ESTIMATION SUR DES PETITS DOMAINES
- APPLICATION A L'ENQUÊTE ÉDUCATION 92 -

Sophie Destandau

Sommaire

Introduction

Les sources statistiques

Pendant longtemps, les seules **sources statistiques** disponibles au niveau national et local furent les **recensements** et les **fichiers administratifs.**

Or, la tendance actuelle est de réaliser des recensements tous les 10 ans; et les champs des fichiers sont fréquemment modifiés en fonction de la politique adoptée dans l'administration associée.

Dans les années 40/50, un nouveau type de méthode de collecte d'information est apparu et s'est rapidement développé pour pallier les inconvénients des deux autres : c'est **l'enquête par sondage**.

Une demande de données locales de plus en plus pressante...

Parallèlement au développement des méthodes d'estimation par sondage, est apparue une nouvelle catégorie de demande : **celle de données locales .**

En effet, ces dernières sont nécessaires à l'élaboration des politiques et des programmes gouvernementaux afin de distribuer des fonds ou pour mettre en place une planification régionale.

Les méthodes par sondage pour obtenir des données localisées

Pour obtenir des résultats sur des domaines particuliers, 2 méthodes sont envisageables utilisant aussi bien l'une que l'autre la technique des enquêtes par sondage :

1 - On peut réaliser **une enquête locale**.

Il faut donc

* une base de sondage locale
* une possibilité de stratification
* des informations globales sur la population à étudier
* et surtout une taille d'échantillon équivalente à celle associée à une enquête nationale ; elle ne dépend pas de la taille du domaine

2 - On peut aussi **utiliser de manière locale des enquêtes globales** ce qui constitue l'objet de mon papier.

Les inconvénients d'une enquête nationale par sondage

a) une enquête par sondage globale a pour but la production d'estimateurs fiables sur des variables d'intérêt générales **à des niveaux d'agrégation supérieurs tel le niveau national.**

b) dans la phase d'échantillonnage, **les zones** sur lesquels au cours d'une exploitation ultérieure, il s'avérerait intéressant d'avoir des données, **ne sont pas initialement prévues** ; la zone en question est **donc** généralement **non planifiée** ('unplanned' en anglais). Car, même si le plan d'échantillonnage est stratifié, la zone intéressante peut être différente de celle définie par une strate ou un groupe de strates. Elle nécessite parfois un redécoupage complet de la population.

Qu'est ce qu'un domaine ?

Les données locales ou localisées trouvent des applications multiples et variées dans tous les secteurs (dotations...). Elles peuvent en effet, ne pas être 'locales' dans le sens géographique mais plutôt dans le sens 'réduit à une zone'. C'est pourquoi, le terme employé dans la littérature sur le sujet est **domaine**.

Types de domaine :

* **géographique** (nommé 'area' ou 'areal domain' en anglais)

Exemple : un département, une région

* **croisement** de variables socio-démographiques ou autres (nommé 'domain' ou 'characteristic domain' en anglais) :

Exemple : tranche d'âge-sexe

Qu'entend-t-on par petit domaine ?

'Like beauty, small is in the eyes of the beholder'

Martin Wilk
International Symposium (86)

Purcell et Kish (1979) dans leur article intitulé 'Postcensal Estimates for Local *Areas (or Domains)'* distinguent eux **4 types de domaines selon N_d la taille du domaine** (c'est à dire la somme des unités contenues dans l'intersection de la population avec le domaine) :

Soit $P_d = \frac{N_d}{N}$

- les grands ('major') domaines pour lesquels $P_d \geq 0{,}1$
- les moyens ('minor') domaines pour lesquels $0{,}01 \leq P_d \leq 0{,}1$
- les petits ('mini') domaines pour lesquels $0{,}0001 \leq P_d \leq 0{,}01$
- et les très petits ('rare') domaines pour lesquels $P_d \leq 0{,}0001$.

Le choix de la méthode

Platek, Rao, Särndal, Singh (International Symposium de *Statistique Canada* - 86) ont déclaré que des méthodes d'estimation différentes doivent être employées selon la taille du domaine.

*'A **small area** is any area for which direct design-based estimates cannot be reliably produced from the current sample survey program or a reasonnable expansion thereof. Small areas thus become areas that need other methods of estimation'.*

Le **choix de l' estimation** sur un domaine est lié à plusieurs facteurs comme

♦ la taille de l'échantillon s : n

♦ la taille du domaine dans l'échantillon : n_d

♦ le plan d'échantillonnage c'est à dire encore :

- la base de sondage

- le mode de tirage de l'échantillon : sondage aléatoire simple, stratification, sondage par grappes, en plusieurs phases, à plusieurs degrés...

♦ les variables auxiliaires disponibles

De nombreux statisticiens insistent sur l'importance du choix des variables auxiliaires. Elles doivent être fortement corrélées avec la variable d'intérêt de l'enquête afin d'améliorer la précision des estimations.

Elles peuvent provenir aussi bien de recensements, des registres que d'autres enquêtes par sondage.

♦ les variables d'intérêt à estimer et leurs fréquences (mensuelles, annuelles ...)

♦ la politique décidée par le responsable d'enquête en matière de coût et de précision des résultats

♦ les modèles ou les hypothèses choisis

♦ l'équilibrage entre biais et variance de l'estimateur.

Les différentes catégories d'estimateurs

M.P. Singh, J. Gambino, et H.J. Mantel, dans leur article intitulé *'Les Petites Régions Problèmes et Solutions' (94),* ont établi une classification des estimateurs de ce type en mesurant les conséquences au niveau du **biais** et de la **variance** inconditionnels.

⇒ les **estimateurs de plan** ou **directs**

Ils seront utilisés **lorsque**
* n_d sera grand
* ou lorsque le domaine est prévu dans le plan d'échantillonnage.
Leur biais est nul et leur variance faible.

⇒ les **estimateurs indirects ou de modèle**.

Ils seront utilisés dès lors que n_d **sera considéré comme faible.**

Beaucoup de méthodologues conseillent d'utiliser cette deuxième catégorie d'estimateurs avec beaucoup de circonspection et seulement lorsque
* tous les estimateurs de plan ont été envisagés
* les données auxiliaires ont fourni le maximum de précision supplémentaire.

D'autres classifications existent mais dans cette étude, trois types d'estimateurs de domaines seront détaillés: les estimateurs directs, et deux types d'estimateurs indirects : les estimateurs synthétiques et les estimateurs combinés...

Dans un deuxième temps, deux applications simples de ces méthodes à partir de l'enquête dite Education réalisée en mai/juin 1992 seront présentées.

Notations

1 - La population U

→ Elle comprend N **individus**. Un individu sera noté k, k variant de 1 à N.

→ Elle est découpée en D **domaines** exhaustifs et disjoints. Chaque domaine sera noté d, d variant de 1 à D. Il comprend N_d individus.

$$N = \sum_{d=1}^{D} N_d$$

→ La **variable d'intérêt** relative à l'individu k est notée Y_k.

La somme, la moyenne et la variance de cette variable sur l'ensemble de la population seront notées :

$$Y = \sum_{k=1}^{N} Y_k \qquad \overline{Y} = \frac{1}{N} \sum_{k=1}^{N} Y_k$$

$$S^2 = \frac{1}{N-1} \sum_{k=1}^{N} (Y_k - \overline{Y})^2$$

tandis que celles relatives au domaine d

$$Y_d = \sum_{k=1}^{Nd} Y_k \qquad \overline{Y}_d = \frac{1}{N_d} \sum_{k=1}^{Nd} Y_k$$

$$S_d^2 = \frac{1}{N_d - 1} \sum_{k=1}^{N_d} (Y_k - \overline{Y}_d)^2$$

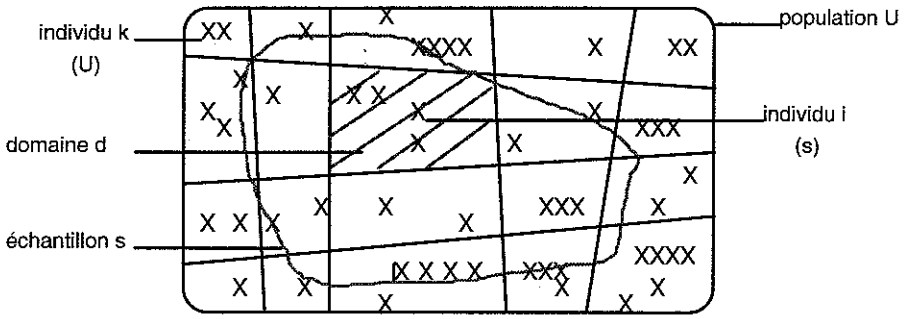
→ La population pourra être aussi subdivisée selon **un autre critère de classification** que celui des domaines ; ce sont G groupes exhaustifs et disjoints.

$$N = \sum_{g=1}^{G} N_g = \sum_{d=1}^{D} \sum_{g=1}^{G} N_{dg}$$

avec N_{dg} le nombre d'individus appartenant à la fois au domaine d et au groupe g.

2 - L'échantillon s

Sa représentation est la suivante :



→ L'échantillon s comprend n **individus**. Un individu appartenant à l'échantillon sera noté i, i variant de 1 à n.

→ Chaque **individu i** appartient aussi à un des D domaines précédemment définis. Chaque **domaine** d comprend n_d individus.

$$n = \sum_{d=1}^{D} n_d$$

→ La **probabilité** de tirage d'un individu i est notée

$$P(i \in s) = p_i$$

→ La **variable d'intérêt** relative à l'individu i est notée: Y_i

3 - Les variables auxiliaires X

→ La **corrélation** entre ces variables auxiliaires choisies et la variable d'intérêt doit être forte.

→ La variable auxiliaire peut être **disponible :**

* sur l'ensemble des individus k de la population

La notation adoptée est X_k , k variant de 1 à N

* sur des groupes g exhaustifs et disjoints différents des domaines

Les totaux sur ces groupes sont notés X_g , g variant de 1 à G

* sur les domaines d

Les totaux sur ces domaines sont notés X_d , d variant de 1 à D

L'objet de l'étude est donc d'estimer le total Y_d
(ou $\overline{Y_d} = \frac{Y_d}{N_d}$)
pour tout domaine d =1 à D

Les trois grandes catégories d'estimateurs

1 - Les estimateurs directs

Principe : les estimateurs directs n'utilisent que les données (de l'enquête ou d'une source extérieure) relatives à un seul domaine à la fois.

Leur caractéristique principale est d'être sans biais.

Pour être les plus performants au niveau biais **et** au niveau variance, ils doivent être utilisés uniquement lorsque n_d **est suffisamment grand** i.e. encore pour des domaines classés dans les 'large areas' ou 'major domains' .

1 - 1 - L'estimateur d'Horvitz Thompson

Cet estimateur est la **référence**
car il utilise seulement les valeurs de la variable d'intérêt et les poids

L'**estimateur d'Horvitz-Thompson ou estimateur par facteur d'extension** (**'expansion estimator'**) du total Y_d est la somme pondérée sur l'échantillon.

$$\hat{Y}_d(h) = \sum_{i=1}^{n_d} \frac{Y_i}{p_i}$$

En subdivisant la population en G **groupes** exhaustifs et disjoints différents des domaines, l'estimateur s'écrit :

$$\hat{Y}_d(h_g) = \sum_{g=1}^{G} \sum_{i=1}^{n_{dg}} \frac{Y_i}{p_i}$$

Biais et Variance :
Inconditionnellement, cet estimateur est sans biais

$$E(\hat{Y}_d(h)) = E(\sum_{k=1}^{Nd} \frac{Y_k}{p_k} \varepsilon_k) = \sum_{k=1}^{Nd} \frac{Y_k}{p_k} E(\varepsilon_k) = Y_d$$

$\varepsilon_k = 1$ si $k \in s_d$

$\varepsilon_k = 0$ sinon

et sa variance s'écrit :

$$V(\hat{Y}d(h)) = \frac{1}{2} \sum_{\substack{k \neq l \\ k,l=1}}^{N_d} \sum (p_k p_l - p_{kl})\left(\frac{Y_k}{p_k} - \frac{Y_l}{p_l}\right)^2$$

$$p_{kl} = P(k \in s \text{ et } l \in s \; k \neq l)$$

$$p_{kk} = p_k$$

➔ **Cas particulier d'un tirage à Probabilités Egales Sans Remise (PESR) :**

Les poids sont alors tous égaux à $\frac{n}{N}$.

$$\boxed{\hat{Y}d(d) = (N/n) \sum_{i=1}^{n_d} Y_i}$$

*** Calcul de la variance de l'estimateur direct $\hat{Y}_d(d)$**

(cf 'Model Assisted Survey Sampling' - C-E Särndal, B. Swenson, J. Wretman - p 392)

$$V(\hat{Y}_d(d)) = N^2 \frac{(1-f)}{n} \frac{(N_d - 1)S^2_{Yd} + N_d(1 - \frac{N_d}{N})\overline{Y}^2_d}{N-1}$$

$$\text{avec } S^2_{Yd} = \frac{1}{N_d - 1} \sum_{k=1}^{N_d} (Y_k - \overline{Y}_d)^2$$

*** Variance estimée de cet estimateur**

$$\hat{V}(\hat{Y}_d(d)) = N^2 \frac{(1 - \frac{n}{N})(n_d - 1)s^2_{yd} + n_d(1 - \frac{n_d}{n})\overline{y}^2_d}{n} \frac{}{n-1}$$

$$\text{avec } s^2_{yd} = \frac{1}{n_d - 1} \sum_{i=1}^{n_d} (y_i - \overline{y}_d)^2$$

$$\overline{y}_d = \frac{1}{n_d} \sum_{i=1}^{n_d} y_i$$

1 - 2 - L'estimateur de post-stratification

Il utilise **en plus la taille du domaine d dans la population :** N_d

L'estimateur de stratification a posteriori ('post-stratified estimator') s'écrit :

$$\hat{Y}_d(pst) = N_d \times \frac{\sum_{i \in sd} \frac{Y_i}{p_i}}{\sum_{i \in sd} \frac{1}{p_i}} = N_d \frac{\hat{Y}_d(h)}{\hat{N}_d(h)}$$

Inconditionnellement, cet estimateur est sans biais.

En subdivisant la population en **groupes** et si on connait N_{dg} , l'estimateur s'écrit :

$$\hat{Y}_d(pst_g) = \sum_{g=1}^{G} N_{dg} \frac{\sum_{i \in sdg} \frac{Y_i}{p_i}}{\sum_{i \in sdg} \frac{1}{p_i}} = \sum_{g=1}^{G} N_{dg} \frac{\hat{Y}_{dg}(h)}{\hat{N}_{dg}(h)}$$

Ces 2 estimateurs sont plus stables que les estimateurs d'Horvitz-Thompson car la connaissance de N_d (ou de N_{dg}) permet un gain de précision.

1 - 3 - L'estimateur par le ratio

Il utilise **les valeurs d'une variable auxiliaire X** bien corrélée avec la variable d'intérêt Y .

L'estimateur par le ratio ('ratio estimator') ressemble à l'estimateur de post-stratification : $\hat{Y}_d(pst)$ avec cette différence près qu'on utilise l'information contenue dans la variable auxiliaire X_d au lieu de l'effectif N_d . Il s'écrit :

$$\hat{Y}_d(r) = X_d \times \hat{R}_d$$

avec $\hat{R}_d$ estimation de $R_d = \frac{Y_d}{X_d}$

En subdivisant la population en **groupes** et à condition de connaître X_{dg} l'estimateur par le ratio s'écrit :

$$\hat{Y}_d(r_g) = \sum_{g=1}^{G} X_{dg} \times \hat{R}_{dg}$$

avec $\hat{R}_{dg}$ estimation de $\frac{Y_{dg}}{X_{dg}}$

Des estimations du ratio : $\hat{R}_d = \frac{\hat{Y}_d(h)}{\hat{X}_d(h)} = \hat{R}_d(h)$ est une des estimations de $R_d = \frac{Y_d}{X_d}$.

Mais cette estimation peut être le rapport d'une estimation directe de Y sur la vraie valeur de X sur le domaine si on en dispose

$$\hat{R}_d(a) = \frac{\hat{Y}_d}{X_d}$$

D'autres estimateurs directs existent comme **l'estimateur de régression** mais ils ne sont pas développés ici compte tenu des estimateurs utilisés dans les deux applications.

Etant donnée la **forte variabilité de la variance de ces types d'estimateurs,** les recherches se sont orientées vers sa diminution parfois au détriment du biais.

C'est ainsi que se sont développées de nouvelles techniques d'estimation fondées sur le principe d'"**emprunter de l'information**" aux autres domaines que ceux sur lesquels se réalise l'estimation en émettant des hypothèses ou en supposant des modèles. Le principe est le suivant :

'to borrow strength from related or similar small areas through explicit or implicit models that connect the small areas via supplementary data'.

Gonzalez (73)

2 - *Les estimateurs synthétiques*

Le qualificatif '**synthétique**' a pour origine l'hypothèse selon laquelle le petit domaine ressemble souvent d'une certaine manière à un autre domaine plus grand dans lequel il est contenu.

Gonzalez (73) qui est à l'origine de la création de ce type d'estimateurs en a donné la définition suivante :

"An unbiaised estimate is obtained from a sample survey for a large area; when this estimate is used to derive estimes for subareas under ***the assumption that the small areas have the same characteristics as the larger area****, we identify these estimates as synthetic estimates."*

2 - 1 - L'estimateur synthétique de moyenne

L'hypothèse la plus simple est de supposer que la moyenne sur un petit domaine est égale à la moyenne globale soit $\overline{Y}_d = \overline{Y}$

L'estimateur synthétique de moyenne ('mean-synthetic estimator') se déduit de l'estimateur direct de post-stratification que l'on peut écrire aussi sous la forme : $\hat{Y}_d(pst) = N_d \times \overline{y}_d$

$$\hat{Y}_d(syn_m) = N_d \times \overline{y}$$

$$\text{avec } \overline{y} = \frac{\sum_{i \in s} \frac{Y_i}{p_i}}{\sum_{i \in s} \frac{1}{p_i}}$$

Lorsque la population est subdivisée en **groupes**, l'hypothèse sous-jacente est $\overline{Y}_{d,g} = \overline{Y}_g$ avec g=1 à G, l'estimateur devient :

$$\hat{Y}_d(syn_m^g) = \sum_{g=1}^{G} N_{d,g} \times \frac{\sum_{i \in s_g} \frac{Y_i}{p_i}}{\sum_{i \in s_g} \frac{1}{p_i}}$$

Calcul du biais :

$$\text{Biais}(\hat{Y}d(syn_m)) = E(N_d \times \frac{\sum_{i=1}^{n} \frac{Y_i}{p_i}}{\sum_{i=1}^{n} \frac{1}{p_i}}) - Y_d = N_d(\overline{Y} - \overline{Y}_d)$$

car l'estimateur sans biais de la moyenne sur l'ensemble des domaines est la moyenne sur l'échantillon $E(\frac{\sum_{i=1}^{n} \frac{Y_i}{p_i}}{\sum_{i=1}^{n} \frac{1}{p_i}}) = \overline{Y}$ et N_d connu.

Sous l'hypothèse $\overline{Y}_d = \overline{Y}$, l'estimateur est **sans biais.**

Application : Dans le cas d'un Sondage Aléatoire Simple, l'estimateur direct de stratification a posteriori a une variance en 1/nd tandis que l'estimateur synthétique associé a une variance en 1/n. Par conséquent, si l'hypothèse est vérifiée, l'estimateur synthétique sera préférable à l'estimateur direct.

2 - 2 - L'estimateur synthétique par le ratio

L'hypothèse ici est la suivante : $R_d = R$,
i.e. encore que le ratio ne dépend pas du domaine

L'estimateur synthétique par le ratio ('ratio-synthetic estimator') se déduit lui aussi d'un estimateur direct, celui par le ratio :

$$\hat{Y}_d(syn_r) = X_d\hat{R}$$
$$\text{avec } \hat{R} \text{ estimation de } R = \frac{Y}{X}$$

Lorsque la population est subdivisée en **groupes**, l'hypothèse sous-jacente est $R_{d,g} = R_g$ avec g=1 à G, l'estimateur devient :

$$\hat{Y}d(syn_r^g) = \sum_{g=1}^{G} X_{dg} \times \hat{R}_g$$

Des estimations du ratio : ici aussi l'estimation de R ou de R_g peut se faire de multiples façons. Citons entre autres $\hat{R} = \frac{\hat{Y}(h)}{\hat{X}(h)} = \hat{R}(h)$. **Singh et Tessier (76)** ont eux utilisé l'information contenue dans la variable auxiliaire X pour estimer R par

$$\tilde{R} = \frac{\sum_{i=1}^{n} \frac{Y_i}{p_i}}{\sum_{\alpha=1}^{N} X_\alpha} = \frac{\hat{Y}(h)}{X}$$

Calcul du biais dans le cas où l'estimation du ratio est celle d'Horvitz-Thompson :

$$\text{Biais}(\hat{Y}_d(\text{syn}_r)) = E(X_d \times \frac{\hat{Y}(h)}{\hat{X}(h)}) - Y_d = X_d(\frac{Y}{X} - \frac{Y_d}{X_d}) = X_d(R - R_d)$$

Si l'hypothèse est vérifiée, le biais de cet estimateur est nul.

Parmi les autres estimateurs synthétiques, on peut aussi citer **l'estimateur synthétique de régression** ou bien les **différents estimateurs ajustés** ou enfin **le Structure PREserving Estimator (SPREE)** de Purcell et Kish (80).

Ainsi, **sous des hypothèses bien précises**, ces estimateurs se révèlent être plus performants pour des petits domaines que les estimateurs directs car ils sont aussi **sans biais** compte tenu des hypothèses et leurs variances sont nettement plus faibles.

Mais, **dès lors que les hypothèses ne sont plus vérifiées**, il faut chercher ailleurs la solution.

3 - Les estimateurs combinés

Dans l'article 'Issues and options in the provision of small area data' paru dans le recueil de la conférence internationale de 1992 à Varsovie dont le thème était 'Small Area Statistics and Survey Design', **Singh, Gambino, Mantel** (93) définissent de manière générale **un estimateur combiné** ('combined estimator') comme étant **la moyenne pondérée de 2 estimateurs quelconques** $\hat{Y}_d(1)$ **et** $\hat{Y}_d(2)$

$$\hat{Y}_d(\text{com}) = a_d \hat{Y}_d(1) + (1 - a_d)\hat{Y}_d(2)$$

avec a_d poids choisi judicieusement

3 - 1 - Le choix de la combinaison des estimateurs

Ici, nous nous intéresserons uniquement à la **combinaison d'un estimateur direct et d'un estimateur synthétique**. Ainsi, l'estimateur combiné *'vise à faire l'équilibre entre la possibilité pour l'estimateur synthétique d'être biaisé (lorsque les hypothèses ne sont pas vérifiées) et l'instabilité de l'estimateur direct'.*

$$\hat{Y}_d(\text{com}) = a_d \hat{Y}_d(\text{dir}) + (1 - a_d)\hat{Y}_d(\text{syn})$$

Rao et Choudhry dans *'Small area estimation : overview and empirical study'* (93) tout comme **Ghosh et Rao** dans *'Small Area Estimation : an appraisal' (94)* nomment ces combinaisons '**composite estimators**'.

3 - 2 - Le choix des poids de la combinaison

1 - des poids fixés à l'avance arbitrairement

Cette méthode simple est trop rigide ; elle ne permet pas de tirer profit de la justesse d'un estimateur direct pour des domaines relativement fournis en données.

2 - des poids dépendant de la taille de l'échantillon : 'sample-size dependent estimators'

L'idée consiste à utiliser :

- à fond l'estimateur direct pour un domaine particulièrement bien représenté dans l'échantillon et
- de profiter des avantages de l'estimateur synthétique dans les autres cas.

2 - 1 - Si on ne dispose pas d'information auxiliaire autre que Nd,

*** Le cas le plus simple** est :

$$\boxed{a_d = \frac{n_d}{N_d}}$$

Si nd est petit par rapport à Nd, alors on privilégiera l'estimation synthétique

➔ Cas particulier : l'estimateur BLUP (Best Linear Unbiased Predictor) d'Hidiroglou

Dans l'article intitulé 'Estimation pour les petits domaines : théorie et pratique à Statistique Canada', **M.A. Hidiroglou** (92) décortique la construction de cet estimateur.

Dans un premier temps, il décompose la population du domaine Ud en 2 :

* l'intersection entre l'échantillon et le domaine : sd
* le complémentaire

$$Y_d = \sum_{i \in s_d} y_i + \sum_{k \in U_d - s_d} Y_k$$

et il estime les 2 termes de manière indépendante.

1 - Sur l'intersection échantillon/domaine d : sd
Il se ramène à l'estimation d'une moyenne sur l'échantillon $\overline{Y}_{s_d}$

car $\sum_{i \in s_d} y_i = n_d \overline{Y}_{s_d}$

$$\hat{\overline{Y}}_{sd} = \frac{\sum_{i \in sd} \frac{y_i}{p_i}}{\sum_{i \in sd} \frac{1}{p_i}}$$

2 - Sur le complémentaire
L'estimation de $\sum_{k \in Ud-sd} Y_k$ se fait sur un modèle de régression en fonction de la disponibilité de variables auxiliaires ou non

On écrira le deuxième terme sous la forme d'une moyenne

$$\sum_{k \in Ud-sd} Y_k = (N_d - n_d)\overline{Y}_{U_d - s_d}$$

$\overline{Y}_{U_d - s_d}$ sera alors estimé en utilisant l'ensemble de l'échantillon (idée du synthétique)

$$\hat{\overline{Y}}_{U_d - s_d} = \frac{\sum_{i \in s} \frac{y_i}{p_i}}{\sum_{i \in s} \frac{1}{p_i}} = \hat{\overline{Y}}(h)$$

D'où l'**estimateur combiné BLUP d'Hidiroglou** s'écrit :

$$\hat{Y}_d(com_H) = nd \frac{\sum_{i \in s_d} \frac{y_i}{p_i}}{\sum_{i \in s_d} \frac{1}{p_i}} + (Nd - nd) \frac{\sum_{i \in s} \frac{y_i}{p_i}}{\sum_{i \in s} \frac{1}{p_i}}$$

Il combine l'estimateur direct de poststratification avec l'estimateur synthétique de moyenne en choisissant le poids égal à $a_d = \frac{n_d}{N_d}$

$$\boxed{\hat{Y}_d(com_H) = \frac{n_d}{N_d} \hat{Y}_d(pst) + (1 - \frac{n_d}{N_d}) \hat{Y}_d(syn_m)}$$

*** Autre poids de combinaison**

On peut choisir

$$\boxed{a_d = \frac{\hat{N}_d(dir)}{N_d}}$$

avec $\hat{N}_d(dir)$ une estimation directe de N_d

en tenant compte du fait que $\frac{\hat{N}_d(dir)}{N_d}$ peut être supèrieur à 1.

➔ Cas particulier : l'estimateur de Drew, Singh et Choudhry (82)

Drew, Singh et Choudhry ont construit un estimateur combinant l'estimateur direct par le ratio et l'estimateur synthétique par le ratio en déterminant les poids à partir de la taille de l'échantillon ; c'est un 'sample-size dependent' estimateur.

$$\hat{Y}_d(com_{DSC}) = a_d \hat{Y}_d(r) + (1 - a_d) \hat{Y}_d(syn_r)$$

$$avec\ a_d = 1 \quad si \quad \frac{\hat{N}_d(h)}{N_d} \geq \delta$$

$$a_d = \frac{1}{\delta} \times \frac{\hat{N}_d(h)}{N_d} \quad si \quad \frac{\hat{N}_d(h)}{N_d} < \delta$$

$$et\ \hat{N}_d(h) = \sum_{i=1}^{n_d} \frac{1}{p_i}\ estimateur\ direct\ et\ sans\ biais\ de\ N_d$$

Le paramètre δ étant choisi de telle manière que l'on contrôle la contribution de l'estimateur synthétique par le ratio. En pratique, il oscille entre 2/3 et 3/2.

Application : Un estimateur du même type est utilisé sur l'enquête canadienne sur la force de travail (Canadian Labour Force Survey) pour estimer annuellement les taux de chômage pour des petites régions à Statistique Canada avec δ =2/3. Pour la majorité des domaines, le poids de l'estimateur synthétique est nul ; pour les autres son poids reste faible entre 10 et 20%.

➔ **Autre cas particulier : l'estimateur de Särndal et Hidiroglou (89)**

Il possède des poids légèrement différents. En effet,

$$\hat{Y}_d(com_{SH}) = a_d \hat{Y}_d(r) + (1-a_d)\hat{Y}_d(syn_r)$$

avec $a_d = 1$ si $\frac{\hat{N}_d(h)}{N_d} \geq 1$

$$a_d = \left(\frac{\hat{N}_d(h)}{N_d}\right)^{t-1} \quad \text{si} \quad \frac{\hat{N}_d(h)}{N_d} < 1$$

et $\hat{N}_d(h) = \sum_{i=1}^{n_d} \frac{1}{p_i}$

Ici, c'est t qui sera choisi de telle sorte que la contribution de l'estimateur synthétique par le ratio soit contrôlée. Ils proposent t=2.

2 - 2 - Si on connait une variable auxiliaire X et N_d

*** un cas simple**

$$\boxed{a_d = \frac{\hat{X}_d(dir)}{X_d}}$$

*** l'estimateur BLUP d'Hidiroglou**

L'idée est la même que dans le cas où on ne connaît que Nd. On estime sur l'échantillon d'une part puis sur le complémentaire.

Justement, $\sum_{k \in U_d - s_d} Y_k$ est estimé par un estimateur synthétique de type par le ratio

$$(N_d \frac{\sum X_k}{N_d} - n_d \frac{\sum \frac{x_i}{p_i}}{\sum \frac{1}{p_i}})\hat{R}$$

$$\text{avec } \hat{R} = \frac{\sum \frac{y_i}{p_i}}{\sum \frac{x_i}{p_i}}$$

D'où l'**estimateur combiné BLUP d'Hidiroglou** s'écrit :

$$\hat{Y}_d(com_H) = n_d \frac{\sum \frac{y_i}{p_i}}{\sum \frac{1}{p_i}} + (N_d \overline{X}_d - n_d \frac{\sum \frac{x_i}{p_i}}{\sum \frac{1}{p_i}}) \frac{\sum \frac{y_i}{p_i}}{\sum \frac{x_i}{p_i}}$$

Ce deuxième estimateur combine l'estimateur direct par le ratio avec l'estimateur synthétique par le ratio en choisissant comme poids $\boxed{a_d = \frac{n_d \hat{X}_d(h)}{\hat{N}_d(h) X_d}}$

$$\boxed{\hat{Y}_d(com_H) = a_d \hat{Y}_d(r) + [1 - a_d]\hat{Y}_d(syn_r)}$$

Remarques sur les 2 estimateurs BLUP d'Hidiroglou:

* Ces 2 estimateurs combinés ont un gros **avantage par rapport aux estimateurs synthétiques** (de moyenne et par le ratio) : ils ont la propriété d'être **"conservateurs".**

En effet, lorsque $n_d = N_d$ (exhaustivité dans le domaine), dans les 2 cas, on a bien $\hat{Y}d(com_H) = Yd$ ce qui n'est pas le cas pour les estimateurs synthétiques.

* Mais, **lorsque** $n_d << N_d$, ce qui est, en définitive le cas le plus courant, le terme $\overline{\overline{Y}}_{s_d}$ estimé par $\hat{\overline{Y}}_{sd} = \overline{y}_d = \frac{\sum_{i \in sd} \frac{y_i}{p_i}}{\sum_{i \in sd} \frac{1}{p_i}}$ de chacun des estimateurs combinés est négligeable.

Le **gain des estimateurs combinés BLUP devient alors minime** par rapport aux estimateurs synthétiques associés $\hat{Y}d(syn_m)$ et $\hat{Y}d(syn_r)$

3 - des poids dépendant des données : 'data dependent estimators'

3 - 1 - Première démarche

Les poids optimaux pour combiner 2 estimateurs sont fonctions de l'erreur quadratique moyenne (Mean Squared Error) des estimateurs combinés et de leur covariance .

En effet, soit l'estimateur combiné suivant : $\hat{Y}_d(com) = a_d \hat{Y}_d(dir) + (1 - a_d)\hat{Y}_d(syn)$

Les poids optimaux $a_d(opt)$ sont obtenus en

$$* \min_{a_d}(MSE(\hat{Y}_d(com))$$

$$* \text{ sous la contrainte } cov(\hat{Y}_d(dir), \hat{Y}_d(syn)) = 0$$

Remarque : La contrainte est assez forte puisque grâce à elle, on tire toute l'information de l'estimateur direct et toute celle de l'estimateur synthétique sans interférence possible

On obtient :

$$a_d(opt) = \frac{MSE(\hat{Y}_d(syn))}{MSE(\hat{Y}_d(syn)) + V(\hat{Y}_d(dir))}.$$

Ces quantités sont généralement inconnues ; il faut donc **les estimer à partir des données**.

Sous l'hypothèse $cov(\hat{Y}_d(dir), \hat{Y}_d(syn)) = 0$, ces poids optimaux peuvent être estimés par

$$\hat{a}_d(opt) = \frac{(\hat{Y}_d(syn) - \hat{Y}_d(dir))^2 - V(\hat{Y}_d(dir))}{(\hat{Y}_d(syn) - \hat{Y}_d(dir))^2}$$

3 - 2 - Démarche de Purcell et Kish (79)

Ils ont cherché plutôt à minimiser la moyenne de MSE c'est à dire

$$\underset{a_d}{Min}\left\{\frac{1}{D}\sum_{d=1}^{D} MSE(\hat{Y}_d(com))\right\}$$

ce qui conduit à une estimation du poids optimal de la forme :

$$\hat{a}_d(opt) = 1 - \frac{\sum_{d=1}^{D} V(\hat{Y}_d(dir))}{\sum_{d=1}^{D} (\hat{Y}_d(syn) - \hat{Y}_d(dir))^2}$$

3 - 3 - Démarche de Fay et Herriot (79)

Réf : Robert E. Fay III et Roger A. Herriot (1979) «Estimates of income for small places : an application of James-Stein procedures to census data».

L'incertitude due à l'échantillonnage s'écrit :

$\hat{Y}_d = Y_d + e_d$ avec e_d erreur d'échantillonnage telle que

$$E(e_d) = 0$$
$$V(e_d) = v_d \text{connue}$$
$$Cov(e_d\, e_{d'}) = 0 \ \forall d \neq d'$$

et Y_d vraie valeur

On modélise une régularité sur les Y_d supposés corrélés aux X_d sous la forme

$$Y_d = aX_d + b + u_d \text{ avec } u_d \text{ erreur de modèle telle que}$$

$$E(u_d) = 0$$
$$V(u_d) = s^2 \text{ inconnue}$$
$$Cov(u_d\ u_{d'}) = 0\ \forall d \neq d'$$

Application :

Soit $\hat{Y}_d = \hat{Y}_d(d)$ l'estimation directe précédemment calculée dans le cas d'un sondage à Probabilité Egale Sans Remise.

Nous avons alors le modèle suivant

$$\hat{Y}_d(d) = aX_d + b + u_d + e_d$$
$$\text{avec } E(u_d + e_d) = 0$$
$$V(u_d + e_d) = V(\hat{Y}_d(d)) + s^2$$
$$Cov(u_d + e_d, u_{d'} + e_{d'}) = 0\ \forall d \neq d'$$

On doit donc estimer a, b, et s^2 par la méthode des moindres carrés quasi généralisés. La régression fournit un estimateur qui sera noté $\hat{Y}_d(reg(s^2))$ dépendant de la variance s^2.

On sait que $E\left\{\sum_{d=1}^{D} \frac{(\hat{Y}_d(d) - \hat{Y}_d(reg(s)))^2}{V(\hat{Y}_d(d)) + s^2}\right\} = D - 2$

On utilise alors la méthode itérative de Newton pour trouver s^2. On notera alors la variance trouvée à la nième itération : $s^2(n)$

$$s^2(n+1) = s^2(n) + \frac{D - 2 - \phi(s^2(n))}{\phi'(s^2(n))}$$

$$\text{avec } \phi(s^2(n)) = \sum_{d=1}^{D} \frac{(\hat{Y}_d(d) - \hat{Y}_d(reg(s^2(n))^2}{V(\hat{Y}_d(d)) + s^2(n)}$$

L'algorithme converge rapidement en moins de 10 itérations en général vers s^2 (*).

L'estimateur de Fay-Herriot s'écrit alors comme un **estimateur combiné de l'estimateur direct et de l'estimateur de régression avec des poids fonction des variances de ces estimateurs.**

$$\hat{Y}_d(com_{FH}) = a_d \hat{Y}_d(d) + (1 - a_d)\hat{Y}_d(reg(s^2(*)))$$

$$\text{avec } a_d = \frac{s^2(*)}{s^2(*) + V(\hat{Y}_d(d))}$$

Application : la méthode de Fay-Herriot est utilisée pour répartir des subventions aux gouvernements locaux et à ceux des états des Etats-Unis à partir des résultats du Recensement de Population et des logements en 1970.

Les deux applications réalisées à partir de l'enquête Éducation 92

1 - Quelques mots sur l'enquête 'Efforts d'Education des familles'

➔ Objectifs

Cette enquête (en abrégé enquête Education) a été conçue et réalisée par l'Insee avec la collaboration de l'Ined (F. Héran et C. Gissot) en mai-juin 1992 de façon à dresser le bilan de l'année scolaire 1991-1992.

Le **thème de l'éducation** y est abordé pour la première fois à l'Insee. En effet, jusqu'ici, les enquêtes sur le système scolaire étaient réalisées auprès des établissements scolaires. Elles ne laissaient pas la parole aux parents, ni même aux enfants.

➔ Organisation de l'enquête

Le **champ de l'enquête** est constitué par les parents d'enfants âgés de 2 à 25 ans scolarisés vivant dans le ménage ou hors ménage.

Les 11703 ménages interrogés ont été sélectionnés par **un plan de sondage stratifié selon la taille de l'unité urbaine à plusieurs degrés**. Tiré dans la base de sondage de l'échantillon maître complétée par la base des logements neufs, il y a eu sur-représentation des ménages ayant déclaré un enfant de cet âge lors du recensement de 1990 et sous-représentation des logements vacants, secondaires et occasionnels.

L'enquête repose sur **plusieurs supports**

- un questionnaire 'parents' qui traite de l'attitude actuelle des parents en matière d'éducation comme le choix de l'établissement, leurs exigences envers l'établissement d'accueil, les dépenses scolaires et extra-scolaires, le temps passé auprès des enfants, les rencontres avec les enseignants.

- un questionnaire 'enfants' (collégiens, lycéens , étudiants) retraçant l'opinion de leurs enfants sur l'école.

➔ La taille de l'échantillon des répondants

Ne disposant pas d'une base de données précise sur la population des parents d'élèves âgés de 2 à 25 ans, il est normal de n'obtenir en fin de compte pour le questionnaire parents que **5265 ménages répondants**.

Mais, les responsables d'enquête ont estimé le **taux de non-réponse** ; il s'élève à 3,6% environ. Ils n'ont donc pas effectué de redressement.

➔ La population des enfants scolarisés et âgés de 2 à 25 ans

La première application porte sur la variable **niveau scolaire de l'enfant** tandis que la seconde porte sur **la dépense scolaire par enfant du ménage**. Or, la dépense scolaire par enfant du ménage tout comme le niveau scolaire de l'enfant est déclarée seulement pour au plus 2 enfants du ménage nommés A et B. Ils sont les 2 premiers parmi un classement par ordre alphabétique des enfants du ménage et hors ménage.

Par conséquent, dans les 2 cas, **un échantillon de 8292 enfants des ménages a été constitué** et les **poids** relatifs à chaque enfant des ménages en fonction des poids des ménages ont été recalculés. Ainsi, si le nombre d'enfants des ménages vivant dans le ménage et hors ménage dans le champ de l'enquête est

* égal à 1, alors le poids de l'enfant sera égal au poids du ménage

* égal à 2, les 2 enfants auront pour poids le poids du ménage

* supérieur à 2, les 2 enfants sélectionnés auront pour poids le poids du ménage multiplié par le nombre d'enfants du ménage dans le champ divisé par 2.

➔ Les publications

* Les dépenses d'éducation des familles - INSEE Première n°261 - juin 1993
* Les efforts éducatifs des familles - INSEE Résultats n°62-63 de septembre 1994
* L'aide au travail scolaire : les mères persévèrent - INSEE Première n°350 - décembre 1994

2 - Les effectifs du préélémentaire et de l'élémentaire en 91/92, par région

➔ le choix du 1er sujet d'application

La variable d'intérêt Y ici est **simple** ; c'est un effectif. D'autre part, les vrais chiffres d'effectifs issus des enquêtes effectuées auprès de tous les établissements scolaires publics et privés du 1er degré et orchestrées par l'Education Nationale étaient disponibles. Des **comparaisons** entre les résultats issus des estimations et la réalité ont pu être effectuées.

2 - 1 - L'étude

* *La population :* les enfants scolarisés pendant l'année scolaire 91/92 âgés de 2 à 25 ans

dans l'échantillon n=8292 enfants
dans la population totale N inconnu

* *Les domaines :* les régions

Nombre : D=22

N_d est le nombre d'enfants dans la population de ménages dans le domaine d

n_d dans l'échantillon

* *Les variables d'intérêt :* le niveau scolaire.

On définit 2 niveaux d'étude : le Préélémentaire et l'Elémentaire.

Soit Y_k et Z_k les variables aléatoires définies ainsi pour chaque enfant k de la population :

$$Y_k = 1 \text{ si } k \in PRE$$
$$Y_k = 0 \text{ sinon}$$

avec PRE = { enfants de la population scolarisés dans le Préélémentaire}

$$Z_k = 1 \text{ si } k \in ELE$$
$$Z_k = 0 \text{ sinon}$$

avec ELE = { enfants de la population scolarisés dans l'Elémentaire}

* *L'objectif*

On cherche à estimer des totaux des variables d'intérêt par domaine d c'est-à-dire encore le **nombre d'enfants scolarisés dans le Préélémentaire et l'Elémentaire par région** à partir des données de l'échantillon et de variables auxiliaires :

$$Y_d = \sum_{k=1}^{N_d} Y_k$$

$$Z_d = \sum_{k=1}^{N_d} Z_k$$

** La variable auxiliaire*

On utilisera comme variable auxiliaire l'effectif réel des enfants scolarisés dans le premier degré par région. Elle est fournie par la la Direction de l'Evaluation et de la Prospective (Ministère de l'Education Nationale). On la notera X_d

** Les vraies valeurs*

La DEP fournit aussi le détail par niveau scolaire c'est-à-dire les effectifs réels des enfants scolarisés dans le Préélémentaire et dans l'Elémentaire par région : Y_d et Z_d que l'on utilisera à des fins de validation des différentes méthodes d'estimation.

On a $X_d = Y_d + Z_d$

2 - 2 - Les estimateurs calculés

La présentation des estimateurs qui suit est faite avec le préélémentaire Y uniquement.

1 - Les estimations directes

➔ Horvitz-Thompson

$$\hat{Y}_d(h) = \sum_{i=1}^{n_d} \frac{y_i}{p_i}$$

➔ Cas particulier

Hypothèse : le tirage de l'échantillon des enfants est supposé être à probabilités égales et sans remise (PESR).

Le poids de chaque individu est donc égal à $p_i = \frac{n}{N}$

$$\hat{Y}_d(d) = \frac{N}{n} \sum_{i=1}^{n_d} y_i$$

Comme N est inconnu, **le taux de sondage** $\frac{n}{N}$ peut être estimé par le rapport du total national du 1er degré dans l'échantillon (enquête Education 92) sur le même total réel (fichier de l'Education Nationale).

$$\hat{f} = \frac{\sum_{d=1}^{D} (y_d + z_d)_{\text{échantillon}}}{\sum_{d=1}^{D} (Y_d + Z_d)_{\text{population}}}$$

➔ par le ratio estimé avec X_d effectifs réels du premier degré en 91/92 par région

$$\hat{Y}_d(rd) = X_d \tilde{R}_{Yd}$$

avec $\tilde{R}_{Yd}$ estimation de $R_{Yd} = \frac{Y_d}{X_d}$

L'estimation du ratio choisie est celle utilisant l'estimation d'Horvitz-Thompson de X_d dans 2 cas :

* *cas général :*

$$\tilde{R}_{Yd}(h) = \frac{\hat{Y}_d(h)}{\hat{X}_d(h)}$$

* *cas particulier d'un tirage PESR :*

$$\tilde{R}_{Yd}(d) = \frac{\hat{Y}_d(d)}{\hat{X}_d(d)} = \frac{\sum_{i=1}^{n_d} y_i}{\sum_{i=1}^{n_d} x_i} = \frac{y_{s_d}}{x_{s_d}}$$

2 - Les estimations synthétiques

Ce sont des **estimations synthétiques uniquement par le ratio.** L'hypothèse sous-jacente est que $R = R_d$

$$\hat{Y}_d(rsd) = X_d \tilde{R}_Y$$

avec $\tilde{R}_Y$ estimation directe de $R_Y = \frac{Y}{X}$

Là aussi on peut distinguer **2 cas** pour l'estimation du ratio :

** cas général*

$$\tilde{R}_Y(h) = \frac{\hat{Y}(h)}{\hat{X}(h)}$$

** cas particulier d'un tirage PESR*

$$\tilde{R}_Y(d) = \frac{\hat{Y}(d)}{\hat{X}(d)} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{y_s}{x_s}$$

3 - Les estimations combinées

Compte tenu du fait que N_d est inconnu, seuls les **estimateurs combinés ayant des poids soit fonction de la variable auxiliaire X, soit fonction de la variance estimée** de l'estimateur direct d'Horvitz-Thompson dans le cas d'un sondage PESR ont pu être construits. L'**estimateur de Fay-Herriot noté** $\hat{Y}_d(FH)$ aussi a pu être calculé.

1 - des poids fonction de X

Les poids sont fonction du total 1er degré X (source Education Nationale) et de son estimation d'Horvitz : $\hat{X}_d(h)$

$$a_d(h) = \frac{\hat{X}_d(h)}{X_d} = \frac{\hat{Y}_d(h) + \hat{Z}_d(h)}{Y_d + Z_d} \text{ si } a_d(h) < 1$$
$$a_d(h) = 1 \text{ si } a_d(h) \geq 1$$

D'où des estimations qui se notent : $\hat{Y}_d(cxh...)$

Dans le cas particulier d'un sondage PESR les poids s'écrivent :

$$a_d(d) = \frac{\hat{X}_d(d)}{X_d} \text{ si } a_d(d) < 1$$
$$a_d(d) = 1 \text{ si } a_d(d) \geq 1$$

D'où des estimations qui se notent : $\hat{Y}_d(cxd...)$

2 - des poids fonction de la variance estimée

Le poids optimal s'écrit - on l'a vu précédemment :

$$a_d = 1 - \frac{\sum_{d=1}^{22} V(\hat{Y}_d(d))}{\sum_{d=1}^{22} (\hat{Y}_d(syn) - \hat{Y}_d(dir))^2} \text{ si } a_d > 0$$

$$a_d = 0 \text{ si } a_d \leq 0$$

Or, on ne connaît la variance d'aucun estimateur direct que nous avons calculé. Par conséquent, il faut l'estimer.

Pour faciliter les calculs, seule la variance de l'estimateur d'Horvitz Thompson dans le cas où le tirage est à Probabilités Egales Sans Remise a été estimée (cf. 1ère partie).

Plusieurs séries d'estimateurs combinés avec des poids fonction de l'estimation de la variance ont été construits. En effet, comme les poids sont aussi fonction des estimations directes et synthétiques, on peut calculer 8 poids différents et toutes les combinaisons possibles.

Ces estimateurs seront notés $\hat{Y}_d(\text{cvh.})$ lorsque ad est fonction de l'estimateur synthétique par le ratio d'Horvitz Thompson (cas général).
Dans le cas particulier où le tirage est PESR, ils seront notés $\hat{Y}_d(\text{cvd.})$.

3 - l'estimateur de Fay-Herriot

$$\hat{Y}_d(FHa) = a_d \hat{Y}_d(d) + (1 - a_d)\hat{Y}_d(rega)$$

$$\text{avec } a_d = \frac{V(\hat{Y}_d(rega))}{V(\hat{Y}_d(rega)) + V(\hat{Y}_d(FHa))}$$

La régression linéaire simple associée à ce modèle s'est effectuée avec la variable indépendante

- du revenu fiscal moyen par région notée FIS

La notation utilisée est $\hat{Y}_d(FHF)$ et $\hat{Y}_d(regF)$ pour les 2 estimateurs (a=F)

- de la population scolaire totale par région notée SCO

La notation utilisée est $\hat{Y}_d(FHS)$ et $\hat{Y}_d(regS)$ pour les 2 estimateurs (a=S)

2 - 3 - Les résultats

1 - Le calage

Compte tenu de l'absence de correction des non-réponses, il s'est avéré nécessaire de **réaliser un calage** en utilisant les totaux du Préélémentaire (Y=2557940) ou de l'Elémentaire (Z=4107640) sur la France métropolitaine (fichier de l'Education Nationale).

Les formules des estimateurs deviennent alors :

$$\hat{Y}C_d(aaa) = \hat{Y}_d(aaa) \times \frac{Y}{\hat{Y}(aaa)}$$

- *Remarque sur le total du premier degré calé* :

$$\hat{X}C(aaa) = \hat{Y}C(aaa) + \hat{Z}C(aaa)$$
$$= \sum_{d=1}^{D} (\hat{Y}_d(aaa) \times \frac{Y}{\hat{Y}(aaa)} + \hat{Z}_d(aaa) \times \frac{Z}{\hat{Z}(aaa)})$$
$$= Y + Z = X$$

Les estimations du 1er degré obtenues aprés calage correcpondent au total réel X, par construction.

2 - Les statistiques de comparaison

Soit l'erreur relative de l'estimateur calé du Préélémentaire

$$EC(\hat{Y}C_d(aaa)) = \frac{\hat{Y}C_d(aaa) - Y_d}{Y_d}$$

avec Y_d vraie valeur

Pour comparer les différentes estimations entre elles, 2 statistiques ont été retenues : **la moyenne et l'écart-type de la valeur absolue de l'écart relatif entre l'estimateur et la vraie valeur, calculés sur les 22 régions.**

3 - Les résultats

1 - Au niveau national n=8292 enfants

Niveau scolaire	Effectif réel Y Z X	Effectif échantillon y z x	Effectif estimé $\hat{Y}(h)$ $\hat{Z}(h)$ $\hat{X}(h)$
Préélémentaire Y	2557940	1546	2318024
Elémentaire Z	4107640	2395	3575087
Premier degré X	6665580	3941	5893111

Grâce à ce tableau, on constate l'utilité d'un calage.

2 - Le Préélémentaire

Les estimateurs combinés avec des poids fonction de la variance sont ici les meilleurs. Sur l'ensemble des 22 régions, ils ont les moyennes les plus faibles et leurs écarts-types sont aussi très faibles.

- **Le classement des estimateurs selon les statistiques de comparaison**

Le **tableau n°1** (p. 341) fournit le classement de tous les estimateurs calculés selon la moyenne sur les 22 régions. On trouvera en tête les **estimateurs qui combinent l'estimateur d'Horvitz Thompson** dans le cas particulier d'un sondage PESR qui seul n'est pourtant pas très performant **avec un estimateur synthétique**. Ce dernier apporte sûrement beaucoup de sa performance.

On peut aussi remarquer que ce sont ces estimateurs combinés qui ont des poids fonctions des estimateurs par le ratio (direct et synthétique).

Viennent ensuite en 5ème position du point de vue de la moyenne, les 2 **estimations synthétiques par le ratio**.

Les **estimateurs combinés dont les poids sont fonction de X** ne sont pas performants puisque le meilleur du point de vue de la moyenne atteint déjà 5,967 et 5,089 en écart-type.

L'estimateur de Fay-Herriot n'est pas non plus 'bon'.

* En utilisant comme variable auxiliaire le revenu fiscal moyen par région, la moyenne de l'écart en valeur absolue s'élève à 11,023 et son écart-type à 9,061 à cause d'une corrélation entre les 2 variables moyenne (0,753) et de la très mauvaise performance de l'estimateur de régression associé. L'estimateur de

Fay-Herriot tire par conséquent sa maigre performance uniquement de l'estimateur direct.

* La variable auxiliaire constituée de la population scolaire totale par région apporte un peu plus d'information que la précédente à l'estimateur de Fay-Herriot (Corr=0,983).

- **L'annexe 1** fournit

☐ les valeurs de quelques estimateurs des effectifs du préélémentaire en 91/92 sur les 22 régions comme
- l'estimateur d'Horvitz Thompson
- l'estimateur synthétique par le ratio
- le meilleur (du point de vue de la moyenne de l'indicateur de comparaison) estimateur combiné dont le poids est fonction de la variance

☐ les écarts associés

☐ ainsi que la taille de l'échantillon par région n_d.

Cette dernière information permet d'avoir un regard différent sur les résultats précédents puis qu'elle permet de constater le rôle important joué par n_d dans les estimations région par région.
En particulier, pour la Corse avec 32 enfants dans l'échantillon, on constate une forte diminution de l'écart relatif à l'estimateur direct à l'écart relatif à l'estimateur synthétique ou combiné (de 41 à 6).

Dans ce tableau, on peut constater à nouveau la bonne performance domaine par domaine de l'estimateur combiné dont le poids est fonction de la variance estimée.

Tableau n°1 : Classement des différents estimateurs d'effectifs d'élèves scolarisés dans le Préélémentaire en 91/92 selon la moyenne de l'écart relatif en valeur absolue

Type	Estimations régionales CALEES	Moyenne	Ecart-type
CV	$a_d(rd,rsd)\hat{Y}_d(d) + (1 - a_d(rd,rsd))\hat{Y}_d(rsd)$	**2,146**	1,767 (4)
CV	$a_d(rd,rsd)\hat{Y}_d(d) + (1 - a_d(rd,rsd))\hat{Y}_d(rsh)$	**2,152**	1,763 (3)
CV	$a_d(rd,rsh)\hat{Y}_d(d) + (1 - a_d(rd,rsh))\hat{Y}_d(rsd)$	**2,156**	1,760(2)
CV	$a_d(rd,rsh)\hat{Y}_d(d) + (1 - a_d(rd,rsh))\hat{Y}_d(rsh)$	**2,162**	1,757 (1)
S	$\hat{Y}_d(rsh)$ et $\hat{Y}_d(rsd)$	**2,286**	1,833
CV	...(56)		
FH	regSCO	**5,471**	7,813
CX	$a_d(Xh)\hat{Y}_d(rd) + (1 - a_d(Xh))\hat{Y}_d(rsh)$	**5,967**	5,089
CX	$a_d(Xh)\hat{Y}_d(rd) + (1 - a_d(Xh))\hat{Y}_d(rsd)$	**5,969**	5,092
CX	...(2)		
D	$\hat{Y}_d(rd)$	**6,962**	5,608
FH	FHSCO	**7,222**	7,556
CX	...(4)		
D	$\hat{Y}_d(rh)$	**8,747**	6,823
CX	...(2)		
CV	...(3)		
CX	...(2)		
CV	...(1)		
FH	FHFIS	**11,023**	9,061
D	$\hat{Y}_d(d)$	**11,474**	8,883
CX	...(4)		
D	$\hat{Y}_d(h)$	**16,661**	9,556
FH	regFIS	**72,487**	71,462

La légende adoptée dans les 2 tableaux est la suivante :

* pour le **type** d'estimateur :

- D comme direct
- S comme synthétique
- CX comme combiné avec un poids de combinaison fonction de la variable auxiliaire X
- CV comme combiné avec un poids de combinaison fonction de la variance estimée
- FH comme Fay-Herriot

* pour signaler l'existence d'estimateurs (sans en indiquer leur moyenne et écart-type) et de leur nombre m, on notera leur type commun et**...(m)**

3 - L'Elémentaire

- **Les estimations réalisées pour l'Elémentaire sont globalement meilleures que celles réalisées pour le Préélémentaire.**

- **Le classement des estimateurs selon les statistiques de comparaison est globalement le même que celui trouvé pour le Préélémentaire. Les estimateurs combinés avec des poids fonction de la variance ainsi que les estimateurs synthétiques viennent toujours en tête.**

- **Le classement des estimateurs selon les statistiques de comparaison (tableau n°2 p. 344)**

Il faut noter la très bonne performance de **l'estimateur de Fay Herriot** avec comme variable auxiliaire le total premier degré DEG due sans doute à l'excellente corrélation entre la variable d'intérêt et la variable auxiliaire (0,994).

Ainsi même si on dispose de peu d'informations : un estimateur direct tout à fait médiocre et une variable auxiliaire très fortement corrélée mais disponible seulement en agrégat sur la région, on obtient un estimateur de Fay Herriot tout à fait honorable ainsi qu'un estimateur de régression.

On trouvera en tête **les estimateurs synthétiques** ainsi que tous **les estimateurs combinés ayant des poids de combinaison fonction de la variance** et de Yrd et de Yrsd. Ces poids a_d sont nuls ce qui entraîne l'égalité de ces estimateurs avec les estimateurs synthétiques.

Tout comme pour le Préélémentaire, on peut remarquer que ce sont les estimateurs combinés dont les poids sont fonction des estimateurs par le ratio (direct et synthétique) qui leur succèdent mais ils combinent l'estimateur direct par le ratio - et non pas le cas particulier d'Horvitz Thompson - avec un estimateur synthétique.

L'estimateur de Fay-Herriot avec pour variable auxiliaire l'effectif scolaire du premier degré suit.

Les estimateurs combinés dont les poids sont fonction de X sont là encore peu performants puisque le meilleur du point de vue de la moyenne atteint déjà 3,739 et 3,171 en écart-type.

- **L'annexe 2** fournit

☐ les valeurs de quelques estimateurs des effectifs de l'élémentaire en 91/92 sur les 22 régions comme
- l'estimateur d'Horvitz Thompson
- l'estimateur synthétique par le ratio
- le meilleur (du point de vue de la moyenne de l'indicateur de comparaison) estimateur combiné dont le poids est fonction de la variance et différent du précédent

☐ les écarts associés.

☐ ainsi que la taille de l'échantillon par région

Tableau n°2 : Classement des différents estimateurs d'effectifs d'élèves scolarisés dans l'élémentaire en 91/92 selon la moyenne de l'écart relatif en valeur absolue

Type	Estimations régionales CALEES	Moyenne	Ecart-type
S	$\hat{Z}_d(rsh)$ et $\hat{Z}_d(rsd)$	**1,380**	1,062
CV	...(16)	**1,380**	1,062
CV	$a_d(rh,rsh)\hat{Z}_d(rd) + (1 - a_d(rh,rsh))\hat{Z}_d(rsh)$	**1,460**	1,084
CV	$a_d(rh,rsh)\hat{Z}_d(rd) + (1 - a_d(rh,rsh))\hat{Z}_d(rsd)$	**1,461**	1,086
CV	$a_d(rh,rsd)\hat{Z}_d(rd) + (1 - a_d(rh,rsd))\hat{Z}_d(rsh)$	**1,472**	1,096
CV	$a_d(rh,rsd)\hat{Z}_d(rd) + (1 - a_d(rh,rsd))\hat{Z}_d(rsd)$	**1,474**	1,098
CV	(24)		
FH	regDEG	**2,303**	2,130
FH	FHDEG	**2,561**	2,665
CV	(6)		
CX	$a_d(Xh)\hat{Z}_d(rd) + (1 - a_d(Xh))\hat{Z}_d(rsd)$	**3,739**	3,171
CX	$a_d(Xh)\hat{Z}_d(rd) + (1 - a_d(Xh))\hat{Z}_d(rsh)$	**3,740**	3,170
CV	...(6)		
CX	...(2)		
D	$\hat{Z}_d(rd)$	**4,345**	3,477
CV	...(4)		
CX	...(2)		
FH	regSCO	**4,952**	7,703
CX	...(2)		
D	$\hat{Z}_d(rh)$	**5,524**	4,256
FH	FHSCO	**8,845**	6,306
CX	...(2)		
CV	...(3)		
CX	...(2)		
CV	...(1)		
CX	...(4)		
FH	FHFIS	**11,107**	7,761
D	$\hat{Z}_d(d)$	**11,424**	7,469
D	$\hat{Z}_d(h)$	**12,70**	9,640
FH	regFIS	**63,427**	56,249

3 - La dépense scolaire par enfant scolarisé dans le premier degré en 91/92 par région

3 - 1 - L'étude

* *La population :* les enfants scolarisés pendant l'année scolaire 91/92 dans **le premier degré**

dans l'échantillon n=3941 enfants
dans la population totale N=6665580 enfants

* *Les domaines :* les régions

Nombre : D=22

N_d est le nombre d'enfants scolarisés dans le premier degré dans la population de ménages dans le domaine d

n_d dans l'échantillon

* *La variable d'intérêt :* la dépense scolaire.

Elle est définie comme la somme des frais de pension - 1/2 pension (PENSI), d'inscription et d'assurances (INASS), des achats de fournitures et de vêtements scolaires (FOVET), des dépenses de loisirs et de sorties dans le cadre de l'école (LOSOR), d'achats de livres scolaires (LIVSC), et de titres de transports (TRANS).

DESCOL=PENSI+INASS+FOVET+LOSOR+LIVSC+TRANS

* *L'objectif*

On cherche à estimer **la dépense scolaire par enfant du premier degré et selon la région d** c'est à dire encore

$$\frac{1}{N_d}\sum_{k=1}^{N_d} DESCOL_k$$

avec $DESCOL_k$ la dépense scolaire totale de l'enfant du premier degré k

et N_d l'effectif d'enfants du premier degré dans la population.

** Les méthodes*

Deux solutions s'offrent alors à nous compte tenu de la formule :

1 - **estimer le total** de la dépense scolaire totale par domaine $\sum_{k=1}^{N_d} DESCOL_k$ puis diviser cette estimation par **Nd**

Car, contrairement à la première application, N_d cette fois ci est connu ; c'est l'effectif d'enfants scolarisés dans le premier degré dans la population totale disponible dans le fichier du ministère de l'Education Nationale.
On notera ces estimations $\hat{MD}_d$

2 - ou bien **estimer le tout** en prenant en compte par conséquent une estimation de Nd. Ces estimations seront notées $\hat{DM}_d$

** L'absence de vraies valeurs*

Contrairement à l'application précédente, nous ne disposons pas des vraies valeurs des dépenses scolaires par enfant scolarisé dans le premier degré.

3 - 2 - Les estimateurs calculés

1 - Les estimations directes

➔ Horvitz-Thompson

1 - La dépense scolaire par enfant connaissant Nd

$$\hat{MD}_d(h) = \frac{\hat{D}_d(h)}{N_d}$$

$$\text{avec } \hat{D}_d(h) = \sum_{i=1}^{n_d} \frac{DESCOL_i}{P_i}$$

2 - On peut aussi l'estimer plus directement par la moyenne

$$\hat{DM}_d(h) = \frac{\hat{D}_d(h)}{\sum_{i=1}^{n_d} \frac{1}{P_i}}$$

➔ **Cas particulier : le plan de sondage est PESR**

1 - Le poids affecté à chaque enfant est constant et est égal à 1/f. Ainsi la dépense scolaire par enfant peut être estimée par

$$\hat{MD}_d(D) = \frac{\hat{D}_d(d)}{N_d}$$

$$\hat{D}_d(d) = \frac{1}{f}\sum_{i=1}^{n_d} DESCOL_i$$

$$\text{avec } f = \frac{n}{N}$$

2 - Mais dans l'hypothèse PESR, on sait que le meilleur estimateur sans biais d'une moyenne est la moyenne sur l'échantillon.

$$\hat{DM}_d(d) = \frac{\sum_{i=1}^{n_d} DESCOL_i}{n_d} = \frac{\hat{D}_d(d)}{n_d}$$

➔ **Post-stratifié**

1 -

$$\hat{MD}_d(pst) = \frac{\hat{D}_d(pst)}{N_d} = \hat{DM}_d(h)$$

$$\text{car } \hat{D}_d(pst) = N_d \frac{\sum_{i=1}^{n_d} \frac{DESCOL_i}{p_i}}{\sum_{i=1}^{n_d} \frac{1}{p_i}} = N_d \frac{\hat{D}_d(h)}{\hat{N}_d(h)}$$

2 - On peut aussi directement estimer la moyenne

$$\hat{DM}_d(pst) = N_d \frac{\hat{DM}_d(h)}{\hat{N}_d(h)}$$

➔ **Post stratifié groupé**

1 - Les groupes sont les 2 niveaux scolaires du premier degré

$$\hat{MD}_d(pst_g) = \frac{\hat{D}_d(pst_g)}{N_d}$$

$$\text{avec } \hat{D}_d(pst_g) = Y_d \frac{\sum_{i=1}^{n_d} \frac{DESCOL * Y_i}{p_i}}{\sum_{i=1}^{n_d} \frac{Y_i}{p}} + Z_d \frac{\sum_{i=1}^{n_d} \frac{DESCOL * Z_i}{p_i}}{\sum_{i=1}^{n_d} \frac{Z_i}{p_i}}$$

en adoptant les notations précédentes Y pour le préélémentaire et Z pour l'élémentaire

2 - On peut là-aussi directement estimer la moyenne à partir des 2 estimations de moyenne sur les 2 groupes

$$\hat{DM}_d(pst_g) = Y_d \frac{\hat{DYM}_d(h)}{\hat{Y}_d(h)} + Z_d \frac{\hat{DZM}_d(h)}{\hat{Z}_d(h)}$$

en notant $\hat{DYM}_d(h)$ et $\hat{DZM}_d(h)$ les estimations d'Horvitz-Thompson des moyennes des dépenses scolaires respectivement pour le Préélémentaire et pour l'Elémentaire.

2 - Les estimations synthétiques

Elles sont sans biais si l'hypothèse de la moyenne est vérifiée

$$\overline{DM_d} = \overline{DM}$$

➔ **simple**

1 -

$$\hat{MD}_d(sn) = \frac{\hat{D}_d(sn)}{N_d} = \hat{DM}(h)$$

$$\text{avec } \hat{D}_d(sn) = N_d \frac{\hat{D}(h)}{\hat{N}(h)}$$

Ainsi la moyenne synthétique revient à celle calculée par Horvitz-Thompson au niveau national et son écart-type sur les 22 régions sera nul.

2 - autre méthode:

$$\hat{DM}_d(sn) = N_d \frac{\hat{DM}(h)}{\hat{N}(h)}$$

➔ groupé

1 - Les groupes sont les niveaux scolaires du premier degré

$$\hat{MD}_d(sn_g) = \frac{1}{N_d}(Y_d \frac{\hat{DY}(h)}{\hat{Y}(h)} + Z_d \frac{\hat{DZ}(h)}{\hat{Z}(h)})$$

2 - autre méthode:

$$\hat{DM}_d(sn_g) = Y_d \frac{\hat{DYM}(h)}{\hat{Y}(h)} + Z_d \frac{\hat{DZM}(h)}{\hat{Z}(h)}$$

3 - Les estimations combinées

1 - des poids fonction de la taille du domaine

Soit le poids $a_d = \frac{\hat{N}_d(h)}{N_d}$ si $a_d \leq 1$

$a_d = 1$ sinon

1 - Première méthode :

On peut alors construire 8 estimateurs combinés compte tenu des 4 estimateurs directs du total : $\hat{D}_d(h)\ \hat{D}_d(d)\ \hat{D}_d(pst)\ \hat{D}_d(pstg)$ et des 2 estimateurs synthétiques $\hat{D}_d(sn)$ et $\hat{D}_d(sn_g)$. Les estimateurs sont alors notés : $\hat{D}_d(cNa)$

Puis on calcule les moyennes connaissant Nd

$$\hat{MD}_d(cNa) = \frac{\hat{D}_d(cNa)}{N_d}$$

avec $a \in [1,8]$

2 - Deuxième méthode

Les estimateurs de la combinaison ne sont pas les dépenses scolaires totales mais déjà des moyennes.

Ainsi, on notera

$$\hat{DM}_d(cNa) = a_d\hat{DM}_d(dir) + (1-a_d)\hat{DM}_d(syn)$$
$$\text{avec } a \in [1,8]$$

2 - les poids fonction de la variance estimée

1 - Première méthode :

L'idée ici est la même que dans l'application précédente. On part du poids optimal

$$a_d = 1 - \frac{\sum_{d=1}^{22} V(\hat{D}_d(dir))}{\sum_{d=1}^{22} (\hat{D}_d(syn) - \hat{D}_d(dir))^2} \text{ si } a_d > 0$$
$$a_d = 0 \text{ si } a_d \leq 0$$

On estime la variance dans le cas le plus simple où l'estimateur direct est l'estimateur d'Horvitz Thompson dans le cas où le tirage est PESR.

Plusieurs séries d'estimateurs combinés avec des poids fonction de l'estimation de la variance peuvent alors être construits.

Ces estimateurs seront notés $\hat{D}_d(cvh.)$ lorsque ad est fonction de l'estimateur synthétique par le ratio d'Horvitz Thompson (cas général).

Enfin, on divise ces dépenses scolaires totales par Nd

$$\hat{MD}_d(cv..) = \frac{\hat{D}_d(cvh.)}{N_d}$$

2 - Deuxième méthode :

La variance qui figure dans le poids est celle d'une estimation directe de la dépense scolaire moyenne. Je l'ai estimée ainsi :

$$V(\hat{D}M_d(d)) = \frac{V(\hat{D}_d(d))}{N_d^2}$$

Le poids optimal est alors estimé par

$$a_d = 1 - \frac{\sum_{d=1}^{22} \hat{V}(\hat{D}M_d(d))}{\sum_{d=1}^{22} (\hat{D}M_d(syn) - \hat{D}M_d(dir))^2} \text{ si } a_d > 0$$

$$a_d = 0 \text{ si } a_d \leq 0$$

Les estimateurs combinés de ce type seront alors notés :

$$\hat{D}M_d(cv..)$$

3 - l'estimateur d'Hidiroglou

Soit le poids $a_d = \frac{n_d}{N_d}$

1 - 1ère méthode :

$$\hat{D}_d(BLUP) = a_d\hat{D}_d(pst) + (1 - a_d)\hat{D}_d(sn)$$

$$M\hat{D}_d(BLUP) = \frac{\hat{D}_d(BLUP)}{N_d}$$

Les poids de la combinaison sont trés faibles. Ainsi,

$$M\hat{D}_d(BLUP) \approx M\hat{D}_d(sn)$$

2 - 2ème méthode :

$$\hat{D}M_d(BLUP) = a_d\hat{D}M_d(pst) + (1 - a_d)\hat{D}M_d(sn)$$

4 - l'estimateur de Fay-Herriot

1 - 1ère méthode :

$$\hat{D}_d(FHa) = a_d\hat{D}_d(d) + (1 - a_d)\hat{D}_d(rega)$$

$$\text{avec } a_d = \frac{V(\hat{D}_d(rega)}{V(\hat{D}_d(rega)) + V(\hat{D}_d(FHa))}$$

- avec le revenu fiscal moyen par région
La notation utilisée est $\hat{D}_d(FHF)$ et $\hat{D}_d(regF)$ pour les 2 estimateurs (a=F)

- avec les effectifs d'enfants scolarisés dans le public par région
La notation utilisée est $\hat{D}_d(FHP)$ et $\hat{D}_d(regP)$ pour les 2 estimateurs (a=P)

Ensuite on calcule la moyenne en divisant par Nd : $\hat{DM}_d(FHP)$

2 - 2ème méthode :

avec la proportion d'effectifs d'enfants scolarisés dans le public par région

$$\hat{DM}_d(FHP) = a_d\hat{DM}_d(d) + (1-a_d)\hat{DM}_d(regP)$$

$$\text{avec } a_d = \frac{V(\hat{DM}_d(regP))}{V(\hat{DM}_d(regP)) + V(\hat{DM}_d(d))}$$

3-3 - Les résultats

1 - Au niveau national

Les dépenses scolaires par enfant estimées sont les suivantes :

Niveau scolaire	$M\hat{D}(d) = \hat{DM}(d)$	$M\hat{D}(h)$	$\hat{DM}(h)$
Préélémentaire	974,6	936,7	848,8
Elémentaire	1473,3	1427,3	1242,3
Premier degré	1277,7	1234,3	1091,3

Les intervalles de confiance des estimations des dépenses scolaires par enfant estimées dans le cas d'un tirage pesr sont les suivants :

Niveau scolaire	<	$M\hat{D}(d) = \hat{DM}(d)$	<
Préélémentaire	884,6	974,6	1064,6
Elémentaire	1398,5	1473,3	1548,2
Premier degré	1219,6	1277,7	1335,8

2 - Au niveau régional

Comment comparer les estimateurs entre eux ne connaissant pas la dépense scolaire des enfants du premier degré réelle?

- comparer leur moyenne ou leur écart-type sur les 22 régions ?

- regarder les différences entre toutes les estimations région par région ?

Le tableau n°3 fournit la moyenne sur les 22 régions des différents estimateurs

Tableau n°3 : Classement des différents estimateurs de la dépense scolaire par enfant du premier degré en 91/92 selon leur moyenne empirique sur les 22 régions

Type	Estimations régionales	Moyenne
D	$\hat{MD}_d(h)$	**964,44**
D	$\hat{MD}_d(d) = \frac{\hat{D}_d(d)}{N_d}$	**1111,06**
D	$\hat{DM}_d(h) = \hat{MD}_d(pst)$	**1121,32**
D	$\hat{MD}_d(pst_g)$	**1133,02**
D	$\hat{DM}_d(d)$	**1141,60**
D	$\hat{DM}_d(pst)$	**1337,23**
D	$\hat{DM}_d(pst_g)$	**2599,71**
S	$\hat{MD}_d(sn) = \hat{DM}(h)$	**1234,33**
S	$\hat{MD}_d(sn_g)$	**1240,96**
CN	$\frac{a_d\hat{D}_d(dir) + (1-a_d)\hat{D}_d(syn)}{N_d}$	**de 1017,10 à 1147,09**
C	$\hat{MD}_d(BLUP)$	**1234,27**
CV	$\frac{a_d\hat{D}_d(dir) + (1-a_d)\hat{D}_d(syn)}{N_d}$	**de 1171,70 à 1224,66**
FH	$\hat{MD}_d(FHP) = \frac{a_d\hat{D}_d(d) + (1-a_d)\hat{D}_d(regF)}{N_d}$	**1096,06**
FH	$\hat{MD}_d(FHF) = \frac{a_d\hat{D}_d(d) + (1-a_d)\hat{D}_d(regF)}{N_d}$	**1093,17**
FH	$\hat{DM}_d(FHP) = a_d\hat{DM}_d(d) + (1-a_d)\hat{DM}_d(regP)$	**1166,96**

Conclusion

Il est bien évident que **le sujet n'est pas épuisé** et que de nombreux estimateurs sont encore à appliquer ne serait ce que sur cette enquête comme les estimateurs de Battese, Harter et Fuller, les estimateurs temporels et l'estimateur hiérarchique de Bayes.

D'autres études sur de nouvelles enquêtes-ménages mais aussi sur des enquêtes-entreprises peuvent être envisagées...

Je vois dans **le logiciel POULPE de calcul de précision** développé par l'Unité Méthodes Statistiques une ouverture intéressante puisqu'il permettrait à la fois de calculer les variances des estimateurs nationaux dans différents cas de sondage plus élaborés que le tirage PESR et d'utiliser ces formules de variance plus appropriées dans le cas des estimateurs combinés. De plus, il pourrait estimer les variances des estimateurs sur des domaines ce qui permettrait une comparaison

En effet, la partie qui me semble intéressante à développer à ce stade de la recherche est celle concernant **la comparaison des performances** de ces différents estimateurs. Cette comparaison de précision peut être faite sur l'ensemble des domaines ou domaine par domaine...

Singh, Gambino et Mantel souligne dans leur dernière étude l'importance de mesurer les performances des estimateurs sur chacun des domaines.

> *'L'élaboration de méthodes qui permettraient d'estimer l'Erreur Quadratique Moyenne pour des domaines pris individuellement devrait figurer parmi les priorités de recherche.'*

L'autre idée intéressante soulignée par **Singh, Gambino, Mantel** est de **prévoir toutes les utilisations d'une enquête avant d'élaborer son plan d'échantillonnage** et par là-même les estimations sur petits domaines. Ainsi, les plans de sondage des grandes enquêtes doivent être établis de telle sorte que les données inférées sur des domaines préétablis qualifiés de 'planifiés' soient fiables. Un arbitrage entre la nécessité de recourir à l'estimation pour domaine et le désir d'obtenir une certaine efficience aux niveaux d'agrégation supérieurs s'impose alors.

> *'On devrait prendre conscience de la question des petites régions dés le début de la conception des plans de sondage pour les grandes enquêtes'*
>
> Singh, Gambino, Mantel
>
> 'Les Petites Régions : Problèmes et Solutions' (94)

BIBLIOGRAPHIE

'Updating Small Area Population Estimates in England and Wales', Stephen Simpson, Ian Diamond, Pete Tonkin - *Royal Statistical Society* (1996).

'Robust Estimation of Mean Squared Error of Small Area Estimators', P. Lahiri, J.N.K. Rao - *Journal of the American Statistical Association*, vol.90 n°430 (juin 1995).

'Generalized sample size dependent estimators for small areas', A.C. Singh, *I.U.H.* Mian - ARC'95.

'Small Area Estimation at Provincial Level in the Italian Labour Force Survey', Pieo D. Falarsi, Stefano Falorsi, Aldo Russo - *ARC 95.*

'Comparaison empirique de méthodes d'estimation pour petites régions pour l'enquête sur la population active italienne', P.D. Falorsi, S. Falorsi, A. Russo - *Techniques d'enquête*, vol20 n°2 pp179-184 (déc 1994).

'Borrowing Strength from past data in Small Domain Prediction by Kalman Filtering - A Case Study', Arijit Chaudhuri, Tapabrata Maiti - 1994.

'Small Area estimation : an Appraisal', M.Ghosh, J.N.K. Rao - *Statistical Science*, vol.9, n°1 - 2 (1994).

'Estimating Activity limitation in the noninstitutionalized population : a method for small areas' - J. Elston Lafata,G.G. Koch, W.G. Weissert - *American Journal of Public Health*, vol.84, n°11 pp1813-1817 (nov.1994).

'Les petites régions : problèmes et solutions', M.P. Singh, J. Gambino et H.J. Mantel - *Techniques d'enquête*, vol.20, n°1 pp3-23 (juin 1994).

'MPLSE à données chronologiques pour petites régions évalués à l'aide de données d'enquête', A.C. Singh, H.J. Mantel, B.W. Thomas - *Techniques d'enquête*, vol.20, n°1 pp35-46 (juin 1994).

'Estimation pour petits domaines dans des plans de sondage avec probabilités inégales', D. Holt, D.J. Holmes - *Techniques d'enquête*, vol.20, n°1 pp25-33 (juin 1994).

'Quelques aspects particuliers des sondages - Estimation sur des domaines', P. Ardilly - *Les sondages*, chapitre IV, Techniques de sondage (1994).

'An application of small area estimation techniques to derive state estimates of health insurance coverage from the 1987 nmes', J.J. Braden, B. Cohen - *Journal of Economic and Social Measurement*, 20, pp193-213 (1994).

'Small Area Estimation : Overview and Empirical Study', J.N.K. Rao, G.H. Choudhry - *ICES Proceedings* (1993).

'Estimation for domains', C.E. Särndal, B. Swensson, J. Wretman - *Model Assisted Survey Sampling*, chapitre 10 pp386-417 Ed.Springer-Verlag (1992).

'Issues and options in the provision of small area data', Singh, Gambino, Mantel - *International Conference on Small Area Statistics and Survey Designs* - Warsaw (sept-oct 1992).

'Design-based approaches in estimation for domains', C.E. Särndal - *International Conference on Small Area Statistics and Survey Designs* - Warsaw (sept-oct 1992).

'Estimation pour les petits domaines : théorie et pratique à Statistique Canada', M.A. Hidiroglou - *Actes des Journées de Méthodologie Statistique 13 et 14 mars 1991*- INSEE-Méthodes n°29-30-31 pp375-401 (déc.1992).

'Méthode d'utilisation d'enquête à un niveau géographique où l'échantillon est faible', F. Jeger - *Actes des Journées de Méthodologie Statistique 13 et 14 mars 1991* - INSEE-Méthodes n°29-30-31 pp363-373 (déc.1992).

'Bayesian prediction in linear models : Applications to small area estimation', G.S.Datta, M. Ghosh - *The Annals of Statistics*, vol.19, n°4 pp1748-1770 (1991).

'Estimation de la production de blé par comté', E.A. Stasny, P.K. Goel, D.J. Rumsey - *Techniques d'enquête*, vol.17, n°2 pp229-244 (décembre 1991).

'Evaluation of procedures for improving population ; Estimates for small areas', K.M.Wolter, B.D. Causey - *Journal of the American Statistical Association*, vol.86, n°414 (juin 1991).

'Estimation robuste pour petits domaines par la combinaison de données chronologiques et transversales', D. Pfeffermann, L. Burck - *Techniques d'enquête*, vol.16, n°2 pp229-249 (décembre 1990).

'The Estimation of the Mean Squared Error of Small-Area Estimators', N.G.N. Prasad, J.N.K. Rao - *Journal of the American Statistical Association*, vol.85, n°409 (March 1990).

'A Bayesian Approach to Small Domain Estimation', Kung-Jong Lui and William G. Cumberland - *Journal of Official Statistics*, vol.5, n°2 (1989).

'Small Domain Estimation : A Conditional Analysis', Carl-Erik Särndal and Michael A. Hidiroglou - *Journal of the American Statistical Association*, vol.84, n°405 (March 1989).

Evaluation of small area estimators : an empirical study, G.H. Choudhry and J.N.K. Rao - (1988), Central Bureau of Statistics OSLO.

'Application of some empirical bayes methods to small area statistics', Emil Spjotvoll and Ib Thomsen - *Invited Paper 29.2 46 th Session of the ISI.*

'An error components model for prediction of county crop areas using survey and satellite data', George E. Battese, Rachel M. Harter, Wayne A. Fuller - april 14 (1986).

'Synthetic estimators, SPREE and best model-based predictors of small area means', J.N.K. Rao - 1986.

'Synthetic estimates for small areas : problems and results of a simulation experiment', Adam Marton - *Journal of the United Nations ECE 4*, 71-80 North-Holland (1986).

'Small area statistics an international symposium', R. Platek, J.N.K. Rao, C.E. Särndal, M.P. Singh - Ottawa May 1985.

'Regression analysis and ratio analysis for domains : a randomization-theory approach', Eva Elvers, Carl Erik Särndal, Jan H. Wretman, Göran Örnberg - *Journal of Statistics*, vol.13, n°2 (1985) (La Revue Canadienne de Statistique).

'An overview of small area estimation techniques', J. Dumais, S. Earwaker, JF. Gosselin, D. Paton, KP. Srinath, R. Verma - march (1985).

'The multivariate components of variance model for small area estimation', Wayne A. Fuller, Rachel M. Harter - November 4 (1985).

'Design-consistent versus model-dependent estimation for small domains', Carl Erik Särndal - *Journal of the American Statistical Association*, vol.79, n° 387, september (1984).

'Une bibliographie pour l'estimation pour les petites régions', Jean Dumais, David Paton, Ravi Verma, Stephen Earwaker, J.F. Gosselin, K.P. Srinath - *Techniques d'enquête*, vol.9, n°2 (1983).

'Postcensal estimates for local areas using current samples with census as the source of sampling frame', S.M. Tam - *International Statistical Review* (1982).

'On estimating population and income for local areas', Evelyn M. Kitagawa and Bruce D. Spencer - May 1981.

'Postcensal estimates for local areas (or domains)', Noel J. Purcell and Leslie Kish - *International Statistical Review* (1980).

'A biometrics invited paper', Noel J. Purcell and Leslie Kish - *Biometrics* 35, 365-384, june 1979.

'Estimates of income for small places : an application of james-stein procedures to census data', Robert E. Fay III and Roger A. Herriot - *Journal of the American Statistical Association*, vol.74, n°366, june 1979.

'A model-Based approach to estimation for small subgroups of a population', D. Holt, TMF. Smith, and T.J. Tomberlin - *Journal of the American Statistical Association*, vol.74, n°366, june 1979.

'A predictive approach to subdomain estimation in finite populations', Petter Laake - *Journal of the American Statistical Association*, vol.74, n°366, june 1979.

'Some estimators for domain totals', M.P. Singh and R. Tessier - *Journal of the American Statistical Association*, vol.71, n°354, june 1976.

'A regression method for estimating population changes of local areas', Eugene P. Ericksen - *Journal of the American Statistical Association*, vol.69, n°348, december 1974.

'Estimation for domains in multistage sampling', Myint tin and Than toe - *Journal of the American Statistical Association*, vol.67, n°340, december 1972.

ANNEXE 1

Quelques estimateurs des effectifs du préélémentaire en 91-92

RG	n_d échantillon	Y_d vraie valeur	$\hat{Y}_d$(h) direct	$\hat{Y}_d$(rsh) synthétique	$\hat{Y}_d$(cvd17) combiné	Ec($\hat{Y}_d$(h)) écart	Ec($\hat{Y}_d$(rsh)) écart	Ec($\hat{Y}_d$(cvd17)) écart
11	1615	487364	446043.35	483263.53	484031.10	8.4784	0.84136	0.68386
21	203	63936	71080.40	63897.27	63998.76	11.1743	0.06058	0.09816
22	289	89583	111296.82	91073.16	91217.81	24.2388	1.66344	1.82491
23	247	80502	89666.63	85834.54	85970.87	11.3844	6.62411	6.79346
24	305	101864	77021.98	103944.64	104109.73	24.3874	2.04256	2.20464
25	198	64105	79468.77	65099.57	65202.97	23.9666	1.55147	1.71276
26	196	66430	81412.39	67980.41	68088.38	22.5536	2.33390	2.49643
31	667	220591	270700.44	213483.25	213822.33	22.7160	3.22214	3.06843
41	334	106471	93979.11	106971.30	107141.21	11.7327	0.46990	0.62947
42	259	68324	59819.16	71820.24	67871.53	12.4478	5.11715	0.66224
43	185	50275	47962.55	51464.42	51546.16	4.5996	2.36583	2.52842
52	517	146545	188179.84	146856.34	147089.60	28.4110	0.21246	0.37162
53	395	133330	120505.70	127630.68	127833.39	9.6185	4.27460	4.12256
54	215	65941	56242.95	66273.85	66379.12	14.7072	0.50478	0.66441
72	316	107771	90903.63	110456.17	110631.61	15.6511	2.49155	2.65434
73	337	96271	111114.46	93982.01	94131.28	15.4184	2.37765	2.22260
74	89	24206	16642.17	24746.35	24785.65	31.2478	2.23229	2.39466
82	781	256818	235810.10	251678.24	252077.98	8.1801	2.00132	1.84567
83	196	49956	42690.43	50998.16	51079.16	14.5439	2.08616	2.24830
91	321	89745	93497.48	89513.20	89655.38	4.1813	0.25828	0.09986
93	595	178514	168378.44	180995.83	181283.30	5.6777	1.39027	1.55131
94	32	9398	5523.19	9976.82	9992.67	41.2301	6.15902	6.32763

moyenne	16,66	2,29	2,15
écart-type	9,56	1,83	1,77

ANNEXE 2

Quelques estimateurs des effectifs de l'élémentaire en 91-92

RG	n_d échantillon	Z_d vraie valeur	$\hat{Z}_d(h)$ direct	$\hat{Z}_d(rsh)$ synthétique	$\hat{Z}_d(cvh31)$ combiné	$EC(\hat{Z}_d(h))$ écart	$EC(\hat{Z}_d(rsh))$ écart	$EC(\hat{Z}_d(cvh31))$ écart
11	1615	771943	739084.76	776043.47	776492.16	4.2566	0.53119	0.58931
21	203	102570	92762.79	102608.73	102668.06	9.5615	0.03776	0.09560
22	289	147739	155452.02	146248.84	146333.40	5.2207	1.00864	0.95141
23	247	143169	141005.79	137836.46	137916.15	1.5109	3.72465	3.66898
24	305	168999	134761.50	166918.36	167014.87	20.2590	1.23115	1.17405
25	198	105534	88944.17	104539.43	103487.12	15.7199	0.94242	1.93955
26	196	110716	93333.50	109165.59	107966.49	15.7001	1.40035	2.48339
31	667	335712	410002.67	342819.75	343017.96	22.1293	2.11722	2.17626
41	334	172279	142687.52	171778.70	171878.02	17.1765	0.29040	0.23275
42	259	118828	135987.22	115331.76	115398.44	14.4404	2.94227	2.88616
43	185	83833	80468.02	82643.58	82691.36	4.0139	1.41880	1.36180
52	517	236139	262596.43	235827.66	235964.01	11.2042	0.13185	0.07411
53	395	199255	204981.51	204954.32	205072.83	2.8740	2.86032	2.91979
54	215	106758	118495.84	106425.15	106486.68	10.9948	0.31178	0.25415
72	316	180060	165000.02	177374.83	177477.39	8.3639	1.49126	1.43431
73	337	148631	158821.42	150919.99	151007.25	6.8562	1.54005	1.59876
74	89	40279	32879.03	39738.65	39761.63	18.3718	1.34151	1.28447
82	781	399015	358178.35	404154.76	404388.43	10.2344	1.28811	1.34667
83	196	82937	85571.51	81894.84	81942.19	3.1765	1.25657	1.19948
91	321	143512	181171.88	143743.80	143826.91	26.2416	0.16152	0.21943
93	595	293132	316040.42	290650.17	290818.22	7.8151	0.84666	0.78933
94	32	16600	9413.63	16021.18	16030.44	43.2914	3.48690	3.43109

moyenne	12,70	1,380	1,46
écart-type	9,64	1,062	1,08