Panorama des principales méthodes d'estimation sur petits domaines Aspects théoriques

Pascal ARDILLY

INSEE - UMS

Table des matières

Introduction

1. Contexte :

Dans toutes les enquêtes par sondage, on rencontre par définition des erreurs d'échantillonnage : elles traduisent le fait que les estimations issues des données collectées dépendent de l'échantillon tiré. L'erreur d'échantillonnage est conditionnée par de nombreux paramètres mais l'un d'eux, fondamental, est la taille de l'échantillon : lorsque cette taille est faible, il y a un fort risque de mauvaise qualité des estimations ! Une telle situation se rencontre dès lors que l'on s'intéresse à des sous-populations, parfois même dans le cadre de grosses enquêtes, car la taille de l'échantillon qui va conditionner la qualité de l'estimation est celle qui concerne la sous-population étudiée. La suite de ce document s'intéresse aux estimations produites à partir de ces sous-populations, appelées « domaines », en particulier lorsque les tailles d'échantillon recoupant ces sous-populations sont faibles. On parle alors de « petits domaines »

Dans la pratique, les exemples de petits domaines sont légion. C'est en particulier le contexte qu'offrent les enquêtes ménages à l'Insee pour tout paramètre défini sur une région[1], un département, une unité urbaine, voire même une commune. La sous-population n'est d'ailleurs pas nécessairement définie par des critères géographiques : ce peut être une catégorie socioprofessionnelle donnée, une tranche d'age, une nationalité, un niveau de diplôme, etc. Ainsi, estimer le revenu moyen des ménages en Bourgogne, le taux de chômage en ZUS, le taux de chômage des moins de 25 ans, le montant total d'une allocation versée aux personnes souffrant d'une déficience mentale, ou encore la proportion de logements locatifs dans le parc des logements de 6 pièces ou plus, sont des problèmes d'estimation sur petits domaines.
.

2. Objectif et premières notations :

On s'intéresse à l'estimation d'un total, d'une moyenne ou d'un pourcentage définis sur le domaine appelé a (le domaine est souvent une zone géographique, d'où le choix de "a" pour "aire"). La variable d'intérêt est notée Y, et l'individu est repéré par son identifiant i (qui va donc indexer Y). On note N la taille de la population totale, et N_a la taille du domaine a. Les paramètres à estimer grâce à l'enquête sont :

$$Y_a = \sum_{i=1}^{N_a} Y_i \qquad \text{et} \qquad \overline{Y}_a = \frac{1}{N_a}\sum_{i=1}^{N_a} Y_i$$

respectivement vrai total et vraie moyenne sur le domaine a. Lorsque la variable Y_i vaut 1 si i vérifie une propriété donnée (exemple : sexe masculin et âge < 50 ans) et 0 sinon, le paramètre devient la proportion des individus du domaine qui vérifient la propriété en question (ici le pourcentage d'hommes de moins de 50 ans au sein du domaine).

Savoir estimer correctement un total va permettre d'estimer toutes les fonctions de totaux, même si elles sont complexes. En effet, une méthode simple consiste à considérer le paramètre complexe et à substituer les estimateurs aux vrais totaux : pour estimer

$$\theta_a = f(Y_a^1, Y_a^2, ..., Y_a^p)$$

on forme

$$\hat{\theta}_a = f(\hat{Y}_a^1, \hat{Y}_a^2, ..., \hat{Y}_a^p)$$

[1] Sauf les très grosses enquêtes, comme l'enquête Emploi - où la région n'est plus assimilée à un petit domaine.

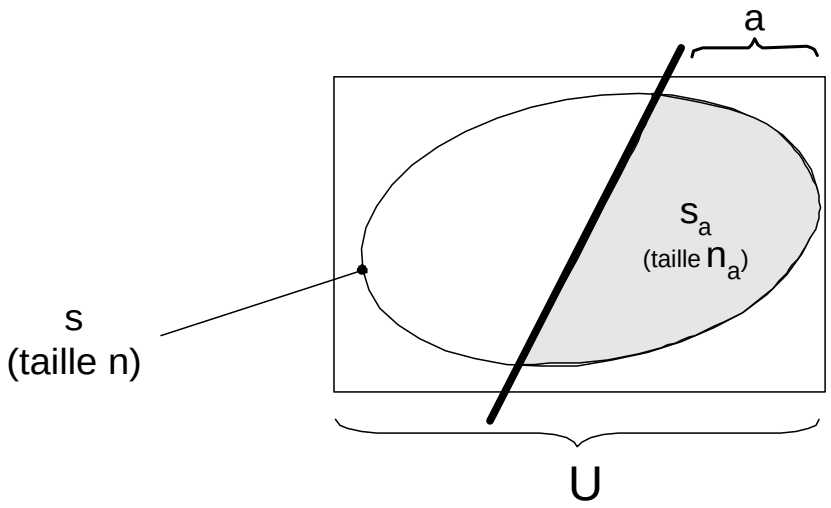
L'opération introduit ainsi un biais qui lui est propre, mais on montre qu'à moins de manipuler des échantillons extrêmement petits, ce biais là reste une composante négligeable de l'erreur totale. Ce qui suit permet donc l'estimation sur petits domaines, par exemple de ratios, de dispersions, de coefficients de corrélation ou encore de coefficients de régression. En revanche, les statistiques qui ne sont pas des fonctions de totaux ne relèvent pas de ce document, comme par exemple les quantiles et leurs satellites

3. Sources d'informations et compléments de notation :

On s'intéresse généralement à la technique d'estimation sur petits domaines lorsqu'on dispose de données d'enquête relatives au phénomène étudié : si tel n'est pas le cas, on doit compter seulement sur d'éventuels fichiers administratifs : soit ces fichiers contiennent l'information individuelle nécessaire et suffisante, auquel cas les méthodes à employer sont de nature comptable, soit ils ne la contiennent pas (ce qui correspond tout de même à la plupart des situations) et on ne peut pas obtenir l'estimation voulue.

On considérera donc qu'on dispose d'un échantillon d'individus interrogés dans le cadre d'une enquête par sondage. Soit il s'agit de la seule source accessible, soit il est possible d'exploiter des sources exhaustives "ou presque" (comme les enquêtes annuelles de recensement de l'Insee) qui fournissent de l'information auxiliaire corrélée aux variables d'intérêt. Dans le premier cas, il y a peu de marge de manœuvre et on devra se contenter d'estimateurs assez "pauvres" - et vraisemblablement peu efficaces. Le second cas est heureusement beaucoup plus commun : c'est en effet en exploitant habilement l'information auxiliaire que l'on va pouvoir augmenter la qualité des estimations sur petits domaines. La réussite des opérations passera donc en premier lieu par le recensement des sources disponibles et par l'examen critique des variables susceptibles d'expliquer le phénomène étudié sur le domaine.

L'échantillonnage dans la population globale U va produire un échantillon s, de taille n. La partie de cet échantillon qui recoupe le domaine a sera noté s_a et sa taille n_a. On a $0 \le n_a \le n$, et on notera qu'il est possible que le hasard conduise à $n_a = 0$ si le domaine est tout petit ou si n est déjà petit. Comme le plan de sondage est a priori quelconque, un individu i de U a une probabilité de sélection notée $\prod_i$. On rappelle que la valeur $\prod_i$ est choisie par le sondeur. En particulier avec un tirage aléatoire simple ou plus généralement avec tout échantillonnage de taille fixe à probabilités égales, on a $\Pi_i = n/N$.



4. Concepts d'erreur en sondage.

La sensibilité d'un estimateur à l'échantillon tiré constitue une préoccupation tout à fait centrale dans la question du choix des estimateurs. Les concepts de qualité essentiels sont au nombre de trois. Si on note $p(s)$ la probabilité de tirer l'échantillon s, si $\hat{\theta}(s)$ est l'estimation obtenue à partir de s et si θ est la vraie valeur (inconnue, à estimer), on distingue :

i) Le biais : $E(\hat{\theta}) - \theta = \sum_s p(s) \cdot \hat{\theta}(s) - \theta$

ii) La variance $V(\hat{\theta}) = E(\hat{\theta} - E\hat{\theta})^2 = \sum_s p(s)(\hat{\theta}(s) - E\hat{\theta})^2$

iii) L'erreur quadratique moyenne $EQM(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = Variance + (biais)^2$

Le biais et la variance sont des indicateurs de natures différentes, puisque le premier mesure une tendance centrale et le second une dispersion. L'erreur quadratique moyenne est un concept général qui « mixte » biais et variance. S'il y a un seul indicateur à fournir pour caractériser la qualité d'un estimateur, c'est probablement ce dernier qu'il faut considérer.

On peut aussi s'intéresser à une précision relative, définie comme le rapport de l'écart type d'échantillonnage (racine carrée de la variance) à la vraie valeur. C'est un indicateur appelé « coefficient de variation » :

$$CV(\hat{\theta}) = \frac{\sqrt{V(\hat{\theta})}}{\theta}$$

On a coutume de considérer qu'un CV inférieur à 5% correspond à une estimation satisfaisante.

5. Estimation de la taille de population d'un domaine avec un sondage aléatoire simple :

Nous avons vu au 2. que l'objectif consistait à estimer un total sur un domaine. L'estimation $\hat{N}_a$ de la taille d'un domaine a est un cas particulier d'estimation de total : il suffit de considérer la variable constante égale à « 1 » sur a pour s'en convaincre, puisque son vrai total sur le domaine vaut N_a. Cette question se traite immédiatement dans le cas d'un sondage aléatoire simple, si on se souvient qu'on estime sans biais toute proportion vraie par la proportion associée dans l'échantillon. Soit

$$\frac{\hat{N}_a}{N} = \frac{n_a}{n}$$

On sait également exprimer la variance de la proportion estimée (mettons que n est négligeable devant N) :

$$V\left(\frac{\hat{N}_a}{N}\right) \approx \frac{1}{n} . \frac{N_a}{N} . \left(1 - \frac{N_a}{N}\right)$$

soit

$$V(\hat{N}_a) \approx N^2 . \frac{1}{n} . \frac{N_a}{N} . \left(1 - \frac{N_a}{N}\right)$$

Si le domaine est (très) petit alors N_a/N sera (très) petit. Comme souvent n est (très) grand, on obtient des valeurs numériques de l'écart type de $\hat{N}_a$ tout à fait modestes. On a donc un sentiment de très bonne précision absolue - et cela est indiscutable (les intervalles de confiance ont effectivement une faible amplitude).

Il n'y a pas de piège ici, en revanche il y a une subtilité : si on raisonne en précision relative, donc en coefficient de variation, on obtient une conclusion inverse. En effet,

$$CV\left(\hat{N}_a\right) = \frac{\sqrt{V\left(\hat{N}_a\right)}}{N_a} \approx \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{N}{N_a}} \cdot \sqrt{1 - \frac{N_a}{N}} \approx \frac{1}{\sqrt{E(n_a)}} \cdot \sqrt{1 - \frac{N_a}{N}}$$

si bien qu'un petit domaine va donner lieu à un (très) grand CV - sauf si la taille totale de l'échantillon n est très grande. Voici quelques tailles n nécessaires (et suffisantes) pour atteindre un CV de 5% (ce qui constitue une précision relative correcte, mais sans plus) en fonction de différentes valeurs de $P_a = N_a/N$ correspondant à des PCS recensées en 1990 (source RP1990) :

PCS	Clergé	Cadres fonction publique	Professions scientifiques	Agriculteurs Exploitants Retraités	Ouvriers Retraités
$\mathbf{P_a}$	0,001	0,005	0,01	0,02	0,05
n	400 000	80 000	40 000	20 000	8 000

6. Qu'est-ce qu'un petit domaine ?

A partir de quand peut-on considérer qu'on a à faire à un « petit domaine » ? Il n'existe pas de définition dans l'absolu qui permette de qualifier indiscutablement de « petit » le domaine auquel on s'intéresse. Il s'agit plutôt d'une appréciation de circonstance, qui est intimement liée à la question de la précision. En fait, lorsqu'il cherche à estimer un paramètre, le statisticien a des concepts pour définir la précision - concepts présentés dans la partie 4 - et une cible quantitative (plus ou moins explicite…) pour chaque estimation calculée. Si l'estimation classique relative au domaine a une précision insuffisante par rapport à son objectif, on se trouve potentiellement en situation d'estimation sur petit domaine.

Concrètement, pour une variable d'intérêt donnée, si on mesure la qualité par le coefficient de variation (qui est un très bon critère), on peut se fixer un seuil de coefficient de variation en deçà duquel l'estimation classique suffira et au-delà duquel on envisagera l'application de méthodes « petits domaines ». Par exemple - mais, encore une fois, cela reste à l'appréciation de chacun en fonction des objectifs de l'enquête - on considérera qu'il faut déclencher une estimation « petits domaines » si le coefficient de variation dépasse 20%. Cela suppose évidemment que l'on soit en mesure d'estimer la précision des estimateurs que l'on construit, ce qui peut être un sérieux obstacle avec des plans de sondage complexes.

7. Grandes catégories d'estimateurs et plan du document :

On distingue trois grands types d'estimateurs :

• Les ***estimateurs directs***, dont la caractéristique est de ne pas faire appel à de l'information (collectée Y ou auxiliaire X) relative à des individus se situant hors du domaine. Cela ne signifie pas qu'il n'y ait pas d'information auxiliaire qui soit utilisée, mais cela signifie que si c'est le cas, cette information ne concerne que les individus du domaine étudié.

• Les ***estimateurs indirects construits à partir de modèles implicites***. Il s'agit là d'estimateurs dont l'expression même s'appuie, d'une façon ou d'une autre sur des hypothèses de comportement reliant le domaine au reste de la population. Ces estimateurs ont une qualité qui dépend de la validité du modèle, lequel traduit toujours une hypothèse du style "du point de vue de tel paramètre, le domaine et la population se comportent de la même façon". Par exemple, on postulera que la moyenne de Y sur le domaine, soit $\overline{Y_a}$, est identique à la moyenne de Y sur l'ensemble de la population. Le modèle est qualifié d'implicite parce qu'il s'applique à un niveau "macro" : il concerne des paramètres et non des comportements individuels, et il ne fait pas intervenir de variable aléatoire. Il a donc pour caractéristique de résulter d'une approche purement descriptive.

• Les ***estimateurs indirects construits à partir de modèles explicites***. Dans cette catégorie, on trouve tout ce qui se fait à partir de la modélisation des paramètres des domaines en fonction, d'une part de variables explicatives conçues au niveau du domaine, et d'autre part d'un aléa (qui n'est pas du tout un aléa de sondage, mais une variable aléatoire de même nature que l'aléa de l'économètre). Par exemple, on postulera que la moyenne $\overline{Y_a}$ est une combinaison linéaire de K variables explicatives X_a^k définies au niveau de chaque aire a, cela à un aléa près. Eventuellement (mais pas nécessairement), on considérera que l'aléa suit une loi de nature connue (loi de Gauss par exemple). On peut aussi modéliser les comportements individuels, - donc à un niveau plus fin que le domaine - que l'on peut relier aux paramètres définis sur le domaine. Par exemple, on postulera qu'une grandeur individuelle quantitative Y_i, où i décrit le domaine d'intérêt, est une combinaison linéaire de variables explicatives individuelles X_i^k, comprenant en particulier un effet propre au domaine, plus un aléa individuel qui regroupe tout ce qui n'est pas expliqué par ces X_i^k.

Cette classification n'est pas la seule envisageable bien sûr, mais nous l'avons retenue pour structurer la suite du document.

Le choix de la méthode dépend ensuite de la nature de l'hypothèse que l'on souhaite faire (donc des risques que l'on est prêt à prendre) et de l'information auxiliaire dont on dispose, sans oublier les moyens et compétences disponibles pour appliquer la théorie associée. L'estimation directe (la voie la plus simple) est toujours techniquement possible, mais on voit bien que dans nombre de cas elle est très fragile, voire fantaisiste, si bien qu'elle fait office de dernier recours. L'estimation indirecte implicite peut également être choisie en toute circonstance mais elle ne prend son sens que si on est en mesure d'étayer (au moins un peu…) l'hypothèse qui sert de base à la méthode. Même si ce n'est pas une nécessité absolue pour appliquer cette approche, de fait on a besoin d'information auxiliaire pour rendre le modèle plus crédible (sans information auxiliaire, le modèle sera presque toujours trop fruste). Quant à l'estimation indirecte explicite, elle ne peut s'envisager qu'avec une information auxiliaire assez riche et actualisée, et des moyens humains et informatiques conséquents pour dérouler toute la procédure. C'est de loin l'approche la plus complexe techniquement et la plus coûteuse. Il n'est pas évident qu'elle soit in fine plus efficace que la méthode implicite, mais elle prend généralement l'avantage lorsqu'il s'agit d'apprécier la pertinence de la modélisation.

Estimation directe

Il s'agit de construire un estimateur « correct » de Y_a (ou de $\bar{Y}_a$) sans utiliser d'information « extérieure » au domaine de a : dans toute cette partie, seuls les individus de l'aire a sont impliqués dans la définition de l'estimateur.

Le problème essentiel posé par cette catégorie d'estimateurs est leur grande variance d'échantillonnage : en effet, lorsque a est petit, n_a sera vraisemblablement petit (sauf si l'échantillonnage a été conçu spécifiquement pour obtenir n_a suffisamment grand, mais ce n'est généralement pas le cas) et la variabilité de l'estimateur sera grande. La justification des extensions locales d'enquête s'appuie directement sur ce principe. Ainsi, pour une enquête nationale de 20 000 ménages par exemple, on trouvera 200 ménages dans un département « moyen », ce qui est très largement insuffisant pour produire n'importe quelle estimation selon l'approche « traditionnelle ». La mise en place d'une extension départementale augmentera de manière ciblée la taille d'échantillon. Même au niveau régional, la situation apparaît bien périlleuse, puisque avec 1000 logements[2] les estimations régionales d'ensemble seront entachées d'une incertitude déjà conséquente[3] et, surtout, on ne pourra pas raisonnablement prétendre faire la moindre exploitation sur des sous-populations au sein de cette région. Donc, là encore, l'extension régionale fournira une solution au problème.

Ce phénomène de grande variance est malheureusement assez incontournable et, sauf si les Y_i ont une structure particulièrement favorable (par exemple une très faible dispersion naturelle alors que l'échantillonnage est équiprobable - prenons le cas de l'estimation du revenu moyen des fonctionnaires d'un grade donné au sein d'une région), il n'y a pas de miracle à attendre du contexte : avec peu d'information, on ne peut pas obtenir une estimation d'une grande qualité.

Ce discours doit cependant resté modéré et on ne peut pas non plus écarter brutalement la piste de l'estimation directe : en effet, entre le tout petit domaine et la très grande population, il y a tout un dégradé de situations et pour des domaines "pas trop petits", l'estimation directe peut être malgré tout adaptée aux objectifs. Cela d'autant qu'il y a deux éléments à prendre en compte : d'une part il est possible, comme on va le voir, d'améliorer l'estimation directe avec de l'information auxiliaire pertinente, et d'autre part il faut raisonner en fonction d'objectifs de précision, lesquels ne sont pas nécessairement très ambitieux. Aussi, si on ne cherche pas une très faible variance mais seulement un ordre de grandeur du paramètre, l'estimation directe peut être une alternative acceptable.

Parmi les estimateurs directs, on distingue essentiellement :

- L'estimateur classique de Horvitz-Thompson ;
- L'estimateur par la régression, et ses nombreuses variantes.

[2] Dont il faudra extraire les seuls répondants !

[3] Pour estimer une proportion de 50% avec un tirage aléatoire simple, on obtient avec 1000 répondants une incertitude de +/- 3 points de pourcentage avec 95 chances sur 100.

1. L'estimateur de Horvitz-Thompson :

1.1 Estimation du total :

Il consiste à retenir seulement les individus de s_a et à conserver les poids d'origine, utilisés, lorsque l'extrapolation porte sur la population tout entière. Pour estimer un total, la pondération « primitive » pour de très nombreux plans de sondage[4] est égale à l'inverse de la probabilité de sélection, soit en l'absence de non-réponse :

$$\boxed{\hat{Y}_a = \sum_{i \in s_a} \frac{Y_i}{\Pi_i}}$$

Cet estimateur est sans biais du vrai total Y_a. Si l'échantillonnage est à probabilités égales et de taille fixe, on a $\Pi_i = n/N$ et donc

$$\hat{Y}_a = \frac{N}{n} \sum_{i \in s_a} Y_i = N \cdot \frac{n_a}{n} \cdot \bar{y}_a$$

où $\bar{y}_a$ est la moyenne simple des n_a valeurs Y_i des individus de s_a. Ces conditions sont assez souvent (approximativement) satisfaites quand on tire une enquête dans l'échantillon-maître ou dans l'EMEX[5]. Il est important de noter que n_a est aléatoire : c'est le hasard qui fixe la valeur de n_a entre 0 et n (techniquement, si le sondage est aléatoire simple, n_a suit une loi hypergéométrique -sinon la loi de n_a est très complexe). S'il y a de la non-réponse totale, on continue à utiliser les poids (corrigés) des individus de s_a. Ainsi, si on a estimé par $\hat{R}_i$ la probabilité de réponse (inconnue) R_i de i, on a :

$$\boxed{\hat{Y}_a = \sum_{i \in r_a} \frac{Y_i}{\Pi_i \cdot \hat{R}_i}}$$

où r_a est l'échantillon de répondants dans a. Cet estimateur est légèrement biaisé en toute généralité parce que la probabilité de réponse vraie est inconnue. Dans le cas où le tirage est à probabilités égales et de taille fixe ($\Pi_i = n/N$) et que R_i est supposé constant (techniquement, R_i suit une loi de Bernoulli), on va estimer R_i par le taux de réponse empirique m/n où m est le nombre total de répondants, d'où :

$$\hat{Y}_a = \frac{N}{m} \sum_{i \in r_a} Y_i$$

Sous la seule condition que r_a ne soit pas vide, on peut donc toujours produire un estimateur sans biais ou peu[6] biaisé du total sur un domaine (petit ou non). On remarquera - ce qui peut être très appréciable - qu'il n'y a pas besoin de connaître la taille N_a du domaine. Concrètement, il faut partir des poids sans biais d'origine, corrigés de la non-réponse. S'il y a eu un redressement (par

[4] Pas pour tous, néanmoins. Par exemple pour le tirage en deux phases, ce n'est pas vrai.

[5] Echantillon-Maitre pour les EXtensions régionales : c'est une base de sondage de logements complémentaire à l'échantillon-maître, qui permet d'augmenter la taille des échantillons régionaux quand on effectue une extension régionale.

[6] On considère ici qu'on parvient à corriger « correctement » la non-réponse, donc à partir d'un modèle de réponse valide sur la population du domaine. En pratique, l'estimation des probabilités de réponse s'effectue d'après un modèle estimé en mobilisant l'ensemble de l'échantillon, et on fait ensuite l'hypothèse que le comportement est le même dans le domaine et dans l'ensemble de la population.

Calmar, par exemple), on considère les poids calés et on fabrique l'estimateur linéaire correspondant en se restreignant à l'échantillon répondant r_a.

Du point de vue de la variance, on se trouve confronté aux problèmes traditionnels de calcul de précision, sur lesquels on ne revient pas ici. La non-réponse complique encore davantage les calculs. En supposant, pour simplifier, qu'il y ait pas de non-réponse, on peut exprimer la vraie variance de $\hat{Y}_a$ selon :

$$(1)\; V\left(\hat{Y}_a\right) = \frac{1}{2} \cdot \sum_{(i,j) \in U} \Delta_{i,j} \cdot \left(\frac{Y_i}{\Pi_i} 1_{i \in a} - \frac{Y_j}{\Pi_j} 1_{j \in a} \right)^2$$

où $\Delta_{i,j} = \Pi_i \Pi_j - \Pi_{i,j}$ ($\Pi_{i,j}$ est la probabilité de sélection conjointe de i et j) et $1_{i \in a}$ vaut 1 si i est dans a et 0 sinon (variable dite "indicatrice" d'appartenance au domaine). Cette expression permet de voir qu'une faible variance s'obtient concrètement en imposant Π_i proportionnel à $Y_i \cdot 1_{i \in a}$, ce qui est fort difficile à assurer : en effet, $Y_i \cdot 1_{i \in a}$ vaut Y_i si i est dans a et 0 sinon, ce qui ne permet pas de bénéficier de conditions favorables de proportionnalité. Le cas « idéal » semble être celui où on choisit un système de probabilités d'inclusion Π_i proportionnelles aux vraies valeurs Y_i (c'est évidemment théorique, puisque les Y_i sont inconnus…). Dans ce cas, en l'absence de non-réponse, $\hat{Y}_a = K \cdot n_a$ où K est la constante de proportionnalité entre Y_i et Π_i. On peut vérifier que K est égal à $Y_a / E(n_a)$, c'est-à-dire :

$$\hat{Y}_a = \frac{n_a}{E(n_a)} \cdot Y_a$$

Si la variable d'intérêt est la constante « 1 », on est dans la cadre de l'estimation de la taille N_a du domaine, et on retrouve l'estimateur bien connu de N_a (voir 5. de l'introduction):

$$\hat{N}_a = \frac{N}{n} \cdot n_a$$

On identifie bien ici l'origine de la variabilité : c'est la taille aléatoire n_a, qui est une cause « incompressible » de perte d'efficacité. La conclusion est donc que l'estimateur direct habituel formé à partir des poids de l'enquête est pénalisé (entre autres) par la variabilité de n_a.

L'estimation de variance s'effectue en utilisant un logiciel de calcul de précision, comme par exemple POULPE. Il n'y a pas de difficulté spécifique aux petits domaines car on constate que l'expression de variance théorique donnée ci-dessus est l'expression classique où Y_i a été remplacé par $Y_i.1_{i \in a}$. Concrètement, face à un problème d'estimation de variance sur un domaine a, on va fabriquer une nouvelle variable $Y_i \cdot 1_{i \in a} = Z_i$ et on va estimer la variance de $\sum_{i \in s} Z_i / \Pi_i$, ce qui est une procédure classique (on se heurtera certes aux problèmes habituels en matière d'estimation de variance, mais il s'agit de difficultés directement causées par la complexité de l'échantillonnage).

Si on a à faire à un sondage aléatoire simple, on obtient, en supposant $n << N$

$$V\left(\hat{Y}_a\right) \approx N_a^2 \cdot \frac{1}{\frac{nN_a}{N}} \cdot \left[S_a^2 + \overline{Y}_a^2 \left(1 - \frac{N_a}{N} \right) \right] \qquad (1)$$

avec S_a^2 dispersion vrai des Y_i dans a. Cette expression (approchée) est très instructive : le terme N_a^2 rappelle que l'on estime un total et le terme entre crochets fait, comme de tradition, intervenir la dispersion des Y_i. Mais il y a deux remarques encore plus intéressantes à faire.

- Primo, la présence au dénominateur de $n\frac{N_a}{N}$, qui est égal à l'espérance de la variable aléatoire n_a (c'est-à-dire au n_a "moyen" attendu) : comme on pouvait s'y attendre - et cela n'apparaît pas si clairement dans (1) lorsqu'il s'agit d'un plan complexe - la précision varie bien comme l'inverse de la taille de l'échantillon recoupant le domaine. Si le domaine est petit, N_a/N sera petit et à moins d'avoir un échantillon national de taille n énorme, nN_a/N sera probablement insuffisant pour que la variance soit faible. Par exemple si on considère un département "moyen", $N_a/N=1\%$, et si l'enquête nationale concerne $n=10000$ logements on aboutira à une taille "utile" égale à 100 sur ce département.
- Secundo, le terme entre crochets ne s'annule pas si les Y_i sont constants. Cela est contre intuitif si on attribue la cause d'une variabilité à la seule hétérogénéité des comportements, mais en la circonstance c'est la variabilité de n_a qui contribue à augmenter la variance de l'estimateur du total.

On peut former l'estimateur sans biais de la variance suivant (si n_a est grand devant 1) , utilisable pour estimer des intervalles de confiance à 95% pour Y_a :

$$\hat{V}\left(\hat{Y}_a\right) \approx N_a^2 \cdot \frac{1}{n_a} \cdot \left(\frac{n_a}{n.\frac{N_a}{N}} \right)^2 \cdot \left[s_a^2 + \bar{y}_a^2 \left(1 - \frac{n_a}{n} \right) \right]$$

Le calcul de précision est en soi une question difficile, qui se trouve encore un peu complexifiée lorsqu'on s'intéresse aux domaines parce qu'il y a en fait deux concepts concurrents. Outre l'approche traditionnelle, on peut en effet s'intéresser à la variance conditionnelle à la taille n_a, c'est-à-dire que l'on peut chercher à mesurer la variabilité de $\hat{Y}_a$ lorsque l'échantillon s_a est composé de toutes les manières possibles, sous la condition express de contenir n_a individus, n_a étant la taille constatée lors de l'enquête. Lorsqu'on pratique le sondage aléatoire simple, on peut exprimer ainsi l'espérance et la variance conditionnels de $\hat{Y}_a$:

$$E\left[\hat{Y}_a \middle| n_a\right] = N \cdot \frac{n_a}{n} \bar{Y}_a$$

$$V\left[\hat{Y}_a \middle| n_a\right] = \left(N \cdot \frac{n_a}{n} \right)^2 \left(1 - \frac{n_a}{N_a} \right) \frac{S_a^2}{n_a}$$

Ces expressions découlent d'un résultat technique remarquable (qui ne se généralise pas lorsque l'échantillonnage n'est plus aléatoire simple) : lorsqu'on conditionne par n_a, tout se passe comme si on pratiquait un sondage aléatoire simple de taille n_a dans une population de taille N_a. Dans cette optique, $\hat{Y}_a$ apparaît biaisé, puisque $Y_a = N_a \cdot \bar{Y}_a$ mais

$$N \cdot \frac{n_a}{n} \neq N_a$$

La variance, dans ce contexte précis, est inadaptée au calcul d'intervalles de confiance parce qu'elle apparaît comme fonction croissante de n_a dès lors que $n_a \leq N_a/2$. Malgré ces imperfections, la notion de qualité conditionnelle doit être considérée avec l'attention qu'elle mérite, car il y a du bon sens à juger que la qualité d'une estimation est liée au n_a constaté plutôt qu'au n_a attendu : si on attend en moyenne 100 interviews dans un département, mais que par malchance il y ait seulement 20 logements tirés dans ce département, on peut penser que la qualité finale sera celle d'une enquête de taille 20 - et non celle d'une enquête de taille 100 !

Du fait du théorème d'équivalence précédent, on voit immédiatement que la vraie dispersion S_a^2 est estimée sans biais par la dispersion s_a^2 calculée dans l'échantillon s_a. Autrement dit

$$E\left[s_a^2 \middle| n_a\right] = S_a^2$$

D'où un estimateur sans biais de la variance, utilisable pour estimer des intervalles de confiance à 95% pour Y_a :

$$\hat{V}\left[\hat{Y}_a \middle| n_a\right] = \left(N \cdot \frac{n_a}{n}\right)^2 \left(1 - \frac{n_a}{N_a}\right) \frac{s_a^2}{n_a}$$

1.2 Estimation de la moyenne :

On rappelle tout d'abord qu'une proportion est une moyenne, donc l'estimation de toute proportion rentre dans le cadre de cette partie. On suppose toujours, pour simplifier, qu'il n'y a pas de non-réponse totale -sinon on repondère les répondants par l'inverse de leur probabilité de réponse estimée (le fait que l'on travaille sur des domaines n'ajoute pas de complexité spécifique en matière de traitement de la non-réponse). On trouve deux cas de figure :

• ***CAS 1 : on ne connaît pas la taille totale N_a du domaine*** **:**

Dans ces conditions, il faut estimer N_a par son estimateur sans biais naturel (généralement celui de Horvitz-Thompson), soit (voir 1.1)

$$\hat{N}_a = \sum_{i \in s_a} \frac{1}{\Pi_i}$$

puis

$$\boxed{\hat{\bar{Y}}_{a,R} = \hat{Y}_a / \hat{N}_a}.$$

L'estimateur est dit de type "ratio" - ce n'est donc plus l'estimateur de Horvitz-Thompson. On montre qu'il est biaisé, et que lorsque n est « grand » (ce qui est toujours le cas) son biais conditionnel est d'ordre de grandeur en $1/n_a$, soit :

$$E\left(\hat{\bar{Y}}_{a,R} \middle| n_a\right) - \bar{Y}_a = O\left(\frac{1}{n_a}\right)$$

On vérifie également que la variance conditionnelle varie en $1/n_a$:

$$V\left(\hat{\bar{Y}}_{a,R} \middle| n_a\right) = O\left(\frac{1}{n_a}\right)$$

De ce fait, l'erreur totale (conditionnelle) - ou erreur quadratique moyenne (EQM, voir introduction) - varie en $1/n_a$:

$$EQM\left(\hat{\bar{Y}}_{a,R}|n_a\right)=O\left(\frac{1}{n_a}\right)$$

On retrouve le problème traditionnel : si a est petit, n_a sera petit et $\hat{\bar{Y}}_{a,R}$ sera de mauvaise qualité. Comme dans le cas du total, on peut hésiter entre l'approche conditionnelle et l'approche non conditionnelle, mais la nature de la conclusion sur les ordres de grandeur restera la même. Cela étant, la ratio s'avère toutes choses égales par ailleurs plus efficace dès lors que Y_i varie peu autour de $\bar{Y}_a$. Si Y_i est constant dans le domaine, égal à $\bar{Y}_a$, alors quel que soit l'échantillon il est clair qu'on aura $\hat{\bar{Y}}_{a,R}=\bar{Y}_a$ l'erreur quadratique moyenne sera nulle, et on aura donc trouvé l'estimateur parfait ! C'est un cas de figure avantageux qui ne se présentait pas dans le cas de l'estimation du total, et qui peut redonner une légitimité à l'estimation directe d'une moyenne sous condition d'une faible dispersion des Y_i au sein du domaine.

Si le tirage est à probabilités égales et de taille fixe, alors

$$\hat{\bar{Y}}_{a,R}=\bar{y}_a$$

moyenne simple sur les n_a individus du domaine. Le cas du sondage aléatoire simple -qui est un cas particulier de tirage à probabilités égales et de taille fixe- est assez remarquable : en effet, un théorème (déjà signalé au 1.a.) nous dit qu'en matière de calcul de biais et de variance, **lorsque le sondage est aléatoire simple et que l'on conditionne par n_a, tout se passe comme si on effectuait un sondage aléatoire simple de taille n_a au sein du domaine.** De ce théorème, il découle immédiatement

$$\underset{SAS}{E}\left(\bar{y}_a\right)=\bar{Y}_a$$

• CAS 2 : on connaît la taille totale N_a du domaine :

On peut, cette fois, utiliser l'estimateur classique (de Horvitz Thompson) :

$$\hat{\bar{Y}}_a=\hat{Y}_a/N_a$$

On a $E\left(\hat{\bar{Y}}_a\right)=\bar{Y}_a$, donc un estimateur sans biais. Par ailleurs

$$V\left(\hat{\bar{Y}}_a\right)=\frac{1}{N_a^2}\cdot V\left(\hat{Y}_a\right)=O\left(\frac{1}{n_a}\right)$$

On obtient un estimateur qui a des propriétés identiques à celles de $\hat{Y}_a$. En particulier, la variabilité de n_a va entraîner un supplément de variabilité de $\hat{\bar{Y}}_a$ par rapport à un échantillonnage de taille fixe.

Cet estimateur $\hat{\bar{Y}}_a$ sera a priori moins performant que le ratio $\hat{\bar{Y}}_{a,R}$ si les Y_i sont peu dispersés, mais au contraire il sera plus efficace que le ratio si les Y_i et les Π_i sont à peu près proportionnels.

2. L'estimateur par la régression :

2.1. Formulation générale

Il s'agit d'exploiter de l'information auxiliaire pour diminuer la variance d'échantillonnage. On appelle information auxiliaire toute variable définie au niveau individuel, disponible pour chaque individu de l'échantillon et dont le vrai total défini sur la population d'inférence est connu. On note X_i cette information : a priori, il s'agit d'un vecteur à p composantes $\left(X_i \in \mathsf{R}^p\right)$, dont le vrai total X_a sur le domaine a est connu : on note

$$X_a = \sum_{i=1}^{N_a} X_i$$

On considère qu'il existe une liaison de type linéaire multivariée entre Y et les composantes de X pour tous les individus i de a (et seulement de a : il est clair que ce modèle ne concerne pas le reste de la population). On la formalise ainsi :

$$\forall i = 1,2,\ldots,N_a : \qquad Y_i = B_a^T \cdot X_i + U_i$$

où $B_a^T = \left(B_a^1, B_a^2, \ldots, B_a^p\right)$ est un vecteur de p coefficients réels inconnus. On prendra le parti de manipuler des vecteurs colonne, B^T désigne donc le vecteur (ligne) transposé de B. La variable U_i est un résidu d'ajustement, auquel on attache une "variance" qui représente la qualité de la liaison entre Y_i et X_i : ce n'est pas traité ici comme une variance de modèle au sens classique de la statistique mathématique (auquel cas Y_i serait aléatoire -et ce n'est pas l'optique retenue), mais disons que c'est « quelque chose » qui est d'autant plus grand que Y_i est susceptible "a priori" de s'éloigner de $B_a^T X_i$, avec X_i fixé (dans un hypothétique exercice de prédiction de Y_i, on pourrait parler d'un indicateur quantifiant l'incertitude). Il est vrai que les U_i peuvent prendre n'importe quelle valeur, de manière non contrainte ; cependant l'approche trouve sa logique et se justifie intuitivement d'autant mieux que les U_i sont petits, c'est-à-dire que les X_i « expliquent bien » Y_i. On se placera donc naturellement dans ce cas.

Ainsi, dans l'esprit qui vient d'être donné, si on considère que $Var\, U_i = \sigma_i^2$, paramètre inconnu, et qu'il y a « indépendance » entre les différents U_i, alors la matrice de (pseudo) variance du vecteur des U_i s'écrit

$$Var\begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_N \end{pmatrix} = \Omega = \begin{pmatrix} \sigma_1^2 & & o \\ & \sigma_2^2 & \\ & \ddots & \\ o & & \sigma_N^2 \end{pmatrix}$$

On montre que le "meilleur" vecteur d'ajustement B (meilleur au sens des moindres carrés)[7], défini dans la population du domaine a, est :

$$\tilde{B}_a = \left(\sum_{i=1}^{N_a} \frac{X_i \cdot X_i^T}{\sigma_i^2}\right)^{-1} \cdot \left(\sum_{i=1}^{N_a} \frac{X_i \cdot Y_i}{\sigma_i^2}\right)$$

[7] On minimise $\left(Y - X^T B\right)^T \Omega^{-1} \left(Y - X^T B\right)$, qui représente une distance entre Y et $X^T B$ où Y est le vecteur colonne des Y_i, et X^T la matrice de taille N X p où les X_i^T constituent les vecteurs ligne.

Cette expression n'a rien d'aléatoire et doit être vue comme la solution d'un problème de géométrie : trouver, dans un sous-espace de $\mathbb{R}^{N_a}$ à p dimensions, le vecteur qui soit « au plus proche » du vecteur Y.

S'agissant d'un ratio de deux sommes inconnues, on l'estime à partir des données collectées sur s_a selon :

$$\hat{B}_a = \left(\sum_{i \in s_a} \frac{X_i \cdot X_i^T}{\Pi_i \hat{\sigma}_i^2} \right)^{-1} \cdot \left(\sum_{i \in s_a} \frac{X_i \cdot Y_i}{\Pi_i \hat{\sigma}_i^2} \right)$$

où $\hat{\sigma}_i^2$ est un estimateur de σ_i^2. En pratique, on rencontre deux cas de figure : ou bien σ_i^2 est constant, ou bien il n'y a qu'une variable explicative $X_i \in \mathbb{R}$ et σ_i^2 est proportionnel à X_i, soit $\sigma_i^2 = \sigma^2 \cdot X_i$.

Dans le premier cas (que l'on appellera « cas de la régression »), on obtient

$$\hat{B}_a = \left(\sum_{i \in s_a} \frac{X_i X_i^T}{\Pi_i} \right)^{-1} \cdot \left(\sum_{i \in s_a} \frac{X_i Y_i}{\Pi_i} \right), \text{ soit } \hat{B}_a \in \mathbb{R}^P$$

Dans le second cas (que l'on appellera « cas du ratio »), on obtient :

$$\hat{B}_a = \frac{\sum_{i \in s_a} Y_i / \Pi_i}{\sum_{i \in s_a} X_i / \Pi_i} = \frac{\hat{Y}_a}{\hat{X}_a}, \text{ soit } \hat{B}_a \in \mathbb{R}$$

L'estimateur par la régression du total Y_a est défini par :

$$\boxed{\hat{Y}_{\mathrm{Re}g,a} = \hat{Y}_a + \hat{B}_a^T \cdot \left(X_a - \hat{X}_a \right)}$$

Dans le cas du ratio, on trouve :

$$\boxed{\hat{Y}_{Reg,a} = X_a \cdot \frac{\hat{Y}_a}{\hat{X}_a}}$$

qui est l'estimateur classique dit « par le ratio ».

L'estimateur par la régression a un biais qui varie en $1/n_a$, donc négligeable si n_a et grand (en revanche son biais peut être substantiel si n_a est petit, c'est-à-dire si on travaille sur de tout petits domaines). On ne sait pas exprimer son biais ni sa variance lorsque n_a est petit. Il n'existe évidemment pas de seuil indiscutable en deçà duquel le biais est négligeable, mais en ordre de grandeur il s'agit de plusieurs dizaines d'individus -mettons que si n_a dépasse 100 il n'y a pas trop à se soucier du biais.

C'est en considérant la variance d'échantillonnage de $\hat{Y}_{Reg,a}$ que l'on mesure bien l'avantage de cet estimateur. En effet, on montre que si n_a est « grand » - en cela il faut comprendre que l'on se situe dans les conditions de biais négligeable, donc n_a atteignant plusieurs dizaine d'individus - alors :

$$\boxed{V\left(\hat{Y}_{Reg,a}\right) \approx V\left(\hat{U}_a\right) = V\left(\sum_{i \in s_a} \frac{U_i}{\Pi_i} \right)}$$

où (rappel) $U_i = Y_i - \widetilde{B}_a^T \cdot X_i$. Ainsi, la variance de l'estimateur par la régression est (presque) égale à celle de l'estimateur classique du total d'une variable U_i définie comme le résidu de la régression linéaire de Y_i sur X_i. Si le vecteur d'informations auxiliaires X_i explique bien la variable d'intérêt Y_i, les résidus U_i seront numériquement petits, et donc mécaniquement la variance sera faible. Dans tous les cas, plan complexe ou pas, la variance varie en $1/n_a$: dans l'erreur totale, la variance prédomine donc (rappel : l'erreur quadratique moyenne est égale à la variance plus le carré du biais).

Si le tirage est aléatoire simple, la variance de $\hat{Y}_{\mathrm{Re}g,a}$ conditionnelle à n_a s'exprime selon :

$$V\left(\hat{Y}_{Reg,a}\middle|n_a\right) \approx \left(\frac{N \cdot n_a}{n}\right)^2 \cdot \left(1 - \frac{n_a}{N_a}\right) \cdot \frac{1}{n_a} \cdot S_{U,a}^2$$

$$\text{où } S_{U,a}^2 = \frac{1}{N_a - 1} \cdot \sum_{i=1}^{N_a} \left(U_i - \overline{U}_a\right)^2 \text{ et } \overline{U}_a = \frac{1}{N_a} \sum_{i=1}^{N_a} U_i$$

Si on connaît N_a, alors il faut inclure la variable qui vaut « 1 » partout dans le vecteur de régresseurs X_i (par construction, cette variable auxiliaire est toujours disponible, mais il faut connaître son vrai total dans le domaine - soit N_a - pour pouvoir la prendre en compte dans l'estimateur par la régression). Dans ces conditions, on peut simplifier l'expression de $S_{U,a}^2$ (c'est également vrai dans le cas du ratio puisque la somme des résidus est nulle) :

$$S_{U,a}^2 = \frac{1}{N_a - 1} \sum_{i=1}^{N_a} U_i^2$$

On peut en tirer immédiatement deux remarques :

- La variance varie en $1/n_a$: si le domaine est de petite taille, on peut s'attendre à une variance assez grande ;
- La variance reste directement liée à l'importance numérique des résidus U_i : si ces résidus sont petits, l'estimateur gagnera en efficacité. De petits résidus s'obtiennent en choisissant des variables X très explicatives.

Un autre résultat fondamental doit être rappelé : la performance d'un estimateur issu d'un calage (par exemple par CALMAR) sur X est, toujours pour n_a grand, identique à celle de l'estimateur par la régression sur X. Il ressort de tout cela que dans un domaine petit « mais pas trop », c'est-à-dire où n_a atteint par exemple une cinquantaine d'individus, un calage peut permettre de produire une estimation acceptable, mais seulement s'il porte sur des informations auxiliaires tout particulièrement explicatives de la variable d'intérêt Y. Si la taille n_a est vraiment petite, ou si on n'a pas une confiance absolue dans l'apport du calage, on obtiendra une estimation $\hat{Y}_{\mathrm{Re}g,a}$ de mauvaise qualité.

Cela étant, si la constante « 1 » fait partie du vecteur X_i (ou s'avère être combinaison linéaire des composantes de X_i), on peut montrer que l'on a toujours[8]

$$S_{U,a}^2 \leq S_a^2$$

si bien que l'estimateur par la régression est systématiquement (toujours sous condition que n_a soit « grand ») meilleur que l'estimateur de Horvitz-Thompson, même si le (pseudo) modèle de liaison entre X et Y donne lieu à des résidus U_i très grands (c'est-à-dire même si Y est extrêmement mal expliquée par les composantes de X) !!!

Dans ce cas, l'estimateur (légèrement biaisé) de la variance s'obtient selon :

$$\hat{V}\left(\hat{Y}_{Reg,a}\middle|n_a\right)=\left(\frac{N.n_a}{n}\right)^2\cdot\left(1-\frac{n_a}{N_a}\right)\frac{1}{n_a}\frac{1}{n_a-1}\sum_{i\in s_a}\hat{U}_i^2$$

où $\hat{U}_i = Y_i - \hat{B}_a^T X_i$

On notera que dans les deux cas habituels qui ont été cités, soit :

- les variances σ_i^2 sont constantes et la constante fait partie des régresseurs ou s'écrit comme combinaison linéaire des régresseurs

ou - les variances σ_i^2 sont proportionnelles à X_i, variable réelle,

on a $\hat{Y}_a = \hat{B}_a^T \cdot \hat{X}_a$, soit

$$\hat{Y}_{Reg,a} = \hat{B}_a^T \cdot X_a = \sum_{i=1}^{N_a}\left(\hat{B}_a^T \cdot X_i\right)$$

L'estimateur par la régression est donc égal à la somme des prédicteurs optimums relatifs aux N_a individus du domaine.

Pour estimer concrètement la variance d'un estimateur par la régression -ou d'un estimateur issu d'un calage- dès lors que n_a est « assez grand », il faut et il suffit de disposer d'un logiciel de calcul de variance et de travailler sur des variables individuelles égales aux résidus estimés

$$\hat{U}_i = Y_i - \hat{B}_a^T X_i$$

Pour terminer, revenons sur le problème du biais de $\hat{Y}_{\mathrm{Re}g,a}$ dans le cas où n_a est petit. Il existe une alternative pour échapper à ce biais, mais au prix d'une augmentation de variance : comme on chercher à éviter les grandes variances, cette piste risque bien de n'être qu'un leurre, aussi il n'est pas du tout évident qu'elle soit intéressante. Nous la signalons tout de même pour être complet. La méthode consiste tout simplement à utiliser l'estimateur par la régression construit à partir de l'ensemble de l'échantillon et appliqué à la variable d'intérêt $Z_i = Y_i \cdot 1_{i\in a}$, variable qui vaut donc Y_i si $i \in a$ et 0 sinon.

$$\tilde{\tilde{Y}}_{\mathrm{Re}g,a} = \hat{Z} + \hat{B}^T \cdot \left(X - \hat{X}\right) = \hat{Y}_a + \hat{B}^T.(X - \hat{X})$$

où $\hat{Z}$ et $\hat{X}$ sont les estimateurs de Horvitz-Thompson élaborés à partir de s et

$$\hat{B} = \left(\sum_{i\in s}\frac{X_i X_i^T}{\Pi_i}\right)^{-1}\cdot\left(\sum_{i\in s}\frac{X_i Z_i}{\Pi_i}\right)$$

[8] Plus exactement, on peut écrire $S_{U,a}^2 = \left(1-R^2\right)S_a^2$ où R est un coefficient dit « de détermination ».

On remarquera que $\hat{Z} = \hat{Y}_a$. Il s'agit de l'estimateur classique par la régression, qui estime le vrai total des Z_i sur la population entière, c'est-à-dire par construction le vrai total des Y_i sur le domaine. Il a un biais négligeable dès que la taille totale de l'échantillon n est grande : cette fois, il s'agit bien de la taille totale et non plus de n_a. On est donc assuré en pratique de l'absence de biais, même si n_a est extrêmement petit. En revanche, la variance est liée à l'importance des résidus $Z_i - B^T \cdot X_i$: or ceux-ci risquent bien d'être fortement négatifs puisqu'il y a une accumulation de valeurs zéro pour Z_i (pour tous les individus hors du domaine). On est donc dans une situation où la qualité de la relation linéaire entre les Z_i et les X_i peut être mauvaise (résidus forts, donc grande dispersion), même si les X_i expliquent bien les Y_i.

2.2 Formulations spécifiques :

L'expression la plus « célèbre » de l'estimateur par la régression est probablement celle qui correspond au modèle suivant, pour tout i du domaine a :

$$\boxed{Y_i = \sum_{h=1}^{H} \lambda_{ah} \cdot 1_{i \in h} + U_i}$$

avec $Var(U_i) = \sigma_h^2$ si $i \in h$. Les λ_{ah} (h varie de 1 à H) sont des réels inconnus caractéristiques de sous-populations clairement identifiées formant une partition du domaine, et les H variables auxiliaires $(1_{i \in h})$ sont les indicateurs d'appartenance à ces H sous-populations : ainsi, pour tout i et pour tout h, $1_{i \in h}$ vaut 1 si $i \in h$ et 0 sinon. Autrement dit, le modèle est : $Y_i = \lambda_{ah} + U_i$ lorsque i est dans h. On a, pour tout i de 1 à N_a :

$$\sum_{h=1}^{H} 1_{i \in h} = 1$$

Par analogie, on note n_{ah} la taille de l'échantillon recoupant à la fois le domaine a et la catégorie h. On note, pour tout h

$$\sum_{i=1}^{N_a} 1_{i \in h} = N_{ah}$$

Ainsi, N_{ah} représente la taille de la sous-population h qui recoupe le domaine a.

Ce modèle peut aussi s'écrire sous forme matricielle, en posant :

$$B_a^T = (\lambda_1, \lambda_2, \ldots, \lambda_H)$$

$$\text{et} \quad X_i^T = (1_{i \in 1}, 1_{i \in 2}, \ldots, 1_{i \in H})$$

Si pour tout h on a $n_{ah} > 0$, on vérifie très facilement que :

$$\forall h \qquad \hat{\lambda}_{ah} = \frac{\sum_{\substack{i \in s_a \\ i \in h}} Y_i / \Pi_i}{\sum_{\substack{i \in s_a \\ i \in h}} 1 / \Pi_i} = \frac{\hat{Y}_{ah}}{\hat{N}_{ah}} = \hat{\bar{Y}}_{ah}$$

$\hat{\lambda}_{ah}$ est l'estimateur naturel de la moyenne $\overline{Y}_{ah}$ dans la population h qui recoupe a. Alors on est dans les conditions où on peut écrire

$$\hat{Y}_{\mathrm{Re}g,a} = \sum_{i=1}^{N_a}\sum_{h=1}^{H} \hat{\lambda}_{ah} \cdot 1_{i\in h} = \sum_{h=1}^{H} N_{ah} \cdot \frac{\hat{Y}_{ah}}{\hat{N}_{ah}}$$

Cet estimateur est connu sous le nom d'estimateur post-stratifié. Il permet un calage sur les valeurs N_{ah}. Les conditions d'utilisation sont cependant un peu contraignantes en ce sens où il faut $n_{ah} > 0$ pour pouvoir estimer λ_h. Cela n'est pas une condition évidente si le domaine a est de petite taille et si H est « assez grand ».

Si la population n'est pas découpée en catégories $(H = 1)$ alors $\hat{Y}_{\mathrm{Re}g,a} = N_a \cdot \frac{\hat{Y}_a}{N_a}$

Dans le cas du sondage aléatoire simple, on obtient $\hat{Y}_{reg,a} = \sum_{h=1}^{H} N_{ah} \cdot \overline{y}_{ah}$, où $\overline{y}_{ah}$ désigne la moyenne simple dans $s_a \cap h$ et sa variance conditionnelle vaut approximativement (n_a « grand ») :

$$V\left(\hat{Y}_{Reg,a} \middle| n_a\right) \approx N_{ah}^2 \cdot \left(1 - \frac{n_a}{N_a}\right)\frac{1}{n_a} \cdot \sum_{h=1}^{H} \frac{N_{ah}}{N_a} S_{ah}^2$$

$$\text{où } S_{ah}^2 = \frac{1}{N_{ah} - 1}\sum_{i=1}^{N_{ah}} \left(Y_i - \overline{Y}_{ah}\right)^2$$

représente la dispersion naturelle des Y_i dans la sous-population h qui recoupe le domaine a. La dispersion des résidus est devenue ici une moyenne pondérée des dispersions des Y_i à l'intérieur de chacune des sous-populations (au sein du domaine, toujours). Si Y_i prend des valeurs caractéristiques des différentes sous-populations impliquées, c'est-à-dire si les Y_i sont peu dispersés au sein de chaque sous-population (recoupant le domaine), l'estimateur précédent sera efficace. L'idéal est constitué par des Y_i constants dans chaque sous-population : imaginons le cas où Y_i vaut α si i est un homme ($h = 1$) et β si i est une femme ($h = 2$). Alors

$$Y_a = N_{a1} \cdot \alpha + N_{a2} \cdot \beta$$

Or on aura toujours

$$\hat{\alpha} = \frac{\sum_{i\in s_a}^{i\ \mathrm{hom}me} \alpha/\Pi_i}{\sum_{i\in s_a}^{i\ \mathrm{hom}me} 1/\Pi_i} = \alpha$$

et de même $\hat{\beta} = \beta$.

Donc finalement $\hat{Y}_a = N_{a1}\hat{\alpha} + N_{a2}\hat{\beta} = Y_a$ et cela quel que soit l'échantillon tiré : l'estimateur est donc parfait dans ce cas (évidemment extrême).

Le modèle précédent peut être enrichi (et donc généralisé) si on dispose, non plus nécessairement des tailles de sous-population N_{ah}, mais de « vrais » totaux X_{ah} définis à partir d'une variable auxiliaire X_i non constante. Il arrive que dans ces conditions, on postule un modèle de proportionnalité entre Y_i et X_i en laissant le coefficient de proportionnalité dépendre de la sous-population et en adoptant l'hypothèse où la variance du résidu est proportionnelle à X_i. Cela donne, pour tout i de a :

$$Y_i = \sum_{h=1}^{H} \lambda_{ah} \cdot \left(X_i 1_{i\in h}\right) + U_i$$

avec $Var(U_i) = \sigma_h^2 . X_i$ si $i \in h$. On remarquera que la variance des U_i fait intervenir des σ_h^2 qui peuvent varier d'une sous-population à l'autre. Alors on vérifie, toujours sous la contrainte que $n_{ah} > 0$ pour tout h,

$$\hat{\lambda}_{a,h} = \frac{\hat{Y}_{ah}}{\hat{X}_{ah}}$$

Donc

$$\boxed{\hat{Y}_{Reg,a} = \sum_{i=1}^{N_a} \sum_{h=1}^{H} \hat{\lambda}_{a,h}\left(X_i 1_{i\in h}\right) = \sum_{h=1}^{H} X_{ah} \cdot \frac{\hat{Y}_{ah}}{\hat{X}_{ah}}}$$

Si $X_i = 1$, on retrouve évidemment l'estimateur post-stratifié. Cet estimateur est une somme d'estimateurs de type ratio, qui cale sur les H valeurs X_{ah}. Il est recommandable si, au sein de chaque sous-population, il y a une proportionnalité « suffisamment forte » entre X et Y (dans le cas contraire, il peut être de qualité médiocre).

3. L'estimateur par calage

Il s'agit de voir dans quelle mesure un calage peut améliorer la qualité d'un estimateur direct : a priori, on part de l'estimateur de Horvitz-Thompson et on détermine de nouveaux poids qui permettent d'estimer parfaitement bien certains totaux connus[9]. L' estimateur par la régression, abordé dans la partie précédente, est un cas particulier d'estimateur par calage. Il y a deux façons de concevoir le calage :

- Soit on cale sur des totaux (structures en effectifs pour des variables qualitatives auxiliaires, vrais totaux pour des variables quantitatives) obtenus au niveau de la population tout entière, et on utilise ensuite les poids calés en se restreignant aux individus du domaine (exactement comme on le fait au 1.1 avec les poids bruts sans biais).
- Soit on se retreint d'emblée aux individus du domaine et on cale sur des marges calculées au niveau du domaine.

On peut d'ailleurs mixer les contextes, et se caler à la fois sur des totaux « nationaux » et sur des totaux « locaux » (il suffit de travailler avec les bonnes variables en utilisant judicieusement des variables indicatrices d'appartenance au domaine).

Le premier contexte est a priori peu efficace : la qualité d'un estimateur calé est en effet étroitement liée aux valeurs des résidus de la régression de la variable d'intérêt sur les variables auxiliaires de calage. Or, la variable d'intérêt intervenant dans les résidus n'est plus Y_i mais $Y_i \cdot 1_{i\in a}$ où $1_{i\in a}$ est la variable indicatrice d'appartenance au domaine. Ces conditions sont

[9] Pour en savoir plus sur le calage, voir un ouvrage généraliste sur les sondages et/ou la documentation utilisateur Calmar sur le site Web de l'Insee.

défavorables à la qualité de la régression parce qu'il y a une forte concentration de valeurs expliquées égales à 0, et donc des résidus extrêmement dispersés, par construction. Il faut donc s'attendre à ce que l'amélioration de variance soit faible, voire inexistante (on peut même imaginer une dégradation de la qualité avec des enquêtes à probabilités fortement inégales). Une illustration concrète de ce phénomène concerne l'enquête « Emploi en continu » : voir l'article (« *Calcul de précision dans l'enquête emploi en continu* », P. Ardilly, G. Osier, Actes des JMS 2004). D'ailleurs, la pratique d'un logiciel comme Calmar laisse percevoir ce risque : on sait que le logiciel distord certains poids de manière importante et imprévisible pour parvenir à satisfaire les équations de calage. Aussi, si le domaine est petit, on peut très facilement imaginer que les poids calés de quelques individus échantillonnés du domaine soient suffisamment éloignés de leurs poids d'origine pour que l'estimation d'ensemble soit fortement modifiée. Notez que l'estimateur calé obtenu avec cette méthode utilise l'ensemble de l'information auxiliaire sur l'ensemble de la population, donc bien au-delà de a. De ce fait, il ne s'agit pas d'un estimateur direct.

Le second contexte est beaucoup plus sympathique, parce qu'il correspond aux conditions « habituelles » du calage et qu'il permet ainsi de tirer profit des ajustements sur des variables bien corrélées à la variable d'intérêt. En effet, les poids d'origine sont sans biais si on se restreint à s_a et les vrais totaux sur lesquels on se cale concernent bien la population d'inférence associée à s_a, soit a. Mais il y a deux obstacles majeurs : d'une part il faut que la taille de l'échantillon n_a ne soit « pas trop petite » (faute de quoi la théorie perd ses fondements - en particulier on récupère un biais incontrôlé), et d'autre part - et surtout - il faut évidemment disposer de marges ou de totaux au niveau du domaine, ce qui est en pratique une situation rare lorsqu'on veut se caler sur une information actualisée[10]. Un calage sur une seule variable X va nous ramener exactement aux conditions de l'estimateur par la régression du 2.1.

Les calages peuvent certes améliorer de manière très sensible la qualité des estimateurs. Cela étant, on ne peut pas leur demander d'éponger toutes les « insuffisances » du plan de sondage initial : si la taille d'échantillon n_a est petite, le miracle ne se produira pas et le calage ne suffira pas à sauver la mise...

[10] Echappatoire : on peut toujours se caler sur des informations anciennes, par exemple issues d'un recensement.

Estimation indirecte avec modélisation implicite

1. Principe de base de cette approche.

Tous les estimateurs directs (partie précédente) ont des variances qui évoluent en $1/n_a$ - y compris l'estimateur par la régression dans les cas, pourtant très favorables, où le (pseudo) modèle reflète bien la réalité. Cette propriété est source d'une imprécision souvent inacceptable, car si le domaine est de petite taille (N_a petit), n_a sera (trop) petit et la variance sera (trop) grande. L'idée générale qui justifie toutes les méthodes de ce chapitre consiste à **postuler une égalité entre un paramètre défini sur le domaine et le paramètre de même nature défini cette fois sur la population globale**. L'exemple le plus simple consiste à croire que la vraie moyenne sur la population complète coïncide avec la vraie moyenne sur le domaine. Estimer la moyenne sur le domaine revient alors à estimer la moyenne sur la population entière. Or l'estimateur de la moyenne globale peut-être considéré a priori comme fiable si la taille de l'échantillon global n est grande (dans toute la suite, on considérera effectivement que n est grand). Cette condition est toujours réalisée en pratique, si bien que le problème de la variance est réglé. Bien évidemment, il y a une contrepartie à ce type d'hypothèse : si elle est fausse (ce qui est probablement le cas, en toute rigueur…) il y aura un biais de l'estimateur, car cet estimateur sera en moyenne égal au paramètre défini sur la population, et non plus au paramètre défini sur le domaine. Sur le plan stratégique, toute la question se ramène donc à un problème de vases communicants : préfère-t-on peu de variance et un risque de biais ou au contraire plus de variance en l'absence de biais ?

L'aspect pratique peut se heurter à un obstacle majeur, qui est celui de la disponibilité d'une information auxiliaire pertinente. En effet, le secret d'une bonne estimation sur petits domaines réside en grande partie dans l'utilisation intelligente d'une information externe, pour une simple raison de bon sens : dans un processus d'estimation, lorsque le contexte initial est pauvre en information, on ne peut (évidemment) pas créer de la précision ex nihilo, il faut bien aller chercher « en dehors de l'enquête » ce que le processus de collecte initial n'a pas été capable d'apporter par lui-même ! C'est pourquoi on a déjà parcouru une grande partie du chemin lorsqu'on est parvenu à rassembler l'information auxiliaire adéquate. Cela crée deux difficultés : d'une part il faut que l'information explicative existe effectivement, et d'autre part il faut des moyens humains experts pour l'exploiter de manière pertinente. Le premier aspect ne va pas de soi : dans le domaine socio-démographique par exemple, les sources externes susceptibles de donner une information « exacte » sur des petits domaines ne sont pas si nombreuses. On pense essentiellement au recensement, mais il comporte peu de variables et comme il s'agit désormais d'enquêtes par sondage, on perd le caractère exhaustif et on ne peut plus le mobiliser à n'importe quel niveau[11]. Le second aspect renvoie aux moyens d'étude, parce que la pertinence du modèle dépendra étroitement de la connaissance que le statisticien aura du phénomène étudié. Il faut donc que ce dernier maîtrise à la fois les aspects explicatifs du phénomène et les aspects techniques de l'estimation.

[11] Par exemple au niveau régional il ne devrait pas y avoir de problème, au niveau d'une grosse agglomération non plus. Mais au niveau départemental par exemple, dans une zone d'emploi ou sur une agglomération « moyenne », c'est beaucoup plus discutable.

2. L'estimation synthétique.

2.1. En l'absence d'information auxiliaire.

Il s'agit du cas où on ne connaît aucun vrai total sur le petit domaine a, sauf éventuellement sa taille N_a. Un estimateur envisageable de la moyenne $\overline{Y}_a$ est

$$\boxed{\hat{\overline{Y}}_{a,SYN} = \hat{\overline{Y}} = \frac{\hat{Y}}{\hat{N}}}$$

où $\hat{Y} = \sum_{i \in s} \frac{Y_i}{\Pi_i}$ et $\hat{N} = \sum_{i \in s} \frac{1}{\Pi_i}$. On notera bien que les sommations portent sur l'échantillon s dans son ensemble (et non plus sur s_a).

Cet estimateur est dit de type « synthétique ». Il peut être calculé pour des domaines de taille aussi petite que l'on veut : il conserve d'ailleurs tout son sens si $n_a = 0$. Evidemment, il est particulièrement « brutal » en ce sens où il ne permet pas de traduire la moindre spécificité locale : s'il y a plusieurs domaines étudiés, l'estimation diffusée sera la même partout.

On rappelle que $\hat{\overline{Y}}$ représente l'estimateur classique de la vraie moyenne $\overline{Y}$ définie sur l'ensemble de la population (estimateur de type « ratio »). Comme n est grand, on a :

$$E\left(\hat{\overline{Y}}\right) \approx \overline{Y}$$

et donc

$$BIAIS = E\left(\hat{\overline{Y}}_{a,SYN}\right) - \overline{Y}_a \approx \overline{Y} - \overline{Y}_a$$

On voit apparaître très clairement l'hypothèse à formuler pour éviter le biais :

$$\overline{Y}_a = \overline{Y} \qquad (1)$$

C'est le modèle implicite qui sous tend l'utilisation de $\hat{\overline{Y}}_{a,SYN}$. Cette condition, portant directement sur les paramètres, peut aussi résulter d'une logique qui considère que tous les Y_i sont des réalisations mutuellement indépendantes d'une variable aléatoire suivant une loi donnée (quelconque). Si on appelle μ l'espérance de cette loi, on aurait donc $E(Y_i) = \mu$ pour tout i de la population, que l'individu soit ou non dans le domaine. C'est une hypothèse équivalente que l'on fait très souvent quand on pratique l'imputation pour corriger la non-réponse partielle, c'est-à-dire qu'on considère que les répondants et les non-répondants ont des comportements semblables[12]. Et bien, dans le cas présent, il s'agit de nouveau de cette philosophie, on considère que les individus du domaine et ceux qui n'y sont pas ont un comportement de même nature. On formalise ainsi :

$$Y_i = \mu + e_i$$

où e_i est une variable aléatoire d'espérance nulle. Par la loi des grands nombres[13], appliquée deux fois :

$$\overline{Y}_a = \frac{1}{N_a} \sum_{i=1}^{N_a} Y_i \approx E\left[Y_i \mid i \in a \right] = EY_i = \mu \approx \frac{1}{N} \sum_{i=1}^{N} Y_i = \overline{Y}$$

[12] On parle alors de mécanisme de réponse ignorable.

[13] Même pour un petit domaine, on considérera N_a comme grand.

où l'espérance E doit être ici comprise par rapport à la loi des Y_i. Notez que par hypothèse, l'espérance conditionnelle au fait que i soit dans le domaine est la même que l'espérance de Y pour tout individu de la population : autrement dit, les individus du domaine ne présentent pas de spécificité liée à Y.

Dans ces conditions, l'erreur quadratique moyenne est

$$EQM\left(\hat{\bar{Y}}_{a,SYN}\right) = \underbrace{V\left(\hat{\bar{Y}}\right)}_{\text{varie en } 1/n} + \underbrace{\left(\bar{Y}_a - \bar{Y}\right)^2}_{\text{ne dépend pas de n}}$$

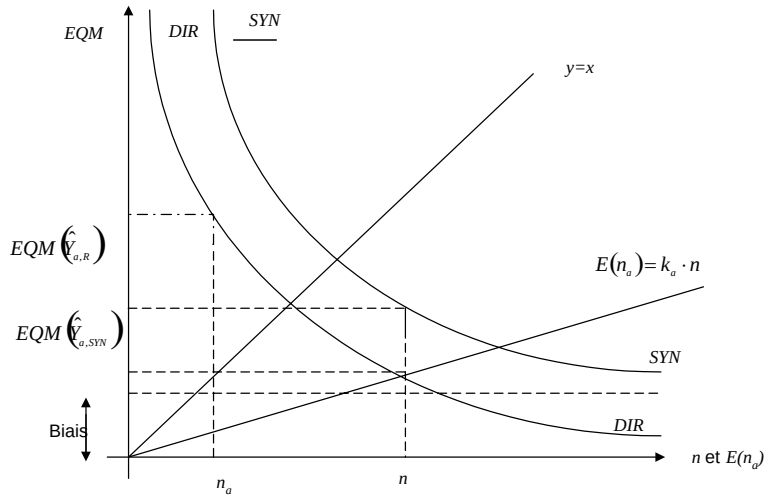
On peut penser, si n est grand et si le modèle (1) est correct (c'est surtout cette seconde condition qui en pratique est cruciale), **que cette erreur sera inférieure à celle des estimateurs directs (qui varie en $1/n_a$).**

Si on pratique un sondage aléatoire simple et si on néglige tous les taux de sondage devant 1, les formules se simplifient : $\hat{\bar{Y}}_{a,SYN}$ est égal à $\bar{y}$, moyenne simple calculé sur l'ensemble des individus de s et

i) $EQM\left(\hat{\bar{Y}}_{a,SYN}\right) \approx \dfrac{S^2}{n} + \left(\bar{Y}_a - \bar{Y}\right)^2$

ii) $EQM\left(\hat{\bar{Y}}_{a,R}\right) = E\left(\dfrac{1}{n_a}\right) \cdot S_a^2 = \dfrac{\delta_a}{E(n_a)}$

où $\hat{\bar{Y}}_{a,R}$ est l'estimateur direct $\hat{Y}_a / \hat{N}_a$ rencontré au 1.2 du chapitre précédent. et δ_a est un coefficient complexe. On peut ainsi visualiser les avantages et les inconvénients offerts par chaque estimateur. On peut en particulier illustrer le dilemme par le graphique suivant, qui représente les évolutions des EQM des deux estimateurs respectifs en fonction de n (pour l'estimateur synthétique -courbe notée SYN) et de n_a (pour l'estimateur direct -courbe notée DIR).



Lecture du graphique : pour un n donné, on en déduit $E(n_a)$ (puisqu'on a une relation du type $E(n_a) = k_a \cdot n$, avec $k_a = N_a / N$), que l'on peut reporter sur l'axe des abscisses grâce à une diagonale $(y = x)$. On lit alors directement en ordonnées les erreurs des deux estimateurs.

Plus le domaine est petit, plus N_a -et donc k_a- sera faible, et plus, à n donné, $E(n_a)$ sera repoussé vers la gauche. Cela jouera en faveur de $\hat{\bar{Y}}_{a,SYN}$. Mais plus le biais sera grand, plus la courbe SYN sera décalée vers le haut par rapport à la courbe DIR : cela jouera, à partir d'un certain seuil, en faveur de $\hat{\bar{Y}}_{a,R}$. Pour un biais fixé, on voit bien qu'il existe « quelque part » une valeur limite n_L de la taille n au-delà de laquelle l'estimateur synthétique sera moins bon que l'estimateur direct, et en deçà de laquelle, au contraire, il vaut mieux utiliser l'estimateur synthétique (n_L serait nettement à droite sur l'axe des abscisses, en dehors de la figure ci-dessus !). Cette valeur n_L croît si le biais décroît, toutes choses égales par ailleurs (cela se comprend bien : à la limite, si le biais est nul -ce qui constitue la situation idéale- l'estimateur synthétique est absolument imbattable puisqu'en terme de variance il est indiscutablement le meilleur). De même n_L croît si k_a diminue, c'est-à-dire si le domaine est de plus en plus petit. Là encore il y a une logique : imaginons k_a très grand, proche par conséquent de 1, on a alors $E(n_a)=n$, et l'estimateur synthétique est systématiquement battu puisque les variances des deux estimateurs sont (à peu près) les mêmes mais l'estimateur synthétique reste pénalisé par son biais.

Evidemment, la grande question que l'on se pose dans ce contexte est : **peut-on évaluer le biais de l'estimateur synthétique** afin de pouvoir faire l'arbitrage en connaissance de cause ? La réponse générale à cette question est clairement **NON** : a priori, on ne peut pas évaluer le biais de $\hat{\bar{Y}}_{a,SYN}$, SAUF si on dispose d'une source externe qui permet d'estimer précisément $\bar{Y}_a$. Cette conclusion tient à deux remarques de simple bon sens. D'une part, si le biais était estimable, on corrigerait $\hat{\bar{Y}}_{a,SYN}$ en lui soustrayant cette estimation, et on travaillerait seulement à partir d'estimateurs sans biais : on ne manipulerait plus que des estimateurs synthétiques corrigés, ce qui n'est manifestement pas le cas… D'autre part et surtout, si un tel biais était estimable, on se demande bien à partir de quoi on pourrait en obtenir une estimation : d'une manière ou d'une autre, il faut bien passer par une estimation de $\bar{Y}_a$… et c'est justement ce que l'on ne peut pas faire de façon fiable. Donc le serpent se mord la queue ! L'espoir réside en fait dans l'utilisation d'une autre source qui, elle, fournirait les conditions adéquates pour estimer $\bar{Y}_a$, autrement dit il faudrait, soit un fichier administratif soit une enquête localisée de grande taille pour pouvoir faire confiance à un estimateur direct de $\bar{Y}_a$ - auquel cas autant utiliser directement cette source à la place de l'enquête, dont l'intérêt devient fort discutable. C'est pourquoi, dans de nombreux cas, l'estimation sur des petits domaines résulte d'un acte de foi.

Il y a néanmoins un espoir d'obtenir, dans certaines circonstances, une appréciation de l'ampleur du biais. Cela est possible si un recensement (ou une enquête de très grosse taille) fournit le vrai total Z_a sur le domaine a d'une variable Z (en général multivariée) reliée à Y par une relation que l'on maîtrise (en général linéaire), et que par ailleurs on estime ce total par $\hat{Z}_a$ selon la technique « petit domaine » (Z est présent dans la base de sondage de l'enquête, ou tout simplement collecté par voie de questionnaire). On peut alors constater le biais $\hat{Z}_a - Z_a$ et en déduire[14] un ordre de grandeur du biais $\hat{Y}_a - Y_a$. Cette technique ne permet en effet que de déplacer la difficulté, parce qu'elle n'a de sens que si on dispose d'une liaison fiable entre Y et Z et que cette liaison est valable sur l'ensemble de la population

[14] Si la liaison est linéaire (ce sera presque toujours le cas), le calcul est immédiat, sinon on utilise un développement limité - sauf si n_a est vraiment tout petit.

L'estimateur du total Y_a, dans cette optique apparaîtra plus sympathique si on connaît la vraie taille N_a. Dans ce cas :

$$\boxed{\hat{Y}_{a,SYN} = N_a \cdot \hat{\bar{Y}}_{a,SYN}}$$

Si on ne connaît pas N_a, on pourra l'estimer par $\hat{N}_a = \sum_{i \in s_a} \frac{1}{\Pi_i}$, qui est sans biais et dont la variance s'exprime en $1/n$ (et non en $1/n_a$, ce qui est fondamental). Par exemple avec du sondage aléatoire simple, $\hat{N}_a = (N/n) \cdot n_a$, puis

$$\hat{Y}_{a,SYN} = \frac{N}{n} n_a \cdot \bar{y}$$

$$EQM\left(\hat{Y}_{a,SYN}\right) = \text{ fonction de } \frac{1}{n} + N_a^2 \left(\bar{Y}_a - \bar{Y}\right)^2$$

Si l'erreur absolue ainsi définie est petite dès lors que le modèle implicite est exact, on aura en revanche une mauvaise surprise si on raisonne en terme d'erreur relative (coefficient de variation -voir introduction) car cette erreur relative sera forte si N_a est petit devant N (cas des petits domaines). Cette faiblesse provient de l'estimation de N_a par $\hat{N}_a$.

Une variante intéressante de toute cette approche consiste à moduler le périmètre de l'échantillon utile pour construire l'estimateur synthétique : en effet, on peut imaginer, si par exemple le domaine est un canton ou une agglomération, que le modèle implicite ne soit pas acceptable France entière mais qu'il soit par contre défendable sur une zone géographique « intermédiaire » entre le domaine proprement dit et l'ensemble du territoire couvert. Ainsi, on remplacerait $\hat{\bar{Y}}$ défini France entière par $\hat{\bar{Y}}_{DEP}$ construit à partir des individus échantillonnés dans le département auquel appartient le domaine, où par $\hat{\bar{Y}}_{REG}$ où ce serait cette fois l'intégralité de l'échantillon de la région qui serait mobilisée. C'est une question de compromis entre biais et variance : plus on étend la zone sur laquelle le modèle implicite est supposé valide, plus on risque d'augmenter le biais, mais plus la variance diminue en contrepartie. Le bon équilibre peut provenir de l'exploitation d'une source externe si on en dispose… et seulement d'une forte conviction dans la pertinence du modèle dans le cas contraire !

2.2. En présence d'information auxiliaire.

2.2.1. L'estimateur synthétique de type régression : formulation générale

Nous avons vu au chapitre I qu'en présence d'information auxiliaire, si on s'en tenait à des estimateurs directs, on avait toutes les chances d'améliorer la précision en utilisant un estimateur par la régression, exprimé ainsi :

$$\hat{Y}_{Reg,a} = \hat{Y}_a + \hat{B}_a^T \left(X_a - \hat{X}_a\right)$$

Cette formulation s'appuyait sur la liaison entre X_i et Y_i, que l'on supposait du type

$$Y_i = \tilde{B}_a^T X_i + U_i$$

pour tout i du domaine a, avec U_i petit. Le coefficient $\tilde{B}_a$ avait une expression spécifique qui lui conférait des propriétés d'optimalité. Imaginons maintenant que **cette liaison soit valable**

en dehors du domaine, c'est-à-dire en fait sur la population complète. Le vecteur $\tilde{B}_a$ est alors remplacé par un vecteur $\tilde{B}$ qui doit en être numériquement proche. Ainsi on distingue

$$\tilde{B}_a = \left(\sum_{i=1}^{N_a} \frac{X_i X_i^T}{\sigma_i^2} \right)^{-1} \cdot \left(\sum_{i=1}^{N_a} \frac{X_i Y_i}{\sigma_i^2} \right) \text{ (rappel)}$$

estimé par $\hat{B}_a$ et $$\tilde{B} = \left(\sum_{i=1}^{N} \frac{X_i X_i^T}{\sigma_i^2} \right)^{-1} \cdot \left(\sum_{i=1}^{N} \frac{X_i Y_i}{\sigma_i^2} \right)$$

naturellement estimé par $$\hat{B} = \left(\sum_{i \in s} \frac{X_i \cdot X_i^T}{\hat{\sigma}_i^2 \Pi_i} \right)^{-1} \cdot \left(\sum_{i \in s} \frac{X_i Y_i}{\hat{\sigma}_i^2 \Pi_i} \right)$$

Si on revient à l'expression $\hat{Y}_{Reg,a}$ re-écrite ainsi :

$$\hat{Y}_{Reg,a} = X_a^T \hat{B}_a + \left(\hat{Y}_a - \hat{B}_a^T \cdot \hat{X}_a \right)$$

on peut se dire que si la liaison linéaire postulée au sein du domaine a est effectivement correcte, le terme $\hat{Y}_a - \hat{B}_a^T \hat{X}_a$ doit être « assez petit », surtout comparativement à $X_a^T \hat{B}_a$ (le premier est de la nature d'une somme de résidus, le second est de la nature d'une somme de Y_i). On rappelle que dans la très grande majorité des cas concrets, soit σ_i^2 est constant et la constante fait partie de X , soit X est une variable réelle et σ_i^2 est proportionnelle à X : or, dans l'un ou l'autre de ces cas, on peut montrer que $\hat{Y}_a - \hat{B}_a^T \hat{X}_a$ est rigoureusement nul. Donc on peut considérer que $\hat{Y}_{Reg,a}$ doit être proche de $X_a^T \hat{B}_a$. Mais cette expression peut être assez instable (c'est-à-dire que sa variance d'échantillonnage peut être grande) à cause de la présence du $\hat{B}_a$ dont la variance s'exprime en $1/n_a$ (c'est un estimateur construit à partir de l'échantillon s_a seulement, donc il a une qualité en conséquence). Le raisonnement s'achève en exploitant l'hypothèse d'égalité des « vrais » coefficients de régression, à savoir

$$\tilde{B}_a = \tilde{B} ,$$

ce qui constitue le modèle implicite. Dans ces conditions, on va naturellement estimer Y_a par :

$$\boxed{\hat{Y}_{a,REGSYN} = X_a^T \hat{B}}$$

puisque le modèle implicite incite à substituer $\hat{B}$ à $\hat{B}_a$. L'estimation va varier avec le domaine considéré : elle prend donc en compte les spécificités locales (contrairement à l'approche du 2.1), mais en considérant que ces spécificités sont entièrement traduites par les variables auxiliaires.

On remarquera que dans les conditions « habituelles », on a $\hat{Y} = \hat{B}^T . \hat{X}$, si bien qu'on peut aussi écrire :

$$\hat{Y}_{a,REGSYN} = \hat{Y} . \frac{\hat{N}_a}{\hat{N}} + \hat{B}^T \left(X_a - \hat{X} . \frac{\hat{N}_a}{\hat{N}} \right)$$

C'est l'estimateur qu'on obtient si on effectue un calage sur les totaux X_a à partir de l'ensemble de l'échantillon s, et après une phase préliminaire où les poids « bruts » d_i du fichier national ont tous été remplacés par les poids $d_i.\frac{\hat{N}_a}{\hat{N}}$. On comparera cette expression avec celle du 2.2.3. L'intérêt de cette approche réside dans le fait qu'on travaille sur le fichier national sans faire à aucun moment de sélection d'individus et qu'elle correspond pratiquement à la mise en œuvre d'un programme de calage (comme Calmar par exemple).

Comme au 2.1., on peut avoir une vision duale en terme de modélisation des comportements individuels. Dans ces circonstances, il faut naturellement poser, pour tout i dans la population (que l'individu soit ou non dans a ne doit pas avoir d'influence sur la formulation du modèle)

$$Y_i = BX_i + e_i$$

où e_i est une variable aléatoire centrée, de variance σ_i^2. Sur la population complète comme sur n'importe quelle sous population, l'estimateur optimum de B (au sens des moindres carrés), non seulement est sans biais (c'est un résultat fondamental bien connu des statisticiens utilisant des modèles), mais également converge vers B lorsqu'il est calculé sur un grand nombre d'individus. Cela est a priori le cas, de manière certaine pour $\tilde{B}$ mais également pour $\tilde{B}_a$ dès lors que N_a est grand. On a donc

$$E\left(\tilde{B}\right) = E\left(\tilde{B}_a\right) = B$$
$$\tilde{B} \underset{N \to +\infty}{\to} B$$
$$\tilde{B}_a \underset{N_a \to +\infty}{\to} B$$

Le risque est néanmoins réel, même avec un modèle exact, lorsque le domaine a une taille N_a petite (donc pour les tout petits domaines).

Que peut-on dire sur la qualité de ce nouvel estimateur synthétique ?

Côté biais, on a :

$$\begin{aligned} BIAIS &= E\left(X_a^T \hat{B}\right) - Y_a \approx X_a^T \tilde{B} - Y_a \\ &= X_a^T \left(\tilde{B} - \tilde{B}_a\right) - \left(Y_a - X_a^T \tilde{B}_a\right) \end{aligned}$$

Cette expression très générale s'appuie sur l'égalité (approchée) $E\hat{B} \approx \tilde{B}$, qui traduit le caractère (approximativement) sans biais des coefficients de régression $\hat{B}$ dès lors que la taille de l'échantillon total s est grande (ce qui est notre cas - on peut donc, en pratique, utiliser une égalité). Le biais est donc composé de deux termes :

• $Y_a - X_a^T \tilde{B}_a$, qui est nul lorsque les variances individuelles sont constantes et que la constante fait partie du vecteur auxiliaire X - voire apparaît comme combinaison linéaire des composantes de X - OU dès que X est une variable réelle et que les variances individuelles sont proportionnelles à X. Ces deux cas correspondent à (pratiquement) tous les cas que l'on peut rencontrer en pratique, si bien qu'abstraction faite de modélisations exceptionnelles, on peut très raisonnablement considérer que ce terme est nul.

• $X_a^T\left(\tilde{B}-\tilde{B}_a\right)$: sa valeur est totalement conditionnée par la pertinence de la modélisation implicite : si on a eu « tort », c'est-à-dire si en réalité $\tilde{B}_a$ diffère sensiblement de $\tilde{B}$, le biais de l'estimateur synthétique de type régression peut être grand.

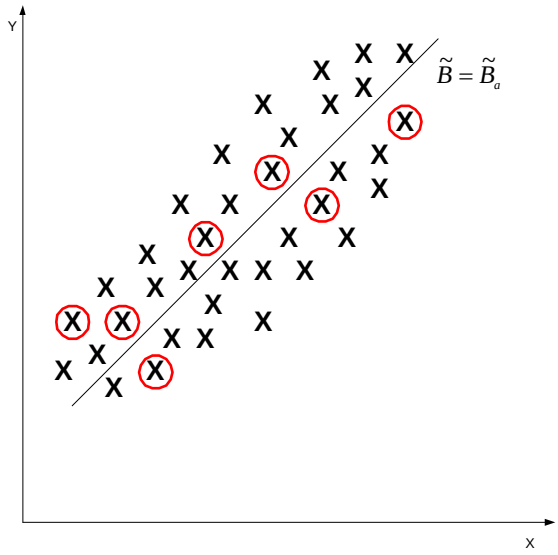
Du point de vue de la variance, on aura

$$V\left(\hat{Y}_{a,REGSYN}\right)=\text{fonction en } 1/n$$

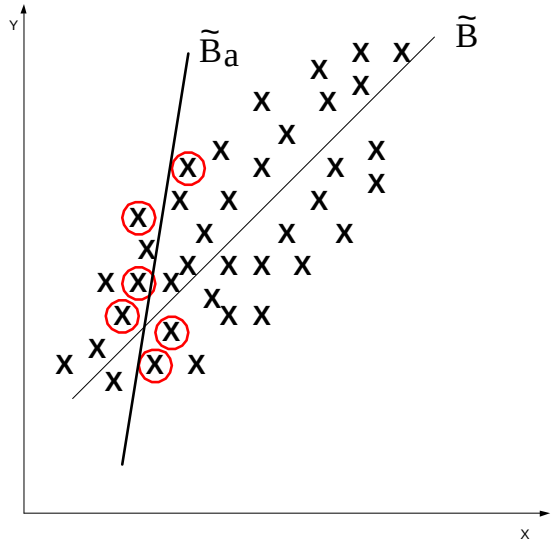
Cette fois, on bénéficie pleinement de la stabilité de $\hat{B}$: contrairement à $\hat{B}_a$, dont on rappelle que la variance est une fonction de $1/n_a$, $\hat{B}$ s'appuie sur l'échantillon s complet et cela va donc fortement limiter sa variabilité. Finalement,

$$EQM\left(\hat{Y}_{a,REGSYN}\right)\approx\left[X_a^T\left(\tilde{B}-\tilde{B}_a\right)\right]^2+\text{fonction de } 1/n$$

à comparer aux erreurs quadratiques moyennes de l'estimateur direct $\hat{Y}_{Reg,a}$ et de l'estimateur synthétique de base $\hat{Y}_{a,SYN}$. Graphiquement, on peut caractériser facilement les contextes à partir du nuage des N points $\left(X_i,Y_i\right)$ des individus de la population, dans lequel on a mis en évidence (points entourés d'un cercle) les N_a individus du domaine a. Le contexte 1 traduit le cas favorable, le contexte 2 le cas défavorable.



Contte 1 : cas favorableex



Contexte 2 : cas défavorable

La question sur l'appréciation a priori de la pertinence du modèle se pose exactement comme au 2.1., et la réponse est la même : seule une source externe permet de juger de l'importance du biais (donc du risque encouru). En l'absence de telle source, la question se résume à un acte de foi !

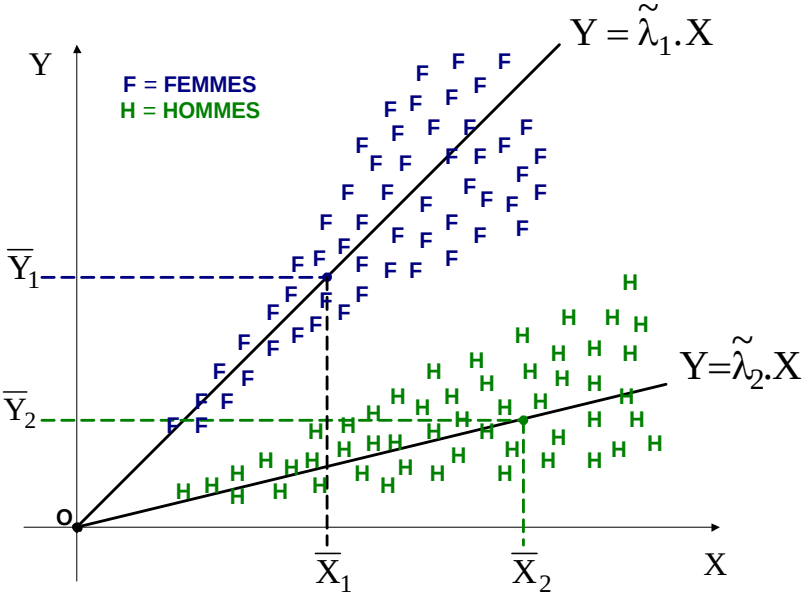
En pratique, comme au 2.1, on examinera avec intérêt toute variante consistant à moduler le périmètre de l'échantillon utile pour appliquer le modèle, afin de trouver le meilleur compromis entre biais et variance (si par exemple le domaine est un canton ou une agglomération, on pourra construire l'estimateur synthétique de type régression sur une zone géographique « intermédiaire » entre le domaine proprement dit et l'ensemble du territoire couvert).

2.2.2. Quelques déclinaisons de l'estimateur synthétique de type régression

On a essentiellement deux écritures particulières à présenter, correspondant au cas où X_i est un réel. Elles découlent - de manière tout à fait parallèle à ce que nous avions vu au 2.2. du chapitre I - d'une liaison de type :

$$Y_i = \sum_{h=1}^{H} \lambda_h \left(X_i 1_{i \in h} \right) + U_i$$

où $Var\, U_i = \sigma_h^2 \cdot X_i$ si $i \in h$. Ce modèle correspond à un partitionnement de la population globale en H catégories. Le schéma ci-dessous donne une image visuelle du contexte dans le cas de deux catégories (hommes et femmes par exemple) :



Dans ce cadre on a

$$\hat{\lambda}_h = \frac{\hat{Y}_h}{\hat{X}_h} = \left(\sum_{\substack{i \in s \\ i \in h}} \frac{Y_i}{\Pi_i} \right) \Big/ \left(\sum_{\substack{i \in s \\ i \in h}} \frac{X_i}{\Pi_i} \right)$$

Donc, si on pose $Z_i^h = X_i \cdot 1_{i \in h}$, on a

$$\hat{Y}_{a,REGSYN} = Z_a^T \hat{B} = \sum_{h=1}^{H} Z_a^h \cdot \hat{\lambda}_h$$

où $Z_a^h = \sum_{i=1}^{N_a} Z_i^h = \sum_{\substack{i=1 \\ i \in h}}^{N_a} X_i = X_{ah}$. D'où la première déclinaison

$$\boxed{\hat{Y}_{a,REGSYN} = \sum_{h=1}^{H} X_{ah} \cdot \frac{\hat{Y}_h}{\hat{X}_h}}$$

Le modèle implicite sous-jacent est le suivant : pour toute catégorie h,

$$\frac{\sum_{\substack{i \in h \\ i \in a}} Y_i}{\sum_{\substack{i \in h \\ i \in a}} X_i} = \frac{\sum_{i \in h} Y_i}{\sum_{i \in h} X_i}$$

Le schéma précédent est un exemple pour lequel il y aura un biais important.

La seconde déclinaison correspond au cas particulier $X_i = 1$: les totaux des X_i sont alors des tailles de population et on débouche sur

$$\boxed{\hat{Y}_{a,REGSYN} = \sum_{h=1}^{H} N_{ah} \frac{\hat{Y}_h}{\hat{N}_h}}$$

où $\hat{N}_h$ est l'estimateur de la taille de la sous-population h construit à partir de s et N_{ah} est la vraie taille de la sous-population croisant la catégorie h et le domaine a. Cette expression est probablement la plus célèbre et la plus utilisée de toutes celles qui permettent l'estimation sur petits domaines. On parle d'estimateur synthétique de type post-stratifié. La spécificité d'un domaine de taille donnée est donc entièrement traduite par un effet de structure : si le jeu de variables catégorielles qui permet de définir les catégories h confère à deux domaines a et b de même taille les mêmes structures, l'estimateur précédent ne les distinguera pas.

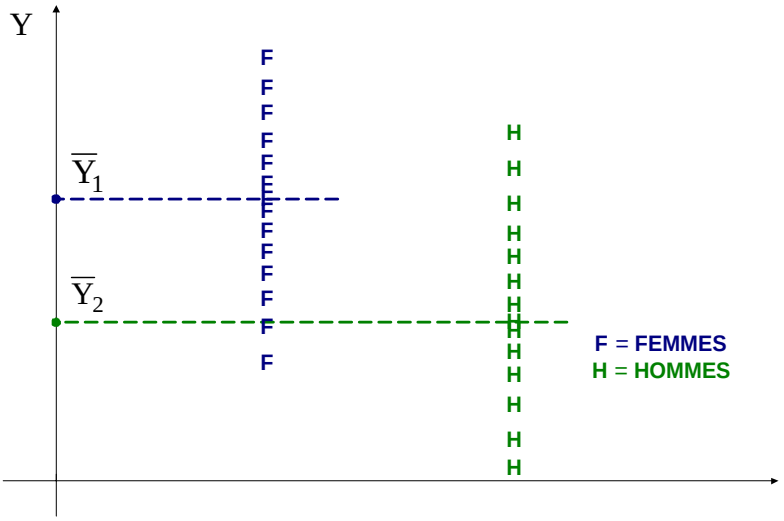
Cette expression suggère une mise en œuvre par un programme de calage. Dans le cas multivarié, on applique très facilement cette philosophie en partant du fichier complet et en effectuant un calage sur les marges **relatives au domaine** N_{ah} en ayant au préalable multiplié tous les poids individuels par $\hat{N}_a / \hat{N}$. Ainsi, on fournit à l'utilisateur final le fichier « national », pondéré de manière spécifique pour chaque domaine : par exemple s'il y a une estimation synthétique par région administrative, on utilisera 22 fois de suite Calmar, avec à chaque fois le fichier national en entrée mais 22 tables de marges différentes. On aboutira à 22 jeux de poids différents, mais les individus in fine pondérés seront toujours les mêmes - en l'occurrence l'ensemble des individus répondants au niveau national[15].

Si $H = 1$, on retrouve l'estimateur du 2.1. Le modèle implicite associé est le suivant : pour toute catégorie h

$$\overline{Y}_{ah} = \overline{Y}_h$$

autrement dit la moyenne des Y_i dans le domaine tout entier est égale à la moyenne des Y_i dans la fraction du domaine qui recoupe la catégorie h. Le schéma ci-dessous pourrait correspondre à un cas réel, distinguant deux catégories (hommes et femmes par exemple). L'estimateur synthétique serait biaisé dans ce contexte.

[15] Cela pourra poser des problèmes lors de la diffusion, car l'utilisateur peu averti ne comprendra pas le contexte ! Dans ce cas, il faudrait au moins supprimer toute trace d'appartenance géographique dans le fichier, afin primo que l'utilisateur ne soit pas surpris de traiter des données individuelles relatives à une région autre que la sienne, et secundo qu'il ne cherche pas à faire une exploitation infra régionale (par exemple départementale) qui aboutirait dans ce contexte à un non-sens.



La modélisation individuelle associée à ce modèle implicite est :

$$Y_i = \mu_h + e_i \text{ si } i \in h$$

C'est un modèle traditionnel d'analyse de la variance à un facteur.

Si l'échantillonnage est aléatoire simple, on obtient l'estimateur synthétique de type post-stratifié :

$$\hat{Y}_{a,REGSYN} = \sum_{h=1}^{H} N_{ah} \cdot \bar{y}_h$$

où $\bar{y}_h$ est la moyenne simple des n_h individus tirés (parmi n) qui se retrouvent dans la catégorie h. Dans ce contexte, l'erreur totale prend une forme assez simple :

$$EQM\left(\hat{Y}_{a,REGSYN}\right) = \left[\sum_{h=1}^{H} N_{ah}\left(\overline{Y}_h - \overline{Y}_{ah}\right)\right]^2 + \sum_{h=1}^{H} N_{ah}^2 \left(E\left(\frac{1}{n_h}\right) - \frac{1}{N_h}\right) S_h^2$$

où S_h^2 désigne la dispersion vraie des Y_i dans l'ensemble de la catégorie h. La condition de qualité se résume en une seule phrase : **l'estimateur synthétique sera d'autant meilleur que, dans chaque catégorie h, les valeurs Y_i seront peu dispersées autour de leur moyenne**. Cette règle simple traduit évidemment et de manière immédiate la condition S_h^2 petit, mais elle va également créer les conditions pour que les $\overline{Y}_{ah}$ soient proches des $\overline{Y}_h$ - conditions qui représentent a priori l'objectif le plus important puisque si n est grand, le risque numérique proviendra bien davantage du biais que de la variance. Pour le comprendre, il faut partir de la modélisation individuelle duale : en effet, avoir de faibles dispersions « descriptives » par catégorie revient à dire que les résidus U_i doivent être proches de zéro, c'est-à-dire que les variances de modèle σ_h^2 doivent être proches de zéro. Dans ce cas, dans chaque catégorie h, il y aura peu d'écart entre les moyennes $\overline{Y}_{ah}$ et $\overline{Y}_h$.

Cette règle de faible variance s'étend bien au modèle général du 2.2.1. Sur le plan théorique, c'est une condition suffisante (mais non nécessaire), mais c'est en pratique une condition quasiment nécessaire et suffisante, la raison étant que les tailles des petits domaines N_a restent limitées et rendent parfois discutable la qualité de la convergence de $\tilde{B}_a$ vers B. Aussi, il est difficile d'imaginer des situations concrètes où $\tilde{B} = \tilde{B}_a$ mais où les Y_i sont mal expliqués par les X_i dont on dispose : cela est bien sûr possible, mais ce serait vraisemblablement un effet curieux du hasard.

Une variante tout à fait intéressante de l'approche par post-stratification consiste à adapter $\hat{Y}_{a,REGSYN}$ de la manière suivante. Supposons qu'on ne dispose pas des totaux N_{ah} (aucune source externe ne permet de les obtenir, en tout cas pas de manière fiable), mais que l'on ait suffisamment de connaissance pour croire que la vraie structure N_{ah} / N_a définie dans le domaine a est très proche de la vraie structure définie sur un domaine D beaucoup plus grand que a et incluant a. Dans ce cas, on peut utiliser l'enquête pour estimer la structure N_{Dh} / N_D définie au sein de D puis estimer N_{ah} par $N_a . \frac{\hat{N}_{Dh}}{\hat{N}_D}$. Comme D est grand, la qualité de l'estimation des structures N_{Dh} / N_D sera bonne. L'estimateur finalement obtenue est :

$$\hat{\hat{Y}}_{a,REGSYN} = N_a . \sum_{h=1}^{H} \frac{\hat{N}_{Dh}}{\hat{N}_D} . \frac{\hat{Y}_h}{\hat{N}_h}$$

L'intérêt essentiel de cette approche est qu'elle ne nécessite aucune source d'information auxiliaire : l'enquête se suffit à elle-même et permet de construire entièrement l'estimateur, si bien qu'on peut définir - en théorie - des sous-populations h très pertinentes pour valider le modèle implicite $\overline{Y}_{ah} = \overline{Y}_h$. L'inconvénient - de taille - c'est qu'on déplace le problème de biais, dont l'ampleur repose désormais sur la pertinence de l'hypothèse $N_{ah}/N_a = N_{Dh}/N_D$ (que l'on ne peut pas davantage tester qu'avec le modèle initial, sauf à s'appuyer sur des éléments provenant d'autres études).

2.2.3. Un nouvel estimateur corrigeant le biais de l'estimateur synthétique

On l'a vu, le risque associé à $\hat{Y}_{a,REGSYN}$ est essentiellement celui d'un biais. On a dit aussi qu'on n'avait pas les moyens (sauf appel à l'aide d'une source externe) de valider le modèle implicite : cela ne signifie pas qu'on ne peut pas estimer le biais (voir 2.3), cela signifie que l'estimation du biais est trop imprécise pour qu'on puisse en tirer une conclusion sur la validité du modèle. On peut donc essayer de soustraire à $\hat{Y}_{a,REGSYN}$ un estimateur de son biais, ce qui donne (on se place dans les conditions - quasi-systématiques - où $\hat{Y}_a = \hat{X}_a^T \hat{B}_a$) :

$$X_a^T \hat{B} - \hat{X}_a^T . \left(\hat{B} - \hat{B}_a \right)$$
$$= \hat{X}_a^T \hat{B}_a + \hat{B}^T \left(X_a - \hat{X}_a \right)$$

On va donc s'intéresser au nouvel estimateur :

$$\boxed{\tilde{\hat{Y}}_{a,REGSYN} = \hat{Y}_a + \hat{B}^T \left(X_a - \hat{X}_a \right)}$$

soit $$\tilde{\hat{Y}}_{a,REGSYN} = \hat{Y}_{a,REGSYN} + \left(\hat{Y}_a - \hat{B}^T . \hat{X}_a \right)$$

On remarque qu'il s'agit de l'expression de l'estimateur direct par la régression (partie 2.1. du chapitre I) dans laquelle on a remplacé $\hat{B}_a$ par $\hat{B}$. Cet estimateur ne peut pas être calculé lorsque $n_a = 0$, contrairement à l'estimateur synthétique.

Comme c'était attendu, cet estimateur est sans biais si n est grand (car $E\hat{B} = \tilde{B}$, $E\hat{X}_a = X_a$ et $E\hat{Y}_a = Y_a$). Côté variance, $\tilde{\hat{Y}}_{a,REGSYN}$ étant égal à la somme de $\hat{Y}_{a,REGSYN}$, dont la variance évolue en

$1/n$, et de $\left(\hat{Y}_a - \hat{B}^T \hat{X}_a\right)$, dont la variance évolue en $1/n_a$ (c'est l'ordre de grandeur des variances de $\hat{X}_a$ et de $\hat{Y}_a$), on aura

$$V\left(\tilde{\tilde{Y}}_{a,REGSYN}\right) \approx V\left(\hat{Y}_a - \hat{B}^T \cdot \hat{X}_a\right)$$
$$\approx V\left(\hat{Y}_a - \tilde{\tilde{B}}^T \hat{X}_a\right)$$
$$\approx V\left(\hat{U}_a\right)$$

où $Y_i = \tilde{\tilde{B}}^T X_i + U_i$. Comme $\hat{U}_a$ fait intervenir les individus de s_a, on confirme que sa variance est en $1/n_a$: on a donc reperdu l'avantage acquis par l'approche synthétique !

Cela étant, les circonstances peuvent amener à préférer $\tilde{\tilde{Y}}_{a,REGSYN}$ à l'estimateur direct par la régression $\hat{Y}_{a,\mathrm{Re}g}$. En effet, ce dernier peut avoir un biais fort (voire très fort) si la taille d'échantillon dans le domaine n_a est vraiment petite (moins de 30 individus par exemple), c'est-à-dire dans le cas de tout petits domaines. Dans ces circonstances, $\tilde{\tilde{Y}}_{a,REGSYN}$ conserve un biais nul ou presque. Mais il faut aussi s'intéresser aux variances respectives de ces deux estimateurs : a priori, celle de l'estimateur direct devrait être plus faible parce qu'elle fait intervenir des résidus construits à partir du coefficient $\tilde{B}_a$, mieux adapté que le $\tilde{B}$ (qui intervient pour sa part dans les résidus associés à $\tilde{\tilde{Y}}_{a,REGSYN}$).

Si on compare cette fois à l'estimateur synthétique, il n'y a espoir de privilégier $\tilde{\tilde{Y}}_{a,REGSYN}$ que si on suspecte un biais fort de $\hat{Y}_{a,REGSYN}$. Ce cas est celui des domaines très « atypiques ».

En conclusion, l'estimateur $\tilde{\tilde{Y}}_{a,REGSYN}$ doit être mis en balance avec ses concurrents plutôt dans le cas où on a à faire à un domaine à la fois tout petit et très atypique.

2.2.4. Estimation d'effectifs par la méthode de préservation des structures

On s'intéresse, dans cette partie, à la problématique suivante. On considère m domaines (de même nature) définis dans une population. A une date t, on cherche, dans le domaine a, à estimer les effectifs vérifiant les modalités u d'une variable qualitative X donnée, alors que l'on dispose :

- Des effectifs, dans chacun des m domaines, qui vérifiaient, à une date antérieure t_0, à la fois la modalité u de X et la modalité v de Z, où Z est une seconde variable qualitative. Eventuellement ces effectifs sont estimés, mais avec une très bonne précision.
- A la date t -donc au moment de l'enquête- des estimations fiables des effectifs croisant u et v, mais seulement sur l'ensemble des m domaines (et non pas sur chaque domaine).

On note $N_{a,uv}$ le nombre d'individus du domaine a qui vérifient à la fois la modalité u de X et la modalité v de Z, à la date t_0. On note $M_{a,uv}$ le nombre d'individus du domaine a qui vérifient à la fois la modalité u de X et la modalité v de Z, à la date t.

On peut dire que $M_{a,uv}$ actualise $N_{a,uv}$: le problème posé est donc bien celui d'une « mise à jour » de statistiques sur petits domaine. Par hypothèse, on connaît les $N_{a,uv}$ (ou une estimation de très bonne qualité), qui sont en pratique issus d'un recensement, ou d'un décompte dans un fichier administratif. Si on note

$$M_{\bullet,uv} = \sum_{a=1}^{m} M_{a,uv} \text{ et } M_{a,u\bullet} = \sum_{v} M_{a,uv} ,$$

on dispose par hypothèse des estimations $\hat{M}_{\bullet,uv}$ et on cherche à estimer $M_{a,u\bullet}$ (l'estimateur sera noté $\hat{M}_{a,u\bullet}$).

Une situation concrète pourrait être la suivante : dans un ensemble de m grandes communes (si besoin, présentant un certain caractère d'homogénéité) on profite de l'enquête annuelle de recensement de l'année t_0 pour attacher un questionnaire complémentaire qui permette de collecter des informations X et Z (par exemple il s'agit de questions inhabituelles sur les déplacements urbains, où X est l'utilisation d'un mode de transport et Z est la possession d'un véhicule. Mais Z peut aussi être une information collectée chaque année, comme la tranche d'âge). Quelques années plus tard, l'année t, on s'intéresse aux nombre d'individus de la commune a prenant une certaine modalité de X (par exemple se déplaçant en vélo). On effectue alors une enquête spécifique sur l'ensemble des m communes, de façon à avoir une taille d'échantillon global assez grande, mais avec une taille d'échantillon par commune qui reste modeste (par exemple 500 questionnaires par commune, si m=10). On cherche à estimer le nombre d'individus, commune par commune, qui vérifient une modalité quelconque de X (dans l'exemple, qui se déplacent en vélo). Dans ce contexte, les $N_{a,uv}$ sont des estimateurs communaux très fiables (compte tenu de la taille d'échantillon importante dans les grandes communes -c'est la taille totale de l'échantillon dans la grande commune a qui compte). Et les $\hat{M}_{\bullet,uv}$ sont également fiables parce qu'ils dépendent en qualité de la taille totale de l'échantillon dans l'ensemble des m communes.

La méthode proposée consiste à chercher des effectifs estimés actualisés $x_{a,uv}$ qui soient « proches » des effectifs $N_{a,uv}$ d'origine mais dont la somme sur l'ensemble des m domaines considérés (les communes dans notre exemple) redonne bien les estimateurs globaux $\hat{M}_{\bullet,uv}$. La mesure de proximité fait intervenir une fonction de distance « au choix », la distance du $chi-2$ étant de ce point de vue un bon critère :

$$\text{Minimiser} \sum_{a=1}^{m}\sum_{u}\sum_{v}\left(N_{a,uv}-x_{a,uv}\right)^2 \Big/ N_{a,uv}$$

$$\text{sous contraintes} \sum_{a} x_{a,uv} = \hat{M}_{\bullet,uv} \quad \forall u \text{ et } \forall v$$

On peut aussi changer le critère de distance et rechercher :

$$\text{Minimiser} \sum_{a=1}^{m}\sum_{u}\sum_{v} N_{a,uv} Log \frac{N_{a,uv}}{x_{a,uv}}$$

$$\text{sous contraintes} \sum_{a} x_{a,uv} = \hat{M}_{\bullet,uv}$$

Quelle que soit la fonction de distance retenue, La solution est la suivante :

$$x_{a,uv} = \frac{N_{a,uv}}{N_{\bullet,uv}} \times \hat{M}_{\bullet,uv}$$

où $N_{\bullet,uv} = \sum_{a=1}^{m} N_{a,uv}$. Cette valeur estime $M_{a,uv}$, on peut donc la noter $\hat{M}_{a,uv}$. Finalement on forme

$$\boxed{\hat{M}_{a,u\bullet} = \sum_{v} \frac{N_{a,uv}}{N_{\bullet,uv}} \times \hat{M}_{\bullet,uv}}$$

On vérifie que si $\hat{M}_{\bullet,uv}$ estime $M_{\bullet,uv}$ sans biais (ce qui est a priori le cas, car l'échantillonnage est en général conçu pour cela) et si, pour tout u et pour tout v :

$$\frac{M_{a,uv}}{N_{a,uv}} \text{ ne dépend pas de } a$$

(en pratique, en dépend peu), alors $\hat{M}_{a,u\bullet}$ est sans biais de $M_{a,u\bullet}$. Cet estimateur $\hat{M}_{a,u\bullet}$ est de type synthétique : en effet, il est construit à partir de données collectées sur d'autres domaines que a. La condition d'absence de biais s'écrit aussi, pour tout couple de modalités (u,v) :

$$\frac{M_{a,uv}}{M_{a',uv}} = \frac{N_{a,uv}}{N_{a',uv}}$$

pour tout couple de domaines a et a'. Cela signifie que, pour tout (u,v), l'importance relative des domaines n'évolue pas (ou évolue peu, en pratique) dans le temps : par exemple si un domaine a' est 3 fois plus gros qu'un domaine a au moment du recensement, cette condition dit qu'il doit toujours être 3 fois plus gros que a au moment de l'enquête.

Comme dans toute approche de type synthétique, la variance de $\hat{M}_{a,u\bullet}$ sera faible si la taille de l'échantillon global concerné par l'estimation $\hat{M}_{\bullet,uv}$ est grande : la variance de $\hat{M}_{a,u\bullet}$ varie comme l'inverse de $\sum_{a=1}^{m} n_a$. Il est donc intéressant d'avoir m grand.

Il existe une extension intéressante de cette méthodologie lorsqu'on dispose d'un effectif global estimé par domaine, soit ici $\hat{M}_{a,\bullet\bullet}$. Il s'agit bien ici d'une « simple » estimation démographique, qui peut être obtenue par n'importe quel procédé ! Dans l'exemple ci-dessus, $\hat{M}_{a,\bullet\bullet}$ estime la taille de la commune a à la date t (le recensement en fournira une estimation tout à fait satisfaisante).

On peut alors résoudre :

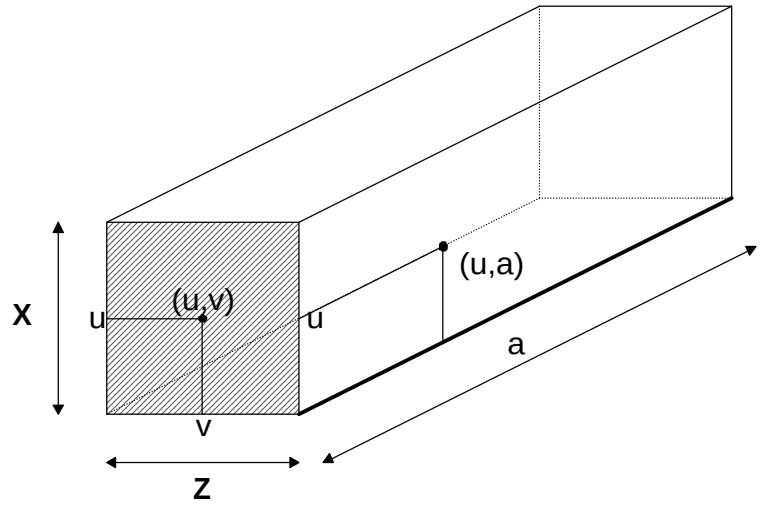
$$\text{Minimiser } \sum_{a=1}^{m} \sum_{u} \sum_{v} N_{a,uv} \cdot Log \frac{N_{a,uv}}{x_{a,uv}}$$

$$\text{sous contraintes } \begin{cases} \sum_{a} x_{a,uv} = \hat{M}_{\bullet,uv} \text{ pour tout } (u,v) \\ \sum_{u,v} x_{a,uv} = \hat{M}_{a,\bullet\bullet} \end{cases}$$

Il n'existe pas de solution analytique, mais on peut résoudre ce problème par un algorithme itératif de type « raking ratio ». (Calmar pourrait convenir si la fonction objectif était programmée -mais ce n'est pas le cas).

On peut illustrer graphiquement la problématique :

Les données sont représentées par un parallélépipède dont les axes correspondent aux modalités des trois variables en jeu, soit X , Z , et les aires a : à la date courante, on connaît la réalité sur l'intégralité de la face avant (hachurée) et sur une arrête (en gras), et on cherche à connaître la réalité de la face latérale (gris clair), sachant que l'intérieur est entièrement connu, mais à une date antérieure.



Dans la littérature, cette méthode est connue sous le nom de SPREE (Structure Preserving Estimation).

2.3 Le problème de l'estimation de la qualité des estimateurs synthétiques.

Nous avons vu en introduction que l'erreur quadratique moyenne (EQM) constitue l'indicateur essentiel pour mesurer la qualité d'un estimateur (quel qu'il soit), du moins en terme « absolu » (en terme relatif, on pourra diviser la racine carrée de l'EQM par l'estimateur synthétique concerné - on obtient alors le coefficient de variation).

L'EQM est la somme de la variance et du biais au carré. L'estimation de variance ne pose pas ici un problème spécifique : elle est certes difficile si l'échantillonnage est complexe, mais ce n'est pas lié à la taille du domaine. En effet, par construction puisque l'estimateur est synthétique, on se ramène de fait à la variance d'un estimateur construit à partir de l'échantillon complet (les tailles d'échantillon sont grandes, donc toutes les approximations habituelles sont permises). En revanche, l'estimation du (carré du) biais pose un vrai problème.

On peut néanmoins espérer obtenir un ordre de grandeur de l'EQM en utilisant habilement un estimateur direct (qui est toujours disponible sous la seule condition $n_a \geq 1$). Le processus découle de l'égalité suivante. Soit $\hat{Y}_a$ un estimateur direct de Y_a (peu importe son expression - ce peut être l'estimateur de Horvitz-Thompson ou l'estimateur par la régression $\hat{Y}_{Reg,a}$ ou une autre expression) et $\hat{Y}_{a,SYN}$ un estimateur synthétique de ce même Y_a (et là encore, peu importe son expression). On montre facilement que

$$E\hat{Q}M\left(\hat{Y}_{a,SYN}\right) = \left(\hat{Y}_{a,SYN} - \hat{Y}_a\right)^2 + \hat{V}\left(\hat{Y}_{a,SYN}\right) - \hat{V}\left(\hat{Y}_{a,SYN} - \hat{Y}_a\right)$$

constitue un estimateur sans biais de $EQM\left(\hat{Y}_{a,SYN}\right)$ où $\hat{V}(\ )$ et $E\hat{Q}M(\ )$ sont des estimateurs sans biais respectivement de la variance et de l'EQM. A ce stade, si le domaine est « assez grand » (mettons si n_a est de l'ordre de 1 000 ou plus), on peut se satisfaire de cette expression. Mais s'il s'agit d'un petit domaine, il y a un obstacle incontournable, et peut-être deux obstacles ou même trois. Avant de les aborder, on peut déjà simplifier l'expression parce que $\hat{Y}_a$ est

beaucoup plus instable que $\hat{Y}_{a,SYN}$. Aussi, la variance de $\hat{Y}_{a,SYN} - \hat{Y}_a$ sera (presque) celle de $\hat{Y}_a$, laquelle variance sera très supérieure à celle de $\hat{Y}_{a,SYN}$. D'où le nouvel estimateur simplifié

$$\boxed{E\hat{Q}M\left(\hat{Y}_{a,SYN}\right) = \left(\hat{Y}_{a,SYN} - \hat{Y}_a\right)^2 - \hat{V}\left(\hat{Y}_a\right)}$$

Si n_a est petit, il est possible que l'on ne sache pas estimer correctement la variance de $\hat{Y}_a$, parce que dans certains cas les expressions de variance ne sont manipulables qu'avec de grandes tailles d'échantillon. C'est le cas par exemple si le tirage est à probabilités inégales (on doit faire appel à des approximations de variance, qui ne se justifient que dans un cadre « asymptotique »). Mais ce problème ne survient pas toujours (avec un sondage aléatoire simple et si $\hat{Y}_a$ est l'estimateur de Horvitz-Thompson, il n'y a plus d'obstacle), alors que, dans tous les cas, le terme $\left(\hat{Y}_{a,SYN} - \hat{Y}_a\right)^2$ sera très instable à cause de la présence de $\hat{Y}_a$. Autrement dit, $E\hat{Q}M$ aura systématiquement une grande variance. Une troisième difficulté, qui, comme la première, ne survient que dans certaines circonstances, est liée au signe de $E\hat{Q}M$. On voit bien qu'il peut être négatif, « par malchance ». On arrive donc à une conclusion assez décevante : il y a bien moyen de produire une formule théorique pour estimer la qualité d'un estimateur synthétique quelconque, mais d'une part il n'est pas du tout sûr que l'on parvienne à la mettre en œuvre, et d'autre part et surtout elle s'avère de toutes façons elle-même de qualité médiocre ! Le miracle ne se produit donc pas : on ne parvient pas à contourner le problème initial dû à la faible quantité d'information dont on dispose sur le domaine qui nous intéresse.

Sur ce constat, on peut re-appliquer la philosophie de l'estimation synthétique et imaginer que dans certaines circonstances et avec certaines hypothèses supplémentaires, on puisse, au prix d'un biais, limiter la variance. C'est en effet possible si on s'intéresse à m petits domaines (a varie de 1 à m), sur chacun desquels on dispose des estimateurs $\hat{Y}_a$ et $\hat{Y}_{a,SYN}$, et si on pense que les EQM des estimateurs des moyennes $\hat{\bar{Y}}_{a,SYN}$ ne sont pas très différentes d'un domaine à l'autre (on applique cette hypothèse à l'estimateur de la moyenne et non pas à l'estimateur du total, tout simplement pour éliminer l'effet de taille). Si les m petits domaines ne sont en effet pas trop différents, l'approche se défend (penser à m départements de même « type » si on veut estimer un taux de chômage par exemple). Dans ces conditions,

$$E\hat{Q}M\left(\hat{\bar{Y}}_{a,SYN}\right) = \frac{1}{N_a^2} E\hat{Q}M\left(\hat{Y}_{a,SYN}\right) \approx \frac{1}{m}\sum_{a=1}^{m} \frac{1}{N_a^2} E\hat{Q}M\left(\hat{Y}_{a,SYN}\right)$$

D'où finalement, pour tout domaine a :

$$\boxed{\frac{1}{N_a^2} E\hat{Q}M\left(\hat{Y}_{a,SYN}\right) \approx \frac{1}{m}\sum_{a=1}^{m} \frac{1}{N_a^2}\left(\hat{Y}_{a,SYN} - \hat{Y}_a\right)^2 - \frac{1}{m}\sum_{a=1}^{m} \frac{1}{N_a^2}\hat{V}\left(\hat{Y}_a\right)}$$

Ce dernier estimateur gagne en stabilité par rapport à la première expression, et ce d'autant plus que m est plus grand. Evidemment, il perd son caractère sans biais en contrepartie. Son utilisation nécessite également un effort pédagogique puisqu'en cas de diffusion des résultats, on affichera la même précision pour tous les estimateurs de tous les petits domaines concernés.

3. L'estimation composite.

3.1. Principe général

L'estimateur direct $\hat{Y}_a^D$ (on utilise ici un index D ostensible pour bien rappeler qu'il peut s'agir de n'importe quelle expression de type direct) est généralement sans biais ou peu biaisé (attention néanmoins à ne pas systématiser cette propriété : ce n'est pas vrai avec une estimation classique de type ratio ou de type régression -comme $\hat{Y}_{Reg,a}$ par exemple- dès lors que n_a est tout petit) mais de forte variance. A l'inverse, un estimateur synthétique $\hat{Y}_a^{SYN}$ sera biaisé, mais de faible variance. D'où l'idée de construire un estimateur d'un nouveau genre, qui soit combinaison linéaire des deux précédents. On parle alors d'estimateur composite :

$$\boxed{\hat{Y}_{a,COMP} = \phi_a \hat{Y}_a^D + (1-\phi_a)\hat{Y}_a^{SYN}}$$

avec ϕ_a réel compris entre 0 et 1. A priori, ce réel dépend du domaine concerné. Il est choisi par l'utilisateur. Plus il est proche de 1, plus l'estimateur direct va peser dans la définition de $\hat{Y}_{a,COMP}$ (et à l'inverse plus ϕ_a est proche de 0, plus l'estimation composite aura des caractéristiques proches de l'estimateur synthétique).

Dans les parties qui suivent, on examine les trois principaux estimateurs composites de la théorie qui ne font pas appel à de la modélisation explicite (car on verra dans le chapitre suivant que tous les estimateurs issus d'un modèle explicite sont de type composite).

3.2. L'estimateur composite optimum.

L'objectif est simple : trouver le coefficient ϕ_a qui minimise l'erreur quadratique moyenne. Si on mène le calcul en supposant

$$E\left(\hat{Y}_a^D - Y_a\right)\cdot\left(\hat{Y}_a^{SYN} - Y_a\right) << E\left(\hat{Y}_a^{SYN} - Y_a\right)\cdot\left(\hat{Y}_a^{SYN} - Y_a\right)$$

(c'est une condition très réaliste en pratique, car le terme de gauche est de l'ordre de la covariance entre $\hat{Y}_a^D$ et $\hat{Y}_a^{SYN}$, qui devrait être relativement faible parce que ces deux estimateurs sont construits selon des logiques fort différentes), alors on trouve

$$\phi_a(OPTI) = \frac{1}{1+F_a} \qquad \text{(et on a bien } \phi_a(OPTI) \in [0,1]\text{)}$$

avec $$F_a = \frac{EQM\left(\hat{Y}_a^D\right)}{EQM\left(\hat{Y}_a^{SYN}\right)}$$

Cette expression est théorique et ne peut pas être calculée exactement : pratiquement, il faut estimer chaque terme à partir des données d'enquête, et on verra plus loin qu'il s'agit là d'une vraie difficulté !

Dans ce cas, l'estimateur composite est noté $\hat{Y}_{a,COMP}^{OPTI}$ et on a

$$\frac{EQM\left(\hat{Y}_{a,COMP}^{OPTI}\right)}{MIN\left(EQM\left(\hat{Y}_a^D\right), EQM\left(\hat{Y}_a^{SYN}\right)\right)} = \begin{cases} \phi_a(OPTI) \text{ si } \phi_a(OPTI) \geq \frac{1}{2} \\ 1-\phi_a(OPTI) \text{ si } \phi_a(OPTI) \leq \frac{1}{2} \end{cases}$$

De ce fait, ce rapport est toujours supérieur ou égal à 1/2. Donc l'EQM de $\hat{Y}_{a,COMP}^{OPTI}$ sera toujours au moins égale à la moitié de l'EQM du meilleur des deux estimateurs parmi $\hat{Y}_a^D$ et $\hat{Y}_a^{SYN}$, ce qui limite le gain de qualité de l'estimateur composite optimum. Cependant, il y a un gain effectif et systématique puisque ce rapport est malgré tout inférieur à 1. De plus, on montre que si,

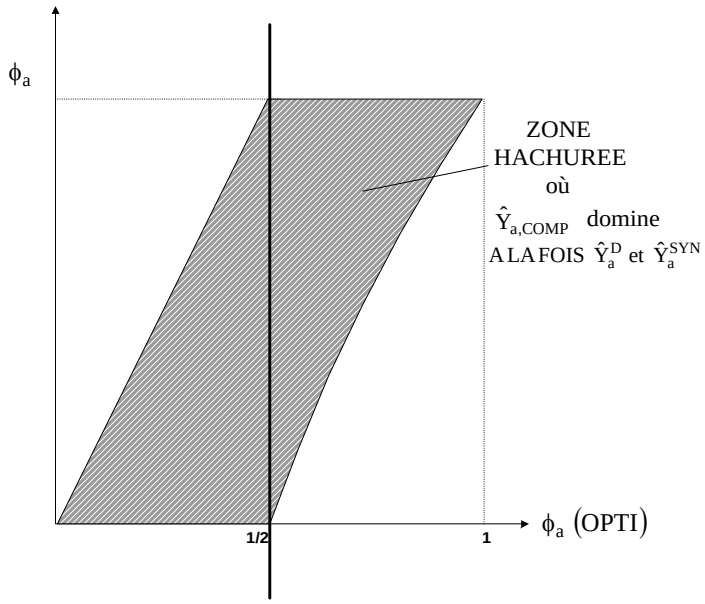
$$Max\left(0,2\cdot\phi_a(OPTI)-1\right)\le\phi_a\le Min\left(1,2\phi_a(OPTI)\right)$$

alors

$$EQM\left(\hat{Y}_{a,COMP}\right)\le Min\left(EQM\left(\hat{Y}_a^D\right),EQM\left(\hat{Y}_a^{SYN}\right)\right)$$

Pour un $\phi_a(OPTI)$ donné, c'est-à-dire pour un ratio F_a donné, il y a donc des conditions sur ϕ_a pour que l'estimateur composite améliore la qualité de l'estimation par rapport au meilleur des deux estimateurs $\hat{Y}_a^D$ et $\hat{Y}_a^{SYN}$.

Si ϕ_a ne vérifie pas la double inégalité ci-dessus, l'estimateur composite sera bien mal choisi et on dégradera la situation par rapport à $\hat{Y}_a^D$ ou à $\hat{Y}_a^{SYN}$. On peut représenter graphiquement les situations possibles en portant $\phi_a(OPTI)$ en abscisse et ϕ_a en ordonnée :



La zone hachurée représente la zone de gain (sous-entendu par rapport au meilleur des deux estimateurs $\hat{Y}_a^D$ et $\hat{Y}_a^{SYN}$). Le risque encouru est très clairement lié à la valeur de $\phi_a(OPTI)$: si $\phi_a(OPTI)$ vaut 1/2, il y a toujours amélioration, mais plus on se rapproche des extrêmes 0 et 1, plus l'intervalle dans lequel ϕ_a doit se trouver est petit. On remarquera que la diagonale $\phi_a=\phi_a(OPTI)$ se situe intégralement à l'intérieur du parallélogramme hachuré (ce qui est réconfortant et qui confirme ce que l'on savait déjà).

On vérifie qu'un estimateur naturel de $\phi_a(OPTI)$ est

$$\hat{\phi}_a(OPTI)=\frac{E\hat{Q}M\left(Y_a^{SYN}\right)}{\left(\hat{Y}_a^D-\hat{Y}_a^{SYN}\right)^2}$$

où $\hat{EQM}$ est une estimation (sans biais) de l'EQM, telle qu'elle a été présentée au 2.3. On retrouve exactement le même genre de problème qu'au 2.3. : la présence de l'estimateur direct rend cette expression très instable. Elle peut d'ailleurs être négative, ou supérieure à 1, auquel cas l'estimateur optimal perd tout son sens. On peut appliquer de nouveau l'idée du moyennage présentée au 2.3., sous condition de disposer de m petits domaines et de supposer que $\phi_a(OPTI)$ ne varie « pas trop » d'un domaine à l'autre. Comme cette dernière hypothèse est un peu hasardeuse, on prend des risques… mais on peut toujours calculer

$$\tilde{\phi}_a(OPTI) = \frac{1}{m}\sum_{a=1}^{m} \hat{\phi}_a(OPTI)$$

Quant à l'estimation de l'erreur quadratique moyenne de $\hat{Y}_{a,COMP}$, on peut reprendre ligne à ligne le développement du 2.3. : c'est exactement la même problématique, et il faut au moins utiliser un estimateur direct pour produire une estimation… instable, hélas.

3.3. Les estimateurs dépendant de la taille de l'échantillon.

Il s'agit d'une catégorie d'estimateurs qui n'apparaît pas de manière particulièrement naturelle -il faut bien le reconnaître. L'idée générale est de trouver un estimateur composite qui coïncide avec l'estimateur direct lorsque la taille d'échantillon n_a est « assez grande » : la condition peut porter sur $\hat{N}_a$, estimateur de Horvitz-Thompson de la taille N_a du domaine, soit (rappel)

$$\hat{N}_a = \sum_{i \in s_a} \frac{1}{\Pi_i}$$

car plus n_a est grand, plus $\hat{N}_a$ est grand (a priori, car il y a plus de termes dans la somme). Cela conduit par exemple à la proposition suivante :

$$\phi_a = \begin{cases} 1 & \text{si } \hat{N}_a > \delta \cdot N_a \\ \dfrac{\hat{N}_a}{\delta N_a} & \text{si } \hat{N}_a \le \delta \cdot N_a \end{cases}$$

où δ est un réel (positif) au choix. Si $\delta = 1$, on retrouve l'estimateur direct dès lors que

$$\hat{N}_a > E\hat{N}_a = N_a$$

donc quand l'estimation de N_a dépasse la valeur attendue « en moyenne ». On notera que cette catégorie d'estimateurs ne peut être envisagée que si on connaît N_a.

Si le sondage est aléatoire simple et si on opte pour $\delta = 1$, on obtient :

- $\hat{Y}_{a,COMP} = \hat{Y}_a^D$, si $n_a \ge n \cdot \dfrac{N_a}{N}$
- $\hat{Y}_{a,COMP} = \left(\dfrac{N}{n}\dfrac{n_a}{N_a}\right)\hat{Y}_a^D + \left(1 - \dfrac{N}{n}\dfrac{n_a}{N_a}\right)\hat{Y}_a^{SYN}$, sinon

Cette approche ne semble vraiment intéressante que si $n \cdot N_a / N$ est « assez grand », c'est-à-dire finalement si le domaine n'est pas trop petit. En effet, le critère qui conduit à l'estimateur direct ne porte pas sur la valeur « absolue » de n_a, mais seulement sur une valeur « relative » (il faut

comprendre relative à son espérance), si bien qu'on peut in fine retomber sur l'estimateur direct avec un n_a petit (et on sait bien que ce n'est pas souhaitable…). Dans la littérature, on trouve en particulier la combinaison

$$\hat{Y}_{a,COMP} = \phi_a \cdot \left(\hat{B}^T \cdot X_a + \frac{N_a}{\hat{N}_a} \left[\hat{Y}_a - \hat{B}^T \hat{X}_a \right] \right) + \left(1 - \phi_a\right) \cdot \left(\hat{B}^T X_a \right)$$

avec $\phi_a = 1 \quad si\ \hat{N}_a \geq N_a$

$$= \left(\frac{\hat{N}_a}{N_a} \right)^2 \quad si\ \hat{N}_a < N_a$$

L'estimateur direct s'inspire de l'estimateur corrigé présenté au 2.2.3. (il est bien de « nature » directe) et l'estimateur synthétique reprend l'expression de base du 2.2.1..
L'estimation de l'EQM peut se faire comme au 2.3.

3.4. Les estimateurs dits « de James-Stein ».

Il s'agit d'une classe d'estimateurs de type composite permettant d'améliorer la qualité de l'estimateur, en un certain sens. On part d'estimateurs directs $\hat{Y}_a^D$ des totaux Y_a et on suppose qu'il existe une fonction g telle que la transformation $g\left(\hat{Y}_a^D\right)$, notée $\hat{\theta}_a$, est une variable aléatoire qui suit une loi normale. On suppose que les conditions permettent d'écrire $E\hat{\theta}_a = Eg\left(\hat{Y}_a^D\right) = g\left(Y_a\right) = \theta_a$. Si g est linéaire, il n'y a en général pas (trop) de problème parce que les estimateurs directs sont sans biais ou peu biaisés dès lors que la taille n_a n'est « pas trop petite ». Si g n'est pas linéaire, il faut des tailles n_a suffisamment importantes pour assurer l'égalité, au moins approximativement (par exemple si on s'intéresse à un revenu moyen, on utilise souvent pour g la fonction logarithme). On note Ψ_a la variance (d'échantillonnage) de $\hat{\theta}_a$, que l'on suppose connue. On a donc

$$\hat{\theta}_a \to \mathcal{N}\left(\theta_a, \Psi_a\right)$$

Les $\hat{\theta}_a$ sont mutuellement indépendants (propriété naturelle non contestable). Par ailleurs, et c'est ce qui fait l'originalité de cette approche, on dispose de valeurs θ_a^o que l'on considère a priori comme proches des paramètres respectifs θ_a (θ_a^o est appelé « guess » en anglais). Toute origine est acceptable : ce peut être une information totalement externe, ou tirée d'une enquête passée. Ce peut être aussi une estimation a priori construite à partir des résultats de l'enquête elle-même, par exemple en régressant les estimations directes $\hat{\theta}_a$ sur un vecteur d'informations auxiliaires z_a de dimension p, soit

$$\theta_a^o = z_a^T \hat{\beta} \quad \text{où} \quad \hat{\beta} = \left(Z^T Z\right)^{-1} Z^T \hat{\theta}$$

avec $Z^T = \left(z_1, z_2, \ldots, z_m\right)$ est une matrice $p \times m$. Cette méthode, de nature prédictive, « stabilise » les estimations a priori θ_a^o (d'autant plus que m est grand) puisque le coefficient $\hat{\beta}$ est calculé à partir de l'ensemble des domaines. La qualité de la prédiction est évidemment directement dépendante de la pertinence de la liaison linéaire postulée entre z_a et θ_a.

Lorsqu'il n'y a pas de raison de penser que les θ_a diffèrent (sensiblement) d'un domaine à l'autre, une méthode simple et naturelle de construction des θ_a^o consiste à choisir $p=1$ et $z_a=1\ (\forall a)$. Ce choix correspond à une situation tout à fait spécifique où, de fait, il n'y a pas de spécificité locale. Ce peut être aussi le choix le plus naturel (et le plus simple) lorsqu'on ne dispose d'aucune information auxiliaire. Dans ce cas, pour tout a de 1 à m :

$$\theta_a^o = \frac{1}{m}\sum_{a=1}^{m}\hat{\theta}_a$$

Décidons de retenir le critère de qualité suivant, notant θ la vraie valeur à estimer et $\tilde{\theta}$ l'estimateur considéré :

$$R\left(\theta,\tilde{\theta}\right)=\sum_{a=1}^{m}E\left(\tilde{\theta}_a-\theta_a\right)^2$$

Cette erreur (R comme Risque) est donc définie comme la somme des erreurs quadratiques moyennes sur chacun des m domaines considérés. Tous les domaines ont donc la même importance. On va, dans un premier temps, considérer que les variances Ψ_a sont indépendantes de a (la valeur commune est notée Ψ). Cette situation peut se présenter si les tailles d'échantillon n_a sont peu sensibles à a et si la distribution des valeurs individuelles Y n'est pas trop différente d'un domaine à l'autre. Sur ce dernier point, dans certains contextes, le choix d'une transformation g judicieuse peut permettre d'y parvenir. Un bon exemple est celui des proportions : si le paramètre d'intérêt θ_a est une proportion P_a (ou un effectif), on peut choisir

$$\theta_a = g\left(P_a\right)=\mathrm{Arcsin}\left(\sqrt{P_a}\right)$$

où Arcsin est la fonction réciproque de la fonction sinus (dite « Arc sinus »). Cette transformation est certes particulièrement exotique, mais il se trouve qu'elle permet de rendre la variance de $\hat{\theta}_a=g\left(\hat{P}_a\right)$ peu sensible[16] à P_a - et cela va donc dans le sens d'une uniformisation des Ψ_a. Cette transformation a également la vertu de rendre plus crédible l'hypothèse de normalité de $\hat{P}_a$, en particulier lorsque P_a se rapproche de 0% ou de 100% (où l'hypothèse est plus fragile, surtout si n_a n'est pas très grand).

Dans ces conditions, on peut construire les estimateurs dits de James-Stein, selon

$$\boxed{\hat{\theta}_{a,JS}=\theta_a^o+\left[1-\frac{K\cdot\Psi}{S}\right]\cdot\left(\hat{\theta}_a-\theta_a^o\right)}$$

avec $$S=\sum_{a=1}^{m}\left(\hat{\theta}_a-\theta_a^o\right)^2$$

et $K=$ $\begin{cases} m-2 \text{ si les } \theta_a^o \text{ sont exogènes} \\ m-p-2 \text{ si les } \theta_a^o \text{ sont issus d'une régression sur } z_a \text{ (de dimension } p\text{).} \end{cases}$

[16] Résultat : si $X \to \mathcal{B}(m,P)$, et si m est « grand », alors $Arc\sin\left(\sqrt{\frac{X}{m}}\right)$ a une variance approximativement égale à 1/ (4 m). Donc P n'apparaît pas dans cette variance.

On remarquera que cela impose une valeur minimale à m, qui vaut 2 dans le cas 1 et $p+2$ dans le cas 2 (ce peut donc être assez contraignant si p est grand -on peut dire aussi qu'à m fixé cela limite le nombre d'informations auxiliaires mobilisables pour construire l'estimateur a priori).

Cet estimateur :

- est (très) facile à calculer
- s'écrit aussi

$$\hat{\phi}_{JS} \cdot \hat{\theta}_a + \left(1 - \hat{\phi}_{JS}\right) \cdot \theta_a^o$$

où $\hat{\phi}_{JS} = 1 - \dfrac{K \cdot \Psi}{S}$ (ne dépend pas de a)

Il a donc une allure typique d'estimateur composite. Néanmoins, si on a bien $\hat{\phi}_{JS} \leq 1$, on peut avoir $\hat{\phi}_{JS} < 0$ (ce qui n'est pas très orthodoxe pour un estimateur composite).

- a la propriété suivante - et c'est bien là son intérêt majeur :

$$R\left(\theta, \hat{\theta}_{JS}\right) \leq R\left(\theta, \hat{\theta}\right), \text{ pour tout } \theta .$$

Cette inégalité signifie que $\hat{\theta}_{JS}$ domine l'estimateur direct $\hat{\theta}$ à partir duquel il est construit : il lui est toujours préférable selon le critère de risque retenu (et cela est vrai pour tout θ).

On peut commenter l'amélioration de qualité à partir des compléments suivants. Le risque de référence est celui de l'estimateur direct, soit

$$R\left(\theta, \hat{\theta}\right) = \sum_{a=1}^{m} E\left(\hat{\theta}_a - \theta_a\right)^2 = \sum_{a=1}^{m} V\left(\hat{\theta}_a\right) = m \cdot \Psi .$$

Or on montre :

$$R\left(\theta, \hat{\theta}_{JS}\right) \leq m\Psi - \frac{(m-2)^2 \Psi^2}{(m-2)\Psi + \sum_{a=1}^{m} \left(\theta_a - \theta_a^o\right)^2}$$

Ceci permet de juger de l'impact des estimations a priori θ_a^o : si on a effectué une estimation a priori très pertinente, c'est-à-dire si on a pu approcher correctement les θ_a (noter que ce n'est qu'un pari : en pratique, on ne peut pas savoir si c'est le cas, puisque θ_a est inconnu), on aura $\sum_{a=1}^{m} \left(\theta_a - \theta_a^o\right)^2$ proche de 0, et donc la majoration du risque de $\hat{\theta}_{JS}$ sera proche de 2Ψ. **Il y aura donc une amélioration notable dès que $m \geq 3$, et d'autant plus grande que m est grand**. Avec le critère de risque retenu (ne pas l'oublier…) l'amélioration peut devenir spectaculaire si m est grand (penser au cas où les domaines sont les départements, avec m proche de 100, ou les ZUS avec m proche de 800). Au contraire, si les estimations a priori sont éloignées des vraies valeurs, la borne supérieure va tendre vers $m\Psi$, et l'amélioration ne sera pas très importante (bien qu'elle existe toujours). Ce résultat est sain : l'amélioration est d'autant plus marquée que l'information utilisée a priori (les z_a) est pertinente pour expliquer les θ_a. En cela, l'estimateur de James-Stein préfigure bien toute l'approche par modélisation explicite qui va constituer les parties qui suivent, et qui est perpétuellement sous-tendue par l'idée qu'un modèle apportant sa contribution explicative va venir appuyer l'estimation directe, en la complémentant d'autant mieux que la modélisation est plus pertinente.

A ce stade, l'estimateur de James-Stein semble être une panacée, d'autant plus qu'il ne s'appuie pas sur une modélisation des θ_a (on ne peut pas dire que l'estimateur dépend d'un modèle) - mais il faut modérer son enthousiasme au vue des points suivants, qui traduisent des faiblesses.

- Si la fonction g n'est pas l'identité (par exemple si on utilise l'arcsinus dans le cas d'une proportion), la dominance de $\hat{Y}_{a,JS} = g^{-1}\left(\hat{\theta}_{a,JS}\right)$ sur l'estimateur direct concurrent $\hat{Y}_a^D$ n'est hélas pas assurée.
- Le critère de qualité est conçu pour une amélioration **globale** sur l'ensemble des domaines, mais **il n'apporte pas de garantie au niveau d'un domaine donné**. Il est d'ailleurs possible qu'il y ait une dégradation de l'erreur de certaines composantes $\hat{\theta}_{a,JS}$ par rapport à $\hat{\theta}_a$ (allant en théorie jusqu'à $m/4$ fois l'erreur de la composante de $\hat{\theta}_a$). Il est néanmoins possible d'adapter $\hat{\theta}_{a,JS}$ afin de limiter ce risque (sans le supprimer, cependant), en construisant l'estimateur suivant :

$$\hat{\theta}^*_{a,JS} = \begin{cases} \hat{\theta}_{a,JS} & \text{si } \left|\hat{\theta}_a - \hat{\theta}_{a,JS}\right| \leq \sqrt{\Psi_a} \\ \hat{\theta}_a - \sqrt{\Psi_a} & \text{si } \hat{\theta}_{a,JS} < \hat{\theta}_a - \sqrt{\Psi_a} \\ \hat{\theta}_a + \sqrt{\Psi_a} & \text{si } \hat{\theta}_{a,JS} > \hat{\theta}_a + \sqrt{\Psi_a} \end{cases}$$

- En pratique et pour mémoire, certaines hypothèses sont formulées pour les besoins théoriques mais ne sont, en réalité, pas vérifiées. Cela ne peut qu'avoir des conséquences sur les propriétés - et donc sur la pertinence - des estimateurs $\hat{\theta}_{a,JS}$ (conséquences que l'on ne maîtrise pas bien semble-t-il, mais qui laissent augurer, évidemment, une perte d'efficacité). Dans cette catégorie, on trouve l'hypothèse de normalité bien entendu, mais aussi le fait que $\hat{\theta}_a$ soit sans biais de θ_a (ce qui est difficilement admissible si g est non linéaire et que n_a est petit), et également le principe de variances d'échantillonnage Ψ_a connues. Cette dernière hypothèse est, en pratique, fantaisiste, car il faut toujours estimer la variance d'échantillonnage, et ce n'est donc pas Ψ_a que l'on peut intégrer dans la formulation de l'estimateur, mais $\hat{\Psi}_a$.
- La théorie s'adapte au cas de variances Ψ_a différentes, mais il faut en contrepartie modifier le critère de qualité, d'une façon qui n'est pas vraiment satisfaisante.

3.5. Les méthodes spécifiques aux estimations de population.

3.5.1. La méthode de modélisation des flux démographiques.

On considère que le petit domaine a est plongé dans une zone plus vaste Ω. Durant une période initiale t_0, on connaît, sur la population Ω, le nombre de "naissances" - noté B_0 - et le nombre de décès - notés D_0. Cette information sur les naissances et les décès est supposée exacte, et provient de fichiers administratifs : par exemple pour une population d'individus physiques, c'est l'état civil, et pour une population de logements c'est un comptage issu des fichiers de permis de construire achevés SITADEL et de permis de démolir. On dispose également d'un estimateur tout à fait fiable de la taille totale de la population sur Ω à la fin de la période t_0, de même qu'à la fin de n'importe quelle période t, noté $\hat{P}_t$. Si Ω contient des individus physiques, $\hat{P}_t$ peut très bien être le résultat des enquêtes de recensement successives. Au sein du domaine a, on peut aussi

exploiter les fichiers administratifs, si bien qu'on dispose du nombre « exact » de naissances dans a durant toute période t (que l'on notera b_t) et du nombre « exact » de décès durant cette même période (que l'on notera d_t). On notera p_t la taille totale de la population de a à la fin de n'importe quelle période t. Cette taille est inconnue - c'est l'objet de ce développement de chercher à l'estimer - mais on suppose que l'on connaît quand même p_0 exactement (par exemple c'est la valeur au RP99 - donc issue d'une source exhaustive).

Pour estimer p_t, on va s'appuyer sur le « modèle implicite » suivant :

$$\frac{\frac{b_t}{p_t}}{\frac{b_0}{p_0}} = \frac{\frac{B_t}{P_t}}{\frac{B_0}{P_0}}$$

Cela revient à considérer que le ratio « poids des naissances dans la population » évolue entre t_0 et t d'une manière identique dans a d'une part, et dans Ω d'autre part. On en déduit immédiatement un estimateur $\hat{p}_t$ de p_t selon :

$$\frac{\frac{b_t}{\hat{p}_t}}{\frac{b_0}{p_0}} = \frac{\frac{B_t}{\hat{P}_t}}{\frac{B_0}{\hat{P}_0}}$$

Le raisonnement symétrique peut être fait avec les décès, si bien qu'on peut finalement retenir :

$$\hat{p}_t = \frac{1}{2}\left(\frac{b_t}{\frac{b_0}{p_0} \cdot \frac{B_t/\hat{P}_t}{B_0/\hat{P}_0}} + \frac{d_t}{\frac{d_0}{p_0} \cdot \frac{D_t/\hat{P}_t}{D_0/\hat{P}_0}} \right)$$

Cette méthode peut être affinée si on applique le modèle sur des sous-populations.

3.5.2. Les méthodes utilisant une régression.

Soit t la date « courante ». On suppose qu'avant t il y a eu (au moins) deux recensements ayant permis de connaître les vraies tailles de population dans un ensemble Ω de m petits domaines a où m est « grand ». On note p_{a0} la population dénombrée dans a lors du premier recensement (date 0, disons) et p_{a1} la population dénombrée dans a lors du second recensement (date 1). Soit

$$P_0 = \sum_{a=1}^{m} p_{a0} \text{ et } P_1 = \sum_{a=1}^{m} p_{a1}$$

On suppose par ailleurs que l'on dispose d'une batterie de variables $\lambda_{at}^{(j)}$ sur les domaines a et pour chaque date t (à partir de la date 0) telles que si on note :

$$\delta_a = \frac{p_{a1}/P_1}{p_{a0}/P_0} \text{ et } z_a^{(j)} = \frac{\lambda_{a1}^{(j)}/\Delta_1^{(j)}}{\lambda_{a0}^{(j)}/\Delta_0^{(j)}}$$

avec $\Delta_t^{(j)} = \sum_{a=1}^{m} \lambda_{at}^{(j)}$

alors on peut lier les δ_a et les $z_a^{(j)}$ par une relation linéaire multivariée du type :

$$\delta_a = \gamma_0 + \gamma_1 z_a^{(1)} + \gamma_2 z_a^{(2)} + \ldots + \gamma_P z_a^{(P)} + u_a$$

où u_a est un résidu « petit » d'espérance nulle. δ_a mesure l'évolution, entre deux recensements consécutifs, de la part de la population totale qui se situe dans le domaine. $z_a^{(j)}$ est construit sur le même principe, mais à partir de variables autres que la taille de population. Les $z_a^{(j)}$ proviennent de fichiers exhaustifs et réguliers, en pratique de fichiers administratifs.

Partant du modèle précédent, on estime les γ_j par la méthode classique des moindres carrés et on utilise les coefficients $\hat{\gamma}_j$ estimés en posant :

$$\frac{\hat{p}_{at}/\hat{P}_t}{p_{a1}/P_1} = \hat{\gamma}_0 + \sum_{j=1}^{P} \hat{\gamma}_j \cdot \frac{\lambda_{at}^{(j)}/\Delta_t^{(j)}}{\lambda_{a1}^{(j)}/\Delta_1^{(j)}}$$

On notera que la procédure est originale en ce sens où les variables - expliquée aussi bien qu'explicatives - ne correspondent pas à celles qui ont permis l'estimation des γ_j ! Cette « manipulation » suppose une certaine stabilité des phénomènes reliant les variables explicatives aux variables expliquées, faute de quoi on introduit un biais.

A ce stade, on connaît les valeurs de tous les régresseurs, on connaît les p_{a1} (et donc P_1) et puisque la réunion Ω des m domaines est de grande taille, l'échantillon global va permettre d'estimer P_t (par $\hat{P}_t$) de manière fiable. Le modèle implicite sous-jacent ne se traduit pas ici par une formule précise, mais il se décrit plutôt par le fait que la relation de type linéaire précédente peut être ré-écrite à l'identique au cours du temps, avec des variables explicatives et expliquée qui évoluent. Cette méthode est bien d'inspiration synthétique parce que les $\hat{\gamma}_j$ sont obtenus à partir de l'ensemble des domaines.

Une autre approche de type régression est envisageable dans le cas où on dispose d'une estimation directe de p_{at} pour certains petits domaines a à la date t, en nombre k $(p \le k \le m)$. Notons $\hat{p}_{at}^{D}$ cette estimation. On pose alors :

$$\frac{\hat{p}_{at}^{D}/\hat{P}_t}{p_{a1}/P_1} = \beta_0 + \sum_{j=1}^{P} \beta_j \cdot \frac{\lambda_{at}^{(j)}/\Delta_t^{(j)}}{\lambda_{a1}^{(j)}/\Delta_1^{(j)}} + u_a$$

Il reste à estimer β_j par $\hat{\beta}_j$ (méthode classique des moindres carrés) et on aboutit aux estimateurs indirects $\hat{\hat{p}}_{at}$ définis selon l'équation :

$$\frac{\hat{\hat{p}}_{at}/\hat{P}_t}{p_{a1}/P_1} = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j \frac{\lambda_{at}^{(j)}/\Delta_t^{(j)}}{\lambda_{a1}^{(j)}/\Delta_1^{(j)}}$$

On notera que les β_j s'estiment à partir de k points mais que l'on obtient bien les estimations $\hat{\hat{p}}_{at}$ sur chacun des m domaines.

Cette approche, comme la précédente, s'appuie sur une technique de régression, mais elle se distingue en revanche par le fait qu'elle fait toujours intervenir les mêmes régresseurs : on abandonne donc ce « tour de passe-passe » qui consistait à changer de régresseurs tout en conservant les coefficients de la régression. Cette stabilité joue plutôt à l'avantage de $\hat{\hat{p}}_{at}$ sur $\hat{p}_{at}$. En contrepartie, la régression qui conduit à $\hat{\hat{p}}_{at}$ s'appuie sur des estimations directes, donc instables (alors que la régression associée à $\hat{p}_{at}$ faisait intervenir des variables expliquées δ_a parfaitement stables par construction). Cet aspect est, cette fois, à l'avantage de $\hat{p}_{at}$.

La comparaison des qualités des trois estimateurs concurrents $\hat{p}_{at}^{D}$, $\hat{p}_{at}$ et $\hat{\hat{p}}_{at}$ est possible sous certaines hypothèses. On montre que si la variance d'échantillonnage de $\dfrac{\hat{p}_{at}^{D}/P_t}{p_{a1}/P_1}$ est supérieure à la variance σ_v^2 du modèle qui suit :

$$\frac{p_{at}/P_t}{p_{a1}/P_1} = \beta_0 + \sum_{j=1}^{P} \beta_j \frac{\lambda_{at}^{(j)}/\Delta_t^{(j)}}{\lambda_{a1}^{(j)}/\Delta_1^{(j)}} + v_a$$

(donc $Var(v_a) = \sigma_v^2$ pour tout a), alors l'EQM de $\hat{p}_{at}$ sera inférieure à celle de $\hat{p}_{at}^{D}$. Cette situation est la plus vraisemblable en pratique. Si le temps n'affecte pas la structure des liaisons qui servent à construire les estimations, autrement dit si on pose

$$\delta_a = \gamma_0 + \sum_{j=1}^{P} \gamma_j \cdot z_a^{(j)} + u_a \qquad Var\, u_a = \sigma_U^2$$

et que, pour tout j de 1 à p, on a $\beta_j = \gamma_j$, alors $\hat{p}_{at}$ sera plus efficace que $\hat{\hat{p}}_{at}$ si et seulement si la variance d'échantillonnage de $\hat{p}_{at}^{D}$ est supérieure à la somme des variances des deux modèles en jeu, soit $\sigma_U^2 + \sigma_V^2$. Cela a de bonnes chances d'être le cas, en pratique.

Estimation indirecte avec modélisation explicite

1. Principe de base de cette approche.

On aborde désormais un ensemble de méthodes relevant d'une autre philosophie. En effet, on va s'appuyer sur des modèles qui relient une variable d'intérêt à des informations auxiliaires en tenant compte de deux sources de variabilité : d'une part une variabilité liée au sondage (puisqu'il y a échantillonnage, il y a « quelque part » une sensibilité à l'échantillon tiré), et d'autre part une variabilité plus classique en statistique inférentielle qui considère le vrai total dans un domaine comme le résultat d'un mécanisme de nature aléatoire.

La modélisation n'a cependant d'intérêt que si elle crée des conditions satisfaisantes pour estimer les paramètres d'intérêt - qui sont toujours ici des totaux par domaine (ou des moyennes ou des proportions, ce qui est de même nature) ou des fonctions (même complexes) de ces totaux. Dans ce dernier cas de figure, on se ramène au problème d'estimation de totaux. Ces conditions se concrétisent

1) par l'hypothèse d'une liaison de type linéaire entre l'espérance (par rapport à l'aléa du modèle) d'un paramètre d'intérêt et une batterie de variables explicatives connues

et

2) par l'hypothèse que cette liaison s'applique de la même façon à l'ensemble des domaines que l'on étudie, la spécificité de chaque domaine étant néanmoins prise en compte dans les composantes du modèle.

Ce sont évidemment les conditions traditionnelles de la modélisation -par exemple celles que l'on trouve en économétrie. Il y a au préalable un effort d'imagination à faire pour voir un vrai total Y_a dans un domaine comme le résultat d'un mécanisme aléatoire : ce n'est en effet pas l'approche habituelle du sondeur. Néanmoins, elle sera dans toute la suite bien pratique car la liaison postulée constitue la base de gains de qualité que l'on peut espérer dans l'estimation de Y_a. L'objectif est d'exploiter au maximum cette liaison pour tirer profit de l'ensemble de l'information dont on dispose : on constatera in fine que c'est effectivement l'ensemble de tous les domaines qui est impliqué dans l'expression de l'estimateur du total relatif à l'un quelconque d'entre eux. Autrement dit, si on ne peut ou si on ne veut travailler que sur un seul domaine, cette partie devient absolument sans objet, parce qu'il n'y a pas de modélisation exploitable possible : si par exemple on veut estimer le taux de chômage dans un département D (qui est un ratio de totaux), il faudra estimer les taux de chômage et récupérer les informations auxiliaires par département, puis relier les premiers aux secondes sur 10, 20, ou peut-être 30 départements - et pas seulement sur « le » département D. Dans le cas contraire, on peut (et on doit, de fait) se raccrocher au chapitre précédent et adopter une approche implicite (qui consistera à postuler qu'au sein de certaines sous-populations, le taux de chômage de D est le même que le taux de chômage régional, voire national - techniquement, on n'a pas besoin de manipuler des taux départementaux en dehors de D). Comme c'est le cas chaque fois qu'il modélise, le statisticien court un risque de biais, mais c'est une contrepartie inévitable à la réduction de la dimension du problème. On dispose néanmoins d'éléments de diagnostic sur la pertinence des modèles (ce dernier point est fort compliqué, et ne sera pas abordé par la suite).

2. Présentation des principaux modèles utilisés

On peut structurer la typologie des modèles de plusieurs manières, mais celle qui nous paraît la plus pertinente s'appuie sur la nature des individus statistiques : soit il s'agit des individus « de base » (personnes physiques ou entreprises), soit il s'agit des domaines eux-mêmes (un individu coïncide avec un domaine, autrement dit la modélisation porte sur des grandeurs définies au niveau du domaine).

2.1. Les modèles conçus au niveau « domaine »

2.1.1. Le modèle de Fay et Herriot

C'est l'approche de base, la plus connue et probablement la plus utilisée. Comme nous venons de le signaler, on travaille avec plusieurs domaines : on notera m le nombre de domaines concernés par les traitements. Un domaine « quelconque » est identifié par a (donc a varie de 1 à m).

On s'intéresse aux moyennes $\overline{Y}_a$ d'une variable d'intérêt réelle Y et on suppose que l'on connaît, pour chacun des m domaines, une information auxiliaire vectorielle (de dimension p) notée[17] z. Pour gagner en généralité, on partira d'une modélisation d'une fonction de la moyenne, notée

$$\theta_a = g\left(\overline{Y}_a\right)$$

où g est une fonction de R dans R parfaitement définie (g est souvent l'identité, mais dans certains cas le logarithme peut être plus adapté, par exemple lorsqu'on manipule des revenus). On va postuler pour tout a une relation du type :

$$\theta_a = z_a^T \cdot \beta + b_a \cdot v_a$$

où β est un vecteur de dimension p inconnu, b_a est un réel connu (fonction du domaine) et v_a est une variable aléatoire d'espérance nulle et de variance σ_v^2 inconnue. Il s'agit de l'expression classique du modèle linéaire, célébrissime en statistique inférentielle et largement manipulée par les économètres. L'aléa de v_a n'a donc rien à voir avec l'aléa de sondage ; il traduit le processus stochastique à l'origine de l'écart entre la vraie valeur θ_a et la partie expliquée $z_a^T\beta$ (c'est ainsi que l'on donne le plus souvent un sens à l'aléa de modèle : représenter et concentrer tout ce qui n'a pas pu être expliqué par l'information dont on dispose). Autrement dit, v_a **concentre toute la variabilité « inter petits domaines » (toute la spécificité locale) qui n'a pas pu être traduite au travers de l'information auxiliaire** z_a. Le coefficient b_a est purement technique : il est là pour créer une variance (celle de $b_a v_a$) qui dépende du domaine (on peut toujours s'en débarrasser en posant $b_a = 1$). On considérera que les m variables aléatoires v_a sont mutuellement indépendantes (ce qui est fort réaliste). On notera qu'il n'y a pas d'hypothèse sur la loi des v_a (en pratique, on fait souvent l'hypothèse d'une loi de Gauss : cela confère certes des propriétés sympathiques à l'estimateur, mais ce n'est pas techniquement indispensable).

Dans notre contexte, qui est centré sur la variabilité des θ_a, on dira par la suite que v_a est un « effet aléatoire » et le modèle un « modèle à effet aléatoire »[18]. La pertinence d'une modélisation par effet aléatoire peut être longuement discutée et on peut entrer dans des débats sans fin sur cette question. Notre sentiment est qu'il s'agit d'un moyen, peut être pas très représentatif de la réalité, mais au moins techniquement pratique pour réduire la dimension du problème et permettre en conséquence de procéder à des estimations en prenant en compte une spécificité locale sans être bloqué par une surabondance de paramètres. Pour bien comprendre l'enjeu, l'effet aléatoire doit être opposé à l'effet fixe. Dans un modèle à effet fixe, ce qui traduit la spécificité v_a du domaine a n'est pas une variable aléatoire mais une variable explicative déterministe, exactement au même titre que les z_a (d'ailleurs on pourrait l'écrire comme une composante de z_a, et ne pas l'isoler dans l'écriture du modèle). Cette vision des choses conduirait à considérer

[17] Par convention, les vecteurs sont toujours des vecteurs « colonne ». Le vecteur « ligne » s'obtient en transposant le vecteur colonne (la transposée est repérée par l'exposant T).
[18] Traditionnellement, c'est au modèle du 2.2.1 qu'on réserve cette appellation.

que l'espérance de θ_a contient cet effet fixe, c'est-à-dire qu'elle s'explique par autre chose que les composantes de z_a : autrement dit, la spécificité d'un domaine apparaitrait comme une composante explicite de l'espérance du paramètre θ_a. Au contraire, dans le modèle à effet aléatoire, l'espérance de θ_a n'a rien de spécifique au-delà de ce qui est pris en compte dans z_a, mais la spécificité de chaque domaine est à mettre au compte d'un phénomène de variance : c'est alors le paramètre de variance σ_v^2 qui encaisse toute cette spécificté. Cette différence d'approche entre effet fixe et effet aléatoire peut sembler byzantine, mais en fait elle traduit une profonde différence de philosophie. Dans le premier cas la spécificité des domaines est structurelle (c'est gravé dans l'espérance du modèle), dans le second elle résulte d'un effet du hasard (c'est un effet de variance qui s'interprète comme la réalisation d'un simple phénomène aléatoire). L'approche par effet fixe nous parait dans le fond beaucoup plus adaptée : si par exemple une grosse usine ferme dans une commune, de manière suffisamment brutale pour qu'aucune information auxiliaire disponible z_a n'ait le temps de traduire le phénomène, et que cela crée un phénomène localisé conduisant à une forte croissance du taux de chômage dans la commune, il n'est pas très naturel de traduire cet effet comme un résulat de loterie et d'imaginer qu'il aurait pu en être autrement ! Hélas, techniquement le modèle par effet fixe est dans ce contexte une impasse, parce qu'il y a autant d'effet fixes que d'observations : comme on ne peut rien estimer, il faut se résoudre à l'abandonner !

En pratique, on peut avoir de sérieuses difficultés à collecter l'information auxiliaire z_a (réputée « exacte », c'est-à-dire qu'il s'agit bien ici d'une vraie valeur relative au domaine et non d'une estimation obtenue par sondage - ou si c'est le cas, il faut que l'estimateur associé soit extrêmement précis). Il est même possible que ce soit finalement la partie la plus délicate, voire la plus décevante, de l'opération. En effet, on se trouve face au problème essentiel déjà évoqué dans la partie 1. du chapitre consacré à la modélisation implicite, et si on ne parvient pas à disposer de sources exhaustives contenant des informations « suffisamment » explicatives, la perspective d'une modélisation efficace disparaîtra.

A ce stade, il est fondamental de noter que θ_a n'EST PAS OBSERVE (il existe… mais on ne le connaît pas, puisque c'est justement ce que l'on cherche !). De ce point de vue, on n'est donc pas du tout dans les circonstances classiques de la modélisation linéaire. En revanche, on a observé un estimateur direct de θ_a, noté $\hat{\theta}_a$, qui diffère de θ_a par une erreur d'échantillonnage notée e_a, soit pour tout a

$$\hat{\theta}_a = \theta_a + e_a$$

$\hat{\theta}_a$ est construit sur le modèle de θ_a, à partir d'un estimateur direct de $\overline{Y}_a$ (noté $\hat{\overline{Y}}_a$). La suite considère que les erreurs e_a sont sans biais - il faut comprendre par rapport à l'aléa d'échantillonnage. C'est un cadre théorique sur lequel on s'appuie par la suite mais qui n'est pas systématiquement vérifié, et qui constitue donc une limitation à la mise en œuvre de ce modèle. Il y a essentiellement deux cas a priori défavorables :

- Une fonction g non linéaire ;
- Un estimateur direct $\hat{\overline{Y}}_a$ biaisé (ce qui est assez courant, penser au ratio ou à l'estimateur par la régression - de façon générale, penser à tout estimateur redressé)

Ces deux limites perdent de leur importance en pratique dès que n_a est « assez grand » (quelques dizaines). Cela étant, on peut toujours contourner la difficulté évoquée dans le second alinéa en utilisant l'estimateur direct de Horvitz-Thompson, qui n'est pas biaisé. On retiendra que si g est non linéaire, la modélisation peut échouer si le domaine est très petit.

La variance d'échantillonnage de $\hat{\theta}_a$ est notée Ψ_a. C'est aussi la variance de e_a. On la suppose connue : s'agissant d'une question directement liée à l'échantillonnage, il n'y a pas d'obstacle théorique incontournable qui empêcherait le calcul de Ψ_a (un logiciel d'estimation de variance comme Poulpe par exemple, fournira dans la plupart des cas une estimation correcte de Ψ_a). Cela étant, le réalisme conduit à modérer ce discours sur deux aspects : d'une part, on obtiendra « seulement » une estimation de Ψ_a et donc des résultats fantaisistes si n_a est vraiment petit, et d'autre part la complexité de certains plans de sondage pourra en pratique constituer une véritable difficulté pour le statisticien qui n'a pas les moyens d'investir en parallèle dans le calcul de variance.

Les e_a sont, quant à eux, supposés mutuellement indépendants dans l'approche de base. La pertinence de cette hypothèse dépend directement de la méthode d'échantillonnage. En effet, certains plans de sondage vont conduire à des corrélations non négligeables entre $\hat{\theta}_a$ et $\hat{\theta}_b$. Cela se produira surtout si on utilise des sondages à plusieurs degrés, dans lesquels l'effet de grappe génère mécaniquement de la corrélation. Par exemple si on échantillonne des districts du recensement et que l'on s'intéresse à un revenu moyen, d'une part sur le (petit) domaine a des hommes de 40 ans, d'autre part sur le (petit) domaine b des femmes de 40 ans, il est fort probable qu'il y aura une corrélation positive[19] entre les estimateurs de revenu $\hat{\theta}_a$ et $\hat{\theta}_b$. Si les domaines a et b sont inclus dans des strates de tirage différentes, la corrélation sera rigoureusement nulle, mais dans tous les autres cas (ne serait-ce qu'avec un banal tirage aléatoire simple dans la population globale) il y aura une covariance théoriquement non nulle. Cependant, dans la plupart des cas, cette covariance sera numériquement négligeable (avec le tirage aléatoire simple, elle varie en $1/N$ où N est la taille de la population globale - donc autant dire qu'elle est nulle).
Finalement, en concaténant les deux écritures précédentes, on obtient, pour tout a :

$$\boxed{\hat{\theta}_a = z_a^T \beta + b_a v_a + e_a} \qquad \textbf{(M1)}$$

Il est tout à fait raisonnable de considérer que v_a et e_a sont des variables aléatoires indépendantes parce qu'on définit les plans de sondage au niveau national (qui conduisent à e_a) d'après des considérations générales de précision et de coût qui sont totalement déconnectées des phénomènes locaux (qui pour leur part conditionnent v_a).

Le modèle obtenu (noté M1) est dit « modèle de Fay et Herriot ». C'est un modèle qui « superpose » donc deux types d'aléa, mais ça n'a en fait aucune importance, ce qu'il faut retenir c'est que le terme aléatoire $b_a \cdot v_a + e_a$ a une espérance nulle et, compte tenu des hypothèses faites, une variance globale égale à $b_a^2 \cdot \sigma_v^2 + \Psi_a$. On rappelle que les v_a sont là pour traduire la spécificité des domaines, alors que ce n'est pas du tout la nature des e_a.

Le paramètre σ_v^2 témoigne de l'existence d'une variabilité de type « inter » domaines : il est non nul dès lors qu'il y a des effets propres aux domaines, au-delà de la partie expliquée par Z. On peut donc interpréter $b_a \cdot v_a$ comme le terme qui traduit la spécificité du domaine a par rapport aux autres, hors effet de Z (spécificité due à un ensemble de causes que l'on n'est pas en mesure de formaliser par une information identifiée et disponible). Les paramètres ψ_a sont au contraire liés à la variabilité « intra » domaines, s'agissant d'une variance d'un estimateur direct.

[19] Compte tenu de l'effet de grappe naturel lié à la proximité imposée par les limites des districts, si le revenu moyen estimé des hommes de 40 ans est supérieur à sa moyenne vraie, il est fort à parier que ce sera également le cas pour le revenu moyen estimé des femmes.

On peut étendre M1 dans deux directions essentiellement :

- la première consiste à modéliser simultanément plusieurs paramètres d'intérêt dans chaque domaine a, soit à manipuler :

$$\theta_a = \left(\theta_a^1, \theta_a^2, \ldots, \theta_a^K\right)^T \qquad (K \geq 2)$$

M1 s'étend alors à un modèle multivarié où, pour tout a :

$$\hat{\theta}_a = Z_a^T \beta + b_a v_a + e_a$$

Z_a est une matrice (contrairement à z_a) et les matrices de variance-covariance de e_a et de v_a ne sont pas diagonales. Ainsi, on peut bénéficier de la structure de corrélation entre les composantes de $\hat{\theta}_a$ pour améliorer la qualité de l'estimateur final : on va en effet introduire quelques paramètres supplémentaires dans le modèle, mais en démultipliant le nombre d'observations (multiplié par K), si bien que l'opération ira globalement dans le sens d'une amélioration de la qualité de l'estimation (même argumentation pour les modèles temporels - voir 2.1.3). Par exemple, il sera probablement bien plus efficace de modéliser le vecteur (revenu total du domaine, patrimoine total du domaine, taux de chômage du domaine) que chaque composante séparément. De même, dans une enquête sur le handicap et la dépendance, on aura intérêt à modéliser des vecteurs regroupant des composantes relatives aux différents handicaps physiques et aux handicaps mentaux.

- La seconde généralisation consiste à accepter une corrélation non nulle entre les e_a. Comme écrit ci-dessus, cette situation survient surtout en cas de tirage à plusieurs degrés, lorsque les différents domaines se retrouvent au sein des unités primaires. Les domaines sont alors généralement des catégories socio-démographiques bien définies, mais a priori, pour les enquêtes Insee en tout cas, ce contexte ne se présente pas lorsque les domaines sont des aires géographiques.

2.1.2. Modèle de corrélation spatiale.

Le modèle (M1), même dans sa version « étendu », s'est toujours appuyé sur des effets locaux v_a mutuellement indépendants. Cela gagne en simplicité, mais il arrive que cette hypothèse soit peu crédible : c'est en particulier le cas lorsque les petits domaines sont géographiquement proches et qu'il y a un effet notable de la géographie sur la détermination des paramètres $\overline{Y}_a$ (donc θ_a). Si on considère deux agglomérations a et b proches l'une de l'autre et que l'on s'intéresse au taux de chômage dans ces agglomérations, on imagine facilement que les effets résiduels v_a et v_b soient corrélés : en effet, il y a probablement des composantes explicatives du taux de chômage non prises en compte dans Z, qui s'avèrent identiques dans a et b.
Un modèle bien adapté consiste à postuler une covariance entre v_a et v_b qui soit une fonction décroissante d'une « distance » d_{ab} entre a et b. Par exemple

$$COV\left(v_a, v_b\right) = \alpha \cdot e^{-\beta d_{ab}} \qquad (\alpha, \beta) \in \mathbb{R}^2$$

On se démarque donc du modèle de Fay et Herriot, dans lequel la covariance entre les résidus du modèle était supposée nulle.

Certains auteurs abordent la question en proposant une loi paramétrée de v_a conditionnelle aux autres effets résiduels $v_b (b \neq a)$. Par exemple, si on définit Ω_a comme étant un « voisinage » de a, on peut poser pour tout a

$$v_a \left| \{v_b, b \neq a\} \rightarrow \mathcal{N}\left(\rho \cdot \sum_{b \in \Omega_a} v_b, \sigma_v^2 \right) \right.$$

où ρ est un paramètre réel inconnu. La notion de voisinage peut s'entendre au sens géographique, mais pas nécessairement : ce peut être selon des critères socio-démographiques par exemple. De cette relation, on peut tirer une forme de covariance (non conditionnelle) entre v_a et v_b, pour tout (a,b).

Toutes ces écritures ne font qu'exprimer la même idée : il y a un effet de « contagion » liée à une forme de proximité entre les petits domaines, c'est-à-dire que deux petits domaines proches vont avoir des effets propres v_a voisins (et d'autant plus voisins qu'ils sont plus proches !).

2.1.3. Modèles temporels.

On rencontre ces modèles lorsqu'on pratique des enquêtes répétées dans le temps. L'échantillonnage peut être indépendant d'une date à l'autre, mais ce peut être aussi un panel, ou un échantillonnage rotatif. On obtient alors des estimations $\hat{\theta}_{at}$ pour m domaines a et pour T dates (t varie de 1 à T). On peut généraliser M1 en posant :

$$\begin{cases} \hat{\theta}_{at} = \theta_{at} + e_{at} \\ \theta_{at} = g\left(\overline{Y}_{at}\right) = z_{at}^T \beta + b_a v_a + u_{at} \end{cases}$$

Les structures sur les erreurs (e_{at}), (v_a), (u_{at}) peuvent être plus ou moins complexes. Il peut aussi y avoir des hypothèses de lois (de Gauss) sur tout ou partie de ces variables aléatoires. Il paraît en tout cas incontournable, dès lors que l'échantillonnage n'est pas conçu indépendamment d'une date à l'autre, de considérer que les e_{at}, pour a fixé et t variant de 1 à T sont corrélés (ce qui traduit une corrélation dans le temps des $\hat{\overline{Y}}_{at}$). La corrélation spatiale des e_{at} est moins claire, et renvoie aux commentaires des parties 2.1.1. et 2.1.2..

A cause de la dimension temporelle et des inerties « naturelles » qui l'accompagnent, on a tendance à considérer que les u_{at} sont corrélés, pour a fixé, lorsque t varie. Une modélisation simple pour traduire cette dépendance temporelle est celle de l'auto régression, soit

$$u_{at} = \rho \cdot u_{a,t-1} + \varepsilon_{at}$$

avec $|\rho| < 1$ (ρ paramètre réel inconnu). Les ε_{at} sont cette fois deux à deux indépendants.
On peut évidemment sophistiquer la modélisation à loisir… les procédures d'estimation seront plus difficiles en conséquence et la technique à mettre en œuvre n'en sera que plus lourde. Mais la philosophie d'ensemble restera toujours la même. Par exemple, le modèle suivant apparaît comme assez général :

$$\hat{\theta}_{at} = \theta_{at} + e_{at}$$

$$\theta_{at} = z_{at}^T \cdot \beta_{at}$$

Le vecteur de coefficients $\beta_{at} = \left(\beta_{at0}, \beta_{at1}, \ldots, \beta_{at,p}\right)^T$ vérifie :
$\forall j = 0,1,2,\ldots,\rho \quad \beta_{atj} = \lambda \cdot \beta_{a,t-1,j} + \mu \cdot \beta_{aj} + \varepsilon_{atj}$, où les β_{aj} sont des réels (ne dépendant donc pas du temps), (λ, μ) sont deux réels connus, et ε_{atj} une variable aléatoire vérifiant, à t et a fixés

$$\forall j \neq k \qquad Cov\left(\varepsilon_{atj}, \varepsilon_{atk}\right) = \sigma_{jk}$$

Les ε_{atj} sont non corrélés dans le temps. On retrouve des modèles classiques avec les cas particuliers $(\lambda,\mu)=(1,0)$, ou $(\lambda,\mu)=(0,1)$ ou encore $(\lambda,\mu)=(\rho,1-\rho)$.

L'intérêt de cette modélisation est de réduire la dimension du problème et d'augmenter a priori la qualité de l'estimation. En effet, l'existence de séries chronologiques va démultiplier les valeurs observées $\hat{\theta}_{at}$ (on aura T fois plus d'observations) au prix d'une augmentation insignifiante du nombre de paramètres (par exemple, dans le premier modèle on ajoute seulement le ρ). Globalement, si le modèle est adapté au contexte (c'est évidemment une condition déterminante qui nécessite une réflexion préalable approfondie), l'opération sera fort rentable et se soldera par une amélioration de la qualité d'estimation de chacun des paramètres, et par conséquent par une diminution de l'erreur quadratique moyenne de l'estimateur « petits domaines »[20] (même argumentation que pour la modélisation multivariée évoquée au 2.1.1).

2.1.4. Modèles pour variables qualitatives ou pour variables de comptage.

Le modèle de comptage le plus courant est le modèle de Poisson. Il s'agit d'un modèle linéaire généralisé. Soit $\hat{N}_a^c$ un estimateur direct de la taille de la sous-population d'intérêt (indexée par c) N_a^c au sein du petit domaine a. On suppose que $\hat{N}_a^c$ suit une loi de Poisson de paramètre λ_a et que l'on dispose de variables auxiliaires z_a définies au niveau du domaine telles que, pour une fonction F connue, on a

$$F(\lambda_a) = z_a^T \beta$$

La fonction F est souvent un logarithme. Les $\hat{N}_a^c$ sont supposés deux à deux indépendants. On peut estimer par maximum de vraisemblance les paramètres composant β, et ainsi proposer l'estimateur de type synthétique

$$\tilde{\tilde{N}}_a^c = \hat{\lambda}_a = F^{-1}(z_a^T \hat{\beta})$$

Dans SAS, la procédure GENMOD permet d'obtenir les composantes de $\hat{\beta}$. La technique d'estimation est celle du maximum de vraisemblance, qui détermine la qualité de l'estimateur synthétique (en supposant que le modèle est « exact »). Or la justification du maximum de vraisemblance est essentiellement asymptotique, ce qui signifie qu'il faut travailler sur un nombre « suffisant » de domaines pour que cette approche soit valide (plusieurs dizaines, disons au moins 20). Si le modèle est exact, il y aura certes un biais (néanmoins faible si on dispose de nombreux domaines) mais la variance sera bien moins forte qu'avec un estimateur direct.
Cette modélisation, de nature typiquement synthétique, est aussi adaptée à l'estimation de proportions, puisqu'il suffit de diviser in fine l'estimation $\hat{N}_a^c$ par la taille N_a (supposée connue) du petit domaine.

2.2. Les modèles conçus au niveau « individu » :

2.2.1. Formulation générale.

L'unité sur laquelle porte la modélisation n'est plus le domaine, mais l'individu « de base » de la population du domaine. On repérera l'individu par l'identifiant a du domaine et par son

[20] Grosso modo, une partie de l'EQM va varier en 1 / mT au lieu de varier en 1 / m.

identifiant i interne au domaine a. On note $Y_{a,i}$ la valeur de Y prise par l'individu (a,i), et on pose

$$\boxed{Y_{a,i} = X_{a,i}^T \cdot \beta + v_a + e_{a,i}} \qquad \textbf{(M2)}$$

où $X_{a,i}$ est un vecteur de variables auxiliaires connues (dont certaines composantes peuvent d'ailleurs être des caractéristiques du domaine, donc des variables de « niveau domaine »), β est un vecteur de paramètres inconnus, v_a reste l'effet, de nature aléatoire, spécifique au domaine a (mêmes caractéristiques exactement que dans M1) et les $e_{a,i}$ sont des variables aléatoires mutuellement indépendantes vérifiant

$$E(e_{a,i}) = 0$$

$$V(e_{a,i}) = (k_{a,i})^2 \cdot \sigma_e^2$$

où les $k_{a,i}$ sont des réels connus et σ_e^2 est un paramètre inconnu. Si $Y_{a,i}$ est un revenu par exemple, ce modèle signifie que le revenu s'explique par une batterie de variables socio-démographiques $X_{a,i}$ (âge, CS, taille du ménage,…) et par un effet v_a propre au domaine, et que tout ce qui reste inexpliqué par la conjonction de ces effets se trouve rassemblé dans l'aléa $e_{a,i}$. Les $k_{a,i}$ sont des coefficients techniques permettant d'introduire une hétéroscédasticité, c'est-à-dire de traiter les cas où la variance (de modèle) de $Y_{a,i}$ varie en fonction de l'individu i de a.
En pratique, il faut de nouveau attirer l'attention sur la question cruciale de la disponibilité de l'information auxiliaire $X_{a,i}$: obtenir suffisamment de variables auxiliaires suffisamment explicatives des Y et dont on connaît les vrais totaux sur le domaine (on verra au 3.4.1 que l'estimation efficace s'appuie sur le vrai total des $X_{a,i}$) est loin d'être gagné ! Cet obstacle est majeur et peut être de nature à faire échouer toute cette approche (voire aussi partie 1 du chapitre consacré à l'estimation implicite).

L'effet v_a a été choisi de nature aléatoire, et de ce fait M2 s'appelle « modèle linéaire mixte ». Au 2.1.1, nous avons tenté d'expliquer la distinction entre effet fixe et effet aléatoire. Le discours peut être repris ici, mais avec une différence fondamentale, c'est que dans M2 il devient a priori possible d'estimer des effets fixes caractérisant le niveau domaine. En effet, sauf cas pathologique, l'échantillon comprendra plusieurs individus dans chaque domaine (notons toujours n_a cet effectif dans le domaine a). Cela étant, si la domaine a est petit, n_a sera généralement petit et donc la qualité de l'estimation de l'effet fixe - qui varie en 1 / n_a - sera médiocre. En pratique, cette raison peut être un argument suffisant pour opter malgré tout pour une modélisation à effets aléatoires !

La piste de la modélisation avec effets fixes (si le domaine est « assez grand ») ne donnera pas lieu ici à des développements spécifiques parce qu'elle renvoie à la théorie classique de la régression multivariée. L'estimateur final du total Y_a serait alors la somme des valeurs prédites.

On retourne désormais au contexte du modèle à effets aléatoires, qui revient à considérer qu'un individu se retrouve « par hasard » dans un domaine : on traite en quelque sorte le domaine comme une agrégation aléatoire d'individus, qui aurait pu avoir une composition autre, ou une frontière différente, de celle qu'il a en réalité (on peut imaginer une « couche d'aléa » supplémentaire préalable, qui vient fabriquer les domaines, et ensuite on génère les $Y_{a,i}$).

Important : le modèle M2 est supposé valable pour tous les individus de tous les domaines. Donc a varie de 1 à m et i varie de 1 à N_a. Il y a un point de fond à garder à l'esprit, qui conditionne la pertinence des estimations : il faut bien comprendre que M2 doit rester valable si on se restreint aux seuls individus de l'échantillon tiré dans les domaines étudiés, c'est-à-dire si on considère seulement $i \in s_a$ (a de 1 à m). Cela n'est pas évident et il y a essentiellement deux contextes en pratique qui le permettent :

- soit, pour tout a donné la probabilité d'inclusion de n'importe quel individu (a,i) dépend des informations constituant $X_{a,i}$ - et seulement de ces informations ;
- soit, pour tout a donné, la probabilité d'inclusion de (a,i) dépend de variables $\Delta_{a,i}$ qui sont indépendantes de $Y_{a,i}$. C'est en particulier le cas pour des tirages à probabilités égales.

Dans ces deux cas, la condition essentielle dans le fond est que le mécanisme d'échantillonnage soit indépendant des valeurs $Y_{a,i}$ (i décrit a) à $X_{a,i}$ fixé, autrement dit le fait d'être ou non dans l'échantillon, conditionnellement aux variables explicatives $X_{a,i}$, ne doit donner aucune information sur les valeurs des $Y_{a,i}$ (on peut parler de mécanisme d'échantillonnage « ignorable »). Le sondage aléatoire simple se place clairement dans ces conditions. Si le tirage est proportionnel à la taille (probabilités inégales), les conditions seront toujours respectées dès lors que la variable de taille, soit est explicative de Y, soit est totalement indépendante de Y. Dans le cas contraire, il y a un biais dit « de sélection », c'est-à-dire que (M2) restreint aux seuls individus tirés va conduire à des estimateurs biaisés. **Manipuler un modèle individuel ne peut raisonnablement se faire que si le mécanisme est ignorable.** Si l'échantillonnage utilise de l'information auxiliaire, il est donc prudent de l'inclure dans la liste des variables explicites du modèle. Concrètement, pour ce qui concerne les enquêtes-ménages de l'Insee, soit il s'agit d'un tirage à probabilités égales (c'est bien la moitié des cas…) et il n'y a pas de souci à se faire, soit on a à faire à un sondage sur-représentant certaines catégories d'individus (tirage dit « en deux phases ») et alors il faut que toutes les variables définissant les sous-populations distinguées par l'échantillonnage soient des variables explicatives dans le modèle de comportement M2. En particulier, les vecteurs $X_{a,i}$ peuvent contenir deux composantes distinctes faisant référence au même concept, mais considéré à deux dates différentes : c'est une traduction du vieillissement de l'information auxiliaire de la base de sondage que l'on peut rencontrer par exemple si un logement est tiré avec une probabilité qui dépend de son caractère locatif ou non au moment du recensement de 1999 (c'est la seule information dont on dispose dans la base de sondage) et que la variable d'intérêt Y attachée à ce logement soit expliquée (entre autres) par son caractère locatif ou non au moment de l'enquête. Dans ce cas, on inclura aussi le statut au recensement de 1999 dans le modèle M2[21].

Notons que M2, sous cette forme, n'est pas adapté au cas des échantillonnages à deux degrés (ou plus) parce qu'il ne traduit pas l'existence d'un effet de grappe (voir 2.2.2 pour tenir compte de cet effet).

On peut étendre le modèle M2 au cas où $Y_{a,i}$ n'est plus une variable réelle, mais un vecteur de variables réelles. Si $Y_{a,i}$ comprend p composantes, on écrit :

$$Y_{a,i} = B \cdot X_{a,i} + v_a + e_{a,i}$$

où B est cette fois une matrice à p lignes, et v_a et $e_{a,i}$ sont des vecteurs de taille p. Cette écriture multivariée permet de tenir compte des corrélations entre les composantes des $Y_{a,i}$.

[21] En espérant qu'il n'y aura pas trop de problèmes de colinéarité. Sinon, c'est que les deux statuts seront très proches et l'intérêt de les distinguer disparaitra.

Le modèle M2 est individuel, mais la valeur à prédire reste θ_a. Il faut donc passer, d'une façon ou d'une autre, à une agrégation de valeurs individuelles. Pour cela, on distingue deux cas :

- Soit N_a est considéré comme grand, et on peut écrire, pour tout a :

$$\overline{Y}_a = \overline{X}_a^T \beta + v_a + \overline{e}_a$$

où $\overline{X}_a$ est la vraie moyenne des $X_{a,i}$ sur l'ensemble du domaine a, soit $\frac{1}{N_a}\sum_{i=1}^{N_a} X_{a,i}$ et $\overline{e}_a$ est la vraie moyenne des $e_{a,i}$ sur l'ensemble du domaine a, soit $\frac{1}{N_a}\sum_{i=1}^{N_a} e_{a,i}$. Or $\overline{e}_a$ est à peu près nul dans ces conditions, parce que la loi des grands nombres montre qu'il s'agit « presque » de son espérance mathématique sous le modèle, égale par définition à zéro.

Donc

$$\overline{Y}_a \approx \overline{X}_a^T \beta + v_a$$

On constate que l'on tombe sur le modèle M1, dont on connaît maintenant la théorie.

- Soit N_a n'est pas considéré comme grand. Dans de telles circonstances, le paramètre d'intérêt $\overline{Y}_a$ ne s'exprime plus directement sous la forme (très pratique) d'une combinaison linéaire de β et de v_a. En revanche, si on pose $f_a = n_a / N_a$, on a :

$$\overline{Y}_a = f_a \cdot \overline{y}_a + (1 - f_a)\overline{Y}_a^*$$

où $\overline{Y}_a^*$ est la vraie moyenne des $Y_{a,i}$ sur les individus de la population de a non échantillonnés. Finalement, la question est ramenée au problème suivant : à partir des $Y_{a,i}$ des individus échantillonnés, prédire la moyenne inconnue $\overline{Y}_a^*$.

2.2.2. Modèle adapté à l'existence d'un effet de grappe.

Une faiblesse de M2 tient à l'hypothèse d'indépendance mutuelle des $e_{a,i}$, pour i dans a. Si l'échantillonnage s'effectue à deux degrés (pour simplifier), il y a de ce fait un découpage naturel du petit domaine a en unités primaires qui regroupent -en général- des individus qui présentent des similarités de comportement Y. On s'attend donc à ce qu'une modélisation adaptée prenne en compte cette réalité. Techniquement, il s'agit d'introduire une corrélation spécifique entre les Y des individus d'une même unité primaire (au-delà de celle qui provient de l'appartenance au même domaine). Pour cela, supposons que le petit domaine a contienne des individus appartenant à M_a unités primaires (indice u) distinctes. L'individu est identifié par le triplet (a,u,i) et sa valeur Y est modélisée selon

$$Y_{a,u,i} = X_{a,u,i}^T \cdot \beta + v_a + \varepsilon_{a,u} + e_{a,u,i}$$

a varie de 1 à m, u varie de 1 à M_a, i varie de 1 à $N_{a,u}$ où $N_{a,u}$ est la taille de l'intersection de a et de l'unité primaire u (une unité primaire donnée n'a en effet aucune raison d'être incluse dans a). Dans ces conditions, si $V(\varepsilon_{a,u}) = \sigma_\varepsilon^2$, on a

- $Cov(Y_{a,u,i}; Y_{a,v,j}) = \sigma_V^2$

- $Cov\left(Y_{a,u,i};Y_{a,u,j}\right)=\sigma_V^2+\sigma_\varepsilon^2$
- $Cov\left(Y_{a,u,i};Y_{a,u,i}\right)=V\left(Y_{a,u,i}\right)=\sigma_V^2+\sigma_\varepsilon^2+\sigma_e^2$

Ainsi, plus les individus sont « proches », plus la covariance est forte. L'effet de grappe se traduit par la covariance plus forte entre $Y_{a,u,i}$ et $Y_{a,u,j}$ qu'entre $Y_{a,u,i}$ et $Y_{a,v,j}$. Dans le contexte d'échantillonnage des enquêtes ménages à l'Insee, on devrait par exemple considérer ce modèle *a priori* dès lors que le domaine est une aire géographique de type région ou même département[22].

2.2.3. Modèle dit « à deux niveaux ».

Le modèle M2 s'écrit aussi, lorsque la première variable est une constante :

$$\begin{aligned} Y_{a,i} &= \beta_1 X_{a,i}^{(1)} + \beta_2 X_{a,i}^{(2)} + \ldots + \beta_p X_{a,i}^{(p)} + v_a + e_{a,i} \\ &= \left(\beta_1 + v_a\right) + \beta_2 X_{a,i}^{(2)} + \ldots + \beta_p X_{a,i}^{(p)} + e_{a,i} \end{aligned}$$

Si on note $\beta_a^{(1)} = \beta_1 + v_a$, on tombe sur une écriture de type régression linéaire avec un coefficient aléatoire, fonction du petit domaine considéré. On peut étendre ce principe à l'ensemble des coefficients de régression (et non plus seulement à celui qui correspond au terme constant), ce qui donne :

$$Y_{a,i} = X_{a,i}^T \cdot \beta_a + e_{a,i}$$

où β_a est un vecteur aléatoire de taille p. On va réduire le nombre de paramètres en modélisant β_a lui-même, selon

$$\beta_a = Z_a \cdot \alpha + v_a$$

où Z_a est une matrice connue, α un vecteur de paramètres fixes mais inconnus, et v_a un vecteur aléatoire centré.

Ce modèle est dit « à deux niveaux », puisque les β_a sont eux-mêmes modélisés. Si on rassemble les deux équations précédentes, on aboutit à :

$$Y_{a,i} = \left(Z_a^T \cdot X_{a,i}\right)^T \alpha + X_{a,i}^T v_a + e_{a,i}$$

qui est un peu plus général que M2 puisque la composante aléatoire « spécifique » au domaine n'est plus v_a mais la combinaison linéaire $X_{a,i}^T \cdot v_a$. De ce fait, ce n'est plus vraiment spécifique au domaine, puisque ce terme dépend de l'individu…

2.2.4. Modèles pour variables qualitatives ou pour variables de comptage.

Les modèles précédents, M2 ou ses satellites, ne conviennent manifestement pas aux variables qualitatives - qui sont par ailleurs les cas les plus fréquemment traités dans les enquêtes-ménages de l'Insee (ces modèles sont conçus pour des variables quantitatives, et même quantitatives continues). Pour s'en convaincre, il suffit de regarder les hypothèses faites sur les

[22] Face à un contexte comme celui-ci - par exemple la région en tant que domaine lorsque l'enquête est l'enquête « Emploi » et que l'on veut estimer un taux de chomage - il faudrait partir de ce modèle et satisfaire le test de nullité de la variance σ_ε^2 pour pouvoir appliquer directement la formulation M2. Si la nullité est refusée, il faut repartir de l'approche générale du 3.1.1 et l'adapter au cas particulier du modèle ci-dessus pour tenir compte de cet effet de grappe.

aléas $e_{a,i}$, qui sont typiques de variables quantitatives continues. Il y a néanmoins un contexte (très fréquent) où $Y_{a,i}$ doit être qualitative : c'est celui de l'estimation d'un effectif ou d'une proportion par domaine (exemple : nombre ou proportion de chômeurs par ZUS, nombre ou proportion d'individus dont le revenu mensuel est inférieur à 1000 euros), puisque si N_a^c est l'effectif et P_a la proportion associée cherchés dans le domaine a, on a :

$$P_a = \bar{Y}_a = \frac{1}{N_a}\sum_{i=1}^{N_a} Y_{a,i} = \frac{N_a^c}{N_a} \quad \text{avec } Y_{a,i} = 1 \text{ ou } 0$$

Une première approche est celle de la modélisation linéaire généralisée « classique », sans effet aléatoire propre aux domaines.

a/ On considère que les $Y_{a,i}$ suivent des lois de Bernoulli $B(1, P_{a,i})$ - deux à deux indépendantes.

Les $P_{a,i}$ sont considérées comme des paramètres et non comme des variables aléatoires.

b/ On dispose par ailleurs de variables auxiliaires $X_{a,i}$ propres à l'individu et on pose :

$$\forall a, \forall i \ \ Log \frac{P_{a,i}}{1-P_{a,i}} = X_{a,i}^T \beta$$

Ce modèle est dit « modèle logistique » : c'est un modèle linéaire généralisé parce qu'une fonction complexe (logistique) de l'espérance $E(Y_{a,i})$ est égale à $X_{a,i}^T \beta$. L'estimation du maximum de vraisemblance (utiliser la procédure LOGISTIC dans SAS) conduit aux estimateurs $\hat{\beta}$. Comme $N_a^c = \sum_{i \in s} 1_{i \in c} + \sum_{i \notin s} 1_{i \in c}$, on va estimer l'effectif N_a^c par

$$\hat{N}_a^c = \sum_{i \in s} 1_{i \in c} + \sum_{i \notin s} \hat{E}(1_{i \in c}) = \sum_{i \in s} 1_{i \in c} + \sum_{i \notin s} \hat{P}_{a,i}$$

$\hat{P}_{a,i}$ est relié à $\hat{\beta}$ via la fonction logistique. L'estimateur final est bien d'inspiration synthétique. Cette modélisation est évidemment adaptée à l'estimation de proportions, puisqu'il suffit de diviser in fine l'estimation $\hat{N}_a^c$ par la taille N_a (supposée connue) du petit domaine.

Une seconde approche traite les $P_{a,i}$ comme des variables aléatoires incluant un effet (aléatoire) propre au domaine. On raisonne alors en trois temps :

a/ On considère que les $Y_{a,i}$ suivent des lois de Bernoulli $B(1, P_{a,i})$ - deux à deux indépendantes.

Les $P_{a,i}$ sont inconnus et fonction de l'individu concerné.

b/ On explique les variables aléatoires $P_{a,i}$ par des variables auxiliaires $X_{a,i}$ propres à l'individu et par un effet (aléatoire) v_a propre au domaine. Un modèle courant est le modèle logistique, qui pose :

$$\forall a, \forall i \ \ Log \frac{P_{a,i}}{1-P_{a,i}} = X_{a,i}^T \beta + v_a$$

où on a besoin de supposer une loi pour v_a - en pratique une loi de Gauss - pour pouvoir estimer le paramètre β :

$$v_a \to \mathcal{N}(0, \sigma_v^2)$$

c/ La grandeur à prédire reste la proportion P_a. Or on a

$$P_a = \frac{1}{N_a}\left(\sum_{i\in s_a} Y_{a,i} + \sum_{i\notin s_a} Y_{a,i}\right)$$

Il faut donc prédire tous les $Y_{a,i}$ pour $i \notin s_a$: la meilleure prédiction de $Y_{a,i}$, compte tenu de l'information dont on dispose, est (voir aussi 4.4) :

$$\begin{aligned} Y_{a,i}^{*} &= E\left(Y_{a,i}\middle|Y_{a,j};\, j\in s_a\right) = \text{Proba}\left(Y_{a,i}=1\middle|Y_{a,j};\, j\in s_a\right) \\ &= E\left(P_{a,i}\middle|Y_{a,j};\, j\in s_a\right) \end{aligned}$$

Ainsi, $Y_{a,i}^{*}$ apparaît aussi comme l'estimateur optimum de $P_{a,i}$. Toute cette approche ressemble à celle du modèle dit « à deux niveaux » (cf. 2.2.3.), mais adaptée à des variables qualitatives.

On peut aussi s'intéresser à des variables de comptage par domaine, notées $\tilde{N}_a$. Un exemple typique est celui du comptage d'individus atteints d'une certaine maladie dans une aire a donnée (ou du comptage d'un effectif de chômeurs). On imagine qu'il existe un taux λ_a réel (inconnu) propre à l'aire a qui s'interprète comme le taux de prévalence de la caractéristique de la population dénombrée (par exemple une probabilité de contracter la maladie, ou une probabilité d'être touché par le chômage). Puis on modélise $\tilde{N}_a$ comme une variable aléatoire suivant une loi de Poisson de paramètre $N_a \cdot \lambda_a$, où N_a est la taille totale de l'aire (rappel). Cela se comprend bien si on considère le cas (simplifié) où les événements « Etre atteint par la maladie » (ou par le chômage) sont indépendants d'un individu à l'autre, car dans ce cas $\tilde{N}_a$ sont a priori une loi binomiale (N_a, λ_a). Il est bien connu que si N_a est grand et λ_a petit, cette loi se confond avec une loi de Poisson de paramètre $N_a \cdot \lambda_a$. Cette approche se trouve ici généralisée au cas où les événements ne sont pas indépendants -ce que l'on imagine comme assez vraisemblable en pratique dans le cas de phénomènes de type maladie ou chômage.

Dans un second temps, on considère λ_a comme une variable aléatoire, un peu selon l'approche du modèle dit « à deux niveaux ». A ce stade, les modélisations potentielles sont multiples. L'approche la plus simple postule une loi paramétrée pour λ_a, qui ne dépend pas de a. Dans la littérature, on trouve par exemple

$$\lambda_a \rightarrow Gamma(\omega_1, \omega_2)$$

On trouve aussi, par exemple

$$Log\ \lambda_a = X_a^T \beta + v_a$$

3. La classe des estimateurs sans biais optimums et linéaires (SBOL et ESBOL)

3.1. Présentation générale de l'estimation « SBOL ».

3.1.1. Formulation de l'estimateur.

Cette partie contient une présentation générale de la prédiction dans le cadre du modèle linéaire mixte, qui dépasse largement le cadre des petits domaines (les notations sont donc très générales et il ne faut pas chercher pour l'instant à les relier avec ce qui précède).

L'approche « SBOL » s'applique lorsqu'on a à faire à une classe de modèles dits « Modèles linéaires mixtes » : il s'agit de modèles expliquant une variable quantitative Y à l'aide d'effets fixes (de type régression linéaire) et d'effets aléatoires. Ce sont des modèles de ce type qui ont été présentés dans le chapitre 2. La formulation générale du modèle linéaire mixte est :

$$\boxed{Y = X\beta + Zv + e} \qquad \textbf{(M3)}$$

Y : vecteur observé de taille n

X : matrice $n \times p$ (connue)

β : vecteur inconnu de taille p

Z : matrice $n \times h$ (connue)

v : vecteur aléatoire de taille h (inobservé)

e : vecteur aléatoire de taille n (inobservé)

On impose à e et v d'être d'espérance nulle. De plus, e et v sont indépendants. Enfin, leurs matrices de variance respectives sont notées :

$$V(v) = G(\delta) \quad \text{(on notera } G\text{)}$$

$$V(e) = R(\delta) \quad \text{(on notera } R\text{)}$$

où δ est un paramètre vectoriel de taille q :

$$\delta = (\delta_1, \delta_2, \ldots, \delta_q)^T$$

Les variances sont donc de formes connues, mais non calculables puisque dépendant d'un paramètre inconnu. La question posée se présente ainsi : on considère la valeur réelle

$$\mu = l^T \beta + m^T v$$

où l et m sont des vecteurs parfaitement connus, et on cherche à la PREDIRE. On notera que μ comprend une composante déterministe (en β) et une composante aléatoire (en v) - ce qui rend μ ALEATOIRE. Un prédicteur de μ, noté $\hat{\mu}$ doit être calculable, et donc il doit s'appuyer sur ce qui est observé, c'est-à-dire sur Y. Pour des raisons de simplicité, on cherchera $\hat{\mu}$ linéaire, donc de la forme :

$$\hat{\mu} = q^T \cdot Y + r$$

où q et r sont des vecteurs parfaitement déterministes. On va imposer à $\hat{\mu}$ d'être sans biais, au sens suivant :

$$E(\hat{\mu} - \mu) = 0$$

Enfin, on va évidemment chercher $\hat{\mu}$ au plus proche de μ « en moyenne », de façon à ce que $\hat{\mu}$ ait un caractère d'optimalité. La proximité entre $\hat{\mu}$ et μ se comprend au sens de l'erreur quadratique moyenne. Donc on minimise :

$$E(\hat{\mu}-\mu)^2$$

où les inconnues sont les vecteurs q et r. Il s'agit d'un problème de minimisation sous contrainte, que l'on sait résoudre. On obtient, après calculs :

$$\boxed{q_{SBOL} = \left(l^T - m^T G\, Z^T V^{-1} X\right)\left(X^T V^{-1} X\right)^{-1} X^T V^{-1} + m^T G\, Z^T V^{-1}}$$

et $$r_{SBOL} = 0$$

avec $$V = V(Y) = V(Z\nu) + V(e) = ZGZ^T + R$$

Le prédicteur optimum est noté $\hat{\mu}^H = q_{SBOL} Y + r_{SBOL}$.

Il est intéressant de présenter autrement ce résultat. On peut en effet écrire :

$$\boxed{\hat{\mu}^H = l^T \tilde{\beta} + m^T \left[G\, Z^T V^{-1}\left(Y - X\tilde{\beta}\right)\right]}$$

où $$\tilde{\beta} = \left(X^T V^{-1} X\right)^{-1} \cdot X^T V^{-1} Y$$

La formulation de $\tilde{\beta}$ est familière, car il s'agit de l'estimateur dit « des moindres carrés généralisés » dans un modèle de régression multiple (puisque V est la matrice de variance de la variable expliquée Y). De ce fait, on interprète le terme entre crochets dans l'expression de $\hat{\mu}^H$ comme « le » prédicteur (optimum) de la variable aléatoire ν.

Si δ est connu, on peut calculer q_{SBOL} et donc $\hat{\mu}^H$.

Il est possible d'étendre cette théorie sans problème au cas où μ est un vecteur aléatoire (et non plus une variable aléatoire réelle), soit

$$\mu = L\beta + M\nu$$

où L et M sont des matrices (et non plus des vecteurs). On obtient

$$\hat{\mu}^H = L\tilde{\beta} + M\left[GZ^T V^{-1}\left(Y - X\tilde{\beta}\right)\right]$$

Cas particulier important :

Un cas particulier de M3, très fréquemment rencontré, est celui du modèle dit « à variance bloc-diagonale ». Il s'agit du cas où on peut décomposer Y, X, Z, ν et e ainsi :

$Y = (Y_1, Y_2, \ldots, Y_m)^T$ où Y_a vecteur de taille n_a

$\nu = (\nu_1, \nu_2, \ldots, \nu_m)^T$ où ν_a vecteur de taille h_a

$e = (e_1, e_2, \ldots, e_m)^T$ où e_a vecteur de taille n_a

Pour reprendre la notation associée à M1, on a

$$n = \sum_{a=1}^{m} n_a \text{ et } h = \sum_{a=1}^{m} h_a$$

On peut toujours poser : $X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{pmatrix}$ matrice $n \times p$ composée de m matrices $n_a \times p$ superposées

$$Z = \begin{pmatrix} Z_1 & & & 0 \\ & Z_2 & & \\ & & \ddots & \\ 0 & & & Z_m \end{pmatrix}$$

Matrice $n \times h$ composée de m matrices $n_a \times h_a$ placées en diagonale

On suppose que G est composée de m matrices. G_a placées sur la diagonale, et R est composé de m matrices R_a placées sur la diagonale -à l'image de la structure de Z (on parle de matrice de type « bloc diagonal »). On en déduit que V est elle-même de type bloc-diagonale, avec

$$V = \begin{pmatrix} V_1 & & & 0 \\ & V_2 & & \\ & & \ddots & \\ 0 & & & V_m \end{pmatrix}$$

et $$V_a = R_a + Z_a \cdot G_a \cdot Z_a^T$$

Dans ces conditions, M3 se décompose en fait en m modèles :

$$\boxed{Y_a = X_a \beta + Z_a v_a + e_a}$$

Cette écriture résulte des hypothèses d'indépendance entre les composantes aléatoires associées aux différents domaines. Il subsiste néanmoins une « liaison » fondamentale entre les différents domaines, concrétisée par le paramètre β.

Si on s'intéresse aux combinaisons linéaires :

$$\mu_a = l_a^T \cdot \beta + m_a^T \cdot v_a \quad (a \text{ varie de 1 à } m)$$

on peut appliquer la théorie générale en remarquant :

$$\mu_a = l_a^T \beta + \left(0,\ldots,0,m_a^T,0,\ldots,0\right) \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_a \\ \vdots \\ v_m \end{pmatrix}$$

$$= l_a^T \beta + (0,\ldots,0,m_a^T,0,\ldots 0)v$$

On aboutit à

$$\boxed{\hat{\mu}_a^H = l_a^T \tilde{\beta} + m_a^T G_a Z_a^T V_a^{-1}\left(Y_a - X_a \tilde{\beta}\right)}$$

où $$\tilde{\beta} = \left(\sum_{a=1}^{m} X_a^T V_a^{-1} X_a\right)^{-1} \left(\sum_{a=1}^{m} X_a^T V_a^{-1} Y_a\right)$$

Il est important de noter que les m modèles distingués ne sont pas « indépendants » - sinon on pratiquerait une estimation pour chacun d'entre eux et on aurait un $\tilde{\beta}$ qui dépend de l'indice a. Cette décomposition en m modèles est un peu « piégeuse » parce qu'on peut oublier qu'il s'agit du même paramètre β qui intervient dans les m équations (alors que v_a, lui, varie d'une équation à l'autre). C'est pourquoi l'expression finale $\tilde{\beta}$ fait intervenir des sommes portant sur les m domaines considérés. On comprend mieux ici ce qui constitue le cœur de la composante synthétique, puisque c'est précisément le fait que la relation entre les Y_a et les X_a soit la même pour tous les domaines a qui permet de stabiliser l'estimation de β.

Le modèle « à variance bloc-diagonale » est très important dans les estimations sur petits domaines parce qu'il s'applique directement dans tous les cas où on considère qu'il n'y a pas de covariance impliquant des variables aléatoires relatives à deux domaines distincts quelconques pris parmi les différents domaines impliqués dans l'étude (l'absence de covariance correspond à la manipulation de matrices de variance-covariance diagonales par bloc).

3.1.2. Expression de l'erreur de l'estimateur SBOL

L'erreur de $\hat{\mu}^H$ est choisie comme étant

$$EQM = E\left(\hat{\mu}^H - \mu\right)^2$$

on montre :

$$EQM = g_1(\delta) + g_2(\delta)$$

où

$$g_1(\delta) = m^T\left(G - GZ^TV^{-1}ZG\right)m$$

$$g_2(\delta) = \left(l^T - m^TGZ^TV^{-1}X\right)\left(X^TV^{-1}X\right)^{-1}\left(l - X^TV^{-1}ZG^Tm\right)$$

L'EQM a donc une expression très compliquée, mais néanmoins numériquement calculable dès lors que δ est connu.

3.2. Présentation générale de l'estimation empirique « ESBOL »

L'estimateur SBOL ne peut être calculé que si on connaît le paramètre δ, autrement dit si on est capable de calculer les matrices de variance de v et de e. Hélas, c'est rarement le cas en pratique. Dans ces conditions, l'approche la plus naturelle consiste à estimer δ à partir de l'information dont on dispose -donc Y- et à substituer l'estimateur de δ (noté $\hat{\delta}$) à δ dans l'expression de $\hat{\mu}^H$. On obtient un estimateur que l'on peut qualifier d'empirique, et que l'on notera $\hat{\mu}_E^H$ (estimateur empirique associé à l'estimateur SBOL -que l'on abrégera en ESBOL).

3.2.1. Estimation du paramètre des matrices de variance-covariance

Le premier - et principal - problème consiste à estimer δ. En toute généralité, on peut utiliser

- la méthode des moments
- la méthode d'ajustement des constantes

On peut aussi utiliser des méthodes moins classiques, comme la méthode MINQU.

Si on postule une loi de Gauss pour Y, on peut utiliser la célébrissime méthode du maximum de vraisemblance (EMV) ou une variante dite du « maximum de vraisemblance restreint » ($EMVR$)[23]. Comme ces hypothèses de normalité sont formulées la plupart du temps, il s'avère que $\hat{\delta}$ est le plus souvent l'EMV. Dans ces conditions, l'estimateur de β est également celui du maximum de vraisemblance (noté $\hat{\beta}$). La théorie et l'algorithme du maximum de vraisemblance sont assez compliqués, et il n'est pas raisonnable de chercher à programmer par soi-même la méthode : mieux vaut faire confiance à un logiciel, comme SAS par exemple qui propose la procédure PROC MIXED pour obtenir $\hat{\delta}$ et $\hat{\beta}$.

3.2.2. Qualité de l'estimateur ESBOL :

En toute généralité, $\hat{\mu}_E^H$ est biaisé. Néanmoins, on montre que si les conditions suivantes sont vérifiées :

i) $E\left(\hat{\mu}_E^H\right)$ finie

ii) $\hat{\delta}$ est une fonction paire de Y, soit $\hat{\delta}(Y) = \hat{\delta}(-Y),\ \forall Y$

iii) $\hat{\delta}(Y - Xb) = \hat{\delta}(Y),\ \forall Y, \forall b$

iiii) Les lois de v et de e sont symétriques autour de 0.

Alors $\hat{\mu}_E^H$ est sans biais de μ (rappel : ce vocabulaire prête à confusion, puisque μ n'est pas un paramètre mais une variable aléatoire : il est plus exact de dire que $\hat{\mu}_E^H - \mu$ est une variable aléatoire d'espérance nulle).

En fait, les deux conditions ii) et iii) portant sur $\hat{\delta}$ sont facilement réalisables : on vérifie que l'EMV, l'$EMVR$, et l'estimateur issu de la méthode d'ajustement des constantes, conduisent à des estimateurs $\hat{\delta}$ qui vérifient ces deux conditions.

L'erreur quadratique moyenne de l'estimateur ESBOL est supérieure à celle de l'estimateur SBOL (cela est logique, il y a une source de variabilité supplémentaire liée à l'estimation de δ). De manière générale, on ne sait pas exprimer l'EQM de $\hat{\mu}_E^H$ sous une forme exploitable (le contexte est beaucoup trop compliqué). Cela étant, dans le cas où v et e suivent les lois de Gauss et si la condition (iii) est vérifiée (encore une fois, cela est la cas avec les estimateurs « habituels »), on peut produire une expression approximative (par un raisonnement dit « heuristique ») de cette EQM.

$$EQM\left(\hat{\mu}_E^H\right) \approx g_1(\delta) + g_2(\delta) + g_3(\delta)$$
$$\approx EQM\left(\hat{\mu}^H\right) + g_3(\delta)$$

où $g_1(\delta)$ et $g_2(\delta)$ ont été donnés au 3.1.2. et on a

$$g_3(\delta) \approx Trace\left[\left(\frac{\partial b^T}{\partial \delta}\right) \cdot V(\delta) \cdot \left(\frac{\partial b^T}{\partial \delta}\right)^T \cdot \overline{V}\left(\hat{\delta}\right)\right]$$

où $b^T = m^T \cdot G(\delta) \cdot Z^T \cdot \left[V(\delta)\right]^{-1}$

[23] Sur cette technique, voir : C.E. McCulloch et S.R. Searle, « Generalised Linear and Mixed Models », 2001, Wiley.

et $\overline{V}\left(\hat{\delta}\right)$ désigne la matrice de variance-covariance asymptotique de $\hat{\delta}$. Si $\hat{\delta}$ est l'*EMV*, c'est l'inverse de la matrice d'information de Fisher.

Tout ce développement reste bien théorique (et ses résultats s'avèrent fort complexes) et n'a d'intérêt que parce qu'il sert de base à l'estimation de l'*EQM* de $\hat{\mu}_E^H$, puisque dans certaines conditions :

$$E\hat{Q}M\left(\hat{\mu}_E^H\right) = g_1\left(\hat{\delta}\right) + g_2\left(\hat{\delta}\right) + 2g_3\left(\hat{\delta}\right)$$

(noter le facteur 2 devant $g_3\left(\hat{\delta}\right)$ - alors qu'on pourrait s'attendre à 1). Cette formule s'avère rigoureuse si $\hat{\delta}$ est l'*EMV* où l'*EMVR* - mais plus « approximative » dans les autres cas. L'estimateur ci-dessus se justifie encore dans le cadre le plus général du modèle M3 - mais il faut quand même supposer que $\hat{\delta}$ estime δ sans biais (ou presque).

La PROC MIXED de SAS fournit l'estimateur $\hat{\mu}_E^H$ à partir des paramètres estimés $\left(\hat{\beta}, \hat{\delta}\right)$ par *EMV*, ainsi que $E\hat{Q}M\left(\hat{\mu}_E^H\right)$, à condition d'utiliser l'option DDFM=KENWARDROGER.

3.3. L'estimation « ESBOL » appliquée au modèle de Fay et Herriot.

3.3.1. Expression de l'estimateur SBOL.

Il s'agit du modèle central M1 conçu au niveau « domaine » (voir 2.1.1). Le paramètre θ_a est en toute rigueur une variable quantitative, continue et non bornée. Cela étant, il ne nous semble pas qu'il y ait de contre indication particulière à l'appliquer à des paramètres plus « contraints »[24], comme par exemple des proportions (comprises nécessairement entre 0 et 1) ou des effectifs (positifs et entiers). On rappelle le modèle M1 :

$$\hat{\theta}_a = z_a^T \beta + b_a v_a + e_a$$

pour a variant de 1 à m.

La remarque préalable essentielle est la suivante : il s'agit d'un modèle linéaire mixte M3 à structure bloc diagonale (voir 3.1.1). On considère que les v_a sont indépendants, de même loi centrée et de variance σ_v^2 et que les e_a sont indépendants, chaque e_a étant centré et de variance connue Ψ_a. On reprend alors les notations de la partie consacrée aux modèles à structure bloc diagonale et on identifie ainsi les deux approches :

Modèle linéaire mixte (M3) bloc diagonal	Modèle de Fay et Herriot (M1)
Y_a	$\hat{\theta}_a$
X_a	z_a^T
Z_a	b_a
G_a	σ_v^2
R_a	Ψ_a
n_a	1

[24] C'est très souvent le cas, ne serait-ce que sur le signe : un revenu moyen par exemple doit être positif.

On notera bien que pour tout a, on a $n_a = 1$: chaque sous modèle issu de M3 porte sur un estimateur $\hat{\theta}_a$ réel (dans l'approche générale, Y_a est vectoriel).

On en déduit

$$V_a = \Psi_a + \sigma_v^2 b_a^2$$

On s'intéresse à

$$\theta_a = z_a^T \beta + b_a v_a$$

c'est-à-dire à $\mu_a = l_a^T \beta + m_a^T \cdot v_a$ avec $l_a = z_a$ et $b_a = m_a$.

L'estimateur SBOL de $\theta_a (= \mu_a)$ est donc, d'après la formule du 3.1.1. :

$$\boxed{\hat{\theta}_a^H = z_a^T \tilde{\beta} + \gamma_a \left(\hat{\theta}_a - z_a^T \tilde{\beta} \right)}$$

avec $$\gamma_a = \frac{b_a^2 \sigma_v^2}{\Psi_a + b_a^2 \sigma_v^2}$$

et $$\tilde{\beta} = \left[\sum_{a=1}^{m} \frac{z_a z_a^T}{\Psi_a + b_a^2 \sigma_v^2} \right]^{-1} \cdot \left[\sum_{a=1}^{m} \frac{z_a \hat{\theta}_a}{\Psi_a + b_a^2 \cdot \sigma_v^2} \right]$$

Le paramètre restant à estimer - et dont la présence interdit le calcul en l'état de $\hat{\theta}_a^H$ - est la variance σ_v^2, soit

$$\delta = \sigma_v^2 \qquad (\delta \in \mathrm{R})$$

On notera l'écriture alternative de $\hat{\theta}_a^H$:

$$\hat{\theta}_a^H = \underbrace{\gamma_a \cdot \hat{\theta}_a}_{\substack{\text{Estimateur} \\ \text{direct}}} + (1 - \gamma_a) \underbrace{z_a^T \tilde{\beta}}_{\substack{\text{Estimateur} \\ \text{synthétique}}}$$

Ainsi, puisque $\gamma_a \in [0,1]$, $\hat{\theta}_a^H$ apparaît comme une combinaison linéaire de l'estimateur direct $\hat{\theta}_a$ et de l'estimateur synthétique $z_a^T \tilde{\beta}$, défini dans l'esprit du chapitre sur l'estimation indirecte avec modélisation implicite. L'estimateur $\hat{\theta}_a^H$ est donc, ni plus ni moins, un estimateur composite de θ_a.

L'interprétation est simple : si b_a est petit (donc si l'influence de v_a est faible) ou/et si σ_v^2 est petit (même effet : l'influence de v_a est faible, puisque v_a tend vers un effet fixe constant, par hypothèse égal à 0), γ_a est proche de 0 et $\hat{\theta}_a^H$ est « presque » l'estimateur synthétique. C'est tout à fait logique : si l'impact de v_a disparaît, le vrai θ_a tend vers sa partie déterministe $z_a^T \beta$, qui représente donc une base d'estimation très fiable - en tout cas plus fiable que $\hat{\theta}_a$, qui résulte d'un sondage et qui subit donc une erreur d'échantillonnage. On remarquera que :

$$\gamma_a = \frac{Variance\left(b_a v_a\right)}{Variance\left(b_a v_a\right) + Variance\left(e_a\right)}$$

Inversement, si Ψ_a est faible, γ_a tend vers 1 et l'estimateur direct reprend l'avantage : sans biais (par hypothèse) et de faible variance, il devient de très bonne qualité, et en tout cas préférable à l'estimateur synthétique qui reste pénalisé par son biais. Cette philosophie rappelle un principe d'estimation bien connu en statistique : quand on dispose de deux estimateurs sans biais et indépendants, la combinaison linéaire sans biais optimale pondère chaque estimateur initial par l'inverse de sa variance. Ici le contexte est tout à fait différent, mais néanmoins on retrouve ce genre de raisonnement.

Il est possible que l'on ait à modifier in fine la valeur de $\hat{\theta}_a^H$ pour tenir compte de contraintes de définition du paramètre vrai : par exemple si θ_a est une proportion, il faudra s'assurer que $\hat{\theta}_a^H$ se trouve entre 0 et 1. Si θ_a est un effectif, il faudra que $\hat{\theta}_a^H$ soit entier[25]. Ces opérations réduisent certes les atouts théoriques de l'estimateur, mais en pratique elles doivent être rares et avoir des conséquences insignifiantes.

L'expression de l'estimateur composite $\hat{\theta}_a^H$ a le grand mérite de réconcilier les deux grandes approches - plutôt concurrentes - de l'estimation / prédiction : d'une part l'approche dans laquelle l'aléa est l'aléa de sondage (donc sans modèle de comportement), d'autre part l'approche classique par modélisation (le modèle de comportement est central). Dans le contexte présent, on n'a pas besoin de prendre parti a priori pour l'une ou l'autre de ces approches : le contexte amène naturellement à les « mixer » de manière optimale. Evidemment, si on raisonne avec l'aléa de sondage seulement, il faut accepter l'existence d'un biais : l'estimateur optimum est construit pour être sans biais si on prend en compte l'ensemble des aléas, mais pas pour être sans biais si on fait abstraction de l'aléa du modèle. Autrement dit, $\hat{\theta}_a^H$ **est bien dépendant du modèle.**

Le développement précédent repose en particulier sur l'hypothèse d'indépendance des erreurs d'échantillonnage e_a, ce qui ne se réalise pas dans tous les contextes. Nous avons déjà signalé (voir 2.1.1) que ce n'était pas réaliste dans le cas d'enquêtes à plusieurs degrés lorsque les domaines sont des sous-populations qui recoupent les unités primaires. Techniquement, l'absence d'indépendance conduit à une matrice R qui n'est pas diagonale (pas plus que V), et donc on n'est plus dans le cadre du modèle de type « bloc diagonal ». Ce n'est pas rédhibitoire parce que la théorie générale présentée au début du 3.1.1 s'applique parfaitement, mais les expressions des estimateurs in fine obtenus sont évidemment beaucoup plus complexes et difficiles à interpréter.

On remarque que si σ_v^2 est nul, alors γ_a est nul et l'estimateur $\hat{\theta}_a^H$ coïncide avec l'estimateur synthétique. Cela permet de dire que l'estimateur synthétique introduit au 2.2.1 est l'estimateur optimum lorsqu'il n'y a pas d'effets aléatoires spécifiques aux domaines considérés. Autrement dit, **on peut résumer ainsi la situation : en présence d'information auxiliaire explicative (c'est z_a), on utilise de manière « basique » l'estimateur synthétique. Si on considère[26] que cette information est insuffisante pour résumer la spécificité des différents domaines, on change d'optique en rajoutant un effet aléatoire et on adopte de fait les modèles dits « à effet aléatoire ».**

[25] Cela étant, si θ_a est une proportion ou un effectif, le modèle de Fay et Herriot n'est pas le plus adapté et on adopte en général un modèle de prédicteur dit « Bayésien empirique » - voir partie 4.

[26] Après un eventuel test de nullité de σ_v. Un tel test est possible dès que m est « grand » puisqu'on dispose d'estimateurs du maximum de vraisemblance (voir 3.3.3), dont le loi est approximativement Gaussienne.

3.3.2. Qualité de l'estimateur SBOL.

On a les propriétés suivantes :

- Lorsque n_a tend vers N_a, Ψ_a tend vers 0 et γ_a tend vers 1 : $\hat{\theta}_a^H$ tend vers $\hat{\theta}_a$, qui tend lui-même vers θ_a : on a une propriété de convergence sympathique.
- Si on ne considère que le biais associé à l'aléa de sondage, puisque $\hat{\theta}_a$ est sans biais de θ_a, on a

$$\underset{P}{E}\hat{\theta}_a^H - \theta_a \approx (1-\gamma_a)\cdot\left(z_a^T \cdot \underset{P}{E}\tilde{\beta} - \theta_a\right)$$

où $\underset{P}{E}$ désigne l'espérance par rapport à l'aléa de sondage. (θ_a étant fixé - c'est donc conditionnel à θ_a). On peut donc dire que $\hat{\theta}_a^H$ est biaisé conditionnellement aux θ_a (c'est-à-dire conditionnellement à l'aléa du modèle). Si on prend l'espérance par rapport au modèle (opérateur $\underset{m}{E}$) de ce biais, on trouve 0, soit

$$\underset{m}{E}\left[\underset{P}{E}\hat{\theta}_a^H - \theta_a\right] = 0$$

Cela est tout à fait normal, puisque $\hat{\theta}_a^H$ est l'estimateur SBOL - donc sans biais (sous entendu par rapport à la superposition des deux aléas en jeu, celui de l'échantillonnage et celui du modèle).

- En appliquant les résultats du 3.1.1., on trouve

$$\boxed{EQM\left(\hat{\theta}_a^H\right) = \gamma_a\Psi_a + (1-\gamma_a)^2 z_a^T\left(\sum_{a=1}^{m}\frac{z_a z_a^T}{\Psi_a + \sigma_v^2 b_a^2}\right)^{-1} z_a}$$

L'EQM est la somme de deux termes. Le second est d'ordre de grandeur $1/m$, si bien que si m est « grand », le terme numériquement prépondérant sera $\gamma_a\Psi_a$. Dans ces conditions, on ne peut pas grand chose sur Ψ_a, mais en revanche si γ_a est petit, $EQM\left(\hat{\theta}_a^H\right)$ sera petite devant Ψ_a, qui est l'EQM de l'estimateur direct $\hat{\theta}_a$. On retrouve bien, au travers de cette formule, ce qui fait la force de l'estimateur composite : si la partie synthétique a de l'influence (donc si γ_a s'éloigne de manière un peu sensible de 1), elle va contribuer à la stabilisation de $\hat{\theta}_a^H$ parce qu'elle possède une faible variance relative dès lors que σ_v^2 est plutôt petit par rapport aux Ψ_a. On remarquera ainsi que plus m est grand, plus l'erreur sera petite (plus on a de points pour l'estimation de modèle, plus on stabilise $\tilde{\beta}$).

Cela conforte bien dans l'idée que $\hat{\theta}_a^H$ **s'emploie a priori lorsque le nombre de domaines étudiés (ici m) est grand.**

3.3.3. Estimateur ESBOL

Il reste, concrètement, à estimer au mieux σ_v^2. On peut utiliser, comme cela a déjà été signalé dans la présentation générale, la méthode des moments (qui s'applique dans n'importe quel contexte) ou la méthode du maximum de vraisemblance (qui nécessite cette fois une hypothèse de loi sur v et e, en pratique une loi de Gauss). On a vu que l'EMV résultait d'un processus fort

compliqué, mais ici il s'avère accessible. Dans le contexte présent, la méthode des moments reste également assez abordable. Pour s'en convaincre, on donne ci dessous des algorithmes permettant d'estimer σ_v^2.

- Deux variantes possibles pour la méthode des moments :

  a) Construire la suite réelle x_n de manière récurrente selon :

$$x_0 = 0$$

$$x_{n+1} = Max\left(x_n + \frac{m - p - A_n}{B_n}, 0 \right)$$

où $A_n = \sum_{a=1}^{m} \frac{\left(\hat{\theta}_a - z_a^T \tilde{\beta}_n\right)^2}{\Psi_a + b_a^2 \cdot x_n}$

$$B_n = -\sum_{a=1}^{m} b_a^2 \frac{\left(\hat{\theta}_a - z_a^T \tilde{\beta}_n\right)^2}{\left(\Psi_a + x_n \cdot b_a^2\right)^2}$$

avec $\tilde{\beta}_n$ obtenu en remplaçant σ_v^2 par x_n dans l'expression de $\tilde{\beta}$.

La suite x_n ainsi définie converge (même rapidement, souvent en moins d'une dizaine d'itérations) : on considère donc la limite empirique (prendre un terme d'ordre suffisamment élevé pour que x_n n'évolue plus), qui constituera l'estimation cherchée.

  b) Calculer :

$$\hat{\tilde{\beta}} = \left(\sum_{a=1}^{m} \frac{1}{b_a^2} z_a z_a^T \right)^{-1} \cdot \left(\sum_{a=1}^{m} \frac{1}{b_a^2} z_a \hat{\theta}_a \right)$$

puis

$$Max\left(0, \frac{1}{m-p} \left[\sum_{a=1}^{m} \left(\frac{\hat{\theta}_a}{b_a} - \frac{z_a^T \hat{\tilde{\beta}}}{b_a} \right)^2 - \sum_{a=1}^{m} \frac{\Psi_a}{b_a^2} \left(1 - \frac{1}{b_a^2} z_a^T \left(\sum_{a=1}^{m} \frac{z_a z_a^T}{b_a^2} \right)^{-1} z_a \right) \right] \right)$$

Cette dernière valeur est une estimation de σ_v^2 de type « moments ». La méthode a) est cependant plus efficace si m est grand.

- Le calcul de l'estimateur du maximum de vraisemblance s'appuie également sur une suite récurrente :

$$y_0 = 0$$

$$y_{n+1} = y_n + \left[\sum_{a=1}^{m} \frac{b_a^4}{\left(\Psi_a + b_a^2 y_n\right)^2} \right]^{-1} \cdot \left[-\sum_{a=1}^{m} \frac{b_a^2}{\Psi_a + b_a^2 y_n} + \sum_{a=1}^{m} b_a^2 \frac{\left(\hat{\theta}_a - z_a^T \tilde{\beta}_n\right)^2}{\left(\Psi_a + b_a^2 y_n\right)} \right]$$

où $\tilde{\beta}_n$ a la même définition que dans la méthode des moments a) (mais σ_v^2 est remplacé par y_n). Cette suite converge vers l'EMV de σ_v^2. Si m est grand, on montre que l'EMV est plus efficace que les méthodes des moments exposées ci-dessus.

L'estimateur du maximum de vraisemblance a le très gros avantage d'avoir une loi asymptotiquement Gaussienne, donc si m est grand on sera en mesure de tester la nullité de σ_v^2 (cas envisageable si les petits domaines sont des ZUS, ou des agglomérations, voire des départements. S'il s'agit de régions, on n'est pas dans les conditions asymptotiques…).

Quelle que soit la méthode choisie, on aboutit finalement à une estimation $\hat{\sigma}_v^2$ de σ_v^2. Il suffit de remplacer σ_v^2 par $\hat{\sigma}_v^2$ dans $\hat{\theta}_a^H$ pour obtenir l'estimateur ESBOL.

$$\boxed{\hat{\theta}_{E,a}^H = \hat{\gamma}_a \hat{\theta}_a + (1-\hat{\gamma}_a) z_a^T \hat{\beta}}$$

où $$\hat{\gamma}_a = \frac{b_a^2 \cdot \hat{\sigma}_v^2}{\Psi_a + b_a^2 \cdot \hat{\sigma}_v^2}$$

et $$\hat{\beta} = \left[\sum_{a=1}^{m} \frac{z_a z_a^T}{\Psi_a + b_a^2 \hat{\sigma}_v^2} \right]^{-1} \cdot \left[\sum_{a=1}^{m} \frac{z_a \hat{\theta}_a}{\Psi_a + b_a^2 \hat{\sigma}_v^2} \right]$$

Cet estimateur est toujours calculable (on rappelle que les Ψ_a sont supposés être connus et que les b_a^2 sont des termes fixés dans le modèle). Fay et Herriot suggèrent d'appliquer la stratégie suivante :

- Si $\hat{\theta}_{E,a}^H \in \left[\hat{\theta}_a - \sqrt{\Psi_a}, \hat{\theta}_a + \sqrt{\Psi_a} \right]$, utiliser $\hat{\theta}_{E,a}^H$
- Si $\hat{\theta}_{E,a}^H < \hat{\theta}_a - \sqrt{\Psi_a}$, utiliser $\hat{\theta}_{E,a}^H - \sqrt{\Psi_a}$
- Si $\hat{\theta}_{E,a}^H > \hat{\theta}_a + \sqrt{\Psi_a}$, utiliser $\hat{\theta}_{E,a}^H + \sqrt{\Psi_a}$

Cela conduit à un estimateur EBLUP adapté, noté $\tilde{\theta}_{E,a}^H$. On achève la procédure en se ramenant à l'estimation de $\overline{Y}_a$ selon :

$$\hat{\overline{Y}}_a = g^{-1}\left(\tilde{\theta}_{E,a}^H \right)$$

3.4. L'estimation « ESBOL » appliquée aux modèles conçus au niveau « individu » :

3.4.1. Expression de l'estimateur SBOL

On considère le modèle M2 présenté au 2.2. soit (rappel) :

$$Y_{a,i} = X_{a,i}^T \cdot \beta + v_a + e_{a,i}$$

avec $E(e_{a,i}) = 0$ et $V(e_{a,i}) = (k_{a,i})^2 \cdot \sigma_e^2$

On note Y_a le vecteur de taille n_a dont la ième composante est $Y_{a,i}$, X_a la matrice dont la ième ligne est $X_{a,i}^T$, e_a le vecteur de taille n_a dont la ième composante est $e_{a,i}$. On note 1_{n_a} le vecteur de taille n_a dont toutes les composantes valent 1. On a alors l'écriture matricielle, pour tout a (de 1 à m) :

$$\boxed{Y_a = X_a \beta + v_a \cdot 1_{n_a} + e_a}$$

Ici, v_a est un réel (rappel). En posant $Z_a = 1_{n_a}$ on obtient un modèle de type « variance bloc diagonale », puisqu'on a (voir notations du 3.1.1.) :

$$G_a = Var(v_a) = \sigma_v^2$$

$$R_a = Var(e_a) = \sigma_e^2 \cdot Diag(k_{a,i})^2$$

où $Diag(k_{a,i})^2$ est la matrice $n_a X\ n_a$ de type diagonale, dont la diagonale est constituée des valeurs $k_{a,i}^2$. Donc

$$V_a = R_a + \sigma_v^2 \left(1_{n_a} \cdot 1_{n_a}^T\right)$$

On a vu que dans ces conditions, l'estimateur SBOL de β et des v_a est :

$$\tilde{\beta} = \left(\sum_{a=1}^{m} X_a^T \cdot V_a^{-1} \cdot X_a\right)^{-1} \cdot \left(\sum_{a=1}^{m} X_a^T \cdot V_a^{-1} \cdot Y_a\right)$$

et

$$\tilde{v}_a = G_a \cdot Z_a^T \cdot V_a^{-1}\left(Y_a - X_a \tilde{\beta}\right)$$

Comme v_a est un réel, on a la chance de pouvoir aboutir à une expression « relativement simple » de V_a^{-1}, puisque

$$V_a^{-1} = \frac{1}{\sigma_e^2}\left[diag(\lambda_{ai}) - \frac{\gamma_a}{\lambda_{a\bullet}} \lambda_a \lambda_a^T\right]$$

en posant

$$\lambda_{ai} = 1/k_{a,i}^2$$

$$\lambda_{a\bullet} = \sum_{i \in s_a} \lambda_{ai}$$

et λ_a est le vecteur de taille n_a dont les coordonnées sont les λ_{ai}, i décrivant s_a.

$$\boxed{\gamma_a = \frac{\sigma_v^2}{\sigma_v^2 + \frac{\sigma_e^2}{\lambda_{a\bullet}}}}$$

On notera que, dans le cas général (celui du modèle M3), on ne sait pas exprimer l'inverse de V_a, et donc $\tilde{\beta}$ et $\tilde{v}_a$ gardent des expressions matricielles qui ne permettent pas de simplifier l'estimation de γ_a. A ce stade, comme cela a été fait au 2.2.1, il faut distinguer deux cas de figure :

- <u>CAS 1</u> :

  C'est le cas où N_a est grand, et le paramètre à estimer est le réel

  $$\overline{Y}_a = \mu_a \approx \overline{X}_a^T \beta + v_a$$

  où $\overline{X}_a$ est la vraie moyenne sur l'ensemble du domaine a, soit $\frac{1}{N_a}\sum_{i=1}^{N_a} X_{a,i}$.

Alors l'estimateur SBOL de $\overline{Y}_a$ est

$$\hat{\mu}_a^H = \overline{X}_a^T \tilde{\beta} + \tilde{v}_a$$

On vérifie facilement que

$$\tilde{\beta} = \left(\sum_{i \in s_a} \lambda_{ai} \cdot X_{a,i} X_{a,i}^T - \gamma_a \cdot \lambda_a \cdot \overline{x}_a^{\lambda} \cdot \overline{x}_a^{\lambda T} \right)^{-1} \cdot \left(\sum_{i \in s_a} \lambda_{a,i} X_{a,i} \cdot Y_{a,i} - \gamma_a \lambda_a \overline{x}_a^{\lambda} \cdot \overline{y}_a^{\lambda} \right)$$

où $\quad \overline{x}_a^{\lambda} = \dfrac{1}{\lambda_{a\bullet}} \sum_{i \in s_a} \lambda_{a,i} \cdot X_{a,i}$

$$\overline{y}_a^{\lambda} = \frac{1}{\lambda_{a\bullet}} \sum_{i \in s_a} \lambda_{a,i} \cdot Y_{a,i}$$

et $\quad \tilde{v}_a = \gamma_a \left(\overline{y}_a^{\lambda} - \overline{x}_a^{\lambda T} \cdot \tilde{\beta} \right)$

On voit donc apparaître des moyennes PONDEREES des $X_{a,i}$ et des $Y_{a,i}$ au sein du domaine a. La pondération est celle des $1/k_{a,i}^2$, où on rappelle que $k_{a,i}^2 \cdot \sigma_e^2$ est la variance de $e_{a,i}$. Dans le cas le plus simple, les $e_{a,i}$ ont tous même variance, et puisque alors $k_{a,i} = 1$, $\overline{x}_a^{\lambda}$ et $\overline{y}_a^{\lambda}$ se présentent comme de simples moyennes arithmétiques (ce que l'on notait auparavant $\overline{x}_a$ et $\overline{y}_a$). Finalement,

$$\hat{\mu}_a^H = \overline{X}_a^T \tilde{\beta} + \gamma_a \left(\overline{y}_a^{\lambda} - \overline{x}_a^{\lambda T} \tilde{\beta} \right)$$

On en tire

$$\boxed{\hat{\mu}_a^H = \hat{\overline{Y}}_a^H = \gamma_a \left[\overline{y}_a^{\lambda} + \left(\overline{X}_a - \overline{x}_a^{\lambda} \right)^T \tilde{\beta} \right] + \left(1 - \gamma_a \right) \overline{X}_a^T \tilde{\beta}}$$

Comme $\gamma_a \in [0,1]$, $\hat{\mu}_a^H$ apparaît comme combinaison linéaire d'un estimateur de type synthétique $\left(\overline{X}_a^T \tilde{\beta} \right)$ et d'un estimateur de type régression, dont le statut est un peu bâtard : ce n'est manifestement pas un estimateur direct parce que $\tilde{\beta}$ se calcule à partir de données collectées sur plusieurs domaines (donc $\hat{\mu}_a^H$ n'est pas un estimateur composite à proprement parler !), mais il a une qualité qui est proche de celle d'un estimateur direct à cause de la présence du $\overline{y}_a^{\lambda}$. De ce point de vue, on se trouve plutôt dans l'esprit de l'estimateur présenté au 2.2.3. Sauf si $k_{ai} = 1$, auquel cas on retombe sur une expression plus familière, cet estimateur de type régression est tout à fait original puisque les moyennes sont pondérées par les $\lambda_{a,i}$. De plus, si l'échantillonnage est complexe (c'est-à-dire en fait si les probabilités de sélection sont inégales), les poids de sondage n'apparaissent pas dans $\hat{\mu}_a^H$. De ce fait, avec un tel plan, $\hat{\mu}_a^H$ est un estimateur biaisé de $\overline{Y}_a$ si on s'en tient à l'aléa classique de sondage. On sait néanmoins adapter l'estimateur au cas de pondération inégales : la théorie n'est pas développée dans ce document parce qu'elle augmente sensiblement la complexité des estimateurs. Le cas des pondérations inégales se rencontre dans les enquêtes ménages de l'Insee lorsqu'on sur-représente certaines catégories de ménages (donc à peu près une fois sur deux). Le lecteur intéressé pourra consulter l'ouvrage de JNK Rao (page 148).

Si σ_v^2 est petit, cela signifie que les v_a sont souvent proches de zéro, et donc que $\overline{Y}_a$ est proche de $\overline{X}_a^T \beta$: le modèle explicatif est très bon, donc l'estimateur du type synthétique s'impose logiquement. Ce raisonnement est cohérent avec l'expression de $\hat{\mu}_a^H$, puisque alors γ_a est proche de 0. De même, si n_a devient grand, λ_a• sera grand et γ_a sera proche de 1 : donc $\hat{\mu}_a^H$ sera presque l'estimateur de type régression. Là encore, c'était attendu puisque plus l'échantillon est de grande taille, plus on a tendance à faire confiance à l'estimateur le « plus proche » de l'estimateur direct.

En conclusion, on retiendra que si $k_{a,i}$ diffère de 1, l'estimateur $\hat{\mu}_a^H$ n'a pas de bonnes propriétés du point de vue du seul aléa de sondage (estimateur biaisé et non convergent, c'est-à-dire ne se rapprochant pas a priori de $\overline{Y}_a$ si n_a devient très grand), en revanche il est par construction sans biais par rapport à l'aléa du modèle. Si $k_{a,i}$ vaut 1, le paysage s'éclaircit mais si l'échantillonnage est complexe, il faut prendre en compte les poids de sondage dans une expression $\hat{\mu}_{a,w}^H$ adaptée afin d'obtenir (au moins) une propriété de convergence (en revanche, l'absence de biais est irréaliste à cause de la présence d'une partie synthétique).

- <u>CAS 2 :</u>

C'est le cas contraire au précédent, celui où N_a n'est pas considéré comme grand. Dans de telles circonstances, il faut passer par l'écriture suivante (voir 2.2.) où $f_a = n_a / N_a$:

$$\overline{Y}_a = f_a \cdot \overline{y}_a + (1 - f_a)\overline{y}_a^*$$

Puisque $\overline{y}_a$ est connu, il reste à obtenir la prédiction SBOL de $\overline{y}_a^*$, qui vaut :

$$\hat{\overline{Y}}_a^H = \overline{X}_a^{*T} \tilde{\beta} + \tilde{v}_a$$

à $\overline{X}_a^*$ représente la vraie moyenne des $X_{a,i}$ sur tous les individus i de a qui ne sont PAS dans s_a, soit

$$\overline{X}_a^* = \frac{N_a \overline{X}_a - n_a \overline{x}_a}{N_a - n_a}$$

Cette valeur est connue. L'estimateur SBOL de $\overline{Y}_a$ est donc

$$\hat{\overline{Y}}_a^H = f_a \overline{y}_a + (1 - f_a)\left[\overline{X}_a^{*T} \cdot \tilde{\beta} + \tilde{v}_a\right]$$

Si on remplace $\tilde{v}_a$ par son expression fonction de $\tilde{\beta}$ et des $X_{a,i}$, on obtient l'estimateur SBOL de $\overline{Y}_a$:

$$\boxed{\hat{\overline{Y}}_a^H = f_a \overline{y}_a + (1 - f_a)\left[\gamma_a \left(\overline{y}_a^\lambda + \left(\overline{X}_a^* - \overline{x}_a^\lambda\right)^T \tilde{\beta}\right) + (1 - \gamma_a)\overline{X}_a^{*T} \cdot \tilde{\beta}\right]}$$

La partie entre crochets est semblable à celle que l'on obtenait dans le cas 1, mais la vraie moyenne $\overline{X}_a$ a été remplacée par la moyenne sur la partie non échantillonnée (qui est une variable aléatoire et non plus une « vraie valeur »). Les expressions de γ_a, $\overline{x}_a^\lambda$, $\overline{y}_a^\lambda$ et de $\tilde{\beta}$ sont exactement celles du cas 1.

3.4.2. Estimation des variances intervenant dans le modèle

Si on suppose que les v_a et e_a suivent des lois de Gauss, on peut utiliser l'estimateur du maximum de vraisemblance, concrètement obtenu par la PROC MIXED de SAS. Sinon (donc dans tous les cas…), on peut procéder ainsi :

- Pour estimer σ_e^2 :
  - Régresser les $\left(Y_{a,i} - \bar{y}_a^{\lambda}\right)/k_{a,i}$ sur les composantes non nulles de $\left(X_{a,i} - \bar{x}_a^{\lambda}\right)/k_{a,i}$, sous condition $n_a > 1$
  - Calculer la somme des carrés des résidus, $SCR(e)$, et former finalement

$$\hat{\sigma}_e^2 = \frac{SCR(e)}{n - m - p_1}$$

où p_1 est le nombre d'individus (a,i) vérifiant $X_{a,i} \neq x_a^{\lambda}$.

- Pour estimer σ_v^2 :
  - Régresser les $Y_{a,i}/k_{a,i}$ sur les composantes de $X_{a,i}/k_{a,i}$
  - Calculer la somme des carrés des résidus, $SCR(v)$ et former finalement

$$\hat{\sigma}_V^2 = Max\left(0, \frac{1}{\mu}\left[SCR(v) - (n-p)\hat{\sigma}_e^2\right]\right)$$

où $$\mu = \sum_{a=1}^{m} \lambda_{a\bullet}\left[1 - \lambda_{a\bullet}\bar{x}_a^{\lambda T}\left(\sum_{a=1}^{m}\sum_{i=1}^{N_a} \lambda_{a,i} X_{a,i} X_{a,i}^T\right)^{-1} \bar{x}_a^{\lambda}\right]$$

3.4.3. Estimateur ESBOL

Il suffit de reprendre l'estimateur SBOL et de remplacer les paramètres inconnus σ_e^2 et σ_v^2 par leurs estimateurs (voir 3.4.2). On obtient l'estimateur ESBOL :

$$\boxed{\hat{\bar{Y}}_{a,E}^H = f_a \bar{y}_a + (1 - f_a)\left[\hat{\gamma}_a \cdot \left(\bar{y}_a^{\lambda} + \left(\bar{X}_a^* - \bar{x}_a^{\lambda}\right)^T \hat{\beta}\right) + (1 - \hat{\gamma}_a)\bar{X}_a^* \hat{\beta}\right]}$$

où $\hat{\beta}$ et $\hat{\gamma}_a$ ont les expressions de $\tilde{\beta}$ et γ_a dans lesquelles σ_v^2 et σ_e^2 ont été remplacés par $\hat{\sigma}_v^2$ et $\hat{\sigma}_e^2$.

Dans certaines conditions « techniques », on sait exprimer un estimateur peu biaisé de l'erreur quadratique moyenne de $\hat{\bar{Y}}_{a,E}^H$. Cela étant, il s'agit d'une expression extrêmement complexe.

4. La classe des prédicteurs optimums (dits « Bayésiens empiriques »).

4.1. Présentation générale du concept et de la méthode.

Dans l'approche SBOL (toute celle de la partie 3), on a pour objectif de prédire une variable aléatoire μ qui représente la grandeur qui nous intéresse. Pour cela, on cherche une fonction linéaire des Y_a (modèle au niveau du domaine) ou des $Y_{a,i}$ (modèle au niveau de l'individu) qui en soit « la plus proche » au sens du risque quadratique moyen. En particulier, le vecteur β défini dans le(s) modèle(s) n'est pas identifié a priori comme étant un paramètre qu'il faut estimer, au contraire de σ_e^2 et de σ_v^2. En effet, le vecteur β est bien une composante de l'estimateur SBOL, mais il n'apparaît que sous forme estimée. Par ailleurs, il n'est pas nécessaire de faire des hypothèses sur la forme des lois des variables aléatoires v et e. Enfin, les modèles linéaires mixtes de niveau « individuel » (M2 ou M3) qui servent de base à toute cette théorie ne sont adaptés qu'aux variables quantitatives continues.

Dans la partie qui suit, on adopte une approche différente.

En premier lieu, on change l'objectif : il ne s'agit plus de chercher un estimateur optimum parmi la classe particulière des estimateurs linéaires et sans biais, mais de chercher un optimum parmi l'ensemble des estimateurs, sans condition (de ce point de vue, l'approche est plus puissante). En second lieu on cherche une méthode qui ne se limite pas aux variables quantitative continues, mais qui puisse aussi s'appliquer aux variables individuelles binaires 0-1 et aux dénombrements (c'est-à-dire à des paramètres d'intérêt qui sont des proportions et des comptages). On se place donc dans le cadre d'objectifs sensiblement plus larges que dans la partie précédente. Il y a possibilité théorique de satisfaire ces objectifs, mais il y a une contrepartie de taille : il faut faire une hypothèse sur les lois des aléas v et e. Traditionnellement, afin que les calculs restent abordables et parce qu'il s'agit le plus souvent d'une bonne approximation de la réalité, on postule des lois de Gauss pour v et pour e.

Le fondement technique essentiel de cette approche est le suivant : lorsqu'on dispose d'une variable aléatoire Y quelconque dont on observe une réalisation et que l'on cherche à prédire une autre variable aléatoire μ au moyen d'une fonction de Y (notée $f(Y)$), le prédicteur optimal au sens de l'écart quadratique moyen est l'espérance de μ conditionnelle à Y, soit :

$$f(Y) = E[\mu|Y]$$

Cette fonction a donc la propriété suivante : quel que soit le prédicteur $g(Y)$,

$$E[g(Y)-\mu]^2 \geq E[f(Y)-\mu]^2$$

De plus, si $Y = X\beta + Zv + e$, si μ s'écrit (comme dans le chapitre précédent) $l^T \cdot \beta + m^T \cdot v$ (β est fixe, inconnu, mais v est une variable aléatoire) et que (v,e) suit une loi de Gauss, alors $f(Y)$ a exactement l'expression de l'estimateur SBOL dans lequel $\tilde{\beta}$ est remplacé par β (qui est un « simple paramètre »), soit

$$f(Y) = l^T\beta + m^T G Z^T V^{-1}(Y - X\beta)$$

Cette propriété est assez remarquable. Dans notre contexte, il faut bien voir qu'il y a des paramètres inconnus (β, au moins, en tant que composant de μ). Il va falloir estimer ces paramètres. La démarche globale sera la suivante :

a/ On considère les deux densités sur lesquelles portent les hypothèses de modélisation, soit

- La densité de μ
- La densité de Y sachant μ

Ces deux densités sont paramétrées. En toute généralité, les paramètres impliqués ne sont pas les mêmes. On note λ_2 le paramètre (vectoriel) qui apparaît dans la densité de μ - qui s'écrit donc $f(\mu\,;\lambda_2)$ - et on note λ_1 le paramètre (vectoriel) qui apparaît dans la densité de Y sachant μ - que l'on notera $f(Y|\mu\,;\lambda_1)$.

b/ On utilise la formule de Bayes pour calculer la densité conditionnelle de μ sachant Y (qui est paramétrée cette fois par λ_1 et par λ_2) :

$$f(\mu|Y\,;\lambda_1,\lambda_2)=\frac{f(Y|\mu\,;\lambda_1)\cdot f(\mu\,;\lambda_2)}{\int f(Y|\mu;\lambda_1)\cdot f(\mu\,;\lambda_2)d\mu}$$

c/ On exprime l'espérance conditionnelle optimale cherchée, soit

$$\hat{\mu}^{OPTI}=E(\mu|Y\,;\lambda_1,\lambda_2)=\int \mu\cdot f(\mu|Y\,;\lambda_1,\lambda_2)d\mu$$

A ce stade, c'est le prédicteur optimum théorique (incalculable car on ne connaît pas λ_1 ni λ_2).

d/ On estime les paramètres (λ_1,λ_2) à partir de la densité $f(Y\,;\lambda_1,\lambda_2)=\int f(Y|\mu\,;\lambda_1)\cdot f(\mu\,;\lambda_2)d\mu$ par une méthode quelconque (par exemple le maximum de vraisemblance). On obtient $(\hat{\lambda}_1,\hat{\lambda}_2)$ et on termine en calculant

$$E(\mu|Y,\hat{\lambda}_1,\hat{\lambda}_2)$$

qui est le prédicteur optimum, dit « empirique » (en toute rigueur, il n'est plus optimum puisqu'on a estimé λ_1 et λ_2). On le notera $\hat{\mu}_E^{OPTI}$ (indice E comme « Empirique »).

Dans la littérature, $\hat{\mu}_E^{OPTI}$ est souvent appelé « Bayésien empirique », mais c'est une appellation abusive parce qu'il n'y a rien de bayésien dans cette mécanique. Cette dénomination provient probablement de l'existence de l'étape b/ qui calcule une loi « a posteriori » à partir des lois « a priori » révélées par l'étape a/ .

4.2. Application au cas du modèle de Fay et Herrict.

On reprend le modèle M1 du 2.1.1., mais en introduisant une hypothèse de normalité, ce qui donne (pour tout a de 1 à m) :

$$\hat{\theta}_a=\theta_a+e_a \text{ avec } e_a\to\mathcal{N}(0,\Psi_a)$$

$$\theta_a=z_a^T\beta+b_a v_a \text{ avec } v_a\to\mathcal{N}(0,\sigma_v^2)$$

Ces hypothèses constituent l'étape initiale a/. On rappelle que les e_a et les v_a sont mutuellement indépendants. On peut alors en déduire la loi θ_a conditionnelle à $\hat{\theta}_a$, les paramètres étant β et σ_v^2 (c'est l'étape b/) :

$$f\left(\theta_a \middle| \hat{\theta}_a \,;\, \beta, \sigma_v^2\right) = \mathcal{N}\left(\hat{\theta}_a^{OPTI}, \gamma_a \Psi_a\right)$$

où $$\gamma_a = \frac{b_a^2 \sigma_v^2}{b_a^2 \sigma_v^2 + \Psi_a}$$

Ψ_a est supposé connu, donc λ_1 est vide et $\lambda_2 = \left(\beta, \sigma_v^2\right)$. Le prédicteur optimum (incalculable) vaut

$$\boxed{\hat{\theta}_a^{OPTI} = E\left[\theta_a \middle| \hat{\theta}_a \,;\, \beta, \sigma_v^2\right] = \gamma_a \hat{\theta}_a + \left(1 - \gamma_a\right) z_a^T \beta \cdot}$$

C'est l'étape c/. Il reste à estimer β et σ_v^2. On a :

$$\hat{\theta}_a \to \mathcal{N}\left(z_a^T \beta, b_a^2 \sigma_v^2 + \Psi_a\right)$$

On est naturellement amené à utiliser le maximum de vraisemblance, ce qui conduit à $\hat{\beta}$ et $\hat{\sigma}_V^2$ (c'est l'étape d/). On en termine en exprimant le prédicteur optimum empirique qui, lui, est calculable :

$$\boxed{\hat{\theta}_{a,E}^{OPTI} = \hat{\gamma}_a \hat{\theta}_a + \left(1 - \hat{\gamma}_a\right) z_a^T \hat{\beta}}$$

C'est le même estimateur que l'ESBOL. Si on est dans le cas particulier de variances d'échantillonnage Ψ_a identiques (égales mettons à Ψ) et que $b_a = 1$, on peut utiliser

$$\hat{\theta}_a^* = \gamma_a^* \hat{\theta}_a + \left(1 - \gamma_a^*\right) z_a^T \hat{\beta}$$

où $$\gamma_a^* = 1 - \frac{m - p - 2}{m - p} \cdot \frac{\Psi}{\Psi + MAX\left(0, \frac{S}{m - p} - \Psi\right)}$$

où p est la taille de β (rappel) et $S = \sum_{a=1}^{m} \left(\hat{\theta}_a - z_a^T \hat{\beta}_{MCO}\right)^2$. On remarquera que si $S \geq (m - p) \cdot \Psi$, l'estimateur $\hat{\theta}_a^*$ est l'estimateur de James-Stein (voir 3.4, chapitre précédent).

Cette approche a l'avantage de permettre un calcul « pas trop compliqué » du prédicteur optimum (empirique) d'une fonction quelconque de θ_a (ce que ne permettait pas l'approche SBOL, car l'optimalité du SBOL ne se transmettait pas quand on considérait une fonction du paramètre). Supposons qu'il s'agisse de $\phi_a = h(\theta_a)$. Le prédicteur empirique optimum est $\hat{\phi}_{a,E}^{OPTI} = E\left[\phi_a \middle| \hat{\theta}_a \,;\, \hat{\beta}, \hat{\sigma}_V^2\right]$, qui peut s'approximer numériquement si on tire K valeurs indépendantes (K grand) dans la loi « a posteriori » de θ_a sachant $\hat{\theta}_a$ - soit pour k de 1 à K, $\hat{\theta}_a^{(k)} \to \mathcal{N}\left(\hat{\theta}_a^{OPTI}, \hat{\gamma}_a \Psi_a\right)$ - puis on forme :

$$\hat{\phi}_{a,E} = \frac{1}{K} \sum_{k=1}^{K} h\left(\hat{\theta}_a^{(k)}\right)$$

qui est presque, d'après la loi des grands nombres (puisque K est grand), l'espérance de $h(X)$ où X suit la loi $f\left(\theta_a \middle| \hat{\theta}_a \,;\, \hat{\beta}, \hat{\sigma}_V^2\right)$, c'est-à-dire $E\left[h(\theta_a)\middle|\hat{\theta}_a \,;\, \hat{\beta}, \hat{\sigma}_V^2\right]$, et donc $\hat{\phi}_{a,E}^{OPTI}$. Cette propriété est très intéressante si on a $\overline{Y}_a = h(\theta_a)$.

L'erreur quadratique moyenne de $\hat{\theta}_{a,E}^{OPTI}$ se calcule exactement comme celle de l'ESBOL -puisqu'il s'agit formellement des mêmes estimateurs. Dans ces conditions, il s'agit d'une erreur conditionnelle à $\hat{\theta}_a$. Il est à noter que dans la littérature on trouve des préconisations de calcul d'erreur par jackknife.

4.3. Cas du modèle linéaire mixte à variance bloc diagonale.

Il s'agit de partir du modèle M3 du 3.1.1. réduit au cas où V est bloc diagonale. Ce modèle généralise le modèle du 4.2. Dans le cas où les variables aléatoires sont normales, ce modèle prend la forme :

$$Y = X\beta + Zv + e$$

ce qui donne après « découpage » :

$$Y_a = \theta_a + e_a \text{, avec } e_a \to \mathcal{N}(0, R_a(\delta))$$

et

$$\theta_a = X_a\beta + Z_a v_a \text{, avec } v_a \to \mathcal{N}(0, G_a(\delta))$$

δ est le paramètre qui intervient dans les matrices de variance de e_a et v_a. Le « paramètre » est $\mu_a = l_a^T\beta + m_a^T v_a$ (donc pas exactement θ_a, mais tout de même une combinaison linéaire de β et de v_a), et son prédicteur optimum est :

$$\hat{\mu}_a^{OPTI} = E\left[\mu_a \middle| Y_a; \delta\right] = l_a^T\beta + m_a^T G_a Z_a^T V_a^{-1}(Y_a - X_a\beta)$$

où

$$V_a(\delta) = R_a + Z_a G_a Z_a^T .$$

On peut en effet déduire la loi de μ_a conditionnelle à Y_a, le paramètre (vectoriel) étant δ :

$$f\left(\mu_a \middle| Y_a \,;\, \delta\right) = \mathcal{N}\left(\hat{\mu}_a^{OPTI}, m_a^T\left(G_a - G_a Z_a^T V_a^{-1} Z_a G_a\right) m_a\right)$$

En estimant β et δ (le maximum de vraisemblance est l'approche la plus tentante) par $\hat{\beta}$ et $\hat{\sigma}$, on aboutit au prédicteur optimum :

$$\boxed{\hat{\mu}_{a,E}^{OPTI} = l_a^T\hat{\beta} + m_a^T\hat{G}_a Z_a^T\hat{V}_a^{-1}\left(Y_a - X_a\hat{\beta}\right)}$$

Le modèle M2 du 2.2. conçu au niveau individu est un cas particulier de modèle linéaire mixte à variance bloc diagonale. Ce type de modèle se traite donc comme ci-dessus.

On pourrait vérifier que ce prédicteur est formellement identique à l'estimateur ESBOL associé au modèle à variance bloc diagonale présenté au 3.1.1.

4.4. Cas des variables qualitatives : paramètre de type « proportion ».

Il s'agit d'estimer des vraies proportions P_a qui traduisent l'importance d'une sous population D (proportion de chômeurs dans la ZUS a par exemple). Un modèle stochastique naturel peut être formulé au niveau individu en partant de la variable individuelle 0 - 1 ainsi définie :

$$Y_{a,i} = 1 \text{ si } (a,i) \in D$$
$$= 0 \text{ sinon}$$

Par exemple, $Y_{a,i}$ vaudra 1 si (a,i) est chômeur et 0 sinon. La loi de $Y_{a,i}$ est donc (nécessairement) de type Bernoulli. Le paramètre de la loi de Bernoulli peut être sensible ou non à l'individu. On distingue

- Soit (cas 1) $Y_{a,i} \to \mathcal{B}(1, P_a)$
- Soit (cas 2) $Y_{a,i} \to \mathcal{B}(1, P_{a,i})$

On voit bien que les lois de Gauss en sont pas adaptées au contexte.
Si on connaît la taille du domaine N_a, l'estimation de l'effectif de la sous population concernée peut être obtenue en calculant $N_a.\hat{P}_a$.

CAS 1

Dans le premier modèle, P_a est notre « paramètre » d'intérêt μ_a. Le modèle suppose que les $Y_{a,i}$ sont indépendants d'un individu à l'autre (dans un domaine donné), ce qui est une hypothèse relativement forte. Dans cette approche, on voit bien que $Y_a = \sum_{i \in s_a} Y_{a,i}$ est une statistique exhaustive (qui contient donc toute l'information), puisqu'il n'y a pas d'information auxiliaire au niveau individuel. On aboutit à

$$Y_a \to \mathcal{B}(n_a, P_a)$$

Y_a (aléatoire) est le nombre total d'individus échantillonnés dans le domaine a qui appartiennent à la sous population D. On remarquera que cette statistique est un simple comptage dans l'échantillon, qui ne prend pas en compte la différenciation des poids de sondage (donc, peu importe la façon dont les individus ont été échantillonnés). Cette première étape fixe donc la loi de Y sachant μ. Une seconde étape doit préciser la loi de μ, c'est-à-dire ici la loi de P_a. A ce niveau, la liberté de choix du modèle semble totale, l'obstacle majeur étant la complexité des calculs associés. On donne ci-après deux exemples relativement classiques, traités dans la littérature :

• *Exemple 1 :* P_a suit une loi bêta (α, β)

Il s'agit de la densité suivante ($\alpha > 0$ et $\beta > 0$) :

$$f(P_a\,;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} P_a^{\alpha-1}(1-P_a)^{\beta-1}$$

Ici, (α, β) joue le rôle du paramètre λ_2. Comme on a

$$f(Y_a|P_a) = \binom{n_a}{y_a} P_a^{Y_a} \cdot (1-P_a)^{n-Y_a}$$

le calcul de l'étape b/ conduit à la densité conditionnelle :

$$f\left(P_a|Y_a\ ;\alpha,\beta\right)=\frac{\Gamma\left(\alpha+\beta+n_a\right)}{\Gamma\left(\alpha+Y_a\right).\Gamma\left(n_a-Y_a+\beta\right)}P_a^{\alpha+Y_a-1}\left(1-P_a\right)^{n_a-Y_a+\beta-1}$$

Autrement dit, P_a conditionnel à Y_a suit une loi bêta $\left(Y_a+\alpha,n_a-Y_a+\beta\right)$. L'étape c/ conduit au prédicteur optimum :

$$\hat{P}_a^{OPTI}=E\left(P_a|Y_a\ ;\alpha,\beta\right)=\frac{Y_a+\alpha}{n_a+\alpha+\beta}$$

parce que l'espérance d'une loi bêta $\left(\alpha,\beta\right)$ vaut $\alpha/\left(\alpha+\beta\right)$. L'étape d/ fournit la loi marginale de Y_a, qui est calculable et se trouve être une loi dite « bêta », c'est-à-dire de densité :

$$f\left(Y_a\ ;\alpha,\beta\right)=\binom{n_a}{Y_a}\frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\cdot\Gamma\left(\beta\right)}\cdot\frac{\Gamma\left(\alpha+Y_a\right)\cdot\Gamma\left(\beta+n_a-Y_a\right)}{\Gamma\left(\alpha+\beta+n_a\right)}$$

A partir de cette densité, on peut obtenir $\hat{\alpha}$ et $\hat{\beta}$, estimateurs du maximum de vraisemblance. Malheureusement, la densité est trop compliquée pour qu'on puisse obtenir des résultats analytiques, et il faut mettre en place des algorithmes qui convergent vers les estimateurs recherchés. On aboutit au prédicteur optimum empirique :

$$\boxed{\hat{P}_{a,E}^{OPTI}=\frac{Y_a+\hat{\alpha}}{n_a+\hat{\alpha}+\hat{\beta}}}$$

Il est possible aussi d'obtenir des estimateurs de α et β selon la méthode des moments.

Rappel : soit une statistique S dont la loi dépend d'un paramètre vectoriel $\alpha=\left(\alpha_1,\alpha_2,\ldots,\alpha_p\right)$. On note $ES^n=f_n\left(\alpha\right)$ avec $\left(n=1,2,3,\ldots,p\right)$, où les f_n sont des fonctions connues. Alors la méthode des moments estime α par $\hat{\alpha}$ solution du système :

$$S^n=f_n\left(\hat{\alpha}\right)\qquad\left(n=1,2,3,\ldots,p\right)$$

Pour cela, on considère les deux statistiques suivantes :

$$\hat{p}=\sum_{a=1}^{m}\frac{n_a}{n}\cdot\hat{p}_a$$

où $\hat{p}_a$ est un estimateur direct (sans biais) de P_a, et

$$s_p^2=\sum_{a=1}^{m}\frac{n_a}{n}\left(\hat{p}_a-\hat{p}\right)^2$$

et on résout le système :

$$\frac{\hat{\alpha}}{\hat{\alpha}+\hat{\beta}}=\hat{p}$$

$$\frac{1}{\hat{\alpha}+\hat{\beta}+1}=\frac{n\,s_p^2-\hat{p}\left(1-\hat{p}\right)\left(m-1\right)}{\hat{p}\left(1-\hat{p}\right)\left[n-\sum_{a=1}^{m}\frac{n_a^2}{n}-m+1\right]}$$

On vérifie que la solution conduit à :

$$\boxed{\hat{P}_{a,E}^{OPTI} = \hat{\gamma}_a \cdot \hat{p}_a + (1-\hat{\gamma}_a)\cdot \hat{p}}$$

où $\hat{\gamma}_a = \dfrac{n_a}{n_a + \hat{\alpha} + \hat{\beta}}$. Ainsi, $\hat{P}_{a,E}^{OPTI}$ a l'allure d'un estimateur composite ($\hat{p}_a$ est direct, $\hat{p}$ peut être qualifié de synthétique puisqu'il implique l'ensemble des domaines). On a bien $\hat{\gamma}_a \in [0,1]$ et l'estimateur direct a d'autant plus d'importance que n_a est grand - ce qui est rassurant. On remarquera que la variance d'échantillonnage n'intervient jamais dans $\hat{\gamma}_a$, ce qui est un atout très appréciable (et qui n'avait pas lieu avec les modèles linéaires mixtes).

• *Exemple 2 :* P_a suit un modèle de type Logit, soit

$$Log\frac{P_a}{1-P_a} = \mu + v_a \text{, où } v_a \text{ suit une loi } \mathcal{N}(0,\sigma^2)$$

Ce modèle est plus complexe que celui de l'exemple 1, car on ne sait plus, cette fois, obtenir une expression analytique de $\hat{P}_a^{OPTI}$ (la densité a posteriori de P_a sachant Y_a est trop complexe, autrement dit on n'arrive pas à franchir l'étape b/). On peut néanmoins traiter la question par une méthode approchée s'appuyant sur des simulations.

Pour cela, on appelle Z_a la variable aléatoire de densité $\mathcal{N}(0,1)$. Puisque $v_a = \sigma.Z_a$, on relie P_a à Z_a selon $P_a = h_1(\mu + \sigma \cdot Z_a)$ avec $h_1(x) = e^x/(1+e^x)$.

Le prédicteur optimum, dans sa forme générale, s'écrit

$$\hat{P}_a^{OPTI} = E\left(P_a \middle| Y_a\ ;\ \mu,\sigma^2\right) = E\left(h_1(\mu+\sigma Z_a)\middle| Y_a\ ;\ \mu,\sigma^2\right)$$

Ce qui rend le développement difficile est le fait que l'espérance est conditionnelle. Par calcul, il se trouve que l'on parvient à progresser un peu en exprimant cette espérance conditionnelle comme un ratio de deux espérances qui ne sont pas des espérances conditionnelles. En effet, on montre que l'on peut ré-écrire $\hat{P}_a^{OPTI}$ selon

$$\hat{P}_a^{OPTI} = \frac{A(Y_a;\mu,\sigma^2)}{B(Y_a\ ;\ \mu,\sigma^2)}$$

où $A = E\left[h_1(\mu+\sigma Z)\cdot e^{h_2(Y_a,\mu+\sigma.Z)}\right] = E[\Delta_1(Z)]$

$B = E\left[e^{h_2(Y_a,\mu+\sigma\cdot Z)}\right] = E[\Delta_2(Z)]$

$h_2(Y_a,x) = x\cdot Y_a - n_a\cdot Log(1+e^x)$

Les espérances se calculent par rapport à la loi $\mathcal{N}(0,1)$ (celle de Z_a). Les calculs de A et B sont ceux d'une intégrale « classique », complexe certes mais dont on peut écrire l'expression. Ensuite, on utilise une quelconque méthode de calcul numérique d'intégrales. Ici, puisque Z suit la loi $\mathcal{N}(0,1)$ de densité $f(z)$, on peut utiliser une méthode de simulation de la manière suivante pour calculer A (idem pour B) :

- Tirer T valeurs indépendantes $Z_1, Z_2, \ldots, Z_T$ dans la loi $\mathcal{N}(0,1)$, avec T grand (T =1000 par exemple) ;
- Former $I = \frac{1}{T}\sum_{t=1}^{T} \Delta_1(Z_t)$

D'après la loi des grands nombre, I converge vers $E[\Delta_1(Z)]$, si bien qu'après avoir estimé les paramètres μ et σ, on peut in fine calculer

$$\hat{P}_a^{OPTI} = \frac{\sum_{t=1}^{T} \Delta_1(Z_t)}{\sum_{t=1}^{T} \Delta_2(Z_t)}$$

L'estimation de μ et σ^2 par maximum de vraisemblance souffre des mêmes problèmes de calcul numérique, aggravés encore par l'opération de maximisation. L'affaire reste cependant gérable, puisqu'on parvient à exprimer la densité marginale de Y_a en fonction de μ et σ :

$$Log\ f(Y_a; \mu, \sigma) = \text{constante} + \sum_{a=1}^{m} Log(B(Y_a; \mu, \sigma))$$

Il est possible aussi d'utiliser des estimateurs de type « moment », par exemple en résolvant le système non linéaire :

$$E[h_1(\hat{\mu} + \hat{\sigma}Z)] = \frac{\sum_{a=1}^{m} Y_a}{\sum_{a=1}^{m} n_a}$$

$$\text{et}\quad E[h_1^2(\hat{\mu} + \hat{\sigma}Z)] = \frac{\sum_{a=1}^{m} Y_a(Y_a - 1)}{\sum_{a=1}^{m} n_a(n_a - 1)}$$

CAS 2

C'est un contexte encore plus compliqué que le cas 1 parce que le modèle ne concerne pas directement les paramètres d'intérêt P_a, mais il introduit des paramètres $P_{a,i}$ qui varient d'un individu à l'autre. Cette variabilité ne prend vraiment un sens qui si on peut expliquer les valeurs individuelles $P_{a,i}$ par un vecteur d'informations auxiliaires $X_{a,i}$. Un façon de le faire consiste à postuler une liaison logistique telle que :

$$Log\frac{P_{a,i}}{1-P_{a,i}} = X_{a,i}^T\beta + v_a \text{ où } v_a \text{ suit une loi } \mathcal{N}(0, \sigma^2)$$

On suppose que l'échantillonnage qui conduit à observer les valeurs $Y_{a,i}$ des individus d'un échantillon s_a est « ignorable », en ce sens où le modèle n'est pas affecté par le processus de sélection (autrement dit, l'inférence sur les paramètres n'a pas à prendre en compte l'échantillonnage). Cela est vrai par exemple avec un sondage aléatoire simple, et, de façon plus générale, avec tout processus de sélection dont les probabilités de sélection font

intervenir des variables incluses dans le vecteur $X_{a,i}$ (voir aussi 2.2.1). Cela étant, le paramètre d'intérêt reste la proportion P_a définie au niveau du domaine. On a

$$P_a = \frac{1}{N_a}\sum_{i=1}^{N_a} Y_{a,i} = \frac{n_a}{N_a}\bar{y}_a + \frac{N_a - n_a}{N_a}\bar{y}_a^C$$

où $\bar{y}_a = \frac{1}{n_a}\sum_{i\in s_a} Y_{a,i}$ représente la proportion d'individus de l'échantillon s_a qui sont dans la sous population D, et $\bar{y}_a^C$ est la proportion d'individus du domaine qui n'appartiennent pas à s_a et qui sont dans D. On notera que $\bar{y}_a$ n'est pas, en général, l'estimateur direct de P_a que recommande la théorie des sondages : il ne l'est que si le tirage est à probabilités égales. Comme $\bar{y}_a$ est calculable, il reste à prédire $\bar{y}_a^C$ par le prédicteur optimum :

$$\hat{\bar{y}}_a^{C,OPTI} = E\left[\bar{y}_a^C \middle| Y_{a,j} \ pour \ j \in s_a \ ; \beta, \sigma^2\right]$$

Or
$$\bar{y}_a^C = \frac{1}{N_a - n_a}\sum_{\substack{i\notin s_a \\ i\in a}} Y_{a,i}$$

Pour un individu (a,i) de a, non échantillonné :

$$E\left[Y_{a,i} \middle| Y_{a,j} \ pour \ j \in s_a \ ; \beta, \sigma^2\right] = \underset{(1)}{E}\, E\left[Y_{a,i} \middle| P_{a,i}, Y_{a,j} \ pour \ j \in s_a \ ; \beta, \sigma^2\right]$$

d'après les règles de l'espérance conditionnelle, où $\underset{(1)}{E}$ désigne l'espérance par rapport à la loi de $P_{a,i}$ conditionnelle aux $Y_{a,j}$ observés. Comme $Y_{a,j}$ suit la loi binomiale $\mathcal{B}(1, P_{a,i})$, la seconde espérance vaut $P_{a,i}$, si bien qu'in fine on aboutit à

$$E\left[P_{a,i} \middle| Y_{a,j} \ pour \ j \in s_a \ ; \beta, \sigma^2\right]$$

puis au prédicteur optimum :

$$\boxed{\hat{P}_a^{OPTI} = \frac{n_a}{N_a}\bar{y}_a + \frac{1}{N_a}\cdot\sum_{\substack{i\notin s_a \\ i\in a}} E\left[P_{a,i} \middle| Y_{a,j} \ pour \ j \in s_a \ ; \beta, \sigma^2\right]}$$

On ne peut pas développer davantage, le calcul de l'espérance conditionnelle ne semblant pas pouvoir déboucher sur des formules analytiques : il faut se résoudre à agir comme dans l'exemple 2 du cas 1, c'est-à-dire en mettant l'espérance conditionnelle sous la forme d'un ratio de deux espérances non conditionnelles, chacune étant en fait une intégrale complexe que l'on calcule par une méthode numérique appropriée. Bien entendu, on aura auparavant estimé les deux paramètres β et σ^2, soit à partir de la densité des $Y_{a,i}$ (approche EMV), soit à partir de la méthode des moments.

4.5. Cas des variables qualitatives : paramètre de type « risque relatif ».

Il arrive que le paramètre d'intérêt soit de type « risque relatif » c'est-à-dire qu'il soit défini comme le rapport de la proportion d'une sous-population donnée au sein du domaine à la proportion de cette même sous-population dans la population entière. Ce type de paramètre est beaucoup utilisé pour caractériser en épidémiologie les zones fortement touchées (relativement) par une maladie donnée, où ayant de forts taux de mortalité associés à une cause donnée, soit :

$$\theta_a = \frac{P_a}{P} = \frac{N_a.P_a}{N_a.P} = \frac{Y_a}{\tilde{Y}_a}$$

P_a = proportion d'une sous-population donnée au sein du domaine ;

P = proportion d'une sous-population donnée au sein de la population entière.

(la sous-population est caractérisée par exemple par les personnes atteintes de la grippe, ou encore par celles qui sont au chômage). Dans l'échantillon de taille n_a, on observe l'effectif $\hat{Y}_a$ égal au nombre d'individus de la sous-population intérêt dans l'échantillon recoupant a. On note :

$$\tau_a = n_a . \frac{\sum_{a=1}^{m} \hat{Y}_a}{\sum_{a=1}^{m} n_a}$$

où τ_a (proche de $n_a.P$) est considéré comme fixé parce que le ratio est construit sur un grand nombre d'individus (il est donc de faible variance par rapport à $\hat{Y}_a$).

Il est assez naturel de considérer que $\hat{Y}_a$ suit une loi binomiale de paramètres n_a et P_a, loi que l'on peut « simplifier » en loi de Poisson, de paramètre $n_a.P_a = \tau_a.\theta_a$. La conjonction de la modélisation et de la simplification débouche sur l'hypothèse suivante :

$$\hat{Y}_a | \theta_a \rightarrow Poisson(\tau_a.\theta_a)$$

Il reste à faire une hypothèse sur la loi du paramètre intérêt Là encore, on peut imaginer plusieurs cas, dont

$\theta_a \rightarrow Gamma(\alpha, \beta)$ (modèle dit « Poisson-Gamma - cas 1) , soit

$$f(\theta_a ; \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} e^{-\beta\theta_a} \theta_a^{\alpha-1}$$

ou $Log\theta_a \rightarrow N(\mu, \sigma^2)$ (modèle dit « Log-normal » - cas 2).

- Avec le cas 1, on montre $\theta_a | \hat{Y}_a \rightarrow Gamma(\alpha + \hat{Y}_a, \beta + \tau_a)$. D'où

$$\hat{\theta}_{a,E}^{OPTI} = \frac{\hat{Y}_a + \hat{\alpha}}{\tau_a + \hat{\beta}}$$

après estimation de α de β par $\hat{\alpha}$ et $\hat{\beta}$ (la loi de $\hat{Y}_a$ est binomiale négative). Si on utilise des estimateurs des moments « bien choisis » et en posant $\hat{\theta}_a = \hat{Y}_a / \tau_a$ estimateur direct de θ_a, on forme :

$$\hat{\theta}_{a,E}^{OPTI} = \hat{\gamma}_a \cdot \hat{\theta}_a + (1-\hat{\gamma}_a)\cdot \hat{\theta}$$

Direct Synthétique

Où $$\hat{\theta} = \frac{\sum_{a=1}^{m} \tau_a . \hat{\theta}_a}{\sum_{a=1}^{m} \tau_a} \quad \text{et} \quad \hat{\gamma}_a = \frac{\tau_a}{\tau_a + \hat{\alpha}} \in [0,1] \quad .$$

- Avec le cas 2, il n'y a pas de formulation analytique de $\hat{\theta}_{a,E}^{OPTI}$. Il faut utiliser une autre technique.

Si on connaît la taille du domaine N_a, l'estimation de l'effectif de la sous population concernée peut être obtenue en calculant $N_a . \hat{P}_a$.

5. La classe des prédicteurs Bayésiens hiérarchiques

Cette méthode est seulement citée pour mémoire. Il s'agit d'une approche qui peut se résumer à celle du chapitre 4 dans laquelle on ajoute une étape : on suppose en effet que les paramètres (λ_1, λ_2) suivent une loi que l'on se fixe A PRIORI, soit $f(\lambda_1, \lambda_2)$. Par la formule de Bayes :

$$f(\mu, \lambda_1, \lambda_2 | Y) = \frac{f(Y, \mu \mid \lambda_1, \lambda_2)\cdot f(\lambda_1, \lambda_2)}{\int f(Y, \mu | \lambda_1, \lambda_2)\cdot f(\lambda_1, \lambda_2) d\mu d\lambda}$$

La densité $f(Y, \mu \mid \lambda_1, \lambda_2)$ est calculée comme au chapitre précédent. Le problème essentiel se rencontre au dénominateur, qui représente la densité de Y : en effet, généralement cette expression est incalculable <u>analytiquement</u> ! On peut s'en sortir en utilisant des techniques de simulation par chaînes de Markov (algorithmes de Gibbs, algorithme de Metropolis). Finalement, on obtient la densité « A POSTERIORI »

$$f(\mu | Y) = \int f(\mu, \lambda_1, \lambda_2 | Y) d\lambda$$

puis on termine en exprimant

$$E(\mu | Y) = \hat{\mu}^{HB}$$

dit estimateur « Bayésien hiérarchique ».

Nous ne développons pas ici ces techniques, qui sont très complexes, et renvoyons à l'ouvrage de J.N.K. Rao (2003) pour plus d'information.

6. L'approche par la prédiction

C'est une façon de pratiquer l'estimation sur petits domaines qui s'appuie sur des modèles de comportement individuel de nature explicite. On se situe donc dans l'esprit de la partie 2.2, mais

en la généralisant (en particulier on manipule des expressions qui ne font plus apparaître d'effet variable associé au domaine). On va poser, pour tout individu de la population :

$$Y_i = X_i^T \cdot B + e_i$$

avec e_i un aléa d'espérance nulle, X_i un vecteur d'informations auxiliaires partout connues et B un vecteur de paramètres inconnus. Le modèle est le même sur le domaine et sur le reste de la population. De même, le modèle s'écrit à l'identique que l'individu soit ou non dans l'échantillon s (sur ce point, voir mise en garde du 2.2.1). On peut mener une théorie complète avec des hypothèses quelconques sur la variance des e_i, mais pour simplifier on considérera que les e_i sont deux à deux indépendants.

Pour prédire le vrai total (aléatoire) Y_a sur le domaine a, on va chercher la fonction linéaire - notée $\tilde{Y}_a$ - des Y_i pour i décrivant s (et non pas seulement s_a) qui soit la plus proche de Y_a au sens suivant :

- Absence de biais, soit $E(\tilde{Y}_a - Y_a) = 0$
- Erreur quadratique minimale, soit $E(\tilde{Y}_a - Y_a)^2$ minimale.

Dans les conditions d'absence de corrélation précédemment posées, la solution à ce problème est :

$$\tilde{Y}_a = \sum_{i \in s_a} Y_i + \sum_{\substack{i \in a \\ i \notin s_a}} \hat{Y}_i$$

où $\hat{Y}_i = \hat{B}.X_i$ est le meilleur prédicteur individuel de Y_i. L'expression de $\hat{B}$ dépend des hypothèses faites sur les variances des e_i. Mais il est à noter que $\hat{B}$ ne tient pas compte des poids de sondage (l'optimalité est conçue par rapport à la loi du modèle seulement : peu importe la façon dont les individus sont échantillonnés). Donc, avec les notations habituelles, et quel que soit le plan de sondage (la présence de moyennes simples $\bar{y}_a$ et $\bar{x}_a$ n'est pas une erreur !) :

$$\boxed{\tilde{Y}_a = n_a.\bar{y}_a + \left(X_a - n_a\bar{x}_a\right)^T \hat{B}}$$

On peut adapter cette expression au cas de modèles traduisant des situations très courantes, par exemple :

- Modèle $Y_i = \mu + e_i$, avec $\mathrm{Var}(e_i) = \sigma^2$.

Il a été introduit au 2.1 du chapitre II, en tant que comportement de base permettant de justifier l'estimateur synthétique construit à partir de la moyenne calculée sur l'ensemble de l'échantillon, soit $\hat{Y}/\hat{N}$. Dans ce cas, on a $\hat{B} = \hat{\mu} = \bar{y}$, moyenne simple sur l'ensemble de l'échantillon s. D'où :

$$\boxed{\tilde{\tilde{Y}}_a = n_a.\bar{y}_a + \left(N_a - n_a\right).\bar{y} = N_a.\bar{y} + n_a\left(\bar{y}_a - \bar{y}\right)}$$

Cette expression est à rapprocher de l'estimateur synthétique classique $N_a.\bar{y}$.

- Modèle $Y_i = \mu_h + e_i$ avec $\mathrm{Var}(e_i) = \sigma_h^2$ si i est dans h (catégorie de la population)

Il a été introduit au 2.2.2 . Il conduit à :

$$\boxed{\tilde{\tilde{Y}}_a = n_a.\bar{y}_a + \sum_h \left(N_{ah} - n_{ah}\right).\bar{y}_h}$$

à comparer avec l'estimateur synthétique classique $\sum_{h=1}^{H} N_{ah} \cdot \bar{y}_h$.

- Modèle $Y_i = BX_i + e_i$ avec $\mathrm{Var}(e_i) = \sigma^2 . X_i$

On obtient

$$\tilde{\bar{Y}}_a = n_a . \bar{y}_a + \left(X_a - n_a \bar{x}_a\right) . \frac{\bar{y}}{\bar{x}} = X_a \frac{\bar{y}}{\bar{x}} + n_a \left(\bar{y}_a - \bar{x}_a \frac{\bar{y}}{\bar{x}} \right)$$

à comparer avec l'estimateur synthétique $X_a \frac{\bar{y}}{\bar{x}}$.

A partir de ces estimateurs de prédiction optimum, on peut faire des calculs de variance assez simplement, et estimer sans biais ces variances.

7. Eléments sur la qualité des estimations

Une des questions essentielles demeure l'évaluation de la qualité des estimations « petits domaines ». S'agissant de comparer une estimation (quelle que soit la méthode qui l'a produite) à une vraie valeur que l'on ne connaît pas (par définition), on ne peut s'en tenir de toute façon qu'à des présomptions. On se trouve donc dans une situation banale en statistique, ni pire ni meilleure que celle que rencontrent par exemple tous les prévisionnistes. En particulier, on ne peut pas tenir compte de spécificités qui ne seraient pas prises en compte d'une façon ou d'une autre au travers d'un modèle. Il ne semble donc pas y avoir de recette infaillible pour qualifier la pertinence de l'estimation, mais néanmoins on dispose d'éléments d'appréciation, en particulier les suivants.

a) L'estimation de l'erreur quadratique moyenne (EQM) de l'estimateur « petits domaines » : ce document a assez largement abordé cette question, et on a vu qu'avec la modélisation implicite il s'agissait plutôt, hélas, d'un pis-aller (cf 2.3 du chapitre III). En revanche, la modélisation explicite fournit des outils plus convaincants si le modèle est juste (chapitre IV).

b) Les diagnostics de qualité des modèles en cas de modélisation explicite. On trouve des indicateurs qui permettent de juger de la pertinence du modèle (choix des variables explicatives par exemple - y compris au niveau d'éventuels effets aléatoires) et des grandeurs qui mesurent le degré « d'ajustement » du modèle aux données collectées (comme le R^2 en régression multivariée classique). Les graphiques de résidus estimés sont fort utiles, à la fois pour valider l'hypothèse d'espérance nulle de ces résidus et pour apprécier la validité de l'hypothèse formulée sur leur variance. Des techniques permettent de détecter les individus ayant une influence particulièrement forte dans la détermination des paramètres du modèle, ce qui permet ensuite de les traiter de manière spécifique.

c) Lorsqu'on dispose d'échantillons dont les tailles sont jugées suffisamment grandes sur quelques domaines (par exemple s'il y a eu une extension d'enquête dans quelques régions, ou dans quelques départements), il est évidemment naturel de juger de la pertinence des estimations petits domaines en les comparant aux estimateurs directs. C'est d'ailleurs la méthode la plus naturelle - et probablement la plus fiable - pour apprécier la qualité de l'estimation « petits domaines ».

d) Même s'ils ont une grande variance, les estimateurs directs sont a priori sans biais (ou faiblement biaisés s'il y a eu un redressement). Donc, si on représente dans un plan les différents domaines avec en abscisse l'estimation « petits domaines » et en ordonnée l'estimation directe, s'il y a un nombre suffisant de domaines, on devrait obtenir un nuage de points repartis autour de la droite d'équation $y = x$ si les estimateurs « petits domaines » ont un faible biais. Si le nuage a une forme spécifique qui s'éloigne

manifestement de ce schéma, on peut soupçonner un problème de biais provenant des estimateurs « petits domaines ». Une technique plus objective que la simple appréciation visuelle consiste à tester l'égalité à 1 de la pente de la droite de régression et l'égalité à 0 de l'ordonnée à l'origine.

e) Il est intéressant de sommer les estimations « petits domaines » pour obtenir des estimations sur des populations de grande taille et apprécier l'écart à l'estimateur direct (réputé fiable sur une population de grande taille). Par exemple, on sommera les estimations départementales pour voir ce que donne l'estimation nationale ainsi reconstituée. On peut d'ailleurs imaginer plusieurs types de sommation. Si l'écart est trop grand, les estimateurs « petits domaines » apparaîtront douteux.

f) Une méthodologie sécurisante consiste de toute façon à mettre en œuvre plusieurs méthodes d'estimation « petits domaines » et à comparer les distributions des estimations relatives aux différents petits domaines. Il est conseillé de produire ne serait-ce que des statistiques descriptives de type « box-plot » et de discuter des résultats avec un expert du thème de l'enquête. On peut aussi voir apparaître graphiquement des structures intéressantes sur des cartes lorsque les petits domaines ont une nature géographique.

Conclusion

On a bien vu que les techniques d'estimation sur petits domaines - au-delà des estimations directes classiques, qui s'avèrent très souvent insuffisantes - sont tout à fait dépendantes de modèles de comportement. Il est donc nécessaire de faire des hypothèses « plus ou moins » proches de la réalité, et dont la pertinence est, dans le meilleur des cas, difficile à apprécier. En particulier, on ne saurait trop insister sur l'importance de la phase préliminaire de sélection de l'information auxiliaire pertinente, qui constitue l'étape clef recelant les gisements d'amélioration de qualité. Si, lors de cette étape, les sources d'information externe apparaissent insuffisamment riches ou inexploitables, voire inaccessibles, on risque bien de se trouver en situation très difficile et de ne pas pouvoir répondre à la demande. Une autre cause d'échec peut être la sous-estimation des moyens humains nécessaires pour mener l'étude, laquelle nécessite de l'expertise technique à la fois en matière de modélisation du phénomène étudié et en terme de technique statistique.

Il y a un aspect un peu troublant dans toute cette problématique, pour la raison suivante. Soit on utilise un estimateur purement synthétique, et c'est une approche brutale qui est de l'ordre de l'acte de foi, soit on utilise un estimateur de type composite qui possède a priori un aspect plus rassurant parce qu'il est en partie bâti sur un « vrai » estimateur local. Mais dans ce dernier cas, on peut penser qu'il n'y a véritablement d'intérêt à utiliser cet estimateur composite que dans le cas où deux conditions simultanées sont vérifiées : d'une part disposer d'un modèle explicatif assez performant, et d'autre part avoir à faire à un domaine ayant tout de même « une certaine taille ». En effet, si le modèle s'ajuste mal, l'estimateur composite sera essentiellement influencé par l'estimateur direct - auquel généralement on ne croit pas - et on risque bien d'obtenir des estimations curieuses, voire fantaisistes. Si la taille de l'échantillon dans le domaine n'atteint pas une taille suffisante, c'est au contraire l'estimateur synthétique qui sera déterminant, mais cela revient à s'appuyer sur le « tout modèle » et on aura peut-être des difficultés à faire croire que les spécificités locales sont correctement prises en compte...C'est la raison pour laquelle il nous semble que la situation la plus favorable dans la mise en œuvre des techniques d'estimation sur petits domaines reste celle qui assure un équilibre entre une contribution « vraiment locale » et une contribution de type modèle, ce qui revient à dire, encore une fois, qu'il faut se résoudre à recourir à des pratiques facilement critiquables dans le cas de tout petits domaines ou lorsqu'on ne dispose pas d'une information auxiliaire au pouvoir explicatif satisfaisant.

La théorie des estimations sur petits domaines est récente, assez compliquée (au moins pour la partie modélisation explicite) et mal diffusée, donc peu mise en œuvre. Cela explique probablement que les outils logiciels adaptés à la problématique soient peu développés à ce jour. En France, on a d'ailleurs très peu d'expérience en la matière. Cela étant, si la conception des poids nécessite une maîtrise technique de la discipline, il n'y a pas de difficulté particulière pour le chargé d'étude non-spécialiste des techniques de sondage à effectuer concrètement les estimations (dès lors qu'il n'a pas l'ambition de faire du calcul de précision…) : en effet, tous les estimateurs manipulés sont des fonctions linéaires des valeurs individuelles collectées et le statisticien peut (doit) fournir au chargé d'études le fichier national de collecte prêt à l'emploi, avec les poids qu'il aura calculés pour les besoins de l'estimation locale. Evidemment, pour un domaine donné, il faut en toute généralité un poids par variable - mais on peut souvent faire des économies d'échelle, en particulier avec une modélisation implicite (avec laquelle les catégories distinguées par le modèle peuvent être les mêmes pour des « paquets » de variables d'intérêt). Néanmoins, globalement, l'opération reste longue et fort coûteuse en moyens humains : sauf exception, le processus ne se prête pas à l'industrialisation et il n'est donc raisonnablement envisageable que pour des enquêtes qui n'ont « pas trop » de variables

Il faut également signaler que la mise en œuvre de ces approches nécessite de gros efforts pédagogiques : il est assez difficile d'expliquer à un financeur local « non-statisticien » qu'il doit mettre des moyens pour que l'on traite une information collectée ailleurs que dans la zone qui l'intéresse mais qui prétend par ailleurs être utile pour estimer les comportements purement locaux. Concrètement, l'utilisateur risque bien d'être perturbé par la nécessité d'exploiter un fichier national pondéré alors que sa logique lui fait attendre un fichier local pondéré..

Enfin, terminons par une évidence qu'il convient de garder en mémoire : toutes ces méthodes ne sont jamais que des palliatifs, qui ne peuvent pas rivaliser en qualité avec une conception du plan de sondage qui prendrait en compte dès l'échantillonnage les objectifs d'estimation sur de petits domaines. Cela veut dire que si on connaît les besoins d'estimation sur des petites populations avant de tirer l'échantillon, il faut réfléchir à des méthodes plus performantes, en particulier à des tirages à probabilités inégales sur-représentant la petite population ou à des compléments de sondage effectués après coup spécifiquement dans la population rare. Cela ne pose pas trop de problème techniques (on sait maintenant traiter facilement les tirages dans des bases de sondage multiples), mais évidemment cela oblige à concevoir un plan de financement adapté à cette ambition.

Bibliographie

On ne cite ici que des ouvrages et articles spécialisés sur le thème des petits domaines.

I) Ouvrages ou assimilés :

[1] J.N.K. Rao, « *Small Area Estimation* », Wiley, 2003 : ouvrage de référence

[2] F. Arnaud, « *Estimation dans les domaines* », Mémoire de DEA sous la direction de JC Deville, 2001

[3] M. Ghosh, C. Sarndal, "*Lecture notes on Estimation for Population Domains and Small Areas* », Reviews Statistics Finland, 2001 /5

[4] N.T. Longford, *"Missing Data and Small Area Estimation"*, Springer, 2005

[5] R. Platek, J.N.K. Rao, C Särndal, M.P. Singh, "*Small Area Statistics*", Wiley, 1987

[6] M.P. Singh, J. Gambino, H.J. Mantel, "*International Conference on Small Area Statistics - Warsaw*", 1992
"Small Area Estimation", IASE Satellite Conference, Riga, Conference Proceedings, 1999

II) Articles :

[7] K. Attal-Toubert, O. Sautory, « *Estimation de données régionales à l'aide de techniques d'analyse multidimensionnelle »*, Document de travail UMS, N° 9807

[8] M.D. Bankier, « *Power allocations: Determining Sample Sizes for Subnational Areas* », The American Statistician, Vol 42, N°3, 1988

[9] C. Couet , P. Mormiche, « *La réalisation d'estimations locales dans le cadre de l'enquête HID* », Actes JMS 2002

[10] G.S. Datta, B. Day, I. Basawa, "*Empirical Best Linear Unbiased and Empirical Bayes Prediction in Multivariate Small Area Estimation*", Journal of Statistical Planning and Inference, 1999, N° 75

[11] G. Decaudin, J-C Labat, " *Une méthode synthétique, robuste et efficace, pour réaliser des restimations locales de population* », Actes JMS 1996

[12] L. Descours, « *Estimation de populations locales par la méthiode de la taxe d'habitation* », Actes JMS 1991

[13] S. Destandau, "*Estimation sur des petits domaines. Application à l'enquête Education 92* », Actes JMS 1996

[14] J.D. Drew, M.P. Singh, G.H. Choudhry, "*Evaluation of Small Area Techniques for the Canadian Labour Force Survey*", Survey Methodology, 1982, N° 8

[15] B. Efron, C. Morris, « Stein's Estimation Rule and its Competitors - An Empirical Bayes Approach », JASA, 1973, N°68

[16] B. Efron, C. Morris, « *Data Analysis using Stein's Estimate and its Generalizations* », JASA, 1975, N°70

[17] P.J. Farrell, B. MacGibbon, T.J. Tomberlin, "*Empirical Bayes Small Area Estimation using Logistic Regression Models and Summary Statistics*", Journal of Business and Economic Statistics, 1997, N° 15

[18] P.D. Falorsi, S. Falorsi, A. Russo, « *Empirical Comparaison of Small Area Estimation Methods for the Italian Labour Force Survey*", Survey Methodology, 1994, N° 20

[19] R.E. Fay, R. Herriot, "*Estimates of Income for Small Places: An application of James Stein Procedures to Census Data*", JASA, 1979, N°74

[20] M. Ghosh, K. Natarajan, T.W.F. Stroud, B.P. Carlin, "*Generalized Linear Models for Small Area Estimation*", JASA, 1998, N° 93

[21] M. Ghosh, J.N.K. Rao, « *Small Area Estimation : an Appraisial* », Statistical Science, 1994, Vol 9, N°1

[22] M.E. Gonzalez, C. Hoza, "*Small Area Estimation with Application to Unemployement and Housing Estimates*", JASA, 1978, N° 73

[23] M. Hidiroglou, « *Estimation pour les petits domaines : théorie et pratique à Statistique Canada* », Actes JMS 1991

[24] M. Hidiroglou, Z. Patak, "*Estimation par domaine par la régression linéaire*", Techniques d'enquête, Vol 30, N°1, 2004.

[25] D. Holt, D.J. Holmes, « *Estimation sur petits domaines dans des plans de sondage avec probabilités inégales* », Techniques d'enquête, 1994, N° 20

[26] D. Holt, T.M.F. Smith, T.J. Tomberlin, « *A Model-Based Approach to Estimation for Small Subgroups of Population*", JASA, 1979, N° 74

[27] F.L. Hulting, D.A. Harville, "*Some Bayesian and Non-Bayesien Procedures for the Analysis of Comparative Experiments and for Small Area Estimation: Computational Aspects, Frequentist Properties and Relationships*", JASA, 1991, N° 86

[28] F. Jeger, "*Méthode d'utilisation d'enquête à un niveau géographique où l'échantillon est faible* », Actes JMS 1991

[29] J. Jiang, P. Lahiri, "*Empirical Best Prediction for Small Area Inference with Binary Data", Annals of the Institute of Mathematical Statistics*", 2001, N° 53

[30] P.S. Kott, "*Robust Small Domain Estimation Using Random Effects Modelling*", Survey Methodology, 1990, N° 15

[31] P. Lahiri, J.N.K. Rao, "*Robust Estimation of Mean Squared Error of Small Area Estimators*", JASA, 1995, N° 82

[32] K.J. Lui, W.G. Cumberland, "*A Model Based Approach: Composite Estimators for Small Area Estimation*", JOS, 1991, N° 7

[33] D. Malec, J. Sedransk, C.L. Moriarity, F.B. Leclerc, "*Small Area Inference for Binay Variables in National Health Interview Survey*" JASA, 1997, N° 92

[34] D.A. Marker, "*Organization of Small Area Estimators Using a Generalized Linear Regression Framework*", JOS, 1999, N° 15

[35] D.A. Marker, "*Producing Small Area Estimates From National Surveys: Methods for Minimizing Use of Indirect Estimators*", Survey Methodology, 2001, N° 27

[36] F.A.S. Moura, D. Holt, « *Small Area Estimation Using Multilevel models* », Survey Methodology, 1999, N° 25

[37] J.L. Pan Ke Shon, H. Vivier, " *Estimation de l'isolement relationnel dans trois ZUS de Bretagne*", Actes JMS 2002

[38] D. Pfeffermann, " *Small Area Estimation - New Developments and Directions*", International Statistical review, 2002, N° 70

[39] D. Pfeffermann, L. Burck, "*Robust Small Area Estimation Combining Time Series and Cross-Sectionnal Data*", Survey Methodology, 1990, N° 27

[40] N.G.N. Prasad, J.N.K. Rao, " *The Estimation of the Mean Squared Error of Small-Area Estimators*", JASA, 1990, N° 85

[41] N.G.N. Prasad, J.N.K. Rao, "*On Robust Small Area Estimation Using a Simple Random Effect Model*"; Survey Methodology, 1999, N° 25

[42] N.J. Purcell, L. Kish, "*Estimates for Small Domains*", Biometrics, 1979, N° 35

[43] N.J. Purcell, L. Kish, "*Postcensal Estimates for Local Areas*", International Statistical Review, 1980, N° 48

[44] J.N.K Rao, G.H. Choudhry "*Small Area Estimation: Overview and Empirical Study*", ICES Proceedings, 1993

[45] J.N.K Rao, M. Yu, "*Small Area Estimation by Combining Time Series and Cross-Sectionnal Data*", Canadian Journal of Statistics, 1994, N° 22

[46] L.P. Rivest, E. Belmonte, "*A Conditional Mean square Error ofSmall Area Estimators*", Survey Methodology, 2000, N° 26

[47] C. Sarndal, "*Design-Consistent Versus Model-Dependent Estimation for Small Domain*", JASA, 1984, vol 79, N° 387

[48] C.E. Sarndal, M.A. Hidiroglou, « *Small Domain Estimation: A Conditional Analysis* ", JASA, 1989, N° 84

[49] C.E. Sarndal, M.A. Hidiroglou, « *An Empirical Study of Some Regression Estimators for Small Domains* ", Survey Methodology, 1985, N° 11

[50] W.A. Schaible, "*Choosing Weights for Composite Estimators for Small Area Statistics*", Proceedings of the Section on Survey Research Methods, American Statistical Association, 1978

[51] A.C. Singh, I.U.H Mian, "*Generalized Sample Size Dependent Estimators for Small Areas*", ARC95 Proceedings

[52] M.P. Singh, J. Gambino, H.J. Mantel, "*Les petites régions : problèmes et solutions*", Techniques d'enquête, 1994, Vol 20, N° 1

[53] E. Stasny, P.K. Goel, D.J. Rumsey, « *Estimation de la production en blé par comté* », Techniques d'enquête, 1991, N° 17

[54] D.M. Stukel, J.N.K. Rao, « *Small-Area Estimation Under Two-fold Nested Errors Regression Models*", Journal of Statistical Planning and Inference, 1999, N° 78

[55] K.M. Wolter, B.D. Causey, "*Evaluation of Procedures for Improving Population: Estimates for Small Area*", JASA, 1991, N° 86

[56] L. Zhang, R.L. Chambers, « *Small Area Estimates for Cross-Classifications*", JRSS B, 2004, Part 2