Correction de la non-réponse totale : par imputation ou par repondération ?

Gwennaëlle BRILHAULT[1] et Nathalie CARON[2]

E2004/01

Insee, Direction des statistiques d'entreprises, Division Harmonisation d'enquêtes auprès des entreprises, Timbre E210
18 Bld Adolphe Pinard - 75675 Paris cedex 14

Résumé :

La correction de la non-réponse totale se fait soit par imputation soit par repondération. Aucun résultat théorique ne permet de donner l'avantage à l'une ou l'autre de ces méthodes de manière générale. Dans ce contexte, il semble intéressant de comparer ces deux méthodes de correction de la non-réponse sur la base de simulations d'enquête, comme le fait R.J.A. Little dans son article de 1986 intitulé "Survey Nonresponse Adjustments for Estimates of Means" : il considère là une enquête fictive dont il fait varier les paramètres pour comparer ces deux méthodes dans des configurations différentes. Cette méthodologie de comparaison est ici adaptée au cadre réel des Enquêtes Annuelles d'Entreprise (EAE), caractérisé notamment par des échantillons de petite taille dans les croisements les plus fins de la stratification. On montre alors que pour les estimations sur des domaines d'intérêt (les régions dans le cas des EAE) la correction de la non-réponse joue un rôle positif, surtout dans le cas de la correction par imputation et de la correction par post-stratification. Les résultats

[1] Ce travail a été réalisé dans le cadre du stage de 2ème année de l'ENSAE de G. Brilhault.
[2] Au moment de la réalisation de cette étude, N. Caron appartenait à la division Harmonisation d'Enquêtes auprès des Entreprises.

numériques obtenus ne sont pas directement généralisables car ils sont indissociables du contexte spécifique retenu et des hypothèses sur lesquels ils reposent.

Mots clés : correction de la non-réponse totale, imputation, repondération, EAE

Les auteurs remercient Christian Hesse, méthodologue rattaché au département "Système Statistique d'Entreprises" de la Direction des Statistiques d'Entreprises de l'INSEE, pour son aide précieuse tout au long de la réalisation de cette étude.

Sommaire

Introduction

La non-réponse constitue un écueil inévitable pour tout responsable d'enquête, qu'il s'agisse de la non-réponse partielle (questionnaire partiellement rempli) ou de la non-réponse totale (absence totale de réponse dans un questionnaire). Deux méthodes sont en général utilisées pour corriger le biais qu'induit la non-réponse dans les estimations sur données incomplètes : il s'agit de la correction par repondération et de la correction par imputation. Bien que la frontière entre non-réponse partielle et non-réponse totale ne soit pas toujours facile à tracer, on peut souligner qu'une fois définie comme telle, la non-réponse partielle ne se prête qu'à la correction par imputation. La non-réponse totale au contraire peut être traitée avec l'une ou l'autre des deux méthodes citées : c'est elle qui nous intéresse dans l'étude présentée ici.

En effet, si l'usage de ces deux méthodes est possible pour corriger la non-réponse totale (une fois constituées des cellules d'ajustement de la non-réponse), il n'existe pas de considération théorique donnant l'avantage à l'une ou l'autre de ces méthodes. On sait simplement qu'elles sont parfois équivalentes quand on réalise des estimations globales, alors qu'elles donnent des résultats différents pour des estimations sur des domaines d'intérêt qui ne coïncident pas avec les cellules d'ajustement de la non-réponse.

En l'absence de résultat théorique sur les avantages relatifs de la repondération et de l'imputation, une solution est de recourir à des simulations d'enquêtes pour les comparer. C'est précisément ce que fait R.J.A. Little dans un article de 1986 intitulé "Survey Nonresponse Adjustments for Estimates of Means" : il y compare des estimateurs corrigés de la non-réponse par repondération et par imputation en se basant sur des simulations de réponses à une enquête fictive. Dans ce cadre artificiel, il est en mesure de faire varier plusieurs caractéristiques de cette enquête fictive (les taux de réponse, la taille de l'échantillon, etc.) pour comparer dans différents cas les avantages relatifs de ces estimateurs.

Cette démarche de comparaison des méthodes de correction de la non-réponse nous a semblé susceptible d'être appliquée à un schéma plus proche de la réalité, afin de faire évoluer la réflexion sur de véritables enquêtes, de manière plus ou moins directe. En particulier, les résultats d'une telle comparaison peuvent être utiles dans le contexte des réflexions sur le futur système des Enquêtes Annuelles d'Entreprise (EAE) que la division "Harmonisation d'Enquêtes auprès des Entreprises" s'apprête à lancer, après avoir dressé en 2003 un bilan de l'actuelle quatrième génération de ces EAE. C'est donc dans cette perspective que s'inscrit l'idée de reprendre la méthodologie développée par Little pour l'appliquer à un cadre proche de celui des EAE, afin de comparer dans ce cadre spécifique la correction de la non-réponse par repondération à celle par imputation. Si Little présente son travail sur données fictives, il s'agit donc ici de le transposer à des données réelles.

Après avoir introduit quelques éléments théoriques sur la correction de la non-réponse et la problématique de notre étude, nous présentons brièvement le travail de Little. Puis nous exposons les choix que nous avons faits pour adapter sa méthodologie au cas particulier des EAE, avant de donner les résultats que nous avons obtenus. Les calculs et résultats détaillés, ainsi que les programmes informatiques réalisés, sont présentés en annexe. Les notations utilisées sont regroupées dans un glossaire en fin de document.

I. Correction de la non-réponse totale par repondération ou par imputation : rappels théoriques et méthodologie de comparaison proposée par Little

Le problème de la non-réponse totale est un sujet incontournable pour un responsable d'enquête. Quels que soient les dispositifs mis en place dans une enquête pour prévenir et limiter au minimum la non-réponse (prise en compte des taux de réponse antérieurement constatés dans la conception du plan de sondage, relance des individus à interroger, etc.), elle ne peut jamais être totalement évitée, ne serait-ce que parce qu'on ne réussit pas forcément à contacter toutes les entités (ménages ou entreprises) sélectionnées dans un échantillon. Le type de collecte de l'enquête (par voie postale, par entretiens individuels, etc.) joue aussi sur le taux de non-réponse enregistré. Dans le cas des enquêtes d'entreprises, le problème de la non-réponse est accru par l'hétérogénéité de la population des entreprises dans laquelle petites et grandes entreprises se côtoient, là où les ménages sont plus interchangeables. Les plus grandes entreprises sont d'ailleurs parfois contactées pour des entretiens en face à face par une équipe d'enquêteurs spécialisés, ce qui n'est pas le cas des petites entreprises.

Or, l'existence de données manquantes détériore la qualité des estimations issues des données collectées. En effet, l'utilisation des estimateurs habituels fondés sur les seuls répondants fournit en général des estimations biaisées. C'est le cas lorsque les non-répondants n'ont pas les mêmes caractéristiques que les répondants, c'est-à-dire la plupart du temps (les caractéristiques des non-répondants expliquent parfois leur absence de réponse à l'enquête : à titre d'exemple, il est connu que les ménages aisés donnent moins volontiers que d'autres des informations sur leur revenu annuel dans une enquête ; il en résulte alors un biais dans l'estimation puisque le revenu moyen calculé sur les seuls répondants est inférieur au vrai revenu moyen). De plus, les estimateurs classiques sont moins précis car basés sur un échantillon de répondants plus petit que l'échantillon initialement tiré.

Il faut donc utiliser des estimateurs alternatifs tenant compte de la non-réponse totale et la corrigeant.

1.1. Survol des principales méthodes de correction de la non-réponse totale

Quelques rappels théoriques sur les différents estimateurs corrigés de la non-réponse nous semblent nécessaires pour la suite de cet exposé. Ils se limiteront aux principaux estimateurs par repondération et par imputation qui apparaîtront par la suite. Leur but est toujours de réduire le biais introduit dans l'estimation par la présence de non-réponse totale.

Le point commun de tous ces estimateurs est que tout traitement de la non-réponse implique une modélisation plus ou moins implicite du phénomène de la non-réponse. La validité du traitement de la non-réponse et la qualité des estimations qui en découlent dépendent donc de l'adéquation du modèle de non-réponse à la réalité. Le modèle de non-réponse postulé reste de toute façon toujours non vérifiable, si bien qu'il persiste une incertitude sur la validité des résultats obtenus.

1.1.1. Première approche face à la non-réponse : "ne rien faire"

Une première solution est de bâtir les estimations sur les seules données des répondants (ce qui est toujours forcément le cas, faute d'autres données), mais sans apporter de correction particulière pour traiter la non-réponse. On estime alors par exemple la moyenne

de la population par la simple moyenne sur les répondants, le total par la moyenne sur les répondants multipliée par la taille de la population. Cette solution, quand elle est justifiée, a l'avantage d'être simple et d'être disponible même en l'absence d'information auxiliaire sur l'échantillon considéré.

Dans cette approche, on fait l'hypothèse implicite que les non-répondants sont "identiques" aux répondants, et on se contente d'un échantillon plus petit que prévu. On ignore alors complètement les non-répondants (Cf. Caron, 1996).

1.1.2. Correction par repondération

Les méthodes de repondération consistent à traiter la non-réponse en modifiant les poids de sondage des individus ayant répondu, afin de combler l'absence de certaines réponses. On attribue alors à chaque individu k une probabilité de réponse c_k, et si celle-ci est connue pour tous les individus, on dispose alors d'estimateurs sans biais en utilisant les réponses des répondants pondérées par les poids de sondage divisés par les probabilités de réponse. Dans le cas de l'estimation d'un total Y et d'un plan de sondage aléatoire simple de n individus parmi N, on obtient l'estimateur sans biais suivant :

$$\hat{Y} = \sum_{k \in R} \frac{Y_k}{\pi_k \times c_k} = \sum_{k \in R} \frac{Y_k}{\frac{n}{N} \times c_k} = \sum_{k \in R} w_k Y_k$$

où R désigne l'ensemble des répondants à l'enquête, π_k est la probabilité d'inclusion de l'individu k dans l'échantillon (égale ici à n/N) et w_k est le poids modifié du répondant k.

Toutefois, ceci suppose d'être en mesure d'estimer les probabilités de réponse individuelles. Les hypothèses faites pour cela conduisent à différents types d'estimateurs par repondération.

Si on dispose d'informations auxiliaires sur la population ou l'échantillon tiré, ces informations peuvent être utilisées pour améliorer l'estimation en présence de non-réponse. Plusieurs solutions sont envisageables en fonction de l'information disponible. Nous en donnons cinq exemples ici.

1.1.2.1. Mécanisme de réponse globalement uniforme (modèle de réponse naïf)

Une première approche est de considérer que tous les individus ont la même probabilité de répondre à l'enquête. Cette probabilité de réponse uniforme peut être estimée par n_R/n où n_R désigne le nombre de répondants. On obtient alors dans le cas d'un sondage aléatoire simple :

$$\hat{Y} = \sum_{k \in R} \frac{Y_k}{\frac{n}{N} \times \frac{n_R}{n}} = \frac{N}{n_R} \sum_{k \in R} Y_k \ .$$

Ceci revient exactement à ne rien faire pour corriger la non-réponse et à appliquer les formules des estimateurs habituels aux seuls répondants sans chercher à les corriger de la non-réponse.

1.1.2.2. Estimation des probabilités de réponse individuelles

Si on dispose de variables auxiliaires connues pour les répondants et les non-répondants à une même date, on peut tenter d'estimer les probabilités de réponse individuelles

par un modèle économétrique de type LOGIT ou PROBIT s'appuyant sur ces variables. Cette technique risque toutefois de conduire à des probabilités de réponse très dispersées, et donc à des estimateurs instables.

1.1.2.3. Mécanisme de réponse homogène à l'intérieur de sous-populations

Parce qu'il peut sembler utopique de parvenir à estimer la probabilité de réponse de chaque individu séparément, on peut chercher à regrouper les individus ayant certaines caractéristiques communes : en supposant que ces caractéristiques communes conduisent à un comportement de réponse proche, on se limite alors à estimer les probabilités de réponse des groupes ainsi formés, dits groupes de réponse homogènes (GRH).

On peut d'ailleurs également utiliser le modèle LOGIT construit sur des variables auxiliaires connues sur les répondants et les non-répondants pour constituer de tels GRH : on divise ainsi la population en sous-populations homogènes vis-à-vis de la non-réponse (le mécanisme de réponse globalement uniforme évoqué précédemment correspond au cas d'un unique GRH).

On estime alors la probabilité de réponse dans le GRH c par n_{cR}/n_c où n_{cR} est le nombre de répondants dans le GRH c et n_c la taille de l'échantillon tiré dans ce GRH. Pour un sondage aléatoire simple, on obtient l'estimateur sans biais suivant dit "estimateur pondéré par classe" :

$$\hat{Y} = \sum_c \sum_{k \in R \cap c} \frac{Y_k}{\frac{n}{N} \times \frac{n_{cR}}{n_c}} = \frac{N}{n} \sum_c n_c \bar{y}_{cR} \quad \text{où } \bar{y}_{cR} = \sum_{k \in R \cap c} \frac{Y_k}{n_{cR}}.$$

1.1.2.4. Utilisation de l'estimateur post-stratifié

On peut également utiliser l'estimateur post-stratifié formé en considérant comme post-strates les groupes de réponse homogènes définis précédemment, ceci dans le cas où les tailles N_c de ces GRH dans la population sont connues. On a alors :

$\hat{Y} = \sum_c N_c \bar{y}_c$ où $\bar{y}_c$ est estimé par $\bar{y}_{cR}$ en présence de non-réponse.

1.1.2.5. Utilisation du calage sur marges

La technique du calage sur marges peut également être utilisée pour obtenir un estimateur par repondération en utilisant de l'information auxiliaire.

On peut par exemple appliquer la méthode du raking-ratio à la correction de la non-réponse, en utilisant comme variables de calage des marges issues de variables de la base de sondage. Ceci suppose qu'on dispose dans la base de sondage de variables connues non seulement sur les répondants et les non-répondants de l'échantillon, mais aussi sur l'ensemble des individus non échantillonnés, et ceci à une même date pour tous. On modifie alors les poids de sondage des individus ayant répondu de manière à se caler sur les marges de ces variables auxiliaires dans toute la population de la base de sondage.

Si ces marges correspondent simplement à des comptages de la population de la base de sondage selon les modalités de deux variables auxiliaires indicées par i et h, on modifiera les poids des répondants de manière à caler le nombre d'individus dans le croisement ih

(estimé à partir des seuls répondants, $\hat{N}_{ih}$) sur les comptages N_i et N_h de la population (supposés connus). On a alors :

$\hat{Y} = \sum_i \sum_h \tilde{N}_{ih} \bar{y}_{ih}$ où $\tilde{N}_{ih}$ peut être obtenu avec le logiciel CALMAR (Cf. Sautory, 1993).

Une autre possibilité, qui sera utilisée également plus loin, est de "caler" les répondants sur l'échantillon complet (plutôt que sur la population). Il est nécessaire pour cela de disposer de variables connues sur tous les individus de l'échantillon, qu'ils aient répondu ou non.

Dans le cas où l'information auxiliaire porte à nouveau sur des comptages, au sein de l'échantillon complet cette fois, on cale n_{ihR} (le nombre de répondants dans le croisement *ih*) sur les comptages n_i et n_h supposés connus sur l'échantillon. Dans ce cas particulier, on n'utilise pas vraiment d'information auxiliaire externe, mais simplement de l'information portant sur l'échantillon.

1.1.3. Correction par imputation

Le principe des méthodes d'imputation est différent de celui des méthodes de repondération : au lieu de "gonfler" les poids des répondants pour compenser l'absence de certaines réponses comme on le fait dans la repondération, l'imputation consiste à remplacer les données absentes par des données "plausibles", en général issues ou estimées à partir de celles des répondants. Ce faisant, on complète la matrice croisant les individus et les variables, ce qui permet de traiter les données imputées comme des données réelles, bien que cette technique de redressement ne soit pas sans conséquence notamment sur la variance des estimateurs.

L'estimateur d'un total pour un échantillon de taille *n* s'écrit alors :

$\hat{Y} = \frac{N}{n} \sum_k \left(Y_k \delta_k + (1 - \delta_k) Y_k^* \right)$ où Y_k^* est la valeur imputée si l'individu *k* n'a pas répondu et δ_k vaut 1 si l'individu *k* a répondu et 0 sinon.

Toutes les méthodes d'imputation ont en commun de reposer sur un même modèle de régression du type : $y = x'\beta + \varepsilon$ où x est un vecteur de variables auxiliaires et ε est un vecteur de résidus. β est alors estimé sur la population des répondants par la méthode des moindres carrés ordinaires et sert à estimer $\hat{y}$ pour les non-répondants (dans la mesure où on dispose des variables auxiliaires x pour les répondants et les non-répondants).

On impute alors pour un individu *k* non-répondant la valeur $y_k^* = \hat{y}_k + \hat{\varepsilon}$.

La différence entre les diverses méthodes d'imputation repose ensuite sur la façon d'imputer, c'est-à-dire sur l'origine des données utilisées pour remplacer les données absentes.

Lorsque les résidus estimés sont arbitrairement posés égaux à zéro, on parle d'imputation déterministe et on a alors $y_k^* = \hat{y}_k$; dans le cas contraire, on parle d'imputation stochastique.

Nous présentons ici deux cas particuliers des méthodes d'imputation.

1.1.3.1. Prédiction par la moyenne

Dans ce premier cas particulier, la donnée manquante est remplacée par la moyenne observée pour les répondants. On se place alors dans un modèle de régression simple du type : $y_k = m + \varepsilon_k$. On procède à une imputation déterministe en attribuant à l'individu k non-répondant la valeur $y_k^* = \hat{m} = \bar{y}_R$ (où $\bar{y}_R$ est la moyenne calculée sur les répondants).

Si la population est répartie en C classes (ou groupes de réponse homogènes), la donnée manquante est remplacée par la moyenne observée parmi les répondants de la classe du non-répondant.

1.1.3.2. Hot-deck

Dans la technique dite du "hot-deck", la donnée manquante est remplacée par la valeur observée pour un répondant choisi au hasard, noté *j* ici : celui-ci est alors appelé le donneur. De manière un peu artificielle, cette méthode peut être présentée comme étant une version stochastique de la prédiction par la moyenne, l'aléa choisi étant égal à l'écart entre la valeur de la donnée du donneur et la moyenne de cette donnée sur les répondants : on attribue ainsi au non-répondant *k* la valeur $y_k^* = y_j = \hat{m} + (y_j - \hat{m})$.

Cette méthode revient en fait à repondérer le donneur *j* : si les poids d'origine des individus *k* et *j* étaient respectivement $\frac{1}{\pi_k}$ et $\frac{1}{\pi_j}$, le nouveau poids de *j* sera $\frac{1}{\pi_k} + \frac{1}{\pi_j}$. On trouve donc ici une similitude avec les méthodes de repondération.

Si la population est répartie en classes, on cherche un donneur de la classe du non-répondant.

Pour conclure sur ce rapide tour d'horizon des principales méthodes de correction de la non-réponse, on peut souligner que nous n'y avons pas évoqué une méthode moins souvent utilisée, la correction par remplacement. Cette méthode consiste à remplacer chaque non-répondant par une unité de la population qui n'avait pas été sélectionnée initialement dans l'échantillon, puis à traiter ce remplaçant comme s'il avait été sélectionné dans l'échantillon de départ. Le but est alors de sélectionner un remplaçant ayant des caractéristiques proches de celle du non-répondant (Cf. Chapman, 2003).

Cette technique, rarement utilisée dans les enquêtes nationales réalisées par l'INSEE, l'est davantage au niveau régional ou dans la sphère privée, lorsque le commanditaire de l'enquête spécifie un nombre de répondants attendus dans le contrat passé avec l'organisme enquêteur.

1.1.4. La problématique de cette étude

Après les quelques rappels théoriques qui précèdent, il est légitime de s'interroger sur les conséquences (en termes de biais et de précision) des deux méthodes "concurrentes" de correction de la non-réponse totale sur les estimations. Notre objectif est de les comparer, dans la perspective actuelle de rénovation des Enquêtes Annuelles d'Entreprise (EAE).

Or le choix entre correction de la non-réponse totale par repondération ou par imputation est difficile, et il n'existe pas de considération théorique permettant de dire que l'une est préférable à l'autre. A l'INSEE, on utilise traditionnellement plutôt la correction par repondération pour les enquêtes réalisées auprès des ménages, et plutôt la correction par

imputation dans les enquêtes réalisées auprès des entreprises (en particulier, le processus de correction de la non-réponse totale des EAE, très complexe, est essentiellement basé sur la technique de l'imputation : en l'absence d'informations sur une entreprise non-répondante dans les autres sources disponibles, on utilise la méthode du hot deck, non étudiée ici). Le but de cette étude est en partie de s'interroger sur la pertinence de cette "tradition" des EAE.

Il faut souligner que dans certains cas, ces deux méthodes sont identiques pour des estimations globales. Ce sera le cas si, pour l'estimation d'une moyenne globale, on corrige la non-réponse totale par une méthode de repondération uniforme ou par une méthode d'imputation par la moyenne globale des répondants (Cf. Caron, 2003).

En revanche, correction par repondération et par imputation ne sont pas équivalentes pour des estimations sur des domaines qui ne coïncident pas avec les groupes de réponse homogènes utilisés pour corriger la non-réponse : l'encadré 1 suivant illustre cela sur un exemple simple.

ENCADRE 1 : Exemple simple sur les effets de la correction de la non-réponse par repondération et par imputation

Dans le cas de l'estimation de la moyenne globale, correction par repondération uniforme et correction par imputation de la moyenne globale des répondants donnent le même résultat. En revanche, ce n'est pas le cas de l'estimation de la moyenne de domaines transversaux par rapport aux groupes de réponse homogènes au niveau desquels on corrige la non-réponse. L'exemple simple suivant illustre cette différence pour l'estimation d'un total.

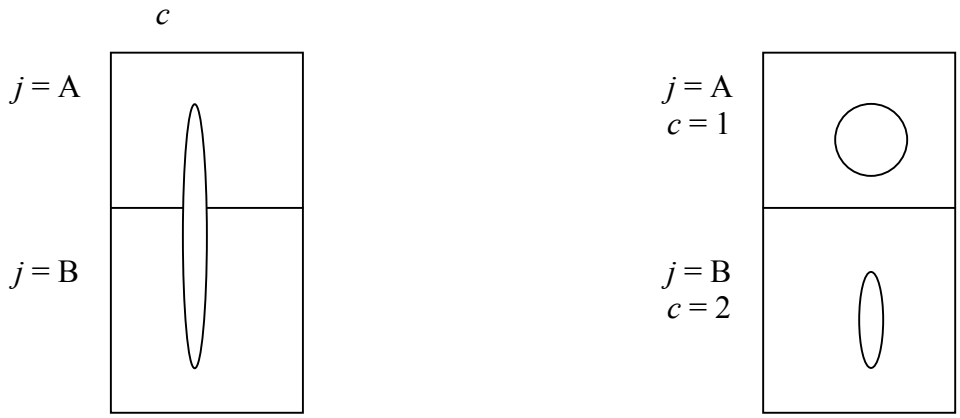
Considérons le cas minimal où on a un seul GRH et une variable de région dont les deux modalités A et B ne coïncident pas avec le découpage en GRH puisqu'elles sont toutes deux présentes dans l'unique GRH. Sur quatre observations, deux sont des non-réponses :

N° d'obs.	Région j	Variables d'intérêt				Poids	Réponse ou Non-rép.
1	A	1	x	x	x	2	R
2	A	1	x	x	x	2	R
3	A	.	.	.	.	2	NR
4	B	.	.	.	.	2	NR

Si on cherche simplement à calculer un estimateur global, i.e. ne distinguant pas les deux régions A et B, on obtiendra la même chose que l'on corrige la non-réponse par repondération ou par imputation. Considérons par exemple l'estimation du nombre d'entreprises (donc une variable d'intérêt valant théoriquement 1 pour chaque observation) : dans le cas de la repondération, on augmentera les poids des observations 1 et 2 afin d'avoir un total de 8 individus ; dans le cas de l'imputation, on utilisera les réponses de 1 ou 2 pour compléter les réponses de 3 et 4, pour aboutir finalement à 8 entreprises également.

En revanche, si on souhaite obtenir un estimateur par région, la correction par imputation donnera ici 2 entreprises dans la région B et 6 dans la région A, alors que la correction par repondération donnera 8 individus dans la région A : avec la repondération, on donne donc plus de poids que réellement à la région A et on perd la région B. La repondération conduit ainsi ici à une perte d'information sur la variable région.

Ceci est lié au fait que le GRH considéré est à cheval sur plusieurs régions comme l'illustre le schéma de gauche suivant :



Il en serait autrement si le GRH et les cellules délimitées par les modalités de la variable région coïncidaient comme sur le schéma de droite ci-dessus.

Or le fait que domaines d'intérêt et groupes de réponse homogènes de correction de la non-réponse ne coïncident pas peut intervenir fréquemment, et ceci pour au moins trois raisons :

- d'abord parce qu'on ne peut pas prendre tous les domaines d'intérêt, en général nombreux, comme cellules de correction de la non-réponse (il faut choisir) ;
- ensuite parce que la correction de la non-réponse n'est pas toujours possible dans les croisements fins de variables que peuvent définir certains domaines (s'il n'y a pas assez de répondants dans ces croisements fins, la correction de la non-réponse nécessitera des regroupements de ces croisements fins) ;
- enfin parce qu'un domaine d'intérêt n'explique pas toujours le comportement de non-réponse (il peut donc n'avoir rien en commun avec les groupes de réponse homogènes).

Dans cette étude, nous nous ciblons sur les enquêtes auprès des entreprises, dans lesquelles le deuxième élément cité ci-dessus (au moins) est susceptible d'intervenir, si bien que domaines d'intérêt et groupes de réponse homogènes n'y coïncident pas toujours ; de ce fait correction de la non-réponse par repondération et correction par imputation au niveau de ces groupes de réponse homogènes ne sont pas équivalentes.

En effet, les enquêtes auprès des entreprises constituent un cadre particulier de la méthodologie d'enquête, souvent moins connu que les enquêtes auprès des ménages et posant des problèmes spécifiques. Sans prétendre citer toutes leurs spécificités ici, rappelons brièvement que la sphère des entreprises se caractérise par une hétérogénéité plus grande que celle des ménages, et dans laquelle un plan de sondage, en général stratifié, conduit la plupart du temps à des échantillons de petite taille au sein des croisements les plus petits de la stratification.

De plus, la population des entreprises étant somme toute assez petite par rapport à celle des ménages (de l'ordre de 2,2 millions), il arrive qu'une même entreprise soit sollicitée pour plusieurs enquêtes dans un laps de temps réduit. Bien que les services d'enquête fassent des efforts de coordination négative de leurs échantillons respectifs pour alléger la charge d'enquête qui pèse sur les entreprises, les enquêtes sont souvent perçues par les entreprises comme des formulaires qui viennent alourdir la paperasse administrative qui leur est imposée. Ce problème est accentué pour les plus petites entreprises qui n'ont pas d'équipe dédiée aux travaux administratifs. Le type de collecte retenu pour les enquêtes auprès des entreprises, par voie postale en général, affecte aussi le nombre de réponses obtenu. Au total, les taux de réponse sont souvent plus faibles dans les enquêtes auprès des entreprises que dans les enquêtes auprès des ménages. Les nombres de répondants dans les croisements les plus fins de la stratification sont donc souvent de fait très petits.

Dans ce cadre, repondération et imputation n'ont pas les mêmes effets pour des estimations sur domaines d'intérêt. Nous nous intéressons ici plus précisément aux Enquêtes Annuelles d'Entreprise (EAE), qui constituent les enquêtes structurelles françaises les plus importantes et entrent tout à fait dans le schéma précédent de petits échantillons et de petits nombres de répondants au niveau le plus fin de la stratification. Notre but est donc de comparer correction de la non-réponse totale par repondération et par imputation dans ce cas précis.

Ce type de comparaison ressemble au travail présenté par R.J.A. Little dans un article publié en 1986 dans l'International Statistical Review sous le titre "Survey Nonresponse Adjustments for Estimates of Means". Notre propre comparaison des différentes méthodes de correction de la non-réponse s'inspire donc très largement de cet article, tant du point de vue méthodologique que du point de vue des notations et des choix faits à différentes étapes : il

nous paraît ainsi nécessaire de le résumer, avant de présenter l'adaptation que nous en avons faite aux Enquêtes Annuelles d'Entreprise dans la partie suivante de ce rapport.

1.2. Une méthodologie de comparaison de la repondération et de l'imputation : l'exemple de Little

Dans cet article de 1986, Little se fixe pour objectif de comparer trois estimateurs de la moyenne en présence de non-réponse, dans une population supposée initialement découpée en C groupes de réponse homogènes (GRH) : l'estimateur par repondération (repondéré par l'inverse du taux de réponse du GRH), l'estimateur post-stratifié utilisant la taille supposée connue de la population dans les GRH et enfin l'estimateur par imputation de la moyenne du GRH. Un quatrième estimateur fait également partie de ses comparaisons : la moyenne des répondants.

1.2.1. Contexte des comparaisons de Little

Little base ses comparaisons sur des simulations d'enquête entièrement fictives, au sens où les variables sont totalement construites (elles ne correspondent à aucune variable réelle), les taux de réponse comme les répartitions de la population par classe sont inventés, etc.

Grâce à ces simulations, les différents estimateurs retenus sont tout d'abord comparés pour l'estimation de la moyenne globale (cas où repondération et imputation sont équivalentes) ; dans un second temps, ils sont comparés pour l'estimation de la moyenne d'un domaine correspondant à la modalité *j* d'une variable d'intérêt dont les classes ne coïncident pas avec les groupes de réponse homogènes (cas où repondération et imputation diffèrent).

Le dispositif léger des simulations fictives permet à Little de faire varier plusieurs paramètres dans ses simulations, pour obtenir des jeux de données différents. Il peut alors calculer les biais et variances des estimateurs retenus dans chaque configuration, et ainsi les comparer.

1.2.2. Les différents estimateurs étudiés par Little et la décomposition du biais

Concernant l'estimation de la moyenne globale pour laquelle l'imputation donne le même résultat que la repondération, Little retient trois estimateurs :

- la moyenne des répondants $\bar{y}_R$ (ceci correspond à "ne rien faire", Cf. paragraphe 1.1.1),
- la moyenne ajustée par l'inverse du taux de réponse dans le GRH, notée $\bar{y}_A$ (c'est un estimateur par repondération comme au paragraphe 1.1.2.1, qui est équivalente à la moyenne corrigée de la non-réponse par imputation de la moyenne $\bar{y}_{cR}$ du GRH)
- et l'estimateur post-stratifié $\bar{y}_S$ (qui est également un estimateur par repondération comme on l'a évoqué au paragraphe 1.1.2.4).

Little calcule pour ces trois estimateurs leurs biais et variances, conditionnels au vecteur des tailles d'échantillon par GRH et au vecteur des nombres de répondants par GRH.

Il démontre alors que le biais conditionnel peut toujours être décomposé en deux termes, le premier étant noté C et le second LSB (pour "large-sample bias"). En effet, dans le cas d'échantillons très grands, le terme C tend vers 0 et seul le terme LSB subsiste. Par contre, si

l'on fait l'hypothèse que le modèle de non-réponse utilisé est correct (c'est-à-dire que $\overline{Y}_{cR} = \overline{Y}_c$), le terme LSB est nul (ceci sera visible dans nos propres formules, Cf. § 2.2).

Pour les estimateurs de la moyenne d'un domaine transversal *j*, cette fois correction de la non-réponse par imputation et correction par repondération ne sont pas équivalentes. Little retient donc, en plus des estimateurs du domaine *j* suivant la même logique que les trois estimateurs présentés précédemment, un estimateur obtenu par correction de la non-réponse par imputation. À nouveau, les biais de ces estimateurs peuvent se décomposer en deux termes C et LSB ayant les mêmes particularités que celles exposées précédemment.

1.2.3. Les simulations de Little

Le dispositif léger mis en place par Little pour ses simulations lui permet de faire varier les paramètres de ces simulations : il fait varier au total six paramètres prenant chacun deux valeurs, d'où 2^6 problèmes différents (64 jeux de données). Ces six paramètres sont :

- le vecteur des taux de réponse dans les GRH *c* $\{B_c\}$
- le vecteur des taux de réponse dans les croisements *cj* pour un GRH *c* $\{B_{cj}\}$
- le vecteur des moyennes de la population dans les GRH *c* $\{\overline{Y}_c\}$
- le vecteur des moyennes de la population dans les croisements *cj* pour un GRH *c* $\{\overline{Y}_{cj}\}$
- la corrélation entre les taux de réponse $\{B_c\}$ et les moyennes de GRH $\{\overline{Y}_c\}$
- et enfin la taille de l'échantillon *n*.

Pour chacun de ces paramètres, deux niveaux sont possibles : faible (1) ou fort (2). Le niveau faible du vecteur des taux de réponse correspond par exemple à de petites variations du taux de réponse entre les différents GRH *c* (entre 60 et 70 %), tandis que son niveau fort correspond à une variation plus importante d'un GRH à l'autre, entre 40 et 90 %. Autre exemple, la taille de l'échantillon est de 240 ou 2400 individus.

Pour chacun des 64 problèmes, 100 vecteurs de tailles d'échantillon dans les croisements *cj* (notés $\{n_{cj}\}$) ont été générés indépendamment (selon une loi multinomiale de paramètres *n* et les N_{cj}/N), ainsi que les 100 vecteurs correspondant aux nombres de répondants dans les croisements *cj* (notés $\{n_{cjR}\}$, tirés avec une loi binomiale de paramètres n_{cj} et B_{cj}). En effet, les expressions des biais et variances des estimateurs proposés par Little peuvent être calculés avec les n_{cj} et les n_{cjR} et ne nécessitent pas que ces 100*64 échantillons soient réellement tirés : il suffit de simuler les tailles d'échantillon obtenues dans les différents croisements.

1.2.4. Les principaux résultats obtenus par Little

Pour comparer les différents estimateurs, Little calcule pour chaque jeu de données la moyenne (sur les 100 échantillons tirés) de la racine de l'erreur quadratique moyenne conditionnelle (ou RMSE). La racine de la somme de la variance et du biais au carré est en effet un critère global de précision. Il calcule également la différence relative des racines d'erreur quadratique moyenne par rapport à un estimateur pris comme référence, la référence étant ici l'estimateur ajusté par repondération : $REL(\overline{y}_R) = 100 * (RMSE(\overline{y}_R) / RMSE(\overline{y}_A) - 1)$. On désignera par la suite cet indicateur sous le nom d'"erreur quadratique moyenne *relative*". Little présente les résultats obtenus pour chacun des jeux de données, ainsi que ce qu'il obtient en faisant la moyenne sur les 64 jeux de données testés. Voici ses principaux résultats :

- $\overline{y}_R$ est légèrement meilleur que $\overline{y}_A$ lorsque l'échantillon est petit et lorsque les moyennes de la population dans les différents GRH sont proches ; dans tous les autres cas, $\overline{y}_A$ est meilleur que $\overline{y}_R$;

- la post-stratification constitue presque toujours un gain en terme d'erreur quadratique moyenne par rapport à l'estimateur ajusté par repondération, sauf dans le cas où tous les paramètres sont au niveau fort, cas où $\bar{y}_A$ est plus performant que $\bar{y}_S$;
- l'estimateur corrigé par imputation donne une erreur quadratique moyenne inférieure à celle de l'estimateur ajusté par repondération dans le cas où les moyennes de la population dans les croisements *cj* pour un GRH *c* sont assez proches les unes des autres. Ceci vient du fait que l'imputation limite la variance alors que ce n'est pas le cas de la repondération. Cet avantage est plus visible quand l'échantillon est petit. Le reste du temps, en revanche, l'estimateur par imputation est moins performant que l'estimateur par repondération.

II. Comparaison de la correction de la non-réponse par repondération et par imputation : une application de la méthodologie de Little aux Enquêtes Annuelles d'Entreprise

La transposition de la méthodologie développée par Little dans un cadre proche de celui des EAE nous a semblé intéressante pour comparer la correction de la non-réponse par repondération et par imputation dans un environnement à la fois moins arbitraire que le schéma fictif de l'article de Little et moins complexe que le véritable plan de sondage des EAE. Cet environnement particulier a nécessité un certain nombre de choix pour adapter les simulations de Little au champ des Enquêtes Annuelles d'Entreprise.

Nous exposons ces choix dans une première sous-partie. Nous donnons ensuite les formules des estimateurs retenus dans ce cadre spécifique, ainsi que celles de leurs biais et variances : en effet, une partie importante de notre travail a été consacrée à l'obtention des formules proposées par Little dans le contexte différent que nous avons retenu. Enfin, nous en venons aux principaux résultats obtenus par nos simulations ; ces simulations ont nécessité un investissement en programmation en macro-langage SAS (une partie des programmes est présentée en Annexe 5) qui a constitué une autre partie importante de cette étude.

Le principal objectif de cette étude d'une durée de deux mois était de fournir les bases pour un travail plus vaste à l'avenir. Ces bases ont été posées d'une part par le travail théorique consacré à la réécriture des estimateurs et d'autre part par le travail de programmation informatique. Le temps imparti étant limité, nous avons dû par contre nous en tenir dans les simulations à des hypothèses un peu frustes (nous utilisons par exemple ici un taux de réponse uniforme) et à un nombre limité de jeux de données. Toutefois, le travail accompli au cours de cette étude fournit tous les outils nécessaires pour le prolonger ultérieurement.

2.1. Adaptation de la méthodologie de Little au cadre particulier des Enquêtes Annuelles d'Entreprise

2.1.1. Présentation générale des particularités des Enquêtes Annuelles d'Entreprise

Les Enquêtes Annuelles d'Entreprise (EAE) constituent un pilier important du dispositif français d'enquêtes auprès des entreprises. Coordonnées par l'INSEE, elles sont menées par les services statistiques ministériels responsables de chaque secteur d'activité qu'elles couvrent. Elles concernent l'industrie, l'agro-alimentaire, les transports, la construction, les services et le commerce, en France métropolitaine et dans les DOM. Leur échantillon annuel compte plus de 200 000 entreprises.

On peut tout de suite préciser que les échantillons des EAE comptent une partie exhaustive (qui concerne les plus grandes entreprises) et une partie sondée (à part pour les EAE Industrie et Agro-alimentaire où n'existe qu'une partie exhaustive car on interroge tout le champ des entreprises de 20 salariés ou plus) ; la partie exhaustive ne nous intéresse pas ici et nous l'avons écartée d'emblée pour nous concentrer sur la partie sondée des EAE.

Concernant cette partie sondée des EAE, on y retrouve en particulier les spécificités des enquêtes auprès des entreprises rappelées au paragraphe 1.1.4, notamment en ce qui concerne la petite taille des échantillons au sein des croisements les plus petits de la stratification et les faibles nombres de répondants dans ces mêmes croisements.

Plus précisément, la stratification des EAE est construite sur le croisement de deux variables : l'activité principale exercée (APE) et la tranche d'effectif salarié. Le système de tirage retenu[3] donne de plus un rôle particulier à la région d'implantation des entreprises dans ce plan de sondage, car il s'apparente à un tirage systématique par région au sein des strates activité*tranche d'effectif (Cf. Cotton et Hesse, 1992). Toutefois, la région ne peut pas réellement être considérée comme un critère de stratification dans les EAE.

Pour adapter la méthodologie de Little à un cadre aussi proche que possible de celui des EAE, il faut donc considérer un plan de sondage stratifié par activité et tranche d'effectif, dans lequel la variable région définit des domaines d'intérêt mais n'est pas une véritable variable de stratification.

Or, les échantillons dans les croisements activité*tranche d'effectif*région étant de taille en général très réduite, la correction de la non-réponse ne peut pas prendre en compte la région, mais doit regrouper plusieurs régions afin d'obtenir des groupes de réponse homogènes comptant suffisamment de répondants. On retrouve donc ici un cadre dans lequel correction de la non-réponse par repondération ou par imputation n'auront pas les mêmes effets pour des estimations sur domaines d'intérêt (par région).

Nous nous situerons par la suite au niveau d'une activité donnée (car il aurait été trop volumineux de travailler sur l'ensemble des activités en même temps), si bien qu'à ce niveau-là notre plan de sondage sera stratifié par tranche d'effectif salarié.

On indicera par *i* la tranche d'effectif et par *h* la région.

On a alors, pour une activité donnée, l'expression suivante de la moyenne :

$\overline{Y} = \sum_i \frac{N_i}{N} \overline{Y}_i = \sum_i P_i \overline{Y}_i$ où $\overline{Y}_i$ est la moyenne de la population dans la tranche d'effectif *i*.

On obtient ainsi comme estimateur de la moyenne :

$\hat{\overline{y}} = \sum_i P_i \overline{y}_i^*$ où $\overline{y}_i^*$ est l'estimateur d'Horvitz-Thompson de $\overline{Y}_i$ en l'absence de non-réponse, et l'estimateur de $\overline{Y}_i$ corrigé de la non-réponse quand il y en a.

2.1.2. Choix effectués pour construire les simulations selon la méthodologie proposée par Little en 1986

2.1.2.1. Choix des activités étudiées

Nous avons tout d'abord décidé de nous concentrer pour nos simulations sur un petit nombre d'activités. En effet, nous n'étions pas en mesure de travailler sur un grand nombre d'activités différentes car ceci aurait été trop fastidieux et aurait nécessité davantage de temps. Mais en même temps, il aurait été dommage de se limiter à une seule activité dans la mesure où considérer différentes activités constitue un moyen de faire varier les conditions de l'expérience (de la même manière que Little fait varier ses six facteurs), grâce aux différences qui existent entre les divers secteurs d'activité.

Pour choisir les activités sur lesquelles nous allions travailler, nous nous sommes limitées délibérément au champ de l'EAE-Commerce sur laquelle d'autres travaux méthodologiques avaient déjà été menés (Cf. Caron et Fuentes, 2002). Au sein du champ du commerce, nous n'avons pas cherché à retenir les activités les plus représentatives de ce secteur, mais plutôt

[3] Ce système de tirage utilise l'outil de tirage et de coordination d'échantillons OCEAN, basé sur les croisements activité*tranche d'effectif*région.

des activités présentant un certain nombre de particularités intéressantes pour les simulations à venir. Nous avons recherché en particulier des activités se caractérisant dans l'EAE à la fois par les échantillons les plus importants (afin de ne pas être complètement bloquées par des échantillons minuscules dans les croisements activité*tranche d'effectif*région) et par des taux de sondage modestes (pour que l'hypothèse de tirage avec remise, utilisée par la suite, reste acceptable). Nous avons trouvé de telles activités dans le commerce de détail, et nous avons retenu les cinq activités suivantes (dont on donne entre parenthèse les codes NAF) :

- les charcuteries (151F)
- les boulangeries (158C)
- l'habillement (524C)
- les librairies-papeteries (524R)
- les fleuristes (524X).

Nous présentons ci-dessous à titre illustratif la taille de la population, la taille de l'échantillon EAE et les taux de sondage de ces cinq activités pour l'EAE réalisée en 2003 et portant sur l'exercice fiscal de 2002.

	Taille de l'échantillon par tranche d'effectif									Taille de l'échan-tillon	Taille de la popu-lation	Taux de sondage	Libellé de l'activité
APE	**00**	**01**	**03**	**11**	**12**	**13**	**21**	**22**	**30**				
151F	34	92	36	50	55	25	14		1	307	7869	3,90%	Charcuteries
158C	52	216	81	64	44	24	26	6		513	32406	1,58%	Boulangeries
524C	324	343	246	239	272	188	80	26	47	1765	36789	4,80%	Habillement
524R	113	266	116	89	74	55	16	6	12	747	18604	4,02%	Livres, journaux, papeterie
524X	71	148	50	191	94	72	19	6	1	652	14804	4,40%	Commerce détail Fleurs

NB : en grisé, on signale la partie de l'activité enquêtée de manière exhaustive, qui correspond à certaines tranches d'effectif ; la partie non grisée est la partie sondée. Les codes des tranches d'effectif signifient : 00="0 salarié", 01="1 à 5 salariés", 03="6 à 9 salariés", 11="10 à 19 salariés", 12="20 à 29 salariés", 13="30 à 49 salariés", 21="50 à 99 salariés", 22="100 à 199 salariés", et 30="200 salariés et plus".

2.1.2.2. Choix des nomenclatures pour la tranche d'effectif et la région

Concernant les deux autres dimensions retenues (tranche d'effectif et région), nous avons procédé à des regroupements avant même de commencer les simulations, car les croisements activité*tranche d'effectif*région contenaient très souvent zéro ou une seule entreprise en l'absence de tout regroupement.

Ainsi, nous avons choisi de retenir la nomenclature des tranches d'effectif utilisée pour le tirage des EAE, qui regroupe les tranches habituelles des "1 ou 2 salariés" et "3 à 5 salariés" en une seule tranche des "1 à 5 salariés".

Pour la région, nous avons préféré à la nomenclature des 22 régions administratives la nomenclature des ZEAT (Zones d'Etudes et d'Aménagement du Territoire) en 8 postes :

1 Ile de France
2 Champagne-Ardenne, Haute-Normandie, Centre, Bourgogne
3 Nord
4 Alsace-Lorraine
5 Pays-de-la-Loire, Bretagne, Poitou-Charente
7 Aquitaine, Midi-Pyrénées, Limousin
8 Rhône-Alpes, Auvergne
9 Languedoc-Roussillon, Provence-Alpes-Côte d'Azur.

2.1.2.3. Restriction du champ d'étude à la partie sondée de l'EAE

Par ailleurs, comme nous l'avons évoqué plus haut, nous avons restreint le champ de nos simulations à la seule partie sondée dans l'EAE-Commerce pour les activités retenues (c'est-à-dire moins de 50 salariés pour les activités 151F et 158C, et moins de 20 salariés pour les activités 524C, 524R et 524X). En effet, pour la partie interrogée de façon exhaustive, c'est-à-dire pour les plus grosses entreprises, la correction de la non-réponse totale ne peut être envisagée qu'en dernier recours : une grosse entreprise non-répondante est trop spécifique pour qu'on puisse la "remplacer" (en repondérant ou en imputant) par une autre (on ne peut qu'essayer d'obtenir son questionnaire en la relançant).

2.1.2.4. Tailles de population issues de la base de sondage de l'EAE

La nomenclature de base étant fixée, nous avons pu obtenir, grâce au dernier échantillon tiré de l'EAE-Commerce et grâce à la base de sondage utilisée pour ce tirage, les tailles réelles de la population des entreprises dans les activités retenues (les N), dans les tranches d'effectif de ces activités (les N_i) et dans les croisements tranche d'effectif*région (ZEAT) de ces activités (les N_{ih}). Nous en avons également tiré les tailles réelles de l'échantillon tiré pour l'EAE portant sur l'exercice fiscal 2002 (les n et les n_i), éléments qui nous ont été utiles pour les simulations d'échantillons proches de la réalité comme on le verra au paragraphe 2.1.3.

*2.1.2.5. Choix de la source pour évaluer les moyennes et variances dans les croisements activité*tranche d'effectif*région*

Restait alors à déterminer ce que serait dans nos simulations la variable d'intérêt notée Y par Little. Toujours dans l'idée de faire varier les conditions de l'expérience, nous avons retenu quatre variables intéressantes a priori pour des entreprises du commerce, et qui font partie des variables les plus importantes de l'EAE-Commerce : l'effectif, le chiffre d'affaires, la marge commerciale et la valeur ajoutée.

Pour calculer les biais et variances conditionnelles sur lesquels se basent nos comparaisons d'estimateurs, il est nécessaire de disposer des vraies moyennes et dispersions dans les croisements activité*tranche d'effectif*région, comme cela apparaîtra dans les formules du paragraphe 2.2. Pour en disposer, on peut soit les estimer (à partir des résultats de l'EAE-Commerce elle-même par exemple), soit trouver une source externe dans laquelle ces grandeurs sont connues sur l'ensemble de la population (solution a priori plus fiable que l'estimation). L'INSEE disposant d'une telle source externe exhaustive, SUSE[4], c'est cette deuxième solution que nous avons retenue (après avoir vérifié qu'on y obtenait des ordres de grandeur comparables aux moyennes et variances calculées grâce aux résultats de l'EAE).

2.1.3. Particularités de cette adaptation par rapport à l'article de Little

Notre adaptation de la méthodologie de Little au cadre inspiré des EAE que nous retenons présente un certain nombre de particularités par rapport à ce que décrit Little dans son article de 1986, particularités que nous soulignons dans ce paragraphe.

[4] Système Unifié de Statistiques d'Entreprises : cette source recense en principe toutes les entreprises françaises, et réunit des données issues des liasses fiscales et des données tirées des EAE.

2.1.3.1. Quatre jeux de données

Comme Little le fait, nous avons souhaité faire varier plusieurs paramètres dans nos simulations afin de pouvoir comparer les performances relatives des différents estimateurs dans des situations diverses. Comme lui également, nous comparons ces estimateurs en calculant une erreur quadratique moyenne *relative* (Cf. *REL* défini au paragraphe 1.2.4), dont l'estimateur de référence est dans notre cas la moyenne simple des répondants. Toutefois, nous n'avons obtenu que quatre jeux de données au sens strict du terme, contre 64 dans le cas de Little (pour un dispositif plus léger dans son cas). Ils sont obtenus en faisant varier deux paramètres, la taille de l'échantillon et le taux de réponse général à l'enquête, chacun prenant deux niveaux :

- le jeu de données A correspond à la taille d'échantillon réelle dans l'EAE actuelle et à un taux de réponse global de 80 %, proche du taux de réponse réel dans l'EAE-Commerce ;
- le jeu de données B correspond à la taille d'échantillon réelle dans l'EAE actuelle et à un taux de réponse global plus faible, égal à 60 % ;
- le jeu de données C correspond à la moitié de la taille d'échantillon réelle dans l'EAE actuelle (pour chaque tranche d'effectif de chaque activité, afin de conserver constante la répartition de l'échantillon par tranche d'effectif) et à un taux de réponse global de 80 % ;
- le jeu de données D correspond à la moitié de la taille d'échantillon réelle dans l'EAE actuelle (pour chaque tranche d'effectif aussi) et à un taux de réponse global de 60 %.

Le nombre de répondants décroît donc entre le jeu de données A et le jeu de données D.

Par ailleurs, en un sens moins strict de l'expression "jeu de données", on peut également considérer que nous faisons varier des paramètres en considérant différentes activités et différentes variables d'intérêt. Dans cette perspective, notre exercice joue sur davantage de paramètres. Nous verrons plus loin dans ce paragraphe que nous faisons aussi varier la méthode de simulation des échantillons.

2.1.3.2. Taux de réponse constant au sein d'un même jeu de données

Par souci de simplicité, nous avons choisi d'adopter un taux de réponse unique quel que soit le croisement activité*tranche d'effectif*région considéré. Ce taux de réponse constant est noté B, et le fait qu'il soit constant joue un rôle important pour les calculs des expressions des biais et variances des estimateurs retenus dans nos simulations (Cf. Annexe 1).

2.1.3.3. Deux méthodes de tirage simulé d'échantillons

Comme Little le fait, nous approximons l'espérance par la moyenne calculée sur 100 échantillons tirés.

De plus, comme Little à nouveau, nous ne tirons pas réellement ces 100 échantillons pour chaque jeu de données : nous nous contentons de simuler le vecteur des tailles d'échantillon dans les croisements ih (les $\{n_{ih}\}$) et le vecteur des nombres de répondants dans les croisements ih (les $\{n_{ihR}\}$).

La plus importante particularité de nos simulations par rapport à la méthodologie proposée par Little réside sans doute dans le fait que nous avons retenu deux méthodes différentes pour simuler le tirage des échantillons (et donc obtenir les $\{n_{ih}\}$). Dans les deux méthodes, afin de rester proche du cadre réel des EAE, ces échantillons sont tirés en respectant les rapports n_i/n

réels constatés dans l'échantillon tiré pour l'EAE portant sur 2002. La taille globale de l'échantillon n variera selon le jeu de données comme on l'a évoqué ci-dessus.

- Dans la méthode a), on réalise un tirage aléatoire simple (TAS) dans chaque tranche d'effectif i, la répartition par région h se faisant selon une loi multinomiale de paramètres (n_i, P_{i1}, …, P_{iH}) (si on considère qu'il y a en tout H régions), où $P_{ih} = \frac{N_{ih}}{N_i}$.

  En utilisant ainsi la loi multinomiale, on fait l'approximation d'un tirage avec remise (permis en théorie par le choix d'activités pour lesquelles les taux de sondage sont assez faibles), car la simulation d'une loi hypergéométrique (équivalent sans remise de la loi multinomiale) serait bien plus complexe. Dans ce cas donc, la répartition de l'échantillon par région est tout à fait aléatoire.

  Nous avons réalisé, pour chaque activité et chaque jeu de données, 100 simulations de ce type, d'où nous avons tiré 100 vecteurs $\{n_{ih}\}$.

- Dans la méthode b), on utilise un tirage aléatoire simple par arrondi systématique dans un fichier trié (pour chaque activité) par tranche d'effectif et par région. L'algorithme de tirage aléatoire simple par arrondi systématique (décrit notamment par Ardilly, 1994) permet de réaliser, pour chaque activité et chaque tranche d'effectif, un tirage à pas systématique égal au rapport N_i/n_i, même si ce rapport n'est pas entier. Ce procédé de tirage est très proche de ce que donne le tirage actuel des EAE. Il se rapproche de plus beaucoup d'un tirage stratifié par région : les tailles d'échantillon dans les croisements activité*tranche d'effectif*région sont moins aléatoires que dans la méthode a), si bien qu'on s'attend à ce que les estimations soient plus stables avec les échantillons issus de la méthode b).

  Nous avons réalisé pour chaque activité et chaque jeu de données 5 simulations de ce type, d'où nous avons tiré 5 vecteurs $\{n_{ihp}\}$ (l'indice p signifie "tiré avec un pas systématique").

Dans un deuxième temps, il faut simuler les nombres de répondants dans les croisements ih (les $\{n_{ihR}\}$). Que le vecteur des tailles d'échantillon dans les croisements ih ait été obtenu par la méthode a) ou par la méthode b), n_{ihR} est obtenu en générant une loi binomiale de paramètres (n_{ih}, B). Nous avons fait cela une fois pour chaque vecteur tiré selon la méthode a) ($\{n_{ih}\}$), et 20 fois pour chaque vecteur tiré selon la méthode b) ($\{n_{ihp}\}$), ce qui donne au total 100 vecteurs de nombres de répondants pour chacune des deux méthodes de tirage.

Comme Little le fait, on calcule ensuite pour chaque échantillon le biais et la variance des estimateurs considérés, conditionnellement aux vecteurs $\{n_{ih}\}$ et $\{n_{ihR}\}$. L'espérance conditionnelle du biais et de la variance est approximée par la moyenne sur 100 échantillons, pour chacune des deux méthodes.

2.1.3.4. Utilisation de regroupements automatiques des régions h

Toutefois, ayant simulé les vecteurs $\{n_{ih}\}$ et $\{n_{ihR}\}$, on n'est pas immédiatement en mesure de calculer les biais et variances conditionnels correspondant. En effet, il est nécessaire d'obtenir au moins un répondant par groupes de réponse homogènes dans lesquels on va procéder à la correction de la non-réponse. Or ce n'est pas toujours le cas dans les croisements ih. Il faut donc procéder à des regroupements de régions h en "supra-régions" g de manière à former des "groupes de réponse homogènes" de remplacement ig

convenablement pourvus en répondants[5]. Nous avons arbitrairement fixé un seuil de 5 répondants minimum par groupe *ig* pour nos simulations.

Pour ce faire, nous avons choisi de pré-définir 3 niveaux de regroupements possibles des régions *h* en supra-régions *g* (selon une logique purement géographique). Les trois niveaux de regroupements pré-définis sont les suivants :

- les supra-régions *g* numérotées 10, 11, 12 et 13 regroupent respectivement les régions *h* 1 et 2, 3 et 4, 5 et 7, 8 et 9 ;
- les supra-régions *g* numérotées 14 et 15 regroupent respectivement les supra-régions 10 et 11, 12 et 13 (14 correspond au nord de la France, 15 au sud de la France) ;
- enfin la supra-région *g* numérotée 16 regroupe les 8 régions *h*.

Ces trois niveaux pré-définis de regroupements possibles des régions *h* ont permis de programmer automatiquement les regroupements : pour un échantillon donné, le principe de ces regroupements automatiques est de passer à des niveaux de supra-régions de plus en plus agrégés jusqu'à ce qu'on trouve le niveau d'agrégation qui assure au moins 5 répondants dans le regroupement de régions obtenu.

On est ainsi en mesure de générer pour chaque échantillon une variable indiquant les regroupements de régions nécessaires pour cet échantillon-là, puis de calculer les biais et variances des estimateurs corrigés de la non-réponse dans les groupes ainsi formés.

2.1.3.5. Hypothèse d'un modèle de non-réponse correct

Autre particularité de notre adaptation de la méthodologie de Little, nous supposons une fois pour toutes que notre modélisation de la non-réponse au niveau des GRH (les croisements *ih*) est bonne, c'est-à-dire qu'on considère que les répondants et les non-répondants ont les mêmes comportements au sein d'un GRH donné.

Il faut à ce propos souligner ici que nous n'avons pas du tout cherché à construire le meilleur modèle de non-réponse possible sur les données retenues, comme on aurait pu le faire à partir d'un modèle LOGIT (Cf. paragraphes 1.1.2.2 et 1.1.2.3) : nous avons plutôt choisi d'emblée de nous baser sur un modèle de réponse proche de celui utilisé dans les EAE, ce qui nous a conduit aux GRH *ih*. Les résultats obtenus sont aussi liés à ce choix.

Dans ce cadre, on peut donc considérer, en reprenant les dénominations introduites par Little (Cf. paragraphe 1.2.2), que le terme LSB du biais est nul par construction et que le biais est égal au seul terme C, terme qui nous intéresse puisqu'on est dans le cas de petits échantillons. C'est ce que nous considérons dans toute la suite.

2.2. Les différents estimateurs retenus dans ce cadre particulier

Une dernière particularité majeure de notre travail par rapport à celui de Little, après celles qui viennent d'être citées, tient au fait qu'il a fallu réécrire entièrement toutes les formules des estimateurs, de leurs biais et de leurs variances, afin d'adapter celles proposées par Little au plan de sondage stratifié par activité et tranche d'effectif que nous avons retenu pour nous rapprocher de la réalité des EAE.

[5] Les GRH sont ici les croisements *ih*. On considère ainsi à ce niveau que les entreprises d'un croisement *ih* ont le même comportement de réponse. Si ces croisements *ih* ne contiennent pas suffisamment de répondants pour permettre la correction de la non-réponse, on est amené à les regrouper en croisements *ig* pour former des groupes de correction de la non-réponse. On considère en revanche que le modèle correct de non-réponse se situe au niveau *ih* et pas au niveau *ig* (Cf. 2.1.3.5) : les croisements *ig* ne sont donc pas des GRH au sens strict.

2.2.1. Les estimateurs d'une moyenne globale pour une activité donnée

Nous donnons ici les expressions des estimateurs de la moyenne globale pour une activité donnée, de leurs biais et variances dans le cadre particulier que nous retenons. Nous introduisons également des estimateurs que Little n'évoquait pas mais qui ont été présentés au paragraphe 1.1. **Ces estimations globales ne nous intéressent pas directement ici, mais l'obtention de leurs formules constitue une première étape pour parvenir aux estimations régionales.**

2.2.1.1. Expressions des estimateurs par activité retenus

Dans le cas des estimations globales pour une activité donnée, nous nous proposons de comparer cinq estimateurs de la moyenne globale. Les trois premiers correspondent à une correction de la non-réponse "pure" car basée uniquement sur les éléments fournis par l'échantillon. Les deux autres (numérotés 4 et 5) utilisent une information auxiliaire pour corriger la non-réponse.

Comme on l'a déjà souligné, dans le cas des estimations globales, l'estimation par imputation de la moyenne des GRH est équivalente à l'estimation par repondération au sein des GRH. Ces cinq estimateurs sont donc présentés ici dans l'optique de la repondération.

Comme le plan de sondage au niveau d'une activité donnée est stratifié selon la tranche d'effectif *i*, on a pour tous les estimateurs :

$$\hat{\bar{y}} = \sum_i \frac{N_i}{N} \bar{y}_i^* = \sum_i P_i \bar{y}_i^* .$$

Nous allons mettre en évidence une forme commune pour ces cinq estimateurs. Les calculs et raisonnements théoriques qui permettent d'aboutir à leurs formules sont donnés en Annexe 1. Nous obtenons en général deux expressions pour chaque estimateur, selon qu'on procède ou non à des regroupements de régions.

- ***Estimateur 1 - l'estimateur par la moyenne des répondants :***

Ce premier estimateur correspond à la moyenne simple sur les répondants, évoquée au paragraphe 1.1.1. On ne corrige alors pas la non-réponse et on considère simplement qu'on peut approximer la moyenne de l'échantillon par celle des seuls répondants, comme si répondants et non-répondants étaient identiques. Cet estimateur est équivalent au $\bar{y}_R$ de Little.

On obtient l'expression suivante de $\bar{y}_i^*$, qu'il y ait des regroupements de régions ou non :

$$\bar{y}_i^* = \sum_h \frac{n_{ihR}}{n_{iR}} \bar{y}_{ihR} .$$

- ***Estimateur 2 - l'estimateur de la moyenne ajustée par le taux de réponse :***

On se place pour ce deuxième estimateur dans la perspective de la recherche de groupes de réponse homogènes du paragraphe 1.1.2.3. Les GRH considérés ici sont les croisements *ih* quand ils contiennent suffisamment de répondants, ou (à défaut) les croisements *ig* après regroupements adéquats de régions dans le cas contraire. Cet estimateur est équivalent au $\bar{y}_A$ étudié par Little.

Dans le cas où aucun regroupement de régions n'est nécessaire, on obtient :

$$\bar{y}_i^* = \sum_h \frac{n_{ih}}{n_i} \bar{y}_{ihR} .$$

Si au contraire il faut procéder à des regroupements g de régions h, on obtient : $\bar{y}_i^* = \sum_g \sum_{h \in g} \frac{n_{ig}}{n_i} \frac{n_{ihR}}{n_{igR}} \bar{y}_{ihR}$.

- ***Estimateur 3 - l'estimateur de la moyenne ajustée sur les marges au sein de l'échantillon par le raking-ratio :***

L'estimateur 3 reprend l'idée évoquée au paragraphe 1.1.2.5 d'utiliser le calage sur marges pour corriger la non-réponse. Plus précisément, il s'agit ici de "caler" les répondants sur l'échantillon complet en termes de nombres d'entreprises : on utilise alors le logiciel CALMAR pour caler les n_{ihR} (ou n_{igR} en cas de regroupements) sur les comptages n_i et n_h (ou n_g) supposés connus pour l'échantillon complet. Ce calage sur marges fournit des $\tilde{n}_{ih}$ (ou des $\tilde{n}_{ig}$ en cas de regroupements), et on obtient :

$\bar{y}_i^* = \sum_g \sum_{h \in g} \frac{\tilde{n}_{ih}}{n_i} \bar{y}_{ihR}$ en l'absence de regroupements

et $\bar{y}_i^* = \sum_g \sum_{h \in g} \frac{\tilde{n}_{ig}}{n_i} \frac{n_{ihR}}{n_{igR}} \bar{y}_{ihR}$ dans le cas contraire.

On peut remarquer que les estimateurs 2 et 3 sont deux façons différentes d'utiliser une même information portant sur la taille de l'échantillon par tranche d'effectif et par région : pour l'estimateur 2, on utilise cette information au niveau de tous les croisements *ih* (ou *ig* s'il y a des regroupements), alors que pour l'estimateur 3 on l'utilise seulement pour les marges. Il peut sembler étrange de n'utiliser l'information que pour les marges si on en dispose pour tous les croisements *ih*. Toutefois, l'estimateur 3 peut gagner en stabilité par rapport à l'estimateur 2, même s'il utilise moins d'information que 2 : en effet, les croisements *ih* (ou *ig*) peuvent s'avérer être de très petite taille, ce qui donnerait à l'estimateur 2 un caractère plus instable que l'estimateur 3.

- ***Estimateur 4 - l'estimateur post-stratifié :***

L'estimateur 4 correspond à l'estimateur post-stratifié évoqué au paragraphe 1.1.2.4, pour lequel les post-strates sont les GRH *ih* (ou *ig*). Il est équivalent au $\bar{y}_S$ présenté par Little.

En l'absence de regroupements de régions, on a dans ce cas :
$\bar{y}_i^* = \sum_h P_{ih} \bar{y}_{ihR}$.

En utilisant des regroupements g de régions h, on obtient :
$\bar{y}_i^* = \sum_g \sum_{h \in g} P_{ig} \frac{n_{ihR}}{n_{igR}} \bar{y}_{ihR}$.

- ***Estimateur 5 - l'estimateur obtenu par calage sur marge :***

L'estimateur 5 reprend à nouveau l'idée évoquée au paragraphe 1.1.2.5 d'utiliser le calage sur marges pour corriger la non-réponse. Cette fois, il s'agit de "caler" les répondants sur la population totale, donc sur une information auxiliaire parfaite (et non réduite au seul échantillon comme dans le cas de l'estimateur 3). De plus, on peut montrer (Cf. Caron et Sautory, polycopié de cours de sondage proposé par le CEPE, 2003) qu'il est équivalent de caler l'échantillon de départ (au niveau des individus) sur les marges de la population ou de

caler sur ces mêmes marges "l'échantillon fictif" formé à partir des comptages d'individus (pondérés par leurs poids de départ) dans les croisements *ih* (ou *ig*).

Le calage consiste donc ici à caler les $\hat{N}_{ih}$ ou les $\hat{N}_{ig}$ (estimés à partir des n_{ihR} et n_{igR} pondérés par N_i/n_i) sur les N_i et les N_h ou N_g supposés connus sur la population de la base de sondage, pour obtenir grâce au logiciel CALMAR les $\widetilde{N}_{ih}$ ou les $\widetilde{N}_{ig}$.

En l'absence de regroupements de régions, on obtient ainsi :

$\bar{y}_i^* = \sum_h \widetilde{P}_{ih} \bar{y}_{ihR}$ où $\widetilde{P}_{ih} = \dfrac{\widetilde{N}_{ih}}{N_i}$.

Dans le cas des regroupements *g* de régions *h*, on a $\widetilde{P}_{ig} = \dfrac{\widetilde{N}_{ig}}{N_i}$ ce qui donne :

$\bar{y}_i^* = \sum_g \sum_{h \in g} \widetilde{P}_{ig} \dfrac{n_{ihR}}{n_{igR}} \bar{y}_{ihR}$.

De la même manière qu'on a noté une proximité entre les estimateurs 2 et 3, on peut noter que les estimateurs 4 et 5 correspondent à deux façons différentes d'utiliser une même information auxiliaire sur l'échantillon. Si l'estimateur 4 l'utilise dans tous les croisements *ih*, l'estimateur 5 ne l'utilise que pour les marges : à nouveau, l'estimateur 4 risque d'être plus instable que l'estimateur 5. Ceci serait surtout le cas si on croisait plus de deux variables, car les croisements deviendraient alors des micro-croisements (ici nous nous limitons aux croisements de deux variables).

- ***Forme commune de ces cinq estimateurs :***

On constate finalement que ces cinq estimateurs peuvent se mettre sous une même forme :

$\hat{\bar{y}} = \sum_i \dfrac{N_i}{N} \bar{y}_i^* = \sum_i P_i \bar{y}_i^*$

où $\bar{y}_i^* = \sum_h w_{ih} \bar{y}_{ihR}$ ou, en cas de regroupements, $\bar{y}_i^* = \sum_g \sum_{h \in g} w_{ih} \bar{y}_{ihR}$.

La présence d'une somme sur *h* dans cette expression commune ne doit pas faire oublier que, dans le cadre considéré, la région *h* n'est pas une véritable variable de stratification. Elle apparaît ici parce qu'elle intervient dans la façon de corriger la non-réponse pour certains estimateurs.

Le tableau suivant résume les formes du poids w_{ih} dans les différents cas :

Estimateur	Expression de w_{ih} en l'absence de regroupements	Expression de w_{ih} en présence de regroupements
1	$w_{ih} = \dfrac{n_{ihR}}{n_{iR}}$	aucun regroupement
2	$w_{ih} = \dfrac{n_{ih}}{n_i}$	$w_{ih} = \dfrac{n_{ig}}{n_i} \dfrac{n_{ihR}}{n_{igR}}$

3	$w_{ih} = \frac{\widetilde{n}_{ih}}{n_i}$	$w_{ih} = \frac{\widetilde{n}_{ig}}{n_i} \frac{n_{ihR}}{n_{igR}}$
4	$w_{ih} = P_{ih} = \frac{N_{ih}}{N_i}$	$w_{ih} = P_{ig} \frac{n_{ihR}}{n_{igR}}$
5	$w_{ih} = \widetilde{P}_{ih} = \frac{\widetilde{N}_{ih}}{N_i}$	$w_{ih} = \widetilde{P}_{ig} \frac{n_{ihR}}{n_{igR}}$

Cette forme générale nous permet en particulier de calculer leurs biais et leurs variances conditionnels de manière générique. On ne donne dans les deux paragraphes suivants que les expressions finales de ces biais et variances. Les calculs menant à ces expressions sont donnés en Annexe 1, et reposent sur l'hypothèse d'un taux de réponse constant quelles que soient l'activité, la tranche d'effectif ou la région.

2.2.1.2. Biais conditionnel de ces estimateurs

Dans tous les cas, on a :

$B(\hat{\bar{y}} / n, n_R) = \sum_i P_i B(\bar{y}_i^* / n, n_R)$ où n et n_R désignent respectivement le vecteur des tailles d'échantillon dans les croisements *ih* (les $\{n_{ih}\}$) et le vecteur des nombres de répondants dans les croisements *ih* (les $\{n_{ihR}\}$), avec :

$$B(\bar{y}_i^* / n, n_R) = \sum_h (w_{ih} - P_{ih}) \overline{Y}_{ihR} + \sum_h P_{ih} (\overline{Y}_{ihR} - \overline{Y}_{ih})$$

On retrouve comme dans l'article de Little la décomposition du biais en deux termes :

- dans le premier terme C, $w_{ih} - P_{ihR}$ tend vers 0 quand la taille de l'échantillon tend vers celle de la population, donc C (traduisant la variabilité liée à l'échantillonnage) tend logiquement vers 0 quand la taille de l'échantillon devient grande (ce n'est pas le cas ici) ;
- si on suppose que $\overline{Y}_{ihR} = \overline{Y}_{ih}$ (c'est-à-dire si on postule que les répondants et les non-répondants ont les mêmes comportements au niveau des croisements *ih* qui sont les GRH de base considérés ici, donc que notre modélisation de la non-réponse est correcte), le deuxième terme LSB est nul.

On supposera par la suite que $\overline{Y}_{ihR} = \overline{Y}_{ih}$ et donc le biais sera égal au seul premier terme C : $B(\bar{y}_i^* / n, n_R) = \sum_h (w_{ih} - P_{ih}) \overline{Y}_{ihR}$.

Dans le cas où on procède à des regroupements *g* de régions *h*, on a :

$$B(\bar{y}_i^* / n, n_R) = \sum_g \sum_{h \in g} (w_{ih} - P_{ih}) \overline{Y}_{ihR} .$$

2.2.1.3. Variance conditionnelle de ces estimateurs

On a dans tous les cas :

$V(\hat{\bar{y}}/n,n_R)=\sum_i P_i^2 V(\bar{y}_i^*/n,n_R)$, avec : $V(\bar{y}_i^*/n,n_R)=\sum_h w_{ih}^2 V(\bar{y}_{ihR})=\sum_h w_{ih}^2(1-\frac{n_{ihR}}{N_{ihR}})\frac{S_{ih}^2}{n_{ihR}}$.

En négligeant le terme de population finie (donc en faisant implicitement l'hypothèse d'un tirage avec remise comme on l'a fait déjà en utilisant la loi multinomiale dans les simulations des tailles d'échantillon), on obtient :

$V(\bar{y}_i^*/n,n_R)=\sum_h w_{ih}^2\frac{S_{ih}^2}{n_{ihR}}$,

et $V(\bar{y}_i^*/n,n_R)=\sum_g\sum_{h\in g} w_{ih}^2\frac{S_{ih}^2}{n_{ihR}}$ dans le cas où on procède à des regroupements *g* de régions *h*.

Le tableau suivant récapitule les estimateurs de la moyenne globale ainsi que leurs biais et variances dans le cas où le modèle de non-réponse est correct (i.e. où LSB est nul) :

	Forme générique de l'estimateur	Biais conditionnel	Variance conditionnelle
Sans regroupement	$\hat{\bar{y}}=\sum_i P_i\sum_h w_{ih}\bar{y}_{ihR}$	$B(\hat{\bar{y}}/n,n_R)=\sum_i P_i\sum_h (w_{ih}-P_{ih})\bar{Y}_{ihR}$	$V(\hat{\bar{y}}/n,n_R)=\sum_i P_i^2\sum_h w_{ih}^2\frac{S_{ih}^2}{n_{ihR}}$
Avec regroupements	$\hat{\bar{y}}=\sum_i P_i\sum_g\sum_{h\in g} w_{ih}\bar{y}_{ihR}$	$B(\hat{\bar{y}}/n,n_R)=\sum_i P_i\sum_g\sum_{h\in g}(w_{ih}-P_{ih})\bar{Y}_{ihR}$	$V(\hat{\bar{y}}/n,n_R)=\sum_i P_i^2\sum_g\sum_{h\in g} w_{ih}^2\frac{S_{ih}^2}{n_{ihR}}$

2.2.2. Les estimateurs d'un total par région pour une activité donnée

Dans le cas des estimations par région, on peut considérer deux estimateurs concurrents de la moyenne régionale : $\hat{\bar{Y}}_h=\frac{\hat{Y}_h}{N_h}$ ou $\hat{\bar{Y}}_h=\frac{\hat{Y}_h}{\hat{N}_h}$ (N_h étant connu dans la base de sondage).

Or comme la région n'est pas une véritable variable de stratification, qu'elle n'est qu'un domaine d'intérêt, on ne maîtrise pas réellement la taille de l'échantillon tiré dans le croisement *ih* (le n_{ih} est lié au hasard). L'estimateur $\hat{\bar{Y}}_h=\frac{\hat{Y}_h}{\hat{N}_h}$ sera donc plus stable que l'estimateur $\hat{\bar{Y}}_h=\frac{\hat{Y}_h}{N_h}$.

Nous avons choisi de nous intéresser ici simplement à l'estimation du total régional $\hat{Y}_h$.

Un estimateur concurrent qui aurait pu être étudié (mais ne l'a finalement pas été) est présenté en Annexe 2.

2.2.2.1. Expression des estimateurs par région

Ici à nouveau, on ne donne que les expressions des estimateurs retenus ; les raisonnements ayant mené à leur obtention sont présentés en Annexe 1.

- ***Cas des estimateurs par repondération :***

En repartant de l'expression générale des cinq estimateurs repondérés exposés dans le cas des estimateurs globaux, on obtient la formule suivante pour l'estimateur du total d'une région *h* :

$$\hat{Y}_h = \sum_i N_i w_{ih} \bar{y}_{ihR} .$$

En cas de regroupements, seuls les w_{ih} changent comme on l'a indiqué dans le tableau récapitulatif des expressions des estimateurs 1 à 5 (Cf. page 22).

- ***Cas de l'estimateur par imputation – estimateur 6 :***

On pratique ici une imputation déterministe en remplaçant les non-réponses par la moyenne du croisement considéré. On appellera par la suite cet estimateur l'estimateur 6.

Dans le cas où on ne procède à aucun regroupement de régions, on remplace les valeurs manquantes par la moyenne calculée sur les répondants dans le croisement *ih*. On a alors :

$$\hat{Y}_h = \sum_i N_i \frac{n_{ih}}{n_i} \bar{y}_{ihR} .$$

Dans le cas où on procède à des regroupements de régions, on remplace les valeurs manquantes dans le croisement *ih* par la moyenne calculée sur les répondants dans le croisement *ig*, d'où : $\hat{Y}_h = \sum_i N_i \frac{n_{ih}}{n_i} \bar{y}_{ihT}$ où $\bar{y}_{ihT} = \frac{n_{ihR}\bar{y}_{ihR} + (n_{ih} - n_{ihR})\bar{y}_{igR}}{n_{ih}}$.

Ceci peut aussi s'écrire sous la forme : $\frac{\hat{Y}_h}{N} = \sum_i P_i \frac{n_{ih}}{n_i} \bar{y}_{ihT} = \sum_i P_i \frac{n_{ig}}{n_i} \frac{n_{ih}}{n_{ig}} \bar{y}_{ihT}$.

On peut alors rapprocher cet estimateur de l'estimateur de la moyenne ajustée par repondération (l'estimateur 2 parmi les estimateurs globaux), dans lequel on aurait modifié les poids $\frac{n_{ihR}}{n_{igR}}$ en $\frac{n_{ih}}{n_{ig}}$ et remplacé $\bar{y}_{ihR}$ par $\bar{y}_{ihT}$.

2.2.2.2. Biais conditionnel de ces estimateurs régionaux

- ***Cas des estimateurs par repondération :***

On obtient dans ce cas (en considérant que $\bar{Y}_{ihR} = \bar{Y}_{ih}$) : $B(\hat{Y}_h / n, n_R) = \sum_i N_i (w_{ih} - P_{ih}) \bar{Y}_{ihR}$.

En cas de regroupements de régions *h*, seuls les w_{ih} changent, comme indiqué précédemment.

- ***Cas de l'estimateur par imputation :***

Dans le cas où il n'y a aucun regroupement, on retrouve : $B(\hat{Y}_h / n, n_R) = \sum_i N_i (p_{ih} - P_{ih}) \bar{Y}_{ihR}$.

Dans le cas où on est amené à regrouper des régions pour la correction de la non-réponse, l'expression du biais est plus complexe car $\bar{y}_{ihT}$ fait intervenir $\bar{y}_{ihR}$ et $\bar{y}_{igR}$ (Cf. Annexe 1).

2.2.2.3. Variance conditionnelle de ces estimateurs régionaux

- ***Cas des estimateurs par repondération :***

$$V(\hat{Y}_h / n, n_R) = \sum_i (N_i w_{ih})^2 V(\bar{y}_{ihR}) = \sum_i (N_i w_{ih})^2 (1 - \frac{n_{ihR}}{N_{ihR}}) \frac{S_{ih}^2}{n_{ihR}}$$

et en négligeant le terme de population finie on obtient : $V(\hat{Y}_h / n, n_R) = \sum_i (N_i w_{ih})^2 \frac{S_{ih}^2}{n_{ihR}}$.

- ***Cas de l'estimateur par imputation :***

En l'absence de regroupement de régions, on obtient en négligeant le terme de population finie : $V(\hat{Y}_h / n, n_R) = \sum_i \left(N_i \frac{n_{ih}}{n_i} \right)^2 \frac{S_{ih}^2}{n_{ihR}}$.

S'il est nécessaire de regrouper des régions, l'expression de la variance est également plus complexe (Cf. Annexe 1 à nouveau).

Le tableau suivant récapitule les estimateurs régionaux du total ainsi que leurs biais et variances conditionnels, dans le cas où on doit regrouper des régions *h* et toujours dans le cas où on suppose $\bar{Y}_{ihR} = \bar{Y}_{ih}$:

	Repondération	Imputation
Forme de l'estimateur $\hat{Y}_h$	$\sum_i N_i w_{ih} \bar{y}_{ihR}$	$\sum_i N_i \frac{n_{ih}}{n_i} \frac{n_{ihR} \bar{y}_{ihR} + (n_{ih} - n_{ihR}) \bar{y}_{igR}}{n_{ih}}$
Biais conditionnel $B(\hat{Y}_h / n, n_R)$	$\sum_i N_i (w_{ih} - P_{ih}) \bar{Y}_{ihR}$	$\sum_i N_i \left(\frac{n_{ih}}{n_i} - P_{ih} \right) \left(B_{ih} \bar{Y}_{ihR} + (1 - B_{ih}) \bar{Y}_{igR} \right) + \sum_i N_i P_{ih} \left(B_{ih} \bar{Y}_{ihR} + (1 - B_{ih}) \bar{Y}_{igR} - \bar{Y}_{ih} \right)$
Variance conditionnelle $V(\hat{Y}_h / n, n_R)$	$\sum_i (N_i w_{ih})^2 \frac{S_{ih}^2}{n_{ihR}}$	$\sum_i \left(N_i \frac{n_{ig}}{n_i} \frac{n_{ih}}{n_{ig}} \right)^2 \left[\left(\frac{n_{ihR}}{n_{ih}} \right)^2 \frac{S_{ih}^2}{n_{ihR}} + \left(\frac{n_{ih} - n_{ihR}}{n_{ih}} \right)^2 \frac{S_{ig}^2}{n_{igR}} + 2 \frac{n_{ihR} (n_{ih} - n_{ihR})}{n_{ih}^2} \frac{S_{ih}^2}{n_{igR}} \right]$

2.2.3. Les indicateurs calculés

Pour chacun des estimateurs retenus, nous avons donc calculé, pour chaque échantillon, son biais, son biais au carré, sa variance, la racine de son erreur quadratique moyenne et son "erreur quadratique moyenne *relative*"[6] par rapport à l'estimateur 1 pris ici comme référence : $REL(\hat{Y}_{hl}) = 100 * (RMSE(\hat{Y}_{hl}) / RMSE(\hat{Y}_{h1}) - 1)$ dans le cas des estimateurs régionaux, pour *l* = 1 à 6, 6 étant le numéro affecté à l'estimateur par imputation.
Nous avons également calculé le coefficient de variation de chaque estimateur : $CV(\hat{Y}_{hl}) = 100 * (RMSE(\hat{Y}_{hl}) / Y_h)$ dans le cas du total régional, pour *l* = 1 à 6,
le dénominateur $\bar{Y}$ et Y_h étant calculé grâce aux données tirées de SUSE qui ont également permis le calcul des $\bar{Y}_{ih}$ et des S_{ih}^2 nécessaires aux calculs des biais et variances (Cf. § 2.1.2.5).

[6] Au sens défini par Little, Cf. paragraphe 1.2.4.

L'intérêt du coefficient de variation par rapport à l'erreur quadratique moyenne *relative* est qu'il autorise les comparaisons entre résultats obtenus par l'une ou l'autre méthode de tirage a) ou b), tandis que le dénominateur de l'erreur quadratique moyenne *relative* dépend de la méthode de tirage utilisée pour simuler les échantillons sur lesquels on calcule $RMSE(\hat{\bar{y}}_1)$.

Dans tous les cas, la moyenne de ces indicateurs sur les 100 échantillons tirés permet d'approximer leur espérance[7]. Pour les estimateurs régionaux, on présentera seulement la moyenne des 8 indicateurs obtenus pour les 8 régions considérées.

2.3. Les principaux résultats obtenus

Comme cela a déjà été souligné précédemment, c'est la comparaison des estimateurs régionaux qui nous intéresse dans cette étude. Nos simulations ont donc porté sur les estimateurs par région et non sur les estimateurs d'une moyenne globale.

L'ensemble des résultats obtenus est présenté dans les tableaux 2 à 5 en Annexe 4.

Pour des raisons pratiques (explicitées en Annexe 4) liées à la réalisation des calages sur marges, les estimateurs 3 et 5 n'ont finalement pas pu être calculés.

La plupart des enseignements que l'on peut tirer des résultats obtenus peuvent être lus à partir du tableau de synthèse 1 qui se trouve à la page 28. Ce tableau présente les moyennes des résultats obtenus pour les 4 jeux de données, la principale différence entre les résultats de ces jeux de données étant que les effets constatés sont de plus en plus importants quand on passe du jeu de données A, au B puis au C et enfin au D, au fur et à mesure que les nombres de répondants dans les croisements *ih* diminuent.

2.3.1. Résultats sur les estimateurs régionaux : la correction de la non-réponse joue un rôle important

Concernant le tableau 1 sur les estimateurs par région, la correction de la non-réponse a un effet positif, surtout dans les cas de l'estimateur par imputation et de l'estimateur post-stratifié

Le tableau 1 des résultats des estimateurs par région compte une majorité des chiffres négatifs : ceci signifie que les estimateurs conçus explicitement pour corriger la non-réponse ont des erreurs quadratiques moyennes inférieures à celle de la simple moyenne des répondants (qui ne fait rien pour la corriger). Il est en effet normal que la correction de la non-réponse constitue une amélioration dans le cadre des estimateurs régionaux, puisque cette correction de la non-réponse est basée sur des GRH liés aux régions (ce serait différent si la correction de la non-réponse se faisait sur un modèle de non-réponse fondé sur l'ancienneté des entreprises par exemple).

Si on s'en tient aux erreurs quadratiques moyennes relatives tous critères confondus, en bas à droite du tableau 1, on constate que celle de l'estimateur 2 est de -3,67 (ce qui signifie que la racine de son erreur quadratique moyenne vaut 96,33 % de celle de l'estimateur 1) ; celle de l'estimateur post-stratifié 4 vaut -8,57 (soit une racine d'erreur quadratique moyenne valant 91,43 % de celle de l'estimateur 1), et celle de l'estimateur par imputation 6 vaut -12,42 (soit une racine d'erreur quadratique moyenne valant 87,58 % de celle de l'estimateur 1).

[7] On peut par ailleurs considérer (dans une optique à la Monte-Carlo) que la moyenne des biais et variances conditionnels obtenue sur l'ensemble de nos simulations est une approximation des biais et variances non conditionnels.

Des différences importantes selon la méthode de simulation des échantillons utilisée

Les estimateurs post-stratifié (4) et par imputation (6) sont donc ceux qui donnent les meilleurs résultats. Leurs performances, légèrement en faveur de l'estimateur 6 globalement, sont en fait comparables : en effet, l'estimateur 6 donne des meilleurs résultats que l'estimateur 4 pour les échantillons simulés par la méthode b) (-18,40 pour 6 contre -7,23 pour 4 en termes d'erreur quadratique moyenne relative), mais c'est le contraire pour les échantillons simulés par la méthode a) (-6,44 pour 6 contre -9,92 pour 4 en termes d'erreur quadratique moyenne relative).

Tableau 1 : Erreurs quadratiques moyennes relatives des estimateurs par région

			méthode de tirage TAS (a)				méthode de tirage systématique avec pas (b)				(a)	(b)	(a) et (b) confondues
estimateur	activité	jeu de données	variable E	variable CA	variable MA	variable VA	variable E	variable CA	variable MA	variable VA	toutes variables	toutes variables	toutes variables
est_reg1	151F	moyenne A-D	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
est_reg2	151F	moyenne A-D	1,64	0,58	-0,08	0,67	-1,98	-2,67	-1,62	-2,61	0,70	-2,22	-0,76
est_reg4	151F	moyenne A-D	0,85	-1,78	-2,19	-1,52	-2,01	-2,70	-1,65	-2,63	-1,16	-2,25	-1,70
est_reg6	151F	moyenne A-D	-7,45	-11,30	-13,78	-10,56	-30,18	-27,40	-23,20	-27,65	-10,77	-27,11	-18,94
est_reg1	158C	moyenne A-D	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
est_reg2	158C	moyenne A-D	2,65	0,60	-0,15	0,40	-9,06	-9,09	-1,10	-8,36	0,88	-6,90	-3,01
est_reg4	158C	moyenne A-D	-7,39	-10,78	-2,27	-10,12	-10,55	-10,19	-1,21	-9,38	-7,64	-7,83	-7,74
est_reg6	158C	moyenne A-D	-2,58	-5,46	-7,95	-5,81	-28,33	-24,40	-9,80	-23,00	-5,45	-21,38	-13,42
est_reg1	524C	moyenne A-D	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
est_reg2	524C	moyenne A-D	-2,43	-1,96	-1,85	-1,83	-18,90	-7,86	-7,25	-7,57	-2,02	-10,39	-6,21
est_reg4	524C	moyenne A-D	-29,62	-14,50	-13,45	-13,77	-19,34	-8,01	-7,40	-7,73	-17,83	-10,62	-14,23
est_reg6	524C	moyenne A-D	-3,79	-3,21	-2,88	-1,31	-20,79	-8,90	-8,05	-6,95	-2,80	-11,17	-6,99
est_reg1	524R	moyenne A-D	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
est_reg2	524R	moyenne A-D	-0,96	-1,93	-1,56	-2,00	-13,32	-7,52	-4,66	-8,47	-1,61	-8,49	-5,05
est_reg4	524R	moyenne A-D	-19,53	-13,11	-8,95	-14,80	-13,93	-7,76	-4,83	-8,78	-14,10	-8,83	-11,46
est_reg6	524R	moyenne A-D	-4,31	-6,85	-6,20	-6,64	-20,00	-14,71	-10,61	-15,66	-6,00	-15,24	-10,62
est_reg1	524X	moyenne A-D	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
est_reg2	524X	moyenne A-D	0,22	-0,48	-0,48	-0,58	-9,54	-4,46	-5,21	-5,86	-0,33	-6,27	-3,30
est_reg4	524X	moyenne A-D	-12,15	-6,91	-8,10	-8,29	-10,09	-4,66	-5,50	-6,16	-8,86	-6,60	-7,73
est_reg6	524X	moyenne A-D	-4,92	-8,21	-7,96	-7,67	-20,96	-15,21	-15,95	-16,26	-7,19	-17,10	-12,14
est_reg1	toutes APE	moyenne A-D	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
est_reg2	toutes APE	moyenne A-D	0,22	-0,64	-0,83	-0,67	-10,56	-6,32	-3,97	-6,57	-0,48	-6,86	-3,67
est_reg4	toutes APE	moyenne A-D	-13,57	-9,42	-6,99	-9,70	-11,18	-6,66	-4,12	-6,94	-9,92	-7,23	-8,57
est_reg6	toutes APE	moyenne A-D	-4,61	-7,01	-7,75	-6,40	-24,05	-18,12	-13,52	-17,90	-6,44	-18,40	-12,42

Il y a donc des différences fortes entre les résultats obtenus selon l'une ou l'autre méthode de simulation des échantillons. Ceci est peut-être lié au fait que les tailles d'échantillon dans les croisements activité*tranche d'effectif*région sont totalement aléatoires dans le cas de la méthode a), alors que la méthode b) s'apparente davantage à un tirage stratifié à la fois par tranche d'effectif et par région. Ainsi, l'effet stabilisant de la post-stratification pourrait être plus fort dans le cas de la méthode a) que dans la méthode b), ce qui donnerait au final un avantage à l'estimateur 4 sur l'estimateur 6 dans le cas de la méthode a) et pas dans le cas de la méthode b).

L'activité 151F (les charcuteries) est particulière en ce qui concerne la comparaison des estimateurs 4 et 6, de même en fait que pour l'ensemble des résultats régionaux. En effet, dans le cas de cette activité, l'estimateur 6 l'emporte toujours largement sur tous les autres estimateurs régionaux. La particularité de cette activité est peut-être liée au fait qu'elle a l'échantillon le plus réduit des cinq activités retenues (moins de 300 entreprises interrogées contre plus de 450 pour les quatre autres activités comme cela est notamment rappelé dans les tableaux 2 et 3 présentés en Annexe 4). Ceci pourrait également être lié au fait que c'est pour cette activité que les moyennes de la population sont les plus dispersées (tant au niveau des moyennes par tranche d'effectif qu'à celui des moyennes par région au sein d'une tranche d'effectif, comme on peut le constater dans les tableaux 6 et 7 en Annexe 4).

Le terme de biais au carré domine ici le terme de variance

Si on rentre dans le détail des tableaux 2 à 5 présentés en Annexe 4, on constate que le biais au carré est souvent supérieur à la variance et est ainsi le terme prépondérant dans l'ereur quadratique moyenne. Il existe de plus des différences importantes entre les biais au carré des divers estimateurs, qui déterminent souvent les positionnements relatifs de leurs erreurs quadratiques moyennes. En ce sens, on peut dire que c'est le biais au carré qui est déterminant pour les estimateurs régionaux.

Ces mêmes résultats se retrouvent dans les jeux de données A, B, C ou D, les biais au carré et variances augmentant progressivement d'un jeu de données à l'autre, au fur et à mesure que le nombre de répondants diminue.

L'explication de l'importance du biais ici est certainement liée au fait qu'on contrôle très peu la taille de l'échantillon tiré pour le domaine d'intérêt, qui ne coïncide pas avec la stratification retenue pour le plan de sondage, surtout dans la méthode de simulation a). De ce fait, le terme w_{ih} est moins stable et estime moins bien P_{ih}, la vraie proportion que représente le croisement *ih* dans la population de la tranche d'effectif *i*. La différence de ces deux termes intervenant dans l'expression du biais (Cf. paragraphe 2.2.2.2), celui-ci est donc important. L'instabilité plus grande de w_{ih} dans le cas des estimations bâties selon la méthode de simulation a) fournit une explication supplémentaire aux moins bons résultats obtenus avec cette méthode par rapport à ceux de la méthode b).

2.3.2. Bilan des résultats obtenus et parallèle avec les résultats de Little

Il est finalement difficile de conclure, au vu de ces résultats, en faveur soit de la repondération soit de l'imputation dans le contexte de l'estimation sur des domaines d'intérêt dans un cadre proche de celui des EAE. La correction de la non-réponse par imputation semble avoir, tous critères confondus, l'avantage sur la correction par repondération, ce qui est rassurant dans la mesure où à l'heure actuelle les EAE utilisent essentiellement la correction de la non-réponse par imputation.

Toutefois, si l'estimateur par imputation domine largement dans certaines configurations, dans d'autres l'estimateur post-stratifié lui est équivalent. Suite à cette étude, on ne peut savoir avec certitude si cela est particulier aux activités retenues, ou s'il en irait de même pour la majorité des autres activités du champ des EAE. En tous cas, nos résultats restent très liés au contexte des EAE (de petits échantillons en particulier), aux paramètres retenus et aux hypothèses faites, sur le modèle de non-réponse notamment. Il faudrait faire varier davantage ces dimensions pour avoir une vision d'ensemble plus large des performances relatives de l'imputation et de la repondération. Une grande partie de la durée de cette étude ayant dû être consacrée à la mise en place des instruments permettant d'évaluer ces performances relatives, nous n'avons finalement pu réaliser qu'un nombre limité de simulations. Mais ces instruments étant désormais construits, ce travail pourra être éventuellement prolongé à l'avenir.

On peut pour finir rapprocher brièvement nos résultats de ceux obtenus par Little dans un cadre fictif.

Pour les estimateurs sur domaine d'intérêt, Little montre que l'estimateur par imputation donne une erreur quadratique moyenne inférieure à celle de l'estimateur ajusté par repondération seulement dans le cas où les moyennes de la population dans les croisements *cj* pour un GRH *c* sont assez proches les unes des autres. Dans notre cas, l'estimation par imputation est souvent meilleure : peut-être avons-nous sélectionné des activités pour lesquelles les moyennes de la population sont assez peu dispersées entre les différentes régions *h* d'une même tranche d'effectif *i*. Pour le vérifier, il faudrait être en mesure de comparer ces moyennes (dont on donne le minimum et le maximum dans le tableau 7 en Annexe 4) à celles d'autres activités.

Conclusion

L'objectif de cette étude était d'adapter la méthodologie proposée par Little pour comparer la correction de la non-réponse par repondération et par imputation à un cadre particulier et plus réaliste que le sien, fortement inspiré de celui des EAE. Basées sur cette méthodologie, les simulations que nous avons réalisées permettent de comparer différents estimateurs corrigés de la non-réponse pour des estimations sur domaines d'intérêt (les régions).

Il apparaît que, dans le contexte particulier retenu, la correction de la non-réponse est bénéfique pour les estimateurs régionaux et donne l'avantage à l'estimateur corrigé par imputation et à l'estimateur repondéré par post-stratification.

Si certains de ces résultats sont cohérents avec ceux obtenus par Little ou les confirment, ils ne permettent en aucun cas de tirer des conclusions générales sur la correction de la non-réponse : en effet, ces résultats sont indissociables du contexte spécifique et des hypothèses sur lesquels ils reposent. En particulier, le cadre retenu se caractérise par des échantillons de petite taille dans les croisements les plus fins de la stratification et par l'hypothèse d'un modèle de non-réponse correct (bâti sur des GRH assimilés aux croisements tranche d'effectif*région).

Au terme de cette étude, il n'est donc pas réellement possible de conclure en faveur de l'imputation ou de la repondération pour corriger la non-réponse dans les EAE. Tout au plus peut-on noter que l'avantage attribué à l'imputation dans le cas des échantillons simulés selon la méthode b), la plus proche du véritable plan de sondage des EAE, semble être un bon signal en faveur des EAE.

De toutes les façons, des simulations pour d'autres configurations de paramètres semblent nécessaires car nos résultats posent plus de questions qu'ils n'en résolvent. Ils donnent en effet un certain nombre de voies de recherche qu'il faudrait suivre pour être en mesure d'avoir une vision d'ensemble plus complète sur la comparaison de la correction de la non-réponse par repondération ou par imputation pour les EAE : il faudrait par exemple étudier des activités différentes, tenter de faire varier le taux de réponse selon les strates, multiplier les tailles d'échantillon testées, etc.

En conclusion, soulignons que l'un des buts premiers de cette étude était de poser les bases permettant de procéder à la comparaison de ces deux méthodes de correction de la non-réponse selon la méthodologie de simulation proposée par Little. Ce travail a permis l'établissement des formules de biais et variance adaptées au cadre particulier des EAE et l'obtention des programmes permettant de réaliser les simulations d'échantillons, les regroupements automatiques, les calages sur marges et les estimations. Ceci étant fait, tout est prêt pour poursuivre l'étude en faisant varier davantage de paramètres pour pouvoir affiner les résultats : un prolongement de ce travail est donc à envisager.

Bibliographie

Ouvrages généraux :

ARDILLY P., *Les techniques de sondage*, Editions Technip, 1994

CARON N., *Séminaire de sondages à l'ENSAE*, polycopié du séminaire, 2003

COCHRAN W.G., *Sampling Techniques* (3rd edition), Wiley, 1977

SÄRNDAL C.E., SWENSSON B., WRETMAN J., *Model Assisted Survey Sampling*, Springer, 1993

SAUTORY O., Polycopié du cours de deuxième année à l'ENSAE, 2003

TILLÉ Y., *Théorie des sondages*, Polycopié du cours de deuxième année à l'ENSAI, 2000

Articles spécifiques :

CARON N., *Les principales techniques de correction de la non-réponse et les modèles associés*, Document de travail de méthodologie statistique de l'INSEE n°9604, 1996

CARON N., *Estimation sur petits domaines*, Note n°051/E210 de la Direction des Statistiques d'Entreprises de l'INSEE, 2002

CARON N., *Le traitement de la non-réponse*, Lettre du SSE, INSEE, 1er trimestre 2003

CARON N., SAUTORY O., Polycopié de cours de sondage proposé par le CEPE, 2003

CHAPMAN D. W., *Remplacer ou ne pas remplacer - Voilà la question*, Le statisticien d'enquêtes, juillet 2003

FORD B. L., *An Overview of Hot-Deck Procedures*, Incomplete Data in Sample Surveys, volume 2, Academic Press, 1983

HAZIZA D., *Inférence en présence d'imputation : un survol*, article présenté aux JMS, 2002

OH H.L., SCHEUREN F.J., *Weighting Adjustment for Unit Nonresponse*, Incomplete Data in Sample Surveys, volume 2, Academic Press, 1983

RANCOURT E. & co., *Le bulletin d'imputation*, volumes 2 et 3, Statistique Canada, 2002 et 2003

Article central :

LITTLE R.J.A., *Survey Nonresponse Adjustments for Estimates of Means*, International Statistical Review, 1986

Articles divers :

CARON N., FUENTES B., *Estimation de précision des estimateurs de l'EAE-Commerce*, Document de travail de la Direction des Statistiques d'Entreprises de l'INSEE n°E2002/03, 2002

COTTON F., HESSE C., *Tirages coordonnés d'échantillons*, Document de travail de la Direction des Statistiques d'Entreprises de l'INSEE n°E9206, 1992

DEVILLE J.-C., SÄRNDAL C.-E., SAUTORY O., *Generalized Raking Procedures in Survey Sampling*, Journal of the American Statistical Association, 1993

LU H., GELMAN A., *A Method for Estimating Design-based Sampling Variances for Surveys with Weighting, Poststratification and Raking*, Journal of Official Statistics, Volume 19, 2003

SAUTORY O., *La macro CALMAR, Redressement d'un échantillon par calage sur marges*, Document de travail de la Direction des Statistiques Démographiques et Sociales de l'INSEE n°F9310, 1993

Glossaire

Nous rappelons ici les principales notations utilisées dans les formules et les calculs :

- Y désigne la variable d'intérêt considérée
- N est la taille de la population considérée
- n est la taille de l'échantillon
- R est l'ensemble des répondants à l'enquête
- k est l'indice des individus de la population
- c_k est la probabilité de réponse à l'enquête de l'individu k
- π_k est la probabilité d'inclusion de l'individu k dans l'échantillon
- c est l'indice des GRH (dans la partie 1.1)
- n_c la taille de l'échantillon tiré dans le GRH c
- n_R est le nombre de répondants à l'enquête
- n_{cR} est le nombre de répondants dans le GRH c
- N_c est la taille de la population dans le GRH c
- N_{cj} est la taille de la population dans le croisement cj
- $\bar{y}_{cR}$ est la moyenne de la variable d'intérêt Y dans le GRH c estimée sur les répondants
- i est l'indice de la tranche d'effectif salarié
- h est l'indice de la région (nomenclature des ZEAT à 8 postes)
- g est l'indice de la supra-région, quand il est nécessaire de regrouper des régions h
- N_i est la taille de la population dans la tranche d'effectif i
- N_h est la taille de la population dans la région h
- N_{ih} est la taille de la population dans le croisement ih
- n_i est la taille de l'échantillon dans la tranche d'effectif i
- n_h est la taille de l'échantillon dans la région h
- n_{ih} est la taille de l'échantillon dans le croisement ih
- n_{ihR} est le nombre de répondants dans le croisement ih
- B_{ih} désigne le taux de réponse dans le croisement ih (on reprend ici une notation de Little)
- $\overline{Y}_i$ est la moyenne de la variable d'intérêt Y dans la tranche d'effectif i dans la population
- $\bar{y}_i^*$ est l'estimateur d'Horvitz-Thompson de $\overline{Y}_i$ corrigé de la non-réponse
- $P_i = \frac{N_i}{N}$ est la proportion de la tranche d'effectif i dans la population
- $P_{ih} = \frac{N_{ih}}{N_i}$ est la proportion du croisement ih dans la population de la tranche d'effectif i
- $P_{ihR} = \frac{N_{ihR}}{N_{iR}}$ est la proportion des répondants du croisement ih dans l'ensemble des répondants de la tranche d'effectif i

Annexes

Annexe 1 : Calculs des estimateurs retenus, de leurs biais et variances

Ces calculs utilisent de manière constante l'élément fondamental suivant : on suppose que le taux de réponse est constant pour une activité donnée, quelle que soit la tranche d'effectif ou la région considérée, i.e. :

$$B_{ih} = B_i = const. = B \Leftrightarrow \frac{N_{ihR}}{N_{ih}} = \frac{N_{iR}}{N_i} \Leftrightarrow \frac{N_{ihR}}{N_{iR}} = \frac{N_{ih}}{N_i} \Leftrightarrow P_{ihR} = P_{ih}$$

Au passage, les grandeurs P_{ihR}, P_{ih}, et P_i sont définies par les rapports suivants :

$$P_{ihR} = \frac{N_{ihR}}{N_{iR}},\ P_{ih} = \frac{N_{ih}}{N_i} \text{ et } P_i = \frac{N_i}{N}.$$

Les biais et variances calculés ici sont des biais et variances conditionnels aux vecteurs $\{n_{ih}\}$ et $\{n_{ihR}\}$ (les vecteurs des tailles d'échantillon et des nombres de répondants dans les croisements les plus fins de notre nomenclature).

Cas des estimateurs de la moyenne globale

Ces estimations globales ne nous intéressent pas directement ici, mais l'obtention de leurs formules constitue une première étape pour parvenir aux estimations régionales.

Expressions des estimateurs retenus

Comme le plan de sondage au niveau d'une activité donnée est stratifié selon la tranche d'effectif *i*, on a pour tous les estimateurs :

$\hat{\bar{y}} = \sum_i \frac{N_i}{N} \bar{y}_i^* = \sum_i P_i \bar{y}_i^*$ où $\bar{y}_i^*$ est l'estimateur d'Horvitz-Thompson corrigé de la non-réponse.

- ***Estimateur 1 - l'estimateur par la moyenne des répondants :***

On a : $\bar{y}_i^* = \sum_{k \in i \cap R} \frac{y_k}{n_{iR}} = \sum_h \sum_{k \in ih \cap R} \frac{y_k}{n_{iR}}$.

Si on recherche une écriture unifiée avec les autres estimateurs qui apparaîtront par la suite, on obtient : $\bar{y}_i^* = \sum_h \frac{n_{ihR}}{n_{iR}} \bar{y}_{ihR}$ avec $n_{ihR}\bar{y}_{ihR} = 0$ si $n_{ihR} = 0$.

- ***Estimateur 2 - l'estimateur de la moyenne ajustée par le taux de réponse :***

On a, dans le cas où aucun regroupement de régions n'est nécessaire :

$$\bar{y}_i^* = \frac{1}{N_i} \sum_{k \in i \cap R} \frac{y_k}{\pi_k \times tauxrep} = \frac{1}{N_i} \sum_{k \in i \cap R} \frac{y_k}{\frac{n_i}{N_i}\frac{n_{ihR}}{n_{ih}}} = \frac{1}{N_i} \sum_h \frac{\sum_{k \in ih \cap R} y_k}{\frac{n_i}{N_i}\frac{n_{ihR}}{n_{ih}}} = \frac{1}{N_i} \sum_h \frac{n_{ihR}\bar{y}_{ihR}}{\frac{n_i}{N_i}\frac{n_{ihR}}{n_{ih}}} = \sum_h \frac{n_{ih}}{n_i} \bar{y}_{ihR}.$$

Si au contraire il faut procéder à des regroupements *g* de régions *h*, on obtient :

$$\bar{y}_i^* = \frac{1}{N_i}\sum_{k\in i\cap R}\frac{y_k}{\pi_k \times tauxrep} = \frac{1}{N_i}\sum_g\sum_{h\in g}\sum_{k\in i\cap R}\frac{y_k}{\pi_k \times tauxrep} = \frac{1}{N_i}\sum_g\sum_{h\in g}\frac{\sum_{k\in ih\cap R} y_k}{\frac{n_i}{N_i}\frac{n_{igR}}{n_{ig}}} = \frac{1}{N_i}\sum_g\sum_{h\in g}\frac{n_{ihR}\bar{y}_{ihR}}{\frac{n_i}{N_i}\frac{n_{igR}}{n_{ig}}} = \sum_g\sum_{h\in g}\frac{n_{ig}}{n_i}\frac{n_{ihR}}{n_{igR}}\bar{y}_{ihR}$$

- ***Estimateur 3 - l'estimateur de la moyenne ajustée sur les marges au sein de l'échantillon par le raking-ratio :***

On a en l'absence de regroupements de régions :

$\bar{y}_i^* = \sum_g\sum_{h\in g}\frac{\tilde{n}_{ih}}{n_i}\bar{y}_{ihR}$, où $\tilde{n}_{ih}$ est obtenu par raking-ratio avec le logiciel CALMAR.

Dans le cas où il faut regrouper des régions *h* en supra-régions *g*, on a :

$$\bar{y}_i^* = \sum_g\sum_{h\in g}\frac{\tilde{n}_{ig}}{n_i}\frac{n_{ihR}}{n_{igR}}\bar{y}_{ihR}.$$

- ***Estimateur 4 - l'estimateur post-stratifié :***

En l'absence de regroupements de régions, on a :
$\bar{y}_i^* = \sum_h P_{ih}\bar{y}_{ihR}$.

En utilisant des regroupements *g* de régions *h*, on obtient :

$\bar{y}_i^* = \sum_g \frac{N_{ig}}{N_i}\bar{y}_{igR}$ où $\bar{y}_{igR} = \frac{1}{n_{igR}}\sum_{h\in g} n_{ihR}\bar{y}_{ihR}$,

d'où : $\bar{y}_i^* = \sum_g\sum_{h\in g} P_{ig}\frac{n_{ihR}}{n_{igR}}\bar{y}_{ihR}$.

- ***Estimateur 5 - l'estimateur obtenu par calage sur marge :***

En l'absence de regroupements de régions, on obtient :

$\bar{y}_i^* = \sum_h \tilde{P}_{ih}\bar{y}_{ihR}$ où $\tilde{P}_{ih} = \frac{\tilde{N}_{ih}}{N_i}$, avec $\tilde{N}_{ih}$ obtenu avec le logiciel CALMAR.

Dans le cas des regroupements *g* de régions *h*, on a $\tilde{P}_{ig} = \frac{\tilde{N}_{ig}}{N_i}$ ce qui donne :

$\bar{y}_i^* = \sum_g\sum_{h\in g}\tilde{P}_{ig}\frac{n_{ihR}}{n_{igR}}\bar{y}_{ihR}$.

Biais

Dans tous les cas, on a :

$B(\hat{\bar{y}}/n, n_R) = \sum_i P_i B(\bar{y}_i^*/n, n_R)$

avec :

$$B(\bar{y}_i^* / n, n_R) = E(\bar{y}_i^*) - \bar{Y}_i$$
$$= \sum_h w_{ih} E(\bar{y}_{ihR}) - \bar{Y}_{iR} + \bar{Y}_{iR} - \bar{Y}_i$$
$$= \sum_h (w_{ih} - P_{ihR}) \bar{Y}_{ihR} + \bar{Y}_{iR} - \bar{Y}_i$$
$$= \sum_h (w_{ih} - P_{ihR}) \bar{Y}_{ihR} + \sum_h (P_{ihR} \bar{Y}_{ihR} - P_{ih} \bar{Y}_{ih})$$
$$= \sum_h (w_{ih} - P_{ih}) \bar{Y}_{ihR} + \sum_h P_{ih} (\bar{Y}_{ihR} - \bar{Y}_{ih})$$

Dans le cas de regroupements g de régions h :

$$B(\bar{y}_i^* / n, n_R) = E(\bar{y}_i^*) - \bar{Y}_i$$
$$= \sum_g \sum_{h \in g} w_{ih} E(\bar{y}_{ihR}) - \bar{Y}_{iR} + \bar{Y}_{iR} - \bar{Y}_i$$
$$= \sum_g \sum_{h \in g} (w_{ih} - P_{ih}) \bar{Y}_{ihR} + \sum_g \sum_{h \in g} P_{ih} (\bar{Y}_{ihR} - \bar{Y}_{ih})$$

Variance

Il n'y a pas de calcul particulier.

Cas des estimateurs par région

Expressions des estimateurs retenus

- ***Cas des estimateurs par repondération :***

En repartant de l'expression générale des cinq estimateurs repondérés exposés dans le cas des estimateurs globaux, on obtient la formule suivante pour l'estimateur du total d'une région h :

$\hat{Y}_h = \sum_i N_i \bar{z}_i$ où $z_k = y_k$ si $k \in h$ et $z_k = 0$ sinon, d'où : $\bar{z}_i = w_{ih} \bar{y}_{ihR}$

et donc : $\hat{Y}_h = \sum_i N_i w_{ih} \bar{y}_{ihR}$.

En cas de regroupements, seuls les w_{ih} changent comme on l'a indiqué dans le tableau récapitulatif des expressions des estimateurs 1 à 5 (Cf. page 22).

- ***Cas de l'estimateur par imputation – estimateur 6 :***

On pratique ici une imputation déterministe en remplaçant les non-réponses par la moyenne du croisement considéré.

Dans le cas où on ne procède à aucun regroupement de régions, on remplace les valeurs manquantes dans les croisements *ih* par la moyenne calculée sur les répondants dans le croisement *ih*. La moyenne de la variable d'intérêt *Y* dans le croisement *ih* après correction de la non-réponse par imputation est donc logiquement :

$\bar{y}_{ihT} = \dfrac{n_{ihR}\bar{y}_{ihR} + (n_{ih} - n_{ihR})\bar{y}_{ihR}}{n_{ih}} = \bar{y}_{ihR}$ (où l'indice T désigne le fait que cette grandeur est calculée sur l'ensemble des données réelles et des données imputées).

On a alors : $\hat{Y}_h = \sum_i N_i \dfrac{n_{ih}}{n_i} \bar{y}_{ihR}$.

Dans le cas où on procède à des regroupements de régions, on remplace les valeurs manquantes par la moyenne calculée sur les répondants dans le croisement *ig*, d'où :

$$\bar{y}_{ihT} = \frac{n_{ihR}\bar{y}_{ihR} + (n_{ih} - n_{ihR})\bar{y}_{igR}}{n_{ih}}.$$

On obtient ainsi : $\hat{Y}_h = \sum_i N_i \dfrac{n_{ih}}{n_i} \bar{y}_{ihT}$.

Biais

- Repondération :

$$B(\hat{Y}_h / n, n_R) = E(\hat{Y}_h) - Y_h$$
$$= \sum_i N_i w_{ih} E(\bar{y}_{ihR}) - \sum_i N_{ih} \bar{Y}_{ih}$$
$$= \sum_i N_i w_{ih} \bar{Y}_{ihR} - \sum_i N_i P_{ihR} \bar{Y}_{ihR} + \sum_i N_i P_{ihR} \bar{Y}_{ihR} - \sum_i N_{ih} \bar{Y}_{ih}$$
$$= \sum_i N_i (w_{ih} - P_{ihR}) \bar{Y}_{ihR} + \sum_i \left(N_i P_{ihR} \bar{Y}_{ihR} - N_i P_{ih} \bar{Y}_{ih} \right)$$
$$= \sum_i N_i (w_{ih} - P_{ih}) \bar{Y}_{ihR} + \sum_i N_i P_{ih} \left(\bar{Y}_{ihR} - \bar{Y}_{ih} \right)$$

- Imputation :

Dans le cas où il n'y a aucun regroupement, on retrouve comme précédemment :

$$B(\hat{Y}_h / n, n_R) = E(\hat{Y}_h) - Y_h$$
$$= \sum_i N_i \frac{n_{ih}}{n_i} E(\bar{y}_{ihR}) - \sum_i N_{ih} \bar{Y}_{ih}$$
$$= \sum_i N_i \frac{n_{ih}}{n_i} \bar{Y}_{ihR} - \sum_i N_i P_{ih} \bar{Y}_{ihR} + \sum_i N_i P_{ih} \bar{Y}_{ihR} - \sum_i N_{ih} \bar{Y}_{ih}$$
$$= \sum_i N_i \left(\frac{n_{ih}}{n_i} - P_{ih} \right) \bar{Y}_{ihR} + \sum_i \left(N_i P_{ih} \bar{Y}_{ihR} - N_i P_{ih} \bar{Y}_{ih} \right)$$
$$= \sum_i N_i (p_{ih} - P_{ih}) \bar{Y}_{ihR} + \sum_i N_i P_{ih} \left(\bar{Y}_{ihR} - \bar{Y}_{ih} \right)$$

si on note p_{ih} la proportion que représente le croisement *ih* dans l'échantillon de la tranche d'effectif *i*.

Dans le cas où on est amené à regrouper des régions pour la correction de la non-réponse, on obtient :

$$B(\hat{Y}_h / n, n_R) = E(\hat{Y}_h) - Y_h$$

$$= \sum_i N_i \frac{n_{ih}}{n_i} E(\bar{y}_{ihT}) - \sum_i N_{ih} \bar{Y}_{ih}$$

$$= \sum_i N_i \frac{n_{ih}}{n_i} \left(\frac{n_{ihR}}{n_{ih}} \bar{Y}_{ihR} + \frac{(n_{ih} - n_{ihR})}{n_{ih}} \bar{Y}_{igR} \right) - \sum_i N_{ih} \bar{Y}_{ih}$$

$$= \sum_i N_i p_{ih} \left(B_{ihR} \bar{Y}_{ihR} + (1 - B_{ihR}) \bar{Y}_{igR} \right) - \sum_i N_i P_{ih} \left(B_{ihR} \bar{Y}_{ihR} + (1 - B_{ihR}) \bar{Y}_{igR} \right) + \sum_i N_i P_{ih} \left(B_{ihR} \bar{Y}_{ihR} + (1 - B_{ihR}) \bar{Y}_{igR} \right) - \sum_i N_i P_{ih} \bar{Y}_{ih}$$

$$= \sum_i N_i (p_{ih} - P_{ih}) \left(B_{ih} \bar{Y}_{ihR} + (1 - B_{ih}) \bar{Y}_{igR} \right) + \sum_i N_i P_{ih} \left(B_{ih} \bar{Y}_{ihR} + (1 - B_{ih}) \bar{Y}_{igR} - \bar{Y}_{ih} \right)$$

où à nouveau on retrouve comme chez Little le fait que le premier terme C tend vers 0 quand la taille de l'échantillon tend vers celle de la population. Concernant le deuxième terme LSB, il n'est nul que si on suppose que $\bar{Y}_{ihR} = \bar{Y}_{igR} = \bar{Y}_{ih} = \bar{Y}_{ig}$, ce qui n'a pas lieu d'être. On aura donc ici un biais "asymptotique" lié aux regroupements.

Variance

- Repondération :

$$V(\hat{Y}_h / n, n_R) = \sum_i (N_i w_{ih})^2 V(\bar{y}_{ihR}) = \sum_i (N_i w_{ih})^2 (1 - \frac{n_{ihR}}{N_{ihR}}) \frac{S_{ih}^2}{n_{ihR}}$$

et en négligeant le terme de population finie on obtient :

$$V(\hat{Y}_h / n, n_R) = \sum_i (N_i w_{ih})^2 \frac{S_{ih}^2}{n_{ihR}}.$$

- Imputation :

En l'absence de regroupement de régions, on obtient :

$$V(\hat{Y}_h / n, n_R) = \sum_i \left(N_i \frac{n_{ih}}{n_i} \right)^2 V(\bar{y}_{ihR}) = \sum_i \left(N_i \frac{n_{ih}}{n_i} \right)^2 (1 - \frac{n_{ihR}}{N_{ihR}}) \frac{S_{ih}^2}{n_{ihR}}$$

soit en négligeant le terme de population finie :

$$V(\hat{Y}_h / n, n_R) = \sum_i \left(N_i \frac{n_{ih}}{n_i} \right)^2 \frac{S_{ih}^2}{n_{ihR}}.$$

S'il est nécessaire de regrouper des régions, on obtient (en négligeant directement le terme de population finie) :

$$V(\hat{Y}_h / n, n_R) = V \left(\sum_i N_i \frac{n_{ig}}{n_i} \frac{n_{ih}}{n_{ig}} \frac{n_{ihR} \bar{y}_{ihR} + (n_{ih} - n_{ihR}) \bar{y}_{igR}}{n_{ih}} \right)$$

$$= \sum_i K^2 \left(\frac{n_{ihR}}{n_{ih}} \right)^2 \frac{S_{ih}^2}{n_{inR}}$$

$$+\sum_i K^2\left(\frac{n_{ih}-n_{ihR}}{n_{ih}}\right)^2\frac{S_{ig}^2}{n_{inR}}$$

$$+2\sum_i K^2\left(\frac{n_{ihR}}{n_{ih}}\right)\left(\frac{n_{ih}-n_{ihR}}{n_{ih}}\right)Cov\left(\bar{y}_{ihR},\bar{y}_{igR}\right)$$

où $K=N_i\frac{n_{ig}}{n_i}\frac{n_{ih}}{n_{ig}}$ et où on a :

$$Cov\left(\bar{y}_{ihR},\bar{y}_{igR}\right)=Cov\left(\bar{y}_{ihR},\sum_{h'\in g}\frac{n_{ih'R}\bar{y}_{ih'R}}{n_{igR}}\right)$$

$$=Cov\left(\bar{y}_{ihR},\frac{n_{ihR}\bar{y}_{ihR}}{n_{igR}}\right)=\frac{n_{ihR}}{n_{igR}}V\left(\bar{y}_{ihR}\right)=\frac{n_{ihR}}{n_{igR}}\frac{S_{ih}^2}{n_{ihR}}=\frac{S_{ih}^2}{n_{igR}}$$

d'où :

$$V(\hat{Y}_h/n,n_R)=\sum_i\left(N_i\frac{n_{ig}}{n_i}\frac{n_{ih}}{n_{ig}}\right)^2\left[\left(\frac{n_{ihR}}{n_{ih}}\right)^2\frac{S_{ih}^2}{n_{ihR}}+\left(\frac{n_{ih}-n_{ihR}}{n_{ih}}\right)^2\frac{S_{ig}^2}{n_{igR}}+2\frac{n_{ihR}\left(n_{ih}-n_{ihR}\right)}{n_{ih}^2}\frac{S_{ih}^2}{n_{igR}}\right].$$

Annexe 2 : Estimateur concurrent du total d'une région

Dans le cas des estimateurs régionaux du total par repondération, on aurait pu envisager un estimateur concurrent de celui qui a été présenté au paragraphe 2.2.2 :

$$\hat{Y}_{hbis} = \frac{\hat{Y}_h}{\hat{N}_h} N_h = \frac{\sum_i N_i w_{ih} \bar{y}_{ihR}}{\sum_i N_i w_{ih}} \cdot N_h .$$

Nous présentons ici les formules de son biais conditionnel et de sa variance conditionnelle.

Biais

$$B(\hat{Y}_{hbis} / n, n_R) = E(\hat{Y}_{hbis} / n, n_R) - Y_h$$

$$= \frac{\sum_i N_i w_{ih} \overline{Y}_{ihR}}{\sum_i N_i w_{ih}} N_h - \sum_i N_{ih} \overline{Y}_{ih}$$

$$= \frac{\sum_i N_i w_{ih} \overline{Y}_{ihR}}{\sum_i N_i w_{ih}} N_h - \frac{\sum_i N_i P_{ihR} \overline{Y}_{ihR}}{\sum_i N_i P_{ihR}} N_h + \frac{\sum_i N_i P_{ihR} \overline{Y}_{ihR}}{\sum_i N_i P_{ihR}} N_h - \sum_i N_{ih} \overline{Y}_{ih}$$

$$= \sum_i N_i \left(\frac{w_{ih}}{\sum_i N_i w_{ih}} - \frac{P_{ihR}}{\sum_i N_i P_{ihR}} \right) \overline{Y}_{ihR} N_h + \frac{\sum_i N_i P_{ih} \overline{Y}_{ihR}}{\sum_i N_i P_{ih}} N_h - \sum_i N_{ih} \overline{Y}_{ih}$$

$$= \sum_i N_i \left(\frac{w_{ih}}{\sum_i N_i w_{ih}} - \frac{P_{ihR}}{\sum_i N_i P_{ihR}} \right) \overline{Y}_{ihR} N_h + \frac{\sum_i N_{ih} \overline{Y}_{ihR}}{\sum_i N_{ih}} N_h - \sum_i N_{ih} \overline{Y}_{ih}$$

$$= \sum_i N_i \left(\frac{w_{ih}}{\sum_i N_i w_{ih}} - \frac{P_{ihR}}{\sum_i N_i P_{ihR}} \right) \overline{Y}_{ihR} N_h + \sum_i N_{ih} \left(\overline{Y}_{ihR} - \overline{Y}_{ih} \right)$$

où on reconnaît :

- un premier terme C qui tend vers 0 quand la taille de l'échantillon tend vers celle de la population et que w_{ih} tend vers P_{ihR}=P_{ih}
- un deuxième terme LSB qui est nul si $\overline{Y}_{ihR} = \overline{Y}_{ih}$.

Variance

En négligeant le terme de population finie, on a approximativement :

$$V(\hat{Y}_{hbis} / n, n_R) = \sum_i \frac{(N_i w_{ih})^2}{\left(\sum_i N_i w_{ih} \right)^2} N_h^2 \frac{S_{ih}^2}{n_{ihR}} .$$

Annexe 3 : Analyse de la variance et ordre de grandeur de la variance inter

A la page 147 de son article de 1986, Little établit une relation entre l'espérance du biais au carré de l'estimateur de la moyenne globale et la variance inter GRH : il montre que l'espérance du biais au carré de l'estimateur ajusté est égale à la variance inter divisée par la taille de l'échantillon. Cette relation nous a semblé intéressante en soi et constitue de plus un moyen de vérification (par le calcul de la variance inter) des ordres de grandeur obtenus dans le calcul des biais au carré. C'est pourquoi nous présentons ici cette formule, adaptée au cas du plan de sondage stratifié qui nous concerne, ainsi que les résultats numériques obtenus lors de la vérification.

On considère comme jusqu'ici que le terme LSB du biais est nul (cas où $\overline{Y}_{ihR} = \overline{Y}_{ih}$), et on a donc :

$$E(Biais^2 / n, n_R) = E\left(\left(\sum_i P_i \sum_h (w_{ih} - P_{ih})\overline{Y}_{ih}\right)^2\right)$$

On se place dans le cas de l'estimateur ajusté 2 où $w_{ih} = p_{ih}$, et dans le cadre de la méthode de tirage a) (tirage aléatoire simple) où, conditionnellement à i, p_{ih} suit une loi multinomiale de paramètres n et les P_{ih}. On a alors :

$$E(Biais^2 / n, n_R) = E\left(\left(\sum_i P_i \sum_h (p_{ih} - P_{ih})\overline{Y}_{ih}\right)^2\right) = E\left(\left(\sum_i P_i \sum_h (p_{ih} - E(p_{ih}))\overline{Y}_{ih}\right)^2\right)$$

$$= \sum_i \sum_h P_i^2 V(p_{ih})\overline{Y}_{ih}^2 + \sum_i \sum_h \sum_{h' \neq h} P_i^2 E((p_{ih} - P_{ih})(p_{ih'} - P_{ih'}))\overline{Y}_{ih}\overline{Y}_{ih'}$$

car pour i et i' distincts, il y a indépendance.

On a donc :

$$E(Biais^2 / n, n_R) = \sum_i \sum_h P_i^2 \frac{P_{ih}(1 - P_{ih})}{n} \overline{Y}_{ih}^2 - \sum_i \sum_h \sum_{h' \neq h} P_i^2 \frac{P_{ih}P_{ih'}}{n} \overline{Y}_{ih}\overline{Y}_{ih'}$$

$$= \frac{1}{n}\sum_i \sum_h P_i^2 P_{ih} \overline{Y}_{ih}^2 - \frac{1}{n}\sum_i \sum_h P_i^2 P_{ih}^2 \overline{Y}_{ih}^2 - \frac{1}{n}\sum_i \sum_h \sum_{h' \neq h} P_i^2 P_{ih} P_{ih'} \overline{Y}_{ih}\overline{Y}_{ih'}$$

$$= \frac{1}{n}\sum_i \sum_h P_i^2 P_{ih} \overline{Y}_{ih}^2 - \frac{1}{n}\sum_i \sum_h \sum_{h'} P_i^2 P_{ih} P_{ih'} \overline{Y}_{ih}\overline{Y}_{ih'}$$

$$= \frac{1}{n}\sum_i \sum_h P_i^2 P_{ih} \overline{Y}_{ih}^2 - \frac{1}{n}\sum_i P_i^2 \left(\sum_h P_{ih}\overline{Y}_{ih}\right)\left(\sum_{h'} P_{ih'}\overline{Y}_{ih'}\right)$$

$$= \frac{1}{n}\sum_i \sum_h P_i^2 P_{ih} \overline{Y}_{ih}^2 - \frac{1}{n}\sum_i P_i^2 \overline{Y}_i^2$$

$$= \frac{1}{n}\sum_i P_i^2 \left(\sum_h P_{ih}\overline{Y}_{ih}^2 - \overline{Y}_i^2\right)$$

$$= \frac{1}{n}\sum_i P_i^2 \left(\sum_h P_{ih}\overline{Y}_{ih}^2 - 2\overline{Y}_i^2 + \overline{Y}_i^2\right)$$

On utilise alors le fait que $\sum_h P_{ih} = 1$:

$$E(Biais^2 / n, n_R) = \frac{1}{n}\sum_i P_i^2 \left(\sum_h P_{ih} \overline{Y}_{ih}^2 - 2\overline{Y}_i \left(\sum_h P_{ih} \overline{Y}_{ih} \right) + \sum_h P_{ih} \overline{Y}_i^2 \right)$$

$$= \frac{1}{n}\sum_i P_i^2 \sum_h P_{ih} \left(\overline{Y}_{ih}^2 - 2\overline{Y}_i \overline{Y}_{ih} + \overline{Y}_i^2 \right)$$

$$= \frac{1}{n}\sum_i P_i^2 \sum_h P_{ih} \left(\overline{Y}_{ih} - \overline{Y}_i \right)^2$$

$$= \frac{1}{n}\sum_i \frac{N_i}{N} P_i \sum_h \frac{N_{ih}}{N_i} \left(\overline{Y}_{ih} - \overline{Y}_i \right)^2$$

Donc : $E(Biais^2 / n, n_R) = \frac{1}{n}\left(\frac{1}{N}\sum_i P_i \sum_h N_{ih} \left(\overline{Y}_{ih} - \overline{Y}_i \right)^2 \right)$.

où $\frac{1}{N}\sum_i P_i \sum_h N_{ih} \left(\overline{Y}_{ih} - \overline{Y}_i \right)^2$ est la variance inter croisements *ih* dans ce plan de sondage stratifié en *i*, tandis que la variance intra s'écrit $\frac{1}{N}\sum_i P_i \sum_h \sum_{k \in U_{ih}} \left(Y_k - \overline{Y}_{ih} \right)^2$ et que la variance totale a pour expression $\frac{1}{N}\sum_i P_i \sum_h \sum_{k \in U_{ih}} \left(Y_k - \overline{Y}_i \right)^2$.

On retrouve donc comme Little que l'espérance du biais au carré est proportionnelle à la variance inter, au terme 1/*n* près.

Les calculs de la variance inter réalisés sur les données de SUSE ont fourni les résultats suivants pour les 4 variables étudiées (exprimées en milliers d'euros sauf la variable effectif) :

APE	n	varinterE/n	varinterCA/n	varinterMA/n	varinterVA/n
151F	292	0,00	0,72	0,30	0,07
158C	481	0,00	0,36	0,11	0,12
524C	1152	0,00	0,24	0,04	0,01
524R	584	0,00	0,26	0,03	0,01
524X	460	0,00	0,25	0,05	0,03

où E désigne l'effectif, CA le chiffre d'affaires, MA la marge commerciale et VA la valeur ajoutée.

La moyenne des biais au carré présentée dans les tableaux de résultats en Annexe 4 est donc bien du même ordre de grandeur que la variance inter divisée par n. Ceci permet de vérifier la validité de nos calculs.

Ceci permet aussi de constater que le rapport Variance inter sur Variance totale est très faible pour les activités étudiées ici :

APE	varinterE/vartotE	varinterCA/vartotCA	varinterMA/vartotMA	varinterVA/vartotVA
151F	1,15%	2,29%	2,70%	1,98%
158C	0,83%	4,18%	1,82%	4,20%
524C	0,38%	0,90%	0,95%	0,36%
524R	0,90%	1,05%	1,44%	0,54%
524X	0,63%	0,95%	1,40%	1,67%

Ceci joue certainement un rôle capital dans les performances modestes des estimateurs globaux, comme cela a été souligné au paragraphe 2.3.1.

Annexe 4 : Résultats détaillés pour chaque jeu de données

NB : Il faut signaler que les calages sur marges ont très mal fonctionné pour les jeux de données C et D. En effet, des cas de multicolinéarité des variables utilisées comme marges (présents en nombre limité pour les jeux de données A et B) se sont multipliés pour les jeux de données C et D. Cette multicolinéarité n'existe pas en réalité au niveau des données individuelles ; toutefois, comme nos calages sur marges portent sur les tableaux de contingence (comptages des individus regroupés au niveau des croisements *ih* ou *ig*), on constate des multicolinéarités apparues de manière fortuite dans les cases des tableaux de contingence[8]. En effet, si on se situe au niveau des données individuelles, on aura colinéarité par exemple si toutes les entreprises de moins de 5 salariés sont dans une même région, ce qui a peu de chances d'arriver ; dans le cas du tableau de contingence par contre, ceci arrivera si le nombre d'entreprises de 0 salariés en région Ile de France est égal au nombre d'entreprises de 1 à 5 salariés en région PACA, ce qui peut arriver plus fréquemment du simple fait du hasard.

Ces cas de multicolinéarité sont d'autant plus susceptibles de se produire que les cases des tableaux de contingence sont peu nombreuses, c'est-à-dire d'autant plus qu'on est amené à regrouper les régions *h*, ce qui est plus fréquent pour les jeux de données C et D où les nombres de répondants sont bien plus petits que dans les cas A ou B.

De ce fait, nous avons dû renoncer aux estimateurs 3 et 5 pour les jeux de données C et D.

De plus, on ne peut pas réellement comparer les estimateurs par imputation et par calage sur marges pour les estimations régionales, car il existe des effets d'interaction que l'imputation prend en compte correctement alors que les estimateurs par calage sur marges n'en tiennent pas compte (Cf. Deville, Särndal et Sautory, 1993, paragraphe 8.1). Nous avons donc choisi de ne pas calculer les estimateurs 3 et 5 pour les estimations par région.

Liste des tableaux de résultats :

Tableau 2 : résultats pour les estimateurs par région pour le jeu de données A

Tableau 3 : résultats pour les estimateurs par région pour le jeu de données B

Tableau 4 : résultats pour les estimateurs par région pour le jeu de données C

Tableau 5 : résultats pour les estimateurs par région pour le jeu de données D

Tableau 6 : minimum et maximum des moyennes de la population par tranche d'effectif pour chaque activité

Tableau 7 : minimum et maximum des moyennes de la population par région pour chaque activité et tranche d'effectif

NB : Pour les tableaux 2 à 5, on présente la moyenne des résultats obtenus pour les 8 régions considérées. Par ailleurs, les chiffres donnés dans les tableaux 6 et 7 ont été calculés à partir de la source SUSE.

[8] On utilise ici les tableaux de contingence car on ne dispose pas des données individuelles : on n'a simulé que les tailles des échantillons et pas les échantillons eux-mêmes.

Explication des abréviations présentes dans les tableaux :

- *Dans les tableaux 2 à 5 :*

biaiscarE est l'espérance du biais au carré de la variable Effectif calculée sur les échantillons obtenus par la méthode a). Idem pour *biaiscarCA*, *biaiscarVA* et *biaiscarMA* pour le chiffre d'affaires, la valeur ajoutée et la marge commerciale (en milliers d'euros).

biaiscarpE est l'espérance du biais au carré de la variable Effectif calculée sur les échantillons obtenus par la méthode b) (p comme tirage à pas systématique). Idem pour *biaiscarpCA*, *biaiscarpVA* et *biaiscarpMA* pour les autres variables d'intérêt.

varE est l'espérance de la variance de la variable Effectif calculée sur les échantillons obtenus par la méthode a). Idem pour *varCA*, *varVA* et *varMA* pour les autres variables d'intérêt.

Idem pour *varpE*, *varpCA*, *varpVA* et *varpMA* pour les échantillons obtenus par la méthode b).

relE est l'espérance de l'"erreur quadratique moyenne *relative*" (telle que définie par Little) de la variable Effectif calculée sur les échantillons obtenus par la méthode a). Idem pour *relCA*, *relVA* et *relMA* pour les autres variables d'intérêt.

Idem pour *relpE*, *relpCA*, *relpVA* et *relpMA* pour les échantillons obtenus par la méthode b).

CVE est l'espérance du coefficient de variation de la variable Effectif calculée sur les échantillons obtenus par la méthode a). Idem pour *CVCA*, *CVVA* et *CVMA* pour les autres variables d'intérêt.

Idem pour *CVpE*, *CVpCA*, *CVpVA* et *CVpMA* pour les échantillons obtenus par la méthode b).

- *Dans le tableau 6 :*

Ebaris_min et *Ebaris_max* sont respectivement le minimum et le maximum des moyennes de la population par tranche d'effectif, pour une activité donnée, pour la variable Effectif, moyennes calculées à partir de la source SUSE (*s* pour SUSE, *i* pour tranche d'effectif, *bar* pour moyenne). Pour l'activité 151F par exemple, on a calculé, sur les données de SUSE, la moyenne de la variable Effectif dans les différentes tranches d'effectif de cette activité, et on présente ici le minimum (0,00) et le maximum (36,95) obtenus parmi ces moyennes par tranche d'effectif.

Idem pour *CAbaris_min*, CA*baris_max*, *VAbaris_min*, VA*baris_max*, *MAbaris_min*, MA*baris_max* pour les autres variables d'intérêt.

- *Dans le tableau 7 :*

Ebarihs_min et *Ebarihs_max* sont respectivement le minimum et le maximum des moyennes de la population par région, pour une activité et une tranche d'effectif données, pour la variable Effectif, moyennes calculées à partir de la source SUSE (*s* pour SUSE, *i* pour tranche d'effectif, *h* pour la région, *bar* pour moyenne). Pour l'activité 151F et la tranche d'effectif 00 par exemple, on a calculé, sur les données de SUSE, la moyenne de la variable Effectif dans les différentes régions de ce croisement activité*tranche d'effectif, et on présente ici le minimum (0,00) et le maximum (0,00) obtenus parmi ces moyennes.

Idem pour *CAbarihs_min*, CA*barihs_max*, *VAbarihs_min*, VA*barihs_max*, *MAbarihs_min*, MA*barihs_max* pour les autres variables d'intérêt.

Tableau 2 : Résultats sur les estimateurs par région pour le jeu de données A

estimateur	activité	txrep	n réel	région	biaiscarE	biaiscarCA	biaiscarMA	biaiscarVA	biaiscarpE	biaiscarpCA	biaiscarpMA	biaiscarpVA	varE	varCA	varMA	varVA	varpE	varpCA	varpMA	varpVA
est_reg1	151F	0,8	292	moyenne	345399	3390493218	498164062	389434854	73578	699062811	101395708	80876900	66346	1238395810	456048446	131858341	66438	1240041882	455205978	132218456
est_reg2	151F	0,8	292	moyenne	306106	2995567485	436200436	341652881	38809	306157785	44617729	36490196	67351	1255428931	462166752	133836084	67504	1257848535	462156479	134177157
est_reg4	151F	0,8	292	moyenne	198507	1726741114	256289998	204926439	37761	301469453	44267151	35832452	70617	1312465400	483596887	140357946	67498	1257677323	461363638	134297184
est_reg6	151F	0,8	292	moyenne	275827	2647442701	388315302	305930369	5179	41587925	6250362	4905520	62830	1132596559	410692997	121633130	64563	1146568462	414083338	123707760
est_reg1	158C	0,8	481	moyenne	5764156	17883918725	1115712678	5183765345	1131747	3519822335	218321320	1015818996	681911	4184271749	3565719363	1422920316	685796	4209468134	3577623934	1431145398
est_reg2	158C	0,8	481	moyenne	4977836	15290750188	943713857	4437270897	232320	579438365	33291602	169911934	689370	4246508878	3627947704	1445994496	693541	4277042285	3637993649	1453934315
est_reg4	158C	0,8	481	moyenne	1512948	3779586475	214690898	1130545087	186557	449992157	26280540	131726485	717780	4466701119	3810226858	1519326358	693867	4281035125	3647357068	1457268445
est_reg6	158C	0,8	481	moyenne	4439988	13917530806	863227562	4032412358	61236	159704340	9023615	47241622	679730	4125076150	3505296652	1407637115	686207	4186689208	3541268411	1425561396
est_reg1	524C	0,8	1152	moyenne	760298	12749387672	1868727815	887168747	178765	3092370524	450529499	211308927	241301	23527329899	3618957951	2984950460	241355	23475033215	3617244985	2983026050
est_reg2	524C	0,8	1152	moyenne	619228	10696302957	1566770935	735990540	2871	54986502	8092404	3707364	242730	23706260569	3643173884	2999961528	242733	23499530690	3628691169	2992665752
est_reg4	524C	0,8	1152	moyenne	9507	125752843	18848191	9012715	0	0	0	0	247961	23965204368	3732345497	3161176540	242770	23441031046	3622818277	2990228404
est_reg6	524C	0,8	1152	moyenne	616481	10661561527	1561522061	733450380	2871	54986502	8092404	3707364	242563	23701444283	3642419533	2999596915	242733	23499530690	3628691169	2992665752
est_reg1	524R	0,8	584	moyenne	180745	3662441671	188290755	332172722	37034	769767302	38648453	72179669	60571	3982954453	331618775	260109376	60142	3952597363	328322752	257763718
est_reg2	524R	0,8	584	moyenne	141618	2878922071	147532861	261834241	2582	68628876	2904738	6310088	60991	3994370360	333933662	262085853	60520	3985722671	331027576	259939598
est_reg4	524R	0,8	584	moyenne	8882	249035650	11964341	20447106	1430	39945995	1558700	3429500	62652	4101858849	342929718	269679148	60533	3980024085	331848677	260413902
est_reg6	524R	0,8	584	moyenne	139620	2804647123	143876981	255812575	1247	33314056	1518417	3247601	60755	3949465454	331556524	260125186	60395	3948946972	329519253	258471352
est_reg1	524X	0,8	460	moyenne	318276	3253775312	634312528	331789188	74161	753473252	148868145	78370069	92624	4080788240	602857237	314813079	92859	4041977447	595171600	327298531
est_reg2	524X	0,8	460	moyenne	268467	2758000695	538905814	281985029	11007	116981500	20043067	11235237	93752	4177643029	613770444	322541184	94097	4101906271	605657710	331726279
est_reg4	524X	0,8	460	moyenne	52061	605455271	107206490	57696310	7893	84473851	13525925	7913869	98564	4436361341	650567049	348902192	94197	4160136174	612940479	329172996
est_reg6	524X	0,8	460	moyenne	256236	2634164855	517230363	269587034	4220	46268200	8657733	4534328	93227	3990102539	591375756	308806714	93865	4011921911	594316302	323579033

estimateur	activité	txrep	n réel	région	relhE	relhCA	relhMA	relhVA	relhpE	relhpCA	relhpMA	relhpVA	CVhE	CVhCA	CVhMA	CVhVA	CVhpE	CVhpCA	CVhpMA	CVhpVA
est_reg1	151F	0,8	292	moyenne	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	24,06	23,89	32,24	24,10	14,70	16,16	25,15	16,11
est_reg2	151F	0,8	292	moyenne	2,91	1,01	-0,03	1,30	-5,57	-5,85	-3,15	-5,77	23,57	23,37	31,85	23,60	13,32	14,86	24,18	14,82
est_reg4	151F	0,8	292	moyenne	-0,38	-5,28	-5,23	-4,50	-5,85	-5,89	-3,24	-5,81	19,98	20,01	28,61	20,23	13,27	14,84	24,14	14,80
est_reg6	151F	0,8	292	moyenne	-0,71	-3,88	-5,69	-3,31	-21,72	-18,66	-13,09	-18,79	22,08	21,79	29,73	22,10	10,62	12,55	21,42	12,46
est_reg1	158C	0,8	481	moyenne	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	16,18	16,26	33,71	16,83	9,41	10,38	30,76	10,99
est_reg2	158C	0,8	481	moyenne	1,26	0,20	-0,43	-0,08	-16,72	-13,27	-1,36	-12,21	15,37	15,54	33,43	16,12	7,16	8,53	30,27	9,22
est_reg4	158C	0,8	481	moyenne	-20,11	-20,64	-4,13	-19,49	-18,52	-14,44	-1,50	-13,25	10,34	10,87	31,72	11,59	6,99	8,41	30,22	9,10
est_reg6	158C	0,8	481	moyenne	-1,07	-2,32	-3,01	-2,66	-24,88	-18,68	-3,74	-17,28	14,69	14,97	32,49	15,53	6,29	7,90	29,48	8,59
est_reg1	524C	0,8	1152	moyenne	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	8,83	10,28	10,35	12,61	6,14	8,58	8,75	10,92
est_reg2	524C	0,8	1152	moyenne	-1,25	-1,46	-1,34	-1,42	-17,85	-7,05	-6,47	-6,80	8,21	9,91	10,00	12,24	4,79	7,83	8,05	10,17
est_reg4	524C	0,8	1152	moyenne	-34,13	-16,17	-14,87	-15,12	-18,36	-7,24	-6,63	-6,94	4,92	7,99	8,22	10,39	4,76	7,81	8,04	10,16
est_reg6	524C	0,8	1152	moyenne	-1,45	-1,57	-1,44	-1,50	-17,85	-7,05	-6,47	-6,80	8,19	9,90	9,99	12,23	4,79	7,83	8,05	10,17
est_reg1	524R	0,8	584	moyenne	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	12,45	13,64	14,61	13,99	8,60	10,82	12,36	10,74
est_reg2	524R	0,8	584	moyenne	-0,77	-1,97	-1,66	-2,01	-13,88	-7,29	-4,50	-8,24	11,68	13,07	14,17	13,31	7,14	9,92	11,71	9,69
est_reg4	524R	0,8	584	moyenne	-27,17	-17,51	-12,36	-19,64	-15,00	-7,85	-4,84	-9,15	7,72	10,45	12,22	10,25	7,03	9,85	11,66	9,57
est_reg6	524R	0,8	584	moyenne	-1,56	-3,26	-2,80	-3,05	-15,33	-9,33	-5,99	-10,18	11,56	12,83	13,98	13,11	6,95	9,62	11,50	9,39
est_reg1	524X	0,8	460	moyenne	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	16,76	19,43	19,02	19,40	11,46	15,83	15,12	15,42
est_reg2	524X	0,8	460	moyenne	0,14	-0,67	-0,73	-0,63	-13,16	-5,91	-6,82	-7,75	15,92	18,94	18,45	18,82	9,53	14,81	13,96	14,12
est_reg4	524X	0,8	460	moyenne	-21,93	-11,47	-13,58	-14,11	-14,46	-6,33	-7,31	-8,44	11,08	16,20	15,32	15,51	9,37	14,76	13,90	13,99
est_reg6	524X	0,8	460	moyenne	-0,73	-2,49	-2,23	-2,14	-15,87	-8,58	-9,00	-10,11	15,61	18,44	18,06	18,39	9,18	14,36	13,61	13,72

Tableau 3 : Résultats sur les estimateurs par région pour le jeu de données B

estimateur	activité	txrep	n réel	région	biaiscarE	biaiscarCA	biaiscarMA	biaiscarVA	biaiscarpE	biaiscarpCA	biaiscarpMA	biaiscarpVA	varE	varCA	varMA	varVA	varpE	varpCA	varpMA	varpVA
est_reg1	151F	0,6	292	moyenne	485564	4523754700	692656730	538339929	195269	1837198912	266213246	213997268	87943	1652058349	607761571	175725186	89696	1676087344	617918327	179064758
est_reg2	151F	0,6	292	moyenne	401821	3694697668	566467893	440135603	140013	1228529860	178221351	143921681	90567	1690120371	622421876	180378020	91905	1715772104	631129753	183556430
est_reg4	151F	0,6	292	moyenne	299877	2661788118	412346423	319863846	139242	1223154646	177643051	143168639	94188	1748793582	643438209	187412556	91936	1717408323	629997386	183445683
est_reg6	151F	0,6	292	moyenne	267880	2505791588	375412309	296219366	8364	76893779	11431469	8697592	74091	1290308585	459827226	139509936	73335	1270168032	450769849	137802870
est_reg1	158C	0,6	481	moyenne	7702093	24136094128	1546502621	7011253180	2995600	9516565265	599040989	2753379260	913012	5640529731	4790758478	1917534387	916137	5626137187	4798282657	1915171834
est_reg2	158C	0,6	481	moyenne	5893043	17646355634	1131198603	5140815462	1066842	2832390159	171166016	849014022	934288	5768516279	4926422112	1961692637	938793	5800010608	4957878486	1974782355
est_reg4	158C	0,6	481	moyenne	2876668	7586219045	467241683	2256134035	1013199	2666329794	158895262	798762611	963523	5961703653	5120501227	2030112143	939382	5805226127	4958559519	1974433858
est_reg6	158C	0,6	481	moyenne	4567096	14175772169	895975124	4107090912	73186	192432932	13072235	57569365	889407	5249785577	4383337632	1794914619	898179	5309849859	4449545463	1820714218
est_reg1	524C	0,6	1152	moyenne	1174331	20177773367	2954256263	1387549171	468029	8190419446	1206480650	559626990	323205	30992938479	4791316589	3931813881	322597	31613002378	4904422777	4183899608
est_reg2	524C	0,6	1152	moyenne	672013	12021528004	1748147371	820934090	5126	83568667	12015764	6011426	328255	31659133005	4941979365	4250136729	327972	31506439065	4861633158	3957680217
est_reg4	524C	0,6	1152	moyenne	13677	213670819	31052046	16610654	1934	25750726	3708675	1971592	336939	32543846022	5041485966	4238922533	328010	31304934681	4807038319	3823098525
est_reg6	524C	0,6	1152	moyenne	663721	11895480900	1730127700	811041818	3561	65021340	9333890	4610671	327905	31573888142	4921006510	4202103033	327618	31486582174	4857727352	3951859074
est_reg1	524R	0,6	584	moyenne	245127	5002554035	253732244	466638731	90273	1887734671	95112131	178501205	80167	5301558027	441184054	345319027	80808	5313576308	442479048	346723978
est_reg2	524R	0,6	584	moyenne	175219	3555289210	185217158	324659540	9744	226922816	11847967	19004103	81568	5388951029	448452063	351466991	82041	5339805163	447843271	351617872
est_reg4	524R	0,6	584	moyenne	30646	761811125	40719636	64087256	8280	208055067	10835119	17265966	83788	5487452945	457790283	360688563	82080	5331871889	447268782	351120438
est_reg6	524R	0,6	584	moyenne	157269	3150739833	162648137	290202527	1576	24350242	1317044	2194569	80184	5133043346	433024138	338268966	80608	5146770296	435855951	341897387
est_reg1	524X	0,6	460	moyenne	395623	4140028766	821186837	424007517	166603	1717700903	340003675	176042393	123115	5395481104	797097643	438561433	124226	5486798901	808559572	433378786
est_reg2	524X	0,6	460	moyenne	293460	3112966942	614118761	318873959	37886	414657916	74758741	40159917	126753	5605628626	831297671	465168710	128548	5614016701	833448720	441968099
est_reg4	524X	0,6	460	moyenne	108078	1170309853	215023640	115225829	37150	412804060	74025643	39837935	131848	5669726700	859260415	476455042	128571	5658527136	834537009	445602125
est_reg6	524X	0,6	460	moyenne	246856	2550864450	506450551	262354502	3203	21786877	4047811	2233993	122783	4900213123	728196538	394411089	125916	4854437763	731518341	385751366

estimateur	activité	txrep	n réel	région	relhE	relhCA	relhMA	relhVA	relhpE	relhpCA	relhpMA	relhpVA	CVhE	CVhCA	CVhMA	CVhVA	CVhpE	CVhpCA	CVhpMA	CVhpVA
est_reg1	151F	0,6	292	moyenne	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	27,27	26,80	36,24	27,12	20,67	21,46	31,47	21,54
est_reg2	151F	0,6	292	moyenne	0,05	-1,06	-1,26	-1,15	-2,36	-3,68	-2,52	-3,49	25,68	25,50	35,14	25,76	18,85	19,86	30,07	19,90
est_reg4	151F	0,6	292	moyenne	-2,26	-4,25	-3,70	-4,26	-2,28	-3,65	-2,55	-3,45	23,44	23,57	33,36	23,77	18,84	19,85	30,06	19,89
est_reg6	151F	0,6	292	moyenne	-10,16	-13,46	-16,04	-13,13	-35,24	-32,22	-26,55	-32,39	21,46	21,20	29,06	21,53	11,41	13,13	21,96	13,08
est_reg1	158C	0,6	481	moyenne	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	18,48	18,58	38,92	19,23	13,04	13,89	36,41	14,58
est_reg2	158C	0,6	481	moyenne	2,88	-0,24	-0,65	-0,44	-13,18	-12,50	-1,79	-11,64	16,85	16,98	38,31	17,68	9,98	11,13	35,53	11,93
est_reg4	158C	0,6	481	moyenne	-12,35	-15,07	-3,39	-14,16	-14,18	-13,20	-1,78	-12,30	13,26	13,59	36,98	14,40	9,85	11,02	35,55	11,82
est_reg6	158C	0,6	481	moyenne	-3,95	-7,30	-7,95	-7,52	-33,47	-28,25	-9,65	-26,60	15,05	15,37	35,30	16,02	7,24	8,87	32,50	9,65
est_reg1	524C	0,6	1152	moyenne	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	10,68	12,28	12,36	14,97	7,96	10,49	10,66	13,28
est_reg2	524C	0,6	1152	moyenne	-5,96	-3,76	-3,64	-3,47	-23,53	-10,58	-9,86	-10,21	9,00	11,24	11,37	14,04	5,59	9,10	9,35	11,77
est_reg4	524C	0,6	1152	moyenne	-34,91	-17,05	-15,89	-16,13	-23,94	-10,92	-10,20	-10,60	5,74	9,30	9,54	12,06	5,57	9,07	9,32	11,68
est_reg6	524C	0,6	1152	moyenne	-6,42	-4,08	-3,93	-3,54	-23,87	-10,70	-9,97	-10,24	8,95	11,20	11,33	14,01	5,57	9,08	9,33	11,76
est_reg1	524R	0,6	584	moyenne	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	14,55	15,95	16,98	16,32	11,06	13,33	14,92	13,34
est_reg2	524R	0,6	584	moyenne	-1,02	-2,94	-2,29	-2,92	-16,70	-10,53	-6,56	-11,61	13,17	14,87	16,19	15,08	8,60	11,57	13,71	11,33
est_reg4	524R	0,6	584	moyenne	-22,09	-15,44	-10,55	-16,90	-17,18	-10,72	-6,61	-11,75	9,75	12,53	14,49	12,42	8,55	11,54	13,69	11,31
est_reg6	524R	0,6	584	moyenne	-5,08	-7,89	-6,78	-7,75	-23,05	-16,61	-11,59	-17,66	12,43	13,93	15,37	14,13	7,81	10,65	12,93	10,38
est_reg1	524X	0,6	460	moyenne	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	18,80	22,00	21,61	22,19	14,72	19,27	18,57	18,83
est_reg2	524X	0,6	460	moyenne	-1,33	-0,81	-1,13	-0,98	-12,67	-6,29	-7,33	-8,14	17,22	21,10	20,59	21,16	11,87	17,67	16,78	16,86
est_reg4	524X	0,6	460	moyenne	-15,29	-8,59	-10,03	-10,28	-12,92	-6,30	-7,39	-8,14	13,80	18,95	18,24	18,63	11,83	17,69	16,76	16,87
est_reg6	524X	0,6	460	moyenne	-5,68	-7,87	-7,80	-7,75	-21,97	-15,39	-16,15	-16,76	16,11	19,31	18,92	19,36	10,47	15,72	15,02	15,10

Tableau 4 : Résultats sur les estimateurs par région pour le jeu de données C

estimateur	activité	txrep	50% n réel	région	biaiscarE	biaiscarCA	biaiscarMA	biaiscarVA	biaiscarpE	biaiscarpCA	biaiscarpMA	biaiscarpVA	varE	varCA	varMA	varVA	varpE	varpCA	varpMA	varpVA
est_reg1	151F	0,8	146	moyenne	650939	6127821543	917916526	714651669	160666	1577889571	233382356	179529634	132104	2462048201	906001221	263853373	133977	2486702646	909256338	265213500
est_reg2	151F	0,8	146	moyenne	621567	5788260053	869362999	679629729	132211	1272404719	188533878	143678020	133791	2493337312	916597841	267373826	135965	2517286958	919046912	269207723
est_reg4	151F	0,8	146	moyenne	522047	4892328709	726784176	570585493	133232	1270751292	189073113	144213661	139469	2594096732	949123249	277658768	136063	2521272333	920270668	269536873
est_reg6	151F	0,8	146	moyenne	511735	4718479276	704615236	556281115	28778	315306474	39172192	32264211	110789	1998633402	728586563	216156481	108208	1955874181	708958416	210380468
est_reg1	158C	0,8	241	moyenne	12125576	39044363740	2486369991	11341715142	2596673	7988185060	481975157	2334249104	1365076	8380667529	7132180098	2858206891	1353031	8304926113	7061659322	2823125599
est_reg2	158C	0,8	241	moyenne	11111849	35034605874	2226862840	10225221761	1561792	4233718851	259240954	1275672178	1385272	8507763688	7248395017	2902654987	1375597	8464745924	7215020811	2882192824
est_reg4	158C	0,8	241	moyenne	6684341	17909919806	1135891585	5374561878	1339732	3534166202	216751377	1068281489	1472282	8959153643	7649100211	3058296527	1377494	8478780417	7233890642	2887384288
est_reg6	158C	0,8	241	moyenne	9511552	30778182604	1943435366	8944901354	361386	1092944854	66722481	320743212	1309255	7750443091	6456079361	2651583241	1307007	7732423606	6439092120	2642221089
est_reg1	524C	0,8	576	moyenne	1498999	26297790717	3823324185	1814767792	328473	5774516374	846489346	394307552	484089	47271874981	7311616203	6152145551	483051	46781257854	7196875158	5880801139
est_reg2	524C	0,8	576	moyenne	1210771	21493371718	3120231444	1476837022	15756	281912075	40494274	20679956	490132	47348608456	7367373716	6251127609	488615	46538086533	7163028718	5732446668
est_reg4	524C	0,8	576	moyenne	123193	1890604255	275336877	139608799	10574	175912671	25657055	13743508	510263	48732323766	7517946935	6126500319	488692	46736513280	7190472870	5773055142
est_reg6	524C	0,8	576	moyenne	1182552	21101139197	3062648252	1446623361	5019	98902565	13867137	6622072	486416	46712838225	7217688027	5916502095	487622	46228217053	7087515996	5554242584
est_reg1	524R	0,8	292	moyenne	374369	8024676645	401939099	758180716	80508	1607033870	81760679	147927020	120259	7889003626	656853153	515830817	120381	7875091653	656749543	515272736
est_reg2	524R	0,8	292	moyenne	312326	6562725581	332310713	614684164	17504	448143536	22918071	38571745	121693	7995842836	664176006	522802681	121831	7982788842	666567702	522163027
est_reg4	524R	0,8	292	moyenne	77519	2127665247	110068291	184850676	15136	417061205	21515067	35665781	127822	8399692076	701393905	548809564	121902	7990546445	665179582	522848245
est_reg6	524R	0,8	292	moyenne	297938	6197911659	311769173	581144819	5754	127173833	5934417	11254986	119406	7542616489	635216736	498519585	119248	7533144520	638194158	497836397
est_reg1	524X	0,8	230	moyenne	668544	6846503119	1356667332	701227336	135837	1363793143	267386567	140546439	186677	8243432149	1214927537	628683512	184750	8071931052	1196639734	646436416
est_reg2	524X	0,8	230	moyenne	584756	6074844320	1193722228	616519053	57077	587584989	108682616	58425677	190443	8409096869	1243818227	634835791	188431	8043779763	1205906324	648029815
est_reg4	524X	0,8	230	moyenne	303814	3114415890	603808774	312900208	54359	538782730	98815795	53461192	203255	8731980841	1299554245	676496135	188614	8090682507	1204290680	654214505
est_reg6	524X	0,8	230	moyenne	522806	5344260327	1060704534	549032842	13488	135627155	24759107	13437902	178871	7334147876	1086697688	560302632	180404	7199753793	1072288007	563006600

estimateur	activité	txrep	50% n réel	région	relhE	relhCA	relhMA	relhVA	relhpE	relhpCA	relhpMA	relhpVA	CVhE	CVhCA	CVhMA	CVhVA	CVhpE	CVhpCA	CVhpMA	CVhpVA
est_reg1	151F	0,8	146	moyenne	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	33,03	32,79	44,54	33,09	21,61	23,57	36,17	23,49
est_reg2	151F	0,8	146	moyenne	0,78	0,03	-0,14	0,13	-0,58	-1,19	-0,77	-1,09	32,33	32,14	43,94	32,43	21,00	23,01	35,74	22,95
est_reg4	151F	0,8	146	moyenne	2,45	0,59	-0,46	0,76	-0,32	-1,18	-0,71	-1,06	30,59	30,83	42,59	31,01	21,03	23,00	35,74	22,94
est_reg6	151F	0,8	146	moyenne	-5,54	-8,75	-10,77	-8,05	-22,39	-20,24	-18,21	-20,52	28,96	28,47	38,66	28,87	15,34	17,79	28,77	17,63
est_reg1	158C	0,8	241	moyenne	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	22,90	22,91	47,57	23,70	13,65	14,91	43,30	15,78
est_reg2	158C	0,8	241	moyenne	2,45	0,29	-0,21	0,23	-4,70	-5,87	-0,64	-5,33	22,17	22,09	47,26	22,93	12,15	13,47	42,86	14,43
est_reg4	158C	0,8	241	moyenne	0,47	-5,75	-1,60	-5,28	-7,00	-7,76	-0,86	-7,10	18,90	18,64	45,97	19,65	11,74	13,12	42,72	14,08
est_reg6	158C	0,8	241	moyenne	1,16	-2,56	-6,24	-2,82	-21,77	-18,91	-7,73	-17,86	20,80	20,85	44,28	21,63	9,45	11,29	39,67	12,20
est_reg1	524C	0,8	576	moyenne	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	12,53	14,80	14,88	18,08	8,44	11,97	12,22	15,24
est_reg2	524C	0,8	576	moyenne	-0,91	-1,02	-0,97	-0,91	-15,15	-5,78	-5,27	-5,54	11,78	14,36	14,47	17,65	6,89	11,12	11,43	14,33
est_reg4	524C	0,8	576	moyenne	-28,17	-14,33	-13,29	-13,56	-15,66	-5,81	-5,28	-5,61	7,67	11,78	12,05	15,00	6,85	11,11	11,41	14,32
est_reg6	524C	0,8	576	moyenne	-1,83	-1,97	-1,81	-0,73	-16,34	-6,55	-5,89	-5,30	11,64	14,23	14,36	17,59	6,78	11,03	11,35	14,31
est_reg1	524R	0,8	292	moyenne	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	17,89	19,82	21,04	20,42	12,15	15,21	17,42	15,06
est_reg2	524R	0,8	292	moyenne	-0,68	-1,93	-1,45	-1,80	-9,99	-5,44	-3,22	-5,93	17,03	19,01	20,45	19,52	10,58	14,21	16,73	13,97
est_reg4	524R	0,8	292	moyenne	-17,10	-12,23	-8,30	-13,60	-10,62	-5,60	-3,42	-6,09	12,82	16,06	18,29	16,10	10,51	14,19	16,70	13,94
est_reg6	524R	0,8	292	moyenne	-2,15	-5,15	-4,69	-4,90	-15,54	-11,42	-8,26	-12,13	16,54	18,25	19,72	18,74	9,79	13,16	15,81	12,85
est_reg1	524X	0,8	230	moyenne	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	24,24	27,84	27,37	27,70	15,98	22,26	21,30	21,56
est_reg2	524X	0,8	230	moyenne	0,65	-0,38	-0,29	-0,61	-6,43	-2,88	-3,17	-3,80	23,41	27,29	26,78	26,98	14,53	21,45	20,45	20,56
est_reg4	524X	0,8	230	moyenne	-7,37	-4,53	-5,71	-4,76	-6,77	-3,05	-3,65	-4,04	19,72	25,12	24,30	24,58	14,44	21,40	20,32	20,51
est_reg6	524X	0,8	230	moyenne	-3,14	-6,62	-6,16	-5,98	-15,87	-12,13	-12,52	-12,97	21,97	25,28	24,90	25,15	12,74	19,23	18,27	18,33

Tableau 5 : Résultats sur les estimateurs par région pour le jeu de données D

estimateur	activité	txrep	50% n réel	région	biaiscarE	biaiscarCA	biaiscarMA	biaiscarVA	biaiscarpE	biaiscarpCA	biaiscarpMA	biaiscarpVA	varE	varCA	varMA	varVA	varpE	varpCA	varpMA	varpVA
est_reg1	151F	0,6	146	moyenne	1001299	9640560378	1404710936	1124911924	410764	4060031498	572549965	457578499	183750	3417329964	1250025241	364629230	179118	3363110863	1234975158	357974566
est_reg2	151F	0,6	146	moyenne	998555	9537304316	1392277752	1115083676	369288	3564472801	513136770	404117656	188808	3494736637	1272791893	373760048	184233	3439068137	1254961056	367192864
est_reg4	151F	0,6	146	moyenne	979615	9179915460	1328195080	1078654337	368328	3557271733	510748935	403273085	192130	3551655001	1287388859	379561931	184282	3437276315	1256889536	367233555
est_reg6	151F	0,6	146	moyenne	535100	5089478228	752193006	595258827	32029	325795584	41712829	34704342	115235	2005768945	725725325	218646578	112957	1957688667	699864647	213231422
est_reg1	158C	0,6	241	moyenne	14436452	45213287699	2866779328	13129582296	6678988	21031979418	1279542522	6080270094	1840216	11354851510	9712749003	3861913874	1843804	11299270811	9571123064	3841011115
est_reg2	158C	0,6	241	moyenne	12985861	39300681657	2490951823	11474309647	4474871	12391393052	739971111	3663629801	1904839	11774972320	10083404975	4011762217	1913683	11673648833	9926169742	3975668108
est_reg4	158C	0,6	241	moyenne	10356776	29098072023	1888374174	8664280267	4330430	11991827213	720996601	3548769678	1962677	12000990034	10314603712	4091376512	1915502	11690505547	9952901291	3983268184
est_reg6	158C	0,6	241	moyenne	7912787	25364219290	1556775360	7301715184	483337	1352072600	72456837	394935994	1638510	9246076243	7441807529	3170928431	1649329	9189235102	7355195630	3154143598
est_reg1	524C	0,6	576	moyenne	2081967	36294201153	5303430774	2504331413	910422	15577970417	2289326329	1071494358	646127	62250654296	9616949926	7898890463	651618	62378256162	9674716329	8041288494
est_reg2	524C	0,6	576	moyenne	1383888	23542530687	3455734194	1637468423	76028	1124378839	163384600	83361637	662770	63913050785	9969125243	8451186794	669262	63446531697	9778782064	7895018599
est_reg4	524C	0,6	576	moyenne	259358	3862202206	562339180	281215996	69849	1014285453	147167710	76202592	684866	67209733052	10371192480	8595364321	669488	63778173479	9813062911	7932236946
est_reg6	524C	0,6	576	moyenne	1244640	21375504917	3141295619	1479711688	5983	107675751	15641597	7370788	645127	60914621166	9282895170	6973539153	655781	61063092223	9204921990	6589872303
est_reg1	524R	0,6	292	moyenne	435041	9203492209	452953966	879180286	210993	4274396454	219808254	395295193	159639	10468427025	873308069	684956387	161502	10600801734	884036411	691594044
est_reg2	524R	0,6	292	moyenne	313301	6918451343	341837298	658641875	57429	1385926461	72127440	120968793	164848	10718660176	895106513	704802029	166051	10861146113	900686350	708962121
est_reg4	524R	0,6	292	moyenne	153818	3615927362	188077225	324171578	56469	1386831076	72116017	120379596	170748	11237443159	938144140	735403125	166188	10879602537	902966152	710910439
est_reg6	524R	0,6	292	moyenne	245242	5276102873	256794004	506999504	5089	114733877	5561024	10633233	154221	9091197232	783625789	611086986	158070	9326795474	802265217	625257831
est_reg1	524X	0,6	230	moyenne	909803	9241795327	1815342602	952556815	354479	3566305113	700701883	364636830	249469	11075349546	1629322413	909770752	248042	11272845239	1641973273	828423569
est_reg2	524X	0,6	230	moyenne	833972	8345237574	1647913484	863157127	199771	2032625744	387871615	201549447	258235	11184484427	1676524166	911750538	258364	11584677214	1703551811	842698742
est_reg4	524X	0,6	230	moyenne	602047	6057505399	1196035070	625066399	193625	1983663595	378088636	196128211	267797	11464280855	1720583210	941197418	258431	11545217397	1702210130	841061246
est_reg6	524X	0,6	230	moyenne	536457	5336651519	1053037699	552643084	13587	149306419	27303884	14470119	203682	7662338924	1148338226	620426453	211899	8030155538	1182786860	598351289

estimateur	activité	txrep	50% n réel	région	relhE	relhCA	relhMA	relhVA	relhpE	relhpCA	relhpMA	relhpVA	CVhE	CVhCA	CVhMA	CVhVA	CVhpE	CVhpCA	CVhpMA	CVhpVA
est_reg1	151F	0,6	146	moyenne	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	39,64	39,16	52,33	39,56	30,43	31,51	45,55	31,65
est_reg2	151F	0,6	146	moyenne	2,81	2,34	1,10	2,39	0,58	0,04	-0,05	-0,09	39,54	39,10	52,19	39,50	29,64	30,88	45,07	30,96
est_reg4	151F	0,6	146	moyenne	3,57	1,84	0,62	1,94	0,42	-0,09	-0,08	-0,22	38,82	38,31	51,34	38,68	29,64	30,89	45,13	30,98
est_reg6	151F	0,6	146	moyenne	-13,36	-19,09	-22,61	-17,73	-41,37	-38,48	-34,95	-38,88	29,64	28,76	38,63	29,45	15,21	17,39	27,92	17,30
est_reg1	158C	0,6	241	moyenne	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	26,28	26,38	55,01	27,30	19,30	20,33	51,78	21,29
est_reg2	158C	0,6	241	moyenne	4,01	2,17	0,69	1,89	-1,64	-4,73	-0,62	-4,26	25,32	25,40	54,95	26,40	17,31	18,18	51,05	19,28
est_reg4	158C	0,6	241	moyenne	2,43	-1,66	0,04	-1,54	-2,50	-5,36	-0,69	-4,89	23,18	23,02	53,89	24,13	17,11	18,01	50,99	19,11
est_reg6	158C	0,6	241	moyenne	-6,46	-9,66	-14,60	-10,23	-33,20	-31,75	-18,09	-30,24	20,71	21,13	46,18	21,96	10,53	12,16	41,68	13,17
est_reg1	524C	0,6	576	moyenne	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	14,37	17,02	17,13	20,72	11,31	14,79	15,03	18,62
est_reg2	524C	0,6	576	moyenne	-1,59	-1,59	-1,46	-1,52	-19,08	-8,02	-7,40	-7,72	12,88	16,06	16,25	19,89	8,42	13,19	13,53	16,92
est_reg4	524C	0,6	576	moyenne	-21,28	-10,44	-9,76	-10,25	-19,39	-8,09	-7,49	-7,79	9,36	13,97	14,26	17,74	8,39	13,16	13,50	16,91
est_reg6	524C	0,6	576	moyenne	-5,47	-5,22	-4,34	0,51	-25,10	-11,28	-9,89	-5,47	12,25	15,51	15,83	19,89	7,80	12,72	13,18	16,96
est_reg1	524R	0,6	292	moyenne	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	20,20	22,28	23,74	23,01	16,29	19,14	21,35	19,26
est_reg2	524R	0,6	292	moyenne	-1,37	-0,86	-0,82	-1,28	-12,72	-6,82	-4,36	-8,09	18,55	21,29	23,01	21,82	13,43	17,39	20,10	17,16
est_reg4	524R	0,6	292	moyenne	-11,77	-7,27	-4,59	-9,06	-12,93	-6,85	-4,45	-8,14	15,92	19,50	21,75	19,62	13,39	17,38	20,07	17,14
est_reg6	524R	0,6	292	moyenne	-8,45	-11,12	-10,51	-10,87	-26,07	-21,47	-16,61	-22,65	16,64	18,71	20,58	19,23	10,93	14,25	17,34	13,95
est_reg1	524X	0,6	230	moyenne	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	28,61	32,59	31,96	32,98	21,47	27,83	26,79	26,73
est_reg2	524X	0,6	230	moyenne	1,40	-0,05	0,21	-0,12	-5,89	-2,76	-3,50	-3,75	27,75	31,96	31,39	32,13	19,15	26,64	25,42	25,14
est_reg4	524X	0,6	230	moyenne	-4,00	-3,05	-3,06	-4,00	-6,19	-2,96	-3,63	-4,01	25,19	30,32	29,63	30,13	19,09	26,60	25,38	25,08
est_reg6	524X	0,6	230	moyenne	-10,15	-15,84	-15,65	-14,80	-30,14	-24,74	-26,13	-25,22	23,04	26,00	25,53	26,37	13,42	19,92	18,83	18,85

Tableau 6 : Minimum et maximum des moyennes de la population par tranche d'effectif

activité	Ebaris_min	Ebaris_max	CAbaris_min	CAbaris_max	MAbaris_min	MAbaris_max	VAbaris_min	VAbaris_max
151F	0,00	36,95	105,45	2780,30	37,01	978,73	28,01	1092,15
158C	0,00	37,19	71,72	1769,03	17,78	453,78	31,56	912,43
524C	0,00	13,20	83,84	1702,38	29,61	633,36	18,49	448,79
524R	0,00	13,41	87,89	1739,61	15,76	466,01	28,65	454,20
524X	0,00	13,69	72,76	1562,62	30,00	592,40	19,32	446,17

Tableau 7 : Minimum et maximum des moyennes de la population par région

activité	tranche d'effectif	Ebarihs_min	Ebarihs_max	CAbarihs_min	CAbarihs_max	MAbarihs_min	MAbarihs_max	VAbarihs_min	VAbarihs_max
151F	00	0,00	0,00	94,85	117,04	32,02	45,44	24,12	29,27
151F	01	1,98	2,51	236,71	274,07	77,46	119,56	67,87	92,50
151F	03	6,98	7,50	490,01	648,00	174,93	284,48	186,57	238,08
151F	11	12,78	14,26	878,93	1786,26	226,11	537,09	349,33	515,26
151F	12	21,40	24,67	1621,00	3310,33	497,67	1208,33	642,67	825,00
151F	13	31,50	41,00	2529,00	3664,00	-2,00	1715,33	940,00	1527,50
158C	00	0,00	0,00	64,22	77,49	12,16	23,08	28,15	35,37
158C	01	2,47	2,81	156,47	215,49	24,17	55,68	77,84	113,92
158C	03	7,08	7,25	335,86	427,90	51,15	104,87	184,92	242,05
158C	11	12,38	13,05	580,56	737,33	85,36	211,61	322,82	425,09
158C	12	22,42	23,89	1000,06	1303,50	160,63	336,94	492,55	761,22
158C	13	33,75	41,47	1467,00	2343,67	220,33	654,17	684,00	1294,47
524C	00	0,00	0,00	72,24	115,59	25,83	43,10	14,99	45,68
524C	01	1,95	2,16	229,20	304,31	87,97	119,15	61,66	74,91
524C	03	7,14	7,43	807,43	1135,88	311,12	457,50	218,37	303,10
524C	11	12,78	13,86	1382,79	2027,55	498,50	774,80	391,64	518,37
524R	00	0,00	0,00	81,78	100,69	12,54	19,22	25,01	33,07
524R	01	1,71	1,98	220,00	280,66	40,60	67,22	70,36	78,97
524R	03	7,10	7,40	755,36	1061,07	135,71	288,74	217,82	259,37
524R	11	12,53	14,42	1463,39	2040,57	337,10	560,00	385,47	533,73
524X	00	0,00	0,00	64,59	83,50	25,45	34,19	17,12	21,43
524X	01	1,86	2,12	171,79	221,64	79,05	104,08	52,99	75,27
524X	03	7,06	7,41	582,98	741,60	230,74	310,14	173,17	243,38
524X	11	12,68	14,37	1297,72	1976,24	512,21	713,11	355,54	512,86

Annexe 5 : Programmes SAS utilisés pour les simulations

Nous présentons ici les principaux programmes utilisés pour réaliser les simulations. Nous avons travaillé tout au long de ces simulations activité par activité : chacun des programmes qui suivent doit donc être utilisé successivement pour chacune des 5 activités retenues (151F, 158C, 524C, 524R, 524X). Nous présentons ces programmes dans l'ordre chronologique de leur utilisation, la table de résultats obtenue par l'un des programmes étant la table utilisée en entrée par le programme suivant.
Par exemple, pour l'activité 151F on a la succession suivante :

- le premier programme concerne la simulation des tirages d'échantillons et fournit une table nommée act151F ;
- le deuxième programme concerne les regroupements de régions *h* en supra-régions *g* au sein d'un croisement activité*tranche d'effectif donné, regroupements nécessaires lorsque le nombre de répondants au niveau des régions *h* est trop faible pour pouvoir y corriger la non-réponse ; il utilise la table act151F en entrée et donne en sortie la table actg151F (g pour régions *g*) ;
- le troisième programme porte sur les calages nécessaires pour le calcul de l'estimateur 3, et aboutit à la table actgc3151F (c3 signifiant calage de type 3). Cette table est ensuite utilisée en entrée par un programme réalisant les calages nécessaires pour le calcul de l'estimateur 5, qu'on ne présente pas ici car il est proche du précédent ; il aboutit à la table actgc3c5151F (c5 signifiant calage de type 5) ;
- toutes ces tables sont utilisées par le quatrième programme présenté ici, qui calcule les différents indicateurs nécessaires à la comparaison des estimateurs retenus dans cette étude, dans le cas des estimateurs régionaux. Les calculs utilisés pour les estimateurs globaux en sont en effet assez proches et nous ne les présentons pas ici.

Programme de simulation des tirages

Bdscomnicr désigne la base de sondage utilisée pour le tirage de l'EAE-Commerce 2003/2002, base de sondage dans laquelle on a rajouté les tailles d'échantillon par tranche d'effectif réellement utilisées pour cette EAE, les n_i (notés *pni3* ici), et qu'on a restreinte à la partie sondée des activités retenues (Cf. paragraphes 2.1.2.1 et 2.1.2.3).

Teff2 désigne la variable de tranche d'effectif salarié (après regroupements de quelques modalités comme évoqué au paragraphe 2.1.2.2) ; *reg2* est la variable de région dans la nomenclature des ZEAT en 8 postes (Cf. paragraphe 2.1.2.2 également).

```
****************************************** ;
******* préparation de la base de sondage *******;
****************************************** ;

data bdscomnicr;
set stage.bdscomnicr;
run;

*** choix de l'activité *****;
%let activite=151F;

*** définition du taux de réponse uniforme :
Bih=Bi=B noté ici bih ***;
data activite;
set bdscomnicr;
if ape="&activite";
bih=0.8; /*fixé une fois pour toute*/
count=1;
run;

*** calcul des Nih notés ici gnih ***;
proc summary data=activite;
var count;
by teff2 reg2;
output out=activite2 sum=gnih;
run;

data activite;
merge activite activite2;
by teff2 reg2;
run;

*** calcul des Ni notés ici gni ***;
proc sort data=activite;
by teff2;
run;

proc summary data=activite;
var count;
by teff2;
output out=activite3 sum=gni;
run;

data activite;
merge activite activite3;
by teff2;
run;

*** calcul du N noté ici gn de l'activité ***;
proc summary data=activite;
var count;
output out=activite4 sum=gn;
run;
```

```
data _null_;
set activite4;
call symput ("gnact",gn);
run;

*** calcul des Pih (ici gpih) et des Pi (ici gpi) ***;
data activite;
set activite;
gn=&gnact;
gpih=gnih/gni;
gpi=gni/gn;
run;



****************************************;
********** "tirage" des échantillons ************;
****************************************;

**** méthode a) : TAS dans chaque i ***********;
**** ceci revient à simuler une loi multinomiale ***;
**** on le fait 100 fois **********************;

proc sort data=activite;
by teff2 reg2;
run;

*** on crée une table par tranche d'effectif ***;
data activ11;
set activite; by teff2 reg2; if teff2='00' and first.reg2;
drop siren isel codenq date poids qm cexan apen
nbetoa siege defenm defens effectif ind anca; run;
data activ12;
set activite; by teff2 reg2; if teff2='01' and first.reg2;
drop siren isel codenq date poids qm cexan apen
nbetoa siege defenm defens effectif ind anca; run;
data activ13;
set activite; by teff2 reg2; if teff2='03' and first.reg2;
drop siren isel codenq date poids qm cexan apen
nbetoa siege defenm defens effectif ind anca; run;
data activ14;
set activite; by teff2 reg2; if teff2='11' and first.reg2;
drop siren isel codenq date poids qm cexan apen
nbetoa siege defenm defens effectif ind anca; run;
data activ15;
set activite; by teff2 reg2; if teff2='12' and first.reg2;
drop siren isel codenq date poids qm cexan apen
nbetoa siege defenm defens effectif ind anca; run;
data activ16;
set activite; by teff2 reg2; if teff2='13' and first.reg2;
drop siren isel codenq date poids qm cexan apen
nbetoa siege defenm defens effectif ind anca; run;
```

```
%macro boucle;

%do i=1 %to 6;

%do j=1 %to 100;

%do h=1 %to 9;
data _null_;
set activ&j&i;
if reg2=&h;
call symput("gpi&h",gpih);
run;
%end;

data _null_;
set activ&j&i (obs=1);
call symput("pni3",pni3);
run;

data simula;
do k=1 to &pni3;
y=rantbl(0,&gpi1,&gpi2,&gpi3,&gpi4,&gpi5,&gpi7,
&gpi8,&gpi9); ** simulation de la multinomiale : la
variable y prend les valeurs _1, _2, …, _8 avec les
probabilités &gpi1, …, &gpi9  **;
output;
end;
keep y;
run;

proc freq data=simula;
tables y / noprint nopercent out=sortie ;
run; ** les nih sont les fréquences de y **;

proc transpose data=sortie out=sortie2;
id y;
var count;
run;

data sortie2;
set sortie2;
call symput("pni1",_1);
call symput("pni2",_2);
call symput("pni3",_3);
call symput("pni4",_4);
call symput("pni5",_5);
call symput("pni7",_6);
call symput("pni8",_7);
call symput("pni9",_8);
run;

%let k=%EVAL(&j+1);

data activ&k&i;
set activ&j&i;
if reg2=1 then pnih&j=&pni1;
else if reg2=2 then pnih&j=&pni2;
else if reg2=3 then pnih&j=&pni3;
else if reg2=4 then pnih&j=&pni4;
else if reg2=5 then pnih&j=&pni5;
else if reg2=7 then pnih&j=&pni7;
else if reg2=8 then pnih&j=&pni8;
else if reg2=9 then pnih&j=&pni9;
run;

%end;

%end;

%mend;

%boucle;
run;

data activ2;
set activ1011 activ1012 activ1013 activ1014
activ1015 activ1016;
run;



**** méthode b) : TAS stratifié par région ****;
**** tirage systématique avec pas non entier ****;
**** on le fait 5 fois **********************;

data activite;
set activite;
poids2=gni/pni3;
run;

proc sort data=activite;
by teff2 reg2;
run;

*** on crée une table par tranche d'effectif ***;
data activite11;
set activite; if teff2='00'; rang=_N_; run;
data activite12;
set activite; if teff2='01'; rang=_N_; run;
data activite13;
set activite; if teff2='03'; rang=_N_; run;
data activite14;
set activite; if teff2='11'; rang=_N_; run;
data activite15;
set activite; if teff2='12'; rang=_N_; run;
data activite16;
set activite; if teff2='13'; rang=_N_; run;


%macro tirpoids;

%do j=1 %to 5;
```

```
%do i=1 %to 6;

data _null_;
set activite&j&i;
alea=1+int(ranuni(0)*gni/pni3);
call symput ("alea", alea);
run;

%let k=%EVAL(&j+1);

data activite&k&i;
set activite&j&i;
if 0<=mod(rang-&alea,gni/pni3)<1 then echp&j=1;
else echp&j=0;
run;

%end;

%end;

%mend ;

%tirpoids;
run;


data activite2;
set activite61 activite62 activite63 activite64
activite65 activite66;
run;

*** calcul des nih avec cette 2ème méthode de tirage
des échantillons ***;
proc sort data=activite2;
by teff2 reg2;
run;

proc summary data=activite2;
var echp1 echp2 echp3 echp4 echp5;
by teff2 reg2;
output out=tailleech2 sum=pnihp1 pnihp2 pnihp3
pnihp4 pnihp5;
run;

data activite3;
merge activite tailleech2;
by teff2 reg2;
drop _TYPE_ _FREQ_;
run;

data activ3;
set activite3;
by teff2 reg2;
if first.reg2;
drop siren isel codenq date poids qm cexan apen
nbetoa siege defenm defens effectif ind anca;
run;


**** réunion des résultats des deux méthodes de
simulation ****;
data activ;
merge activ2 activ3;
by teff2 reg2;
run;



****************************************;
******** tirage des répondants ***************;
****************************************;

data activR1;
set activ;
run;

%macro tirrep;

%do j=1 %to 100;

%let k=%EVAL(&j+1);

data activR&k;
set activR&j;
if pnih&j=. then pnihR&j=0;
else pnihR&j=ranbin(0,pnih&j,bih);
run; **utilisation d'un générateur de loi binomiale**;

%end;

%mend;

%tirrep;
run;


%macro tirrepp;

%let t=%EVAL(101);

%let ind=1;

%do j=1 %to 5;

%do l=1 %to 20;

%let k=%EVAL(&t+1);

data activR&k;
set activR&t;
```

```
pnihpR&ind=ranbin(0,pnihp&j,bih);
run;

%let t=%EVAL(&k);

%let ind=%EVAL(&ind+1);

%end;

%end;

%mend;

%tirrepp;
run;


****** sauvegarde des résultats des simulation et des tirages des répondants ******;

data stage.act&activite;
set activR201;
run;
```

```
***** on ajoute finalement les moyennes et variances des variables d'intérêt tirés de SUSE *****;

data tabsusec2;
set stage.tabsusec2;
if ape="&activite";
drop temp temp2;
run;

data act&activite;
set stage.act&activite;
drop Ebarihs CAbarihs MAbarihs VAbarihs S2Eihs S2CAihs S2MAihs S2VAihs grandNihs;
run;

data stage.act&activite;
merge act&activite (in=ina) tabsusec2;
by ape teff2 reg2;
if ina;
run;
```

Regroupements des régions h en régions g contenant un nombre suffisant de répondants

La macro SAS suivante génère une variable *g* (respectivement *pg*) pour chaque échantillon simulé selon la 1ère méthode (respectivement selon la 2ème méthode). Cette variable est par la suite utilisée comme indicateur du niveau de nomenclature de régions agrégée qu'il faut utiliser pour chaque échantillon.

```
*** à faire pour chaque activité ***;

%let activite=151F;

** on définit le nombre minimal de répondants qu'on
veut avoir dans les regroupements de régions **;
%let min=5;

data regroup1;
set stage.act&activite;
run;

proc sort data=regroup1;
by teff2 reg2;
run;

data regroup1;
set regroup1;
if reg2 in (1,2) then temp=12;
if reg2 in (3,4) then temp=34;
if reg2 in (5,7) then temp=57;
if reg2 in (8,9) then temp=89;
if reg2 in (1,2,3,4) then temp2=1234;
if reg2 in (5,7,8,9) then temp2=5789;
run;

proc sort data=regroup1;
by teff2 temp;
run;

proc summary data=regroup1;
var pnihR1-pnihR100 pnihpR1-pnihpR100;
by teff2 temp;
output out=regroup2 min=min_pnihR1-
min_pnihR100 min_pnihpR1-min_pnihpR100;
run;

data regroup1;
merge regroup1 regroup2;
by teff2 temp;
drop _TYPE_ _FREQ_;
run;

proc summary data=regroup1;
var pnihR1-pnihR100 pnihpR1-pnihpR100;
by teff2 temp;
output out=regroup3 sum=sum_pnihR1-
sum_pnihR100 sum_pnihpR1-sum_pnihpR100;
run;

data regroup3;
set regroup3;
if temp in (12,34) then temp2=1234;
if temp in (57,89) then temp2=5789;
run;

proc summary data=regroup3;
var sum_pnihR1-sum_pnihR100 sum_pnihpR1-
sum_pnihpR100;
by teff2 temp2;
output out=regroup4 min=min_pnigR1-
min_pnigR100 min_pnigpR1-min_pnigpR100;
run;

data regroup1;
merge regroup1 regroup4;
by teff2 temp2;
drop _TYPE_ _FREQ_;
run;

proc summary data=regroup1;
var pnihR1-pnihR100 pnihpR1-pnihpR100;
by teff2 temp2;
output out=regroup5 sum=sum_pnigR1-
sum_pnigR100 sum_pnigpR1-sum_pnigpR100;
run;

proc summary data=regroup5;
var sum_pnigR1-sum_pnigR100 sum_pnigpR1-
sum_pnigpR100;
by teff2;
output out=regroup6 min=min_pniggR1-
min_pniggR100 min_pniggpR1-min_pniggpR100;
run;

data regroup1;
merge regroup1 regroup6;
by teff2;
drop _TYPE_ _FREQ_;
run;
```

```
%macro boucleg;

%do j=1 %to 100;

%let k=%EVAL(&j+1);

data regroup&k;
set regroup&j;
if min_pniggR&j<&min then g&j=16;
else if reg2 in (1,2,3,4) and (min_pnigR&j<&min)
then g&j=14;
else if reg2 in (5,7,8,9) and (min_pnigR&j<&min)
then g&j=15;
else if reg2 in (1,2) and (min_pnihR&j<&min) then
g&j=10;
else if reg2 in (3,4) and (min_pnihR&j<&min) then
g&j=11;
else if reg2 in (5,7) and (min_pnihR&j<&min) then
g&j=12;
else if reg2 in (8,9) and (min_pnihR&j<&min) then
g&j=13;
else g&j=reg2;

if min_pniggpR&j<&min then gp&j=16;
else if reg2 in (1,2,3,4) and (min_pnigpR&j<&min)
then gp&j=14;
else if reg2 in (5,7,8,9) and (min_pnigpR&j<&min)
then gp&j=15;
else if reg2 in (1,2) and (min_pnihpR&j<&min) then
gp&j=10;
else if reg2 in (3,4) and (min_pnihpR&j<&min) then
gp&j=11;
else if reg2 in (5,7) and (min_pnihpR&j<&min) then
gp&j=12;
else if reg2 in (8,9) and (min_pnihpR&j<&min) then
gp&j=13;
else gp&j=reg2;
run;

%end;

%mend;

%boucleg;
run;

data regroup101;
set regroup101;
drop temp temp2 min_pnigR1-min_pnigR100
min_pnihR1-min_pnihR100
min_pniggR1-min_pniggR100 min_pnigpR1-
min_pnigpR100 min_pnihpR1-min_pnihpR100
min_pniggpR1-min_pniggpR100;
run;


*** sauvegarde des résultats dans la librairie ***;

data stage.actg&activite;
set regroup101;
run;
```

Calage pour l'estimateur 3

On présente ici le programme utilisé pour réaliser les calages nécessaires au calcul de l'estimateur 3 (calage du nombre de répondants dans les croisements *ig* (n_{igR}) sur les comptages obtenus pour l'échantillon aux niveaux des tranches d'effectif *i* et des supra-régions *g* (n_i et n_g)). Le programme donnant les calages réalisés pour le calcul de l'estimateur 5 n'est pas présenté ici car son principe est assez proche de celui du programme qui suit.

```
*** à faire pour chaque activité ***;

%let activite=151F;

data act;
set stage.actg&activite;
run;


*** marge pour le premier critère teff2 ***;

proc sort data=act;
by teff2;
run;

proc summary data=act;
var pni3;
by teff2;
output out=marge1 min=pni3;
run;

data marge1;
set marge1;
drop _TYPE_ _FREQ_;
run;

proc transpose data=marge1 out=marge1;
run;

data marge1;
set marge1;
rename _name_=var col1=mar1 col2=mar2
col3=mar3 col4=mar4 col5=mar5
col6=mar6;
ape="&activite";
if ape in ('151F' '158C' '502Z') then N=6;
else N=4;
drop ape;
run;

data marge1;
retain var N mar1 mar2 mar3 mar4 mar5 mar6 ;
set marge1;
run;
```

```
*** macro de calage de type 3 ***;

%macro calage3;

%do e=1 %to 100;

%let l=%eval(%sysfunc(int(%sysevalf((&e-
1)/20)))+1);

*** pour la table en entrée de la macro CALMAR, on
a besoin de récupérer des variables indicatrices dont
les modalités sont numérotées de 1 à … ***;

proc freq data=act;
tables teff2 / noprint out=marg1;
run;

data marg1;
set marg1;
x=_N_;
keep teff2 x;
run;

proc freq data=act;
tables g&e / noprint out=marg2;
run;

data marg2;
set marg2;
y=_N_;
keep g&e y;
run;

proc freq data=act;
tables gp&e / noprint out=margp2;
run;

data margp2;
set margp2;
y=_N_;
keep gp&e y;
run;
```

```
**** cas des échantillons de la méthode a) ****;

** obtention des n_igR **;
proc sort data=act;
by teff2 g&e;
run;
proc summary data=act;
var pnihR&e;
by teff2 g&e;
output out=nigR sum=pnigR&e;
run;

** obtention des n_g (les n_i sont les pni3) **;
proc sort data=act;
by g&e;
run;
proc summary data=act;
var pnih&e;
by g&e;
output out=ng sum=png&e;
run;

** marge pour le second critère g&e **;
data ng;
set ng;
count=_N_;
drop _TYPE_ _FREQ_;
run;

proc summary data=ng;
var count;
output out=N max=N;
run;
data _null_;
set N;
call symput ("N",N);
run;
data ng;
set ng;
drop count g&e;
run;

proc transpose data=ng out=marge2t;
run;
data marge2t;
set marge2t;
rename _name_=var col1=mar1 col2=mar2
col3=mar3 col4=mar4 col5=mar5
col6=mar6 col7=mar7 col8=mar8 col9=mar9
col10=mar10 col11=mar11
col12=mar12 col13=mar13 col14=mar14
col15=mar15;
N=&N;
run;
```

```
data marge2t;
retain var N mar1 mar2 mar3 mar4 mar5 mar6 mar7
mar8 mar9 mar10 mar11
mar12 mar13 mar14 mar15;
set marge2t;
run;

** mise en forme de la table en entrée de la macro
CALMAR **;

data tab;
set nigR;
identi=_N_;
keep identi teff2 g&e pnigR&e;
run;

proc sort data=tab; by teff2; run;
data tab2;
merge tab (in=ina) marg1;
by teff2;
if ina;
keep identi x g&e pnigR&e;
run;
proc sort data=tab2; by g&e; run;
data tab3;
merge tab2 (in=ina) marg2;
by g&e;
if ina;
keep identi x y pnigR&e;
run;
data tab3;
retain identi x y pnigR&e;
set tab3;
run;

proc sort data=tab3;
by x y;
run;

data marge2;
set marge2t;
run;

data marges;
set marge1 marge2;
if var='pni3' then var='x';
else var='y';
run;

** et on lance finalement la macro CALMAR **;

%calmar(data=tab3,poids=pnigR&e,ident=identi,
datamar=marges,m=2,editpoi=oui,obseli=oui,
datapoi=sortie&e,poidsfin=pnigtilde&e);
run;
```

```
**** cas des échantillons de la méthode b) ****;
**** même chose avec gp&e à la place de g&e ****;

** obtention des n_igR **;
proc sort data=act;
by teff2 gp&e;
run;
proc summary data=act;
var pnihpR&e;
by teff2 gp&e;
output out=nigR sum=pnigpR&e;
run;

** obtention des n_g (les n_i sont les pni3) **;
proc sort data=act;
by gp&e;
run;
proc summary data=act;
var pnihp&l;
by gp&e;
output out=ng sum=pngp&e;
run;

** marge pour le second critère gp&e **;
data ng;
set ng;
count=_N_;
drop _TYPE_ _FREQ_;
run;

proc summary data=ng;
var count;
output out=N max=N;
run;
data _null_;
set N;
call symput ("N",N);
run;
data ng;
set ng;
drop count gp&e;
run;

proc transpose data=ng out=marge2t;
run;
data marge2t;
set marge2t;
rename _name_=var col1=mar1 col2=mar2
col3=mar3 col4=mar4 col5=mar5
col6=mar6 col7=mar7 col8=mar8 col9=mar9
col10=mar10 col11=mar11
col12=mar12 col13=mar13 col14=mar14
col15=mar15;
N=&N;
run;

data marge2t;
retain var N mar1 mar2 mar3 mar4 mar5 mar6 mar7
mar8 mar9 mar10 mar11
mar12 mar13 mar14 mar15;
set marge2t;
run;

*mise en forme de la table en entrée de la macro
CALMAR *;

data tabp;
set nigR;
identi=_N_;
keep identi teff2 gp&e pnigpR&e;
run;

proc sort data=tabp; by teff2; run;
data tabp2;
merge tabp (in=ina) marg1;
by teff2;
if ina;
keep identi x gp&e pnigpR&e;
run;
proc sort data=tabp2; by gp&e; run;
data tabp3;
merge tabp2 (in=ina) margp2;
by gp&e;
if ina;
keep identi x y pnigpR&e;
run;
data tabp3;
retain identi x y pnigpR&e;
set tabp3;
run;

proc sort data=tabp3;
by x y;
run;

data marge2;
set marge2t;
run;

data marges;
set marge1 marge2;
if var='pni3' then var='x';
else var='y';
run;

** et on lance finalement la macro CALMAR **;

%calmar(data=tabp3,poids=pnigpR&e,ident=identi,
datamar=marges,m=2,editpoi=oui,obseli=oui,
datapoi=sortiep&e,poidsfin=pnigptilde&e);
run;
```

```
*** récupération des résultats ***;

data tab3;
merge tab3 sortie&e;
by identi;
drop identi;
run;

data tab3;
merge tab3 marg1;
by x;
drop x;
run;

proc sort data=tab3; by y; run;
data tab3;
merge tab3 marg2;
by y;
drop y pnigR&e;
run;


*même chose pour les échantillons de la méthode b)*;

data tabp3;
merge tabp3 sortiep&e;
by identi;
drop identi;
run;

data tabp3;
merge tabp3 marg1;
by x;
drop x;
run;

proc sort data=tabp3; by y; run;
data tabp3;
merge tabp3 margp2;
by y;
drop y pnigpR&e;
run;

*** et on réunit les résultats obtenus pour les
échantillons des deux méthodes ***;

proc sort data=tab3; by teff2 g&e; run;
proc sort data=tabp3; by teff2 gp&e; run;

proc sort data=act; by teff2 g&e; run;
data act;
merge act tab3;
by teff2 g&e;
run;

proc sort data=act; by teff2 gp&e; run;
data act;
merge act tabp3;
by teff2 gp&e;
run;

%end;

%mend;



*** lancement de la macro précédente ***;

%calage3;
run;

*** et sauvegarde dans la librairie de travail ***;

data stage.actgc3&activite;
set act;
run;
```

Calcul des performances des estimateurs régionaux

```
*** à faire pour chaque activité ***;

%let activite=151F;


****************************************;
***** estimateur 1 : moyenne des répondants *****;
****************************************;

*** calcul des biais et variances au niveau ih ***;

%macro estim1;

data est11;
set stage.act&activite;
gni2=gni*gni;
run;

%do j=1 %to 100;

%let k=%EVAL(&j+1);

proc summary data=est1&j;
var pnihR&j pnihpR&j;
by teff2;
output out=pniR sum=pniR&j pnipR&j;
run;

data est1&j;
merge est1&j pniR;
by teff2;
run;


data est1&k;
set est1&j;

biaisih1E&j=(pnihR&j/pniR&j-gpih)*Ebarihs;
biaisih1CA&j=(pnihR&j/pniR&j-gpih)*CAbarihs;
biaisih1MA&j=(pnihR&j/pniR&j-gpih)*MAbarihs;
biaisih1VA&j=(pnihR&j/pniR&j-gpih)*VAbarihs;
if pnihR&j NE 0 then do;
varih1E&j=(pnihR&j/pniR&j)*(pnihR&j/pniR&j)*
S2Eihs/pnihR&j;
varih1CA&j=(pnihR&j/pniR&j)*(pnihR&j/pniR&j)*
S2CAihs/pnihR&j;
varih1MA&j=(pnihR&j/pniR&j)*(pnihR&j/pniR&j)*
S2MAihs/pnihR&j;
varih1VA&j=(pnihR&j/pniR&j)*(pnihR&j/pniR&j)*
S2VAihs/pnihR&j;
end;
else do;
varih1E&j=0; varih1CA&j=0;
varih1MA&j=0; varih1VA&j=0;
end;

biaisih1pE&j=(pnihpR&j/pnipR&j-gpih)*Ebarihs;
biaisih1pCA&j=(pnihpR&j/pnipR&j-gpih)
*CAbarihs;
biaisih1pMA&j=(pnihpR&j/pnipR&j-gpih)
*MAbarihs;
biaisih1pVA&j=(pnihpR&j/pnipR&j-gpih)
*VAbarihs;
if pnihpR&j NE 0 then do;
varih1pE&j=(pnihpR&j/pnipR&j)*
(pnihpR&j/pnipR&j)*S2Eihs/pnihpR&j;
varih1pCA&j=(pnihpR&j/pnipR&j)*
(pnihpR&j/pnipR&j)*S2CAihs/pnihpR&j;
varih1pMA&j=(pnihpR&j/pnipR&j)*
(pnihpR&j/pnipR&j)*S2MAihs/pnihpR&j;
varih1pVA&j=(pnihpR&j/pnipR&j)*
(pnihpR&j/pnipR&j)*S2VAihs/pnihpR&j;
end;
else do;
varih1pE&j=0; varih1pCA&j=0;
varih1pMA&j=0; varih1pVA&j=0;
end;

%end;

%mend;

%estim1;
run;


*** agrégation au niveau régional h ***;
*** (différente pour les estimations globales) ***;

proc sort data=est1101;
by reg2;
run;

proc summary data=est1101;
var biaisih1E1-biaisih1E100 biaisih1CA1-
biaisih1CA100 biaisih1MA1-biaisih1MA100
biaisih1VA1-biaisih1VA100
biaisih1pE1-biaisih1pE100 biaisih1pCA1-
biaisih1pCA100 biaisih1pMA1-biaisih1pMA100
biaisih1pVA1-biaisih1pVA100;
```

```
by reg2;
weight gni;
output out=est1 sum=biais1E1-biais1E100
biais1CA1-biais1CA100
biais1MA1-biais1MA100 biais1VA1-biais1VA100
biais1pE1-biais1pE100 biais1pCA1-biais1pCA100
biais1pMA1-biais1pMA100
biais1pVA1-biais1pVA100;
run;

proc summary data=est1101;
var varih1E1-varih1E100 varih1CA1-varih1CA100
varih1MA1-varih1MA100 varih1VA1-varih1VA100
varih1pE1-varih1pE100 varih1pCA1-varih1pCA100
varih1pMA1-varih1pMA100
varih1pVA1-varih1pVA100;
by reg2;
weight gni2;
output out=est1v sum=var1E1-var1E100 var1CA1-
var1CA100 var1MA1-var1MA100 var1VA1-
var1VA100 var1pE1-var1pE100 var1pCA1-
var1pCA100 var1pMA1-var1pMA100
var1pVA1-var1pVA100;
run;

data est1;
merge est1 est1v;
by reg2;
run;


*** calcul des biais au carré, EQM, RMSE, REL ***;

%macro estim1b;

data est11;
set est1;
run;

%do j=1 %to 100;

%let k=%EVAL(&j+1);

data est1&k;
set est1&j;
biaiscar1E&j=biais1E&j*biais1E&j;
eqm1E&j=biais1E&j*biais1E&j+var1E&j;
rmse1E&j=sqrt(eqm1E&j);
rel1E&j=0;
biaiscar1CA&j=biais1CA&j*biais1CA&j;
eqm1CA&j=biais1CA&j*biais1CA&j+var1CA&j;
rmse1CA&j=sqrt(eqm1CA&j);
rel1CA&j=0;
biaiscar1MA&j=biais1MA&j*biais1MA&j;
eqm1MA&j=biais1MA&j*biais1MA&j+var1MA&j;
rmse1MA&j=sqrt(eqm1MA&j);
rel1MA&j=0;
biaiscar1VA&j=biais1VA&j*biais1VA&j;
eqm1VA&j=biais1VA&j*biais1VA&j+var1VA&j;
rmse1VA&j=sqrt(eqm1VA&j);
rel1VA&j=0;

biaiscar1pE&j=biais1pE&j*biais1pE&j;
eqm1pE&j=biais1pE&j*biais1pE&j+var1pE&j;
rmse1pE&j=sqrt(eqm1pE&j);
rel1pE&j=0;
biaiscar1pCA&j=biais1pCA&j*biais1pCA&j;
eqm1pCA&j=biais1pCA&j*biais1pCA&j
+var1pCA&j;
rmse1pCA&j=sqrt(eqm1pCA&j);
rel1pCA&j=0;
biaiscar1pMA&j=biais1pMA&j*biais1pMA&j;
eqm1pMA&j=biais1pMA&j*biais1pMA&j
+var1pMA&j;
rmse1pMA&j=sqrt(eqm1pMA&j);
rel1pMA&j=0;
biaiscar1pVA&j=biais1pVA&j*biais1pVA&j;
eqm1pVA&j=biais1pVA&j*biais1pVA&j
+var1pVA&j;
rmse1pVA&j=sqrt(eqm1pVA&j);
rel1pVA&j=0;

%end;

%mend;

%estim1b;
run;


*** calcul des moyennes sur les 100 échantillons ***;

data est1;
set est1101;
biais1E=mean(of biais1E1-biais1E100);
biais1CA=mean(of biais1CA1-biais1CA100);
biais1MA=mean(of biais1MA1-biais1MA100);
biais1VA=mean(of biais1VA1-biais1VA100);
biaiscar1E=mean(of biaiscar1E1-biaiscar1E100);
biaiscar1CA=mean(of biaiscar1CA1-
biaiscar1CA100);
biaiscar1MA=mean(of biaiscar1MA1-
biaiscar1MA100);
biaiscar1VA=mean(of biaiscar1VA1-
biaiscar1VA100);
var1E=mean(of var1E1-var1E100);
var1CA=mean(of var1CA1-var1CA100);
var1MA=mean(of var1MA1-var1MA100);
var1VA=mean(of var1VA1-var1VA100);
rmse1E=mean(of rmse1E1-rmse1E100);
rmse1CA=mean(of rmse1CA1-rmse1CA100);
rmse1MA=mean(of rmse1MA1-rmse1MA100);
```

```
rmse1VA=mean(of rmse1VA1-rmse1VA100);
rel1E=mean(of rel1E1-rel1E100);
rel1CA=mean(of rel1CA1-rel1CA100);
rel1MA=mean(of rel1MA1-rel1MA100);
rel1VA=mean(of rel1VA1-rel1VA100);

biais1pE=mean(of biais1pE1-biais1pE100);
biais1pCA=mean(of biais1pCA1-biais1pCA100);
biais1pMA=mean(of biais1pMA1-biais1pMA100);
biais1pVA=mean(of biais1pVA1-biais1pVA100);
biaiscar1pE=mean(of biaiscar1pE1-biaiscar1pE100);
biaiscar1pCA=mean(of biaiscar1pCA1-
biaiscar1pCA100);
biaiscar1pMA=mean(of biaiscar1pMA1-
biaiscar1pMA100);
biaiscar1pVA=mean(of biaiscar1pVA1-
biaiscar1pVA100);
var1pE=mean(of var1pE1-var1pE100);
var1pCA=mean(of var1pCA1-var1pCA100);
var1pMA=mean(of var1pMA1-var1pMA100);
var1pVA=mean(of var1pVA1-var1pVA100);
rmse1pE=mean(of rmse1pE1-rmse1pE100);
rmse1pCA=mean(of rmse1pCA1-rmse1pCA100);
rmse1pMA=mean(of rmse1pMA1-rmse1pMA100);
rmse1pVA=mean(of rmse1pVA1-rmse1pVA100);
rel1pE=mean(of rel1pE1-rel1pE100);
rel1pCA=mean(of rel1pCA1-rel1pCA100);
rel1pMA=mean(of rel1pMA1-rel1pMA100);
rel1pVA=mean(of rel1pVA1-rel1pVA100);

keep reg2 biais1E biais1CA biais1MA biais1VA
biaiscar1E biaiscar1CA biaiscar1MA biaiscar1VA
var1E var1CA var1MA var1VA
rmse1E rmse1CA rmse1MA rmse1VA
rel1E rel1CA rel1MA rel1VA
biais1pE biais1pCA biais1pMA biais1pVA
biaiscar1pE biaiscar1pCA biaiscar1pMA
biaiscar1pVA var1pE var1pCA var1pMA var1pVA
rmse1pE rmse1pCA rmse1pMA rmse1pVA
rel1pE rel1pCA rel1pMA rel1pVA;
run;


*** mise en forme ***;

data est1bis;
set est1;
estimateur='est_reg1';
regroupements='non';
activite="&activite";
txrep=0.8;
rename reg2=region biais1E=biaishE
biais1CA=biaishCA biais1MA=biaishMA
biais1VA=biaishVA biaiscar1E=biaiscarhE
biaiscar1CA=biaiscarhCA
biaiscar1MA=biaiscarhMA
biaiscar1VA=biaiscarhVA var1E=varhE
var1CA=varhCA var1MA=varhMA
var1VA=varhVA
rmse1E=rmsehE rmse1CA=rmsehCA
rmse1MA=rmsehMA rmse1VA=rmsehVA
rel1E=relhE rel1CA=relhCA rel1MA=relhMA
rel1VA=relhVA biais1pE=biaishpE
biais1pCA=biaishpCA biais1pMA=biaishpMA
biais1pVA=biaishpVA biaiscar1pE=biaiscarhpE
biaiscar1pCA=biaiscarhpCA
biaiscar1pMA=biaiscarhpMA
biaiscar1pVA=biaiscarhpVA var1pE=varhpE
var1pCA=varhpCA var1pMA=varhpMA
var1pVA=varhpVA rmse1pE=rmsehpE
rmse1pCA=rmsehpCA rmse1pMA=rmsehpMA
rmse1pVA=rmsehpVA rel1pE=relhpE
rel1pCA=relhpCA rel1pMA=relhpMA
rel1pVA=relhpVA;
run;

data stage.result1_reg&activite;
retain estimateur activite txrep regroupements region
biaishE biaishCA biaishMA biaishVA biaishpE
biaishpCA biaishpMA biaishpVA
biaiscarhE biaiscarhCA biaiscarhMA biaiscarhVA
biaiscarhpE biaiscarhpCA biaiscarhpMA
biaiscarhpVA
varhE varhCA varhMA varhVA varhpE varhpCA
varhpMA varhpVA
rmsehE rmsehCA rmsehMA rmsehVA rmsehpE
rmsehpCA rmsehpMA rmsehpVA
relhE relhCA relhMA relhVA relhpE relhpCA
relhpMA relhpVA;
set est1bis;
run;
```

```
*****************************************;
******** estimateur 2 : moyenne ajustée *********;
*****************************************;

*** calcul des biais et variances au niveau ih ***;

%macro estim2;

data est21;
set stage.actg&activite;
gni2=gni*gni;
run;


%do j=1 %to 100;

%let k=%EVAL(&j+1);
%let l=%eval(%sysfunc(int(%sysevalf((&j-
1)/20)))+1);

proc sort data=est2&j;
by teff2 g&j;
run;

proc summary data=est2&j;
var pnih&j pnihR&j ;
by teff2 g&j;
output out=nivg sum=pnig&j pnigR&j;
run;

data est2&j;
merge est2&j nivg;
by teff2 g&j;
run;

proc sort data=est2&j;
by teff2 gp&j;
run;

proc summary data=est2&j;
var pnihp&l pnihpR&j;
by teff2 gp&j;
output out=nivgp sum=pnigp&j pnigpR&j;
run;

data est2&j;
merge est2&j nivgp;
by teff2 gp&j;
run;


data est2&k;
set est2&j;

biaisih2E&j=((pnig&j/pni3)*(pnihR&j/pnigR&j)
-gpih)*Ebarihs;
biaisih2CA&j=((pnig&j/pni3)*(pnihR&j/pnigR&j)
-gpih)*CAbarihs;
biaisih2MA&j=((pnig&j/pni3)*(pnihR&j/pnigR&j)
-gpih)*MAbarihs;
biaisih2VA&j=((pnig&j/pni3)*(pnihR&j/pnigR&j)
-gpih)*VAbarihs;
if pnihR&j NE 0 then do;
varih2E&j=(pnig&j/pni3)*(pnig&j/pni3)*(pnihR&j
/pnigR&j)*(pnihR&j/pnigR&j)*S2Eihs/pnihR&j;
varih2CA&j=(pnig&j/pni3)*(pnig&j/pni3)*(pnihR&j
/pnigR&j)*(pnihR&j/pnigR&j)*S2CAihs/pnihR&j;
varih2MA&j=(pnig&j/pni3)*(pnig&j/pni3)*(pnihR&j
/pnigR&j)*(pnihR&j/pnigR&j)*S2MAihs/pnihR&j;
varih2VA&j=(pnig&j/pni3)*(pnig&j/pni3)*(pnihR&j
/pnigR&j)*(pnihR&j/pnigR&j)*S2VAihs/pnihR&j;
end;
else do;
varih2E&j=0; varih2CA&j=0;
varih2MA&j=0; varih2VA&j=0;
end;

biaisih2pE&j=((pnigp&j/pni3)*(pnihpR&j/pnigpR&j)
-gpih)*Ebarihs;
biaisih2pCA&j=((pnigp&j/pni3)*(pnihpR&j
/pnigpR&j)-gpih)*CAbarihs;
biaisih2pMA&j=((pnigp&j/pni3)*(pnihpR&j
/pnigpR&j)-gpih)*MAbarihs;
biaisih2pVA&j=((pnigp&j/pni3)*(pnihpR&j
/pnigpR&j)-gpih)*VAbarihs;
if pnihpR&j NE 0 then do;
varih2pE&j=(pnigp&j/pni3)*(pnigp&j/pni3)
*(pnihpR&j/pnigpR&j)*(pnihpR&j/pnigpR&j)
*S2Eihs/pnihpR&j;
varih2pCA&j=(pnigp&j/pni3)*(pnigp&j/pni3)
*(pnihpR&j/pnigpR&j)*(pnihpR&j/pnigpR&j)
*S2CAihs/pnihpR&j;
varih2pMA&j=(pnigp&j/pni3)*(pnigp&j/pni3)
*(pnihpR&j/pnigpR&j)*(pnihpR&j/pnigpR&j)
*S2MAihs/pnihpR&j;
varih2pVA&j=(pnigp&j/pni3)*(pnigp&j/pni3)
*(pnihpR&j/pnigpR&j)*(pnihpR&j/pnigpR&j)
*S2VAihs/pnihpR&j;
end;
else do;
varih2pE&j=0; varih2pCA&j=0;
varih2pMA&j=0; varih2pVA&j=0;
end;

%end;

%mend;

%estim2;
run;
```

```
*** agrégation au niveau régional h ***;

data est2101;
set est2101;
drop _TYPE_ _FREQ_;
run;

proc sort data=est2101;
by reg2;
run;

proc summary data=est2101;
var biaisih2E1-biaisih2E100 biaisih2CA1-
biaisih2CA100 biaisih2MA1-biaisih2MA100
biaisih2VA1-biaisih2VA100
biaisih2pE1-biaisih2pE100 biaisih2pCA1-
biaisih2pCA100 biaisih2pMA1-biaisih2pMA100
biaisih2pVA1-biaisih2pVA100;
by reg2;
weight gni;
output out=est2 sum=biais2E1-biais2E100
biais2CA1-biais2CA100
biais2MA1-biais2MA100 biais2VA1-biais2VA100
biais2pE1-biais2pE100 biais2pCA1-biais2pCA100
biais2pMA1-biais2pMA100
biais2pVA1-biais2pVA100;
run;

proc summary data=est2101;
var varih2E1-varih2E100 varih2CA1-varih2CA100
varih2MA1-varih2MA100 varih2VA1-varih2VA100
varih2pE1-varih2pE100 varih2pCA1-varih2pCA100
varih2pMA1-varih2pMA100 varih2pVA1-
varih2pVA100;
by reg2;
weight gni2;
output out=est2v sum=var2E1-var2E100 var2CA1-
var2CA100 var2MA1-var2MA100 var2VA1-
var2VA100 var2pE1-var2pE100 var2pCA1-
var2pCA100 var2pMA1-var2pMA100 var2pVA1-
var2pVA100;
run;

data est2;
merge est2 est2v est1101; *pour avoir le RMSE de
l'estimateur 1 pour le calcul du REL*;
by reg2;
run;


*** calcul des biais au carré, EQM, RMSE, REL ***;

%macro estim2b;

data est21;
set est2; run;


%do j=1 %to 100;

%let k=%EVAL(&j+1);

data est2&k;
set est2&j;
biaiscar2E&j=biais2E&j*biais2E&j;
eqm2E&j=biais2E&j*biais2E&j+var2E&j;
rmse2E&j=sqrt(eqm2E&j);
rel2E&j=100*((rmse2E&j/rmse1E&j)-1);
biaiscar2CA&j=biais2CA&j*biais2CA&j;
eqm2CA&j=biais2CA&j*biais2CA&j+var2CA&j;
rmse2CA&j=sqrt(eqm2CA&j);
rel2CA&j=100*((rmse2CA&j/rmse1CA&j)-1);
biaiscar2MA&j=biais2MA&j*biais2MA&j;
eqm2MA&j=biais2MA&j*biais2MA&j+var2MA&j;
rmse2MA&j=sqrt(eqm2MA&j);
rel2MA&j=100*((rmse2MA&j/rmse1MA&j)-1);
biaiscar2VA&j=biais2VA&j*biais2VA&j;
eqm2VA&j=biais2VA&j*biais2VA&j+var2VA&j;
rmse2VA&j=sqrt(eqm2VA&j);
rel2VA&j=100*((rmse2VA&j/rmse1VA&j)-1);

biaiscar2pE&j=biais2pE&j*biais2pE&j;
eqm2pE&j=biais2pE&j*biais2pE&j+var2pE&j;
rmse2pE&j=sqrt(eqm2pE&j);
rel2pE&j=100*((rmse2pE&j/rmse1pE&j)-1);
biaiscar2pCA&j=biais2pCA&j*biais2pCA&j;
eqm2pCA&j=biais2pCA&j*biais2pCA&j
+var2pCA&j;
rmse2pCA&j=sqrt(eqm2pCA&j);
rel2pCA&j=100*((rmse2pCA&j/rmse1pCA&j)-1);
biaiscar2pMA&j=biais2pMA&j*biais2pMA&j;
eqm2pMA&j=biais2pMA&j*biais2pMA&j
+var2pMA&j;
rmse2pMA&j=sqrt(eqm2pMA&j);
rel2pMA&j=100*((rmse2pMA&j/rmse1pMA&j)-1);
biaiscar2pVA&j=biais2pVA&j*biais2pVA&j;
eqm2pVA&j=biais2pVA&j*biais2pVA&j
+var2pVA&j;
rmse2pVA&j=sqrt(eqm2pVA&j);
rel2pVA&j=100*((rmse2pVA&j/rmse1pVA&j)-1);

%end;

%mend;

%estim2b;
run;


*** calcul des moyennes sur les 100 échantillons ***;

data est2;
set est2101;
```

```
biais2E=mean(of biais2E1-biais2E100);
biais2CA=mean(of biais2CA1-biais2CA100);
biais2MA=mean(of biais2MA1-biais2MA100);
biais2VA=mean(of biais2VA1-biais2VA100);
biaiscar2E=mean(of biaiscar2E1-biaiscar2E100);
biaiscar2CA=mean(of biaiscar2CA1-
biaiscar2CA100);
biaiscar2MA=mean(of biaiscar2MA1-
biaiscar2MA100);
biaiscar2VA=mean(of biaiscar2VA1-
biaiscar2VA100);
var2E=mean(of var2E1-var2E100);
var2CA=mean(of var2CA1-var2CA100);
var2MA=mean(of var2MA1-var2MA100);
var2VA=mean(of var2VA1-var2VA100);
rmse2E=mean(of rmse2E1-rmse2E100);
rmse2CA=mean(of rmse2CA1-rmse2CA100);
rmse2MA=mean(of rmse2MA1-rmse2MA100);
rmse2VA=mean(of rmse2VA1-rmse2VA100);
rel2E=mean(of rel2E1-rel2E100);
rel2CA=mean(of rel2CA1-rel2CA100);
rel2MA=mean(of rel2MA1-rel2MA100);
rel2VA=mean(of rel2VA1-rel2VA100);

biais2pE=mean(of biais2pE1-biais2pE100);
biais2pCA=mean(of biais2pCA1-biais2pCA100);
biais2pMA=mean(of biais2pMA1-biais2pMA100);
biais2pVA=mean(of biais2pVA1-biais2pVA100);
biaiscar2pE=mean(of biaiscar2pE1-biaiscar2pE100);
biaiscar2pCA=mean(of biaiscar2pCA1-
biaiscar2pCA100);
biaiscar2pMA=mean(of biaiscar2pMA1-
biaiscar2pMA100);
biaiscar2pVA=mean(of biaiscar2pVA1-
biaiscar2pVA100);
var2pE=mean(of var2pE1-var2pE100);
var2pCA=mean(of var2pCA1-var2pCA100);
var2pMA=mean(of var2pMA1-var2pMA100);
var2pVA=mean(of var2pVA1-var2pVA100);
rmse2pE=mean(of rmse2pE1-rmse2pE100);
rmse2pCA=mean(of rmse2pCA1-rmse2pCA100);
rmse2pMA=mean(of rmse2pMA1-rmse2pMA100);
rmse2pVA=mean(of rmse2pVA1-rmse2pVA100);
rel2pE=mean(of rel2pE1-rel2pE100);
rel2pCA=mean(of rel2pCA1-rel2pCA100);
rel2pMA=mean(of rel2pMA1-rel2pMA100);
rel2pVA=mean(of rel2pVA1-rel2pVA100);

keep reg2 biais2E biais2CA biais2MA biais2VA
biaiscar2E biaiscar2CA biaiscar2MA biaiscar2VA
var2E var2CA var2MA var2VA
rmse2E rmse2CA rmse2MA rmse2VA
rel2E rel2CA rel2MA rel2VA
biais2pE biais2pCA biais2pMA biais2pVA
biaiscar2pE biaiscar2pCA biaiscar2pMA
biaiscar2pVA var2pE var2pCA var2pMA var2pVA
rmse2pE rmse2pCA rmse2pMA rmse2pVA
rel2pE rel2pCA rel2pMA rel2pVA;
run;


*** mise en forme ***;

data est2bis;
set est2;
estimateur='est_reg2';
regroupements='oui';
activite="&activite";
txrep=0.8;
rename reg2=region biais2E=biaishE
biais2CA=biaishCA biais2MA=biaishMA
biais2VA=biaishVA biaiscar2E=biaiscarhE
biaiscar2CA=biaiscarhCA
biaiscar2MA=biaiscarhMA
biaiscar2VA=biaiscarhVA var2E=varhE
var2CA=varhCA var2MA=varhMA
var2VA=varhVA
rmse2E=rmsehE rmse2CA=rmsehCA
rmse2MA=rmsehMA rmse2VA=rmsehVA
rel2E=relhE rel2CA=relhCA rel2MA=relhMA
rel2VA=relhVA biais2pE=biaishpE
biais2pCA=biaishpCA biais2pMA=biaishpMA
biais2pVA=biaishpVA biaiscar2pE=biaiscarhpE
biaiscar2pCA=biaiscarhpCA
biaiscar2pMA=biaiscarhpMA
biaiscar2pVA=biaiscarhpVA var2pE=varhpE
var2pCA=varhpCA var2pMA=varhpMA
var2pVA=varhpVA rmse2pE=rmsehpE
rmse2pCA=rmsehpCA rmse2pMA=rmsehpMA
rmse2pVA=rmsehpVA rel2pE=relhpE
rel2pCA=relhpCA rel2pMA=relhpMA
rel2pVA=relhpVA;
run;

data stage.resultg2_reg&activite;
retain estimateur activite txrep regroupements region
biaishE biaishCA biaishMA biaishVA biaishpE
biaishpCA biaishpMA biaishpVA
biaiscarhE biaiscarhCA biaiscarhMA biaiscarhVA
biaiscarhpE biaiscarhpCA biaiscarhpMA
biaiscarhpVA varhE varhCA varhMA varhVA
varhpE varhpCA varhpMA varhpVA
rmsehE rmsehCA rmsehMA rmsehVA rmsehpE
rmsehpCA rmsehpMA rmsehpVA relhE relhCA
relhMA relhVA relhpE relhpCA relhpMA relhpVA;
set est2bis;
run;
```

```
*****************************************;
****** estimateur 3 : calage sur échantillon ******;
*****************************************;

*** calcul des biais et variances au niveau ih ***;

%macro estim3;

data est31;
set stage.actgc3&activite;
gni2=gni*gni;
run;


%do j=1 %to 100;

%let k=%EVAL(&j+1);
%let l=%eval(%sysfunc(int(%sysevalf((&j-
1)/20)))+1);

proc sort data=est3&j;
by teff2 g&j;
run;

proc summary data=est3&j;
var pnihR&j;
by teff2 g&j;
output out=nivg sum=pnigR&j;
run;

data est3&j;
merge est3&j nivg;
by teff2 g&j;
run;

proc sort data=est3&j;
by teff2 gp&j;
run;

proc summary data=est3&j;
var pnihpR&j;
by teff2 gp&j;
output out=nivgp sum=pnigpR&j;
run;

data est3&j;
merge est3&j nivgp;
by teff2 gp&j;
run;

data est3&k;
set est3&j;

if pnigtilde&j=. then do;
biaisih3E&j=.; biaisih3CA&j=.;
biaisih3MA&j=.; biaisih3VA&j=.;
varih3E&j=.; varih3CA&j=.;
varih3MA&j=.; varih3VA&j=.;
end;

else do;

biaisih3E&j=((pnigtilde&j/pni3)
*(pnihR&j/pnigR&j)-gpih)*Ebarihs;
biaisih3CA&j=((pnigtilde&j/pni3)
*(pnihR&j/pnigR&j)-gpih)*CAbarihs;
biaisih3MA&j=((pnigtilde&j/pni3)
*(pnihR&j/pnigR&j)-gpih)*MAbarihs;
biaisih3VA&j=((pnigtilde&j/pni3)
*(pnihR&j/pnigR&j)-gpih)*VAbarihs;
if pnihR&j NE 0 then do;
varih3E&j=(pnigtilde&j/pni3)*(pnigtilde&j/pni3)
*(pnihR&j/pnigR&j)*(pnihR&j/pnigR&j)
*S2Eihs/pnihR&j;
varih3CA&j=(pnigtilde&j/pni3)*(pnigtilde&j/pni3)
*(pnihR&j/pnigR&j)*(pnihR&j/pnigR&j)
*S2CAihs/pnihR&j;
varih3MA&j=(pnigtilde&j/pni3)*(pnigtilde&j/pni3)
*(pnihR&j/pnigR&j)*(pnihR&j/pnigR&j)
*S2MAihs/pnihR&j;
varih3VA&j=(pnigtilde&j/pni3)*(pnigtilde&j/pni3)
*(pnihR&j/pnigR&j)*(pnihR&j/pnigR&j)
*S2VAihs/pnihR&j;
end;
else do;
varih3E&j=0; varih3CA&j=0; varih3MA&j=0;
varih3VA&j=0;
end;

end;

if pnigptilde&j=. then do;
biaisih3pE&j=.; biaisih3pCA&j=.;
biaisih3pMA&j=.; biaisih3pVA&j=.;
varih3pE&j=.; varih3pCA&j=.;
varih3pMA&j=.; varih3pVA&j=.;
end;

else do;

biaisih3pE&j=((pnigptilde&j/pni3)
*(pnihpR&j/pnigpR&j)-gpih)*Ebarihs;
biaisih3pCA&j=((pnigptilde&j/pni3)
*(pnihpR&j/pnigpR&j)-gpih)*CAbarihs;
biaisih3pMA&j=((pnigptilde&j/pni3)
*(pnihpR&j/pnigpR&j)-gpih)*MAbarihs;
biaisih3pVA&j=((pnigptilde&j/pni3)
*(pnihpR&j/pnigpR&j)-gpih)*VAbarihs;
if pnihpR&j NE 0 then do;
varih3pE&j=(pnigptilde&j/pni3)*(pnigptilde&j/pni3)
*(pnihpR&j/pnigpR&j)*(pnihpR&j/pnigpR&j)
*S2Eihs/pnihpR&j;
```

```
varih3pCA&j=(pnigptilde&j/pni3)
*(pnigptilde&j/pni3)*(pnihpR&j/pnigpR&j)
*(pnihpR&j/pnigpR&j)*S2CAihs/pnihpR&j;
varih3pMA&j=(pnigptilde&j/pni3)
*(pnigptilde&j/pni3)*(pnihpR&j/pnigpR&j)
*(pnihpR&j/pnigpR&j)*S2MAihs/pnihpR&j;
varih3pVA&j=(pnigptilde&j/pni3)
*(pnigptilde&j/pni3)*(pnihpR&j/pnigpR&j)
*(pnihpR&j/pnigpR&j)*S2VAihs/pnihpR&j;
end;
else do;
varih3pE&j=0; varih3pCA&j=0;
varih3pMA&j=0; varih3pVA&j=0;
end;

end;
%end;

%mend;

%estim3;
run;


*** les étapes suivantes sont identiques à celles qui
sont présentées ci-dessus pour l'estimateur 2 ***;



****************************************;
******** estimateur 4 : poststratification *********;
****************************************;

*** calcul des biais et variances au niveau ih ***;

%macro estim4;

data est41;
set stage.actg&activite;
gni2=gni*gni;
run;


%do j=1 %to 100;

%let k=%EVAL(&j+1);

proc sort data=est4&j;
by teff2 g&j;
run;

proc summary data=est4&j;
var gpih pnihR&j;
by teff2 g&j;
output out=nivg sum=gpig&j pnigR&j;
run;

data est4&j;
merge est4&j nivg;
by teff2 g&j;
run;

proc sort data=est4&j;
by teff2 gp&j;
run;

proc summary data=est4&j;
var gpih pnihpR&j;
by teff2 gp&j;
output out=nivgp sum=gpigp&j pnigpR&j;
run;

data est4&j;
merge est4&j nivgp;
by teff2 gp&j;
run;


data est4&k;
set est4&j;

biaisih4E&j=(gpig&j*pnihR&j/pnigR&j
-gpih)*Ebarihs;
biaisih4CA&j=(gpig&j*pnihR&j/pnigR&j
-gpih)*CAbarihs;
biaisih4MA&j=(gpig&j*pnihR&j/pnigR&j
-gpih)*MAbarihs;
biaisih4VA&j=(gpig&j*pnihR&j/pnigR&j
-gpih)*VAbarihs;
if pnihR&j NE 0 then do;
varih4E&j=(gpig&j*pnihR&j/pnigR&j)
*(gpig&j*pnihR&j/pnigR&j)*S2Eihs/pnihR&j;
varih4CA&j=(gpig&j*pnihR&j/pnigR&j)
*(gpig&j*pnihR&j/pnigR&j)*S2CAihs/pnihR&j;
varih4MA&j=(gpig&j*pnihR&j/pnigR&j)
*(gpig&j*pnihR&j/pnigR&j)*S2MAihs/pnihR&j;
varih4VA&j=(gpig&j*pnihR&j/pnigR&j)
*(gpig&j*pnihR&j/pnigR&j)*S2VAihs/pnihR&j;
end;
else do;
varih4E&j=0; varih4CA&j=0;
varih4MA&j=0; varih4VA&j=0;
end;

biaisih4pE&j=(gpigp&j*pnihpR&j/pnigpR&j
-gpih)*Ebarihs;
biaisih4pCA&j=(gpigp&j*pnihpR&j/pnigpR&j
-gpih)*CAbarihs;
biaisih4pMA&j=(gpigp&j*pnihpR&j/pnigpR&j
-gpih)*MAbarihs;
biaisih4pVA&j=(gpigp&j*pnihpR&j/pnigpR&j
-gpih)*VAbarihs;
```

```
if pnihpR&j NE 0 then do;
varih4pE&j=(gpigp&j*pnihpR&j/pnigpR&j)
*(gpigp&j*pnihpR&j/pnigpR&j)*S2Eihs/pnihpR&j;
varih4pCA&j=(gpigp&j*pnihpR&j/pnigpR&j)*
(gpigp&j*pnihpR&j/pnigpR&j)*S2CAihs/pnihpR&j;
varih4pMA&j=(gpigp&j*pnihpR&j/pnigpR&j)*
(gpigp&j*pnihpR&j/pnigpR&j)*S2MAihs/pnihpR&j;
varih4pVA&j=(gpigp&j*pnihpR&j/pnigpR&j)*
(gpigp&j*pnihpR&j/pnigpR&j)*S2VAihs/pnihpR&j;
end;
else do;
varih4pE&j=0; varih4pCA&j=0;
varih4pMA&j=0; varih4pVA&j=0;
end;

%end;

%mend;

%estim4;
run;


*** les étapes suivantes sont identiques à celles qui
sont présentées ci-dessus pour l'estimateur 2 ***;



*****************************************;
******** estimateur 5 : calage sur marge
*********;
*****************************************;

*** calcul des biais et variances au niveau ih ***;

%macro estim5;

data est51;
set stage.actgc3c5&activite;
gni2=gni*gni;
run;


%do j=1 %to 100;

%let k=%EVAL(&j+1);

proc sort data=est5&j;
by teff2 g&j;
run;

proc summary data=est5&j;
var pnihR&j;
by teff2 g&j;
output out=nivg sum=pnigR&j;
run;

data est5&j;
merge est5&j nivg;
by teff2 g&j;
run;

proc sort data=est5&j;
by teff2 gp&j;
run;

proc summary data=est5&j;
var pnihpR&j;
by teff2 gp&j;
output out=nivgp sum=pnigpR&j;
run;

data est5&j;
merge est5&j nivgp;
by teff2 gp&j;
run;


data est5&k;
set est5&j;

if gpigtilde&j=. then do;
biaisih5E&j=.; biaisih5CA&j=.;
biaisih5MA&j=.; biaisih5VA&j=.;
varih5E&j=.; varih5CA&j=.;
varih5MA&j=.; varih5VA&j=.;
end;

else do;

biaisih5E&j=(gpigtilde&j*pnihR&j/pnigR&j
-gpih)*Ebarihs;
biaisih5CA&j=(gpigtilde&j*pnihR&j/pnigR&j
-gpih)*CAbarihs;
biaisih5MA&j=(gpigtilde&j*pnihR&j/pnigR&j
-gpih)*MAbarihs;
biaisih5VA&j=(gpigtilde&j*pnihR&j/pnigR&j
-gpih)*VAbarihs;
if pnihR&j NE 0 then do;
varih5E&j=(gpigtilde&j*pnihR&j/pnigR&j)
*(gpigtilde&j*pnihR&j/pnigR&j)*S2Eihs/pnihR&j;
varih5CA&j=(gpigtilde&j*pnihR&j/pnigR&j)*
(gpigtilde&j*pnihR&j/pnigR&j)*S2CAihs/pnihR&j;
varih5MA&j=(gpigtilde&j*pnihR&j/pnigR&j)*
(gpigtilde&j*pnihR&j/pnigR&j)*S2MAihs/pnihR&j;
varih5VA&j=(gpigtilde&j*pnihR&j/pnigR&j)*
(gpigtilde&j*pnihR&j/pnigR&j)*S2VAihs/pnihR&j;
end;
else do;
varih5E&j=0; varih5CA&j=0;
varih5MA&j=0; varih5VA&j=0;
end;
```

```
end;

if gpigptilde&j=. then do;
biaisih5pE&j=.; biaisih5pCA&j=.;
biaisih5pMA&j=.; biaisih5pVA&j=.;
varih5pE&j=.; varih5pCA&j=.;
varih5pMA&j=.; varih5pVA&j=.;
end;

else do;

biaisih5pE&j=(gpigptilde&j*pnihpR&j/pnigpR&j
-gpih)*Ebarihs;
biaisih5pCA&j=(gpigptilde&j*pnihpR&j/pnigpR&j
-gpih)*CAbarihs;
biaisih5pMA&j=(gpigptilde&j*pnihpR&j/pnigpR&j
-gpih)*MAbarihs;
biaisih5pVA&j=(gpigptilde&j*pnihpR&j/pnigpR&j
-gpih)*VAbarihs;
if pnihpR&j NE 0 then do;
varih5pE&j=(gpigptilde&j*pnihpR&j/pnigpR&j)
*(gpigptilde&j*pnihpR&j/pnigpR&j)
*S2Eihs/pnihpR&j;
varih5pCA&j=(gpigptilde&j*pnihpR&j/pnigpR&j)
*(gpigptilde&j*pnihpR&j/pnigpR&j)
*S2CAihs/pnihpR&j;
varih5pMA&j=(gpigptilde&j*pnihpR&j/pnigpR&j)
*(gpigptilde&j*pnihpR&j/pnigpR&j)
*S2MAihs/pnihpR&j;
varih5pVA&j=(gpigptilde&j*pnihpR&j/pnigpR&j)
*(gpigptilde&j*pnihpR&j/pnigpR&j)
*S2VAihs/pnihpR&j;
end;
else do;
varih5pE&j=0; varih5pCA&j=0;
varih5pMA&j=0; varih5pVA&j=0;
end;

end;

%end;

%mend;

%estim5;
run;


*** les étapes suivantes sont identiques à celles qui
sont présentées ci-dessus pour l'estimateur 2 ***;
```

```
*****************************************;
********* estimateur 6 : imputation ************;
*****************************************;

*** calcul des biais et variances au niveau ih ***;

%macro estim6;

data est61;
set stage.actg&activite;
gni2=gni*gni;
run;


%do j=1 %to 100;

%let k=%EVAL(&j+1);
%let l=%eval(%sysfunc(int(%sysevalf((&j-
1)/20)))+1);

proc sort data=est6&j;
by teff2 g&j;
run;

proc summary data=est6&j;
var pnih&j pnihR&j;
by teff2 g&j;
output out=nivg sum=pnig&j pnigR&j;
run;

data est6&j;
merge est6&j nivg;
by teff2 g&j;
run;

proc sort data=est6&j;
by teff2 gp&j;
run;

proc summary data=est6&j;
var pnihp&l pnihpR&j;
by teff2 gp&j;
output out=nivgp sum=pnigp&j pnigpR&j;
run;

data est6&j;
merge est6&j nivgp;
by teff2 gp&j;
run;
```

```
data est6&k;
set est6&j;

if g&j in (1,2,3,4,5,7,8,9) then do;
EbarigR=Ebarihs; CAbarigR=CAbarihs;
MAbarigR=MAbarihs; VAbarigR=VAbarihs;
S2EigR=S2Eihs; S2CAigR=S2CAihs;
S2MAigR=S2MAihs; S2VAigR=S2VAihs;
end;
else if g&j in (10,11,12,13) then do;
EbarigR=Ebarigs; CAbarigR=CAbarigs;
MAbarigR=MAbarigs; VAbarigR=VAbarigs;
S2EigR=S2Eigs; S2CAigR=S2CAigs;
S2MAigR=S2MAigs; S2VAigR=S2VAigs;
end;
else if g&j in (14,15) then do;
EbarigR=Ebariggs; CAbarigR=CAbariggs;
MAbarigR=MAbariggs; VAbarigR=VAbariggs;
S2EigR=S2Eiggs; S2CAigR=S2CAiggs;
S2MAigR=S2MAiggs; S2VAigR=S2VAiggs;
end;
else if g&j in (16) then do;
EbarigR=Ebaris; CAbarigR=CAbaris;
MAbarigR=MAbaris; VAbarigR=VAbaris;
S2EigR=S2Eis; S2CAigR=S2CAis;
S2MAigR=S2MAis; S2VAigR=S2VAis;
end;

biaisih6E&j=(pnih&j/pni3-gpih)*(bih*Ebarihs
+(1-bih)*EbarigR);
biaisih6CA&j=(pnih&j/pni3-gpih)*(bih*Cabarihs
+(1-bih)*CAbarigR);
biaisih6MA&j=(pnih&j/pni3-gpih)*(bih*Mabarihs
+(1-bih)*MAbarigR);
biaisih6VA&j=(pnih&j/pni3-gpih)*(bih*Vabarihs
+(1-bih)*VAbarigR);
if pnihR&j NE 0 then do;
varih6E&j=(1/(pni3*pni3))*((pnihR&j+2*pnihR&j
*(pnih&j-pnihR&j)/pnigR&j)*S2Eihs + ((pnih&j
-pnihR&j)*(pnih&j-pnihR&j)/pnigR&j)*S2EigR);
varih6CA&j=(1/(pni3*pni3))*((pnihR&j+2*pnihR&j
*(pnih&j-pnihR&j)/pnigR&j)*S2CAihs + ((pnih&j
-pnihR&j)*(pnih&j-pnihR&j)/pnigR&j)*S2CAigR);
varih6MA&j=(1/(pni3*pni3))*((pnihR&j+2*pnihR&j
*(pnih&j-pnihR&j)/pnigR&j)*S2MAihs + ((pnih&j
-pnihR&j)*(pnih&j-pnihR&j)/pnigR&j)*S2MAigR);
varih6VA&j=(1/(pni3*pni3))*((pnihR&j+2*pnihR&j
*(pnih&j-pnihR&j)/pnigR&j)*S2VAihs + ((pnih&j
-pnihR&j)*(pnih&j-pnihR&j)/pnigR&j)*S2VAigR);
end;
else do;
varih6E&j=0; varih6CA&j=0;
varih6MA&j=0; varih6VA&j=0;
end;
```

```
if gp&j in (1,2,3,4,5,7,8,9) then do;
EbarigpR=Ebarihs; CAbarigpR=CAbarihs;
MAbarigpR=MAbarihs; VAbarigpR=VAbarihs;
S2EigpR=S2Eihs; S2CAigpR=S2CAihs;
S2MAigpR=S2MAihs; S2VAigpR=S2VAihs;
end;
else if gp&j in (10,11,12,13) then do;
EbarigpR=Ebarigs; CAbarigpR=CAbarigs;
MAbarigpR=MAbarigs; VAbarigpR=VAbarigs;
S2EigpR=S2Eigs; S2CAigpR=S2CAigs;
S2MAigpR=S2MAigs; S2VAigpR=S2VAigs;
end;
else if gp&j in (14,15) then do;
EbarigpR=Ebariggs; CAbarigpR=CAbariggs;
MAbarigpR=MAbariggs; VAbarigpR=VAbariggs;
S2EigpR=S2Eiggs; S2CAigpR=S2CAiggs;
S2MAigpR=S2MAiggs; S2VAigpR=S2VAiggs;
end;
else if gp&j in (16) then do;
EbarigpR=Ebaris; CAbarigpR=CAbaris;
MAbarigpR=MAbaris; VAbarigpR=VAbaris;
S2EigpR=S2Eis; S2CAigpR=S2CAis;
S2MAigpR=S2MAis; S2VAigpR=S2VAis;
end;

biaisih6pE&j=(pnihp&l/pni3-gpih)*(bih*Ebarihs
+(1-bih)*EbarigpR);
biaisih6pCA&j=(pnihp&l/pni3-gpih)*(bih*Cabarihs
+(1-bih)*CAbarigpR);
biaisih6pMA&j=(pnihp&l/pni3-gpih)*(bih*Mabarihs
+(1-bih)*MAbarigpR);
biaisih6pVA&j=(pnihp&l/pni3-gpih)*(bih*Vabarihs
+(1-bih)*VAbarigpR);
if pnihpR&j NE 0 then do;
varih6pE&j=(1/(pni3*pni3))*((pnihpR&j
+2*pnihpR&j*(pnihp&l-pnihpR&j)/pnigpR&j)
*S2Eihs + ((pnihp&l-pnihpR&j)*(pnihp&l
-pnihpR&j)/pnigpR&j)*S2EigpR);
varih6pCA&j=(1/(pni3*pni3))*((pnihpR&j
+2*pnihpR&j*(pnihp&l-pnihpR&j)/pnigpR&j)
*S2CAihs + ((pnihp&l-pnihpR&j)*(pnihp&l
-pnihpR&j)/pnigpR&j)*S2CAigpR);
varih6pMA&j=(1/(pni3*pni3))*((pnihpR&j
+2*pnihpR&j*(pnihp&l-pnihpR&j)/pnigpR&j)
*S2MAihs + ((pnihp&l-pnihpR&j)*(pnihp&l
-pnihpR&j)/pnigpR&j)*S2MAigpR);
varih6pVA&j=(1/(pni3*pni3))*((pnihpR&j
+2*pnihpR&j*(pnihp&l-pnihpR&j)/pnigpR&j)
*S2VAihs + ((pnihp&l-pnihpR&j)*(pnihp&l
-pnihpR&j)/pnigpR&j)*S2VAigpR);
end;
else do;
varih6pE&j=0; varih6pCA&j=0;
varih6pMA&j=0; varih6pVA&j=0;
end;
```

```
%end;

%mend;

%estim6;
run;


*** les étapes suivantes sont identiques à celles qui
sont présentées ci-dessus pour l'estimateur 2 ***;



*****************************************;
********* rassemblement des résultats **********;
*****************************************;


data stage.resultg_reg;
set stage.result1_reg151F stage.resultg2_reg151F
stage.resultg3_reg151F stage.resultg4_reg151F
stage.resultg5_reg151F stage.resultg6_reg151F
stage.result1_reg158C stage.resultg2_reg158C
stage.resultg3_reg158C stage.resultg4_reg158C
stage.resultg5_reg158C stage.resultg6_reg158C
stage.result1_reg524C stage.resultg2_reg524C
stage.resultg3_reg524C stage.resultg4_reg524C
stage.resultg5_reg524C stage.resultg6_reg524C
stage.result1_reg524R stage.resultg2_reg524R
stage.resultg3_reg524R stage.resultg4_reg524R
stage.resultg5_reg524R stage.resultg6_reg524R
stage.result1_reg524X stage.resultg2_reg524X
stage.resultg3_reg524X stage.resultg4_reg524X
stage.resultg5_reg524X stage.resultg6_reg524X;
run;
```