i, n)$ Pr (le départ s est sélectionné$|n_i)$

$= \sum_{n_i} \Pr(n_i|N, N_i, n)\ (kn_i/N_i)$

$= E(n_i)(k/N_i) = nk/N.$

Il n'est pas facile d'obtenir les probabilités conjointes de sélection parce que les ensembles se chevauchent. Par exemple, pour une grappe *i* avec N_i = 7 départs et *k* = 2 on a sept ensembles possible de deux départs, soit $\{d_1,d_2\}$, $\{d_2,d_3\}$, $\{d_3,d_4\}$, $\{d_4,d_5\}$, $\{d_5,d_6\}$, $\{d_6,d_7\}$ et $\{d_7,d_1\}$ (le départ d_j correspond au j-ième départ selon l'ordre de sélection établi par la séquence des valeurs de S2). Pour le cas particulier où, dans chaque grappe, le nombre de départs serait un multiple de *k* l'échantillon de *n* ensembles de départs obtenu ressemble conditionnellement à un ÉAS sans remise de taille *n* d'une population de taille *N/k*. Ce résultat est obtenu en conditionnant sur les ensembles de départs présents dans chaque grappe (par exemple, si N_i = 6 et *k* = 2 et que le prochain départ à être choisi est d_1 , d_3 ou d_5 on conditionne sur les ensembles $\{d_1,d_2\}$, $\{d_3,d_4\}$ et $\{d_5,d_6\}$, sinon, on conditionne sur les ensembles $\{d_2,d_3\}$, $\{d_4,d_5\}$,et $\{d_6,d_1\}$).

Le processus de sélection a ses inconvénients. Tout d'abord, en présence d'une forte corrélation intragrappe, ce processus peut perdre de l'efficience puisque les grappes sont sélectionnées avec remise. De plus, à moins que toutes les enquêtes ne se servent de la même stratification et ne sélectionnent qu'un seul départ par grappe (*k* = 1), il se peut que des départs soient rééchantillonnés trop vite parce que leur grappe a été sélectionnée trop souvent. Le tableau 1 donne des probabilités que ces situations se produisent pour divers scénarios. Les scénarios représentent différentes combinaisons de nombres de grappes et de départs dans une strate. Dans tous les cas présentés, on suppose que le nombre de départs par grappe, N_i , est le même dans chaque grappe. Les scénarios sont basés sur le plan de sondage de l'EPA. L'EPA sélectionne un départ dans une grappe chaque mois. La fraction de sondage inverse (FSI) de l'EPA représente une accumulation de six mois d'échantillons.

Si, à chaque sélection d'une grappe *i*, on prend *k* de ses N_i départs le nombre maximum de fois que la grappe peut être choisie sans avoir à en rééchantillonner des départs est $n_{i(max)} = [N_i/k]$. La ligne $Pr(n_i > n_{i(max)}$ si *k* = *) donne les probabilités qu'une grappe soit sélectionnée plus de $n_{i(max)}$ fois en moins de 5 années quand *k* est égal à *. Pour les scénarios présentés, cette probabilité est très petite sauf lorsque *k* = 3 et la FSI est égale à 40. Seulement 3,5 % des strates de l'EPA ont une FSI inférieure à 60.

Tableau 1 : Probabilités d'évènements donnés sous divers scénarios

Fraction de sondage inverse (FSI) = *N*/6	40	60	40	60	80	120	72	90	180	120	240	180	240
N (n^{bre} de départs)	240	360	240	360	480	720	432	540	1080	720	1440	1080	1440
N_i (tailles de grappes)	16	24	12	18	24	36	18	18	36	18	36	18	24
N/N_i (n^{bre} de grappes)	15	15	20	20	20	20	24	30	30	40	40	60	60
Probabilités qu'une grappe soit sélectionnée plus de $n_{i(max)}$ fois en moins de 5 années													
$n_{i(max)}$ quand *k* = 2	8	12	6	9	12	18	9	9	18	9	18	9	12
$n_{i(max)}$ quand *k* = 3	5	8	4	6	8	12	6	6	12	6	12	6	8
$Pr(n_i > n_{i(max)}$ si *k* = 2)	0,006	0,000	0,012	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
$Pr(n_i > n_{i(max)}$ si *k* = 3)	0,182	0,009	0,152	0,018	0,001	0,000	0,007	0,002	0,000	0,000	0,000	0,000	0,000
Probabilités d'obtenir 6, 5 ou moins de 5 grappes distinctes avec un échantillon de 6 grappes													
Pr(6 distinctes)	0,337	0,330	0,464	0,455	0,450	0,445	0,525	0,603	0,595	0,689	0,682	0,783	0,781
Pr(5 distinctes)	0,474	0,474	0,426	0,429	0,431	0,433	0,392	0,342	0,347	0,279	0,284	0,202	0,204
Pr(<5 distinctes)	0,189	0,196	0,110	0,116	0,119	0,122	0,083	0,055	0,058	0,032	0,034	0,015	0,015

Les trois dernières lignes donnent les probabilités d'obtenir différents nombres de grappes distinctes lorsqu'on en sélectionne six. Les probabilités dépendent en grande partie du nombre de grappes dans la strate (N/N_i). Si ce nombre est entre 20 et 60, les probabilités de sélectionner six grappes distinctes

varient entre 45 % et 78 %, et les probabilités d'obtenir au plus quatre grappes distinctes varient entre 1,5 % et 12 %. Des simulations indiquent qu'avec des corrélations intraclasses allant jusqu'à 0,24, l'effet de plan obtenu n'est que légèrement plus élevé que celui obtenu lorsque les grappes sont sélectionnées sans remise. Bien entendu, l'impact de la sélection avec remise serait atténué si on augmentait la taille des strates.

Pour avoir un échantillon qui est quasiment sans remise, il est souvent suggéré d'utiliser des nombres aléatoires S1 équidistants. Pour une grappe de N_i départs on peut mettre S1 = $a + bN_i^{-1}$, où a est un nombre aléatoire entre 0 et N_i^{-1} et b = 0, 1, ... , N_i –1. Les probabilités conjointes de sélection seront affectées et on n'aura plus l'équivalent d'un ÉAS sans remise de départs, mais les départs sélectionnés proviendront presque toujours de grappes distinctes.

5. Coordination de la collecte pour les enquêtes à deux degrés

5.1. Coordination générale

Dans la section 2, on a noté qu'il serait souhaitable que la stratégie d'échantillonnage facilite la coordination des enquêtes se servant d'échantillons à deux degrés. On a donné un bref aperçu de la manière dont les échantillons des trois grandes enquêtes à deux degrés, l'EPA, l'ESCC et l'EDM étaient coordonnés. Cependant, si on permettait aux enquêtes de stratifier les grappes selon leurs besoins, il faudrait ajuster la manière dont les échantillons seront coordonnés. Aux fins d'illustration supposons que l'ESCC se serve d'une stratification différente de celle de l'EPA. La coordination d'échantillons aura lieu au niveau des intersections des strates.

Soient n'_h le nombre de grappes qui ont été sélectionnées par l'EPA dans ses strates h = 1, 2, … , H, et n''_g le nombre de grappes qui ont été sélectionnées par l'ESCC dans ses strates g = 1, … , G. Si n'_{gh} et n''_{gh} représentent les nombres de grappes de ces enquêtes qui ont été sélectionnées dans le croisement de leurs strates h et g, alors on pourrait remplacer jusqu'à n^*_{gh} = min{ n'_{gh} ,n''_{gh} } des grappes échantillonnées par l'ESCC par des grappes échantillonnées par l'EPA dans le croisement de strates gh. Ce serait comme si l'échantillon initial de l'ESCC ne servait qu'à déterminer le nombre d'unités à prendre dans chaque croisement de strates gh et qu'ensuite on choisissait ces unités en maximisant le chevauchement avec l'échantillon de l'EPA.

On pourrait même aller plus loin et coordonner la collecte pour les deux enquêtes afin que le plus grand nombre possible de grappes soient visitées par l'ESCC le même mois où leur échantillon serait renouvelé par l'EPA. La coordination de la collecte deviendrait un problème d'optimisation. On attribuerait à chaque combinaison de grappes de l'ESCC et période de collecte une variable dichotomique qui serait égale à un si la grappe était visitée pendant la période, 0 sinon. On attribuerait aussi à chaque combinaison une cote qui serait positive si cette combinaison correspondait à une période de collecte pour l'EPA. Il suffirait de maximiser le total des cotes sous les contraintes de respecter les tailles d'échantillon de l'ESCC, de ne visiter une grappe qu'une seule période de collecte (si une grappe était sélectionnée deux fois, on la traiterait comme deux grappes ici), et que le nombre de grappes visitées à chaque période de collecte soit constant pour chaque strate ou groupe de strates de l'ESCC.

5.2. Cas particulier de la création de nouveaux départs

La création de nouveaux départs dans les grappes qui ont subi une certaine croissance peut nuire à la coordination des échantillons si les nouveaux départs ne sont pas traités en même temps par toutes les enquêtes. Les enquêtes annuelles pourraient s'accommoder d'une mise à jour annuelle de la base de sondage, mais ce n'est pas le cas pour les enquêtes courantes comme l'EPA. Pour l'EPA il serait préférable d'ajouter les nouvelles unités d'une façon plus ou moins uniforme à travers l'année. Actuellement pour l'enquête, chaque mois les nouveaux départs sont ajoutés dans le sixième des grappes dont l'échantillon est renouvelé ce mois.

Au pire, l'EPA pourrait s'accommoder d'une mise à jour trimestrielle de sa base. Mais les autres enquêtes, surtout l'ESCC, ne pourraient pas subir une mise à jour plus fréquente que tous les six ou douze mois. La figure 5, qui reprend en partie la séquence des valeurs de S1 de la figure 3, illustre comment la coordination pourrait être affectée par les nouveaux départs. Pour permettre la coordination des échantillons, l'EPA sélectionnerait ses grappes, et départs, jusqu'à six mois à l'avance. Disons qu'il s'agisse des grappes 2, 1, 2, 3, 1 et 4 représentées par les six valeurs soulignées de S1 dans la séquence initiale de la figure 5. On utilise des lignes pointillées pour indiquer que la sélection pour les trois derniers mois est provisoire. Une fois cette sélection accomplie, les autres enquêtes pourraient s'arranger pour sélectionner leurs échantillons de grappes en fonction des mois de collecte de l'EPA (puisque les grappes et les départs échantillonnés dans une strate sont interchangeables).

Séquence initiale : 046_3 $\underline{051}_2$ $\underline{130}_1$ $\underline{144}_2$ $\underline{151}_3$ $\underline{192}_1$ $\underline{218}_4$ 223_2 275_3 309_4 ...
3 mois plus tard : 017_2 046_3 $\underline{051}_2$ $\underline{130}_1$ $\underline{144}_2$ $\underline{151}_3$ $\underline{168}_2$ $\underline{192}_1$ 218_4 223_2 ...

Figure 5 : Séquences des valeurs de S1 indiquant l'ordre de sélection de grappes pour l'EPA

Supposons que, trois mois plus tard, une mise à jour dans la strate donne lieu à de nouveaux départs dans la grappe 2. Les valeurs de S1 de ces nouveaux départs seront insérées dans la liste en cours comme indiqué sur la deuxième ligne de la figure 5 (pour simplifier la discussion on ignore l'ajustement proposé dans la section 3.2). Étant donné que l'EPA voudrait inclure cette mise à jour dans son échantillon, il en résulterait un effet de domino pour l'EPA et pour toute enquête qui puiserait son échantillon après celui de l'EPA (c.-à-d. en commençant par la valeur S1 de 223). L'insertion de la valeur S1=017 n'affecterait pas l'échantillon puisqu'elle aurait eu lieu « dans le passé ». Mais la valeur S1=168, par contre, retarderait probablement la période de collecte pour la grappe 1 (S1=192) et pousserait la grappe 4 (S1=218) hors de l'échantillon.

Pour éviter ces perturbations, surtout la dernière, et maintenir une certaine coordination entre les enquêtes, on se servirait d'intervalles de valeurs S1 pour réserver les échantillons à venir. Par exemple, pour l'échantillon correspondant aux trois derniers mois de l'EPA, on identifierait un intervalle de valeurs S1 qui irait de la moyenne des troisième et quatrième valeurs (soit 147,5) à la moyenne des sixième et septième valeurs (221,5). Les nouvelles valeurs de S1 à l'intérieur de cet intervalle pourraient affecter l'échantillon de l'EPA, et donc la coordination de la collecte d'autres enquêtes avec celle de l'EPA, mais n'auraient aucune incidence sur les échantillons tirés ultérieurement. L'ajout de la valeur S1=168 pousserait la grappe 4 hors de l'échantillon de l'EPA, certes, mais sans ajouter cette grappe à un échantillon à venir. Du moins, cela ne se produirait pas à cause de la valeur S1=218.

6. Reproduction du plan de sondage de l'EPA

Il a été noté dans la section 2 que la stratégie, ou son application, devrait permettre de reproduire la méthode de l'ancien plan de sondage de l'EPA. Cette méthode, qui retient les grappes dans l'échantillon pour de longues périodes, est souvent préférable en milieu rural. En dehors des villes, l'absence d'adresses de type municipal rend difficiles la création et le maintien d'une base d'adresses à partir de sources administratives. Il faut souvent effectuer un listage sur place pour avoir une bonne couverture des logements. La rétention de grappes dans l'échantillon réduit considérablement les coûts de listage. De plus, elle permet une meilleure planification du travail des intervieweurs.

Pour permettre à la nouvelle stratégie de reproduire le plan de sondage représenté dans la figure 1, il suffit de traiter les grappes dans chaque strate au niveau de leurs groupes de renouvellement et de faire en sorte à ce que les valeurs de S1 soient regroupées par grappe tel qu'illustré par la figure 6. L'ordre des grappes, et les valeurs de S1, sont déterminés de façon aléatoire tandis que l'attribution de valeurs S1 aux grappes dépend de leur ordre et de leurs tailles.

046_3 051_3 130_3 144_3 151_3 192_3 218_3 223_3 275_2 309_2 382_2 405_2 409_2 427_2 482_2
518_2 535_2 631_1 659_1 667_1 714_1 738_1 739_1 760_1 773_4 814_4 878_4 887_4 919_4 944_4

Figure 6: Séquence des valeurs de S1 regroupées pour quatre grappes dans un groupe de renouvellement

7. Une alternative à l'utilisation de départs

Selon la stratégie d'échantillonnage proposée, les enquêtes à deux degrés doivent se partager les mêmes ensembles de départs et de grappes, ce qui peut présenter quelques inconvénients. Premièrement, la croissance inégale des grappes affectera l'homogénéité des tailles des grappes et pourrait accroître la variance intergrappes. Ce problème pourra être contrôlé par la création de nouveaux départs dans une grappe chaque fois que sa taille augmentera de façon significative.

Deuxièmement, la taille moyenne des départs ne conviendra pas nécessairement à toutes les enquêtes. Des enquêtes devront peut-être choisir plus d'un départ à la fois ou sous-échantillonner des départs pour éliminer un excédent d'échantillon. Ce suréchantillonnage non souhaité pourrait épuiser les unités de certaines grappes et les réintroduire trop tôt dans l'échantillon.

Troisièmement, du financement additionnel ou des réductions budgétaires pourront nécessiter de légères modifications aux tailles d'échantillon. Le sous-échantillonnage pourra facilement accommoder une légère coupure d'échantillon. Mais, pour toute enquête ayant des contraintes d'échantillon au premier degré (par exemple, pour l'EPA le nombre de grappes par strate devra être un multiple de six) une légère augmentation pourrait amener à doubler le nombre de grappes ou de départs, quitte à faire un fort usage du sous-échantillonnage pour ramener les tailles d'échantillon aux niveaux voulus. Il en résulterait un suréchantillonnage non souhaité comme mentionné au paragraphe précédent.

Au lieu d'allouer les valeurs S2 aux départs, on pourrait aussi les donner directement aux unités de chaque grappe. Chaque enquête pourrait sélectionner le nombre exact d'unités voulues en respectant la séquence des valeurs de S2. Un nombre S2 aléatoire serait donné à chaque unité ou, pour mieux répartir les unités dans une grappe, les valeurs S2 pourraient incorporer un numéro de départ. Par exemple, S2 pourrait être la somme du numéro de départ (attribué aléatoirement) et d'un nombre aléatoire entre 0 et 1.

Une telle approche compliquerait l'estimation de la variance, mais des méthodes approximatives devraient donner d'assez bons résultats – on traite souvent l'échantillon au premier degré comme un échantillon avec remise. Il faudrait aussi effectuer l'ajustement pour la croissance présentée dans la section 3.2.

Il est à noter que l'attribution des valeurs S2 directement aux unités n'empêche pas d'allouer les valeurs de S1 à l'équivalent des départs des grappes. Par exemple, chaque grappe pourrait recevoir un nombre de valeurs S1 équivalent à sa taille divisée par dix ou vingt. Cette approximation éviterait de modifier la séquence des valeurs de S1 suite à chaque changement mineur dans la taille des grappes.

8. Contrôle du chevauchement avec les enquêtes à un degré

La stratégie proposée contrôle automatiquement le chevauchement des échantillons pour les enquêtes à deux degrés. La méthode des microstrates[1] peut servir à contrôler le chevauchement des enquêtes à un degré. Il reste donc à développer une méthode pour contrôler le chevauchement d'échantillons entre ces deux types d'enquêtes.

La méthode des microstrates a été conçue pour permettre la coordination d'échantillons aléatoires stratifiés simples sans remise à partir d'une base de sondage d'unités. La microstratification d'enquêtes consécutives est une partition de la population telle que ses strates, appelées microstrates, correspondent à l'intersection de toutes les strates utilisées dans les enquêtes antérieures à celles en cours (y compris les strates à tirage nul, correspondant aux populations hors champ et aux naissances). La méthode se sert de nombres aléatoires non permanents et d'une fonction de sélection d'échantillons faisant usage de ces nombres aléatoires, par exemple, pour en sélectionner les plus petites valeurs. Ces nombres aléatoires sont permutés entre les unités de chaque microstrate afin que les unités ayant le plus gros fardeau de réponse cumulatif soient les dernières à être échantillonnées. La permutation est faite avant la stratification des unités par l'enquête en cours. Finalement, pour éviter de créer de petites microstrates, ce qui nuirait à l'efficacité de la méthode, il est préférable de limiter le nombre d'enquêtes traitées par microstrate.

Le contrôle du chevauchement entre les enquêtes à un et à deux degrés est difficile. On pourrait se servir de la méthode des microstrates à l'intérieur de chaque grappe pour permettre aux enquêtes à deux degrés d'éviter les échantillons des enquêtes à un degré. Mais cette solution est loin d'être pratique compte tenu de la petite taille des grappes. Et elle ne traiterait pas du cas inverse, où l'on voudrait qu'une enquête à un degré évite les échantillons d'enquêtes à deux degrés.

La solution proposée consiste à « réserver », dans chaque grappe, une proportion fixe, mais mobile, d'unités pour les enquêtes à deux degrés. Cette proportion pourrait être établie pour de grandes zones géographiques qui représenteraient des niveaux de stratification communs entre les enquêtes à un et à deux degrés. Lors de l'application de la méthode des microstrates aux enquêtes à un degré, les unités dans la partie réservée de chaque grappe, sans distinction de grappe, seront traitées comme les unités ayant le fardeau de réponse le plus élevé. On agirait en sorte comme si les enquêtes à deux degrés formaient une seule enquête à un degré n'ayant qu'une strate par grande zone géographique. Le fait d'ignorer les grappes n'est pas néfaste puisque l'échantillon « réservé » sera en quelque sorte mieux réparti qu'un échantillon aléatoire simple.

La partie réservée de chaque grappe serait constituée d'étendues des valeurs de S2 correspondant à des échantillons récents et à venir pour des enquêtes à deux degrés. L'utilisation d'étendues pour les valeurs de S2 permet d'incorporer automatiquement les nouvelles unités dans le traitement du chevauchement. Si la valeur de S2 d'une naissance est dans la partie réservée, cette naissance n'est pas échantillonnée par une enquête à un degré. Il n'est donc pas nécessaire de créer des catégories à part pour traiter des naissances, comme avec les microstrates.

9. Comparaison récapitulative de trois approches

Nous avons proposé un mécanisme simple et flexible pour la sélection et la coordination d'échantillons pour les enquêtes à un et à deux degrés. La figure 7 contraste la stratégie au plan de sondage de l'EPA et à un ÉAS par rapport aux caractéristiques souhaitables présentées à la section 2. Comme prévu, la stratégie proposée répond mieux aux besoins exprimés. En présence d'une forte corrélation intragrappe, il serait préférable pour les enquêtes à deux degrés d'utiliser des valeurs de S1 qui donneraient plutôt des échantillons sans remise au premier degré.

Caractéristiques souhaitables	Plan de l'EPA	Plan à un degré (ÉAS)	Stratégie proposée
a. Simple/flexible, accepte plans de sondage divers	×××	✓✓×	✓✓✓
b. Permet de modifier les tailles d'échantillon	✓××	✓✓✓	✓✓×
c. S'inscrit dans le cadre des systèmes généralisés	✓××	✓✓×	✓✓✓
d. Utilisation efficace d'une base de sondage à jour	✓××	✓✓✓	✓✓×
e. Peut reproduire le plan de sondage de l'EPA	✓✓✓	×××	✓✓✓
f. Produit des échantillons à un et à deux degrés	×××	×××	✓✓?
g. Produit des échantillons coordonnés non biaisés	✓?	×?	✓?
h. Permet la coordination de la collecte	✓?	×××	✓?
i. Facilite la transition après un remaniement	×××	✓✓✓	✓?

Figure 7. Évaluation de trois approches selon les caractéristiques souhaitables (répond : ✓– oui, × – non, ? –adaptation requise)

10. Prise en considération des tendances

10.1. Impact de la stratégie sur les tendances

Un facteur ajouté tardivement à l'évaluation de la nouvelle stratégie d'échantillonnage est son impact sur l'estimation des tendances pour une enquête continue comme l'EPA. Quoique l'estimation de tendances ne figure pas dans les objectifs de l'EPA, il va sans dire que les variations mensuelles des taux de chômage diffusés par cette enquête sont scrutées attentivement. L'impact sur les tendances mensuelles de la rotation d'un sixième de l'échantillon chaque mois est largement atténué par l'utilisation d'un estimateur composite qui accorde en partie les estimations mensuelles à celles du mois précédent[3]. Mais l'estimation des tendances profite également du fait que, la plupart du temps, la rotation de l'échantillon se fait à l'intérieur des grappes. En général, seulement 15 % des grappes de l'échantillon sont remplacées chaque année.

Sous le plan proposé la rotation de l'échantillon aura plutôt l'effet de remplacer près du sixième des grappes à chaque mois, soit presque toutes les grappes après six mois. En présence d'une forte corrélation intragrappes l'impact sur les tendances sera marqué. On a évalué l'impact de la nouvelle approche sur l'estimation des tendances et les résultats ont été inattendus[4]. Comme il était difficile d'incorporer l'estimateur composite aux évaluations, on s'est contenté de calibrer les estimations mensuelles à des projections mensuelles des tailles de population. Pour les tendances mensuelles, le plan proposé occasionnait des augmentations de variance de 15 à 25 % par rapport au plan actuel de l'EPA. L'estimation de la tendance annuelle (entre les mois de janvier 2001 et janvier 2002) subissait des pertes d'efficacité semblables. On a également évalué l'estimation des tendances avec l'approche ÉAS et les résultats étaient presque aussi mauvais qu'avec la stratégie proposée.

Les résultats auraient peut-être été différents si on avait réussi à mieux reproduire le plan de sondage de l'EPA actuel, y compris son utilisation de l'estimateur composite. Cependant, l'ampleur de la perte d'efficacité des estimations de tendances nous a menés à préférer l'approche actuelle de l'EPA, quitte à la reproduire avec un choix approprié de nombres aléatoires S1 et S2, comme décrit dans la section 6.

Pour les autres enquêtes à deux degrés, il sera nécessaire de choisir entre le maintien du plan de sondage de l'EPA et l'approche proposée. Le plan de l'EPA permettra de coordonner leurs échantillons avec celui de l'EPA, ce qui peut entraîner des économies de coûts de collecte et, peut-être même d'améliorer leurs estimations de tendances annuelles (quoique cela ne semble pas être le

cas pour l'ESCC). Opter pour l'approche proposée permettrait à ces enquêtes de profiter des avantages de cette approche y compris plus de flexibilité quant à la stratification et à la répartition de l'échantillon, et une meilleure gestion de la croissance.

10.2. Ajustements à la stratégie

L'adoption de l'approche de l'EPA par les autres enquêtes nuirait à la coordination des échantillons avec ceux des enquêtes à un degré. L'EPA a tendance à « épuiser » les grappes avant de les remplacer dans l'échantillon. Pour gérer le chevauchement, on pourrait adopter une approche de compromis qui consisterait à réserver une proportion fixe, mais mobile, des départs de chaque groupe de renouvellement pour les enquêtes à deux degrés. Les départs réservés viendront des grappes qui sont ou étaient récemment dans l'échantillon ainsi que de départs des grappes à venir dans l'échantillon. Si les taux d'échantillon sont assez faibles, cette proportion pourrait l'être également, ce qui gênera moins la sélection d'échantillons à un degré.

Si les autres enquêtes à deux degrés optent pour ne pas utiliser le plan de l'EPA, il est alors suggéré de diviser les départs de chaque grappe en deux parties. Une portion des départs (qui sera fixe dans chaque grande zone géographique commune) sera réservée à l'EPA et le reste servira aux autres enquêtes à deux degrés. En opérant de la sorte, les autres enquêtes ne risqueront pas d'être biaisées par la forte utilisation de certaines grappes par l'EPA, et l'EPA ne risquera pas de sélectionner des départs utilisés par d'autres enquêtes. La portion de départs réservée à l'EPA influencera directement le taux de remplacement des grappes pour cette enquête. Si elle est trop petite, les grappes seront remplacées trop vite et les estimations de tendances pourraient en souffrir. Par contre, si elle est trop grande, ceci pourrait occasionner aux départs réservés aux autres enquêtes de revenir trop tôt dans leurs échantillons.

Avec cette approche hybride, il serait préférable aux autres enquêtes d'ignorer les grappes et de sélectionner directement des unités (logements) dans leurs portions de chaque grappe, comme suggéré dans la section 7. Ceci permettrait aux probabilités de sélection de grappes d'être mises à jour sans avoir à créer de nouveaux départs. Les autres enquêtes utiliseraient des nombres aléatoires S1* qui dépendraient de la taille des grappes et qui seraient donc différents en quantité et en valeur des nombres aléatoires S1 utilisés par l'EPA. L'attribution d'unités aux départs servira principalement à l'EPA et aussi à déterminer quelles unités seront réservées à l'EPA et lesquelles seront réservées aux autres enquêtes.

Les enquêtes à un degré pourraient éviter tous les départs réservés à l'EPA ou elles pourraient éviter une portion fixe des départs réservés à l'EPA dans chaque groupe de renouvellement, comme expliqué ci-dessus. Elles éviteront également une portion fixe des départs de chaque grappe qui sont réservés aux autres enquêtes.

11. Conclusion

Comme il a été mentionné dans l'introduction, la stratégie d'échantillonnage est principalement un processus mécanique de sélection et de coordination d'échantillons qui repose sur une utilisation de nombres aléatoires permanents. On a démontré qu'un tel mécanisme offrait beaucoup de flexibilité aux enquêtes à un et à deux degrés tout en gérant le chevauchement, mais qu'il y avait également des lacunes. Certaines lacunes ont pu être traitées par de simples ajustements, d'autres nécessitaient des remèdes plus radicaux, ce qui pourrait remettre en cause l'adoption de l'approche. Quoique la stratégie développée possède la plupart des caractéristiques souhaitables présentées dans la section 2, il est difficile de les satisfaire toutes en même temps. De plus, l'ajout de nouvelles considérations, comme l'estimation de tendances, peut grandement influencer son application. Somme toute, il s'agit surtout d'une méthode en développement. Sans doute, d'autres modifications seront apportées avant sa mise en application finale.

Bibliographie

[1] Keyfitz, N., « Sampling with Probabilities Proportional to Size: Adjustment for Changes in the Probabilities », *Journal of the Americal Statistical Association*, vol. 46, n° 253, p. 105-109, mars 1951.
[2] Rivière, P., « Coordination d'échantillons par la méthode des microstrates », *Recueil du Symposium 2001 de Statistique Canada*, n° 11-522-XIF au catalogue, juillet 2002.
[3] Statistique Canada, *Méthodologie de l'Enquête sur la population active du Canada*, publication hors série, n° 71-526-X au catalogue, juin 2008.
[4] Lafortune, Y., « Revisiting the Use of the Mini-Cluster Approach for CAPI Surveys », Rapport technique présenté à la réunion n° 53 du Comité consultatif sur les méthodes statistiques de Statistique Canada, 31 octobre et 1er novembre 2011.