# Comment pondérer une enquête auprès des personnes sans domicile ?

Pascal Ardilly, David le Blanc

## 1 Introduction

L'INSEE réalisera en 2001 une enquête auprès des personnes fréquentant des services destinés aux personnes sans domicile. Compte tenu de l'absence de base de sondage permettant d'atteindre directement les personnes sans domicile, le principe de l'enquête est d'échantillonner des prestations et d'interroger les individus qui fréquentent ces prestations. Evidemment, une personne peut fréquenter une ou plusieurs prestations de la base de sondage pendant la période de référence considérée. Pour pouvoir estimer des paramètres d'intérêts relatifs à la population visée par ces prestations, il faut passer d'un jeu de poids des prestations échantillonnées à un jeu de poids des individus qui ont fréquenté ces prestations. Le but de cet article est de présenter le plus rigoureusement possible une méthode pour faire ce calcul. Le plan est le suivant : on rappelle d'abord le principe de l'enquête et son plan de sondage. On définit ensuite la population de référence, les paramètres d'intérêt, après quoi on dérive les estimateurs de ces paramètres. Certaines hypothèses sur les paramètres d'intérêt permettent de se placer dans le contexte de la méthode du partage des poids. Le principe de celle-ci et son application dans ce cadre sont brièvement exposés. On insiste particulièrement sur les implications en termes de données à collecter, et sur la mise en oeuvre pratique de calculs de poids " journaliers ", permettant d'estimer sans biais le nombre de personnes qui fréquentent un des services un " jour moyen ", et de poids " hebdomadaires ", permettant d'estimer sans biais le nombre de personnes qui fréquentent un des services une " semaine

moyenne ". Enfin, on donne des considérations pratiques sur la correction de la non-réponse.

# 2 L'enquête " sans domicile ".

## 2.1 définition des termes spécifiques utilisés

Le principe de l'enquête consiste à échantillonner dans une population de prestations destinées à des populations en difficulté, correspondant à plusieurs types de services : hébergement, repas, etc. Ces prestations sont fournies sur des bases temporelles qui varient selon leur nature : les repas sont fournis chaque jour midi et soir, les nuitées une fois par jour, les services d'accueil de jour pendant certaines plages horaires de la journée, etc.

La caractéristique principale des services considérés est qu'ils sont fournis dans un lieu précis ; ce lieu est appelé par la suite *centre*. A un centre donné correspond un ou plusieurs types de services.

L'unité statistique échantillonnée, que nous appellerons par la suite *prestation*, sera définie comme un triplet (service, jour, horaire), où on entend par service, un service de type donné dans un centre donné, et par horaire, un intervalle de temps (à préciser) un jour donné. Une personne peut bien sûr être bénéficiaire de plusieurs prestations la même journée, et *a fortiori* une semaine donnée ou pendant le mois d'enquête.

Les personnes fréquentant les services sont donc atteintes par échantillonnage indirect. Pour cette raison, la méthode du partage des poids est un candidat " naturel " pour le calcul de pondérations individuelles dans cette enquête, sous des hypothèses qu'il faudra préciser.

## 2.2 le plan de sondage de l'enquête

Le premier degré du plan de sondage consiste à tirer des agglomérations, proportionnellement à un critère de taille défini comme une combinaison de la population des agglomérations et des capacités d'accueil (incomplètes) telles qu'elles ont pu être recensées dans les fichiers des associations et les fichiers du Ministère de la Santé. Ce premier degré de tirage est effectué plusieurs mois avant les autres. Ce décalage s'impose car le recensement exhaustif des centres et des informations les concernant (capacité moyenne, jours d'ou-

verture,...) est entrepris sur les agglomérations tirées. Cette opération est réalisée en deux fois : une enquête lourde l'année précédant la collecte, et une mise à jour l'année suivante juste avant le début de la collecte. On obtient ainsi une base de sondage de centres ; pour chaque centre, on connaît sa capacité moyenne, le type de service rendu, et d'autres informations. Ces données ne suffisent pas en général à constituer une base de sondage de prestations. Ce sont néanmoins ces prestations qui doivent être échantillonnées.

### 2.2.1 Tirage des centres et des jours

Pour des raisons pratiques, il n'est pas possible d'enquêter l'ensemble des centres et de maintenir sur le terrain, dans un centre donné, un enquêteur durant une journée entière. Enfin, on ne peut interroger toutes les personnes dans un centre. Il est donc incontournable d'échantillonner :

- des centres dans les agglomérations tirées
- des intervalles de temps pendant la période de collecte : on les repérera par l'indice $t$, qui désigne dans toute cette section le croisement d'un jour et d'une période de temps pendant ce jour.
- des prestations au sein d'un (centre,intervalle de temps) tiré.

Pour des raisons pratiques toujours, les intervalles de temps sont à calibrer de façon qu'un individu ne puisse pas fréquenter deux prestations différentes durant cet intervalle de temps. En effet, la mesure des liens avec la base de sondage ne peut raisonnablement s'effectuer qu'en permettant aux personnes interrogées de repérer facilement dans le temps et l'espace les prestations qui lui ont été servies au cours de la période d'enquête. Pour les centres offrant des repas par exemple, un intervalle de temps recouvrira les repas du midi et un intervalle les repas du soir : on considère qu'une personne ne peut fréquenter qu'un seul centre durant l'intervalle de temps correspondant au repas de midi, faute de quoi il faudrait lui demander si elle n'a pas déjà pris un repas ailleurs, ou si elle ne mange pas deux fois dans le même centre.

Il se trouve par ailleurs que la largeur d'un intervalle assurant une telle propriété correspond à la durée au cours de laquelle on peut raisonnablement demander à un enquêteur d'interroger sur place (soit 3 à 4 heures au maximum).

On échantillonne donc des intervalles de temps. En fait, il n'y a pas de différence fondamentale entre l'échantillonnage des centres et l'échantillonnage des périodes de temps : les unités pertinentes à considérer sont les couples

$(c, t)$ correspondant au croisement d'un centre et d'un intervalle de temps. Certaines cases du tableau croisant " temps " et " centres " seront éliminées *a priori* avant le tirage, soit parce que le centre est fermé durant le créneau horaire considéré, soit parce que la fréquentation y est manifestement très faible (dans ce dernier cas, il faut prendre garde à l'éventuelle restriction du champ couvert, s'il s'avérait que des personnes ne fréquentent que ce centre et ne sont présentes que dans ce créneau horaire).

Des calculs présentés dans Ardilly et le Blanc (1999) montrent que pour minimiser la variance, on doit échantillonner un couple (centre, intervalle de temps) avec une probabilité d'autant plus forte que :

- sa capacité d'accueil est grande,
- les liens des individus qui fréquentent le centre pendant cet intervalle de temps avec les prestations du champ de l'enquête pendant la période d'enquête sont ténus.

En pratique, il semble difficile d'aller au-delà, et même de se servir concrètement de ces considérations, si on n'a pas une idée assez précise des comportements des individus fréquentant un type de centre en matière d'utilisation des services[1]. Dans l'hypothèse où l'on n'a pas ce genre d'informations, postuler l'homogénéité des populations qui s'adressent aux différents centres en terme de fréquence de passage dans les centres du champ de l'enquête conduit à retenir des probabilités proportionnelles à la capacité, c'est-à-dire aux nombre de prestations fournies durant l'intervalle de temps $t$ dans le centre $c$. En présence de variations " saisonnières " (selon le jour de la semaine d'une part, selon l'heure de la journée d'autre part) du nombre de prestations servies, l'idéal serait d'avoir pour chaque centre des profils de fréquentation heure par heure s'étendant sur une semaine. La probabilité d'inclusion du centre $c$ durant l'intervalle de temps $t$ serait alors prise proportionnelle à la fréquentation durant cet intervalle de temps.

En réalité, ces profils ne sont pas disponibles. Une solution de repli consiste à postuler une relation du type, modélisant la fréquentation par un " effet centre " (proportionnel à la taille " moyenne " du centre) et un " effet intervalle de temps ". On peut alors tirer indépendamment les intervalles de temps et les centres. Un tirage " réaliste " consiste à tirer au hasard des couples (centres, intervalles de temps) proportionnellement à la taille moyenne des centres. On peut stratifier par type de centre ; toutefois, la stratification,

[1] Obtenir des informations à ce sujet sera un des buts des tests de l'enquête.

ne portant pas directement sur les unités d'observation, n'a d'intérêt que si le comportement des personnes diffère sensiblement selon le type de centre où on les trouve.

### 2.2.2 Tirage des prestations

Ce dernier degré de tirage consiste à tirer des prestations dans un centre sélectionné dans un intervalle de temps donné. Dans certains centres comme les centres d'hébergement, il peut exister des listes ; c'est le cas le plus favorable, un tirage des prestations pouvant être conduit à partir de ces listes. En revanche, dans un point-soupe, on ne connaît pas a priori le nombre de personnes qui vont se présenter durant un intervalle de temps donné : on ne peut donc pas faire de base de sondage des prestations. L'échantillonnage des prestations s'effectue *a priori* à probabilités égales. Comme traditionnellement dans les sondages à plusieurs degrés, tirer un nombre constant de prestations (dernier degré) permet d'assurer des probabilités de tirage constantes, et donc de limiter les risques d'explosion de variance.

En pratique, la méthode de tirage retenue peut varier d'un type de centre à un autre, selon la topographie des lieux : liste existante, file d'attente, arrivées espacées dans le temps, population " groupée " sans ordre dans un même lieu au même moment, etc. Elle doit aussi tenir compte du nombre maximal d'interviews raisonnablement assurables par le ou les enquêteurs pendant l'intervalle de temps de l'enquête, et du fait qu'il n'est pas souhaitable de " retenir " des personnes échantillonnées trop longtemps après la fermeture d'un centre où l'arrêt de distribution de repas, sous peine d'augmenter la non-réponse.

On peut toutefois proposer une méthode assez générale :

- un " dénombreur " compte pendant la période d'échantillonnage le nombre N de prestations servies. Ce rôle est essentiel pour déterminer la probabilité de tirage des prestations échantillonnées.
- un " superviseur " procède à un tirage de type systématique.

Deux variantes de tirage sont possibles :

1) on tire $n$ prestations, $n$ étant fixé avant l'enquête. Cette méthode est naturelle dans tous les endroits où une liste est disponible. Mais elle peut être appliquée dans d'autres types de centres, pour des raisons pratiques (nombre d'interviews par enquêteur limité).

2) on tire les prestations avec un taux de sondage $f$ fixé. $f$ est déterminé

selon le nombre de prestations attendues $\widetilde{N}$ et le nombre de prestations que l'on désirerait échantillonner $n$. Dans ce cas, la taille de l'échantillon est inconnue *a priori*, car l'effectif $N$ n'est pas connu avant la fin de la période d'échantillonnage. C'est le cas dans les points soupe et les centres où la fréquentation suit un processus de type " file d'attente ".

Nous pouvons maintenant préciser la population et les paramètres d'intérêt.

# 3 la population de référence et les paramètres d'intérêt

La période de référence de l'enquête s'étend sur un mois (de manière indicative du 15 janvier au 15 février 2001). Le champ géographique de l'enquête est celui des agglomérations de plus de 20 000 habitants.

Les prestations dans le champ de l'enquête sont celles qui relèvent d'un des trois types de services retenus : repas, hébergement, autres services, rendus pendant la période de l'enquête dans le champ géographique retenu.

La population d'intérêt est constituée des personnes qui ont fréquenté au moins une prestation du champ de l'enquête pendant la période de référence. La base de sondage des centres, constituée par l'ensemble des centres repérés avant la collecte, joue donc un rôle fondamental : des personnes qui ne fréquenteraient qu'un seul centre non recensé seraient de fait non échantillonnables.

Pour la commodité de l'exposé, nous ne faisons pas apparaître explicitement tous les degrés de tirage. Le lecteur trouvera un exposé complet des formules dans Ardilly et le Blanc (1999). Nous nous plaçons au niveau d'une agglomération échantillonnée au premier degré du tirage.

On note :

$J$ : l'ensemble des jours d'enquête, repérés par l'indice $j$

$C_N$ : ensemble des centres de l'agglomération ouverts le jour $j$, repérés par l'indice $n$

$P_c$ : ensemble des prestations servies dans le centre $c$, repérés par l'indice $i$.

On définit une unité de temps comme un couple $(j, t)$, où $j$ désigne un jour de la période de collecte et $t$ désigne un intervalle de temps (par exemple, de 13 à 15 heures). Dans la pratique, cet intervalle de temps est défini de telle manière qu'une personne ne puisse pas fréquenter deux prestations différentes

durant cet intervalle de temps.

Repérons par l'indice $i, i \in \{1, ...P\}$, les prestations servies dans le centre $C_n$ le jour $j$ pendant l'intervalle de temps $t$. Il ressort de la définition de $t$ qu'à chaque individu $k$ appartenant à la population des personnes qui ont fréquenté un des centres de la base pendant l'intervalle de temps considéré, correspond une et une seule prestation $i$. Ainsi, pour un triplet $(C_n, j, t)$ donné, on a une correspondance biunivoque entre une prestation et un individu. Si on note alors $P_{n,j,t}$ l'ensemble des personnes se présentant au cours de l'unité de temps $(j, t)$ dans le centre $C_n$ , la population d'intérêt est définie par

$P(J) = \underset{n,j,t}{\cup} P_{n,j,t}$.

Notons que de par la définition de $t$, pour tout couple $(j, t)$, les $P_{n,j,t}$ sont disjointes. En revanche, $P_{n,j,t}$ et $P_{n^*,j^*,t^*}$ peuvent avoir une intersection non vide, dès que $t \neq t^*$.

Pour résumer, la population d'intérêt croît avec la longueur de la période d'enquête. Mais son effectif croît " moins vite " que le temps : en effet, d'un jour sur l'autre, on retrouve certaines personnes dans les centres. $P(J)$ dépend donc fondamentalement de la période de collecte. Bien entendu, l'évolution de $P(J)$ avec $J$ est *a priori* très complexe : on peut penser que $P(J)$ acquiert une certaine stabilité si $J$ est une période assez longue (par exemple 30 jours), mais une période très longue ferait en revanche perdre son sens à ce concept. Deux phénomènes distincts interviennent, dont on peut penser qu'ils ont des temps caractéristiques différents :

1) la population " sans-abri " à un moment donné ne fréquente qu'épisodiquement les centres de la base : pour prétendre la couvrir, il faut donc enquêter sur une période de temps où toutes les personnes de cette population ont au moins une fois recours à des services ; cette période est sans doute de l'ordre de la semaine ou du mois ;

2) la population des sans-abri se renouvelle dans le temps. D'une année sur l'autre, des entrées et des sorties, sans doute nombreuses, interviennent.

La question de la détermination de $J$ revient finalement à savoir si on s'intéresse plutôt à une notion de sans-domicile " à un instant donné " ($J$ plutôt court), approche qui s'attache plutôt à cerner le noyau de la population sans domicile, les " sans-domicile permanents ", ou à une notion de sans-domicile au cours d'une période donnée ($J$ plutôt long), qui va intégrer davantage de " sans domiciles occasionnels ".

On suppose dans un premier temps que l'on s'intéresse à l'estimation d'un total relatif à une variable $Y$ définie sur la population $P(J)$ , $T_J = \sum_{k \in P(J)} Y_k$.

Par exemple, $Y$ peut être l'âge auquel l'individu a terminé ses études, où le nombre de centres qu'il a fréquentés le jour de l'enquête. Ces deux exemples montrent qu'il faut distinguer deux types de variables :

- les variables qui varient au cours du temps ($Y_k = Y_k(j)$). Le nombre de centres fréquentés le jour de l'enquête appartient à cette catégorie. Dans ce cas, la variable $Y$ ne prend pas la même valeur pour un même individu ayant fréquenté deux prestations correspondant à deux jours distincts.

- les variables fixes au cours du temps (par exemple, l'âge de fin d'études), plus précisément au cours de la période de référence de l'enquête. Dans ce cas, la valeur de la variable $Y$ est attachée à l'individu, et sera constante au cours de la période d'enquête.

# 4 Estimation d'un total dans le cas où la variable d'intérêt est constante sur la période d'enquête

Nous traitons d'abord le cas des variables fixes au cours de la période de référence de l'enquête.

## 4.1 Formule fondamentale

On définit l'application $K$, qui à toute prestation $i$ servie durant la période de référence $J$ dans l'ensemble des centres du champ de l'enquête, associe l'individu bénéficiaire de cette prestation.

$K : \{prestations\} \rightarrow \{individus\}$
$i \rightarrow K(i)$

La population d'intérêt $P(J)$ est l'image par $K$ de l'ensemble des prestations servies durant la période de référence dans l'ensemble des centres du champ de l'enquête. Pour tout $k \in P(J)$ , on définit $R_k(J) = card(K^{-1}(k))$, le nombre d'antécédents de $k$ au cours de la période d'enquête.

Dans ce cas, on a l'égalité fondamentale :

$$T_J = \sum_{k \in P(J)} Y_k = \sum_{j=1}^{J} \left( \sum_{c \in C_N} \left( \sum_{i \in P_c} \frac{Y_{K(i)}}{R_{K(i)}(J)} \right) \right) \quad (1)$$

En effet, la variable $Y$ prenant la même valeur pour toutes les prestations $i$ " pointant " sur l'individu $k$, c'est-à-dire telles que $K(i) = k$, le membre de droite peut s'écrire $\sum_{k \in P(J)} \frac{Y_k}{R_k(J)} \left[ \sum_{j=1}^{J} \left[ \sum_{c \in C_N} \left[ \sum_{i \in P_c ; K(i)=k} \right] \right] \right]$

Mais la quantité entre crochets est le nombre de prestations servies à l'individu $k$ durant la période $J$ dans l'ensemble des centres du champ de l'enquête, soit $R_k(J)$, ce qui prouve l'égalité.

On peut alors voir $Y_{K(i)}$ comme attaché à la prestation $i$ correspondante et noter $Y_i$ au lieu de $Y_{K(i)}$, et $Y_i(J)$ au lieu de $R_{K(i)}(J)$. On est ramené à un problème d'estimation du total de la variable sur les prestations.

## 4.2 la méthode du partage des poids appliquée au problème

Rappelons brièvement le principe de la méthode du partage des poids. Pour un exposé plus complet, le lecteur pourra consulter Lavallée (1995), ou Deville (1999) dont nous reprenons les notations.

1. On dispose d'une population $U$ de $n$ unités, et d'une population $V$ de $m$ unités. Ici, les unités de $U$ sont les prestations dans le champ de l'enquête. Les unités de $V$ sont les personnes ayant bénéficié d'au moins une prestation pendant la période de l'enquête (autrement dit dans le cas présent $V = P(J)$ avec les notations précédentes).

2. On suppose qu'il existe des liens entre les unités des deux populations. Ces liens peuvent s'écrire sous la forme d'une matrice $(r_{ik})_{\substack{1 \le i \le n \\ 1 \le k \le m}}$, où $r_{ik} = 1$ si l'unité $k$ de $V$ est reliée à l'unité $i$ de $U$, $r_{ik} = 0$ sinon. Ici, les liens relient les prestations aux personnes ayant fréquenté ces prestations : $r_{ik} = 1$ si la personne $k$ a fréquenté la prestation $i$ de $U$, $r_{ik} = 0$ sinon.

3. Toutes les unités de $U$ ont au moins un lien avec une unité de $V$. Cela est évidemment réalisé ici, par la définition de la population $V$. De plus, ici, chaque unité de la population $U$ pointe sur une unité et une seule de $V$. En particulier, $card(V) \le card(U)$.

Dans le cas général, on s'intéresse au total d'une variable d'intérêt $y$ sur $V$, $Y = \sum_{k=1}^{m} y_k$. Si par exemple on prend $y \equiv 1$ , le total d'intérêt est le nombre de personnes ayant fréquenté un service du champ de l'enquête pendant le mois de l'enquête.

On note $r_k = \sum_{i \in U} r_{ik}$.

L'identité $Y = \sum_{i \in U} \sum_{k \in V} \frac{r_{ik}}{r_k} y_k$ permet de définir pour tout $i \in U$ la variable $z_i = \sum_{k \in V} \frac{r_{ik}}{r_k} y_k$ et on a : $Z = \sum_{i \in U} z_i = \sum_{k \in V} y_k = Y$.

On suppose maintenant que l'on dispose d'un échantillon $s_U$ issu de la population $U$, auquel est associé un jeu de poids $(w_i)_{i \in s_U}$. Cet échantillon définit implicitement un échantillon dans $V$, $s_V$ , précisément

$s_V = \{k \in V ; \exists i \in U, r_{ik} = 1\}$.

On collecte les $r_{ik}$ sur $s_U$, et les $r_{ik}$ sur $s_V$.

Le total $Z = Y$ est estimé par $\widehat{Z} = \sum_{s_U} w_i z_i = \widehat{Y}$.

Et donc, si les poids sont sans biais (c'est-à-dire, établis de manière que $\widehat{Z}$ est sans biais), $\widehat{Y}$ estime sans biais $Y$.

On peut réécrire $\widehat{Z} = \sum_{s_U} w_i \sum_{k \in V} r_{ik} \frac{y_k}{r_k} = \widehat{Y}$. La deuxième somme ne porte que sur $s_V$ par définition, et donc $\widehat{Y} = \sum_{s_V} y_k \left( \sum_{s_U} \frac{w_i r_{ik}}{r_k} \right) = \sum_{s_V} y_k \tilde{w}_k$ , où l'on a posé pour tout $k \in s_V$ :

$$\tilde{w}_k = \frac{1}{r_k} \sum_{s_U} w_i r_{ik} \qquad (2)$$

Ici, $r_k$ est le nombre de liens, c'est-à-dire le nombre de services fréquentés par la personne interrogée pendant la période de référence de l'enquête. C'est la quantité qui était notée $R_k(J)$ dans les sections précédentes. Ce nombre se déduit des données de fréquentation collectées à l'enquête. La formule 2 énonce simplement que le poids d'un individu est égal à la somme des poids des prestations qui ont servi à " l'attraper ", divisée par le nombre de liens avec la base de sondage. On peut donc travailler directement sur les individus échantillonnés : pour chaque individu $k$, on calcule le poids $\tilde{w}_k$, et on estime le total d'une variable d'intérêt $Y$ par $\widehat{Y} = \sum_{s_V} y_k \tilde{w}_k$.

# 5 Estimation sans biais d'un total ou d'un ratio

## 5.1 Estimation d'un total

Dans la partie précédente, on a montré que le total d'intérêt $T_J$ s'écrivait comme un total sur l'ensemble des prestations du champ. Laissant pour l'instant de côté la correction de la non-réponse, supposons maintenant que l'on dispose d'un échantillon de prestations répondantes (au sens où la personne bénéficiant de la prestation a accepté de répondre à l'enquête), auquel est associé un jeu de poids $(w_i)$. Ces poids sont supposés sans biais. Notons $s$ l'échantillon d'individus correspondant aux prestations répondantes échantillonnées, et $s_k$ l'ensemble des prestations échantillonnées qui renvoient à l'individu $k$.

D'après la section précédente, disposant d'un jeu de poids sans biais pour les prestations répondantes, *si on connaissait les* $R_k$ , on estimerait $T_J$ sans biais par

$\widehat{T}_J = \sum_{k \in s} Y_k \widetilde{w}_k$, où l'on a posé $\widetilde{w}_k = \frac{1}{R_k} \sum_{i \in s_k} w_i$ .

Dans ce cas où la variable étudiée ne varie pas au cours de la période d'enquête, il est en théorie " indifférent " (pour le biais de l'estimateur) d'identifier les personnes fréquentant les prestations. Considérons en effet un individu " attrapé " par deux prestations différentes. Deux cas peuvent se produire en pratique :

- on repère que l'individu est le même ; la pondération associée à cet individu sera égale à $\frac{w_1+w_2}{R_k}$, et le terme correspondant à l'individu dans l'estimateur sera égal à $Y_k \frac{w_1+w_2}{R_k}$.

- on ne repère pas que l'individu a déjà été interrogé ; on comptera deux individus différents ; les pondérations associées à ces individus seront égales à $\frac{w_1}{R_k}$ et $\frac{w_2}{R_k}$, et le terme correspondant à ces deux pseudo-individus dans l'estimateur sera encore égal à $Y_k \frac{w_1+w_2}{R_k}$.

Cela suppose bien sûr (et c'est pour cela que la remarque ne vaut pas en pratique) que les informations données par la même personne enquêtées à deux endroits/jours différents soient les mêmes, ce qui est loin d'être acquis. D'autre part, le repérage des individus peut s'avérer crucial si l'on veut correctement estimer la variance des estimateurs produits : cette variance serait sous-estimée si l'on ne considère pas les doublons dans l'échantillon.

Enfin, on peut penser qu'un individu échantillonné deux fois répondra lors du premier tirage, mais pas lors du second. Le second tirage générerait alors une " fausse non-réponse " qu'il est préférable d'éviter.

## 5.2 Estimation d'un ratio

On suppose maintenant que l'on s'intéresse à l'estimation de la moyenne dans la population de référence d'une variable $Y$, $\overline{Y} = \frac{T_J}{N_J} = \frac{1}{N_J} \sum_{k \in P(J)} Y_k$.

Le total $N_J$ est inconnu, il peut s'écrire comme le total d'une variable $X$ valant identiquement 1 pour tous les individus de $P(J)$.

Si l'on connaît les $R_k$ , on estimera $\overline{Y}$ par l'estimateur de Hajek, $\widehat{\overline{Y}} = \frac{\widehat{T}_J}{\widehat{N}_J}$, où $\widehat{N}_J = \sum_{k \in s} \tilde{w}_k$.

# 6 Problèmes pratiques d'estimation

Dans les formules présentées plus haut, la connaissance des liens $R_k(J)$ des personnes avec l'univers des prestations est indispensable. Or, d'un point de vue pratique, on ne connaît pas ces quantités, pour plusieurs raisons :

- une raison théorique : parce que la collecte est étalée dans le temps, et qu'un individu interrogé en début de période ne peut pas prévoir les services qu'il va fréquenter après la date d'interview[2],

- des raisons pratiques : parce que la mémoire des personnes interrogées fait défaut au-delà de quelques jours, et parce que le repérage par l'enquêteur ou le concepteur d'enquête des prestations appartenant à la base de sondage peut s'avérer problématique.

En pratique, il est donc impossible d'estimer un total d'intérêt sur la période de l'enquête (un mois) sans faire des hypothèses *a priori* (nous reviendrons par la suite sur le type d'hypothèses qu'il est possible de faire).

[2]Notons que la collecte doit nécessairement être étalée dans le temps, si l'on veut atteindre toute la population visée ; une collecte synchrone, même si elle était techniquement réalisable, n'attendrait pas l'ensemble de la population-cible mais seulement les personnes qui fréquentent les services à cette date.

## 6.1 Estimation " un jour moyen ", " une semaine moyenne "

On est donc amené à s'intéresser à des quantités qui font intervenir les liens $R_k$ sur une période courte, par exemple le jour ou la semaine.

Introduisons les quantités suivantes:

- le total de la variable $Y$ sur la population des personnes qui fréquentent les services du champ de l'enquête un jour $j$ donné

$\Theta_j = \sum_{k \in P_j} Y_k$ , où $P_j = \bigcup_{n,t} P_{n,j,t}$, un cas particulier étant le nombre de personnes qui fréquentent les services du champ de l'enquête un jour $j$ donné

$N_j = \sum_{k \in P_j} 1 = card(P_j)$ .

- la moyenne de la variable $Y$ sur la population des personnes qui fréquentent les services du champ de l'enquête un jour $j$ donné, $\bar{Y}_j = \frac{\Theta_j}{N_j}$

- $\tau$ , nombre de jours de la période de référence.

Nous définissons les paramètres d'intérêt suivants :

- le total de la variable $Y$ sur la population des personnes qui fréquentent les services du champ de l'enquête un jour " moyen ", dans le sens suivant :

$$\Phi = \frac{1}{\tau} \sum_{j=1}^{\tau} \Theta_j$$

Un cas particulier est le nombre de personnes qui fréquentent les services du champ de l'enquête un jour " moyen " , $\bar{N} = \frac{1}{\tau} \sum_{j=1}^{\tau} N_j$ .

- la moyenne de la variable $Y$ sur la population des personnes qui fréquentent les services du champ de l'enquête un jour " moyen " donné, dans le sens suivant :

$$\Psi = \frac{\Phi}{\bar{N}} = \frac{\sum_{j=1}^{\tau} \Theta_j}{\sum_{j=1}^{\tau} N_j}$$

- les mêmes totaux ou moyennes, mais une semaine donnée ou une " semaine moyenne ", avec les mêmes notations.

Pour estimer ces paramètres, il suffit d'adapter les formules du paragraphe précédent, en constatant que *les $R_k$ comptent maintenant le nombre de services du champ de l'enquête que la personne échantillonnée a fréquentés le jour (resp. la semaine) d'enquête.*

Notons $R_k(j)$ le nombre de prestations de l'univers reçues par l'individu $k$ le jour $j$ uniquement.

$\Theta_j$ sera estimé par $\hat{\Theta}_j = \sum_{k \in s_j} Y_k \tilde{w}_k$ , où $\tilde{w}_k = \frac{1}{R_k(j)} \sum_{i \in s_k(j)} w_i$ ,

$s_j$ est l'échantillon des personnes interrogées le jour $j$,

$s_k(j) = \left\{ i \in \bigcup_t s_{j,t}; K(i) = k \right\}$

De même, $\bar{Y}_j$ sera estimé par $\bar{Y}_j = \frac{\sum_{k \in s_j} Y_k \tilde{w}_k}{\sum_{k \in s_j} \tilde{w}_k}$ .

Ici, les poids dépendent du jour $j$. On mesure l'importance de la définition correcte du paramètre d'intérêt, du fait de la non-constance dans le temps de la population observée : pour s'en convaincre, il suffit de noter que $\Phi$ est différent de $T_J$ défini plus haut. Considérons l'exemple numérique suivant. On suppose que l'enquête dure trois jours. Soit deux individus, notés 1 et 2. Le premier fréquente trois prestations le premier jour, et aucune prestation les deux jours suivants. Il est échantillonné le premier jour par une prestation ayant un poids $w$. Le second fréquente une prestation chaque jour ; il est échantillonné également le premier jour par une prestation ayant un poids $w$.

Les contributions de individus en question aux totaux d'intérêt ainsi qu'à leurs estimateurs sont les suivantes :

| | individu 1 | individu 2 |
|---|---|---|
| $\Phi$ | $\frac{Y}{9}$ | $\frac{Y}{3}$ |
| $\hat{\Phi}$ | $\frac{wY}{9}$ | $\frac{wY}{3}$ |
| $T_J$ | $\frac{Y}{3}$ | $\frac{Y}{3}$ |
| $\hat{T}_J$ | $\frac{wY}{3}$ | $\frac{wY}{3}$ |

Pour donner une analogie commode du comptage des liens, on peut imaginer un guichet où chaque personne qui arrive doit remplir un dossier. Le cas de $T_J$ correspond à un fonctionnement où une personne remplit un dossier la première fois où elle se présente au guichet, et n'en remplit plus les fois suivantes ; le cas du " jour moyen " correspond à un fonctionnement où chaque jour, la personne se présentant doit remplir un dossier, qu'elle soit déjà venue un jour précédent ou pas. On voit bien qu'au bout d'une semaine par exemple, l'analyse des caractéristiques des personnes ayant rempli des dossiers sera très différente dans les deux cas : dans le deuxième cas, *les personnes qui viennent souvent au guichet seront surreprésentées par rapport au premier cas.*

Il est possible de formaliser cette approche. Nous renvoyons le lecteur intéressé à Ardilly et le Blanc (1999).

## 6.2 Estimation pratique des liens avec la base de sondage

Même si l'on se restreint à estimer des quantités de type "semaine moyenne" ou "jour moyen", se pose un problème de connaissance des liens. Il n'est pas en général possible de connaître les liens avec la base de sondage un jour donné (et *a fortiori* une semaine donnée ou sur toute la période de l'enquête).

### 6.2.1 Estimation " un jour moyen "

Pour partager les poids, il faut estimer les liens relatifs au jour de l'enquête ; le cas le plus fréquent est celui de personnes enquêtées en début ou en milieu de journée ; pour ces personnes, les centres fréquentés le soir même ne sont pas connus.

- il est possible que le questionnaire inclue des questions du type " Où allez-vous dîner (resp. dormir) ce soir ? ". Dans ce cas, les réponses peuvent être utilisées pour imputer des liens. La question est bien entendu de savoir si on peut faire confiance aux réponses à ces questions pour refléter les vrais liens, et d'autre part si la non-réponse à cette question ne sera pas trop élevée.

- d'un point de vue plus statistique, on peut utiliser (en faisant l'hypothèse d'une certaine " régularité " des comportements, dans un sens qu'il faudrait définir) des informations portant sur la période correspondante du jour précédent. Les liens correspondants sont sans doute des proxys convenables des vrais liens. Le problème pratique concerne l'éventuelle existence de différentiation des jours de la semaine en ce qui concerne la fréquentation des centres : par exemple, certains centres ne sont pas ouverts en fin de semaine, d'autres n'ouvrent que certains jours précis.

### 6.2.2 Estimation " une semaine moyenne "

Pour partager les poids, on garde tous les liens relatifs à la semaine. La première solution décrite plus haut est évidemment à proscrire. Le questionnaire étant prévu pour récolter les liens sur la semaine précédant l'enquête, pour les estimations une semaine donnée, on peut prendre comme proxy

pour les services fréquentés le jour $j$ de la semaine de référence les services fréquentés par l'individu le jour $(j-7)$. Cela est cohérent si l'on suppose qu'il existe une certaine saisonnalité des services fréquentés selon le jour de la semaine. Cela revient à remplacer dans les estimateurs la semaine civile de référence par une semaine glissante, c'est-à-dire les sept derniers jours à compter de la date d'interview.

## 6.3 Estimation sur l'ensemble de la période d'enquête

Estimer un nombre de personnes fréquentant les services d'intérêt pendant l'ensemble de la période de collecte pourrait apparaître comme un des objectifs de l'enquête. Cependant, comme on l'a montré plus haut, cette estimation du total fait intervenir les liens des individus échantillonnés avec les prestations du champ de l'enquête pendant l'ensemble de la période de collecte. Ces liens ne sont pas connus, la collecte des liens ne portant que sur une semaine. Pour procéder à l'estimation de ce paramètre, il est donc nécessaire de modéliser l'évolution des liens au-delà d'une semaine, ou, ce qui revient au même, de modéliser le comportement de passage des individus dans les centres.

La solution à adopter n'est pas simple. Par exemple, l'hypothèse qui peut venir à l'esprit, à savoir poser $R_k(J) = T.R_k(S)$, où $T$ est le nombre de semaines de l'enquête et $R_k(S)$ le nombre de liens de l'individu $k$ avec les prestations du champ de l'enquête pendant une semaine $S$, conduit à des estimateurs sur l'ensemble de la période identiques aux estimateurs sur une semaine moyenne.

En effet, un estimateur " semaine moyenne " pondère l'individu $k$ par $\sum_{i \in s_k(J)} \frac{w_i}{T.R_k(S_i)}$, où $S_i$ est la semaine durant laquelle la prestation $i$ lui est servie, alors qu'un estimateur théorique " ensemble de la période " pondère l'individu $k$ par $\sum_{i \in s_k(J)} \frac{w_i}{R_k(J)}$.

Une condition suffisante d'égalité de ces estimateurs est donc $R_k(J) = T.R_k(S)$, pour tout individu $k$. Cette condition est notamment satisfaite si pour tout $j$ et tout $k$

$$R_k(J) = J.R_k(j) \tag{3}$$

et *a fortiori* si le nombre de liens journaliers ne dépend pas de $j$.

Ce genre d'hypothèse est certainement trop fort. Sur ce point, l'enquête elle-même permet d'obtenir des informations sur le comportement des individus en matière de fréquentation des centres. Dans la section qui suit, nous considérons un modèle de comportement où la relation 3 est vérifiée seulement en espérance, et nous montrons que les estimations de la population totale réalisées à l'aide de liens portant sur des périodes différentes sont biaisées, d'autant moins que la période de collecte des liens est longue.

Les auteurs conseillent donc de ne pas s'engager dans la voie d'une estimation "ensemble de la période", et de s'en tenir à des estimations une semaine moyenne.

## 6.4 Un modèle probabiliste simple de comportement des individus

Nous cherchons maintenant à répondre à la question suivante : " de combien réduit-on le biais dans l'estimation du total de la population $T_J$ lorsqu'on élargit la fenêtre d'observation? ", c'est-à-dire, en passant d'une collecte des liens sur un jour, à une collecte des liens sur deux jours, trois jours, etc.

Pour cela, nous considérons un modèle probabiliste simpliste de comportement des individus. Considérons qu'il n'existe qu'un centre, fournissant un type de services et une seule prestation par jour et par individu. Considérons la population $P_\infty$, ensemble des personnes se présentant " un jour ou l'autre " dans ce centre. On suppose cette population bien définie et fixe (les entrées et sorties de la population cible sont nulles), et de cardinal fini $N$.

On suppose que pour tous les individus $k$ de $P_\infty$, la fréquentation du centre le jour $j$ est une variable de Bernoulli $F_k(j)$ valant 1 avec une probabilité $\alpha$. On suppose les $F_k(j)$ indépendantes et identiquement distribuées à la fois selon les individus et selon les jours. On définit $R_k(J) = \sum_{j=1}^{J} F_k(j)$, nombre de jours où la personne $k$ est passée dans le centre. Il est clair que $R_k(J)$ croît avec $J$, et qu'on a pour tout $j = 1, ..., 7$ les égalités:

$P\{R_k(J) = 0\} = (1-\alpha)^J$

$E(R_k(J)) = J\alpha = J.R_k(1)$

La relation 3 est donc vérifiée en espérance.

Le nombre de personnes se présentant dans le centre au moins un des jours $1, ..., J$ a pour espérance $E(N_J) = N\left[1-(1-\alpha)^J\right]$, ou, ce qui revient

au même, la fraction de la population d'intérêt se présentant pendant une période d'observation de $J$ jours est d'espérance $\left[1-(1-\alpha)^J\right]$. Dès que le paramètre $\alpha$ est différent de 1 (c'est-à-dire, dès que les individus ne passent pas de manière certaine une fois par jour dans le centre), les populations observées sur un intervalle de temps de $J$ jours sous-estiment $N$. Pour l'application numérique, on considère les cas $\alpha = 0.2$, $\alpha = 0.3$, $\alpha = 0.5$. Les résultats sont représentés dans le graphique 1.

L'objectif principal de l'enquête est de " couvrir " le mieux possible la population $P_\infty$. Selon la probabilité de fréquentation un jour donné, une collecte sur une semaine peut donner des résultats très acceptables (plus de 99 % de la population est atteinte dans le cas d'une probabilité 0.5) ou laisser une part non négligeable de la population d'intérêt dans l'ombre. Dès que $\alpha$ est différent de 1, on a intérêt à étendre la fenêtre de collecte, d'où l'intérêt de considérer des estimateurs " sur une semaine moyenne " par rapport à des estimateurs " sur un jour moyen ".

De l'enquête, on tire des estimateurs $\widehat{N}_1, \widehat{N}_2, ..., \widehat{N}_J$ des totaux $N_1, N_2, ..., N_J$. On a donc :

$\underset{\pi}{E}\,(\widehat{N}_J) = N_J$ et $\underset{\varepsilon}{E}\,(N_J) = N\left[1-(1-\alpha)^J\right]$, où les symboles $\underset{\pi}{E}$ et $\underset{\varepsilon}{E}$ désignent respectivement l'espérance par rapport au plan de sondage et l'espérance par rapport au modèle de comportement, de sorte que l'on a :

$\underset{\varepsilon}{E}\underset{\pi}{E}\,(\widehat{N}_J) - N = -N(1-\alpha)^J \underset{J\longrightarrow+\infty}{\longrightarrow} 0.$

Pour estimer $N$ et $\alpha$ à partir des estimateurs $\widehat{N}_1, \widehat{N}_2, ..., \widehat{N}_J$, on pourrait écrire par exemple

$\frac{\widehat{N}_j}{\widehat{N}_1} = \frac{1-(1-\alpha)^j}{\alpha} + \varepsilon_j$, $j = 2, ...7$, ce qui donne un estimateur $\widehat{\alpha}$ de $\alpha$ ,

puis obtenir un estimateur $\widehat{N}$ de $N$ en faisant la moyenne des $\frac{\widehat{N}_j}{1-(1-\widehat{\alpha})^j}$.

Ce modèle simpliste ne reflète sans doute pas la réalité. La première raison en est que la population considérée n'est pas homogène. Il existe certainement dans la population-cible des sous-groupes ayant des comportements de fréquentation très différents ; par exemple, certains groupes dont les représentants viendraient chaque jour dans au moins un centre, alors que les représentants d'autres groupes viendraient très épisodiquement. Il n'y aurait donc pas un paramètre $\alpha$, mais plusieurs. La deuxième raison tient au fait que la population d'intérêt n'est pas, contrairement à ce que suppose ce modèle, constante dans le temps : elle est sujette à des entrées et des sorties.

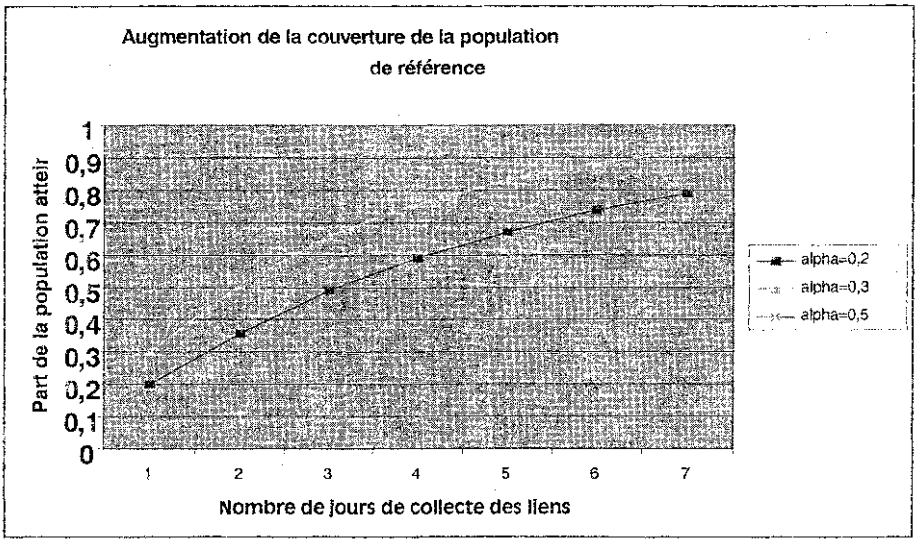
Augmentation de la couverture de la population de référence
Part de la population atteir
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
1
2
3
4
5
6
7
Nombre de jours de collecte des liens
alpha=0,2
alpha=0,3
alpha=0,5

Figure 1:

# 7 Cas où les variables d'intérêt ne sont pas constantes au cours de la période d'enquête

Certaines variables d'intérêt de l'enquête dépendent de la date d'observation, et ne sont pas constantes au cours de la période d'enquête. Ce peut être le cas de réponses à des questions portant sur la journée précédant l'interview, par exemple " Combien de repas avez-vous pris hier ? ", " Combien de fois avez-vous dormi dans la rue la semaine dernière ? ", etc. Les questions sur les liens sont également dans ce cas de figure. Il est donc important de voir dans quelle mesure on peut adapter le formalisme précédent à des estimations portant sur ce type de variables. Soit donc $Y$ une telle variable d'intérêt.

Si nous revenons à l'expression 1, il est facile de voir que la constance des $Y_k$ au cours de la période d'enquête est la condition qui permet de factoriser $Y_k$ et de faire apparaître les liens $R_k(J)$. On en déduit que *le type de calcul mené ci-dessus est toujours valable pour des estimations portant sur des périodes plus courtes que la période sur laquelle les $Y_k$ sont constants.*

Ainsi, pour des variables constantes sur un jour, on pourra parfaitement utiliser des estimateurs " un jour moyen ". Pour des variables constantes sur la semaine, on pourra utiliser des estimateurs " un jour moyen " ou " une semaine moyenne ".

Dans le cas contraire, par exemple si $Y_k$ dépend du jour $j$ et que l'on cherche à estimer le total des $Y$ sur une semaine en utilisant des pondérations relatives à la semaine, la relation 1 n'est plus valable.

# 8 Correction de la non-réponse

Pour décrire complètement l'opération, il reste à préciser comment passer d'un jeu de probabilités d'inclusion (et donc de poids initiaux des prestations incluses dans l'échantillon) à un jeu de poids sur les prestations répondantes. En effet, certaines personnes vont accepter l'interview, d'autres non. On parlera dans le premier cas de prestation répondante, dans le deuxième de prestation non répondante. A partir du poids de sondage initial d'une prestation, on désire obtenir un poids d'extrapolation pour les prestations répondantes. Nous suggérons une correction de la non-réponse par sous-groupes homogènes (pour une description de la méthode, voir par exemple Chambaz et Legendre, 1999).

Concrètement, la difficulté majeure tient au fait qu'il n'y a pas de base de sondage d'individus, et donc pas d'information *a priori* sur les non-répondants. Dans un monde probablement très hétérogène, c'est un handicap considérable. On modélise donc le comportement de réponse des prestations. On sait depuis les enquêtes expérimentales de l'INED que la non-réponse varie fortement selon le type de centre (Marpsat et Firdion, 1997). D'autres variables de la base de sondage peuvent être utilisées pour constituer des groupes homogènes (jour de la semaine, période du jour, groupes d'agglomérations,...).

Une repondération des prestations répondantes conduit à des poids pour les prestations répondantes du type

$w_i = \frac{1}{\delta_i \pi_i}$, où

$\pi_i$ est la probabilité d'inclusion de la prestation $i$ dans l'échantillon

$\delta_i$ est la probabilité *a posteriori* que la prestation $i$ donne lieu à réponse.

On obtient ainsi un jeu de poids pour les prestations répondantes.

En fait, certaines non-réponses viendront du fait qu'un même individu est échantillonné plusieurs fois (la fréquence d'occurrence de cet événement n'est pas connue pour l'instant). Ce type de non-réponse n'est pas une " vraie " non-réponse. Dans ce cas, la procédure de correction de la non-réponse totale amène à repondérer à tort, alors que la " vraie valeur " peut être récupérée dans un questionnaire déjà rempli. On gagnerait en qualité à identifier l'individu non-répondant et la raison de la non-réponse (" j'ai déjà été interrogé "). L'idéal serait de disposer d'un identifiant des répondants. Les impératifs de confidentialité et la prise en compte de l'accueil d'une telle mesure par les personnes interrogées conduisent à ne pas retenir cette idée. Dans la pratique, il sera très difficile de savoir si un individu a déjà été interrogé. L'expérience de l'INED montre qu'il est très difficile, à la fois d'identifier des " doublons " sur la base des variables remplies dans deux questionnaires (les réponses données par les doublons potentiels n'étant jamais strictement identiques), et par ailleurs de vérifier qu'un individu qui déclare avoir déjà été interrogé l'a effectivement déjà été[3].

[3]Même si l'individu est de bonne foi, il peut avoir été interrogé quelques jours auparavant pour une toute autre enquête que l'enquête INSEE...la maîtrise de la charge d'enquête pesant sur les sans-domicile n'étant pas encore à l'ordre du jour.

# 9 Dénombrer les sans-domicile ?

Au-delà des questions de comptage des liens déjà abondamment évoquées, le dénombrement des personnes fréquentant les services se heurte à plusieurs insuffisances de la base de sondage ainsi qu'au caractère indirect de l'échantillonnage.

1. Le risque d'oublier certaines structures lors du dénombrement des centres est important. Même si l'inventaire est exhaustif, le décalage temporel entre cet inventaire et l'enquête à proprement parler rend probable l'apparition de nouvelles structures non recensées dans la base de sondage. Cela peut générer un biais dans la mesure où certains des individus qui fréquenteraient ces structures ne fréquenteraient par ailleurs aucun service de la base de sondage. Par ailleurs, l'absence de biais est conditionnée par un calcul correct des liens, les passages dans des centres non recensés ne devant pas être comptabilisés dans ces liens.

2. Les individus qui fréquenteraient des centres uniquement en dehors des heures " classiques " (concrètement, celles où on se sera donné les moyens de compter les prestations) sont hors champ de l'enquête.

3. Une autre source de biais peut provenir du délicat comptage du nombre total de prestations servies dans les centres lors de l'enquête, ces nombres servant à calculer la probabilité pour une prestation d'être échantillonnée. Pour des raisons budgétaires, une seule personne assurera concrètement le comptage des prestations et l'échantillonnage, ce qui peut poser des problèmes de rigueur d'échantillonnage.

4. Au niveau des concepts, il demeure une difficulté puisque l'enquête doit se dérouler sur un mois et que la population-cible évolue au cours de la période.

L'estimation de la taille de la population est donc particulièrement fragile. Pour cette raison, on peut s'attendre à ce que les erreurs commises soient plus importantes pour les totaux que pour les moyennes.

# References

[1] Ardilly, P., D. le Blanc (1999) : Enquête auprès des personnes sans-domicile : éléments techniques sur l'échantillonnage et le calcul de

pondérations individuelles, une application de la méthode du partage de poids, *document de travail INSEE*, n° F9903.

[2] Chambaz, C., N. Legendre (1999) : Calcul des pondérations dans le panel européen de ménages, Actes des journées de méthodologie statistiques, INSEE Méthodes n° 84-85-86.

[3] Deville, J. C. (1999): Les enquêtes par panel : en quoi diffèrent-elles des autres enquêtes ? suivi de : comment attraper une population en se servant d'une autre, Actes des journées de méthodologie statistiques, INSEE Méthodes n° 84-85-86.

[4] Firdion, J. M., M. Marpsat (1997) : Comptes rendus du groupe " pondérations " de l'enquête auprès des personnes sans-domicile, mimeo.

[5] Lavallée, P. (1995) : Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids, Techniques d'enquête vol. 21, p.27-35.

[6] RTI (1993) : Prevalence of Drug Use in the Washington DC metropolitan area, Homeless and transient population : 1991, technical report # 2.