Bootstrap dans les enquêtes par sondage
Application à l'enquête Histoire de Vie et Patrimoine

Guillaume Chauvet
Travail joint avec Emmanuel Gros, Olivier Guin et Jean Rubin (Insee)

École Nationale de la Statistique et de l'Analyse de l'Information

23/02/2023

Le bootstrap, pour quoi faire?

Dans le cadre d'une enquête, on aimerait assortir les estimations produites d'une mesure de précision (estimation de variance, coefficient de variation, intervalle de confiance).

C'est une tâche difficile, où il faut prendre en compte le plan de sondage, les corrections apportées à l'estimation (redressement de la non-réponse, calage), la forme du paramètre.

L'utilisateur de données d'enquête a généralement une connaissance limitée du plan de sondage (stratification, unités primaires) et de l'estimation (traitement de la non-réponse, calage). Parfois, cette connaissance se résume à une variable de pondération intégrant les différents traitements.

Le bootstrap, pour quoi faire?

Technique computationnelle (Efron, 1979) qui permet (en théorie) d'estimer la distribution d'un estimateur, en reproduisant de façon répétée le mécanisme de sélection et la procédure d'estimation utilisée.

Dans le cadre des enquêtes, l'objectif est souvent plus simplement de produire un estimateur de variance. Le bootstrap consiste alors :

- à répliquer B fois la création des poids d'extrapolation par rééchantillonnage + réestimation,
- à obtenir ainsi un jeu de B poids bootstrap,
- à les utiliser pour calculer les statistiques bootstrappées.
  Leur dispersion est alors utilisée comme estimateur de variance.

Procédure très simple d'un point de vue utilisateur.
Du point de vue du bootstrappeur, c'est un peu plus compliqué.

Le plan

1. Le bootstrap sur données i.i.d.
2. Les méthodes de bootstrap en Sondages
3. Enquête Histoire de Vie et Patrimoine

Le bootstrap sur données i.i.d.

Estimateur par substitution

Supposons que l'on s'intéresse à une distribution, que l'on assimile à sa fonction de répartition $F(\cdot)$. On veut estimer un paramètre $\theta(F) \equiv \theta$ de la distribution.

On suppose dans cette partie que l'on dispose d'un échantillon de n valeurs $(X_1, \ldots, X_n)^\top$ générées indépendamment selon F, et donc indépendantes et identiquement distribuées. On calcule la fonction de répartition empirique

$$F_n(\cdot) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} 1(X_i \leq \cdot),$$

qui donne une proba. identique $(1/n)$ aux observations de l'échantillon.

On l'utilise pour produire un estimateur $\theta(F_n) \equiv \hat{\theta}$, selon un principe de substitution. Sous des conditions générales (e.g., Shao, 1992), c'est un estimateur asymptotiquement sans biais et consistant du paramètre θ.

Principe du bootstrap

On veut assortir l'estimateur $\theta(F_n)$ d'une mesure de précision. Le bootstrap permet de le faire en produisant une estimation empirique de la distribution de $\theta(F_n) - \theta(F)$.

Le principe consiste à reproduire approximativement et de façon répétée le mécanisme d'échantillonnage et le mécanisme d'estimation :

$$F \xrightarrow{\text{Echant.}} F_n \xrightarrow{\text{Estim.}} \theta(F_n) \quad \Rightarrow \quad F_n \xrightarrow{\text{Rééchant.}} F_n^* \xrightarrow{\text{Réestim.}} \theta(F_n^*)$$

Procédure de base :

Procédure bootstrap :

Principe du bootstrap

On veut assortir l'estimateur $\theta(F_n)$ d'une mesure de précision. Le bootstrap permet de le faire en produisant une estimation empirique de la distribution de $\theta(F_n) - \theta(F)$.

Le principe consiste à reproduire approximativement et de façon répétée le mécanisme d'échantillonnage et le mécanisme d'estimation :

$$F \xrightarrow{\text{Echant.}} F_n \xrightarrow{\text{Estim.}} \theta(F_n) \quad \Rightarrow \quad F_n \xrightarrow{\text{Rééchant.}} F_n^* \xrightarrow{\text{Réestim.}} \theta(F_n^*)$$

Procédure de base :

- Paramètre d'intérêt : $\theta(F)$

Procédure bootstrap :

- Paramètre d'intérêt : $\theta(F_n)$

Principe du bootstrap

On veut assortir l'estimateur $\theta(F_n)$ d'une mesure de précision. Le bootstrap permet de le faire en produisant une estimation empirique de la distribution de $\theta(F_n) - \theta(F)$.

Le principe consiste à reproduire approximativement et de façon répétée le mécanisme d'échantillonnage et le mécanisme d'estimation :

$$F \xrightarrow{\text{Echant.}} F_n \xrightarrow{\text{Estim.}} \theta(F_n) \quad \Rightarrow \quad F_n \xrightarrow{\text{Rééchant.}} F_n^* \xrightarrow{\text{Réestim.}} \theta(F_n^*)$$

Procédure de base :

- Paramètre d'intérêt : $\theta(F)$
- Echantillonnage : tirage de $(X_1, \ldots, X_n) \sim_{iid} F$. Calcul de $F_n(\cdot) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \leq \cdot)$.

Procédure bootstrap :

- Paramètre d'intérêt : $\theta(F_n)$
- Echantillonnage : tirage de $(X_1^*, \ldots, X_m^*) \sim_{iid} F_n$ où $m = n - 1$. Calcul de $F_n^*(\cdot) = \frac{1}{m} \sum_{i=1}^{m} 1(X_i^* \leq \cdot)$.

Principe du bootstrap

On veut assortir l'estimateur $\theta(F_n)$ d'une mesure de précision. Le bootstrap permet de le faire en produisant une estimation empirique de la distribution de $\theta(F_n) - \theta(F)$.

Le principe consiste à reproduire approximativement et de façon répétée le mécanisme d'échantillonnage et le mécanisme d'estimation :

$$F \xrightarrow{\text{Echant.}} F_n \xrightarrow{\text{Estim.}} \theta(F_n) \quad \Rightarrow \quad F_n \xrightarrow{\text{Rééchant.}} F_n^* \xrightarrow{\text{Réestim.}} \theta(F_n^*)$$

Procédure de base :

- Paramètre d'intérêt : $\theta(F)$
- Echantillonnage : tirage de $(X_1, \ldots, X_n) \sim_{iid} F$. Calcul de $F_n(\cdot) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \leq \cdot)$.
- Estimation : calcul de $\theta(F_n)$

Procédure bootstrap :

- Paramètre d'intérêt : $\theta(F_n)$
- Echantillonnage : tirage de $(X_1^*, \ldots, X_m^*) \sim_{iid} F_n$ où $m = n - 1$. Calcul de $F_n^*(\cdot) = \frac{1}{m} \sum_{i=1}^{m} 1(X_i^* \leq \cdot)$.
- Estimation : calcul de $\theta(F_n^*)$

Principe du bootstrap : exemple

i	Y_i	Z_i					
1	4.51	8.88					
2	4.27	9.30					
3	2.62	8.16					
4	3.64	7.53					
5	4.23	7.14					
6	5.03	8.83					
...	...	...					
$\bar{Y}_n$	3.86						

1. Tirage d'un rééchant. $(X_1^*, \ldots, X_{n-1}^*) \sim_{iid} F_n$. On résume par un jeu de poids bootstrap W.

Principe du bootstrap : exemple

i	Y_i	Z_i		W_i^1			
1	4.51	8.88		$\frac{50}{49} \times 2$			
2	4.27	9.30		0			
3	2.62	8.16		$\frac{50}{49} \times 1$			
4	3.64	7.53		$\frac{50}{49} \times 1$			
5	4.23	7.14		$\frac{50}{49} \times 2$			
6	5.03	8.83		0			
...	...	...		...			
$\bar{Y}_n$	3.86		$\bar{Y}_n^*$				

1. Tirage d'un rééchant. $(X_1^*, \ldots, X_{n-1}^*) \sim_{iid} F_n$. On résume par un jeu de poids bootstrap W.

Principe du bootstrap : exemple

i	Y_i	Z_i		W_i^1			
1	4.51	8.88		$\frac{50}{49} \times 2$			
2	4.27	9.30		0			
3	2.62	8.16		$\frac{50}{49} \times 1$			
4	3.64	7.53		$\frac{50}{49} \times 1$			
5	4.23	7.14		$\frac{50}{49} \times 2$			
6	5.03	8.83		0			
...	...	...		...			
$\bar{Y}_n$	3.86		$\bar{Y}_n^*$	4.04			

1. Tirage d'un rééchant. $(X_1^*, \ldots, X_{n-1}^*) \sim_{iid} F_n$. On résume par un jeu de poids bootstrap W.
2. Calcul de l'équivalent bootstrap :

$$\theta_1(F_n) = \bar{Y}_n \quad \Rightarrow \quad \theta_1(F_n^*) = \bar{Y}_n^* = \frac{\sum_{i=1}^n W_i Y_i}{\sum_{i=1}^n W_i}$$

Principe du bootstrap : exemple

i	Y_i	Z_i		W_i^1	W_i^2	$\cdots$	W_i^B
1	4.51	8.88		$\frac{50}{49} \times 2$	$\frac{50}{49} \times 1$	$\cdots$	$\frac{50}{49} \times 1$
2	4.27	9.30		0	0	$\cdots$	$\frac{50}{49} \times 2$
3	2.62	8.16		$\frac{50}{49} \times 1$	$\frac{50}{49} \times 1$	$\cdots$	0
4	3.64	7.53		$\frac{50}{49} \times 1$	$\frac{50}{49} \times 1$	$\cdots$	$\frac{50}{49} \times 2$
5	4.23	7.14		$\frac{50}{49} \times 2$	0	$\cdots$	$\frac{50}{49} \times 1$
6	5.03	8.83		0	$\frac{50}{49} \times 1$	$\cdots$	$\frac{50}{49} \times 3$
$\cdots$	$\cdots$	$\cdots$		$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\bar{Y}_n$	3.86		$\bar{Y}_n^*$	4.04	3.87	$\cdots$	3.95

1. Tirage d'un rééchant. $(X_1^*, \ldots, X_{n-1}^*) \sim_{iid} F_n$. On résume par un jeu de poids bootstrap W.
2. Calcul de l'équivalent bootstrap :

$$\theta_1(F_n) = \bar{Y}_n \quad \Rightarrow \quad \theta_1(F_n^*) = \bar{Y}_n^* = \frac{\sum_{i=1}^n W_i Y_i}{\sum_{i=1}^n W_i}$$

3. On répète B fois les étapes 1-2 pour obtenir les stats bootstrappées.

Principe du bootstrap : exemple

i	Y_i	Z_i		W_i^1	W_i^2	$\cdots$	W_i^B
1	4.51	8.88		$\frac{50}{49} \times 2$	$\frac{50}{49} \times 1$	$\cdots$	$\frac{50}{49} \times 1$
2	4.27	9.30		0	0	$\cdots$	$\frac{50}{49} \times 2$
3	2.62	8.16		$\frac{50}{49} \times 1$	$\frac{50}{49} \times 1$	$\cdots$	0
4	3.64	7.53		$\frac{50}{49} \times 1$	$\frac{50}{49} \times 1$	$\cdots$	$\frac{50}{49} \times 2$
5	4.23	7.14		$\frac{50}{49} \times 2$	0	$\cdots$	$\frac{50}{49} \times 1$
6	5.03	8.83		0	$\frac{50}{49} \times 1$	$\cdots$	$\frac{50}{49} \times 3$
$\ldots$	$\ldots$	$\ldots$		$\ldots$	$\ldots$	$\ldots$	$\ldots$
$Corr_{F_n}(Y, Z)$	0.52		$Corr_{F_n^*}(Y, Z)$	0.33	0.52	$\cdots$	0.58

1. Tirage d'un rééchant. $(X_1^*, \ldots, X_{n-1}^*) \sim_{iid} F_n$. On résume par un jeu de poids bootstrap W.
2. Calcul de l'équivalent bootstrap de $\theta_2(F_n)$

$$\theta_2(F_n) = Corr_{F_n}(Y, Z) \quad \Rightarrow \quad \theta_2(F_n^*) = \frac{\sum_{i=1}^n W_i (Y_i - \bar{Y}_n^*)(Z_i - \bar{Z}_n^*)}{\sqrt{\sum_{i=1}^n W_i (Y_i - \bar{Y}_n^*)^2 \times \sum_{j=1}^n W_j (Z_j - \bar{Z}_n^*)^2}}$$

3. On répète B fois les étapes 1-2 pour obtenir les stats bootstrappées.

Estimation de variance bootstrap

On estime $V(\hat{\theta} - \theta)$ par la dispersion des estimateurs bootstrappés $\hat{\theta}^* - \hat{\theta}$.
On obtient l'estimateur de variance bootstrap

$$v_{boot}^{B}(\hat{\theta}) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}^{b*} - \frac{1}{B} \sum_{c=1}^{B} \hat{\theta}^{c*} \right)^2 .$$

Si l'estimateur $\hat{\theta}$ suit asymptotiquement une loi normale, on obtient l'intervalle de confiance basé sur la normalité :

$$IC_{95\%}^{nor}(\theta) = \left[\hat{\theta} - 1.96\sqrt{v_{boot}^{B}(\hat{\theta})}, \hat{\theta} + 1.96\sqrt{v_{boot}^{B}(\hat{\theta})} \right]$$

Retour sur l'exemple (coef. de corrélation) :
On obtient $\hat{\theta}_2 = 0.52$, et $\hat{\theta}_2^* \in \{0.33, 0.52, ..., 0.58\}$ $(B = 200)$.
On obtient $v_{boot}^{B}(\hat{\theta}) = 0.0076$, puis

$$IC_{95\%}^{nor}(\theta) = [0.52 \pm 0.17] = [0.35, 0.69] .$$

Les méthodes de bootstrap en Sondages

Cadre général

Population finie $U = \{1, \ldots, N\}$. On utilise un plan de sondage $p(\cdot)$ respectant des probas d'inclusion $\pi_k > 0$ pour tirer un échant. aléatoire S.

On s'intéresse à un paramètre $\theta = f(t_y)$, où la fonction $f(\cdot)$ est connue mais le p-vecteur des totaux $t_y = \sum_{k \in U} y_k$ est inconnu.

Cadre général

Population finie $U = \{1, \ldots, N\}$. On utilise un plan de sondage $p(\cdot)$ respectant des probas d'inclusion $\pi_k > 0$ pour tirer un échant. aléatoire S.

On s'intéresse à un paramètre $\theta = f(t_y)$, où la fonction $f(\cdot)$ est connue mais le p-vecteur des totaux $t_y = \sum_{k \in U} y_k$ est inconnu. On l'estime par substitution

$$\hat{\theta} = f(\hat{t}_{y\pi}) \quad \text{avec} \quad \hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

Cadre général

Population finie $U = \{1, \ldots, N\}$. On utilise un plan de sondage $p(\cdot)$ respectant des probas d'inclusion $\pi_k > 0$ pour tirer un échant. aléatoire S.

On s'intéresse à un paramètre $\theta = f(t_y)$, où la fonction $f(\cdot)$ est connue mais le p-vecteur des totaux $t_y = \sum_{k \in U} y_k$ est inconnu. On l'estime par substitution

$$\hat{\theta} = f(\hat{t}_{y\pi}) \quad \text{avec} \quad \hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

Comparaison avec le cadre i.i.d. :

$$\text{Cadre i.i.d. : } F \xrightarrow{\text{Echant.}} F_n \xrightarrow{\text{Estim.}} \hat{\theta}$$

$$\text{Cadre sondages : } U \xrightarrow{\text{Echant.}} S \xrightarrow{\text{Estim.}} \hat{\theta}$$

Cadre général

Population finie $U = \{1, \ldots, N\}$. On utilise un plan de sondage $p(\cdot)$ respectant des probas d'inclusion $\pi_k > 0$ pour tirer un échant. aléatoire S.

On s'intéresse à un paramètre $\theta = f(t_y)$, où la fonction $f(\cdot)$ est connue mais le p-vecteur des totaux $t_y = \sum_{k \in U} y_k$ est inconnu. On l'estime par substitution

$$\hat{\theta} = f(\hat{t}_{y\pi}) \quad \text{avec} \quad \hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

Comparaison avec le cadre i.i.d. :

$$\text{Cadre i.i.d. : } F \xrightarrow{\text{Echant.}} F_n \xrightarrow{\text{Estim.}} \hat{\theta} \quad \Rightarrow \quad F_n \xrightarrow{\text{Rééchant.}} F_n^* \xrightarrow{\text{Réestim.}} \hat{\theta}^*$$

$$\text{Cadre sondages : } U \xrightarrow{\text{Echant.}} S \xrightarrow{\text{Estim.}} \hat{\theta}$$

Cadre général

Population finie $U = \{1, \ldots, N\}$. On utilise un plan de sondage $p(\cdot)$ respectant des probas d'inclusion $\pi_k > 0$ pour tirer un échant. aléatoire S.

On s'intéresse à un paramètre $\theta = f(t_y)$, où la fonction $f(\cdot)$ est connue mais le p-vecteur des totaux $t_y = \sum_{k \in U} y_k$ est inconnu. On l'estime par substitution

$$\hat{\theta} = f(\hat{t}_{y\pi}) \quad \text{avec} \quad \hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

Comparaison avec le cadre i.i.d. :

$$\text{Cadre i.i.d. : } F \xrightarrow{\text{Echant.}} F_n \xrightarrow{\text{Estim.}} \hat{\theta} \quad \Rightarrow \quad F_n \xrightarrow{\text{Rééchant.}} F_n^* \xrightarrow{\text{Réestim.}} \hat{\theta}^*$$

$$\text{Cadre sondages : } U \xrightarrow{\text{Echant.}} S \xrightarrow{\text{Estim.}} \hat{\theta} \quad \Rightarrow \quad ?? \xrightarrow{??} S^* \xrightarrow{\text{Réestim.}} \hat{\theta}^*.$$

Une approche pragmatique

Devant la difficulté d'étendre de façon naturelle le bootstrap au cas d'enquêtes complexes, les auteurs ont été pragmatiques.

On abandonne l'idée d'avoir une procédure bootstrap qui estime bien la distribution de $\hat{\theta} - \theta$.
On cherche des procédures qui garantissent :

$$\begin{aligned} E^*(\hat{t}^*_{y\pi}) &= \hat{t}_{y\pi}, \\ V^*(\hat{t}^*_{y\pi}) &\simeq \hat{V}(\hat{t}_{y\pi}). \end{aligned}$$

Autrement dit, on cherche à reproduire l'estimateur du total et son estimateur de variance.
On peut montrer que dans le cas d'une fonction lisse de totaux $\theta = f(t_y)$, on reproduit approximativement son estimateur de variance par linéarisation (e.g., Rao et Wu, 1988).

Une approche pragmatique (2)

La plupart des méthodes de bootstrap proposées en Sondages visent à vérifier ces deux égalités. On trouve plusieurs approches :

1. Appliquer le bootstrap i.i.d., et recaler les poids bootstrap obtenus sur $\hat{t}_{y\pi}$ et $\hat{V}(\hat{t}_{y\pi})$: méthode du rescaled bootstrap (Rao, Wu et Yue, 1992).
2. Trouver une méthode de rééchantillonnage conduisant directement à des poids bootstrap respectant ces deux équations (e.g., Gross, 1980; Sitter, 1992; Antal et Tillé, 2011).
3. Générer directement les poids bootstrap selon une distribution avec des moments appropriés (e.g., Bertail et Combris, 1997; Beaumont et Patak, 2012).

Bootstrap avec remise des unités primaires

Cas particulier de la méthode de Rao, Wu et Yue (1992), mais qui semble la plus utilisée en pratique (Rust and Rao, 1996; Yeo et al., 1999).
Supposons S sélectionné selon un plan à plusieurs degrés, avec sélection d'un échantillon S_I de n_I unités primaires (UP).

La méthode de bootstrap procéde ainsi :

1. On sélectionne un échantillon avec remise et à probabilités égales de $n_I - 1$ unités dans S_I.
2. On donne à l'UP u_i le poids bootstrap de sondage :

$$d^*_{Ii} = \frac{n_I}{n_I - 1} \times \{\text{Nb de rééchantillonnages de } u_i\} \times d_{Ii}$$

3. On reproduit la chaîne d'estimation selon les deux principes suivants :
   - Les étapes d'échant./NR post 1er degré ne sont pas bootstrappées.
   - Les étapes d'estim. (correction de la NR, calage) sont bootstrappées.

Bootstrap avec remise des unités primaires

Cette méthode donne un estimateur sans biais de la variance si les UP sont sélectionnées avec remise: cf Bessonneau et al. (2021), où nous expliquons quel est l'estimateur de variance de référence que l'on cherche à reproduire.

Quand les UP sont sélectionnées sans remise, cette méthode :

- surestime la variance du premier degré,
- estime correctement la variance due aux étapes suivantes de tirage et de non-réponse.

Enquête Histoire de Vie et Patrimoine

Enquête Histoire de Vie et Patrimoine

L'enquête a pour objectif de décrire les actifs financiers, immobiliers et professionnels des ménages français. Elle s'inscrit depuis 2010 dans un cadre européen : partie française du Household Finance and Consumption Survey (HFCS).

L'enquête a lieu tous les trois ans, avec une réinterrogation sur plusieurs vagues d'une partie des ménages (panel rotatif). Le dispositif d'enquête a été panélisé entre 2014 et 2017 : 1ère réinterrogation en 2017-18, 4ème et dernière réinterrogation en 2023 des premiers individus panélisés de 2014.

Le réseau HFC demande que l'estimation de variance puisse être faite par bootstrap. Ce travail (en cours) vise à expliquer comment le rescaled bootstrap (Rao and Wu, 1988) peut être utilisé pour des estimations transversales de l'enquête HVP. La méthode fait l'objet d'une double programmation SAS/R.

Le plan de sondage en deux mots

Pour une estimation transversale en $t+3$ (disons en 2023), l'échantillon HVP est obtenu à partir de 4 sous-échantillons sélectionnées lors de 4 vagues différentes $t, \ldots, t+3$. Chaque sous-échantillon est interrogé 4 fois.

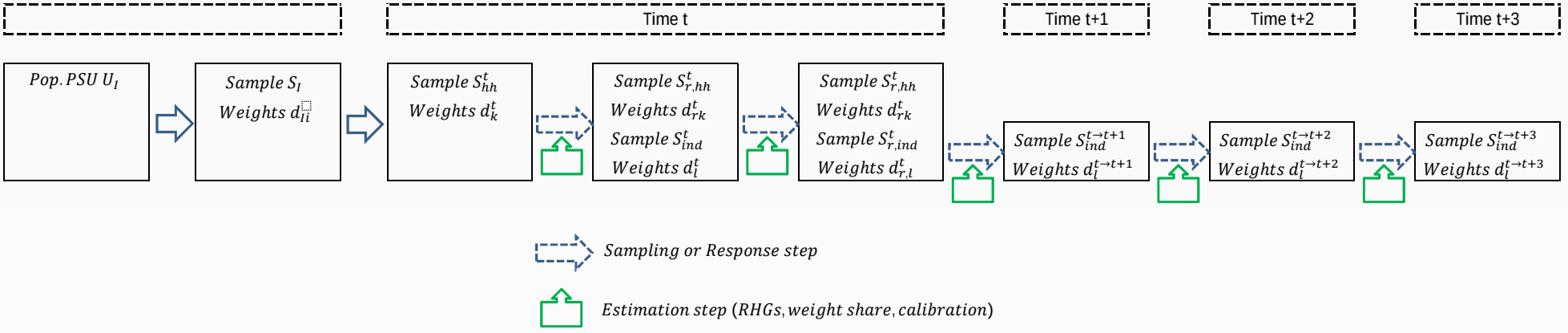
Le sous-échant. au temps t est sélectionné et enquêté ainsi :

- Un échantillon de logements est sélectionné dans l'échantillon-maître OCTOPUSSE (tirage à 2 degrés).
- Tous les ménages et les individus de ces logements sont sélectionnés au temps t.
- Les individus sont suivis et ré-enquêtés 3 fois. Aux temps $t+1$, $t+2$, $t+3$, leur ménage est enquêté. Leurs cohabitants sont également enquêtés, mais pas suivis dans le temps.

On procède de la même façon pour les sous-échantillons sélectionnés aux temps $t+1$, $t+2$ et $t+3$. Lors de l'estimation transversale en $t+3$, les ménages associés aux 4 sous-échantillons d'individus sont réunis en utilisant la méthode de partage des poids (Deville and Lavallée, 2006).

Sous-échantillon sélectionné au temps t

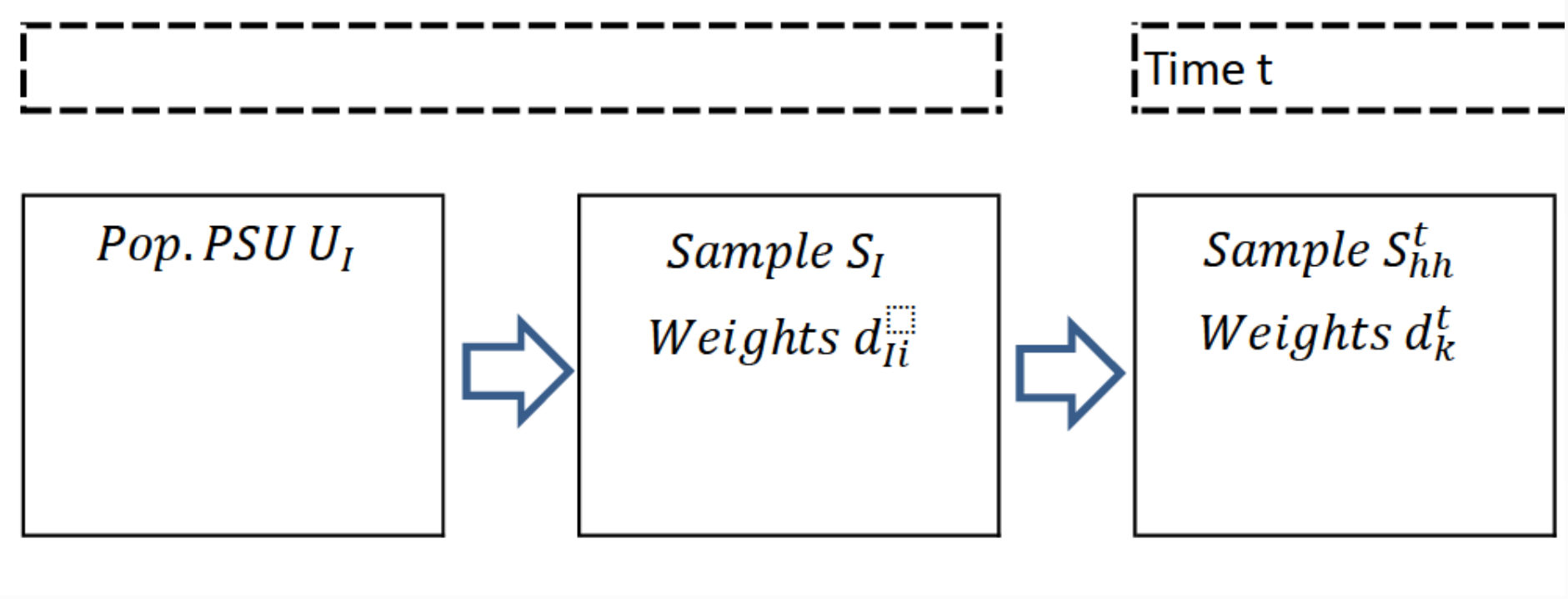
Chaîne de traitements entre t et $t+3$



Flèche bleue horizontale : une étape d'échantillonnage (traits pleins) ou de non-réponse (pointillés).
Flèche verte verticale : une étape d'estimation.
Ce sera important pour modéliser le bootstrap.

Sous-échantillon sélectionné au temps t

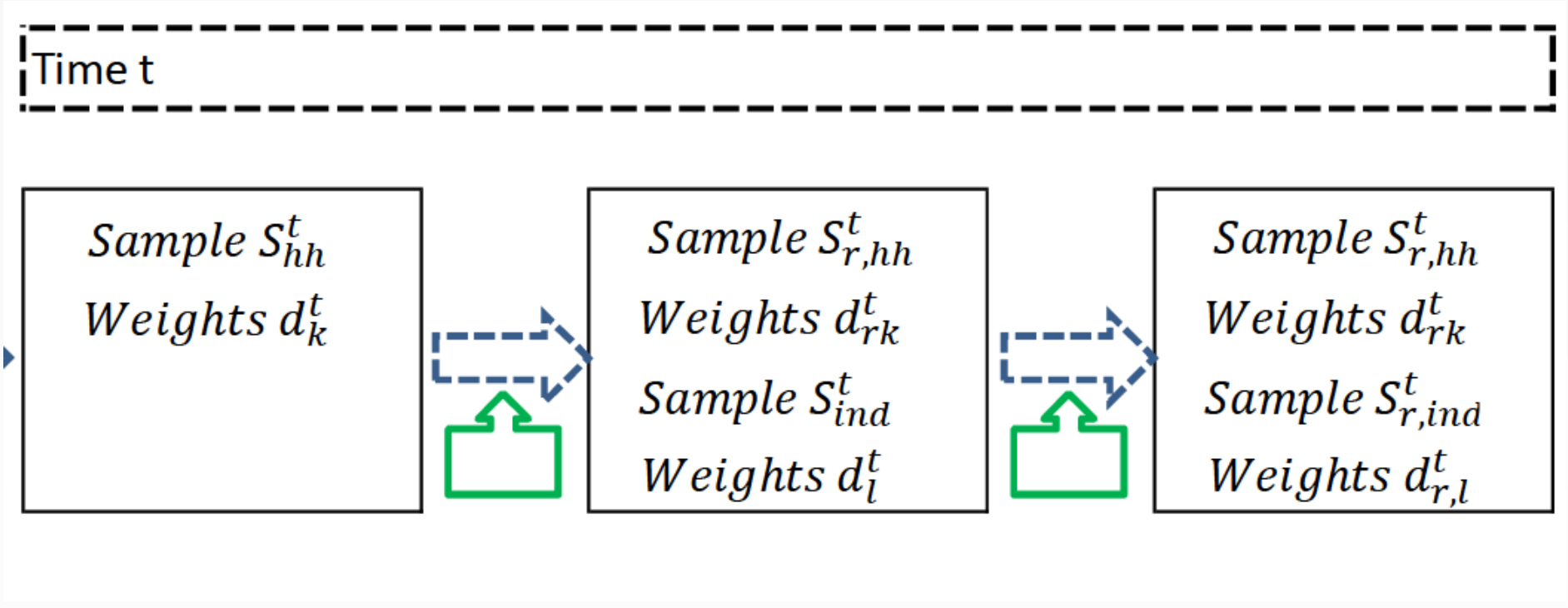
Echantillonnage au temps t



Tirage de l'échantillon d'unités primaires S_I.
Tirage du sous-échantillon de ménages S_{hh}^t (hh=household).

Sous-échantillon sélectionné au temps t

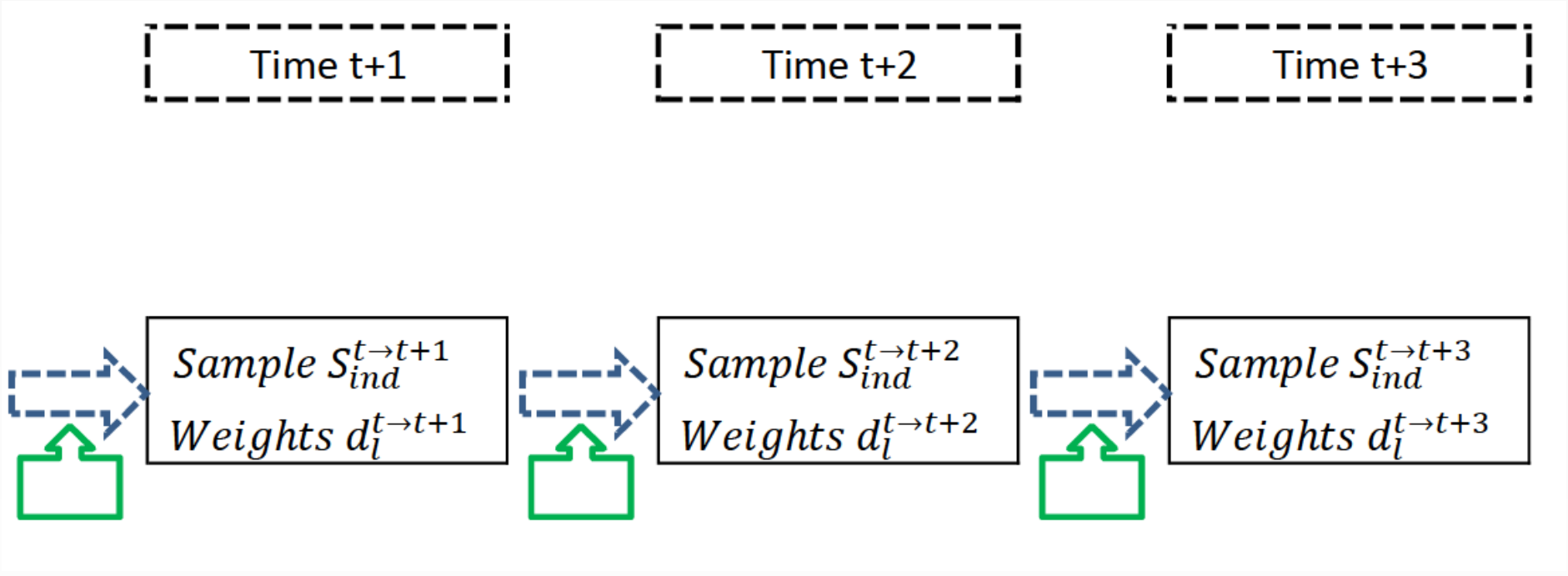
Correction de la non-réponse au temps t



Correction de la non-réponse dans l'échant. de ménages répondants $S^t_{r,hh}$.
Correction de la non-réponse dans l'échant. d'individus répondants $S^t_{r,ind}$.

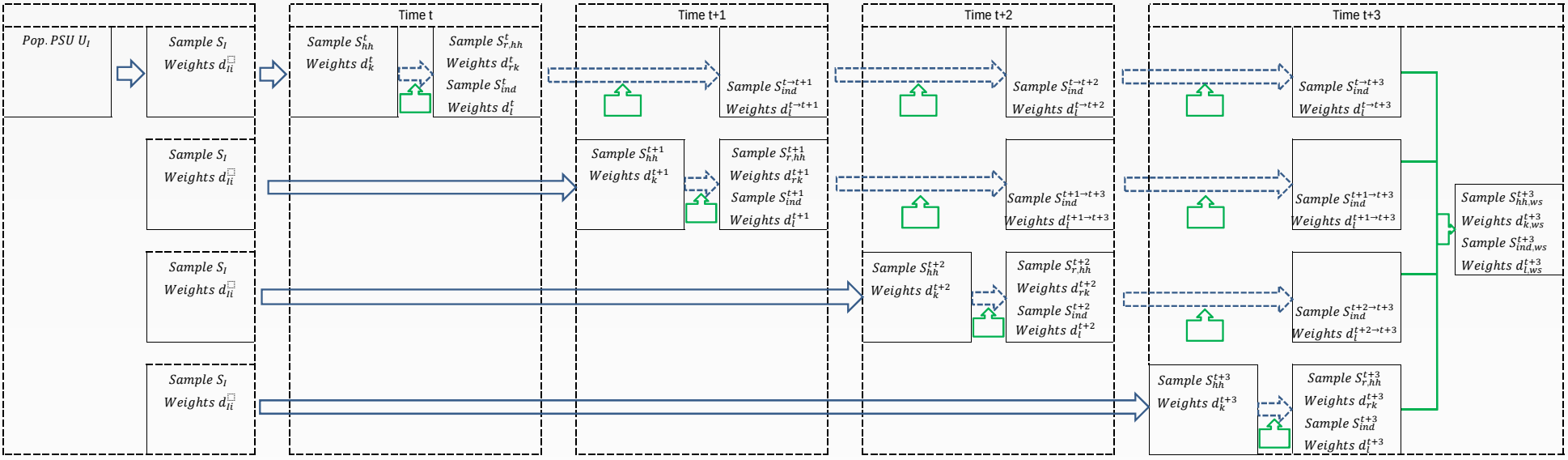
Sous-échantillon sélectionné au temps t

Suivi aux temps $t+1$ à $t+3$



Correction de l'attrition dans l'échant. d'individus répondants $S_{ind}^{t\to t+1}$.
Correction de l'attrition dans l'échant. d'individus répondants $S_{ind}^{t\to t+2}$.
Correction de l'attrition dans l'échant. d'individus répondants $S_{ind}^{t\to t+3}$.

Processus global pour une estimation transversale en $t+3$



Au temps $t+3$, on réunit quatre sous-échantillons :

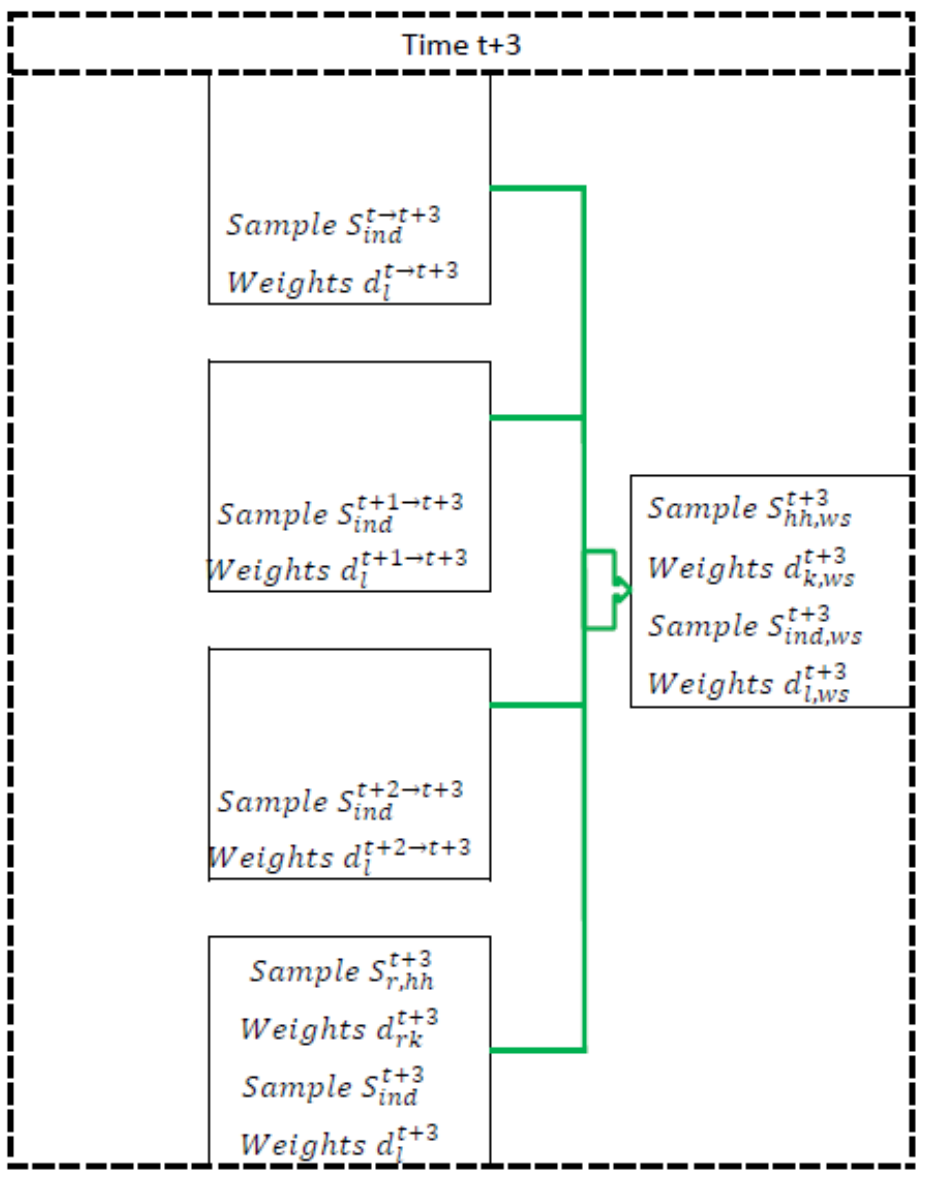
$S_{ind}^{t\to t+3}$ suivi depuis le temps t,

$S_{ind}^{t+1\to t+3}$ suivi depuis le temps $t+1$,

$S_{ind}^{t+2\to t+3}$ suivi depuis le temps $t+2$,

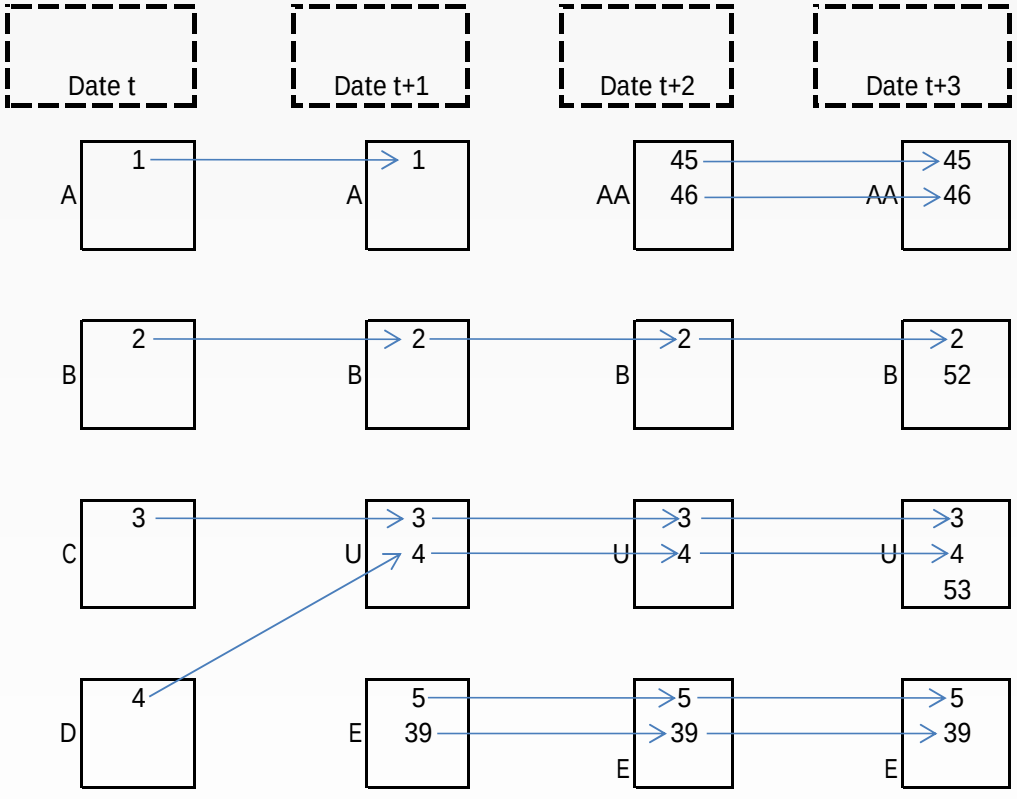
S_{ind}^{t+3} sélectionné au temps $t+3$.

Partage des poids pour une estimation transversale en $t+3$



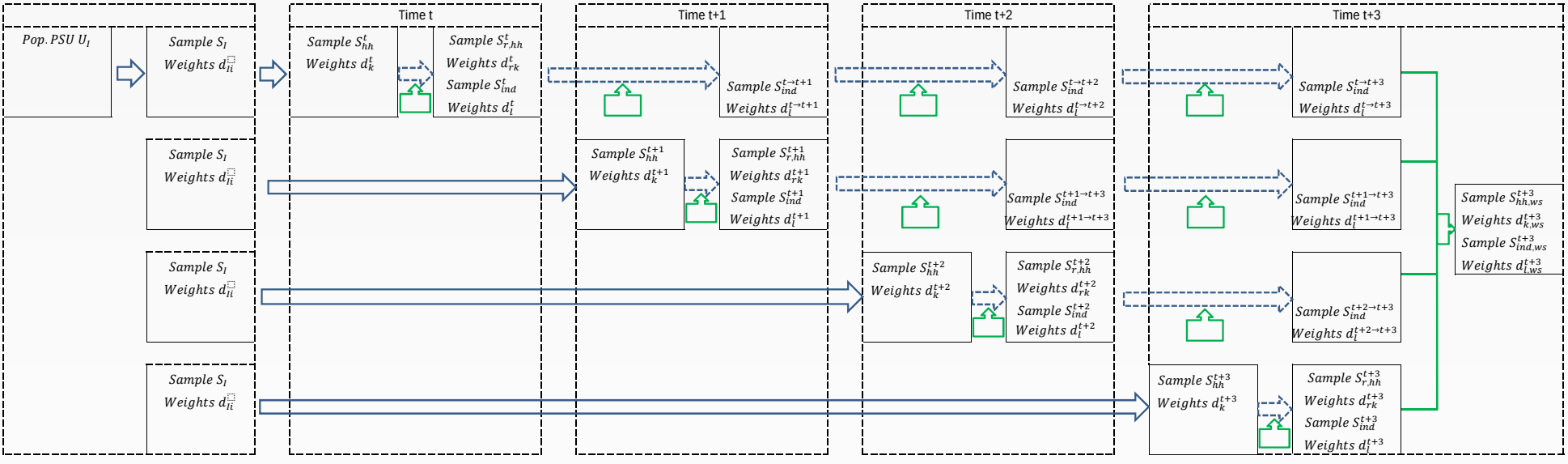
Partage des poids pour une estimation transversale en $t + 3$

Un petit exemple

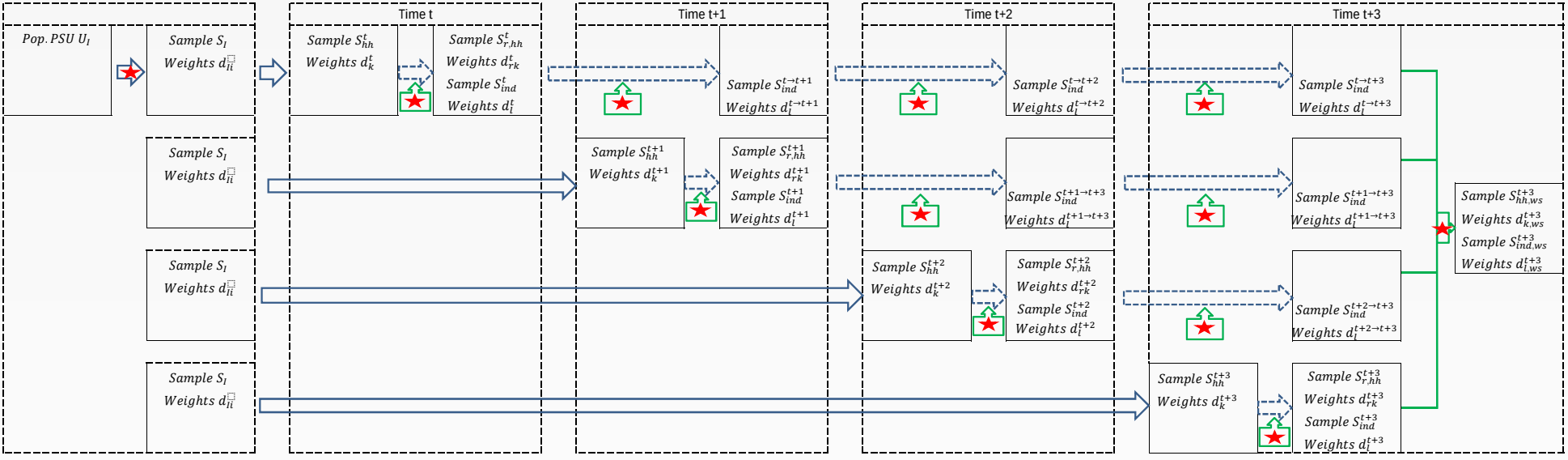


Household	Number of links between the household and the stacked pop. of individuals
	2
	2
AA	4
	4
	1
B	5
	4
	4
	1
U	9
	3
	3
E	6

Processus global pour une estimation transversale en $t+3$



Bootstrap pour une estimation transversale en $t+3$



A chaque itération :

1. On bootstrappe le tirage des unités primaires, mais pas les autres étapes d'échantillonnage/non-réponse.
2. On bootstrappe toute l'estimation $\Rightarrow$ estim. transversal bootstrappé.
3. On répète 1 et 2 un grand nb de fois $\Rightarrow$ estimation de variance.

Etude par simulations

Etude par simulations

Population U_I de $N_I = 2,000$ UP contenant (en moyenne) 40 ménages. Au temps t, pop. U_{hh}^t de 80 520 ménages contenant (en moyenne) 2 individus; pop. U_{ind}^t de 161 027 individus.

Les populations évoluent de la façon suivante. Aux temps subséquents :

- 4 % des individus sortent du champ de l'enquête,
- 5 % d'individus entrent dans le champ dans les ménages existants,
- 1 % de nouveaux ménages sont créés.

Dans la population des ménages au temps $t + 3$, deux variables auxiliaires z_1 et z_2 sont générées selon des distributions Gamma. Pour chaque UP i et ménage j, trois variables y_{ij}^m, $m = 1, \ldots, 3$ sont générées selon le modèle

$$y_{ij}^m = 5 + 2z_{1,ij} + 2z_{2,ij} + \epsilon_i + \sigma_m \epsilon_{ij}. \quad (1)$$

Etude par simulations

Plan de sondage

Un échantillon S_l de $n_l \in \{20, 50, 100, 200\}$ UP est tiré selon la méthode de Rao-Sampford. Au temps t:

- Un sous-échantillon de $n_0 = 10$ ménages est sélectionné, et tous les individus du ménage sont sélectionnés et suivis dans le temps.
- Non-réponse d'environ 20% à chaque temps (groupes de réponse homogènes)

Même principe pour les sous-échantillons sélectionnés aux temps $t+1$, $t+2$, $t+3$. Les quatre sous-échantillons sont réunis au temps $t+3$, et la méthode de partage des poids est utilisée pour produire une estimation transversale sur U_{hh}^{t+3}.

L'échantillonnage est répété $D_1 = 2,000$ fois pour obtenir une approximation Monte Carlo de la variance. Nous calculons également $D_2 = 100$ fois un estimateur de variance bootstrap avec $B = 50$ itérations bootstrap (faible, travail préliminaire).

Résultats : biais relatif de l'estimateur bootstrap de variance

Table: Biais relatif (en pourcentage) de l'estim. de variance bootstrap

	Taille de l'échantillon d'UP n_I			
Var. d'intérêt	20	50	100	200
y_1	-1.6%	-1.2%	$+6.0\%$	$+5.0\%$
y_2	$+1.9\%$	-3.6%	$+4.6\%$	$+2.8\%$
y_3	$+0.9\%$	$+4.5\%$	$+6.7\%$	$+16.3\%$

Commentaires/Travaux en cours

Commentaires/Travaux en cours

Travail en cours, qui nécessite de prendre en compte des aspects spécifiques de l'enquête.

- L'échantillon d'UP est sélectionné selon un tirage équilibré avec la méthode du Cube (Deville and Tillé, 2004).
  $\Rightarrow$ actuellement, pas d'algorithme de bootstrap adapté
  $\Rightarrow$ pris en compte via un calage des poids bootstrap des UP (en test)
- En plus de la non-réponse, problème de statut de réponse inconnu à chaque temps (*l'individu est-il ou non dans le champ de l'enquête?*).
  $\Rightarrow$ Traité comme une phase de non-réponse.

Commentaires/Travaux en cours

- Certaines UP sont tirées de façon certaine ($\pi_{Ii} = 1$).
  $\Rightarrow$ La procédure de bootstrap doit être appliquée directement sur les ménages à l'intérieur de ces UP.
  $\Rightarrow$ Nécessaire de trouver un seuil pour les UP "quasi-certaines".
- Changement d'échantillon-maître : en 2019, le vieil échantillon-maître (OCTOPUSSE) a été remplacé par NAUTILE. Les estimations transversales en 2023 mélangent donc des échantillons issus de deux EM.
  $\Rightarrow$ Procédure de bootstrap appliquée séparément sur chaque EM, puis réunion des échantillons (à tester).

C'est tout pour aujourd'hui
Merci pour votre attention :)