Tirages coordonnés d'échantillons poissoniens

Desislava NEDYALKOVA,
Lionel QUALITÉ,
Yves TILLÉ
Institut de Statistique,
Université de Neuchâtel, Suisse.

Introduction

De nombreuses méthodes de tirage coordonné d'échantillons ont été développées au sein des instituts nationaux de statistique [voir par exemple à l'Insee 1, 2]. On peut trouver dans [3] une description de certaines d'entre elles et dans [4] une description plus formelle des méthodes de tirage d'échantillons coordonnés. Chacune de ces méthodes fonctionne parfaitement dans le cas idéal d'une population constante, de coordinations toutes positives ou toutes négatives. Et pour chacune, des adaptations existent pour résoudre les problèmes posés par les évolutions de population. Il reste cependant certaines rigidités qui font que l'on ne voit pas bien comment les utiliser pour organiser, par exemple tout le système d'enquêtes auprès des entreprises d'un institut. Avec certaines méthodes, il parait difficile de coordonner une collection d'enquêtes en mélangeant coordinations positives et coordinations négatives, ce qui est le cas par exemple lorsqu'on veut coordonner négativement deux panels rotatifs. Pour d'autres méthodes, qui fournissent des plans simples ou stratifiés de taille fixe à chaque vague d'enquête, ce sont les évolutions de la population qui sont problématiques. Il faut intégrer les naissances et les décès et réellement obtenir un plan simple. La stratification doit souvent être la même pour toutes les enquêtes, ou du moins constituée de blocs fixes, ou alors la méthode doit être apliquée à l'intersection de toutes les strates, ce qui devient vite ingérable.

Une grande partie des problèmes techniques rencontrés avec les méthodes existantes vient du fait que celles-ci cherchent à produire, au moins approximativement, des plans transversaux stratifiés. L'utilisation de plans de Poisson pour la coordination, comme proposé par [5], permet de concevoir un système beaucoup plus simple. La naissance ou le décès d'une unité se traduit par une probabilité d'inclusion qui devient strictement positive, ou bien qui devient nulle, les strates n'en sont plus, mais sont éventuellement remplacées par des domaines où les individus ont les mêmes probabilités d'inclusion. Le prix à payer est que la taille des échantillons n'est plus fixe, et qu'il faut prendre garde à choisir des probabilités d'inclusion de sorte à obtenir des échantillons suffisants dans les domaines d'intérêt. Cela est de toute façon le cas, même avec un plan simple ou stratifié, du fait de la non-réponse.

L'Office Fédéral de la Statistique (Suisse) souhaite disposer d'un système unique pour organiser toutes ses enquêtes auprés des entreprises. Ce système repose sur des échantillons transversaux poissonniens. Il doit permettre de tirer des enquêtes uniques, des panels, des panels rotatifs, d'obtenir de bonnes coordinations positives et/ou négatives, et être adapté à une population dynamique : naissances, décés, fusions, scissions.

1 Charge d'enquête

L'Office Fédéral de la Statistique organise chaque année plusieurs enquêtes auprès des entreprises. Certaines de ces entreprises sont enquêtées à plusieurs reprises. Parfois cela ne peut être évité, par exemple pour les grandes entreprises. Il est malgré tout souhaitable de pouvoir garantir aux entreprises que l'on ne les interroge à nouveau que lorsque cela devient inévitable. Il y a deux aspects à la charge d'enquête des entreprises : le nombre de fois que celles-ci vont être enquétées et le temps qui s'écoule entre les enquêtes. Le nombre moyen de fois qu'elles sont enquétées ne dépend que de leurs probabilités d'inclusion, et on ne peut pas forcément leur promettre de le diminuer. En effet, pour chaque enquête, les entreprises reçoivent des probabilités d'inclusion qui sont calculées de manière à obtenir la meilleure précision possible. Une entreprise qui reçoit les probabilités d'inclusion $\pi^1, \ldots, \pi^r$ pour les enquêtes d'une année donnée sera enquêtée en moyenne $\pi^1 + \cdots + \pi^r$ fois cette année là, de n'importe quelle façon que l'on s'y prenne. Le seul moyen de diminuer ce paramètre est de réduire une ou plusieurs probabilités d'inclusion, et donc de sacrifier la précision des enquêtes.

Néanmoins, dans le cas d'enquêtes réparties dans le temps, on peut essayer de garantir aux entreprises qui sont sélectionnées à une enquête qu'elles vont être "dispensées" d'enquêtes pendant un moment. La régularité avec laquelle les unités sont sélectionnées peut être contrôlée en choisissant une bonne méthode de tirage. Par exemple, si l'on veut faire cinq enquêtes et qu'une entreprise reçoit la probabilité d'inclusion 0.2 à chacune de ces enquêtes, elle sera en moyenne enquêtée une fois. La méthode naïve serait de la sélectionner ou non de manière indépendante à chaque enquête. Elle pourrait donc être échantillonnée 0 fois, 1 fois, 2 fois, etc, jusqu'à 5 fois. En sélectionnant les cinq échantillons de manière coordonnée, on peut garantir à cette entreprise qu'elle va être sélectionnée à une et une seule de ces enquêtes. Lorsqu'elle est échantillonnée, on peut donc lui assurer qu'elle ne participera pas à une autre enquête. Plus généralement, lorsque la somme des probabilités d'inclusion d'une entreprise aux différentes enquêtes est comprise entre deux entiers j et $j+1$, un tirage coordonné permet d'être certain qu'elle va être sélectionnée soit j fois, soit $j+1$ fois. Avec des enquêtes indépendantes, la même unité aurait pu être sélectionnée à toutes les enquêtes. Les tirages coordonnés sont particulièrement intéressants pour les entreprises dont la somme des probabilités d'inclusion est inférieure ou égale à 1 car dans ce cas on peut leur assurer qu'elles seront enquétées une fois au maximum. La méthode que nous développons dans ce document permet d'organiser tous les types d'enquêtes (ponctuelle, panel, panel rotatif...) dans une population dynamique (naissances, décés, scissions et fusions d'unités) en offrant la garantie aux entreprises qu'elles seront enquêtées le moins fréquemment possible, sous contrainte des probabilités d'inclusion qui leur ont été attribuées.

Dans certains cas, la différence entre tirages indépendants et tirages coordonnés est faible. Par exemple, lorsque les probabilités d'inclusion sont très faibles, des échantillons indépendants seront naturellement disjoints. Dans d'autres cas, la coordination aura des effets très nets. Par exemple, lorsque l'on effectue deux enquêtes, dont l'une est très grosse et l'autre très petite (penser à des taux de sondage de 0.9 et 0.1), un tirage indépendant conduira à sélectionner des échantillons qui se recouvrent fortement, alors qu'ils auraient pu être disjoints en faisant un tirage coordonné.

2 Méthode - Objectifs

Nous proposons une méthode permettant de tirer des échantillons coordonnés négativement (recouvrement des échantillons le plus petit possible) ou positivement (recouvrement des échantillons le plus grand possible) avec des enquêtes précédentes. Un ordre de priorité pour ces coordinations entre les enquêtes doit être déterminé avant chaque nouvelle enquête. Par exemple, lors d'une enquête unique, la coordination avec la dernière enquête réalisée pourra être favorisée par rapport à celle avec l'avant-dernière enquête et ainsi de suite. La coordination peut être positive avec certaines enquêtes et négative avec d'autres. Pour une enquête répétée dans le temps, le tirage de l'échantillon pour une année pourra être positivement coordonné avec les tirages pour les vagues précédentes de la même enquête, et négativement avec le tirage de toutes les autres enquêtes.

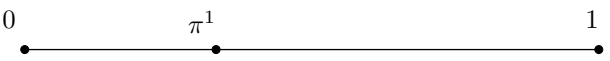
Cette méthode permet de traiter les naissances et les décés dans la base de sondage et de selectionner des panels rotatifs. Elle repose sur des numéros aléatoires permanents, et fournit des échantillons

transversaux sélectionnés selon un plan de Poisson. Une première programmation en SAS a été réalisée, et permet déjà de traiter en quelques minutes au minimum une trentaine d'enquêtes sur une cadre de sondage assez grand (typiquement autour de 400'000 unités pour des sondages en entreprise en Suisse). Les unités de la population sont traitées indépendamment les unes des autres. Les priorités et sens de coordination peuvent être les mêmes pour toutes les unités ou bien être différents.

2.1 Fonctionnement général

Chaque unité de la population est traitée indépendamment des autres unités. Pour comprendre l'algorithme, prenons une unité pour exemple :

– Lors de la première enquête, l'unité a une probabilité de sélection π^1, et reçoit son numéro aléatoire permanent u compris entre 0 et 1. On place alors π^1 sur le segment $[0, 1]$.



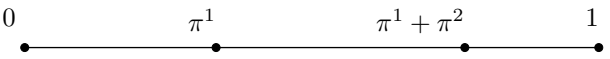
Le segment $[0, 1]$ est ainsi divisé en deux parties (dont l'une pourrait éventuellement etre vide si π^1 est égal à 0 ou 1). Si u est dans la première partie, c'est à dire compris entre 0 et π^1, l'unité est sélectionnée à la première enquête, et sinon elle n'est pas sélectionnée. Il s'agit de l'algorithme usuel pour tirer un échantillon selon plan de Poisson.

– Lors de la deuxième enquête, on peut choisir de coordonner positivement par rapport à la première, ou bien négativement. Ici encore, on va définir une *zone de sélection* constituée d'un ou plusieurs intervalles inclus dans $[0, 1]$ et de longueur totale π^2, la probabilité d'inclusion de l'unité à la deuxième vague. Si l'on veut obtenir une coordination positive, il faudra choisir une zone qui ait la plus grosse intersection possible avec la zone de sélection de la première enquête $[0, \pi^1]$, et au contraire, si l'on veut une coordination négative, il faudra choisir une zone qui ait la plus petite intersection possible avec $[0, \pi^1]$. Ainsi, dans le cas de la coordination positive, la zone de sélection de l'unité à la deuxième enquête sera incluse à $[0, \pi^1]$ si cela est possible, i.e. si π^2 est plus petit que π^1.

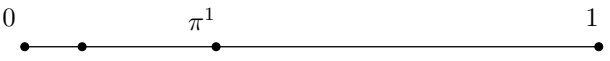
Si π^2 est plus grand que π^1, la zone de sélection contiendra tout l'intervalle $[0, \pi^1]$, et une partie de l'intervalle $[\pi^1, 0]$.

La sélection ou non de l'unité aux deux enquêtes est déterminée par l'appartenance de u aux zones de sélection : si u est dans l'intersection de ces zones, l'unité sera sélectionnée les deux

fois. Si u est dans l'une des zones de sélection mais pas dans l'autre, l'unité ne sera sélectionnée qu'à l'enquête correspondante, et enfin si u n'est dans aucune de ces deux zones, l'unité n'est pas sélectionnée. Le segment $[0,1]$ est maintenant découpé en trois intervalles, et à chacun de ces intervalles correspond un tirage de l'unité soit pour les deux enquêtes, soit pour l'une des enquêtes soit dans aucune des deux enquêtes.
Le deuxième cas possible est celui de la coordination négative. Dans ce cas, il faut déterminer une zone de sélection pour la deuxième enquête qui recoupe le moins possible celle de la première enquête. Il s'agit en fait exactement du même problème que précédemment mais les roles de $[0,\pi^1]$ et $[\pi^1,1]$ ont été intervertis. Si $\pi^1+\pi^2$ est inférieur à 1, la zone de sélection de la deuxième enquête sera $[\pi^1,\pi^1+\pi^2]$ (on prend de préférence la zone de sélection dans $[\pi^1,1]$).



Si $\pi^1+\pi^2$ est superieur à 1, la zone de sélection sera $[\pi^1,1]\cup[0,\pi^2+\pi^1-1]$.



$\pi^1+\pi^2 \bmod 1$

Ici encore, le segment $[0,1]$ est découpé en trois intervalles, et l'appartenance de u à ces intervalles détermine le tirage de l'unité aux deux enquêtes.

– De manière générale, une fois que t enquêtes sont passées, le segment $[0,1]$ est divisé en $t+1$ intervalles, dont certains peuvent être vides. L'appartenance de u à l'un de ces intervalles détermine exactement à quelles enquêtes l'unité fait partie de l'échantillon et à quelles autres elle n'en fait pas partie. Pour déterminer la zone de sélection de l'unité pour la $t+1$-ième enquête, il faut déterminer lesquels de ces $t+1$ intervalles doivent faire partie de cette zone, ou posséder la plus grande intersection possible avec elle, et dans quel ordre de priorité. Si l'on dispose de cette information, la zone de sélection sera composée des intervalles ayant la plus forte priorité tant que la longueur totale de ces intervalles ne dépasse pas la probabilité d'inclusion π^t, et le cas échéant d'une partie de l'intervalle suivant dans l'ordre de priorité de façon à ce que la zone de sélection ait exactement une longueur π^t.

Tout repose donc sur la création et la mise à jour pour chaque individu d'une liste d'intervalles qui forment une partition de $[0,1]$, et des indicateurs de sélection ou non aux différentes vagues pour chacun de ces intervalles. Ces données permettent de réaliser le tirage de l'échantillon à la date t et sont nécéssaires pour calculer le plan de sondage à la date $t+1$. Les données à créer et à conserver pour un individu se présentent sous la forme :

Intervalle	e_1	e_2	$\dots$	e_t
a_1	1	0	$\dots$	0
a_2	1	1	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_{t+1}	0	0	$\dots$	1

avec $\sum_{i=1}^{t+1} a_i = 1$, le premier intervalle étant $[0,a_1]$, le deuxième $[a_1,a_1+a_2]$, etc. Il faut également conserver le numéro aléatoire u.

2.2 Traitement des naissances et des décès

Cette méthode est aisément adaptable au cas d'une population dynamique. En effet, les unités sont traitées indépendamment les unes des autres, et la naissance ou le décés d'une unité n'influe pas sur la sélection des autres (autrement que dans le calcul des probabilités d'inclusion pour les enquêtes futures). De plus, le traitement de chaque unité est séquentiel : la décision de la sélectionner à un temps donné ne dépend pas de ses caractéristiques futures, ni de sa durée de vie. En pratique, pour ajouter une unité qui vient de naitre, il suffit de lui affecter une probabilité d'inclusion nulle à toutes les vagues précédentes. De cette manière, elle est introduite dans l'enquête avec un passé vierge : un certain nombre d'intervalles fictifs de longueur nulle. Son premier 'vrai' intervalle est créé à la première enquête où elle a une probabilité non nulle d'être sélectionnée. Pour prendre en compte le décés d'une unité, on peut soit la conserver dans la base de sondage avec une probabilité d'inclusion nulle pour les vagues suivantes, soit la supprimer de la base de sondage (mais ce n'est pas une grande économie, et pas forcément judicieux).

2.3 Traitement des fusions et scissions d'unités

Dans une population dynamique, il est possible que deux (ou plus) unités fusionnent, ou qu'une unité se scinde en deux (ou plus) nouvelles unités. L'exemple le plus fréquent est celui des entreprises, mais cela peut aussi se produire dans le cas de ménages ou de logement. Le cas des scissions ne pose naturellement pas de problèmes, le seul choix à faire est de conserver le passé de ces unités ou bien de l'effacer.

Dans le cas des fusions par contre, l'organisateur d'enquêtes coordonnées va devoir tenir compte de l'impossibilité d'enquêter une des unités qui ont fusionné sans enquêter les autres. Par exemple, si deux entités qui appartiennent à deux panels différents fusionnent, il faudra décider si l'on considère que l'entité résultant de la fusion appartient aux deux panels, ou bien seulement à l'un ou l'autre, ou bien même à aucun des deux. Voici quelques choix possibles :

- la nouvelle unité est considérée comme vraiment nouvelle. On ne se préoccupe pas de son passé, tout se passe comme si les unités qui fusionnent étaient décédées et le résultat de la fusion est une naissance. Comme vu précédemment, cela ne pose pas de problème.

- Deux (ou plus) unités fusionnent et l'une de ces unités est considérée comme dominante par rapport aux autres. Il est alors naturel de ne retenir que le passé de cette unité, et de ne pas tenir compte du passé des autres unités de la fusion.

- L'unité résultant de la fusion hérite des caractéristiques des anciennes unités qui la composent. Par exemple, si deux unités k et ℓ avec des indicatrices d'appartenance s_k^t et s_ℓ^t se fondent en une unité m, on veut pouvoir décréter que $s_m^t = min(s_k^t, s_\ell^t)$ ou bien au contraire que $s_m^t = max(s_k^t, s_\ell^t)$. Cela permet entre autres de choisir d'inclure m dans deux panels auxquels appartiennent respectivement k et ℓ, ou bien de s'assurer que le fardeau de réponse est limité au maximum.

Seul le dernier cas nécéssite un développement (complexe) du point de vue des problèmes de sélèction et de coordination. Il s'agit de fusionner deux entités avec leur deux plans de sondages longitudinaux et leur deux nombres aléatoires permanents. Prenons le cas de la fusion de deux unités, dont les intervalles de tirage sont respectivement $a_1, \ldots, a_{t+1}$ et $b_1, \ldots, b_{t+1}$, et leurs numéros aléatoires u et v. On utilisera les notations a_i et b_j indifféremment pour désigner les intervalles ou leur longueur.

A chaque intervalle a_i correspond un échantillon longitudinal pour l'unité k, et à chaque intervalle b_j correspond un échantillon longitudinal pour l'unité ℓ. Les couples d'échantillons longitudinaux $(\mathbf{s}_k, \mathbf{s}_l)$ sont en bijection avec les rectangles $a_i \times b_j$ de la Figure 1. La probabilité de tirer un couple $(\mathbf{s}_k, \mathbf{s}_l)$ est égale à l'aire $a_i \cdot b_j$ du rectangle qui lui correspond.

Notre objectif est de nous ramener à une situation comparable à celle que l'on a avec une unité "normale" : $t+1$ échantillons longitudinaux possibles (au maximum) représentés par leur probabilité de tirage $(c_j)_{j=1,\ldots,t+1}$ et un numéro aléatoire.

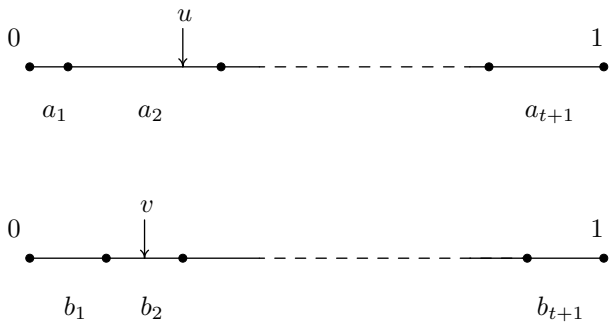
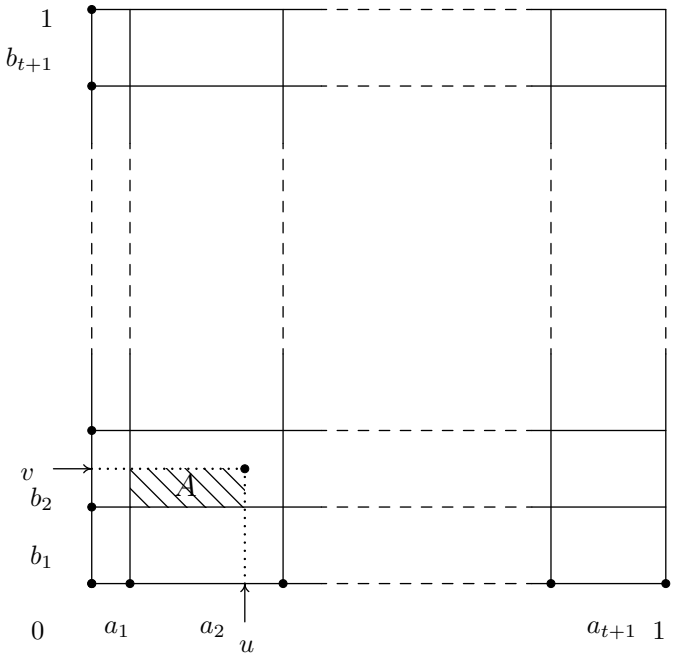


FIGURE 1 – Fusion de deux unités : plan joint



Une première étape qui ne présente pas de difficulté est de se ramener à une partition du segment $[0, 1]$. On peut par exemple poser $c_1 = a_1 \cdot b_1$, $c_2 = a_1 \cdot b_2$, et ainsi de suite jusqu'à $c_{t+1} = a_1 \cdot b_{t+1}$, puis $c_{t+1} = a_2 \cdot b_1$, et cœtera, et enfin $c_{(t+1)^2} = a_{t+1} \cdot b_{t+1}$. L'ordre dans lequel on place les aires des rectangles n'a aucune importance, les segments $c_1, \ldots, c_{(t+1)^2}$ peuvent être réordonnés au besoin, pour peu que le nombre aléatoire soit modifié en conséquence. Le couple de nombres aléatoires (u, v) peut être remplacé par n'importe quel point w du segment correspondant au rectangle dans lequel il se trouve. Si ce segment est noté $[a, b]$, on peut naturellement choisir $w = a + A$ (voir Figure 1). On est donc ramené à une partition de $[0, 1]$ en $(t + 1)^2$ segments. Chaque segment correspond à un couple d'échantillons longitudinaux $(\mathbf{s}_k, \mathbf{s}_\ell)$ pour les unités à fusionner. Un nouveau nombre aléatoire w correspond au tirage donné par le couple (u, v).

Pour pouvoir intégrer la nouvelle unité dans le système de coordination, il faut être capable de calculer des scores pour chaque intervalle de sélection, il faut donc transformer les couples $(\mathbf{s}_k, \mathbf{s}_\ell)$ en des vecteurs $\mathbf{s}_m \in \{0, 1\}^t$. Cette transformation, qui entraine une perte d'information, est nécéssaire pour calculer des scores pour chaque échantillon et poursuivre la coordination des enquêtes. Les

informations relatives aux unités k et ℓ doivent être conservées à part car elles sont nécessaires aux procédures d'estimation mais pour la coordination et le plan de sondage, on ne se sert plus que des échantillons $\mathbf{s}_m$. On définit donc une "fonction de fusion" :

$$f = \begin{pmatrix} \{0,1\}^t \times \{0,1\}^t & \longrightarrow & \{0,1\}^t \\ (\mathbf{s}_k, \mathbf{s}_\ell) & \longmapsto & \mathbf{s}_m \end{pmatrix}.$$

Par exemple, on peut choisir :
- $f(\mathbf{s}_k, \mathbf{s}_\ell) = \mathbf{s}_k$,
- $f(\mathbf{s}_k, \mathbf{s}_\ell) = \mathbf{s}_\ell$,
- $f(\mathbf{s}_k, \mathbf{s}_\ell) = (max(s_k^1, s_\ell^1), \ldots, max(s_k^t, s_\ell^t))$,
- $f(\mathbf{s}_k, \mathbf{s}_\ell) = (min(s_k^1, s_\ell^1), \ldots, min(s_k^t, s_\ell^t))$.

N.B. : Des probabilités d'inclusion de l'unité m dans chaque vague d'enquête peuvent être calculées a posteriori. Ces probabilités dépendent du choix de la fonction f lors de la fusion. Pour les fonctions données en exemple, on aura respectivement $\pi_m^i = \pi_k^i$, $\pi_m^i = \pi_\ell^i$, $\pi_m^i = \pi_k^i + \pi_\ell^i - \pi_k^i \cdot \pi_\ell^i$, et $\pi_m^i = \pi_k^i \cdot \pi_\ell^i$.

A ce stade, on peut calculer un score pour chaque segment en fonction des sens de coordination et des priorités choisies pour la $t+1^{\text{ième}}$ enquête. Cependant l'unité m se distingue des autres par le fait qu'elle a typiquement $(t+1)^2$ échantillons longitudinaux possibles au temps t, alors que les autres n'en ont que $t+1$. Bien qu'il soit possible de traiter les unités résultant d'une fusion à part, cela risquerait de rendre le système instable. En effet, si par la suite cette unité fusionnait encore avec une autre unité, il faudrait envisager $(t+1)^3$ échantillons, et ainsi de suite. La quantité de données à stocker et à traiter risquerait de rendre rapidement la méthode inutilisable.

Le système sera largement simplifié si l'on ne retient que $t+1$ échantillons longitudinaux pour les unités provenant d'une fusion, de manière à pouvoir les traiter comme des unités normales de la population. Une unité pour laquelle il n'y a déjà pas plus de $t+1$ échantillons longitudinaux possibles, par exemple une très grande entreprise qui est toujours enquêtée, peut être laissée telle quel. Pour les autres unités, il faut accepter de perdre encore de l'information si l'on veut une méthode qui puisse s'appliquer à une période longue ou à un grand nombre d'enquêtes. On doit donc choisir une nouvelle fonction de compression

$$g_m = \begin{pmatrix} \{0,1\}^t & \longrightarrow & \{0,1\}^t \\ \mathbf{s}_m & \longmapsto & \tilde{\mathbf{s}}_m \end{pmatrix}$$

qui prenne au plus $t+1$ valeurs distinctes quand $\mathbf{s}_m$ parcourt l'ensemble des $(t+1)^2$ échantillons longitudinaux possibles pour l'unité m. Cette fonction doit être choisie de manière à perdre le moins d'information possible.

Si l'on ne considérait que des coordinations négatives et que le seul but était de maximiser le temps passé hors échantillon une fois que l'unité à été tiré, on pourrait se contenter de ne retenir que la dernière fois que m a été "enquêtée" :

$$g_m : \mathbf{s}_m = (s_m^1, \ldots, s_m^{i-1}, 1^i, 0, \ldots, 0) \mapsto \tilde{\mathbf{s}}_m = (0, \ldots, 0, 1^i, 0 \ldots, 0).$$

Malheureusement cette solution simple n'est pas acceptable lorsqu'il y a des panels ou des panels rotatifs, pour lesquels l'information essentielle n'est pas la dernière date d'enquête mais l'appartenance à une vague précédente du panel. Il faut donc chercher une meilleure fonction g_m. Les aspects les plus importants me semblent être :

1. Conserver l'échantillon $\mathbf{s}_m(w)$ effectivement tiré, associé au segment dans lequel se situe w. De cette manière, les échantillons longitudinaux futurs pour l'unité m sont compatibles avec les données qui ont réellement été recueillies (modulo le choix de f).

2. Conserver les probabilités d'inclusion $\pi_m^1, \ldots, \pi_m^t$ qui résultent du choix de f. Ainsi, par la suite, la coordination avec la vague ayant la plus forte priorité sera réalisée correctement.

3. Retenir une méthode qui puisse être appliquée. Pour certains problèmes d'énumération on ne connait pas d'algorithme efficace (de complexité polynômiale). Pour d'autres il existe des algorithmes mais qui ne sont pas libres, ou implémentés en SAS.

4. Une fois ces points vérifiés, choisir un algorithme qui permet de dégrader le moins possible l'information. En particulier,
   - choisir de manière raisonnée la longueur de l'intervalle correspondant à $\mathbf{s}_m(w)$, par exemple égale à la valeur de départ $p(\mathbf{s}_m(w))$. Cela pourrait être nécessaire pour assurer que la coordination négative est la meilleure possible, et s'assurer qu'une unité avec une probabilité d'inclusion assez faible restera bien hors échantillon aussi longtemps que souhaité. Ou bien choisir une longueur la plus petite ou au contraire la plus grande possible,
   - ne pas forcément choisir un plan à support minimal, mais au contraire retenir $t+1$ échantillons distincts (s'il y en avait plus de $t+1$ au départ),
   - avoir une méthode adaptée au vecteur de probabilités d'inclusion, par exemple s'il contient des 1 ou des 0,
   - utiliser un critère d'information pour le choix (entropie, compression avec perte ?).

Il est théoriquement possible de ne conserver que $\mathbf{s}_m(w)$ et t autres échantillons longitudinaux $\mathbf{s}_m(1), \ldots, \mathbf{s}_m(t)$ parmi ceux obtenus après la fusion, et de définir un plan de sondage sur ces échantillons de manière à respecter les probabilités d'inclusion à posteriori $\pi_m^1, \ldots, \pi_m^t$. Il n'est par contre pas toujours possible de garder à la fois le vecteur de probabilités d'inclusion, $\mathbf{s}_m(w)$, et la probabilité d'origine $p(\mathbf{s}_m(w))$. On peut toutefois choisir la nouvelle probabilité $\tilde{p}(\mathbf{s}_m(w))$ parmi un ensemble fini de valeurs. Pour chacune de ces valeurs il y a également plusieurs choix possibles pour les t échantillons complémentaires $\mathbf{s}_m(1), \ldots, \mathbf{s}_m(t)$.

Une solution beaucoup plus simple du point de vue algorithmique est de garder $\mathbf{s}_m(w)$ avec sa probabilité $p(\mathbf{s}_m(w))$ et de choisir t autres échantillons parmi tous les échantillons possibles, et non plus seulement parmi ceux qui étaient réalisables du fait de la coordination des unités impliquées dans la fusion.

Quant à la fusion de plus de deux unités... on peut faire le même développement que précédemment, ou bien faire une fusion de deux unités puis fusionner la troisième avec le résultat, etc.

...à suivre.

2.4 "Oubli" d'anciennes enquêtes

Il est possible de ne plus vouloir faire intervenir explicitement certaines enquêtes dans la coordination des enquêtes futures. Cela ne veut pas dire qu'elles ne seront plus coordonnées avec ces enquêtes (dans le sens où la taille de l'intersection serait en moyenne égale à celle que l'on aurait si les enquêtes étaient indépendantes). En effet, si l'une de ces 'enquêtes oubliées' est coordonnée avec une enquête qui reste dans le système, alors il y aura toujours coordination même si celle-ci n'est pas contrôlée.

En pratique, si l'on est à la $t^{\text{ième}}$ enquête dans la situation décrite précédemment, avec des intervalles $a_1, \ldots, a_{t+1}$ et des indicatrices d'appartenance aux différents échantillons $e_1, \ldots, e_t$,

FIGURE 2 – Plan de sondage pour une unité

Intervalle	e_1	e_2	$\ldots$	e_t
a_1	1	0	$\ldots$	0
a_2	1	1	$\ldots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_{t+1}	0	0	$\ldots$	1

et si l'on veut oublier la première enquête, il suffit de supprimer la colonne correspondante du tableau de la Figure 2, de permuter les lignes de manière à rassembler les intervalles qui ont les mêmes valeurs pour $e_2, \ldots, e_t$, et enfin de fusionner ceux-ci. Le numéro aléatoire devra être modifié au moment de la permutation pour 'suivre' l'intervalle auquel il appartient.

2.5 Traitement des panels rotatifs

La méthode décrite permet de tirer des panels rotatifs. Supposons que l'on ait déjà tiré $t-1$ enquêtes dans la population et que l'on veuille intégrer une nouvelle enquête par panel rotatif, avec un taux de renouvellement de un cinquième par exemple, coordonnée avec les enquêtes précédentes. Les probabilités d'inclusion de ce panel sont notées π_k^p. On va tirer cinq échantillons $s^t, \cdots, s^{t+1}$ qui constitueront l'échantillon de départ du panel s^p.

On sélectionne en premier, par la méthode décrite ci-avant, s^t avec les probabilités d'inclusion $\pi_k^p/5$ et les règles de coordination souhaitées avec les $t-1$ enquêtes précédentes. Puis on sélectionne s^{t+1}, toujours avec les probabilités d'inclusion $\pi_k^p/5$ et des règles de coordination légèrement différentes : en première priorité on demande la coordination négative avec l'enquête s^t, puis on introduit les coordinations voulues avec les enquêtes précédentes. Puis on tire s^{t+2} coordonné en priorité négativement avec s^t et s^{t+1}, ainsi de suite. Une fois les cinq échantillons sélectionnés, on pose

$$s^p = s^t \sqcup \cdots \sqcup s^{t+4}.$$

La partie s^t qui est tirée en premier respecte le mieux la coordination demandée initialement, tandis que s^{t+4} est tirée en dernier et donc respecte le moins bien la coordination. Enfin ces cinq parties sont tirées avec des probabilités d'inclusion au plus égales à 0.2 et sont négativement coordonnées. Elles sont automatiquement disjointes avec notre méthode.

Une fois le moment venu de renouveler le panel, si l'on a déjà effectué $u-1$ enquêtes, commence par mettre à jour la partie du panel que l'on veut conserver pour compenser son érosion. On tire donc s^u coordonné positivement avec s^t et négativement avec les autres morceaux, toujours avec les probabilités d'inclusion $\pi^k/5$. Puis $s^{u+1}, \ldots, s^{u+3}$ pour renouveler $s^{t+1}, \ldots, s^{t+3}$. Enfin, on tire un nouvel échantillon s^{u+4} avec les probabilités d'inclusion $\pi_k^p/5$, coordonné négativement avec $s^u, \ldots, s^{u+3}$ et avec la partie à renouveler s^{t+4}.

$$s^{p+1} = s^u \sqcup \cdots \sqcup s^{u+4}.$$

De cette manière s^t, qui est la partie la mieux coordonnée avec les autres enquêtes reste dans le panel le plus longtemps, tandis que s^{t+4} est renouvelée en premier.

2.6 Propriétés du plan obtenu

- Les échantillons transversaux sont des échantillons poissonniens, et les probabilités d'inclusion jointes entre les enquêtes sont faciles à calculer. Les variances sont donc très simples à calculer pour les estimateurs transversaux et longitudinaux.

- Dans le cas de coordinations toutes négatives, le plan longitudinal pour chaque unité est un plan systématique. D'après [4], c'est un plan qui a de bonnes propriétés pour cet usage. En particulier, lorsque le taux de sondage est égal à $1/p$ pour chaque enquête, il garantit à chaque unité d'être sélectionnée exactement chaque p enquêtes.

- La méthode fournit une coordination optimale dans le sens suivant : la coordination entre une enquête (par exemple S^t) et celle avec laquelle elle devait être coordonnée en priorité (par exemple S^u) est effectivement optimale. C'est à dire que les probabilités d'inclusion jointes $\pi_k^{t,u}$ des unités à ces enquêtes sont égales aux bornes optimales données par exemple dans [3], pour la coordination positive comme pour la coordination négative. Puis, *parmi les plans pour lesquels S^t est coordonnée optimalement avec S^u*, celui obtenu est optimalement coordonné avec la deuxième enquête dans l'ordre de priorité, et ainsi de suite.

Conclusion

Nous proposons une méthode de sélection coordonnée d'échantillons poissoniens. La méthode est adaptée à une population dynamique, c'est à dire qu'elle permet d'avoir des naissances, des décés, des

scissions et des fusions d'unités dans la base de sondage (ce dernier point n'est pas encore totalement résolu). Elle permet de coordonner positivement ou négativement des enquêtes, de gérer des enquêtes uniques mais aussi des panels et des panels rotatifs. Enfin, elle fournit, dans un certain sens, une coordination optimale.

Bibliographie

[1] Cotton, F. et Hesse, C. (1992). Tirages coordonnes d'échantillons. Document de travail de la Direction des Statistiques Économiques E9206. Technical report, INSEE, Paris.

[2] Rivière, P. (2001). Coordinating samples using the microstrata methodology. Proceedings of Statistics Canada Symposium 2001.

[3] Hesse, C. (1999). Sampling co-ordination : a review by country. Technical Report E9908, INSEE.

[4] Nedyalkova, D., Qualité, L. et Tillé, Y. (2009). General framework for the rotation of units in repeated survey sampling. *Accepté pour publication dans Statistica Neerlandica.*

[5] Brewer, K., Early, L., and Joyce, S. (1972). Selecting several samples from a single population. Australian Journal of Statistics, 3 :231–239.