Estimation de variance simplifiée pour les enquêtes à plusieurs degrés

Guillaume Chauvet

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

15/03/2018

Contexte de ce travail

Premier travail : Panel Politique de la Ville

L'enquête "Panel Politique de la Ville" (PPV) a été mise en place pour étudier les conditions de vie des habitants des Zones Urbaines Sensibles (ZUS). Quatre vagues d'enquête entre 2011 et 2014 pour étudier :

- la mobilité résidentielle entre les quartiers,
- la perception des politiques publiques,
- l'impact des politiques publiques sur les bénéficiaires.

L'échantillon d'origine est sélectionné par un sondage à deux degrés (Couvert et al., 2016) :

- tirage d'un échantillon stratifié de quartiers,
- dans les quartiers, tirage d'un échantillon de ménages à partir d'une base de sondage Recensement,
- dans les ménages, tous les individus sont enquêtés.

Second travail : Household Finance and Consumption Survey (HFCS)

L'enquête HCFS est conduite de façon décentralisée, avec quelques directives données par le réseau HCF + coordination de la Banque Centrale Européenne. La méthodologie varie selon le pays : de 1 à 3 degrés de tirage.

L'enquête fournit des données détaillées au niveau des ménages sur différents éléments de leur bilan ainsi que sur des variables économiques et démographiques connexes (revenus, pensions de retraite, emploi, mesures de la consommation).

La priorité de l'enquête HCFS est de produire des estimations transversales (à un temps donné).
Le réseau HCF recommande également l'introduction d'une composante panel pour mesure les changements individuels dans le temps.

Des caractéristiques et des défis communs

Au temps initial t, l'échantillon est tiré selon un plan à plusieurs degrés. Aux temps suivants, ajout d'un échantillon de rafraîchissement :

- suivi d'une partie des unités (individus ou ménages) enquêtés au temps précédent,
- sélection d'un échantillon de rafraîchissement de ménages et d'individus pour représenter les naissances.

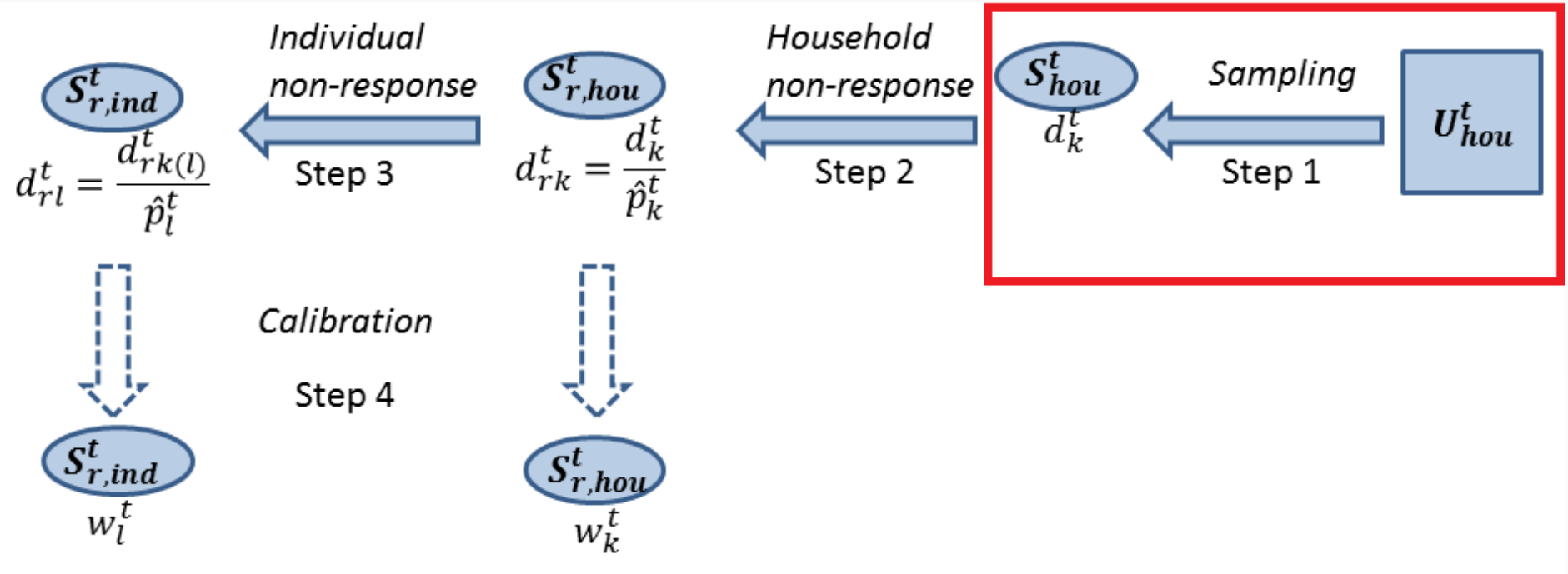
But de mon travail :

- pondération pour des estimations transversales (Sala et Chauvet, 2018),
- pondération pour des estimations longitudinales,
- Estimation de variance associée :
  - par linéarisation pour l'enquête PPV,
  - par bootstrap pour l'enquête HCFS.

Dans cette présentation : estimations transversales au temps t + bootstrap.

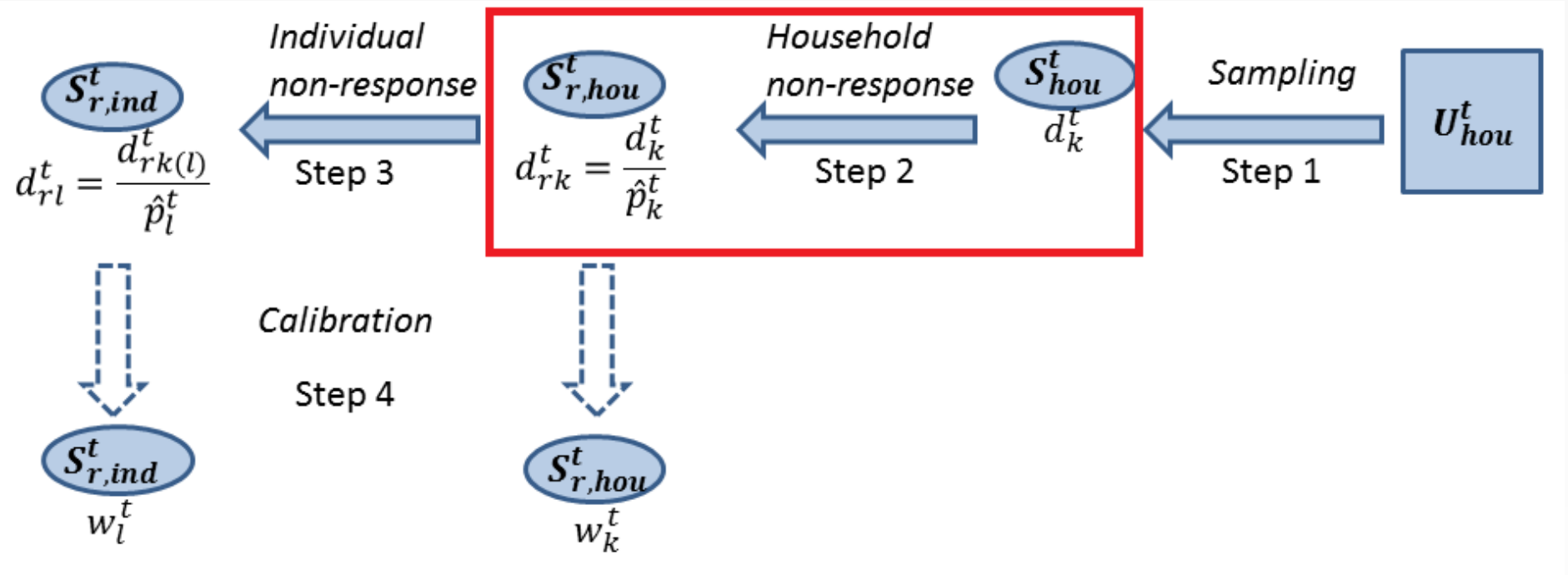
Estimation au temps initial

Estimation transversale au temps initial



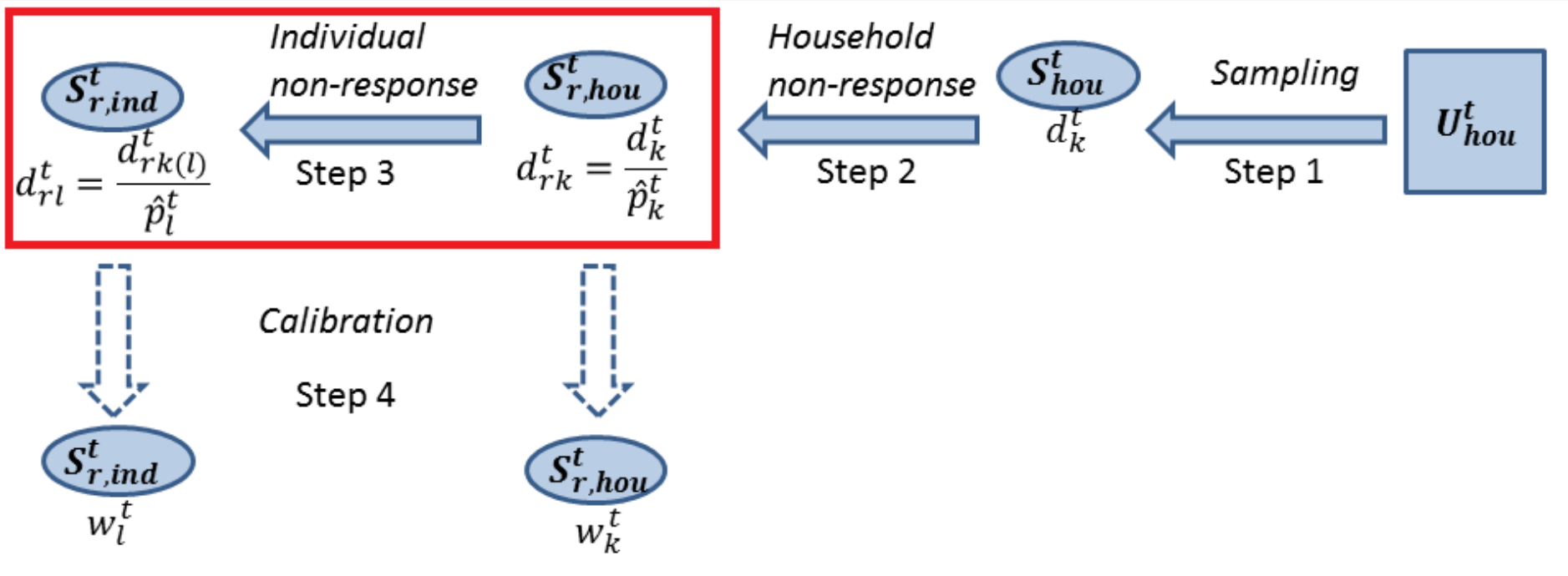
Echantillon de ménages tiré par sondage à plusieurs degrés : n_I Unités Primaires (e.g., communes) et sous-échantillonnage de ménages.
On obtient un échantillon de ménages S^t_{hou} avec des poids de sondage d^t_k.

Estimation transversale au temps initial



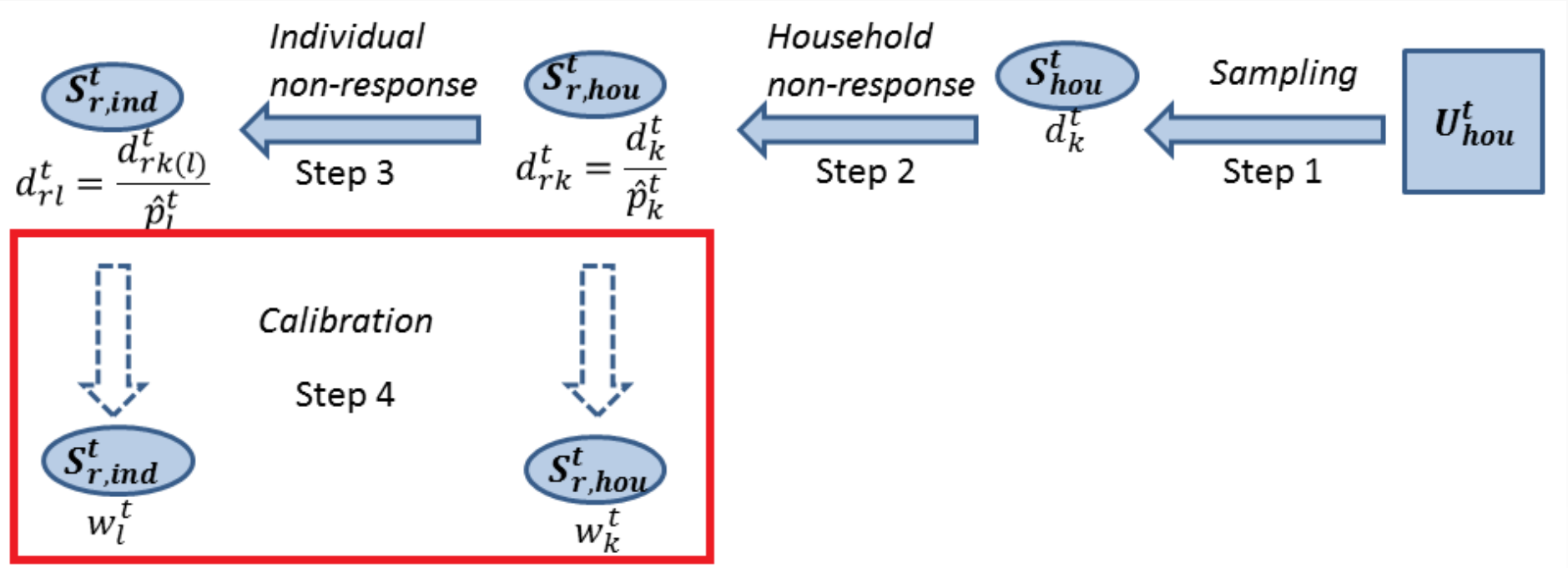
On corrige de la non-réponse des ménages, par exemple selon la méthode des Groupes Homogènes de Réponse (GHR).
Estimation de la probabilité de réponse $\hat{p}_k^t \Rightarrow$ poids ménage d_{rk}^t.

Estimation transversale au temps initial



On corrige de la non-réponse individuelle, par exemple selon la méthode des Groupes Homogènes de Réponse (GHR).
Estimation de la probabilité de réponse $\hat{p}_l^t \Rightarrow$ poids individuels d_{rl}^t.

Estimation transversale au temps initial



Les poids ajustés de la non-réponse sont calés sur de l'information auxiliaire au niveau ménage et individu.
Conduit à des poids calés w_k^t pour les ménages et w_l^t pour les individus.

Estimation de variance Bootstrap

Calcul des poids Bootstrap

Dans le cas de la Household Finance and Consumption Survey (HFCS), nécessité de produire un estimateur de variance bootstrap.

Bootstrap = technique computationnelle permettant de produire des estimations de variance en répliquant de façon répétée le processus d'échantillonnage + les étapes d'estimation.

Bootstrap proposé = bootstrap des Unités Primaires (UP) :

- ne nécessite pas de bootstrapper l'échantillonnage dans les UP
  $\Rightarrow$ très simple à implémenter,
- surestimation de la variance du premier degré
  $\Rightarrow$ approche conservatrice,
- la variance des degrés suivants (e.g., non-réponse) est correctement prise en compte.

Calcul des poids bootstrap

Les poids bootstrap sont obtenus de la façon suivante :

- Sélection d'un échantillon avec remise de $n_I - 1$ UP dans l'échantillon d'origine, avec des **probabilités égales** de tirage (**Bootstrap des UP**). Poids bootstrap d'une UP u_i :

$$W_{Ii} = \frac{n_I}{n_I - 1} \times \text{Nombre de sélections de } u_i.$$

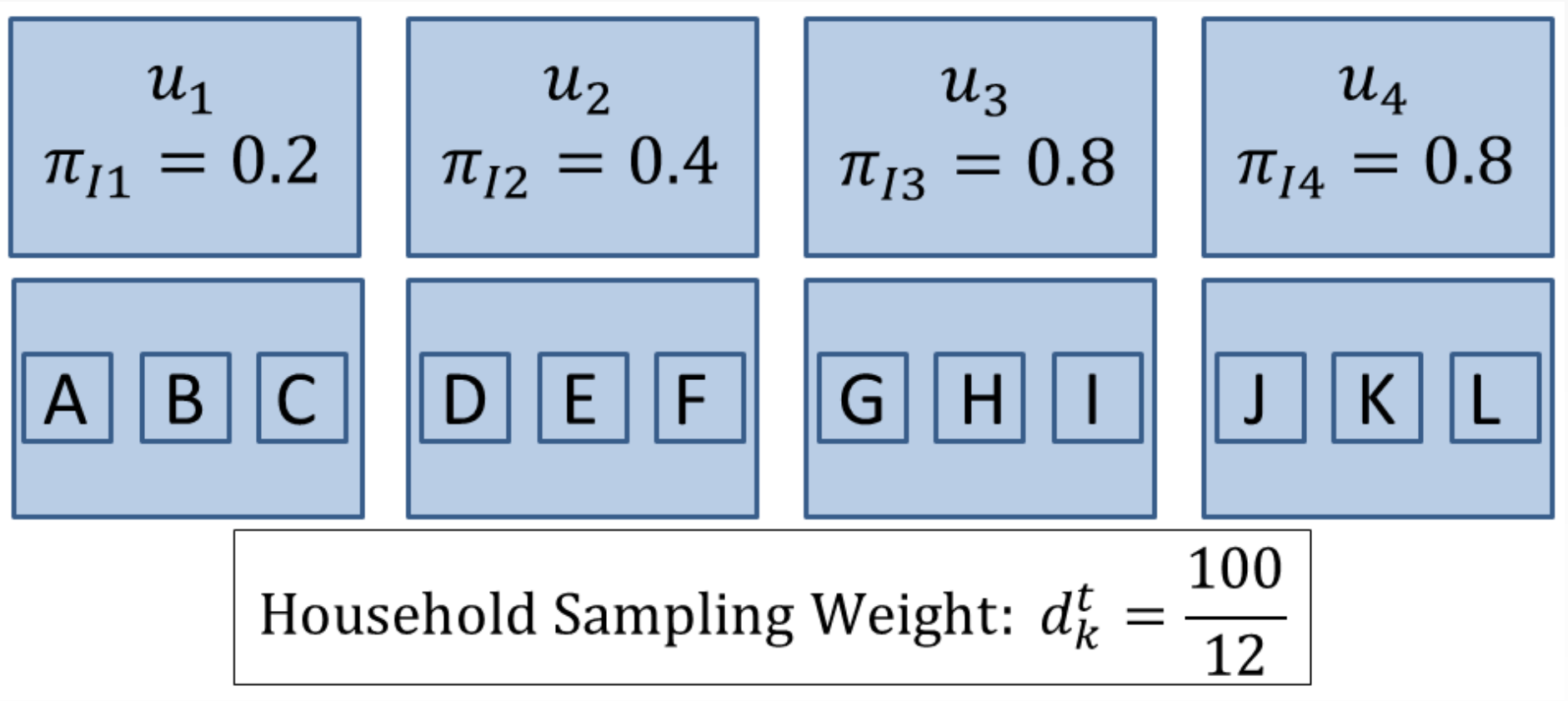
- Poids bootstrap d'échantillonnage d'un ménage $k \in u_i$:

$$d_{k*} = d_k \times W_{Ii}.$$

- Ces poids bootstrap sont ensuite ajustés pour les étapes suivantes d'estimation :
  - Non-réponse au niveau ménage,
  - (Non-réponse au niveau individuel),
  - Calage.

Exemple avec un plan à deux degrés

Sélection des ménages

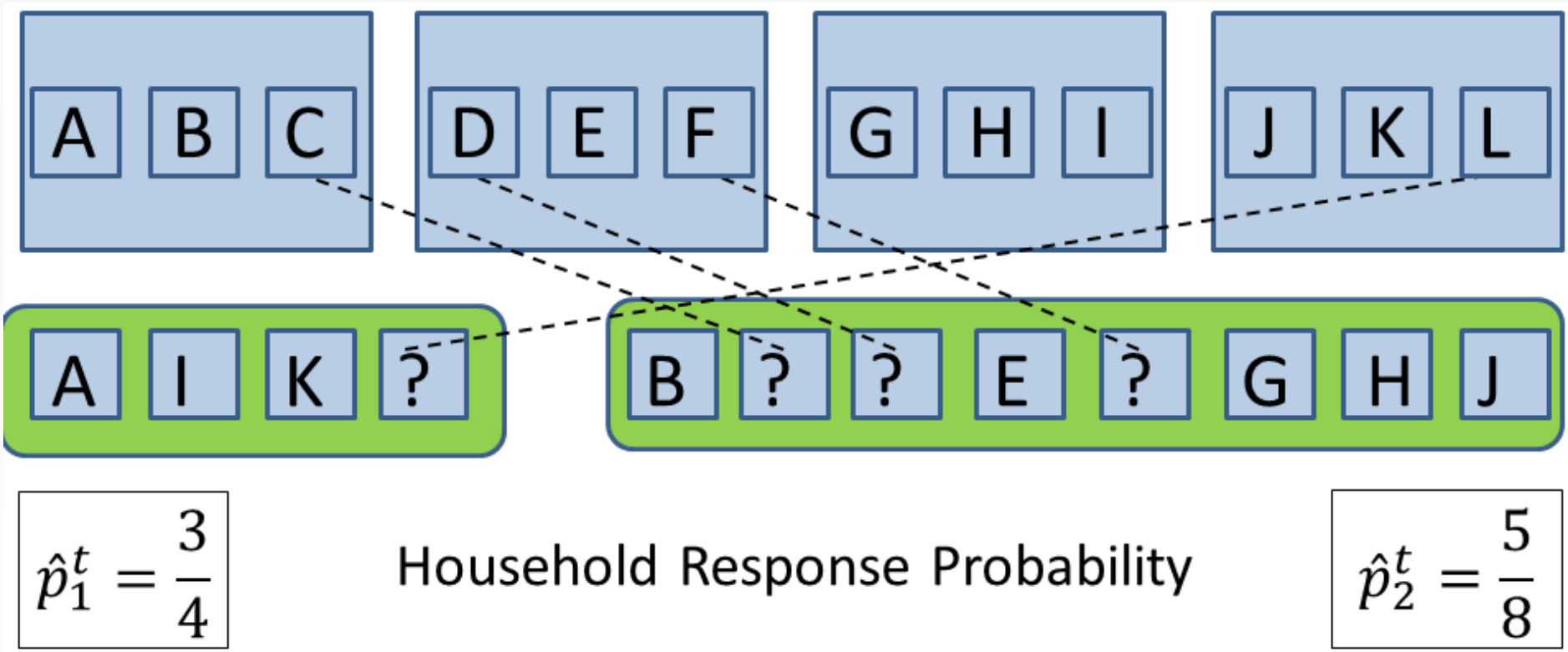


Echantillon de $n_I = 4$ UP et de $n = 12$ ménages sélectionné selon un plan autopondéré.

On obtient des poids de sondage d_k^t égaux pour les ménages.

Exemple avec un plan à deux degrés

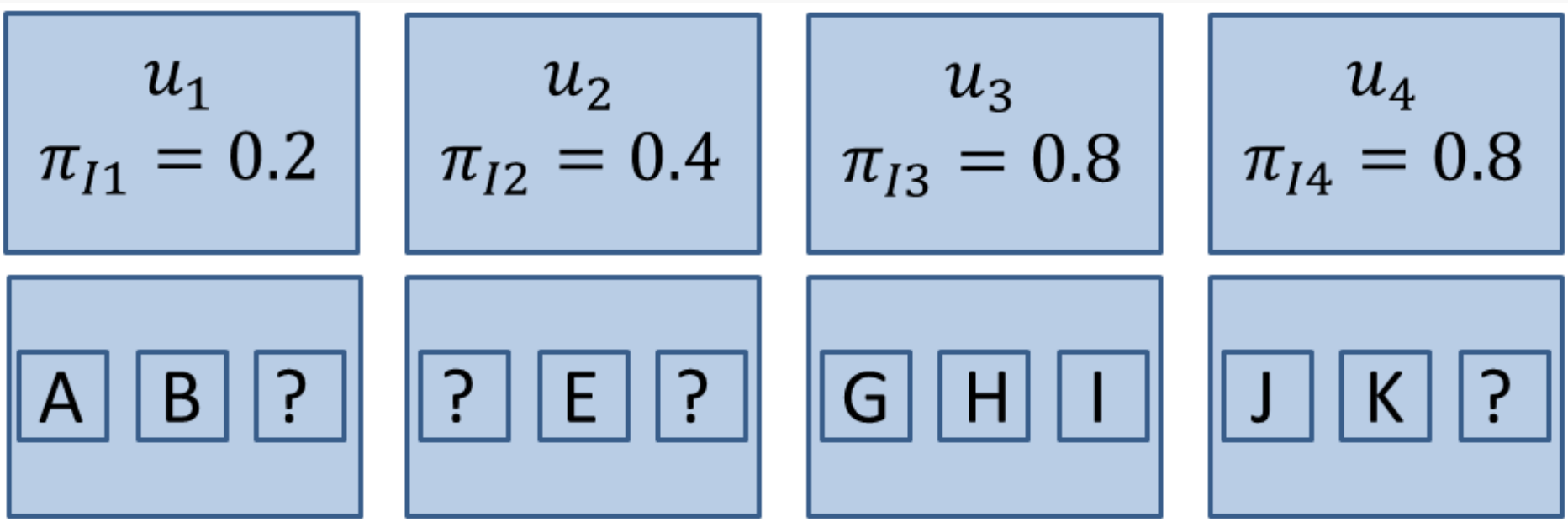
Correction de la non-réponse de ménage



Non-réponse corrigée à l'aide de 2 Groupes Homogènes de Réponse.
Probas de réponse estimées par la fréquence de réponse dans chaque GHR.
Les poids de sondage sont corrigés de la non-réponse.

Exemple avec un plan à deux degrés

Calage

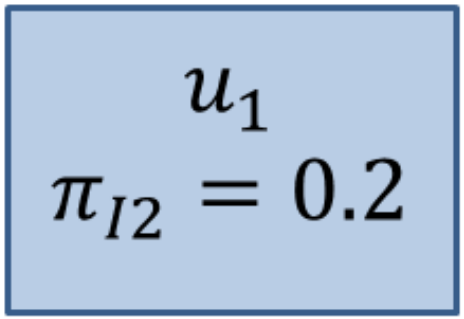


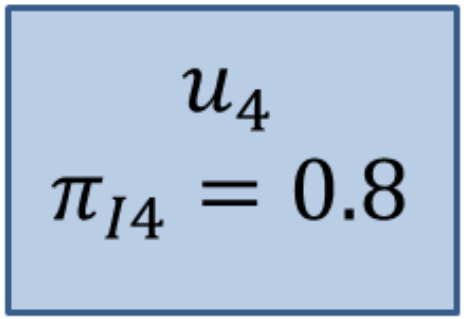
$$\text{Household Adjusted weights } d_{rk}^t = \frac{d_k^t}{\hat{p}_k^t} \Rightarrow \text{Calibrated weights } w_k^t$$

Les poids sont finalement calés sur les totaux $t_x = \sum_{k \in U} x_k$.

Exemple avec un plan à deux degrés

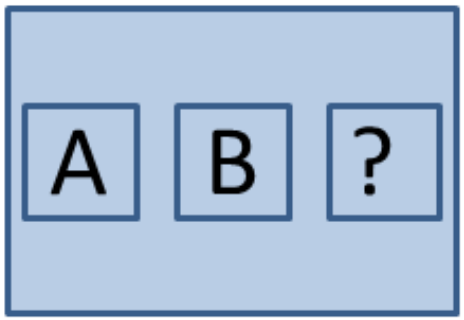
Bootstrap des UP

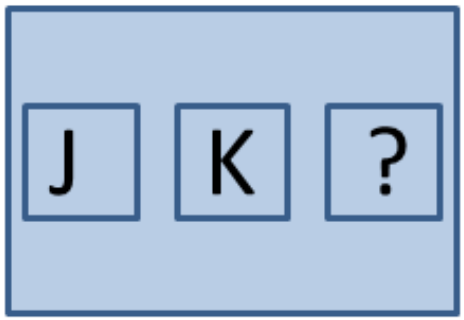




$$W_{I1} = \frac{4}{3}$$

$$W_{I4} = \frac{8}{3}$$



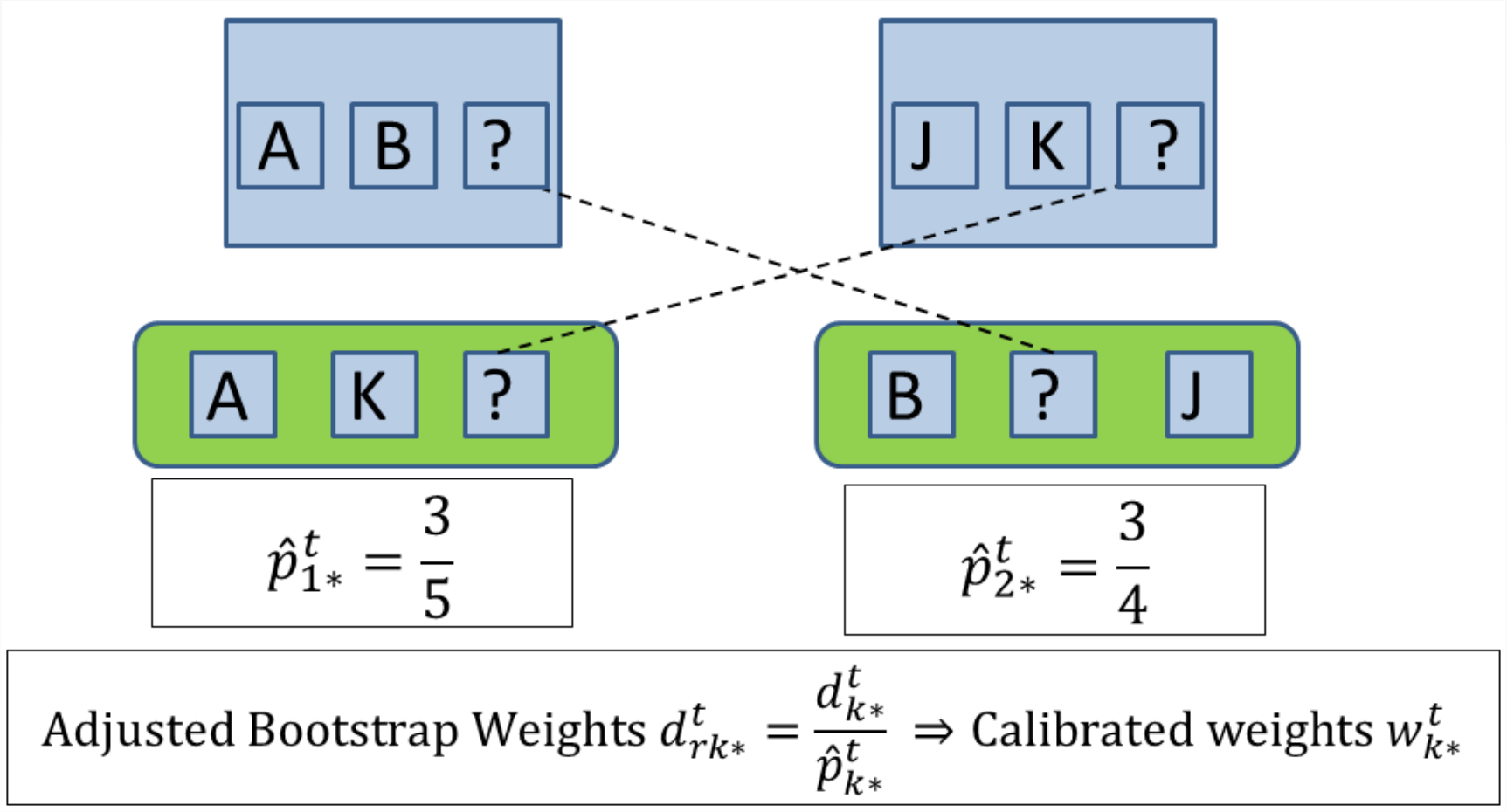


Bootstrap Household Sampling Weights $d^t_{k*} = W_{Ii} \times d^t_k$

Un rééchantillon de $n_I - 1 = 3$ UP est tiré à probabilités égales.

Exemple avec un plan à deux degrés

Bootstrap de la non-réponse et du calage



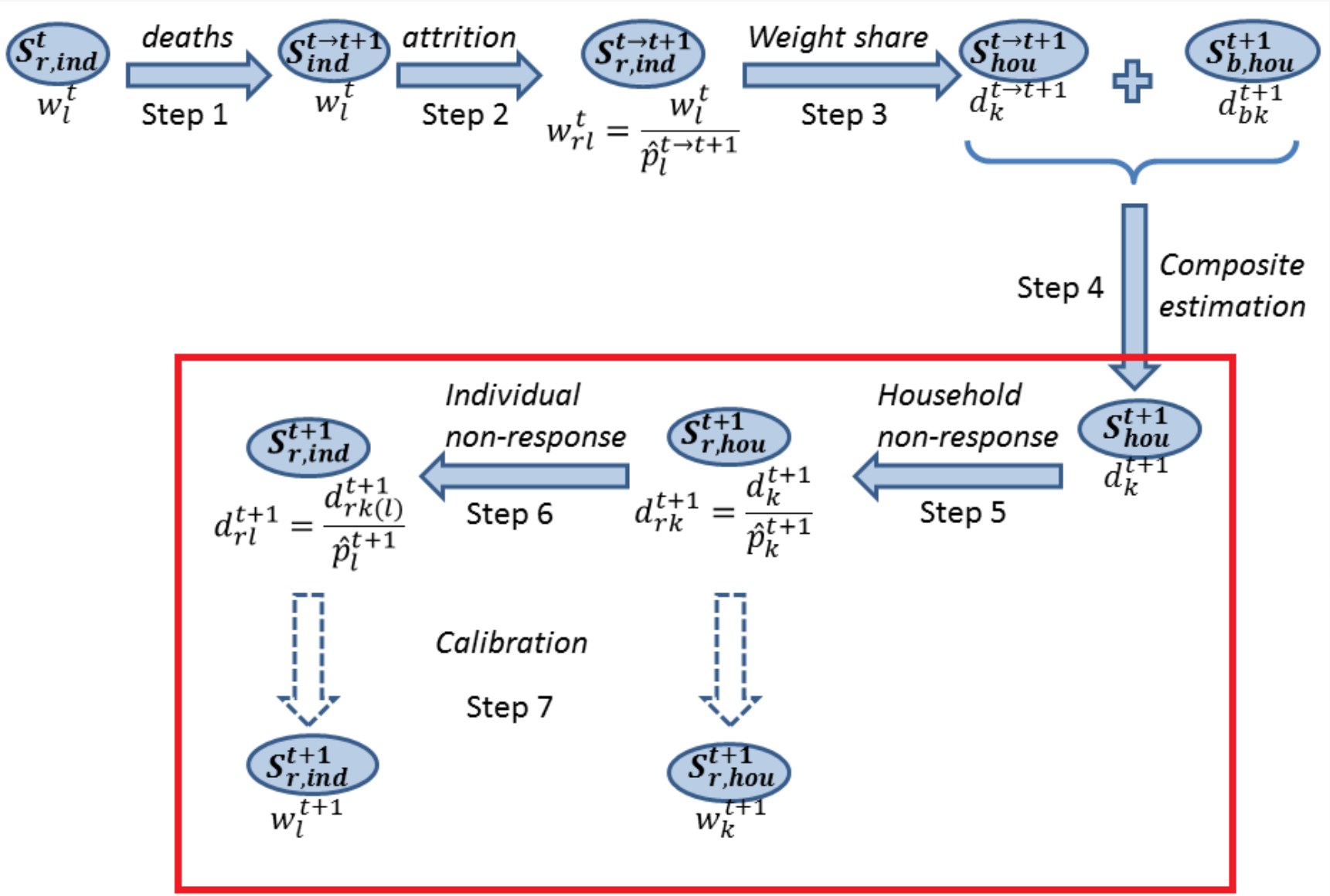
Poids de sondage bootstrap corrigés de la non-réponse comme initialement.
Les poids sont finalement calés sur les totaux $\hat{t}_{xr} = \sum_{k \in S^t_{r,hou}} d_{rk} x_k$.

Travail en cours

Cadre théorique

- Avoir une asymptotique solide pour le tirage à deux degrés (Chauvet et Vallée, 2018).
- Avoir une asymptotique solide pour les estimateurs par substitution (en cours).
- Consistance des estimateurs de variance par linéarisation/bootstrap pour les plans à plusieurs degrés.

Estimation transversale avec une composante panel



Une approche similaire est possible (un peu plus difficile).

Annonce : SMURF Workshop

"Survey Methods and their Use in Related Fields" : lien entre la théorie des Sondages et le reste de la statistique.

Thèmes : big data, fusion de bases de données, méthodes MCMC, processus déterminantaux, statistique spatiale.

Neuchâtel (Suisse), du 20 au 22/08/2018.

SMURF
Survey Methods and their Use in Related Fields

Annonce : Colloque Francophone sur les Sondages

Ouverture : J-F. Beaumont (Statistique Canada) : Les enquêtes probabilistes sont-elles vouées à disparaître pour la production de statistiques officielles ?
Clôture : J. Chiche (CEVIPOF) : Enquêtes et estimations le jour du vote.

Conférence Waksberg 2018 : Jean-Claude Deville.

Lyon, du 24 au 26/19/2018.