ESTIMATION LOCALISEE DU CHOMAGE : UNE APPLICATION DES TECHNIQUES D'ESTIMATION SUR PETITS DOMAINES

Pascal ARDILLY

Insee, Unité des méthodes statistiques

Objectif et contexte

L'estimation de données sur le chômage constitue un enjeu local fort. Actuellement l'Insee diffuse, à rythme trimestriel, des effectifs de chômeurs estimés au niveau Zone d'emploi (ZE), ainsi que des taux de chômage par ZE. Ces statistiques s'appuient sur l'enquête trimestrielle emploi (EE) en ce sens où les estimations du nombre de chômeurs par ZE, lorsqu'elles sont sommées au niveau national, par trimestre et avant la phase de désaisonnalisation, sont par construction égales à l'estimation nationale[1] obtenue par l'EE. Pour cela, l'Insee utilise une méthode simple qui consiste, pour chaque croisement sexe X âge, à répartir l'effectif national de chômeurs entre les ZE proportionnellement au nombre de demandeurs d'emploi en fin de mois (DEFM). Il applique donc déjà une technique d'estimation "petits domaines".
La ZE est la partition du territoire la plus fine qui donne actuellement lieu à une diffusion officielle d'effectifs trimestriels de chômeurs. Le zonage utilisé dans cette étude est un zonage qui a été défini en 1990 (la géographie des ZE a été modifiée fin 2011). Il y a exactement 348 ZE en France métropolitaine. Toute commune appartient à une et une seule ZE - la ville de Paris semblant être la seule commune coïncidant exactement avec une ZE. Les ZE doivent respecter les frontières régionales, mais pas nécessairement les frontières départementales. En terme de taille de population des individus de 15 ans ou plus vivant en ménage ordinaire, on constate une très grande disparité de situations puisque, d'après le recensement de 2007 et en arrondissant, la ZE de taille médiane comprend environ 83 000 individus, le premier décile vaut 32 000, le neuvième décile vaut 293 000 (minimum : 8000 ; maximum 1 847 000 - s'agissant de la ville de Paris) .

On cherche à estimer l'effectif de chômeurs dans chaque ZE au sens du BIT, en moyenne sur le premier trimestre de 2007. Seule la France métropolitaine est ici concernée. Même si actuellement il n'y a pas de norme en matière de précision pour décider de la diffusion ou non d'un indicateur local, il apparaît manifeste que l'EE ne saurait à elle seule produire les estimations par ZE parce que dans les standards de l'Insee, les tailles d'échantillon par ZE restent - à quelques exceptions près peut-être - petites ou très petites. C'est pourquoi il a paru important d'appliquer des méthodes spécifiquement adaptées à l'estimation sur petits domaines. Ce document fournit des résultats numériques associés à différentes méthodes concurrentes. En partie 1, on présente les données disponibles, en partie 2 on explique comment on effectue une présélection de variables explicatives susceptibles d'être utilisées avec les méthodes "petits domaines", et en partie 3, après un bref rappel théorique, on donne les principaux résultats obtenus pour chaque méthode.

1. Les données

On distingue les données provenant de l'enquête Emploi - parmi lesquelles figure la variable d'intérêt chômeur / non-chômeur au sens du BIT, et les autres sources de données, qui sont susceptibles de fournir une information auxiliaire explicative du chômage, soit à un niveau individuel, soit à un niveau agrégé - donc au niveau ZE.

[1] Cette propriété en en fait vérifiée par croisement sexe - tranche d'âge.

1.1. L'échantillon Emploi

Le champ de l'enquête est constitué par les personnes ayant 15 ans ou plus au 31 décembre de l'année en cours (donc au 31 décembre 2007). On a restreint les estimations aux individus résidant en ménage ordinaire. Le plan de sondage est un plan complexe de type rotatif, stratifié, à plusieurs degrés, donnant lieu *in fine* au tirage de grappes (grappes d'une quarantaine d'individus physiques - soit environ 20 logements). La pondération finale est obtenue après un calage dit "en une étape", qui à la fois corrige la non-réponse et réduit la variance d'échantillonnage. La pondération moyenne de l'échantillon considéré vaut 680, mais la dispersion des poids est appréciable : 308 pour le premier décile, 1087 pour le neuvième décile. On vérifie que 90% des individus se trouvent dans un rapport de poids variant de 1 à 6 (ce qui déjà assez fort).

La taille de la population de 15 ans et plus résidant en ménage ordinaire en métropole est estimée à 49 745 000 individus au premier trimestre 2007. Parmi ces personnes, on estime à 2 416 400 le nombre total de chômeurs au sens du BIT (estimation nationale en métropole, après calage).

L'échantillon global de personnes physiques répondantes (entretien complet ou partiel[2]), dans le champ ici concerné, comprend 73 153 individus. Les tailles d'échantillon répondant par ZE varient beaucoup, si on exclut les ZE non couvertes par l'enquête :

Quantile		Taille d'échantillon répondant
100%	**Max**	2301
99%		1233
95%		693
90%		520
75%	**Q3**	227
50%	**Median**	**128**
25%	**Q1**	72
10%		38
5%		23
1%		12
0%	**Min**	1

Ainsi, il y a exactement 3 ZE comprenant entre 1 et 10 répondants, et 7 ZE comprenant entre 11 et 20 répondants. A l'autre bout de l'échelle, on trouve 3 ZE ayant plus de 1500 répondants. Par ailleurs, on trouve exactement 10 ZE non couvertes par l'enquête, c'est-à-dire qui ne contiennent aucun individu échantillonné dans l'EE, par le fait du hasard. Ce sont, logiquement, des petites ZE (toutes comprennent moins de 65 000 personnes au RP 1999).

A noter qu'une sérieuse difficulté apparait au niveau du calcul de variance des estimateurs transversaux car l'échantillonnage fait apparaître, à certains degrés, des tailles d'échantillon égales à un.

1.2. Les sources auxiliaires

1.2.1. Le recensement de la population

Il s'agit d'une source contenant énormément d'observations (environ 35 700 000 en 2007), mais pauvre en variables (une dizaine de variables sociodémographiques individuelles potentiellement utiles pour notre problématique). Le recensement a la caractéristique d'être une enquête par sondage portant sur des adresses (unités d'échantillonnage). Si on cumule les données sur 5 années

[2] La proportion d'entretiens partiels est de 0.4% du nombre total d'entretien partiels ou complets.

consécutives, dans les communes de moins de 10 000 habitants (dites "petites communes") on obtient l'exhaustivité des individus (modulo les perturbations dues aux décalages temporels), mais en revanche dans les communes de plus de 10 000 habitants (dites "grandes communes"), l'opération reste une vraie enquête par sondage avec un taux de sondage de 40%. Les individus statistiques (individus, logements, familles) sont donc pondérés (la plupart des poids sont compris entre 1 et 4). L'agrégation des collectes sur 5 années consécutives a pour effet de rendre plus complexes les concepts de "vraie valeur" que l'on estime à partir des micro-données, puisqu'il s'agit d'une combinaison savante de cinq paramètres annuels. Ainsi, la source recensement labélisée 2007 associe en réalité les collectes annuelles 2005, 2006, 2007, 2008 et 2009 - ce n'est donc pas un paramètre 2007 au sens strict qui est estimé.

Techniquement, en toute rigueur, on ne devrait considérer dans les grandes communes aucune donnée agrégée du recensement comme "exacte", puisque chacune de ces données est entachée d'une erreur d'échantillonnage - sans parler du biais dû au mécanisme d'agrégation. Néanmoins, par la suite, on fera "comme si" on avait à faire à un recensement exhaustif traditionnel - cette simplification n'ayant aucune conséquence significative dans les ZE peuplées, et une conséquence qui reste très limitée dans les ZE les moins peuplées.

Pour se situer exactement sur le champ de l'enquête Emploi, on a filtré les seuls individus de 15 ans et plus au 31 décembre 2007 et résidant en ménage ordinaire. On a sélectionné les (quelques) variables individuelles qui semblaient *a priori* susceptibles d'apparaitre en rapport avec la situation individuelle en matière d'emploi. Seules les données en cumul sur 5 ans ont servi pour produire les vrais totaux : les enquêtes annuelles de recensement (EAR) n'ont pas été exploitées.

1.2.2. Les Demandes d'Emploi en Fin de Mois

Les statistiques DEFM sont produites par Pôle Emploi (auparavant ANPE) qui a distingué jusqu'en 2009 huit catégories de demandeurs d'emploi. Les DEFM dites "1, 2 et 3", regroupent les personnes sans emploi immédiatement disponibles, tenues d'accomplir des actes positifs de recherche d'emploi. Il est possible de distinguer les personnes pratiquant une activité réduite[3].

Nous disposons de données au niveau individuel, après anonymisation. La base de données individuelles DEFM ne renseigne pas toujours la commune : au cours du T1 de 2007, environ 16 700 individus, assez bien répartis dans l'ensemble des régions, sont concernés en moyenne chaque mois par ce problème. Ils ont été purement et simplement exclus, sans compensation, avec l'argument que leur faible effectif (environ 0.8% des DEFM totales) ne devrait pas générer de biais significatif dans les estimations localisées qui s'appuieraient sur les DEFM. Les variables communiquées distinguent le sexe, l'âge et le niveau de formation. Aussi, des dénombrements exhaustifs ont été effectués par ZE en distinguant les croisements tranche d'âge X sexe X niveau de formation (en deux modalités : bac + 3 ans ou plus d'une part, niveau inférieur d'autre part).

Noter que dans l'enquête Emploi, on demande à la personne interrogée si elle est inscrite à Pole Emploi / ANPE (sur une période glissante d'un mois).

1.2.3. Les autres sources de données

D'autres sources de données auxiliaires ont été mobilisées :

- Une source nationale ad-hoc[4] fournit un ensemble très diversifié de variables de nature économique et/ou sociale agrégées par ZE. Les sources alimentant cette base sont, ou bien liées aux personnes (recensements, caisse d'allocations familiales), ou bien relatives aux entreprises (CLAP, Sirene, Lifi), ou bien encore se situent à l'intersection des deux univers (DADS, données fiscales, données touristiques). Certaines variables dont la corrélation avec le dynamisme du marché du travail apparaît *a priori* naturelle concernent la démographie d'entreprises et caractérisent plutôt l'aspect "offre d'emploi". S'ajoutent des informations

[3] Par définition, dont la durée mensuelle est inférieure à un certain seuil.
[4] Base constituée par Laurence Labosse, SED, DR Rhône-Alpes.

issues de la Base permanente des équipements. Ce sont (presque) toutes des sources que l'on a coutume de qualifier d'exhaustives.

- La (célèbre) typologie "Tabard", qui est une typologie socioprofessionnelle des quartiers et communes où seuls les hommes actifs ayant un emploi et les chômeurs sont pris en compte, et qui associe *in fine* à chaque unité géographique retenue un des 27 codes distingués. On l'utilise pour calculer pour chaque ZE un "profil moyen" socio démographique en considérant la proportion d'individus de la ZE affectés à chaque code de la typologie. Pour chaque ZE, on construit donc une structure composée de 27 pourcentages. Cette typologie s'appuie certes sur les données du recensement 1999 (exploitation au quart) et n'a pas été mise à jour depuis. Comme il est difficile d'obtenir, pour chaque individu de l'échantillon Emploi, l'identifiant de l'unité géographique "Tabard" dans lequel il se trouve, cette typologie, dans sa conception initiale, paraît peu exploitable dans un modèle individuel. Mais il y un moyen de contourner cette difficulté car existe aussi une typologie définie au niveau communal. Cette typologie connexe a également donné lieu à des profils par ZE.

- Nous disposons d'une base de données RMI. Par ZE, on distingue d'une part l'effectif d'allocataires, d'autre part l'effectif de personnes couvertes par la prestation (dites aussi "bénéficiaires"). Le second concept inclut, outre les allocataires, toutes les personnes qui bénéficient de l'allocation RMI dans l'entourage de l'allocataire (il y a grosso modo deux fois plus de bénéficiaires que d'allocataires).

- il y a environ 850 ZUS en France. L'implantation en ZUS du lieu de résidence d'un individu est un facteur susceptible d'influer sur son statut d'activité. La configuration des ZUS ne répond à aucune règle précise : par exemple, il existe des petites communesqui sont coupées en deux par une frontière de ZUS. L'affectation des individus recensés aux ZUS a été conçue pour refléter la situation de l'année 2006.

1.2.4. La base de données définitive au niveau ZE

Dans la méthodologie d'estimation "petits domaines", la connaissance pour chaque domaine des vrais totaux des variables auxiliaires susceptibles d'expliquer le chômage est un élément central. Ces totaux par ZE sont ceux qui résultent des sources de données auxiliaires présentées ci-dessus et s'utilisent de deux manières : soit la modélisation du nombre de chômeurs s'effectue directement au niveau ZE et ces totaux jouent le rôle de variables auxiliaires dans les modèles explicatifs, soit la modélisation s'effectue à un niveau individuel, et lorsqu'on utilise une modélisation linéaire[5] on a besoin de ces valeurs pour former les estimateurs "petits domaines".

Au-delà de ces informations explicatives, il convient de construire des variables expliquées pour chaque ZE. L'enquête Emploi classe chaque individu répondant selon trois modalités : actif occupé / chômeur / inactif. La notion de chômeur est bien celle qui répond à la définition du BIT. L'information essentielle expliquée est, pour chaque ZE, le rapport $\hat{P}$ de l'estimateur pondéré du nombre total de chômeurs à l'estimateur pondéré de la taille totale de la ZE - toujours limitée aux individus du champ. Noter qu'on s'intéresse ici au nombre total de chômeurs et non au taux de chômage par ZE (qui est un paramètre plus complexe puisque le dénominateur doit être lui-même estimé). Voici la distribution $\hat{P}$:

Quantile		Taux $\hat{P}$ (%)
100%	**Max**	20.7
99%		14.2
95%		10.0
90%		8.2
75%	**Q3**	6.2
50%	**Median**	**4.2**

[5] Ce ne sera pas le cas dans ce document - voir partie 6 - mais le modèle individuel (logistique) impliqué exige plus que la simple connaissance des totaux auxiliaires par ZE.

Quantile		Taux $\hat{P}$ (%)
25%	**Q1**	2.4
10%		0.0
5%		0.0
1%		0.0
0%	**Min**	0.0

Les données sur le chômage (en volume) au niveau ZE peuvent prendre la forme d'au moins 6 variables :

- l'effectif officiel trimestriel, publié par l'Insee, qui est un effectif CVS
- l'effectif brut - c'est-à-dire non CVS - ayant servi à établir la statistique officielle
- le nombre de personnes DEFM (toutes catégories confondues) (T1 de 2007)
- le nombre de DEFM catégories 1, 2, 3 hors activité réduite (T1 de 2007)
- l'estimateur direct issu de l'enquête trimestrielle Emploi
- le nombre de chômeurs estimé à partir des déclarations individuelles[6] du recensement 2007

Le tableau suivant, constitué au T1 de 2007 à partir de 10 ZE quelconques (plus la totalisation France métropolitaine), donne une idée de la disparité de ces variables - bien logiquement puisqu'il s'agit de concepts différents les uns des autres.

ZE	Chômage officiel	Série brute	DEFM totales	DEFM 1,2,3 HAR(*)	Estimateur EE	Estimateur recensement
1	2220	2311	2218	1436	2133	3268
2	12428	12590	12600	8583	12143	16511
3	4471	4645	4106	2798	3909	6327
4	1845	1945	1724	1198	571	2561
5	2000	2141	1955	1330	2766	2865
6	19855	20634	18153	12209	12951	25347
7	4671	5036	4277	2961	7356	6325
8	1536	1619	1483	995	2084	2106
9	7749	7912	6775	4580	5489	10029
10	2606	2672	2622	1660	1402	3227
France (hors DOM)	**2 352 000**	**2 408 000**	**2 096 000**	**1 464 000**	**2 416 000**	**3 145 000**

(*) Hors Activité Réduite.

La base de données finale, au niveau ZE, contient 348 observations et 275 variables - essentiellement des rapports de totaux à la taille totale de population de la ZE, exprimés directement en pourcentage.

2. La présélection des variables explicatives

2.1. Pour expliquer une proportion de chômeurs

Pour estimer la vraie proportion P de chômeurs par ZE, définie comme le ratio du nombre total de chômeurs BIT à la taille totale de la population du champ, on utilise naturellement l'estimateur direct "ratio" $\hat{P}$ du 1.2.4. Sur l'ensemble des ZE de métropole, l'estimation $\hat{P}$ vaut 4,86% (soit 2 416 400 /

[6] Question 10 du BI, modalité 4.

49 745 000). Noter que P n'a rien à voir avec le taux de chômage, plus compliqué, qui comprend au dénominateur le nombre d'actifs du champ. Dans chaque ZE, à partir de $\hat{P}$, on obtient immédiatement un effectif estimé de chômeurs dans la ZE, puisqu'on connaît les tailles de population des ZE grâce au recensement (de fait, c'est un estimateur calé sur la taille de la population appartenant au champ). Selon le modèle retenu, c'est soit la variable $\hat{P}$, soit le nombre total de chômeurs estimé, qui constituera la variable expliquée par ZE. On cherche désormais à sélectionner des variables explicatives pertinentes au niveau ZE.

A titre préliminaire, en utilisant des ellipses de prédiction, deux ZE au comportement atypique sont éliminées (le nombre de chômeurs estimé par l'EE y est anormalement grand). Par la suite, on effectue le choix de modèle à partir de toutes les ZE ayant au moins 50 individus répondants (pour éviter d'éventuelles valeurs vraiment trop fantaisistes de $\hat{P}$*).* Sont alors concernées 285 ZE.
La liste des ZE étant stabilisée, notre parti pris a été de faire une première sélection de variables en se basant sur les corrélations bivariées, entre d'une part la variable d'intérêt $\hat{P}$ et d'autre part un candidat régresseur. Cette étape est une façon simple et finalement assez naturelle de limiter le nombre de variables tout en permettant une interprétation facile des variables qui subsisteront en fin de course. Il ne faut pas être exigeant, car contrairement à la problématique habituelle des modèles linéaires, on se trouve ici en présence d'une forte erreur d'observation sur la variable expliquée. Il s'agit aussi d'une façon préventive de limiter le risque d'introduire des régresseurs en réalité inutiles qui pourraient augmenter la variance des effets aléatoires locaux propres aux domaines dans les modèles. Sur le plan théorique, il est néanmoins clair que ce n'est pas optimum : un "bon" caractère explicatif d'une variable d'intérêt par une combinaison linéaire de variables auxiliaires ne veut pas dire que mathématiquement toutes les corrélations linéaires entre la variable d'intérêt et chacun des régresseurs pris isolement soient nécessairement au-delà d'un certain seuil.

A ce stade, pour une plus grande visibilité et un équilibrage des volumétries[7], on décide d'effectuer une sélection séparée, d'une part sur les variables issues du recensement, des DEFM, du RMI et de la base ZUS (1ère liste), d'autre part sur les variables économiques et les profils Tabard (2nde liste). On considère les variables de la première liste. En première étape, on retient toutes celles dont le coefficient de Pearson est inférieur à -10% ou supérieur à +20%, avec la condition supplémentaire que la p-value du test de nullité de ce coefficient de corrélation soit inférieure à 5% (nota : ce test s'appuie sur la transformée de Fisher).

Les variables les plus corrélées positivement à $\hat{P}$ sont :

- Recherche d'un emploi (déclaration au recensement) : $\rho = 42{,}7\,\%$
- Déclaration spontanée de chômage au recensement : $\rho = 42{,}2\,\%$
- Inscription DEFM, catégories 1,2,3 et HAR : $\rho = 39{,}5\,\%$
- Allocataire RMI : $\rho = 38{,}8\,\%$
- Inscription DEFM, catégories 1,2,3 et HAR et appartenir à la catégorie des hommes de 30 à 49 ans, peu diplômés[8] : $\rho = 38{,}2\,\%$
- Ne pas vivre en couple : $\rho = 35{,}4\,\%$
- Inscription DEFM, toutes catégories et appartenir à la catégorie des hommes de 50 à 64 ans, peu diplômés : $\rho = 35{,}0\,\%$

Les variables les plus corrélées négativement à $\hat{P}$ sont

- être marié : $\rho = -32{,}8\,\%$
- avoir une profession d'indépendant : $\rho = -24{,}8\,\%$
- avoir un niveau de diplôme égal au CAP / BEP : $\rho = -19{,}7\,\%$

[7] On a au total 285 observations : il faut quand même que le nombre d'observations soit sensiblement supérieur au nombre de régresseurs …
[8] DEFM, catégories 1,2,3 et HAR ou DEFM toutes catégories confondues, au choix, car cela ne change pas la corrélation linéaire.

Cela conduit à sélectionner 70 variables sur 255.

La seconde étape de choix de variables est importante mais sans être déterminante, en ce sens où il s'agit d'une étape de simplification, préparatoire à l'estimation d'un modèle qui offrira le moment venu les moyens de juger de la pertinence de retenir (ou non) telle ou telle variable[9]. C'est pourquoi il faut un processus qui ne soit pas trop compliqué et qui utilise des outils logiciels existants.
La stratégie ici appliquée consiste à oublier toute éventualité d'une structure complexe de la variance stochastique et à procéder, pour cette étape seulement, à une sélection "comme si" on était en présence d'un modèle ordinaire avec une variance de type $\sigma^2 \cdot Id$. Dans ces conditions, on utilise la PROC GLMSELECT de SAS qui propose différentes méthodes de sélection de variables explicatives. Avec une stratégie STEPWISE, on a testé trois options :

- *une méthode basée sur la valeur du F de Fisher*

  Dans le processus de construction du modèle, on ajoute un régresseur lorsque sa p-value (Significance level SL) est inférieure à un seuil donné SLE. La p-value est la probabilité qu'une variable aléatoire de Fisher dépasse la valeur du Fisher calculée. De même, on retire un régresseur lorsque sa p-value est supérieure à un seuil donné SLS. A chaque étape du processus Stepwise, on calcule le R^2 ajusté et on retient le modèle qui correspond à un minimum local de la valeur R^2 ajusté. Cela signifie que le processus se déroule tant que R^2 ajusté diminue, et lorsqu'on arrive au modèle définitif, pousser le Stepwise à l'étape itérative suivante conduirait à augmenter le R^2 ajusté.

  ```
  SELECTION = stepwise (select=SL SLE=0.20 SLS=0.20 stop=ADJRSQ)
  ```

- *une méthode basée sur la valeur du R^2 ajusté*

  Cette fois, on base le processus STEPWISE sur le critère du R^2 ajusté et on recherche un minimum local du Cp de Mallow. Ainsi, l'ajout d'un régresseur ne se fait que si R^2 ajusté augmente - de même pour le retrait éventuel d'un régresseur. L'arrêt du processus s'effectue lorsqu'on atteint un minimum local du Cp de Mallow.

  ```
  SELECTION = stepwise (select=ADJRSQ stop=CP)
  ```

- *une méthode de cross-validation*

  Le principe de la cross-validation consiste à scinder un échantillon en deux morceaux : l'un des morceaux est utilisé pour ajuster le modèle (échantillon d'apprentissage noté a), l'autre pour évaluer la qualité de l'ajustement (échantillon de test noté t). Le critère utilisé est une forme d'erreur de prédiction :

  $$PRESS = \sum_{i \in t} (Y_i - \hat{Y}_i)^2$$

  où $\hat{Y}_i$ est obtenu à partir des paramètres estimés via l'échantillon a . On peut raffiner en définissant plusieurs façons de constituer l'échantillon t (et donc a) : nous avons ainsi défini des blocs de 5 ZE consécutives, dans chaque bloc 4 des 5 ZE servent pour construire a , et la cinquième ZE sert pour construire t . On a ainsi 5 façons de construire le couple (a,t) : pour chaque couple on peut produire un jeu d'estimateurs des paramètres du modèle, ainsi on a une appréciation intéressante de la variabilité de ces coefficients.

  ```
  CVMETHOD=BLOCK(5) CVDETAILS=ALL SELECTION = STEPWISE (select=CV)
  ```

[9] En fait, le processus est asymétrique : éliminer à tort en première phase une variable explicative occasionne une perte d'efficacité non compensée, alors que retenir une variable qui s'avèrera ultérieurement non explicative n'a pas de conséquence.

Le critère de sélection est la statistique $CV - PRESS$ qui est définie comme la somme des cinq $PRESS$ associées à chacun des blocs : on retient le modèle qui en donne un minimum local. On voit apparaître des variables nouvelles, mais la p-value des tests de Student reste faible et donc leur significativité n'est pas convaincante. Voici la table de sortie sélectionnant les variables explicatives (la partie droite fournit la variabilité des coefficients).

Parameter Estimates									
					Cross Validation Estimates				
Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**1**	**2**	**3**	**4**	**5**
Intercept	1	-7.585163	2.555916	-2.97	-8.725	-8.199	-7.2249	-7.1332	-6.7812
t_natc_afri	1	-0.103727	0.096369	-1.08	-0.143	-0.125	-0.0892	-0.0912	-0.0869
t_age15_19HdiplBIT	1	114.316350	53.259219	2.15	112.391	110.770	93.8495	122.0480	130.4282
t_couple_2	1	0.205983	0.066111	3.12	0.237	0.228	0.1972	0.1830	0.1867
t_allocRMI	1	-0.320170	0.368771	-0.87	-0.279	-0.340	-0.2754	-0.3526	-0.3882
t_rech_oui	1	0.720946	0.235373	3.06	0.709	0.704	0.7147	0.7817	0.7152

Avec les variables de la seconde liste, on procède de manière équivalente, avec des options éventuellement un peu différentes (par exemple on adapte les seuils des coefficients de corrélation de l'étape initiale).

Les variables les plus corrélées positivement à $\hat{P}$ sont :

- Part de la population ayant des bas revenus en 2005 : $\rho = 35{,}4\,\%$
- Part de la fonction "Administration publique" dans l'emploi total en 2006 : $\rho = 21{,}2\,\%$
- Part des étudiants dans la population des plus de 15 ans : $\rho = 18{,}5\,\%$

Les variables les plus corrélées négativement à $\hat{P}$ sont :

- Taux d'activité dans la tranche d'âge 25-54 ans selon le RP 2006 : $\rho = -26{,}9\,\%$
- Part des agriculteurs dans la population de plus de 5 ans (RP 2006) : $\rho = -24{,}4\,\%$
- Taux d'activité dans la tranche d'âge 15-24 ans selon le RP 2006 : $\rho = -23{,}8\,\%$

Finalement, cela conduit à présélectionner 38 variables sur 90.

L'application des différents critères précédents (plus d'autres qui n'ont pas été mentionnés ici) ne fournissent pas (hélas, mais on pouvait s'y attendre) des sorties identiques. Néanmoins, on peut déjà retenir toutes les variables qui apparaissent dans au moins une sortie dès lors que leur significativité est "suffisante". De plus, on constate que certaines d'entre elles se retrouvent dans différentes sorties, ce qui est réconfortant. Il y a là certes une phase un peu empirique de sélection mais *de visu* on arrive à un stade où 15 variables ressortent comme potentiellement explicatives, à savoir :

- Taux de personnes qui recherchent un emploi (déclaration au recensement) ;
- Part de la population ayant des bas revenus en 2005 ;
- Taux d'allocataires RMI ;
- Taux de personnes DEFM catégories 1,2,3 et HAR et appartenant à la catégorie des hommes de 15 à 19 ans, diplômés ;
- Taux de personnes DEFM catégories 1,2,3 et HAR et appartenant à la catégorie des hommes de 30 à 49 ans, peu diplômés ;
- Taux de personnes DEFM toutes catégories et appartenant à la catégorie des hommes de 50 à 64 ans, peu diplômés ;
- Taux de personnes ne vivant pas en couple (déclaration au recensement) ;
- Taux d'étrangers hors Europe
- Part de la fonction "BTP" dans l'emploi total en 2006 ;

- Part de la fonction "Santé-action sociale" dans l'emploi total en 2006 ;
- Part de la fonction "Fabrication" dans l'emploi total en 2006 ;
- Part de la fonction "Gestion" dans l'emploi total en 2006 ;
- Part des agriculteurs dans la population des plus de 15 ans
- Taux de solde des établissements entre 2000 et 2006 = (arrivées - départs) divisé par stock d'établissements au 1/1/2006;
- Proportion de la population en catégorie Tabard dite "SEMAG02" (Hôtellerie, restauration).

Lorsqu'on procède à un ultime ajustement utilisant l'ensemble de ces variables, on constate que certaines d'entre elles perdent totalement leur significativité, du fait des corrélations entre les régresseurs : par exemple, la part de la population ayant des bas revenus en 2005 disparaît (elle est très corrélée au Taux de personnes qui recherchent un emploi : $\rho = 0{,}89$). Finalement, le modèle définitif implique 7 variables (plus la constante), dont voici la liste :

- Taux de personnes qui recherchent un emploi (déclaration au recensement - variable *t_rech_oui*) ;
- Taux de personnes ne vivant pas en couple (déclaration au recensement - variable *t_couple_2*);
- Taux de personnes DEFM catégories 1,2,3 et HAR et appartenant à la catégorie des hommes de 15 à 19 ans, diplômés (variable *t_age15_19HdiplBIT*);
- Taux de personnes DEFM catégories 1,2,3 et HAR et appartenant à la catégorie des hommes de 30 à 49 ans, peu diplômés (variable *t_age30_49HnondiplBIT*);
- Taux de personnes DEFM toutes catégories et appartenant à la catégorie des hommes de 50 à 64 ans, peu diplômés (variable *t_age50_64HnondiplBIT)*;
- Taux de solde des établissements entre 2000 et 2006 = (arrivées - départs) divisé par stock d'établissements au 1/1/2006 (variable *c02_txsoldetab_0006*);
- Proportion de la population en catégorie Tabard dite "SEMAG02" (Hôtellerie, restauration - variable *part_depcom_24*).

Toutes les variables sont significatives ou à peu près (à l'exception du solde sur les établissements, mais on décide de la garder car c'est une variable qui a bonne allure !) :

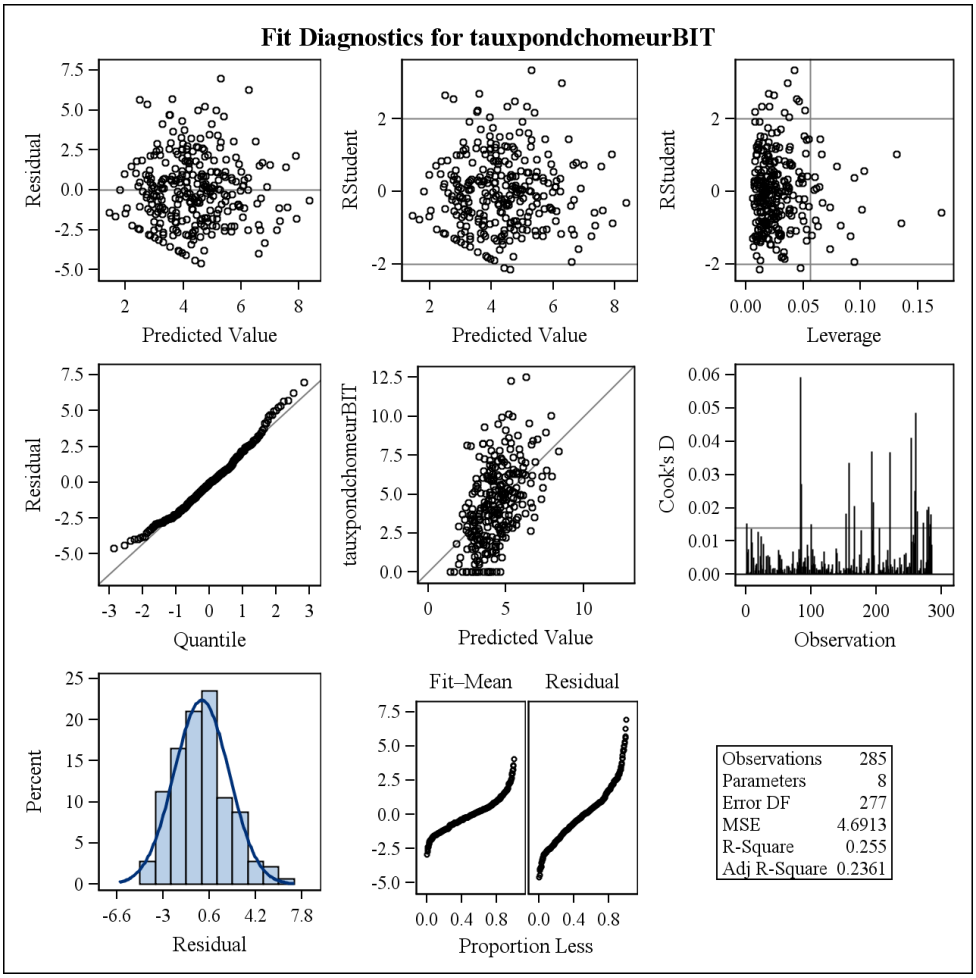
Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	- 6.12	2.03	- 3.02	0.0028
t_rech_oui	0.81	0.26	3.07	0.0024
t_couple_2	0.18	0.06	2.99	0.0030
t_age15_19HdiplBIT	137.70	53.26	2.59	0.0102
t_age50_64Hnondipl	4.08	1.88	2.17	0.0308
t_age30_49HnondiplBIT	- 7.65	2.91	- 2.63	0.0090
c02_txsoldetab_0006	- 0.11	0.07	- 1.67	0.0956
part_depcom_24	0.02	0.01	2.20	0.0286

La très grande valeur du coefficient associé à la variable *t_age15_19HdiplBIT* (qui tranche avec les autres coefficients) tient vraisemblablement au fait que les valeurs de la variable en question sont très faibles - la population concernée ayant certes une influence sur la proportion de chômeurs, mais étant indiscutablement une population rare si on la compare à celle qui sous-tend les autres variables. On obtient le jeu de diagnostics dans la figure ci-dessous[10]. Les résidus ont une distribution qui a plutôt bonne allure en ce sens où il ne se dégage aucune allure particulière du nuage de points[11]. Le

[10] *TauxpondchomeurBIT* est le nom donné à la variable SAS représentant l'estimateur direct $\hat{P}$.

[11] On voit seulement se profiler une droite $Y = -X$ en bas à gauche : c'est dû à la présence d'un certain nombre de ZE où $\hat{P} = 0$, auquel cas le résidu est égal à moins la valeur prédite.

nombre de résidus studentizés anormalement élevés reste par ailleurs raisonnable. Une douzaine d'observations (sur 285) ont des valeurs "leverage" / D de Cook élevées (elles ont donc une influence forte sur l'estimation des coefficients de régression), ce qui est un peu désagréable. Néanmoins, la distribution empirique des résidus apparaît compatible avec une loi de Gauss. On peut imaginer - éventuellement - améliorer la qualité de l'ajustement dans les modèles "petits domaines" en modulant la variance des effets locaux aléatoires des quelques ZE atypiques (en l'augmentant pour les domaines concernés par un leverage élevé, intuitivement on devrait voir leur influence diminuer dans l'estimation des coefficients de régression du modèle définitif). Noter que le graphique central (valeurs prédites versus valeurs expliquées) préfigure le phénomène que l'on va retrouver - plus ou moins - dans les méthodes "petits domaines", à savoir que les estimateurs de type synthétique ont une dispersion plus faible que celle de l'estimateur direct ("shrinkage").



2.2. Pour expliquer un statut individuel

La partie 2.1. concerne des modèles dont la variable expliquée est une proportion de chômeurs par ZE, donc une grandeur quantitative agrégée, comprise entre 0 et 1. Mais on peut aussi concevoir des modèles au niveau individuel dont la variable expliquée est de nature qualitative, prenant seulement deux valeurs : 1 si l'individu est chômeur BIT, et 0 sinon.

Formellement, les modèles individuels sur variables qualitatives sont *a priori* des modèles linéaires généralisés (mixtes ou non). Cela a pour conséquence de conduire à des estimateurs "petits domaines" par ZE qui s'appuient sur les estimations des probabilités de chômage calculées individu par individu. L'estimateur final "petit domaine" du nombre de chômeurs par ZE utilise la somme formée sur l'ensemble de la ZE des probabilités individuelles d'être chômeur. Or le modèle n'est pas linéaire, si bien que cette somme de probabilités ne s'écrit pas en fonction de la somme des valeurs individuelles des variables explicatives (ce qui nous aurait bien arrangé, car les différentes sources

auxiliaires à notre disposition fournissent les effectifs agrégés par ZE correspondant aux différentes modalités des variables explicatives ...). En terme d'information requise, cela a une conséquence considérable : il est nécessaire de connaître, pour chaque individu de la population, les valeurs des variables explicatives. De ce fait, le fichier auxiliaire central est le fichier du recensement, parce que c'est le seul fichier individuel qui permet de procéder à une extrapolation correcte à la population entière (grâce au système de pondération adéquat, puisqu'on rappelle qu'en réalité c'est une très grosse enquête par sondage dans la catégorie des grosses communes). Le seul cas de figure qui permettrait d'échapper à cette logique est celui d'une unique variable explicative qui serait aussi la variable indicatrice de la base auxiliaire. Exemple : utiliser uniquement la variable DEFM avec la base auxiliaire des demandeurs d'emploi, parce qu'il y a alors deux probabilités seulement à calculer (celle qui caractérise les demandeurs d'emploi, et celle qui caractérise les autres personnes), que l'on connaît par ZE la taille de population totale, que l'on connaît par ZE la taille de population totale DEFM, et donc par différence que l'on connaît par ZE la taille de population non demandeuse d'emploi. Là, effectivement, les ingrédients permettent de former une prédiction exhaustive par ZE. Mais ce cas de figure est extrême et, par ailleurs, il revient de fait à utiliser un estimateur synthétique dans un modèle au niveau ZE dont la variable explicative est la proportion de demandeurs d'emploi dans la population totale.

Sur le principe, la conclusion est sévère si on s'en tient aux variables individuelles par nature et si on veut tirer partie de la richesse des corrélations entre ces variables : non seulement ces variables explicatives doivent être présentes dans le fichier échantillon Emploi (donc dans le questionnaire Emploi), mais encore elles doivent être aussi présentes dans le fichier individuel du recensement (donc dans le questionnaire RP). Cela est très restrictif et limite sensiblement les possibilités. Les variables *a priori* susceptibles de liaison avec le statut d'activité et disponibles à la fois dans le questionnaire Emploi et dans le questionnaire du recensement sont les suivantes :

- La situation déclarée pour le mois en cours (déclaration spontanée de l'état de chômage ou non)
- La recherche ou non d'un emploi
- L'âge
- Le sexe
- La nationalité
- Le diplôme le plus élevé
- L'indicateur de vie en couple ou non
- Le statut matrimonial
- Le statut d'occupation du logement (récupéré de la feuille de logement)

Certes, on peut espérer améliorer le modèle en rajoutant des variables explicatives définies au niveau communal (voire supra communal) - comme par exemple n'importe quelle typologie de communes, ou une variable caractérisant le tissu économique de la commune. Mais cela n'a pas été fait dans cette étude. En particulier, on perd toute possibilité d'exploiter une éventuelle inscription à Pôle Emploi : cette information existe bien dans le questionnaire Emploi mais on ne la retrouve pas dans le recensement. Finalement et malheureusement, il apparait qu'en dehors du recensement, aucune des sources auxiliaires citées au 1.2 ne pourra contribuer à la modélisation individuelle …

Il existe dans le questionnaire de l'enquête Emploi une question qui permet de collecter la déclaration spontanée de chômage, au sens où chaque individu concerné le perçoit. Le concept n'est pas bien défini puisqu'il s'agit du ressenti de l'enquêté, mais l'important n'est pas là, ce qui est fondamental est de mesurer le même concept qu'au recensement. Autrement dit, il faut croire que toute personne se déclarant "chômeur" dans le bulletin individuel du recensement répondrait exactement "chômeur" à la question concernée par le questionnaire Emploi si elle faisait partie de l'échantillon - deux individus distincts pouvant par ailleurs se positionner de manière différente s'ils ne comprennent pas la question de la même façon. Cette piste pourrait être prometteuse, mais hélas cette question n'est posée dans l'enquête Emploi de 2007 qu'à la première interrogation, ce qui touche finalement un sous-échantillon de l'enquête qui est de taille plus modeste : nous ne l'avons donc pas retenue dans les régresseurs potentiels pour ne pas dégrader l'ajustement du modèle par rapport aux autres variables. En outre, l'expérience décrite en partie 6, avec la difficulté spécifique rencontrée sur la variable de recherche d'emploi (qui souffre du même risque) ne nous a pas encouragé à persister dans cette voie. La variable "situation déclarée pour le mois en cours" doit de toute façon sortir de notre liste.

Si on fait totalement abstraction des variables définies à un niveau communal ou supra communal, on ne peut donc compter que sur 8 variables individuelles potentiellement explicatives - ce qui occasionne un contraste majeur avec les modèles agrégés.

3. Le calage sur des structures locales

3.1. Éléments de théorie

Il s'agit ici d'utiliser un estimateur direct - c'est-à-dire qui n'utilise que les informations propres à la ZE considérée - qui repondère l'estimateur de Horvitz-Thomson de façon à ce que certains totaux (ou moyennes) soient estimés de manière parfaite, c'est-à-dire que le nouvel estimateur redonne exactement les vraies valeurs connues de ces totaux quel que soit l'échantillon. La théorie du calage généralisé permet de construire une telle repondération, mais il y a plusieurs options possibles qui conduisent à différents estimateurs, tous asymptotiquement équivalents[12]. L'un des estimateurs calés auquel cette théorie conduit est l'estimateur par la régression - dit GREG.

On note d l'identifiant de la ZE, s l'échantillon national d'individus répondants à EE, et $\hat{Y}_d$ l'estimateur direct du vrai total Y_d dans la ZE. On a $\hat{Y}_d = \sum_{\substack{i\in s\\ i\in d}} w_i \cdot Y_i$ et $\hat{N}_d = \sum_{\substack{i\in s\\ i\in d}} w_i$. On rappelle que N_d désigne, dans notre cas, la taille de la population de la ZE d ayant 15 ans ou plus au 31 décembre 2007 et vivant en ménage ordinaire. S'agissant de calage sur un vecteur de moyennes (locales) $\overline{X}_d$, si on cherche à estimer la moyenne vraie $\overline{Y}_d$ on forme :

$$\hat{\overline{Y}}_{d,GREG} = \frac{\hat{Y}_d}{\hat{N}_d} + \hat{B}_d^t \cdot \left(\overline{X}_d - \frac{\hat{X}_d}{\hat{N}_d} \right)$$

où $\hat{B}_d$ est le coefficient de régression vectoriel estimé dans la régression multiple des Y_i sur les vecteurs individuels X_i (dont la première coordonnée est la constante 1). On peut vérifier que si la taille d'échantillon n_d est "assez grande", si on définit $\forall i \in d$: $U_i = Y_i - B_d^t \cdot X_i$ le (vrai) résidu de la régression linéaire de Y sur le vecteur X, alors

$$Var\left(\hat{\overline{Y}}_{d,GREG}\right) \approx \frac{1}{N_d^2} Var\left(\sum_{i\in s\cap d} w_i \cdot U_i \right)$$

où on rappelle que w_i est le poids associé à i dans l'estimateur direct formé avant calage.

On rappelle que le plan de sondage de l'enquête Emploi est complexe et qu'il est particulièrement délicat de procéder à une estimation analytique des variances des estimateurs directs, d'autant plus que, par nature, on a à faire à des échantillons de grappes de (très) petite taille - et de taille aléatoire - dans chaque ZE. On opte pour une estimation de variance qui utilise les effets de sondage $deff_d$ calculés au niveau de la région à partir de l'intégralité de l'échantillon Emploi (voir partie 4.2). Ainsi, dès lors que $n_d \geq 1$, on postule

$$Var\left(\sum_{i\in s\cap d} w_i \cdot U_i \right) = deff_d \cdot \underset{SAS}{Var}\left(\frac{N_d}{n_d} \sum_{i\in s\cap d} U_i \right)$$

[12] Le contexte rend douteuse la validité des résultats asymptotiques, puisqu'on se trouve précisément dans un contexte de petites tailles d'échantillon. En fait, on a coutume de pratiquer les redressements avec des échantillons de quelques dizaines d'unités, les études pratiques montrant qu'on atteint assez vite les conditions dites "asymptotiques".

où $\underset{SAS}{Var}$ est la variance d'un échantillonnage aléatoire simple de taille n_d, où n_d représente le nombre d'individus répondants constaté en ZE d. En considérant la taille n_d comme fixée (raisonnement en variance conditionnelle à la taille de l'échantillon), l'estimateur le plus recommandable - malgré le biais dont il peut être affecté - de la variance de l'estimateur calé s'écrit donc :

$$\hat{V}\left(\hat{\bar{Y}}_{d,GREG}\right) = deff_d \cdot \frac{1}{n_d^2} \sum_{i \in s \cap d} \hat{U}_i^2$$

avec $\hat{U}_i = Y_i - \hat{B}_d^t X_i$ (la constante fait partie des régresseurs). Le fait que Y soit une variable qualitative (prenant donc exclusivement les valeurs 0 et 1) ne change rien à tous ces calculs.

3.2. Les principaux résultats

L'estimation par calage utilise des données auxiliaires individuelles $X_{d,i}$. Ces données servent en particulier à estimer le coefficient $\hat{B}_d$. Les conditions de mise en œuvre du calage local sont spécifiques et diffèrent de celles qui prévalent dans le cas de l'utilisation d'un modèle stochastique individuel (cf. partie 2.2.). Le pseudo-"modèle" (donc le calcul de $\hat{B}_d$) s'ajuste à partir de l'échantillon et de lui seul - donc pour cette étape, il faut et il suffit de disposer des variables auxiliaires au travers du questionnaire (de ce point de vue, pas de différence avec les modèles stochastiques). Pour former l'estimateur "petits domaines", il reste à connaître le vrai total de ces variables auxiliaires ZE par ZE (les marges locales, donc), mais à ce stade ultime il est essentiel de remarquer que l'on a pas du tout besoin d'en obtenir le détail pour chaque individu de la population ! Du point de vue de la quantité d'information nécessaire, le calage local est donc sensiblement moins exigeant que le modèle stochastique. Concrètement, le gain essentiel à ce niveau est le suivant : il n'est pas nécessaire que les fichiers auxiliaires contiennent l'identifiant du recensement - ce qui lève un obstacle majeur qui avait été mis en avant dans la partie 2.2. (il faut et il suffit d'une indication géographique permettant d'affecter chaque individu de la source auxiliaire à une ZE - concrètement on passe par le niveau commune, qui est toujours disponible en pratique).

Cela étant, à des fins de comparaison des méthodes, en particulier avec ce qui provient des modèles définis au niveau domaine, nous sommes partis des sept variables auxiliaires qui ont été isolées à la fin de la partie 2.1, plus la constante. Dans une optique individuelle, l'exploitation d'une variable de démographie d'entreprise ou celle d'un code Tabard n'est pas du tout naturelle et ne paraît pas opportune (à la limite, on pourrait imaginer l'utilisation d'une typologie à un niveau géographique très fin, mais ce n'est pas notre contexte). Par ailleurs, il n'y a qu'un seul individu dans tout l'échantillon Emploi qui ait moins de 20 ans, soit diplômé et demandeur d'emploi dans les catégories adéquates - ce qui ne permet pas un ajustement de modèle.
In fine, il y a donc 4 variables raisonnablement exploitables (plus la constante). Au travers du questionnaire, on sait si la personne interrogée est inscrite ou non à Pole Emploi, en revanche, on ne sait pas dire dans quelle catégorie elle a été classée ni si elle est en activité réduite. Les informations collectées par ailleurs dans le questionnaire permettraient probablement de déduire tout ou partie de ces informations complémentaires, mais nous ne nous sommes pas risqué à reconstruire une telle variable, qui serait de toute façon fragile : nous avons simplement abandonné toute référence aux catégories 1 à 3 et à l'activité réduite. Les informations sur la recherche d'un emploi et sur la vie en couple sont quant à elles récupérées à partir de questions ad hoc simples.
Finalement, cela conduit à considérer les 4 variables de calage suivantes - outre la constante :

- Recherche d'un emploi ;
- Ne pas vivre en couple ;
- Inscription DEFM et appartenir à la catégorie des hommes de 30 à 49 ans, peu diplômés ;
- Inscription DEFM et appartenir à la catégorie des hommes de 50 à 64 ans, peu diplômés ;

Néanmoins, il y a un grand nombre de ZE où l'une ou l'autre des variables de DEFM est intégralement nulle, c'est-à-dire que la valeur de la variable auxiliaire est 0 pour chaque individu de l'échantillon dans la ZE. S'agissant de populations relativement rares, ce n'est pas surprenant, d'autant plus que n_d est petit. Dans ce cas, le programme ne peut pas calculer de coefficient de régression $\hat{B}_d$ parce qu'il y a une matrice qui n'est pas inversible. On peut, soit limiter la production d'un estimateur calé au seul cas des ZE pour lesquels le calcul est possible, soit supprimer au cas par cas, en fonction de la ZE, les variables de calage à la source de ces problèmes[13].

Pour des raisons de simplicité dans la programmation, tout autant que par curiosité pour la mise en œuvre d'un programme nouveau, nous avons utilisé non pas Calmar mais une macro SAS produite par l'ONS (Grande-Bretagne) : ZE par ZE, nous avons ainsi obtenu l'estimateur par la régression GREG présenté au 3.1.

On considère les 348 ZE de France métropolitaine. On écarte en amont 61 d'entre elles pour cause d'effectif répondant jugé insuffisant, manifestement trop en contradiction avec le contexte asymptotique qui sous-tend la théorie du calage (moins de 50 répondants - y compris 10 ZE non couvertes). Sur les 287 restantes, on constate que 186 ne permettent pas de calage à partir des 4 variables initiales, dont 170 pour une raison claire de nullité complète d'un (au moins) des régresseurs[14]. Parmi les 101 ZE où un calage a pu être mené à terme avec la programmation utilisée, on obtient la distribution suivante de la variance d'échantillonnage de l'estimateur calé de la proportion de chômeurs par ZE présentée en partie 3.1.

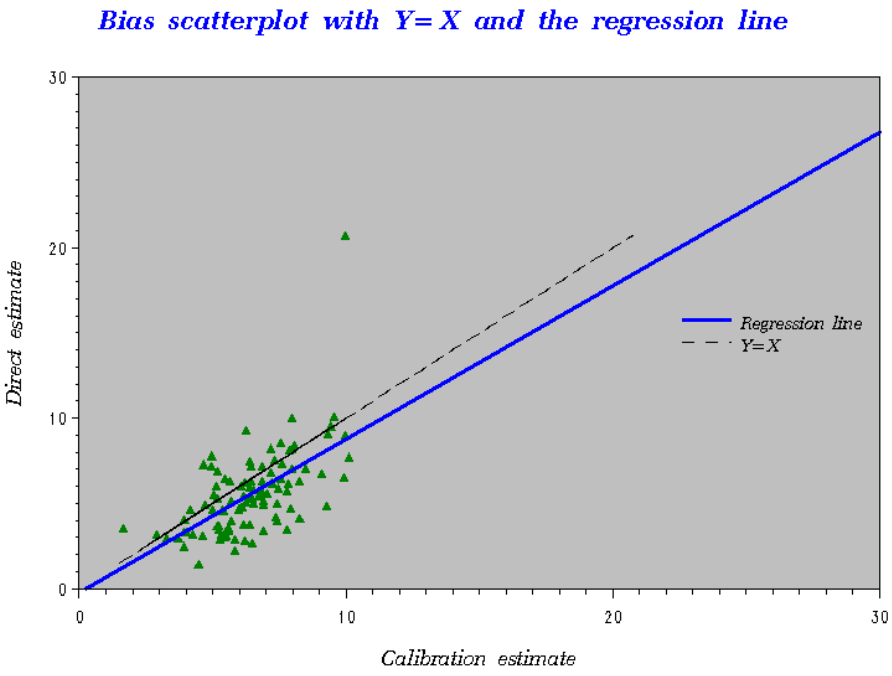
Quantile		Variance x 10^5
100%	**Max**	53.17
99%		46.33
95%		36.45
90%		20.12
75%	**Q3**	8.67
50%	**Median**	**4.18**
25%	**Q1**	2.46
10%		1.46
5%		1.09
1%		0.65
0%	**Min**	0.64

Cette distribution est à comparer avec celle de l'estimateur direct non calé - qui est disponible en partie 4.2 - mais sans oublier qu'on travaille ici sur un petit tiers de la population des ZE, lequel doit correspondre essentiellement aux ZE ayant les plus petites variances directes. Malgré cette limite, on constate, comme prévu, des améliorations très importantes de variance : dans certaines conditions, le calage sur des structures locales bien "explicatives" est susceptible d'améliorer beaucoup la précision. C'est une première technique d'estimation sur petits domaines qu'il ne faut pas oublier ! Par opposition à toutes les autres techniques (parties suivantes), le calage produit des estimations dites "directes", qui ne font donc jamais appel, en aucune manière, à des données collectées en dehors de la ZE concernée.

Partant de ces 101 ZE, chacune représentée par un point, le graphique portant en abscisse l'estimation calée et en ordonnée l'estimation directe laisse apparaître un léger décalage de la droite de régression par rapport à la droite $Y = X$. S'agissant ici d'un ensemble d'estimateurs ayant vocation à être asymptotiquement sans biais - même s'il est vrai que l'on peut douter d'être dans les conditions asymptotiques en question - ce résultat apparaît un peu sévère. Un facteur explicatif peut

[13] Il faudrait alors mettre en œuvre une programmation spécifique qui détecte en amont les régresseurs identiquement nuls et les exclut automatiquement du calage : nous ne l'avons pas fait.
[14] Ce qui se passe pour les 16 autres ZE est un peu plus mystérieux - très probablement y a-t-il des colinéarités d'une autre nature dans les régresseurs puisque l'inversion de la matrice ne peut pas se faire.

être lié au fait que l'estimateur direct auquel on se compare est en soi déjà légèrement biaisé, puisqu'il ne s'agit pas de l'estimateur de Horvitz-Thompson (post-stratification sur la taille de la ZE, poids nationaux issus d'un calage en une étape).



Toujours sur les 101 ZE où il y a eu calcul d'un estimateur calé, la sommation des estimations calées relatives au nombre total de chômeurs[15] apparaît sensiblement supérieure à la sommation des estimations directes (+ 20%). On obtient en effet :

Estimateur utilisé	Estimation totale
Calage	1 997 000
Enquête Emploi	1 667 000
Méthodologie actuelle Insee	1 542 000

L'ampleur de ces différences peut surprendre car les écarts de distribution entre estimations directes et estimations calées paraissent limités - du moins peut-on le ressentir comme tel à première vue. En réalité, les ZE qui se situent sous la diagonale, d'une part sont nettement plus nombreuses (72 sous la diagonale, 29 au-dessus), d'autre part sont les plus peuplées, ce qui éclaire le phénomène. Une explication plus profonde pourrait bien être liée à un phénomène d'hétérogénéité de l'information : le calage met en regard l'information sur la recherche d'un emploi telle qu'elle figure dans le questionnaire Emploi avec des structures locales issues du recensement. Or, ces informations ne sont pas aussi concordantes qu'on pourrait l'imaginer. Nous renvoyons à la partie 6.2.1 pour mieux comprendre : un souci similaire est en effet apparu avec une modélisation individuelle stochastique et il y a lieu de penser qu'un problème de même nature se pose ici. D'ailleurs, la surestimation affichée du calage par rapport à l'estimation directe joue bien en ce sens.

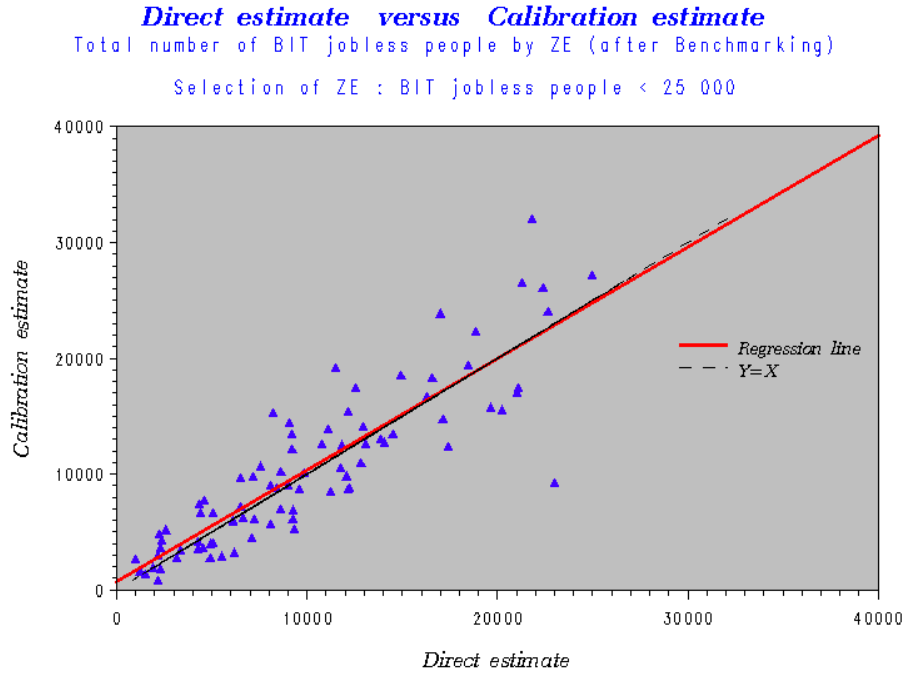
Après benchmarking - opération quelque peu violente mais nécessaire compte tenu du décalage de volume enregistré avec l'approche directe, à considérer comme une opération de correction de l'hétérogénéité - on est en mesure d'évaluer l'ampleur des écarts relatifs avec l'estimation (brute) de l'Insee :

$$100 \cdot \frac{\hat{Y}_{ZE,calage} - \hat{Y}_{ZE,INSEE}}{\hat{Y}_{ZE,INSEE}}$$

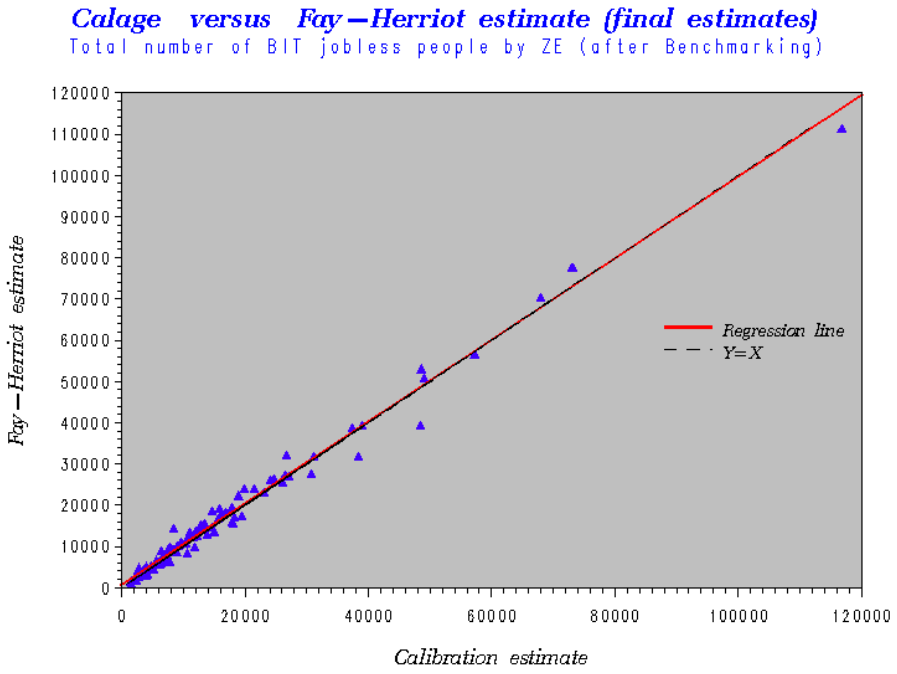
[15] L'estimation calée du nombre total de chômeurs s'obtient en partant de l'estimation calée de la proportion que l'on multiplie par la "vraie" taille de population de la ZE dans le champ (modulo ce qui a été dit sur l'aléa dû au recensement).

Quantile		Relative Gap
100%	**Max**	59.8
99%		38.9
95%		22.4
90%		17.7
75%	**Q3**	6.8
50%	**Median**	**-0.4**
25%	**Q1**	-11.3
10%		-20.1
5%		-25.0
1%		-46.6
0%	**Min**	-56.0

Après benchmarking, le test conclut qu'il n'y a plus d'écart significatif entre la diagonale et la droite de régression : on peut considérer que les estimateurs calés ayant subi un benchmarking en phase finale ne sont pas (ne sont plus…) biaisés. Le graphique suivant limite les risques d'effet optique en éliminant les ZE ayant plus de 25 000 chômeurs BIT, donc les plus grandes ZE.



Il nous a semblé intéressant de rapprocher les estimations obtenues avec le modèle de Fay et Herriot présenté en partie 4 et les estimations calées, après benchmarking de part et d'autre.



Si on supprime les deux variables de DEFM, dont on peut douter de la valeur ajoutée compte tenu de la rareté des populations qu'elles caractérisent, on peut effectuer cette fois les calages sur 209 ZE - ce qui au moins constitue une amélioration considérable en terme de champ couvert.

Quantile		PSI x 10^5
100%	**Max**	134.76
99%		74.83
95%		44.48
90%		35.75
75%	**Q3**	16.75
50%	**Median**	**6.46**
25%	**Q1**	3.50
10%		2.00
5%		1.36
1%		0.68
0%	**Min**	0.64

Les variances ont sensiblement augmenté du fait du retrait de ces deux variables de DEFM, surtout dans le haut de la distribution. La raison essentielle de ce phénomène vient de l'introduction de nombreuses ZE plutôt petites en taille (on supprime en effet les régresseurs qui sont identiquement nuls, ce qui survient plus couramment parmi les ZE où la taille d'échantillon est petite). Pour juger vraiment de la détérioration de qualité, on peut raisonner à champ constant : si on reprend les 101 ZE ayant participé au calage initial et qu'on produit la distribution de leurs variances obtenues à partir du modèle en conservant seulement 2 variables explicatives, alors on constate qu'il y a certes une augmentation de la variance, conformément à la théorie, mais que celle-ci reste faible, voire symbolique pour bon nombre de ZE.

Sur les 209 ZE traitées, concernant les effectifs de chômeurs estimés indépendamment par chaque méthode on trouve toujours des écarts importants entre l'approche directe et l'approche par calage, du même ordre de grandeur qu'avec le modèle complet (+ 22%).

Estimateur utilisé	Estimation totale
Calage	2 649 000
Enquête Emploi	2 170 000
Méthodologie actuelle Insee	2 078 000

Cela conforte l'hypothèse de l'hétérogénéité de l'information entre la source Emploi et la source recensement.

4. Le modèle de Fay et Herriot

4.1. Éléments de théorie

On note d l'identifiant de la ZE et $\hat{\overline{Y}}_d$ l'estimateur direct de la moyenne vraie $\overline{Y}_d$ dans la ZE d. Dans le cas présent, cet estimateur prend la forme d'un ratio $\frac{\hat{Y}_d}{\hat{N}_d}$ avec $\hat{Y}_d = \sum_{\substack{i \in s \\ i \in d}} w_i \cdot Y_i$ et $\hat{N}_d = \sum_{\substack{i \in s \\ i \in d}} w_i$.

On rappelle que N_d désigne, dans notre cas, la taille de la population de la ZE d ayant 15 ans ou plus au 31 décembre 2007 et vivant en ménage ordinaire. On dispose, pour chaque ZE d, d'informations (pseudo) exactes $\overline{X}_d$ constituant un vecteur à p dimensions et issues de sources externes décrites en partie 1.2. (recensement, DEFM, RMI, Sirene, etc.). On peut écrire, d'une part $\hat{\overline{Y}}_d = \overline{Y}_d + \varepsilon_d$ où ε_d est l'erreur d'échantillonnage, et d'autre part $\overline{Y}_d = B^t \cdot \overline{X}_d + v_d$ où v_d est une variable aléatoire s'interprétant comme un effet propre au domaine d. Cela nous conduit au modèle suivant (dit de Fay et Herriot) :

$$\hat{\overline{Y}}_d = B^t \cdot \overline{X}_d + v_d + \varepsilon_d$$

où B est un paramètre vectoriel inconnu, avec les hypothèses : $Ev_d = E\varepsilon_d = 0$, $Var(v_d) = \sigma_v^2$ et $Var(\varepsilon_d) = \psi_d$. Les aléas v_d et ε_d sont supposés être mutuellement indépendants, c'est-à-dire que toutes les matrices de variance sont diagonales. On considère par ailleurs que les variances d'échantillonnage ψ_d sont connues (en pratique, il faut utiliser un estimateur…). Matriciellement, ce modèle s'écrit

$$\hat{\overline{Y}} = \overline{X} \cdot B + Z \cdot v + \varepsilon$$

La théorie de la prédiction linéaire sans biais optimale conduit à retenir l'estimateur suivant :

$$\hat{\overline{Y}}_d^H = \hat{B}^t \cdot \overline{X}_d + \hat{v}_d = \hat{\gamma}_d \cdot \hat{\overline{Y}}_d + (1 - \hat{\gamma}_d) \cdot \hat{B}^t \cdot \overline{X}_d$$

où $\hat{B}$ est l'estimateur des moindres carrés généralisés, soit $\hat{B} = \left(\sum_d \frac{\overline{X}_d \cdot \overline{X}_d^t}{\hat{\sigma}_v^2 + \psi_d} \right)^{-1} \cdot \sum_d \frac{\overline{X}_d \cdot \hat{\overline{Y}}_d}{\hat{\sigma}_v^2 + \psi_d}$ et

$$\hat{\gamma}_d = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \psi_d}$$

Si on ajoute l'hypothèse de normalité des aléas, l'estimateur $\hat{\sigma}_v^2$ du paramètre σ_v^2 peut être obtenu par maximum de vraisemblance (sinon on peut l'estimer par la méthode des moments).

Cet estimateur est une combinaison linéaire d'une part de l'estimateur direct, sans biais (ou peu biaisé) mais de forte variance, et d'autre part de l'estimateur synthétique $\hat{B}^t \cdot \overline{X}_d$, biaisé mais de faible variance. On est en mesure de fournir une estimation de l'erreur de $\hat{\overline{Y}}_d^H$ définie comme $E\left(\hat{\overline{Y}}_d^H - \overline{Y}_d\right)^2$. Attention, sa définition associe les deux natures d'aléa en jeu. On montre :

$$E\left(\hat{\overline{Y}}_d^H - \overline{Y}_d\right)^2 = \gamma_d \cdot \psi_d + O\left(\frac{1}{m}\right)$$

où m désigne le nombre total de ZE participant à l'estimation du modèle (donc $m = 338$ si on retient toutes les ZE où $n_d > 0$). Ainsi, lorsque m est grand, le gain de précision par rapport à l'estimateur direct est de l'ordre de grandeur de γ_d. Ce gain peut être considérable.

4.2. Les principaux résultats

On mobilise les 7 variables auxiliaires $\overline{X}_d$ listées en fin de partie 2.1 (plus la constante).

La fixation des variances d'échantillonnage ψ_d pose un véritable problème, parce que l'échantillonnage de l'enquête Emploi est fort complexe, qu'il fait intervenir des degrés au sein desquelles les tailles d'échantillon valent 1, que les tailles d'échantillon de logements par ZE sont très sensibles au hasard, et même qu'il y a une probabilité non négligeable que les petits domaines ne soient pas du tout couverts ! Nous n'avons pas cherché à estimer de manière directe les variances d'échantillonnage au niveau ZE, d'une part du fait de ces complications techniques qui nécessiteraient un investissement spécifique - d'ailleurs avec un grand risque d'échec - d'autre part parce que l'estimation de variance a déjà été faite aux niveaux national et régional. Aussi, nous avons repris les effets de sondage ($deff$) estimés au niveau de chaque région[16] en supposant qu'ils s'appliquent en l'état à chaque ZE de la région. C'est évidemment une approche discutable que de reprendre le *deff* régional, en l'absence de véritable justification technique formelle. Il y a en premier lieu un souci de comparabilité des conditions d'échantillonnage région / ZE au niveau de la variabilité des tailles d'échantillon : contrairement à la région, la ZE n'est pas une strate de tirage de l'échantillonnage national, c'est véritablement un domaine et donc on ne controle pas les tailles d'échantillon dans la ZE comme on les controle au niveau de la région. Cette considération laisse penser que l'on surestime le $deff$ propre à la ZE. Un autre motif de confusion tient au fait que notre estimateur direct est un ratio qui cale "seulement" sur la population de la ZE alors que le calcul des $deff$ régionaux porte sur un estimateur complexe du nombre total de chômeurs qui inclut un calage plus riche (de ce point de vue, peut-être-a-t-il une légère tendance à sous-estimer le *deff* propre à la ZE).
Néanmoins, cette approche a le mérite de la simplicité et surtout de la stabilité numérique : ce dernier point et déterminant, car il offre une forme de robustesse très appréciable alors même qu'on a à faire à des (très) petites tailles échantillon (d'autant plus, on le rappelle, qu'il s'agit d'un échantillonnage à plusieurs degrés et *in fine* en grappes - quelques grappes sont tirées par ZE, souvent moins de 10). Ainsi, une estimation de variance "directe" donne lieu à une variabilité d'estimation des vraies variances considérable au niveau local - ce qui conduit à un échec. Il faut ajouter que la théorie s'appuie sur une vraie variance et non une variance estimée, ce qui relativise les erreurs qui sont imputables spécifiquement à la méthodologie d'estimation de variance.

Précisément, on a affecté à chaque ZE le *deff* moyen de sa région (noté $deff_d$) obtenu à partir des estimations trimestrielles portant sur différents trimestres entourant le T1 de 2007, soit du T4 de 2005 compris au T2 de 2008 compris. La (vraie) variance que l'on aurait avec un sondage aléatoire simple de même taille a pour sa part été calculée en considérant la proportion $\tilde{p}_d$ déduite du nombre total de chômeurs estimés par ZE avec la méthodologie actuelle de l'Insee - et seulement lorsque ce taux est compris entre 3% et 16%. Dans les ZE où ce taux s'avère inférieur à 3% ou supérieur à 16%, on a

[16] Estimation effectuée par Karim Moussallam, UMS.

retenu le taux national. Cette méthode permet de bénéficier d'un estimateur de nature synthétique, ce qui évite d'introduire une variabilité pénalisante dans les estimations de variance (utiliser les estimations directes $\hat{\bar{Y}}_d$ donne des résultats catastrophiques - dont des variances nulles - et l'ajustement mécanique du modèle conduit d'ailleurs à $\hat{\sigma}_v^2 = 0$!). Ainsi :

$$\hat{\psi}_d = deff_d \times \frac{\tilde{p}_d(1-\tilde{p}_d)}{n_d}$$

Voici des éléments sur la distribution des $\hat{\psi}_d$ si on se limite aux ZE ayant au moins 50 individus répondants :

Quantile		Variance X 10^4
100%	**Max**	**19.66**
99%		14.94
95%		10.76
90%		9.51
75%	**Q3**	6.75
50%	**Median**	**4.10**
25%	**Q1**	2.31
10%		1.13
5%		0.83
1%		0.44
0%	**Min**	**0.39**

Ce tableau affiche directement la variance exprimée en "points de pourcentage" (4,1 points de pourcentage pour une ZE médiane : soit P connu à $\pm 2.\sqrt{4.1} \approx \pm 4$ points de pourcentage, avec 95 chances sur 100 - sous réserve de validité des hypothèses de loi de Gauss…).
La valeur minimale correspond à la plus grande ZE, qui est celle de la ville de Paris. Dans une ZE, le CV (estimé) associé à cette variance s'obtient en divisant la racine carrée du $\hat{\psi}_d$ ainsi calculé par le taux estimé $\tilde{p}_d$, exprimé directement en pourcentage (exemple : si $\tilde{p}_d$ vaut 6 et $\hat{\psi}_d$ vaut 9, le CV vaut 50%). Numériquement, les CV ainsi estimés varient entre 11,2% et 616,6 %, avec $Q1 = 33{,}8\%$, une médiane de $51{,}1\%$ et $Q3 = 72{,}5\%$ - ce qui justifie bien la mise en œuvre de techniques spécifiques "petits domaines".

La procédure informatique de SAS utilisée est la PROC GLIMMIX, spécialisée dans le traitement des modèles linéaires mixtes généralisés. Cette procédure offre de nombreuses options mais elle est extrêmement gourmande en mémoire vive. L'ajustement du modèle a été effectué en se restreignant aux ZE ayant au moins 50 individus répondants, afin de limiter la dispersion des variances d'échantillonnage (ce point sera développé plus loin). Sont alors concernées 287 ZE. L'ajustement a utilisé la technique du maximum de vraisemblance restreint, qui permet d'effectuer *in fine* une correction réduisant le biais des estimateurs obtenus. Nous avons également demandé une correction du nombre de degrés de liberté dans les statistiques de test appelée correction de Kenward et Roger, qui apparaît recommandable sur le plan théorique (bien qu'elle soit grande consommatrice de ressources informatiques).

Un premier ajustement du modèle a été effectué en mobilisant l'ensemble des variables jugées potentiellement explicatives. On constate qu'en fait, dans le contexte du modèle complet, deux (ou trois) d'entre elles n'apportent rien de significatif si on en juge d'après le test de Student associé à chaque variable explicative. On a ensuite engagé un processus de réduction du nombre de variables explicatives, en supprimant à chaque étape la variable la moins significative du modèle ajusté à l'étape précédente (en "sautant" le modèle à 6 variables). On peut finalement juger de la qualité de

l'ajustement d'après différents critères classiques, en fonction du nombre de variables explicatives participant au modèle :

Critère de qualité utilisé	7 variables	5 variables	4 variables	3 variables	2 variables	1 variable
-2. Log vraisemblance restreinte	1280.2	**1268.1**	1273.3	1280.1	1292.9	1297.4
AIC	1282.2	**1270.1**	1275.3	1282.1	1294.9	1299.4
BIC	1285.9	**1273.1**	1279.0	1285.7	1298.6	1303.0
Chi-2 généralisé	**281.8**	284.1	286.9	288.7	292.2	295.6

Pour chacun de ces critères, on peut dire que le modèle est d'autant préférable que le critère est numériquement petit. *In fine*, il apparaît que sur 3 des 4 critères repris ici, 3 sont minimum pour le modèle à 5 variables - le critère du chi2-généralisé étant pour sa part minimum pour le modèle complet (7 variables).

L'algorithme du maximum de vraisemblance restreint donne les estimations $\hat{\sigma}_v^2$ (on fait donc l'hypothèse de résidus gaussiens). Le tableau ci-dessous résume les valeurs obtenues en fonction du modèle ajusté. L'estimateur s'avère relativement stable, ce qui est rassurant. L'écart-type (estimé) de cet estimateur est égal à 0.33 ou 0.34 en toutes circonstances, ce qui le rend significatif mais rappelle néanmoins que la qualité de l'estimation de $\hat{\sigma}_v^2$ ne peut pas être considérée comme très bonne (CV de l'ordre de 30%).

Statistique	7 variables	5 variables	4 variables	3 variables	2 variables	1 variable
Estimateur $\hat{\sigma}_v^2$	1.137	**1.111**	1.090	1.088	1.085	1.197
Ecart-type estimé de $\hat{\sigma}_v^2$	0.33	**0.33**	0.33	0.33	0.33	0.34

On s'en remet à l'arbitrage mécanique des critères classiques et on choisit donc de conserver 5 variables explicatives (plus la constante) en abandonnant définitivement le taux de solde des établissements entre 2000 et 2006 et la proportion de la population en catégorie Tabard dite "SEMAG02". Dans ces conditions, le modèle définitif donne les résultats suivants :

Solutions for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	-4.70	1.64	134.7	**-2.86**	0.0050
t_rech_oui	0.65	0.25	223.2	**2.64**	0.0089
t_couple_2	0.12	0.049	126.3	**2.56**	0.0117
t_age15_19HdiplBIT	111.22	50.77	207.1	**2.19**	0.0296
t_age50_64Hnondipl	3.09	1.85	217.7	**1.67**	0.0961
t_age30_49HnondiplBIT	-3.27	2.76	190.9	**-1.19**	0.2370

On constate qu'il est possible de rencontrer une situation où deux variables sont non significatives du point de vue du test de Student (p-value seuil à 5%) mais où néanmoins les critères de choix de modèle concluent à l'intérêt de conserver chacune d'elles. La valeur décimale du nombre de degrés de liberté est explicable par l'utilisation de l'option de correction de Kenward et Roger. On rappelle que la très grande valeur du coefficient associé à la variable *t_age15_19HdiplBIT* tient au fait que les valeurs de la variable en question sont numériquement très faibles.

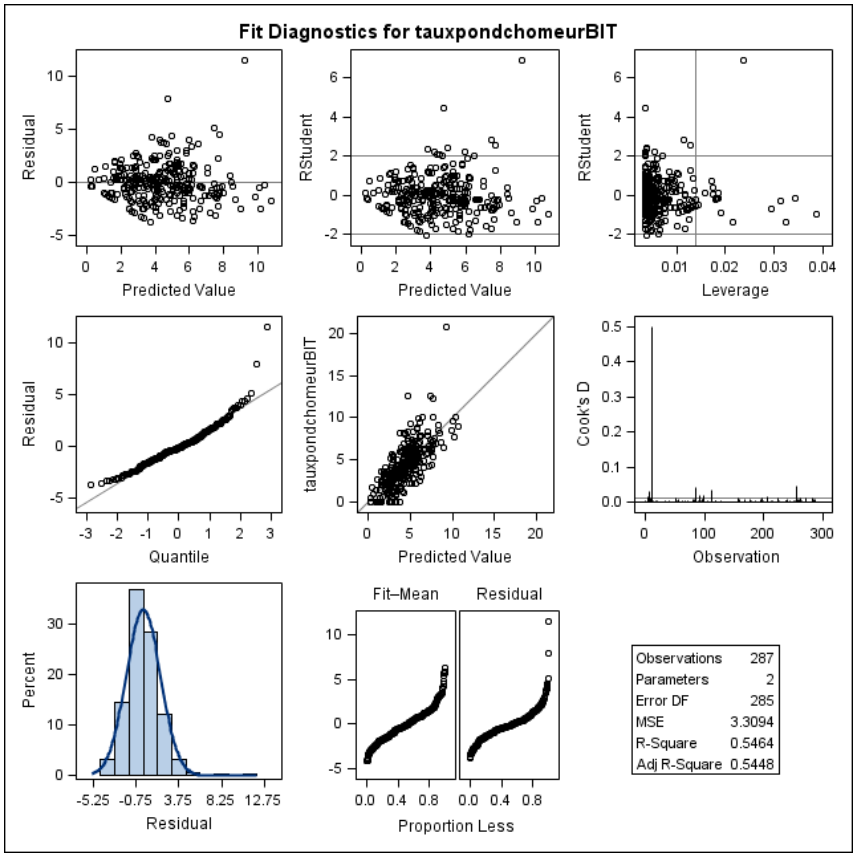
Le système itératif de la Proc Glimmix consiste à estimer le vecteur des paramètres $\hat{B}$ et le paramètre réel $\hat{\sigma}_v^2$ par un algorithme itératif de type Newton où la fonction objectif à minimiser est - 2 fois la log vraisemblance restreinte. La convergence a lieu très rapidement, après 4 itérations seulement.

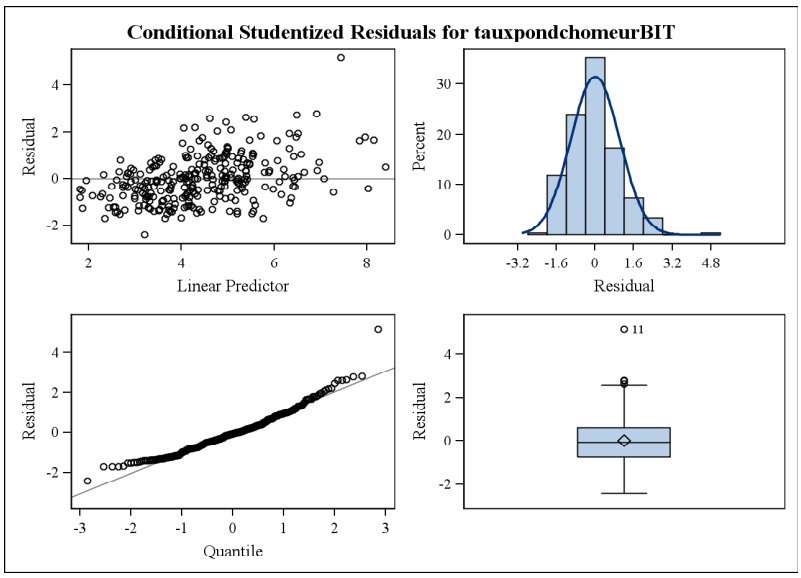
L'estimateur de Fay et Herriot donné ici en pourcentage (avant toute opération de benchmarking) a la distribution suivante (on rappelle qu'on se limite aux 287 ZE ayant au moins 50 répondants).

Quantile		Estimation FH
100%	**Max**	8.40
99%		8.02
95%		**6.51**
90%		6.16
75%	**Q3**	5.21
50%	**Median**	**4.34**
25%	**Q1**	3.47
10%		2.85
5%		**2.58**
1%		1.86
0%	**Min**	1.82

Ainsi, 90% des taux estimés se situent entre 2,6% et 6,5%. L'estimation minimale (1,8%) est celle de la ZE de Rodez et l'estimation maximale (8,4%) celle de la ZE de Roubaix-Tourcoing.

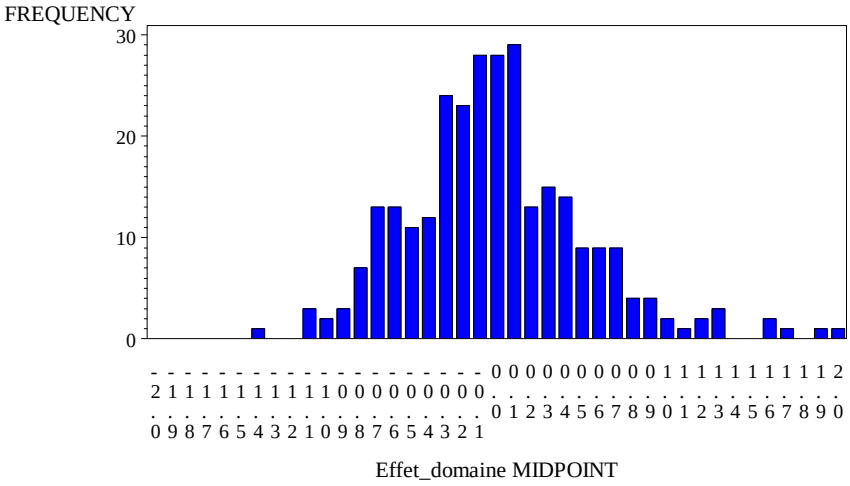
Pour apprécier la qualité de l'ajustement, on peut s'appuyer sur les graphiques suivants (produisant respectivement des résidus "bruts" et des résidus studentizés). D'après ces résultats, il nous semble qu'il n'y a pas lieu de rejeter le modèle à 5 variables. Les quelques résidus atypiques qui ressortent ne remettent pas le modèle en cause mais rappellent qu'effectivement il y a quelques ZE dans lesquelles l'enquête Emploi enregistre un nombre tout particulièrement fort de chômeurs - ce que le modèle n'arrive pas bien à reproduire. La forme évasée du nuage des résidus en fonction du taux prédit est en partie liée au fait que pour certaines ZE, on a une proportion de chômeurs constatée égale à 0, auquel cas $\hat{U} = p - \hat{p} = -\hat{p}$, d'où cette droite de pente -1 qui se distingue dans le graphique.





La ZE qui possède le résidu le plus atypique et qui se distingue très clairement dans les graphiques ci-dessus (résidu studentisé égal à 5,2) affiche une proportion de chômeurs estimée par l'enquête Emploi considérable (20,7% - pour un CV estimé de 45,6%) et le modèle propose un gamma égal à 0,13, ce qui conduit à une estimation localisée finale de 7,4% dans cette ZE.

Voici la distribution des effets domaines prédits, soit $\hat{v}_d$:



Les effets domaines se situent manifestement entre -1,5 et 2 points de pourcentage. L'allure générale de la distribution semble compatible avec l'hypothèse initiale de normalité de ces effets aléatoires.

La pondération γ_d est un "output" particulièrement intéressant. Le tableau qui suit donne lieu à la distribution du coefficient γ_d correspondant au modèle définitif . Dans un peu plus de 90% des cas, il apparaît que la priorité est donnée à l'estimateur synthétique. Dans la moitié des ZE, l'estimateur direct issu de l'enquête Emploi contribue à plus de 20% dans la valeur de l'estimation composite finale de Fay et Herriot - ce qui est loin d'être négligeable. La contribution maximale de l'estimateur direct, soit 74%, est celle de la ZE de Paris, ce qui n'est pas surprenant compte tenu de la taille de l'échantillon Emploi répondant dans cette ZE ($n_d = 2301$), pour un CV de 11,3% ($\hat{\psi}_d = 0{,}4$). La contribution minimale de l'estimation directe est celle d'une ZE pour laquelle on a (comme on pouvait s'y attendre) $n_d = 50$ et CV = 71% ($\hat{\psi}_d = 19{,}7$).

Quantile		Gamma
100%	**Max**	0.74
99%		0.71
95%		**0.57**
90%		0.49
75%	**Q3**	0.33
50%	**Median**	**0.21**
25%	**Q1**	0.14
10%		0.10
5%		**0.09**
1%		0.07
0%	**Min**	0.05

On peut se risquer à un calcul de coin de table pour apprécier la réduction d'erreur "en moyenne" par rapport à l'estimateur direct : une ZE "médiane" aura une variance d'échantillonnage de l'ordre de 4 points de pourcentage et disons un gamma de 20%. La variance de l'estimateur de Fay et Herriot sera donc de l'ordre de 0.8, soit un écart-type de 1 point de pourcentage, en arrondissant. La vraie valeur devrait donc être connue "en moyenne" à 2 points de pourcentage près - ainsi on divise par deux la largeur des intervalles de confiance par rapport à l'estimateur direct. *Attention* : nous rappelons que l'erreur au sens du modèle associe l'aléa de sondage et l'erreur stochastique de modélisation. Elle n'est donc pas de même nature que l'erreur d'échantillonnage !!!

Il est intéressant de noter que la sensibilité de l'estimateur $\hat{\sigma}_v^2$ aux estimations $\hat{\psi}_d$ est forte - ce qui n'est pas rassurant, et en tout cas cela nous incite fortement à stabiliser les variances estimées. Par simple curiosité, on peut imposer - fictivement - une valeur constante à tous les $\hat{\psi}_d$ et regarder comment évoluent $\hat{\sigma}_v^2$ et $\hat{\gamma}_d$ (qui est alors constant) quand on fait varier cette constante. En se limitant par exemple à 150 ZE, on obtient :

Valeur de la constante $\hat{\psi}_d$	$\hat{\sigma}_v^2$	$\hat{\gamma}_d$
8	0.07	0.01
7.5	0.57	0.07
7.25	0.82	0.10
7	1.07	0.13
6.75	1.31	0.16

Pratiquement, cela signifie que s'il y a une proportion trop importante des ZE qui ont des variances d'échantillonnage élevées, toute la variabilité des $\hat{\bar{Y}}_d$ est concentrée dans l'aléa d'échantillonnage ε_d - donc dans les $\hat{\psi}_d$ - et le paramètre σ_v^2 du modèle est alors numériquement (très) faible. Il apparaît que si les variances d'échantillonnage sont trop dispersées, l'effet est le même. C'est pourquoi il faut ajuster le modèle en excluant d'emblée les ZE dans lesquelles la taille d'échantillon est trop faible. Nous avons mis une limite à 50 observations : avec ce seuil, on obtient des résultats probants et si d'aventure on ne fait pas cette présélection et qu'on retient l'ensemble des ZE, l'ajustement par SAS conduit à $\hat{\sigma}_v^2 = 0$ parce que la variabilité du modèle s'avère être (presque) entièrement attribuable à l'erreur d'échantillonnage.

On trouvera en partie 7 (modèle EBLUP-B) les résultats d'une autre modélisation qui permettent de conforter ces conclusions et qui en constituent, de fait, une forme de validation.

Le tableau suivant fournit la distribution des écarts absolus entre l'estimateur direct de la proportion de chômeurs et la proportion estimée par le modèle de Fay et Herriot. Comme on pouvait s'y attendre, il y a des corrections très fortes apportées par le modèle.

Quantile		Ecart absolu
100%	**Max**	13.36
99%		6.2
95%		3.5
90%		2.2
75%	**Q3**	0.9
50%	**Median**	**-0.1**
25%	**Q1**	-1.2
10%		-2.4
5%		-2.8
1%		-3.6
0%	**Min**	-4.0

Par ailleurs, puisqu'il existe une méthode en place appliquée par l'Insee, on s'interroge naturellement sur l'importance des perturbations qu'occasionnerait la mise en œuvre d'une estimation de Fay et Herriot. On trouvera ci-dessous la distribution des écarts, au sens relatif $\frac{\hat{\bar{Y}}_{ZE,FH} - \hat{\bar{Y}}_{ZE,Insee}}{\hat{\bar{Y}}_{ZE,Insee}}$:

Quantile		Ecart relatif
100%	**Max**	0.55
99%		0.54
95%		0.25
90%		0.17
75%	**Q3**	0.08
50%	**Median**	**-0.04**
25%	**Q1**	-0.14
10%		-0.22
5%		-0.27
1%		-0.39
0%	**Min**	-0.44

Il y a environ 10% des ZE pour lesquelles la méthode de Fay et Herriot amènerait à faire évoluer la proportion de chômeurs (donc le nombre total de chômeurs) de plus de 25% en valeur absolue par rapport à la statistique officielle.

Il est important de comparer la somme au niveau national des estimations des effectifs de chômeurs issues du modèle de Fay et Herriot (en fait au niveau pseudo national puisqu'on n'a pas toutes les ZE - on rappelle que les plus petites sont manquantes) et l'estimation directe issue de l'enquête Emploi. L'estimation directe par ZE issue de l'enquête Emploi a été effectuée de manière post-stratifiée en calant sur la taille de la population des ZE dans le champ emploi[17] (d'après les données du recensement 2007), soit

[17] Rappel : ensemble des personnes ayant au moins 15 ans au 31 décembre 2007 et vivant en ménage ordinaire.

$$N_{ZE} \cdot \frac{\hat{Y}_{ZE}}{\hat{N}_{ZE}}$$

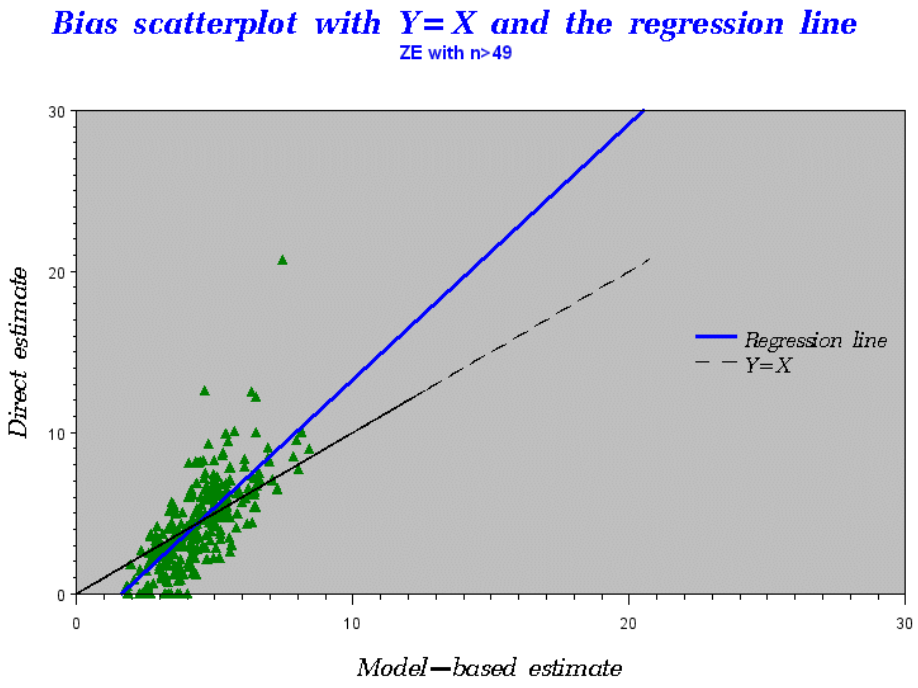
Les poids utilisés résultent certes d'un calage, mais c'est un calage conçu au niveau national, lequel n'a aucune vertu de réduction de variance quand on se place à un niveau local. Au niveau ZE, c'est donc un estimateur très simple qui reste fruste, mais qui est néanmoins sensiblement plus stable que l'estimateur de Horvitz-Thompson utilisant la pondération "nationale". Quand on somme au niveau (pseudo) national, l'estimateur devient évidemment très stable. On obtient, sur les 287 ZE concernées

Estimateur (pseudo) national	Estimation totale
Fay et Herriot	2 339 000
Enquête Emploi	2 340 000
Méthodologie actuelle Insee	2 305 000

La proximité des estimations respectives de Fay et Herriot et de l'estimateur direct est ici exceptionnelle - probablement faut-il y voir un heureux coup du sort pour atteindre ce niveau de similitude, mais en tout état de cause c'est une forme de validation de l'approche de Fay et Herriot.

Le graphique suivant permet de juger du biais de l'estimateur du seul point de vue de l'aléa d'échantillonnage. On compare la distribution de l'estimateur direct, sans biais, avec celle de l'estimateur de Fay et Herriot, dont la propriété théorique d'absence de biais ne tient que si on prend en compte à la fois l'aléa de modèle et l'aléa d'échantillonnage - et qui plus est sous l'hypothèse fondamentale que le modèle est "juste". Si le nuage de points se répartit de manière "harmonieuse" et équilibrée le long de la droite $Y = X$, il y a une forte présomption d'absence de biais.

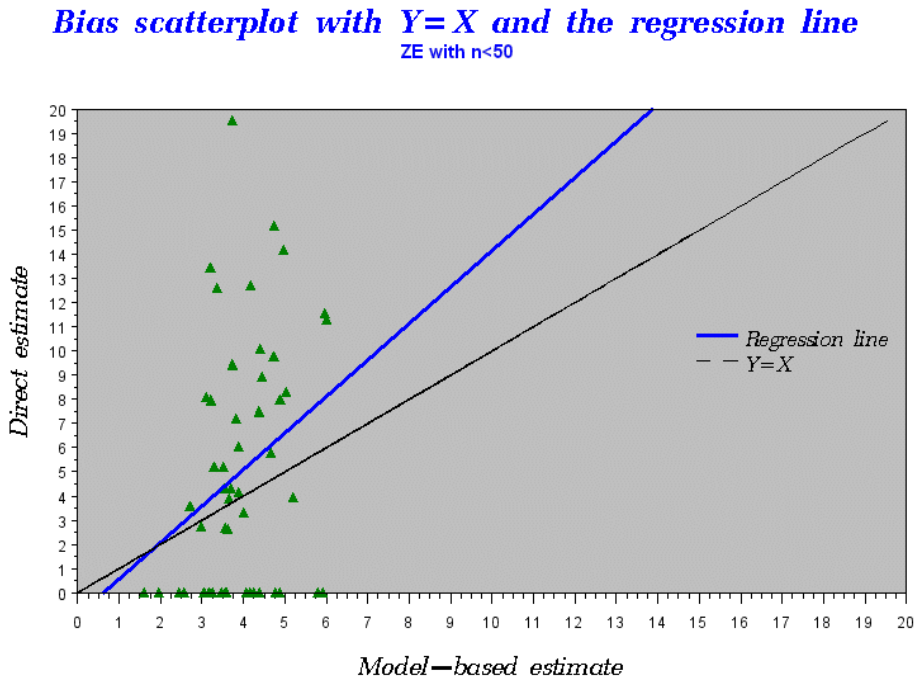
En l'occurrence, dans notre contexte, il y a un décalage, qui n'est certes pas considérable mais que l'on ne peut ignorer. Un test classique conclut d'ailleurs à une différence significative entre la droite $Y = X$ et la droite de régression. On a à faire à un phénomène de resserrement de la distribution, dit "Shrinkage", qui est assez classique dans les procédures d'estimation utilisant des estimateurs ayant une composante synthétique.



L'estimation dans les 61 ZE non traitées par le modèle s'effectue tout simplement en optant pour l'estimateur synthétique, qui s'obtient en imposant $\hat{v}_d = 0$. Ainsi, la proportion de chômeurs dans la ZE d dans laquelle $n_d < 50$ est égale à

$$\hat{\bar{Y}}_d^{SYN} = \hat{B}^t \cdot \bar{X}_d$$

Cette stratégie concerne en particulier (mais sans qu'on ne les distingue spécifiquement) les 10 ZE dans lesquelles on avait initialement $n_d = 0$. Par construction, les estimateurs synthétiques sont sensiblement moins dispersés que les estimateurs directs, et donc le phénomène de shrinkage est très marqué, comme on le voit dans le graphique ci-dessous réservé aux 61 ZE en question :



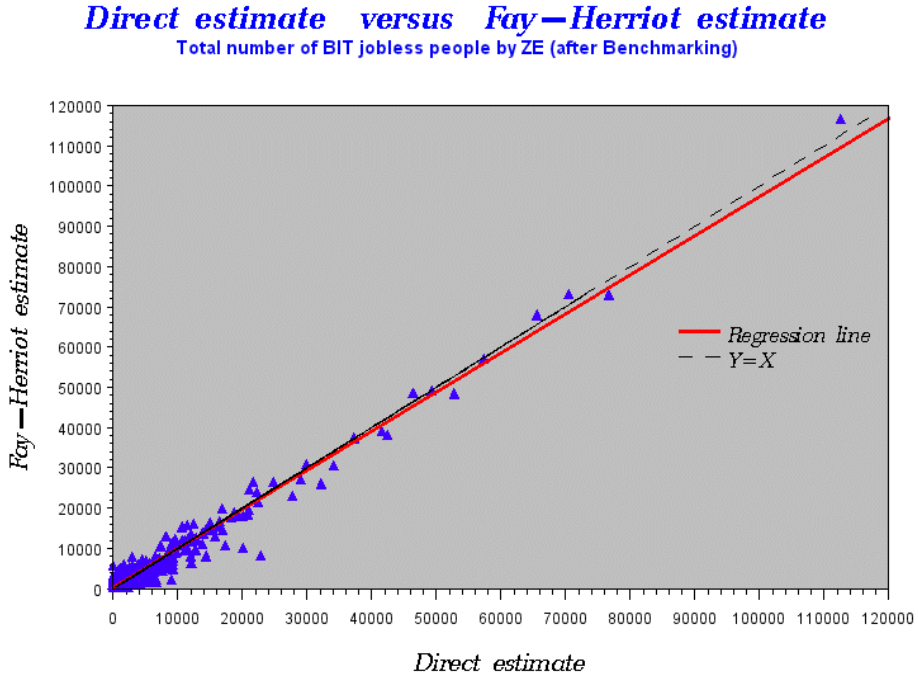
Lorsqu'on considère l'intégralité des 348 ZE de France métropolitaine, et que l'on s'intéresse de nouveau à l'estimation nationale obtenue à partir des différentes méthodes concurrentes, on trouve :

Estimateur (pseudo) national	Estimation totale
Fay et Herriot (ou synthétique)	2 432 000
Enquête Emploi	2 436 000
Méthodologie actuelle Insee	2 408 000

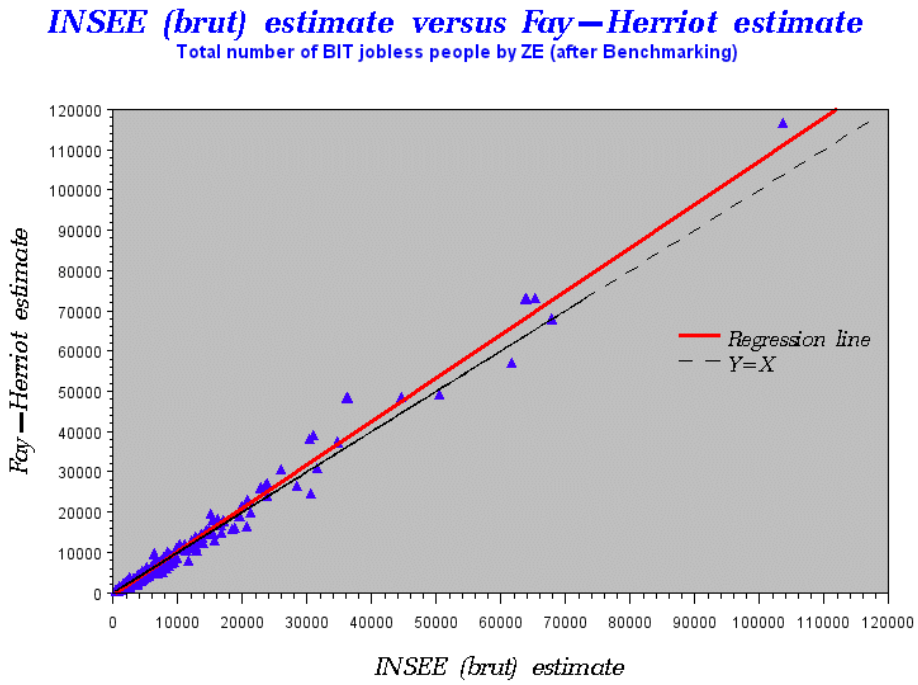
On rappelle que l'estimateur direct est ici post-stratifié sur la taille de population de chaque ZE- l'estimation nationale "officielle" - après correction de non-réponse et calage - étant de 2 416 000 chômeurs. On retrouve donc une proximité très satisfaisante avec l'estimation directe. Cela étant, pour des raisons de cohérence dans la diffusion, nous avons voulu que l'on retrouve exactement l'estimation nationale de l'enquête Emploi telle qu'elle ressort de l'exploitation du fichier diffusé avec les poids définitifs. Pour ce faire, un calage par simple ratio a été effectué *in fine* sur l'effectif cible de 2 416 000 chômeurs : si la procédure précédente a conduit à des estimations $\hat{Y}_d$, l'estimation finale est

$$\hat{\hat{Y}}_d = 2\,416\,000 \cdot \frac{\hat{Y}_d}{\sum_{ZE} \hat{Y}_{ZE}}$$

Le graphique suivant porte de nouveau sur la comparaison entre l'estimation directe locale - sans biais et de forte variance - et l'estimation de Fay et Herriot ou synthétique après benchmarking $\hat{\hat{Y}}_{ZE}$ - biaisée en toute rigueur mais de plus faible variance. Ce graphique concerne l'ensemble des 348 ZE.



Enfin, il a paru naturel de comparer les estimations finales de Fay et Herriot (ou synthétiques) avec celles qui résultent de l'application par l'Insee de la méthodologie actuelle d'estimation du chômage localisé (ensemble des 348 ZE). On vérifie là encore que la relation entre les deux distributions reste visuellement proche d'une relation linéaire, ce qui est rassurant - en rappelant qu'un nuage de points ayant cette allure n'empêche pas que l'on trouve des écarts relatifs sensibles pour certaines ZE (cf. supra).



Puisqu'on a estimé des variances d'échantillonnage $\hat{\psi}_d$ pour l'estimateur direct, on peut estimer des intervalles de confiance théoriques à 95% du nombre de chômeurs en formant $\left[\hat{Y}_d - 2 \cdot N_d \cdot \sqrt{\hat{\psi}_d} \;\;,\;\; \hat{Y}_d + 2 \cdot N_d \cdot \sqrt{\hat{\psi}_d}\right]$ et constater l'appartenance ou non de l'estimateur de Fay et Herriot à cet intervalle. On vérifie, pour le modèle à 5 variables, que sur les 338 ZE permettant un calcul d'estimateur direct, 7 ZE donnent lieu à une estimation inférieure à la borne basse, mais aucune ZE ne se trouve estimée au-dessus de la borne haute. Le tableau suivant liste les ZE hors de l'intervalle théorique.

psi	petitn	poptot2007	Nbchomeur_FH	Nbchomeur_direct	Binf	Bsup
7.6227	129	110795	8245	22975	16857	29093
5.4673	82	46559	3018	5708	3531	7885
10.6929	61	52739	2437	6654	3205	10103
4.0394	74	27571	1487	2748	1640	3856
25.4167	29	49685	1590	6687	1677	11697
21.3970	25	71815	2417	9070	2426	15714
11.1477	41	12212	454	2386	1571	3201

On constate qu'il s'agit de petites ou toutes petites ZE. Cet exercice doit être considéré avec beaucoup de prudence et trouve ses limites. En effet, d'une part on rappelle que l'estimation de variance est en soi particulièrement délicate - surtout dans ces conditions extrêmes - ensuite il est fort périlleux de postuler une loi de Gauss quand la taille de l'échantillon est trop petite et/ou la vraie valeur trop proche de zéro - ce qui conduit par exemple à une borne inférieure de l'intervalle estimé négative (ce qui est hélas souvent le cas...). La dissymétrie du résultat (entre borne inf et borne sup) est surprenante, surtout dans ce sens. L'ensemble de ces considérations laisse augurer une inadéquation de l'intervalle de confiance, le plus critiquable étant l'hypothèse de la loi de Gauss. Néanmoins, dans une version plus conforme à la réalité, la borne inférieure et la borne supérieure devraient avoir plutôt tendance à augmenter : on aurait donc davantage d'estimations "petits domaines" en dessous de la borne inférieure, et pas davantage au-dessus…ce qui accroitrait encore cette curieuse dissymétrie.

Néanmoins, la méthode devrait être correcte pour les "grosses" - voire les moyennes - ZE : et force est de constater que pour ces ZE là, l'estimateur FH se situe bien dans l'intervalle de confiance (qui plus est, il presque toujours sensiblement éloigné des deux bornes) - ce qui est un point tout à fait positif et encourageant pour l'estimation de Fay et Herriot.

Voici, pour information, le cas des 10 ZE ayant les plus grandes tailles d'échantillon :

psi	petitn	poptot2007	Nbchomeur_FH	Nbchomeur_direct	Binf	Bsup
0.39630	2301	1846612	117523	113485	90235	136735
0.40748	1927	1416269	68405	66105	48024	84186
0.44426	1721	1375232	73485	77292	58959	95625
0.84573	1233	898256	57590	57805	41284	74326
0.60790	1215	622351	48759	53207	43502	62912
0.82991	1177	975347	49420	49682	31911	67453
0.48860	1118	592621	38580	42801	34516	51086
1.05513	1077	918988	73720	71031	52151	89911
0.74452	1065	693852	37636	37557	25583	49531
0.84029	1057	818088	48879	46787	31789	61785

5. Le modèle de Poisson

5.1. Éléments de théorie

Dans une ZE d, le vrai nombre de chômeurs N_d^c est toujours une grandeur entière, strictement positive. Son estimation naturelle $\hat{N}_d^c$, tel qu'elle ressort de l'enquête Emploi, est réelle positive ou nulle - en tout cas lorsqu'elle est calculable, donc dès lors que $n_d \geq 1$. On peut considérer N_d^c comme la résultante d'un phénomène aléatoire. S'y ajoute l'aléa dû à l'enquête Emploi, ce qui donne à $\hat{N}_d^c$ un statut de variable aléatoire complexe. On peut alors modéliser la loi de $\hat{N}_d^c$ par une loi de Poisson. Cette loi a un unique paramètre, que l'on explique par une combinaison linéaire de variables déterministes connues au niveau domaine (ZE) et éventuellement on y ajoute un effet aléatoire propre au domaine. Cela étant, ce n'est pas directement le paramètre de Poisson qui est expliqué, mais une fonction $g(\,)$ au choix de ce paramètre. Ainsi, on pose

$$\hat{N}_d^c \rightarrow P(\mu_d) \quad \text{avec}$$
$$g(\mu_d) = X_d^t \cdot \beta + v_d$$

On fait une hypothèse de loi de Gauss pour les v_d et d'indépendance mutuelle. On introduit un paramètre de variance supplémentaire puisque

$$Var(v_d) = \sigma_v^2$$

Le modèle complexe finalement obtenu est un modèle linéaire mixte généralisé (GLMM). Si on ne souhaite pas faire apparaître d'effet local aléatoire, alors on impose $\sigma_v^2 = 0$ et on a alors à faire à un modèle linéaire généralisé (GLM).

L'estimation des paramètres utilise une technique de maximum de vraisemblance restreinte qui porte sur un modèle approché. Ce modèle entraine la construction d'une pseudo variable P_d (calculable) qui intervient dans un modèle approché linéaire mixte , à savoir

$$P_d = X_d^t \cdot \beta + v_d + \varepsilon_d$$

où ε_d est une variable aléatoire d'espérance nulle. Lorsque $\hat{N}_d^c$ suit effectivement une loi de Poisson conditionnellement à v_d, on a égalité entre la variance et l'espérance, c'est-à-dire que $Var\left(\hat{N}_d^c | v_d\right) = \mu_d$. En réalité, on a souvent à faire à une réalité qui diffère de ce modèle théorique et dans laquelle la variance est supérieure (parfois inférieure) à μ_d. Dans un tel cas, on dit qu'il y a "overdispersion" et la formalise en introduisant un nouveau paramètre noté ϕ tel que

$$Var\left(\hat{N}_d^c | v_d\right) = \phi \cdot \mu_d$$

Ce paramètre intervient alors dans le pseudo modèle car on vérifie que $Var\varepsilon = \phi \cdot M(\beta)$ où ε est le vecteur des ε_d et $M(\beta)$ est une matrice complexe, dont la forme est connue mais qui dépend de β. Le paramètre ϕ, lorsqu'on choisit de l'introduire, est *in fine* estimé par maximum de vraisemblance, au même tire que le vecteur β et le paramètre réel σ_v^2.

Finalement, après l'ajustement du pseudo modèle linéaire, on obtient un estimateur EBLUP $\hat{\beta}$ et un prédicteur EBLUP $\hat{v}_d$ pour chaque ZE d. Ces estimateurs / prédicteurs prennent en compte l'estimateur $\hat{\sigma}_v^2$ et éventuellement le paramètre d'overdispersion $\hat{\phi}$. On conclut en formant l'estimateur suivant, qui est l'estimateur "petit domaine" :

$$\hat{\hat{N}}_d^c = \hat{\mu}_d = g^{-1}\left(X_d^t \cdot \hat{\beta} + \hat{v}_d\right)$$

Si le modèle est un GLM, donc si on a choisi d'imposer $\sigma_v^2 = 0$, alors $\hat{v}_d = 0$. Cette méthodologie s'apparente à celle du modèle linéaire de Fay et Herriot (voir partie 4) en ce sens où les variables observées utilisées pour l'ajustement sont des estimateurs directs pondérés, donc tenant compte du plan de sondage. En revanche, il y a une différence majeure au sens où on ne distingue plus les deux étapes qui structuraient la modélisation linéaire, à savoir d'une part une "vraie valeur" modélisée comme variable aléatoire et d'autre part l'introduction explicite d'une erreur d'échantillonnage : cette fois, la modélisation mêle de manière indiscernable les deux natures d'aléa.

5.2. Les principaux résultats

L'unité statistique est le domaine. Les variables explicatives sont donc définies au niveau ZE. La modélisation initiale n'étant plus linéaire, la présélection présentée en partie 2 perd beaucoup en pertinence. Néanmoins, il n'est pas raisonnable d'effectuer un ajustement dans lequel on entasse toutes les variables disponibles; par ailleurs, on ne dispose pas d'un outil informatique capable d'optimiser le choix d'un tel modèle. Partant d'un ensemble de p variables, où p est grand dans notre contexte, on ne peut pas raisonnablement tester les 2^p modèles possibles - d'autant plus que les critères de sélection sont plutôt destinés à comparer des modèles emboités, et pas vraiment deux modèles dont l'un n'est pas inclus dans l'autre. C'est pourquoi il nous est paru satisfaisant de repartir de la liste - déjà plutôt large - des 15 variables qui ont été mentionnées dans la partie 2, sans plus.

La fonction de lien choisie est le logarithme : $g(\mu_d) = Log(\mu_d)$. Donc pour toute ZE d ayant participé à l'ajustement, on a $\hat{\hat{N}}_d^c = \exp\left(X_d^t \cdot \hat{\beta} + \hat{v}_d\right) = \exp\left(X_d^t \cdot \hat{\beta}\right) \cdot e^{\hat{v}_d}$. La modélisation en logarithme nous incite à écrire la proportion de chômeurs comme un produit de proportions élevées à certaines puissances - en faisant "comme si" on décomposait la probabilité d'être chômeur en un produit de probabilités d'avoir telle ou telle caractéristique (aux puissances près). Ainsi, si $poptot2007$ désigne la population totale dans le champ couvert et X_d^k la k^{ième} variable explicative (concrètement, ces variables sont des taux, donc des probabilités…), on pose

$$\frac{\hat{N}_d^c}{poptot2007} = \prod_{k=1}^{p}\left(X_d^k\right)^{\beta_k}$$

soit encore

$$Log(\hat{N}_d^c) = Log(poptot2007) + \sum_{k=1}^{p}\beta_k \cdot Log\left(X_d^k\right)$$

La variable explicative $Log(poptot2007)$ participe à l'ajustement mais son coefficient est maintenu égal à 1 (variable dite "offset"). A ce stade, trois variables doivent être modifiées :

- la part de population en catégorie Tabard "SEMAG02" a été purement et simplement supprimée, car trop souvent nulle ;
- Le taux de personnes DEFM catégories 1,2,3 et HAR et appartenant à la catégorie des hommes de 15 à 19 ans, diplômés - très faible - n'a pas été pris en logarithme ;
- Le taux de solde des établissements entre 2000 et 2006 = (arrivées - départs) divisé par stock d'établissements au 1/1/2006 - parfois négatif - n'a pas été pris en logarithme ;

On peut alors ajuster un modèle impliquant 14 régresseurs (plus la constante), et éventuellement réduire cette liste en fonction des critères d'ajustement disponibles. Une première option consiste à introduire ou pas d'effet aléatoire propre au domaine, c'est-à-dire imposer $\sigma_v^2 = 0$ (GLM) ou estimer σ_v^2 (GLMM). Une seconde option consiste à introduire ou pas un paramètre ϕ d'overdispersion. Elle peut d'ailleurs être exercée indépendamment de la première. Dans tous les cas, il est nécessaire d'exclure les ZE pour lesquelles il n'y a pas de données, soit $n_d = 0$. Cela étant, une estimation est possible dès lors que $n_d \geq 1$. Il est également possible de retenir les ZE pour lesquelles $\hat{N}_d^c = 0$, même si intuitivement on peut avoir quelque interrogation sur la pertinence de la prise en compte de telles ZE pour participer à l'ajustement.

Que l'on soit dans le cas GLM ou dans le cas GLMM, quelle que soit la règle d'exclusion, les ZE n'ayant pas participé à l'ajustement donnent lieu à une estimation "petit domaine" finale purement synthétique, définie par $\hat{\hat{N}}_d^c = \exp\left(X_d^t \cdot \hat{\beta}\right)$.

L'examen de toutes les configurations possibles en croisant toutes les options possibles demanderait trop de temps. Il a donc fallu choisir une stratégie. Dans un premier temps, l'ajustement sur le modèle complet (14 variables plus la constante) a été tenté avec différentes sélections préliminaires des ZE :

- sélection des ZE ayant $n_d > 0$
- sélection des ZE ayant $n_d > 49$
- sélection des ZE ayant $n_d > 49$ et $\hat{N}_d^c > 0$
- sélection des ZE ayant $n_d > 49$ et $0 < \hat{N}_d^c < 25000$

On s'est intéressé à l'estimation nationale obtenue par sommation des estimations par ZE, au paramètre d'overdispersion estimé, et à la régression linéaire simple des estimations "petits domaines" sur les estimations directes (R^2, coefficient de la pente de régression $\hat{b}$, conclusion du test d'égalité entre la droite de régression et le droite $y = x$). On obtient le tableau suivant, en rappelant que l'estimation directe nationale EE du nombre de chômeurs BIT vaut 2 416 000, ou encore 2 436 000 si on utilise un estimateur direct calé sur la taille de la ZE.

Sélection pour l'ajustement du modèle	**Somme des $\hat{\hat{N}}_d^c$**	$\hat{\phi}$	R^2	$\hat{b}$	**Conclusion test**
$n_d > 0$	2 431 000	1751	0.91	1.01	égalité
$n_d > 49$	2 463 000	1711	0.91	1.01	égalité
$n_d > 49$ et $\hat{N}_d^c > 0$	2 513 000	1598	0.91	1.03	égalité
$n_d > 49$ et $0 < \hat{N}_d^c < 25000$	2 322 000	1463	0.90	1.23	Écart significatif

Il apparaît que plus on sélectionne, plus la distribution des estimations "petits domaines" s'éloigne de celle des estimations directes - donc plus on introduit *in fine* de biais - mais qu'en revanche le paramètre d'overdispersion diminue (un peu). C'est une tendance assez naturelle, le biais et la variance évoluent en opposition de phase. Le cas d'une sélection originelle conditionnée seulement par $n_d > 0$ nous apparaît finalement assez satisfaisante : contrairement au cas de Fay et Herriot, nous n'éprouvons pas le besoin de nous limiter à une sous-population pour procédre à l'ajustement du modèle (sauf à éliminer les 10 ZE où $n_d = 0$, mais cela s'impose).

Sur le plan pratique, l'ajustement s'effectue avec la Proc GLIMMIX de SAS, qui traite les modèles linéaires mixtes généralisés.

5.2.1. Modélisation classique

On rappelle que la modélisation classique conduit à un estimateur de type synthétique. Si on part du modèle ajusté avec les 14 variables explicatives initialement retenues (modèle dit "complet"), on constate que deux variables[18] ne sont pas significatives. On les supprime donc. On obtient une situation où toutes les variables sont significatives - à l'exception de la part des bas revenus mais on se situe à la limite de significativité et on choisit donc de conserver cette variable.

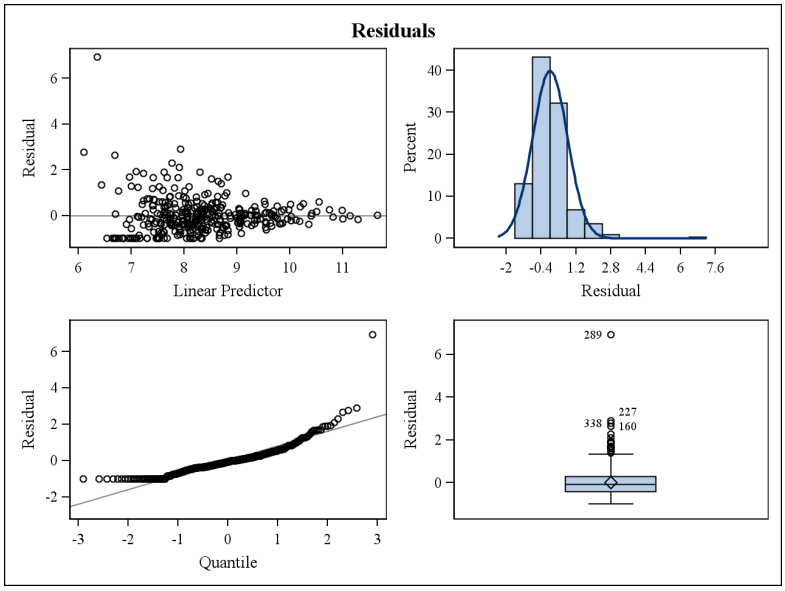
Critères de qualité - modèle classique	
-2 Log Likelihood	580231.5
AIC (smaller is better)	580257.5
BIC (smaller is better)	580307.2
Pearson Chi-Square	565773.5

Estimations - modèle classique					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	-2.8401	0.07061	325	-40.22	<.0001
log_t_rech_oui	1.2594	0.01202	325	104.79	<.0001
log_t_couple_2	-0.1865	0.01658	325	-11.25	<.0001
log_t_age30_49Hnondi	-0.2398	0.009844	325	-24.37	<.0001
log_t_age50_64Hnondi	0.1705	0.006534	325	26.10	<.0001
log_b05_partbasrev	0.01324	0.007453	325	1.78	0.0765
log_a06_partagri06	0.01781	0.000877	325	20.31	<.0001
log_c08_partslbtp06	-0.1746	0.005822	325	-29.99	<.0001
log_c08_partslsante0	0.01255	0.005127	325	2.45	0.0149
log_c08_partslfabri0	-0.2042	0.002581	325	-79.12	<.0001
log_c08_partslgest06	-0.3828	0.005696	325	-67.20	<.0001
c02_txsoldetab_0006	-0.02422	0.000396	325	-61.16	<.0001
t_age15_19HdiplBIT	-3.3256	0.2740	325	-12.14	<.0001

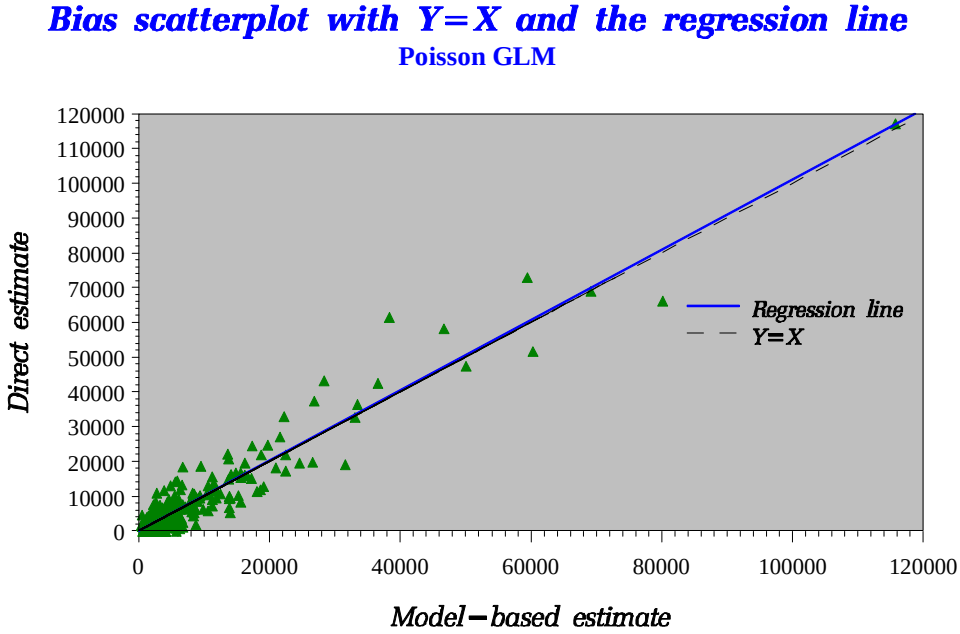
Il est remarquable de constater qu'il y a de nombreuses variables expliquant $\hat{N}_d^c$ de manière significative, en particulier cette liste est beaucoup plus riche que celle que l'on obtenait avec le modèle linéaire - avec lequel on rappelle qu'il avait été assez laborieux de détecter une information explicative convaincante. De ce point de vue, les deux modèles - linéaire versus non linéaire - réagissent donc de manière très différente.

Le coefficient estimé proche de 1 (puisque égal à 1,26) de la variable *log-t-rech-oui* - alors que les coefficients estimés des autres variables en logarithme sont proches de zéro - est plutôt rassurant, parce que grosso modo la proportion de chômeurs est proche de la proportion d'individus recherchant du travail (la constante paraît fortement négative mais c'est parce que les taux explicatifs présents dans la base sont exprimés en pourcentage, donc multiplié par 100 - pour apprécier la constante il faut ajouter quelque chose de l'ordre de log100, et là on retrouve effectivement un ordre de grandeur raisonnable). On voit en particulier que les résidus ont bonne allure et qu'il n'y a pas de phénomène de "shrinkage", si bien que les estimateurs peuvent être considérés comme sans biais (ou à peu près).

[18] Allocataires RMI + Effectif d'étrangers hors Europe



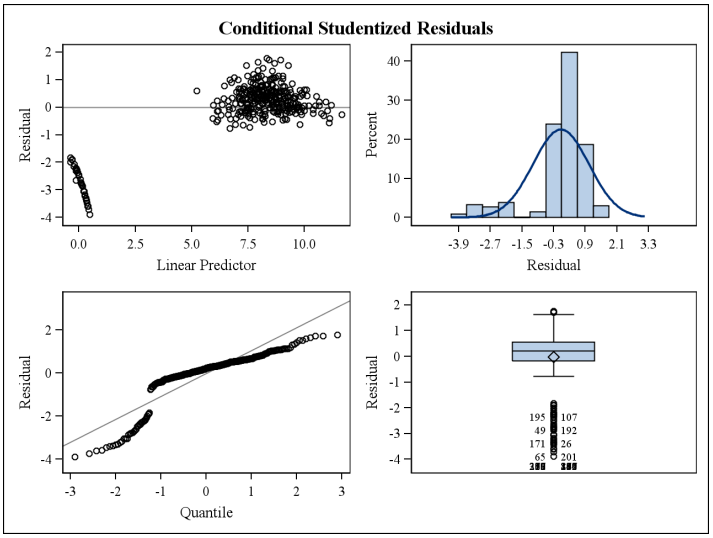
A noter que le programme utilisant la Proc Glimmix a été passé une fois avec le modèle classique de Poisson sans overdispersion et une fois avec ce même modèle mais en déclarant une overdispersion : on peut vérifier que les deux estimateurs de Poisson sont rigoureusement identiques ZE par ZE. Cela est rassurant parce que dans ces conditions la matrice de variance-covariance du pseudo modèle est $\phi \cdot \Delta$ où Δ est une matrice diagonale, si bien que $\hat{\phi}$ n'intervient pas dans l'expression $\hat{\beta}$ et n'a donc pas d'impact sur les estimations numériques. Cette propriété disparaîtra lorsqu'on considèrera un modèle mixte (partie 5.2.2).



On notera, à titre de curiosité, que le benchmarking, de manière inattendue (?), dégrade la situation en introduisant un biais significatif (la droite de régression est significativement distincte de la première bissectrice).

5.2.2. Modélisation mixte

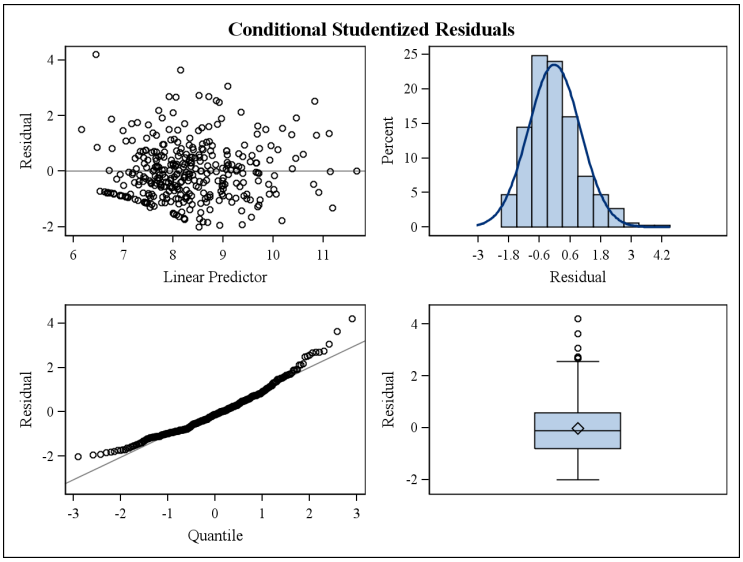
La modélisation mixte introduit des effets locaux : elle est donc, en principe, plus sensible aux spécificités locales que la modélisation classique, qui ne reproduit que les effets (locaux) venant de variables prédéterminées. Lorsqu'on ajuste le modèle mixte complet (14 variables) sans overdispersion, on trouve $\hat{\sigma}_v^2 = 4{,}96$ (écart-type de 0,44). Cette valeur conduit a de nombreux effets locaux v_d numériquement très forts, compris par exemple entre 3 et 4, donc à des effets locaux dans la formation de $\hat{N}_d^c$ de l'ordre de e^3 ou e^4, ce qui est bien difficile à croire ... Par ailleurs on obtient un résultat d'ajustement fort curieux, avec *in fine* trois variables explicatives significatives et des résidus peu sympathiques de ce type :



Enfin, l'estimation nationale issue de la sommation des estimations locales serait de 4 124 000 chômeurs BIT : on obtient donc manifestement des résultats aberrants, ce schéma n'est clairement pas acceptable.

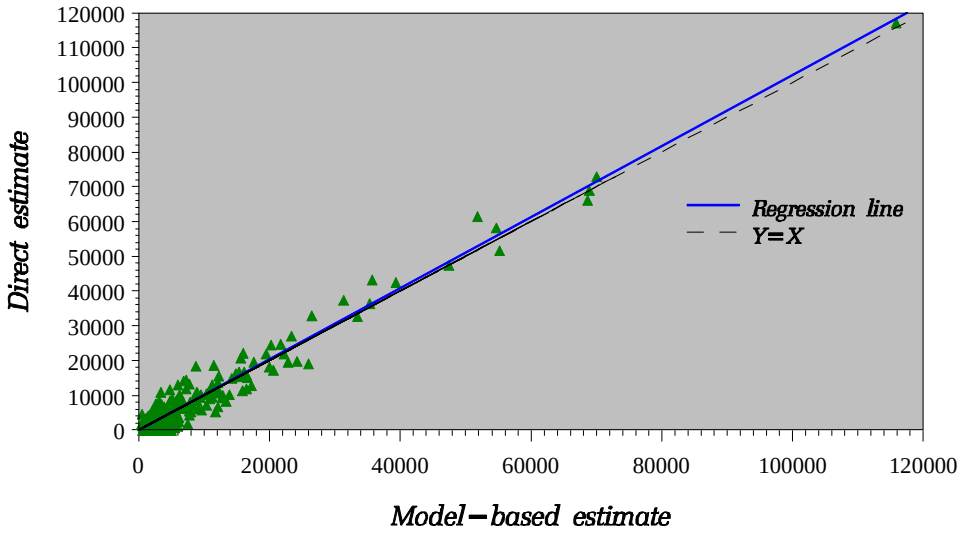
En revanche, toujours dans le modèle complet, si on introduit un paramètre d'overdispersion, on obtient cette fois $\hat{\sigma}_v^2 = 0{,}039$ avec un écart-type de 0,024 tandis que $\hat{\phi} = 1434$ avec un écart-type de $169{,}9$. De fait, $\hat{\sigma}_v^2$ n'est pas significativement différent de zéro. Par ailleurs, l'estimation $\hat{\phi}$ apparaît numériquement très grande par rapport à celle de $\hat{\sigma}_v^2$ et cela peut surprendre (noter qu'on retrouve les ordres de grandeur des $\hat{\phi}$ obtenus avec le modèle classique en présence d'overdispersion). Manifestement, le modèle attribue la quasi-totalité de la variance des $\hat{N}_d^c$ à l'effet d'overdispersion et non à l'effet local, puisque $Var\left(\hat{N}_d^c | v_d\right) = \left(\phi \cdot e^{v_d}\right) \cdot \exp(X_d^t \cdot \beta)$.

Par contre, avec ce paramètre d'overdispersion, l'estimation nationale mixte est globalement satisfaisante puisqu'on aboutit à 2 449 000 chômeurs BIT et, surtout, les résidus ont cette fois des allures beaucoup plus conformes. De plus, le tracé du nuage de poids rapportant l'estimation GLMM avec overdispersion à l'estimation directe conduit à conclure que la droite de régression n'est pas significativement différente de la droite $y = x$, autrement dit l'estimateur "petits domaines" apparaît non biaisé.



Bias scatterplot with Y=X and the regression line

Poisson GLMM



Cela étant, de nombreuse variables sont (très nettement) non-significatives : parmi les 14 variables intiales ayant participé à l'ajustement, plus la constante, seules trois variables se distinguent :

Estimations - Modèle mixte avec overdispersion					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	-2.2569	3.6216	207.4	-0.62	0.5339
log_t_rech_oui	1.2478	0.5823	249.9	2.14	0.0331
log_t_couple_2	-0.2376	0.8199	166	-0.29	0.7723
log_t_age30_49Hnondi	-0.2852	0.4419	276	-0.65	0.5192
log_t_age50_64Hnondi	0.2735	0.2921	187.9	0.94	0.3504
log_b05_partbasrev	0.03203	0.3976	254.7	0.08	0.9359
log_t_allocRMI	-0.01473	0.2435	292.3	-0.06	0.9518

Estimations - Modèle mixte avec overdispersion					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
log_t_natc_afri	0.002322	0.05515	299.9	0.04	0.9664
log_a06_partagri06	0.000826	0.04370	122.8	0.02	0.9850
log_c08_partslbtp06	-0.1835	0.2676	221	-0.69	0.4937
log_c08_partslsante0	-0.00526	0.2263	301	-0.02	0.9815
log_c08_partslfabri0	-0.2172	0.1128	283.2	-1.93	0.0552
log_c08_partslgest06	-0.4635	0.2558	258.7	-1.81	0.0711
c02_txsoldetab_0006	-0.02418	0.01693	295.1	-1.43	0.1544
t_age15_19HdiplBIT	-3.8832	12.0684	209	-0.32	0.7480

Si on se restreint maintenant aux trois variables explicatives qui ressortent comme significatives (et encore, de manière limite pour deux d'entre elles…) dans le modèle complet, on obtient un premier modèle réduit :

Estimation - Modèle mixte réduit N°1 avec overdispe rsion					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	-4.4401	0.5224	176.5	-8.50	<.0001
log_t_rech_oui	1.1907	0.1506	144.9	7.91	<.0001
log_c08_partslfabri0	-0.1219	0.08028	240.9	-1.52	0.1302
log_c08_partslgest06	-0.2258	0.1398	85.6	-1.62	0.1098

Si on supprime de nouveau les variables non significatives à l'étape précédente - donc en ne conservant que *log_t_rech_oui* - on aboutit à un second modèle réduit :

Estimation - Modèle mixte réduit N°2 avec overdispe rsion					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	-5.2258	0.2740	117.5	-19.07	<.0001
log_t_rech_oui	1.1643	0.1433	95.45	8.12	<.0001

Avec le premier modèle réduit, $\hat{\sigma}_v^2 = 0{,}021$ avec un écart-type de 0,02 - donc $\hat{\sigma}_v^2$ est toujours non significative - tandis que $\hat{\phi} = 1548$ avec un écart-type de 166. L'estimation nationale du nombre de chômeurs BIT est d'environ 2 447 000. Le test habituel conclut à une absence de biais des estimateurs "petits domaines".

Avec le second modèle réduit, $\hat{\sigma}_v^2 = 0{,}009$ avec un écart-type de 0,017 - $\hat{\sigma}_v^2$ est de plus en plus non significative - tandis que $\hat{\phi} = 1692$ avec un écart-type de 170.L'estimation nationale du nombre de chômeurs BIT vaut dans ce cas environ 2 432 000. Le test habituel conclut de nouveau à une absence de biais des estimateurs "petits domaines".

Ainsi, lorsqu'on réduit la liste des variables explicatives, on renforce le caractère non significatif de $\hat{\sigma}_v^2$ et on a tendance à augmenter en contrepartie $\hat{\phi}$ - mais dans une proportion limitée qui permet de dire que $\hat{\phi}$ est une estimation relativement stable.

Si on s'intéresse aux critères de qualité habituels produits à l'occasion des ajustements, on trouve :

Critères de qualité	Modèle complet	Modèle réduit 1	Modèle réduit 2
-2 Res Log Pseudo-Likelihood	666.52	669.52	684.38
Generalized Chi-Square	463 178.9	516 992.4	568 416.1

Ainsi, si on s'en tient aux critères d'ajustement de modèle - en la circonstance la vraisemblance optimisée et le chi2 de Pearson - c'est clairement le modèle complet qui l'emporte. Mais il faut accepter la présence de très nombreuses variables manifestement non significatives ... Autrement dit, si on veut ne conserver que des variables significatives, alors la qualité des critères d'ajustement se dégrade. Ces résultats sont donc quelque peu déroutants !!!

La mémoire des phénomènes constatés dans l'ajustement du modèle de Fay et Herriot (partie 4) peut laisser penser qu'on aurait peut-être dû, en amont, exclure de l'ajustement les ZE potentiellement "perturbatrices", c'est-à-dire au moins les plus petites, peut-être aussi les plus grosses (qui peuvent avoir des comportements trop influents). Dans cet esprit, si on reprend les 14 variables explicatives initiales (modèle complet), mais cette fois en restreignant la population des ZE participant à l'ajustement du modèle afin que $n_d > 49$ et $0 < \hat{N}_d^c < 25000$, alors on trouve $\hat{\sigma}_v^2 = 0{,}13$ avec un écart-type de 0,048 : $\hat{\sigma}_v^2$ devient donc (de peu…) significative et on assiste en parallèle à un effondrement de la valeur $\hat{\phi}$ puisque $\hat{\phi} = 650$, avec un écart-type de 228. Il se produit donc bien "quelque chose", qui ressemble à un système de vases communicants entre les composantes de la variance. En contrepartie, de manière surprenante, on constate que la variable *log_t_rech_oui* perd très nettement sa significativité (p-value de 0.20), et le test habituel de la droite de régression conclut à un biais des estimateurs "petits domaines". Cette nouvelle expérience laisse un paysage assez déconcertant : que convient-il de faire, d'une part pour définir la population sur laquelle on ajuste le modèle et d'autre part pour fixer les variables explicatives à retenir, alors même que les résultats d'estimation finaux "petits domaines" par ZE apparaissent fort vraisemblables dans la plupart des alternatives ?

A l'examen de ces résultats consacrés à l'estimation mixte, on peut avoir le sentiment que, finalement, la prise en compte d'effets locaux n'est pas convaincante, ce qui revient à renvoyer dos à dos le modèle mixte et le modèle classique (partie 5.2.1). Pour éclairer un peu plus cette question, les tableaux qui suivent permettent d'apprécier les écarts relatifs en valeur absolue par ZEAT et avant toute procédure de benchmarking, c'est-à-dire à un niveau géographique où l'enquête Emploi peut être considérée comme fiable, et l'estimateur direct peut donc être pris comme référence. On rappelle que l'estimateur Insee est par construction calé sur l'estimateur direct national, la comparaison est donc quelque peu "biaisée" en sa faveur (il faudrait donc faire également la comparaison après benchmarking). Avec le modèle complet :

Modèle complet

ZEAT	Poisson classique	Poisson mixte	Direct	Insee	Ecart Poisson classique	Ecart Poisson mixte	Ecart Insee
1	514 891	500 040	514 275	459 889	0,1 %	2,8 %	10,6 %
2	406 869	408 705	394 334	403 680	3,2 %	3,6 %	2,4 %
3	202 598	214 479	218 159	200 330	7,1 %	1,7 %	8,2 %
4	201 926	199 104	181 870	201 225	11,0 %	9,5 %	10,6 %
5	259 729	261 852	257 199	277 946	1,0 %	1,8 %	8,1 %
7	242 207	251 550	247 656	248 665	2,2 %	1,6 %	0,4 %
8	234 547	246 667	241 831	250 347	3,0 %	2,0 %	3,5 %
9	367 980	366 851	361 093	365 490	1,9 %	1,6 %	1,2 %
Total	**2 431 000**	**2 449 000**	**2 416 000**	**2 408 000**	**Sigma = 29,5**	**Sigma =24,6**	**Sigma =45,0**

L'indicateur "sigma" donne la somme des écarts en valeur absolue. Il n'a pas d'interprétation particulière au niveau global mais il permet d'apprécier l'importance des erreurs relatives indépendamment des populations en jeu : c'est un indicateur qui pourrait être utilisé pour classer les méthodes. Le modèle de Poisson, bien qu'il soit défectueux en ZEAT 4, semble dominer nettement la méthode utilisée actuellement à l'insee.

Le même exercice avec le modèle réduit à la variable explicative *log_t_rech_oui* donne :

Modèle réduit N°2

ZEAT	Poisson classique	Poisson mixte	Direct	Insee	Ecart Poisson classique	Ecart Poisson mixte	Ecart Insee
1	506 513	504 222	514 275	459 889	1,5 %	2,0 %	10,6 %
2	403 805	403 871	394 334	403 680	2,4 %	2,4 %	2,4 %
3	205 225	208 493	218 159	200 330	5,9 %	4,4 %	8,2 %
4	204 099	202 942	181 870	201 225	12,2 %	11,6 %	10,6 %
5	264 961	264 305	257 199	277 946	3,0 %	2,8 %	8,1 %
7	240 348	242 272	247 656	248 665	3,0 %	2,2 %	0,4 %
8	247 991	250 048	241 831	250 347	2,5 %	3,4 %	3,5 %
9	356 769	356 342	361 093	365 490	1,2 %	1,3 %	1,2 %
Total	**2 430 000**	**2 432 000**	**2 416 000**	**2 408 000**	**Sigma = 31,7**	**Sigma = 30,1**	**Sigma =45,0**

6. Le modèle logistique

6.1. Éléments de théorie

Il s'agit, contrairement aux méthodes présentées en parties 4 et 5, d'une approche individuelle. L'unité statistique n'est plus la ZE mais l'individu physique, identifié par sa ZE (notée d) et un identifiant interne à la ZE (noté i). On va construire la variable aléatoire $Y_{d,i}$ égale à 1 si l'individu d,i est chômeur BIT et égale à 0 sinon. Cette variable suit donc une loi binomiale (dite de Bernoulli)

$$Y_{d,i} \rightarrow B(1, P_{d,i})$$

En toute généralité, on considère que $P_{d,i}$, qui s'interprète comme la probabilité de l'individu d'être au chômage, est la réalisation d'une variable aléatoire. On l'explique essentiellement par des variables auxiliaires $X_{d,i}$ formant un vecteur de taille p (effets fixes), selon

$$g(P_{d,i}) = \beta^t \cdot X_{d,i} + \nu_d$$

où $g()$ est une fonction connue, β est un paramètre vectoriel de dimension p inconnu, et ν_d une variable aléatoire, que l'on peut éventuellement faire disparaitre, qui dépend de d et qui représente donc l'effet propre à la ZE. Lorsqu'elle est introduite, ν_d est une variable aléatoire que l'on suppose gaussienne, d'espérance nulle et de variance σ_ν^2. Dans notre contexte de modèle dit logistique, on opte pour la fonction de lien $g(x) = Log\dfrac{x}{1-x}$.

De manière similaire au traitement du modèle de Poisson, on construit un pseudo modèle linéaire mixte approché[19] - cette fois au niveau individuel - et on estime par maximum de vraisemblance - pondéré ou non - les paramètres β et σ_v^2 (s'il y a lieu). On peut ainsi estimer (cas du modèle standard) ou prédire (cas du modèle mixte) $P_{d,i}$ par $\hat{P}_{d,i} = g^{-1}\left(\hat{\beta}^t \cdot X_{d,i} + \hat{v}_d\right)$. Comme on veut prédire $N_d^c = \sum_{i=1}^{N_d} Y_{d,i}$, on va former pour chaque ZE d de taille N_d :

$$\hat{N}_d^c = \sum_{i \in s_d} w_{d,i} \cdot \hat{Y}_{d,i} = \sum_{i \in s_d} w_{d,i} \cdot \hat{P}_{d,i}$$

où s_d désigne l'ensemble des unités présentes dans le fichier du recensement de la ZE d (en cumul sur 5 années) et $w_{d,i}$ est le poids de sondage de l'unité (d,i) , tel qu'il figure dans ce même fichier.

Compte tenu de la fonction de lien retenue, dans le cas classique où il n'y a pas d'effet local (donc lorsqu'on impose $\sigma_v^2 = 0$), si emp désigne l'ensemble des identifiants des individus de l'enquête Emploi ayant participé à l'ajustement du modèle[20], alors le maximum de vraisemblance non pondéré vérifie le système d'équations

$$\sum_{d,i \in emp} X_{d,i} \cdot \hat{P}_{d,i} = \sum_{d,i \in emp} X_{d,i} \cdot Y_{d,i}$$

Lorsqu'on l'applique à la seule composante constante du vecteur $X_{d,i}$, on obtient

$$\sum_{d,i \in emp} \hat{P}_{d,i} = \sum_{d,i \in emp} Y_{d,i}$$

Ainsi, en modélisation standard (non mixte), la somme des probabilités prédites sur l'ensemble des individus de l'échantillon Emploi est mécaniquement égale au nombre de chômeurs dans l'échantillon Emploi[21]. Si on opte pour un maximum de vraisemblance pondéré, il suffit de rajouter les poids de l'échantillon Emploi dans ces équations (le terme de droite devient l'estimation nationale du nombre total de chômeurs BIT, mais la somme pondérée des probabilités prédites - donc le terme de gauche - n'a pas grande signification, et ne doit évidemment pas être assimilée à la somme des probabilités prédites définie à partir du recensement - même si elle devrait en être numériquement très proche).

6.2. Les principaux résultats

Pour les 10 ZE qui n'ont pas de données Emploi, on ne peut (évidemment) pas estimer d'effet aléatoire v_d . On convient d'imposer, pour ces ZE, $\hat{v}_d = 0$ (qui est la seule valeur raisonnablement imputable). Les variables potentiellement explicatives du phénomène de chômage doivent être présentes à la fois dans le questionnaire Emploi et dans le questionnaire du recensement. Nous renvoyons sur cette question importante à la partie 2.2 . Contrairement aux cas des modèles définis au niveau ZE, le faible nombre de variables individuelles nous dispense de toute phase préalable de sélection de variables - qui sont initialement au nombre de 8 (voir 2.2). Compte tenu de quelques valeurs manquantes dans les variables explicatives, l'ajustement des différents modèles logistiques s'effectue ici sur 72 141 individus (ce qui est considérable).

[19] La démarche du maximum de vraisemblance appliquée au modèle initial est inextricable : il faudrait intégrer la densité conditionnelle de la loi binomiale par rapport à la loi Gaussienne de $\mathcal{V}$. Pour éviter cet exercice, soit on simplifie l'intégrale, soit on passe par un modèle approché.

[20] Dans notre contexte, c'est l'échantillon des 73 153 individus répondants moins un tout petit ensemble d'individus pour lesquels au moins une des variables explicatives est manquante - soit in fine 73 096 individus.

[21] Soit 3630.

En matière informatique, il y a une difficulté pratique toute spécifique aux modèles individuels parce qu'on doit calculer $\hat{N}_d^c = \sum_{i \in s_d} w_{d,i} \cdot \hat{P}_{d,i}$. Cela amène à manipuler des tables énormes, comprenant environ 35 millions d'observations - donc à employer des ruses informatiques pour éviter des temps de traitement prohibitifs.

6.2.1. Modélisation standard non pondérée

Une première option consiste à ajuster un modèle sans effet aléatoire et sans faire intervenir les poids des individus tirés par l'enquête Emploi dans les procédures d'ajustement de ce modèle (maximum de vraisemblance non pondéré). Un tel modèle produit une estimation de 3 062 000 chômeurs BIT au niveau national - donc très au-dessus de ce qui ressort des autres méthodes. Cette estimation n'est donc pas acceptable.

De notre point de vue, l'explication la plus vraisemblable est celle d'une hétérogénéité dommageable entre certaines variables explicatives, selon que l'on considère l'une ou l'autre des deux sources en jeu. En effet, la variable essentielle, la plus explicative dans ce modèle, est clairement la déclaration spontanée de recherche d'emploi. Cette information provient, d'une part de l'enquête Emploi (pour ce qui concerne l'ajustement du modèle), d'autre part du recensement (pour ce qui concerne le calcul des prédictions des probabilités individuelles, que l'on somme pour obtenir l'estimation finale). Le bon sens voudrait qu'une personne interrogée qui répond "oui" à la question "Recherchez-vous un emploi ?" posée au cours de l'entretien Emploi (question A.10) réponde également "oui" à la question "Cherchez-vous un emploi ?" du bulletin individuel du recensement (question 16). Il semble bien s'agir du même concept dans les deux cas, mais les conditions de collecte sont évidemment très différentes et on peut craindre que pour décrire une situation ressentie par un individu, même s'il est cohérent, cette variable ne prenne malheureusement des modalités différentes selon que l'individu répond au questionnaire Emploi ou qu'il remplit (lui-même) son bulletin individuel de recensement. Un tel phénomène n'est pas surprenant, on sait combien le mode de collecte et le contexte dans lequel se déroule la collecte ont de l'influence sur les réponses des enquêtés (plus ou moins cohérents…), même pour des questions qui apparaissent assez simples.
S'ajoutent des perturbations dues aux filtrages, qui compliquent la visibilité : au sens du recensement, il n'y a pas de recherche d'emploi dès lors que l'individu a un emploi occasionnel ou de très courte durée. A l'enquête Emploi, on peut déclarer en A10 rechercher un emploi alors même qu'on a travaillé durant la semaine de référence - mais dès lors qu'on n'a pas un emploi considéré comme "régulier". Cela étant, l'effet de ces filtrages vient plutôt modérer les décalages numériques qui nous affectent : selon l'enquête Emploi, une personne qui recherche un emploi mais qui a un travail précaire durant la semaine de référence ne sera pas chômeur BIT, donc elle appartient à une sous-population qui contribue à diminuer l'impact d'une recherche d'emploi sur la probabilité d'être chômeur BIT. On a donc un effet de limitation du nombre de chômeurs BIT quand on extrapole à la population entière : autrement dit, si on appliquait le filtrage du recensement à l'enquête Emploi, on trouverait bien plus de 3 062 000 chômeurs prédits. Par ailleurs, si on n'appliquait pas le filtrage du recensement, on trouverait au recensement des personnes qui ont un travail précaire mais qui déclarent spontanément chercher un emploi. Leur probabilité prédite d'être chômeur BIT se trouverait considérablement augmentée et là encore on aboutirait à bien plus de 3 062 000 chômeurs prédits.

Cette piste trouve une traduction numérique qui la rend plus convaincante. En effet, à partir de l'échantillon Emploi exploité, on estime à 2 512 000 le nombre de personnes recherchant un emploi. Avec les données du recensement 2007, on estime cette fois à 3 296 000 le nombre d'individus déclarant spontanément rechercher un emploi (+ 31,2 % en changeant de source) - malgré le filtrage évoqué supra. Cette piste est renforcée par un fait parallèle : il y a sensiblement plus de déclarations spontanées de chômage au recensement qu'il y a de chômeurs BIT au sens de l'enquête Emploi : c'est la raison pour laquelle l'utilisation éventuelle dans ce modèle logistique de la variable de déclaration spontanée d'un état de chômage conduirait au même décalage excessif[22]. Si on divise l'effectif estimé en utilisant le modèle par 1,312 (pour corriger de l'effet d'erreur de mesure, en quelque sorte), on trouve une estimation nationale "petits domaines" de 2 334 000 chômeurs, qui est cette fois beaucoup plus proche des 2 416 000 chômeurs que donne l'enquête Emploi (et encore plus des 2 436 000 chômeurs obtenus en sommant les estimations directes calées sur la taille de la ZE).

[22] Nous n'avons donc pas cherché à nous engager dans cette voie.

Il faut noter que ce problème d'hétérogénéité n'avait pas de conséquence fâcheuse dans tous les modèles constitués au niveau ZE : certes l'information sur la recherche d'emploi était déjà une variable explicative fondamentale, mais l'important était de s'assurer de la corrélation entre la proportion de chômeurs BIT (au sens de l'enquête Emploi) et la proportion de chercheurs d'emploi (au sens du recensement), ni plus ni moins - en tout cas on n'avait pas besoin d'une similitude des concepts. De ce point de vue, il apparait que la modélisation individuelle est plus délicate, plus exigeante, que la modélisation au niveau agrégé.

L'exploitation des données de l'échantillon Emploi fournit la table de contingence suivante :

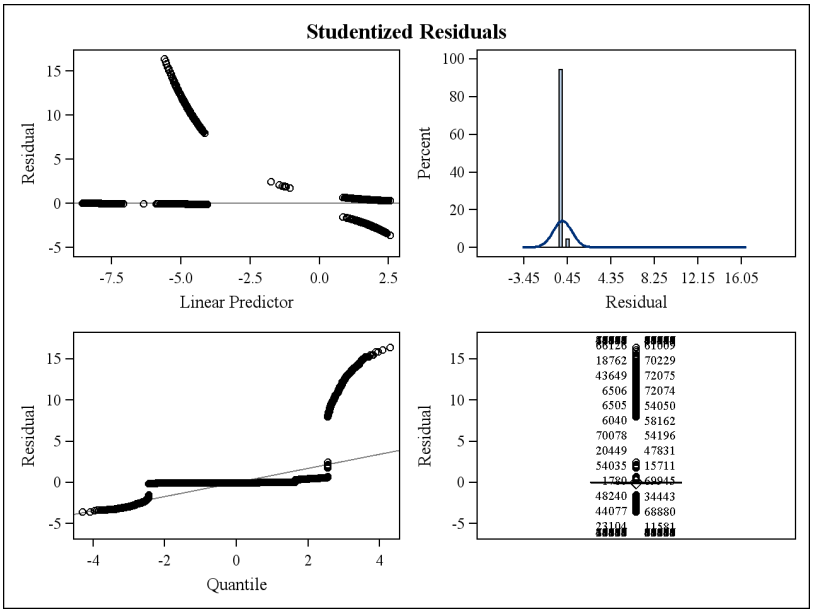
	Recherche un emploi	Ne recherche pas d'emploi	*% marginal*
Chômeur BIT	3 254	**382**	4,97%
N'est pas chômeur BIT	**534**	68 983	95,03%
% marginal	5,18 %	94,82 %	73 153

Il est donc indéniable qu'il existe une très forte corrélation entre les deux variables, même s'il y a aussi des individus, en nombre non négligeable, qui recherchent un emploi sans être chômeur BIT (cette situation se comprend bien) ainsi que des individus déclarés chômeurs BIT mais qui apparaissent *in fine* dans l'enquête comme ne recherchant pas d'emploi (ce qui est, pour le non-spécialiste, beaucoup plus surprenant…). L'examen des tables produites en sortie du modèle montre clairement que si on considère un individu qui déclare au recensement ne pas rechercher d'emploi, sa probabilité estimée d'être chômeur BIT est extrêmement faible (elle est comprise le plus souvent entre 0,5% et 1%) - quelles que soient les autres variables. Si cet individu déclare rechercher du travail (dans le questionnaire Emploi), cette probabilité devient brusquement proche de 1 (le plus souvent comprise entre 80% et 90%) - là aussi quelles que soient les autres variables. Il était donc extrêmement tentant d'expérimenter la prise en compte de la variable "Recherche d'emploi" dans le modèle !

La modélisation (modèle complet initial) permet d'effectuer des tests de significativité qui montrent que toutes les variables individuelles sont significatives, sauf la nationalité et l'indicateur de vie en couple. La plus significative de toutes ces variables est … la recherche d'emploi, naturellement ! Le résidu studentizé $\hat{U}$ se trouve alors prendre des valeurs assez caractéristiques du croisement des deux variables clé citées ci-dessus. On vérifie :

	Recherche un emploi	Ne recherche pas d'emploi
Chômeur BIT	$0.3 \le \hat{U} \le 0.5$	$8 \le \hat{U} \le 15$
N'est pas chômeur BIT	$-3 \le \hat{U} \le -2$	$-0.10 \le \hat{U} \le -0.01$

Cela est compatible avec les éléments d'appréciation graphique des résidus, qui distinguent bien les quatre sous-groupes de population :



On va donc s'intéresser désormais à un modèle qui ne contient pas la variable de recherche d'emploi, puisqu'elle est fortement soupçonnée d'être à l'origine du déséquilibre d'estimation nationale. Hélas, on ne peut pas la remplacer par une variable alternative très corrélée au chômage provenant d'une autre source que le recensement, comme l'inscription à Pole Emploi, puisqu'on ne dispose pas de fichier individuel correctement pondéré pour inférer à la population entière. On a donc testé un ajustement qui mobilise six variables explicatives (plus la constante) :

- statut d'occupation du logement
- sexe
- âge
- diplôme (2 modalités, après regroupement)
- nationalité
- état matrimonial (2 modalités, après regroupement)

La variable sexe n'est pas significative (test de Fisher : p-value = 0.47). Si on la supprime, on obtient un modèle réduit :

Effect	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-6.1625	0.3845	-16.03	<.0001
stoc_loc	-0.7554	0.03843	-19.66	<.0001
stoc_loc	0	.	.	.
AGE	3.5158	0.3888	9.04	<.0001
AGE	5.3603	0.3828	14.00	<.0001
AGE	5.2693	0.3829	13.76	<.0001
AGE	5.0095	0.3798	13.19	<.0001
AGE	4.5242	0.3804	11.89	<.0001
AGE	0	.	.	.
dipl_binaire	-0.5705	0.03813	-14.96	<.0001
dipl_binaire	0	.	.	.
nat	-0.5125	0.05949	-8.62	<.0001
nat	0.1269	0.09095	1.40	0.1628
nat	0	.	.	.

Effect	Estimate	Standard Error	t Value	Pr > |t|
matr_celib	-0.5323	0.04400	-12.10	<.0001
matr_celib	0	.	.	.

Il n'est pas surprenant que nombre de variables apparaissent significatives, car la taille de l'échantillon Emploi est énorme (dans ces conditions, les variances des coefficients sont suffisamment faibles pour qu'on considère ces coefficients comme non nuls). D'après les traditionnels critères de qualité (tableau ci-dessous), les performances de la modélisation sont manifestement très inférieures à celle que l'on obtient si on inclut la variable de recherche d'emploi, ce qui n'est pas pour nous surprendre.

Fit Statistics	Modèle initial	Modèle réduit
-2 Res Log Pseudo-Likelihood	7 518	25 136
Critère d'Akaike (AIC)	7 556	25 158
Generalized Chi-Square	52 084	75 904

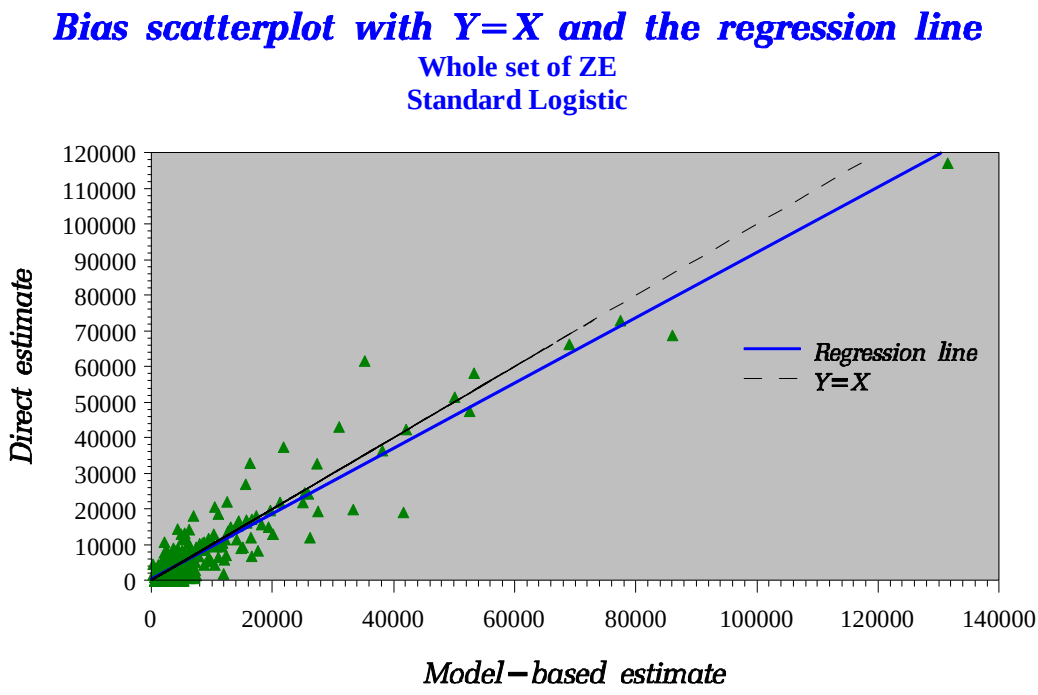
Mais la contrepartie de cette perte est le retour à une estimation nationale beaucoup plus conforme : on estime désormais à 2 504 000 le nombre total de chômeurs en France métropolitaine, ce qui redonne un caractère vraisemblable à la modélisation logistique.

Les tableaux suivants donnent les distributions des estimations petits domaines à l'issue d'un benchmarking respectivement par ZEAT et au niveau national. On constate que le niveau d'application du benchmarking, sans être très influent, n'est malgré tout pas anodin.

Quantiles (benchmarking par ZEAT))		
Quantile		**Estimate**
100%	**Max**	116 677
99%		61256
95%		23 565
90%		15 286
75%	**Q3**	6 496
50%	**Median**	3 494
25%	**Q1**	1 983
10%		1 232
5%		983
1%		577
0%	**Min**	393

Quantiles (benchmarking national)		
Quantile		**Estimate**
100%	**Max**	126 990
99%		66 670
95%		24 600
90%		15 144
75%	**Q3**	6 657
50%	**Median**	3 557
25%	**Q1**	2 095
10%		1 241
5%		949
1%		649
0%	**Min**	429

La comparaison des estimations directes et des estimations petits domaines conduit à conclure à la présence d'un biais (test). On constate un phénomène de shrinkage, mais "à l'envers" (les estimations directes sont plutôt un peu moins dispersées que les estimations logistiques) :



Notons que le shrinkage reprend sa forme habituelle si on pratique un benchmarking national.

6.2.2. Modélisation mixte non pondérée

Dans toute cette partie, on ne tient toujours pas compte des poids de sondage de l'enquête Emploi dans la phase d'estimation des paramètres du modèle. Dans un premier temps, si on inclut la variable explicative traduisant la recherche d'emploi malgré les critiques qui ont pu être formulées en 6.2.1, on doit faire face - de manière inattendue - à des soucis de convergence. En effet, le paramétrage par défaut utilisé par SAS pour la maximisation de la vraisemblance impose un changement de valeur de la fonction objectif (la log densité restreinte) inférieur à 10^{-8} entre deux étapes successives de l'algorithme. Or cette valeur est bien trop petite, le logiciel conclut qu'il n'y a pas de convergence. Il faut donc assouplir ce paramètre technique et le porter à 10^{-5} pour que la convergence ait lieu. On trouve $\hat{\sigma}_v^2 = 0.063$ mais l'écart-type de cet estimateur est 0.035, ce qui le rend à la limite de la significativité. On obtient *in fine*, exactement comme dans le cas du modèle classique (6.2.1), probablement pour la raison qui a déjà été soulevée, une surestimation manifeste du nombre total de chômeurs. Cette piste n'est donc pas la bonne.

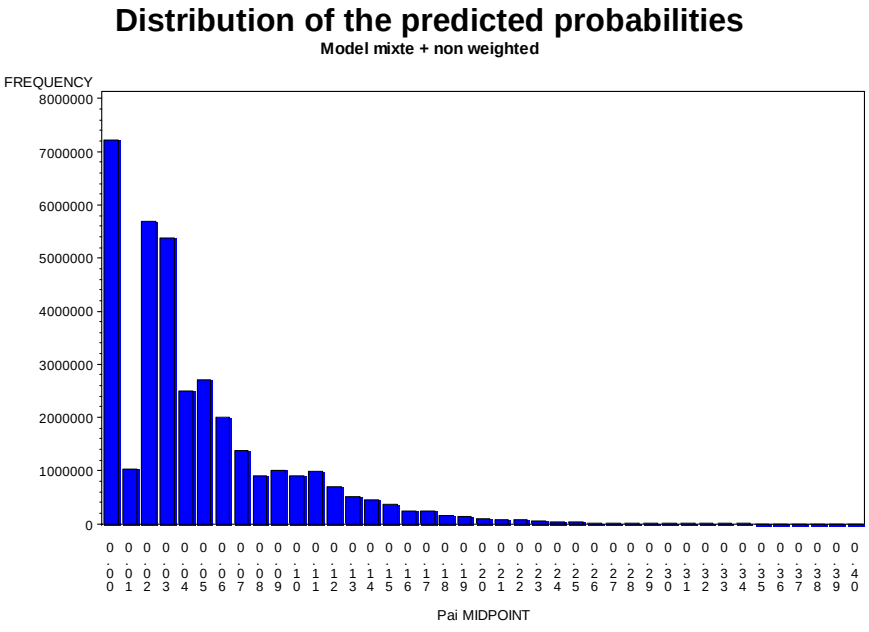
Si on reprend les variables *in fine* retenues pour l'estimation du modèle classique, donc après l'abandon de la variable "Recherche d'emploi", on ne parvient pas à se débarrasser du problème de manque de convergence, il faut donc maintenir le paramètre technique à 10^{-5}. L'estimation nationale du nombre de chômeurs BIT est désormais égale à 2 484 000 chômeurs, ce qui est tout à fait satisfaisant, surtout si on réalise que toute cette procédure d'estimation logistique "petits domaines" est extrêmement éloignée de la procédure d'estimation de l'enquête Emploi, qui relève pour sa part de la théorie classique des sondages (ces deux approches relèvent d'esprits différents, et utilisent des outils différents).

Les effets locaux v_d ont une variance estimée $\hat{\sigma}_v^2$ = 0.082. L'écart-type de $\hat{\sigma}_v^2$ est estimé à la valeur 0.017, ce qui rend le paramètre σ_v^2 très significativement différent de zéro : il y a donc lieu de prendre en compte des effets ZE.

Toutes les variables retenues sont significatives (encore une fois, ce n'est pas surprenant compte tenu de la taille de l'échantillon) :

Effect	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-6.1460	0.3855	-15.94	<.0001
stoc_loc	-0.7615	0.03919	-19.43	<.0001
stoc_loc	0	.	.	.
AGE	3.4938	0.3889	8.98	<.0001
AGE	5.3303	0.3829	13.92	<.0001
AGE	5.2553	0.3830	13.72	<.0001
AGE	4.9974	0.3798	13.16	<.0001
AGE	4.5184	0.3805	11.88	<.0001
AGE	0	.	.	.
dipl_binaire	-0.5320	0.03863	-13.77	<.0001
dipl_binaire	0	.	.	.
nat	-0.5434	0.06131	-8.86	<.0001
nat	0.1085	0.09149	1.19	0.2356
nat	0	.	.	.
matr_celib	-0.5539	0.04428	-12.51	<.0001
matr_celib	0	.	.	.

La distribution des probabilités individuelles prédites par le modèle et considérée sur l'ensemble des individus recensés en France appartenant au champ de l'enquête (donc sur 35 millions d'individus) est la suivante :

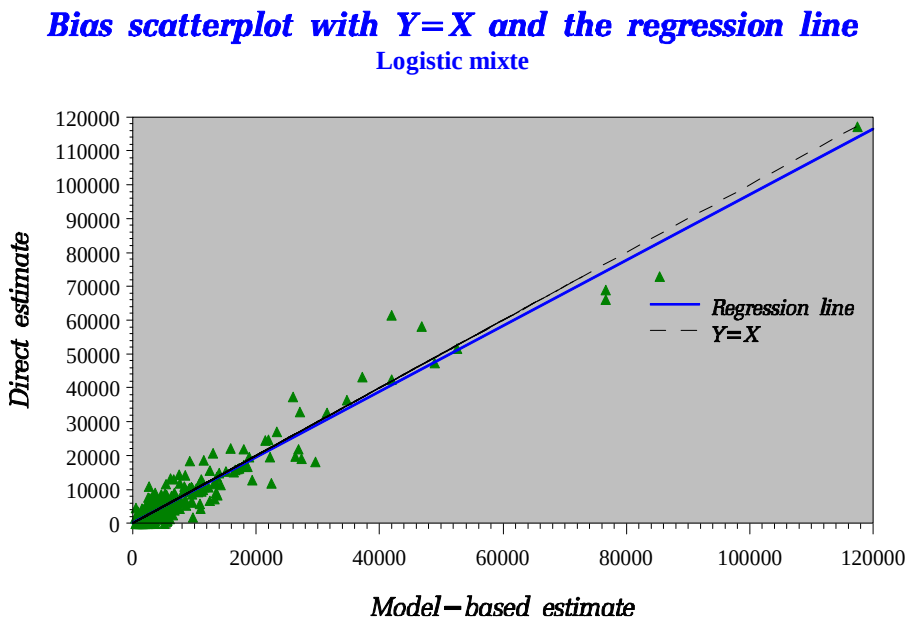


La distribution de l'estimateur logistique avec effet ZE, respectivement avant et après benchmarking national, est la suivante :

Quantiles Avant benchmarking		
Quantile		Estimate
100%	Max	117 561
99%		76 723
95%		26 050
0%		15 188
75%	Q3	6 839
50%	Median	3 421
25%	Q1	2 127
10%		1 281
5%		933
1%		616
0%	Min	492

Quantiles Après benchmarking		
Quantile		Estimate
100%	Max	114 345
99%		74 624
95%		25 337
90%		14 772
75%	Q3	6 652
50%	Median	3 328
25%	Q1	2 069
10%		1 246
5%		907
1%		599
0%	Min	479

Le test de biais habituel conclut à l'absence de biais de l'estimateur logistique mixte (pas de différence significative entre la droite de régression et la droite y=x).



Le tableau suivant fournit des estimations par ZEAT (regroupement de régions - donc de ZE) en comparant à ce niveau 3 estimateurs "petits domaines" avec l'estimateur direct, que l'on peut considérer comme étant de bonne qualité au niveau ZEAT : on calcule ainsi les erreurs relatives de l'estimateur logistique sans effet aléatoire (classique_zeat), de l'estimateur logistique avec effet aléatoire (mixte_zeat) et de l'estimateur synthétique utilisé actuellement par l'Insee (est_Insee). Attention, l'estimateur Insee est par nature "presque" calé[23] sur l'estimateur direct national - mais ce n'est pas du tout le cas des deux estimateurs logistiques.

[23] "Presque" parce que les séries brutes ont été recalculées pour les besoins de l'exercice, et il n'y a pas eu de calage final dédié à l'opération.

ZEAT	est_direct	classique_zeat	mixte_zeat	est_Insee	Err classique	Err mixte	Err Insee
1	514275	580130	536288	459889	+12.8 %	+ 4.3 %	- 10.6 %
2	394334	407733	404248	403680	+ 3.4 %	+ 2.5 %	+ 2.4 %
3	218159	161315	198601	200330	- 26.1 %	- 9.0 %	- 8.2 %
4	181870	211113	198639	201225	+ 16.1 %	+ 9.2 %	+ 10.6 %
5	257199	289244	279685	277946	+ 12.5 %	+ 8.7 %	+ 8.1 %
7	247656	249932	239576	248665	+ 0.9 %	- 3.3 %	+ 0.4 %
8	241831	292738	283396	250347	+ 21.1 %	+ 17.2 %	+ 3.5 %
9	361093	312268	343956	365490	- 13.5 %	- 4.7 %	+ 1.2 %
TOTAL	**2 416 417**	**2 504 473**	**2 484 389**	**2 407 572**	**Sigma = 106,4**	**Sigma = 58,9**	**Sigma = 45,0**

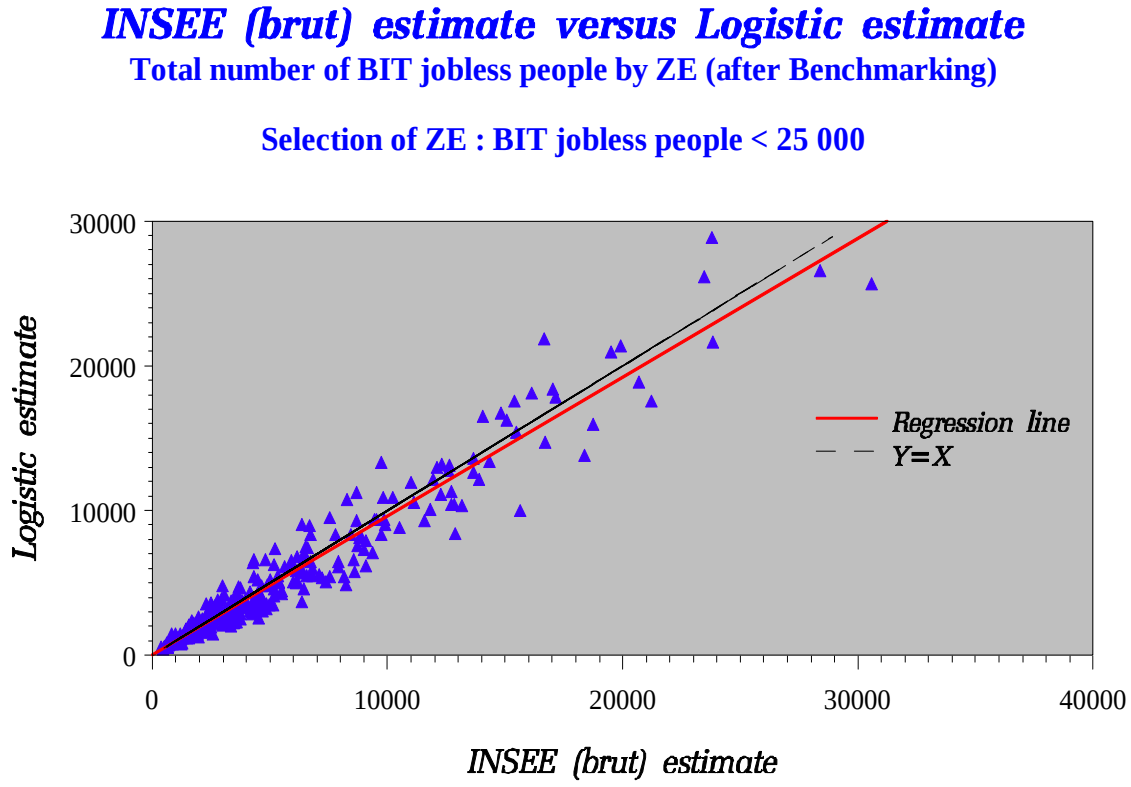
Si on effectue un benchmarking national sur les estimateurs logistiques et sur l'estimateur Insee, les résultats évoluent ainsi :

ZEAT	est_direct	classique_zeat	mixte_zeat	est_Insee	Err classique	Err mixte	Err_Insee
1	514275	559730	521617	461578	+ 8.8 %	+ 1.4 %	- 10.2 %
2	394334	393393	393188	405165	- 0.2 %	- 0.3%	+ 2.7 %
3	218159	155643	193167	201066	- 28.7 %	-11.5 %	- 7.8 %
4	181870	203692	193205	201962	+ 12.0 %	+ 6.2 %	+ 11.0 %
5	257199	279075	272034	278968	+ 8.5 %	+ 5.8 %	+ 8.5 %
7	247656	241140	233023	249580	- 2.6 %	- 5.9 %	+ 0.8 %
8	241831	282444	275643	251265	+ 16.8 %	+ 14.0 %	+ 3.9 %
9	361093	301291	334543	366833	- 16.6%	- 7.4 %	+ 1.6 %
TOTAL	**2 416 417**	**2 416 408**	**2 416 420**	**2 416 417**	**Sigma = 94,2**	**Sigma = 52,5**	**Sigma = 46,5**

Après benchmarking, on constate en particulier que, partant d'un estimateur national par construction égal au direct, un ensemble d'estimations formées sur une partition pourtant assez grossière du territoire (la ZEAT, en la circonstance) suffit pour donner lieu à des écarts parfois importants avec l'estimation de référence directe. Une piste d'amélioration éventuelle consisterait à ajuster des modèles différents par groupe de ZEAT. Il apparaît que le benchmarking améliore (un peu) la qualité des estimations logistiques (ce qui est naturel), que l'estimation mixte est nettement préférable à l'estimation non mixte, mais qu'*in fine* les estimateurs actuels Insee apparaissent plutôt meilleurs que les estimateurs logistiques mixtes (et *a fortiori* que les estimateurs logistiques classiques). Néanmoins, il convient de souligner la belle performance de l'estimateur mixte, qui ne perd que modestement par rapport à l'estimateur Insee, alors même que le modèle sur lequel il est formé implique des variables individuelles qui apparaissent spontanément bien moins pertinentes que les DEFM[24] pour expliquer un phénomène aussi complexe que le chômage. Il est donc bien dommage que l'on ait perdu les variables individuelles les plus prometteuses, à savoir la recherche d'emploi et le positionnement spontané par rapport à l'état de chômage : on aurait pu espérer une amélioration très sensible de la méthode actuelle si ces variables avaient affiché une plus grande homogénéité entre les sources …

La comparaison estimateur INSEE / estimateur logistique mixte après benchmarking (donnée ici en excluant les plus grandes ZE) garde une allure d'ensemble satisfaisante :

[24] On rappelle (voir introduction) qu'il s'agit de la variable auxiliaire utilisée par la méthode Insee actuelle.



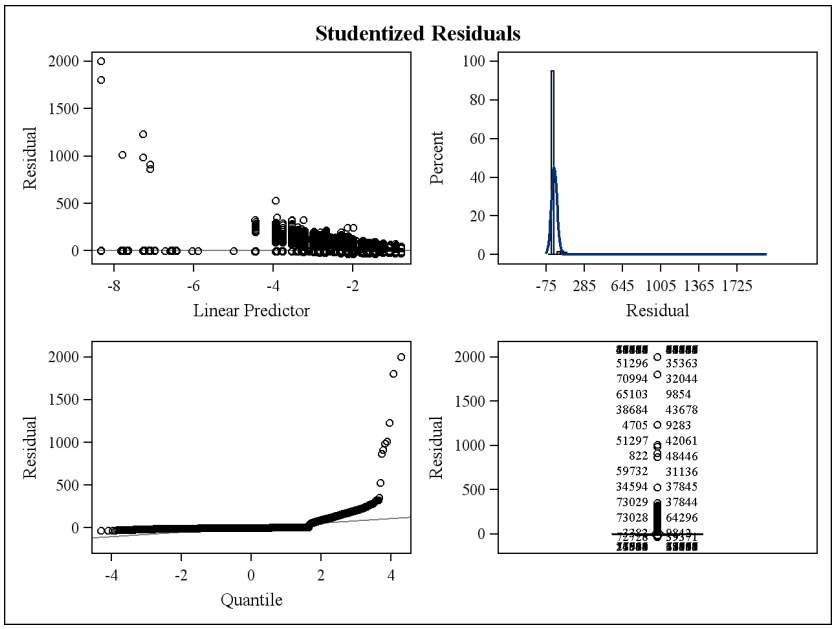
6.2.3. Modélisation standard pondérée

Désormais, on pondère les données individuelles par les poids de l'enquête Emploi. Les estimateurs des paramètres du modèle stochastique sont donc des estimateurs du maximum de la vraisemblance pondérée. Cela change totalement le contexte numérique parce que toute la procédure se déroule désormais comme si on traitait un échantillon environ 680 fois plus gros que dans l'approche non pondérée (mettons, puisque le poids moyen dans l'enquête Emploi vaut 680). On obtient

Effect	Weighted Estimate	Standard Error	t Value	Pr > |t|	Rappel estimateur NON pondéré
Intercept	-6.0200	0.01408	-427.67	<.0001	-6.1625
stoc_loc	-0.6936	0.001471	-471.66	<.0001	-0.7554
stoc_loc	0	.	.	.	0
AGE	3.3343	0.01427	233.59	<.0001	3.5158
AGE	5.1042	0.01401	364.24	<.0001	5.3603
AGE	4.9687	0.01402	354.45	<.0001	5.2693
AGE	4.7973	0.01389	345.45	<.0001	5.0095
AGE	4.3940	0.01391	315.90	<.0001	4.5242
AGE	0	.	.	.	0
dipl_binaire	-0.5431	0.001482	-366.37	<.0001	-0.5705
dipl_binaire	0	.	.	.	0
nat	-0.5152	0.002307	-223.30	<.0001	-0.5125
nat	0.1407	0.003527	39.89	<.0001	0.1269
nat	0	.	.	.	0
matr_celib	-0.5496	0.001683	-326.46	<.0001	-0.5323
matr_celib	0	.	.	.	0

En particulier, la présence des poids augmente considérablement les valeurs des statistiques de test puisque les écart-types estimés des coefficients du modèle s'effondrent. Ainsi, on a encore plus tendance à considérer les variables auxiliaires en jeu comme explicatives. En revanche - et conformément au bon sens - les coefficients de la régression sont numériquement peu perturbés par la pondération : c'est un point fondamental, qui signifie que les probabilités individuelles prédites ne seront (heureusement !) que peu modifiées par rapport au cas non pondéré.

Dans cette partie, on abandonne tout effet aléatoire local au niveau ZE. Les résidus évoluent ainsi :



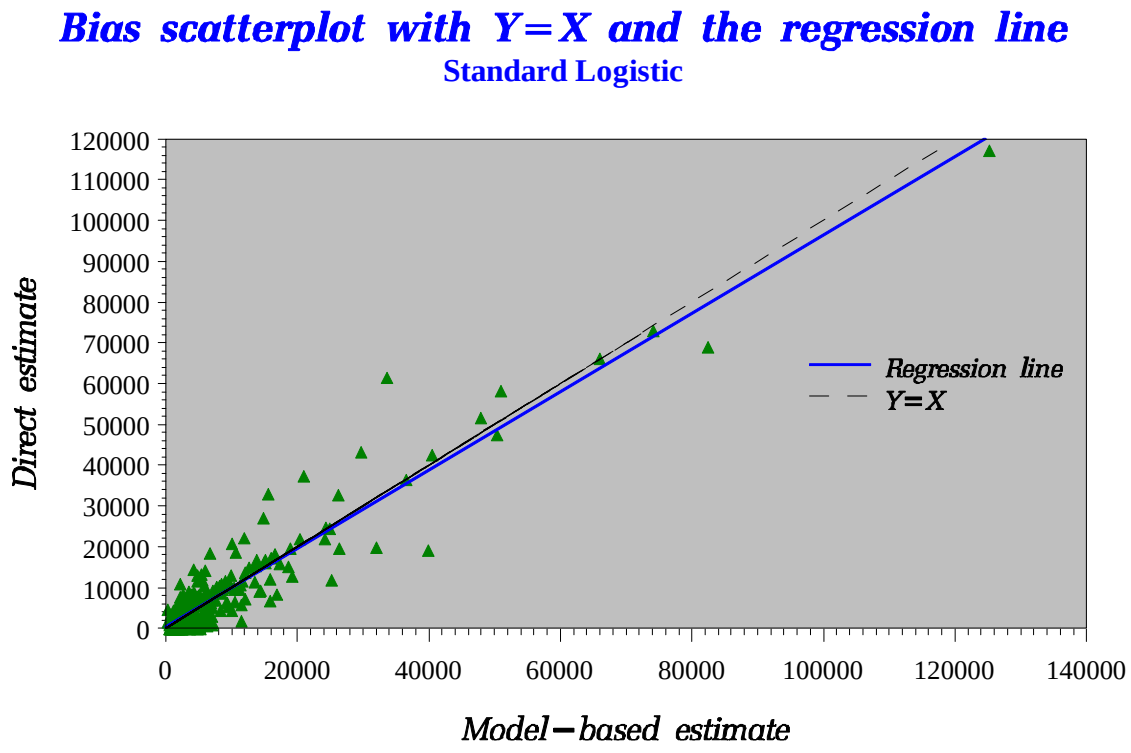
Le jeu de coefficients résultant de la prise en compte des poids conduit à une estimation nationale de 2 409 000 chômeurs, ce qui est évidemment très satisfaisant. On peut y voir une traduction de la propriété théorique énoncée en fin de partie 6.1 (l'estimation nationale pondérée des $\hat{P}_{d,i}$ dans l'échantillon Emploi - qui est d'après cette propriété l'estimation nationale du nombre de chômeurs - devrait être proche de l'estimation nationale pondérée dans le recensement).

Les tableaux suivants donnent les distributions des estimations petits domaines à l'issue d'un benchmarking, respectivement par ZEAT et au niveau national. On compare à la distribution de l'estimateur non pondéré - laquelle apparaît très voisine, quel que soit le niveau du benchmarking.

Quantiles (benchmarking par ZEAT)			
Quantile		Modèle pondéré	Modèle non pondéré
100%	Max	115 961	116 677
99%		61 082	61 256
95%		23 618	23 565
90%		15 286	15 286
75%	Q3	6 516	6 496
50%	Median	3 519	3 494
25%	Q1	1 997	1 983
10%		1 237	1 232
5%		984	983
1%		584	577
0%	Min	397	393

Quantiles (benchmarking national)			
Quantile		Modèle pondéré	Modèle non pondéré
100%	Max	125 324	126 990
99%		66 014	66 670
95%		24 445	24 600
90%		14 980	15 144
75%	Q3	6 650	6 657
50%	Median	3 553	3 557
25%	Q1	2 107	2 095
10%		1 246	1 241
5%		963	949
1%		648	649
0%	Min	426	429

La comparaison des estimations directes et des estimations "petits domaines" conduit à conclure à l'absence de biais (test). On obtient



6.2.4. Modélisation mixte pondérée

On utilise de nouveau les poids de l'enquête Emploi, et on ajoute dans le modèle un effet aléatoire propre à chaque ZE. Les effets locaux v_d ont une variance estimée $\hat{\sigma}_v^2$ = 4.839. L'écart-type de $\hat{\sigma}_v^2$ est estimé à la valeur 0.426, ce qui rend le paramètre σ_v^2 très nettement différent de zéro : il y a donc lieu, sans réserve, de prendre en compte des effets ZE. On constate par ailleurs combien cet

estimateur de variance a augmenté numériquement par rapport au cas non pondéré (il était alors égal à 0.082).

Effect	Estimate	Standard Error	t Value	Pr > |t|	Rappel estimateur NON pondéré
Intercept	-6.7575	0.1215	-55.62	<.0001	-6.1460
stoc_loc	-0.6881	0.001547	-444.74	<.0001	-0.7615
stoc_loc	0	.	.	.	0
AGE	3.3074	0.01429	231.48	<.0001	3.4938
AGE	5.0960	0.01403	363.31	<.0001	5.3303
AGE	4.9900	0.01403	355.68	<.0001	5.2553
AGE	4.8003	0.01389	345.50	<.0001	4.9974
AGE	4.3974	0.01391	316.02	<.0001	4.5184
AGE	0	.	.	.	0
dipl_binaire	-0.4967	0.001529	-324.93	<.0001	-0.5320
dipl_binaire	0	.	.	.	0
nat	-0.5342	0.002428	-219.97	<.0001	-0.5434
nat	0.1236	0.003582	34.50	<.0001	0.1085
nat	0	.	.	.	0
matr_celib	-0.5796	0.001713	-338.37	<.0001	-0.5539
matr_celib	0	.	.	.	0

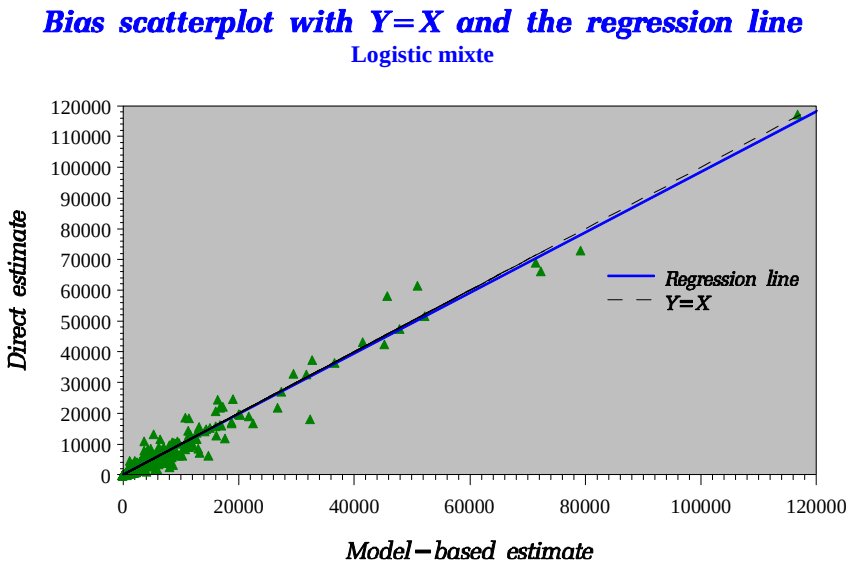
L'estimation nationale du nombre de chômeurs BIT est égale à 2 409 000 chômeurs[25]. La distribution de l'estimateur logistique avec effet ZE est la suivante, respectivement sans benchmarking et après un benchmarking au niveau national :

Quantiles Avant benchmarking		
Quantile		**Estimate**
100%	**Max**	125 324
99%		66 014
95%		24 445
90%		14 980
75%	**Q3**	6 650
50%	**Median**	3 553
25%	**Q1**	2 107
10%		1 246
5%		963
1%		648
0%	**Min**	426

Quantiles Après benchmarking		
Quantile		**Estimate**
100%	**Max**	125 690
99%		66 207
95%		24 516
90%		15 024
75%	**Q3**	6 669
50%	**Median**	3 563
25%	**Q1**	2 113
10%		1 250
5%		966
1%		650
0%	**Min**	427

Le test de biais habituel conclut à l'absence de biais de l'estimateur logistique mixte pondéré (pas de différence significative entre la droite de régression et la droite y=x).

[25] C'est, de notre point de vue, un pur hasard de retrouver exactement l'effectif du modèle classique.

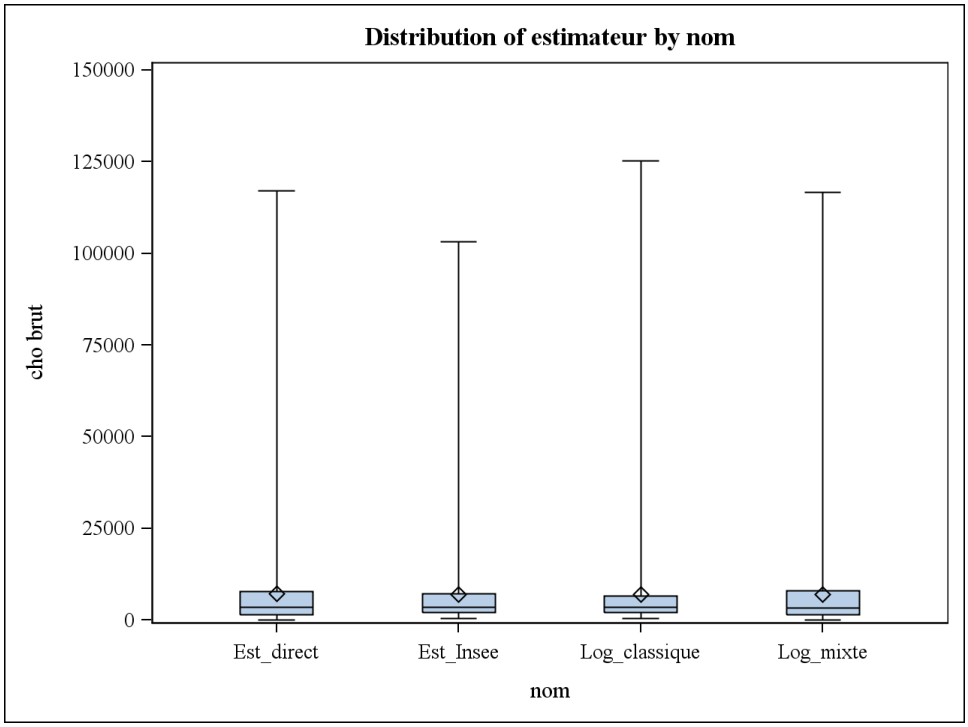


.
Le tableau suivant fournit des estimations par ZEAT, sans aucun benchmarking, dans les conditions exposées en partie 6.2.2 :

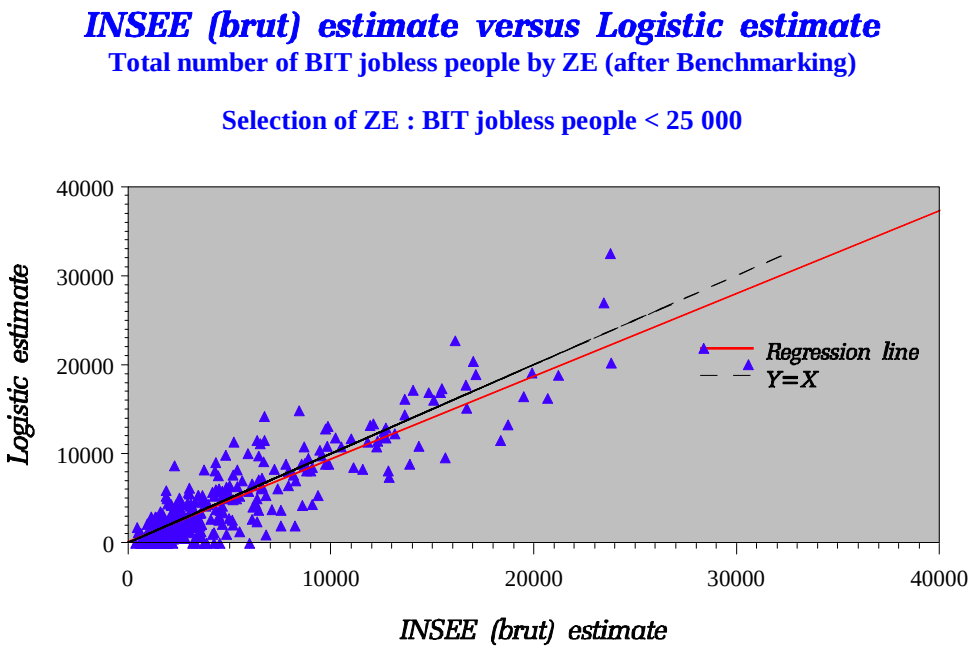
ZEAT	est_direct	classique_zeat	mixte_zeat	est_Insee	Err classique	Err mixte	Err Insee
1	514275	555801	518037	459889	+ 8.1 %	+ 0.7 %	- 10.6 %
2	394334	392664	397832	403680	- 0.4 %	+ 0.9 %	+ 2.4 %
3	218159	154755	217117	200330	- 29.1 %	- 0.5 %	- 8.2 %
4	181870	203284	176332	201225	+ 11.8 %	- 3.0 %	+ 10.6 %
5	257199	279330	258491	277946	+ 8.6 %	+ 0.5 %	+ 8.1 %
7	247656	241240	231881	248665	- 2.6 %	- 6.4 %	+ 0.4 %
8	241831	281922	248592	250347	+ 16.6 %	+ 2.8 %	+ 3.5 %
9	361093	300394	360625	365490	- 16.8 %	- 0.1 %	+ 1.2 %
TOTAL	**2 416 417**	**2 409 390**	**2 408 907**	**2 407 572**	**Sigma = 94,0**	**Sigma = 14,9**	**Sigma = 45,0**
RAPPEL *Cas non pondéré*	//	**2 504 473**	**2 484 389**	//	**Sigma = 106,4**	**Sigma = 58,9**	**Sigma = 45,0**

On vérifie que le benchmarking national sur ces estimateurs logistiques ne génère pas de gain de qualité. On retiendra l'excellente performance de l'estimateur pondéré mixte - au-delà de toute espérance, même. Il faut probablement mettre une part de son efficacité au crédit d'un heureux hasard, car il reste assez difficile de croire à la pertinence des variables explicatives retenues pour expliquer le chômage - même si la présence d'un effet aléatoire apporte un gain substantiel à la modélisation. En tout cas, la prise en compte de la pondération semble avoir conduit à une estimation du σ_v^2 tout à fait pertinente, sensiblement plus forte et (en partie pour cette raison) bien meilleure que celle que donnait le cas non pondéré : l'inefficacité relative des effets fixes se trouve alors compensée par l'existence d'effets aléatoires bien adaptés. A ce stade, nous ne pouvons néanmoins pas en tirer de généralité sur les vertus comparées de l'approche non pondérée versus l'approche pondérée (au-delà des traditionnels discours sur la question …).

Le graphique ci-dessous compare les distributions avant benchmarking - où l'on voit que la distribution la plus proche de celle de l'estimateur direct est bien la distribution de l'estimateur mixte (le phénomène de shrinkage de l'estimateur synthétique "officiel" de l'Insee est ici assez net) :



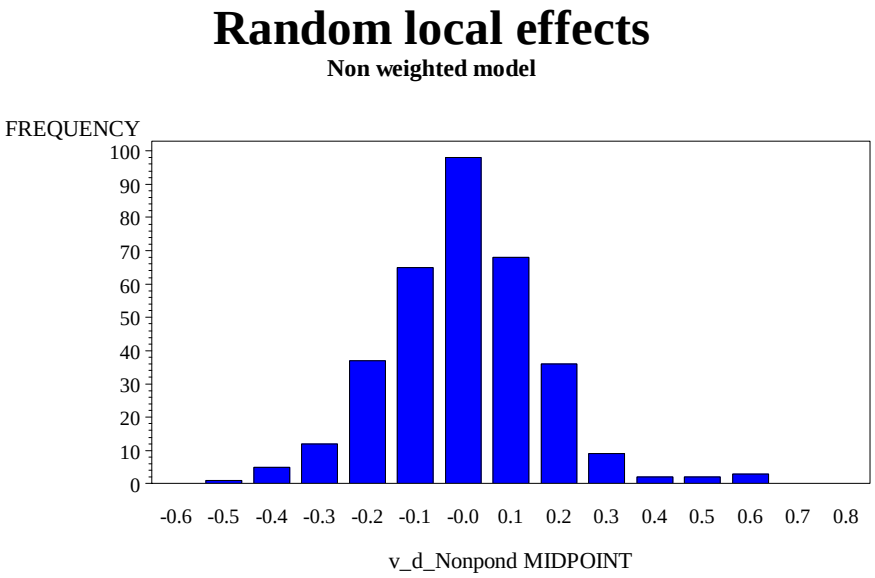
La comparaison estimateur INSEE / estimateur logistique mixte avant benchmarking (en excluant les plus grandes ZE) garde une allure d'ensemble satisfaisante. Cela n'empêche évidemment pas la présence d'estimations "petits domaines" aberrantes, car fort éloignées de la statistique officielle.

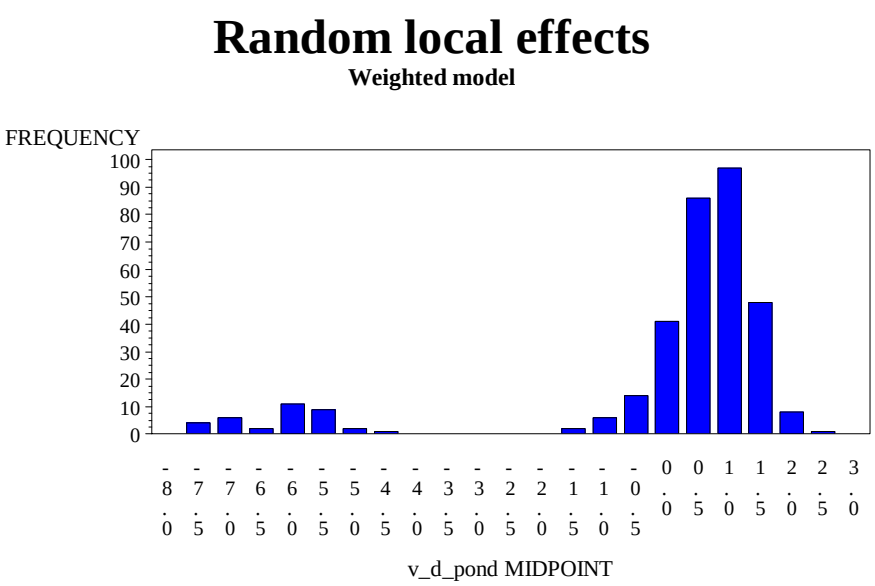


Ce qui suit tente d'éclairer l'effet de la pondération sur la procédure d'estimation.

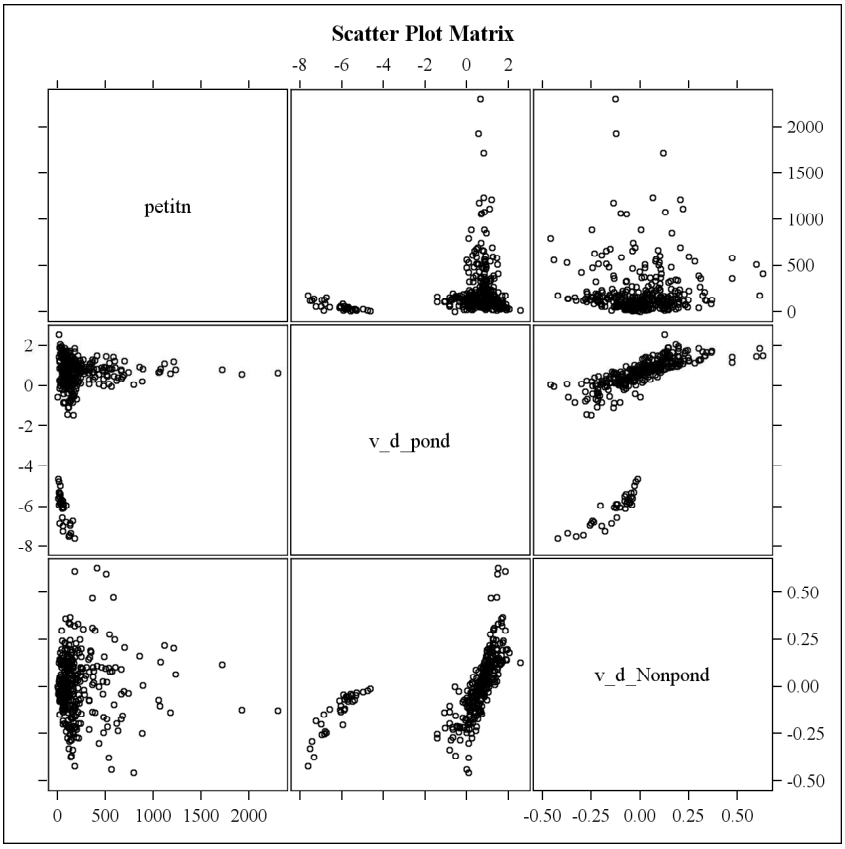
On donne ci-dessous la distribution de l'effet aléatoire $\hat{v}_d$ du modèle mixte, respectivement pour le modèle non pondéré (variation de - 0.46 à + 0.63) et pour le modèle pondéré (variation de - 7.6 à + 2.6), ce qui confirme bien la forte disparité des situations. L'apparition de deux groupes de valeurs

dans le cas pondéré est - à ce jour - une curiosité. On rappelle ici que les poids de l'enquête Emploi apparaissent *in fine* relativement dispersés.





L'importance numérique de l'effet local du modèle pondéré apparait liée de manière non linéaire à la taille de l'échantillon, comme le montre le graphique qui suit : en particulier, on ne trouve des effets aléatoires fortement négatifs que dans les ZE à petite taille d'échantillon (par contre, dans la sous-population des ZE de petite taille, on trouve aussi des effets faibles ou parmi les plus forts des effets positifs). Ce type de relation disparaît dans le cas du modèle non pondéré.



Il est intéressant de constater que les ZE ayant de toutes petites tailles d'échantillon (disons en dessous de 50 répondants, jusqu'au cas de la ZE comprenant 1 seul répondant) ne se distinguent pas particulièrement dans tout ce processus d'estimation : en tout cas, pour chacune d'entre elles le modèle estime sans difficulté un effet aléatoire, lequel prend manifestement une valeur numérique tout à fait acceptable et qui se fond dans la distribution de l'ensemble. Par ailleurs, on vérifie qu'il ne semble pas y avoir de relation entre la valeur de l'effet aléatoire et la géographie.

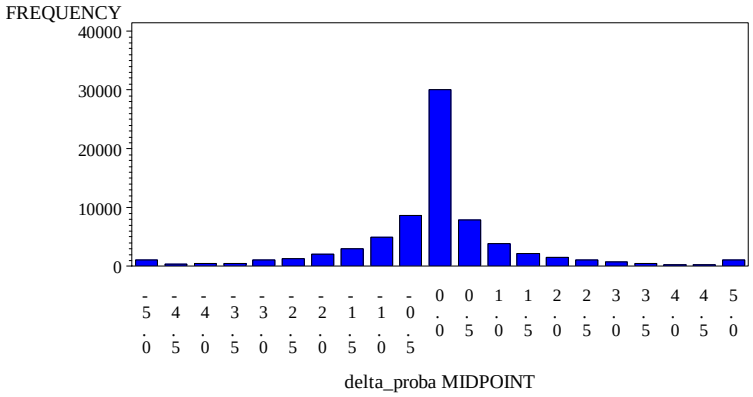
On peut s'intéresser à la distribution de l'écart entre la probabilité individuelle prédite issue du modèle pondéré (mixte par exemple) et la probabilité individuelle prédite issue du modèle non pondéré. On trouve essentiellement de très petites valeurs en écart absolu, le plus souvent moins de 2 points de pourcentage d'écart (ce qui est rassurant), mais les variations relatives sont parfois considérables :

Quantiles : écart absolu		
Quantile		Estimateur
100%	Max	0.250
99%		0.059
95%		0.024
90%		0.013
75%	Q3	0.003
50%	Median	0.00004
25%	Q1	-0.005
10%		-0.017
5%		-0.027
1%		-0.056
0%	Min	-0.240

Quantiles : écart relatif (%)		
Quantile		Estimateur
100%	Max	588.5
99%		116.8
95%		59.3
90%		39.3
75%	Q3	18.5
50%	Median	1.7
25%	Q1	-14.19
10%		-35.6
5%		-63.0
1%		-99.9
0%	Min	-100.0

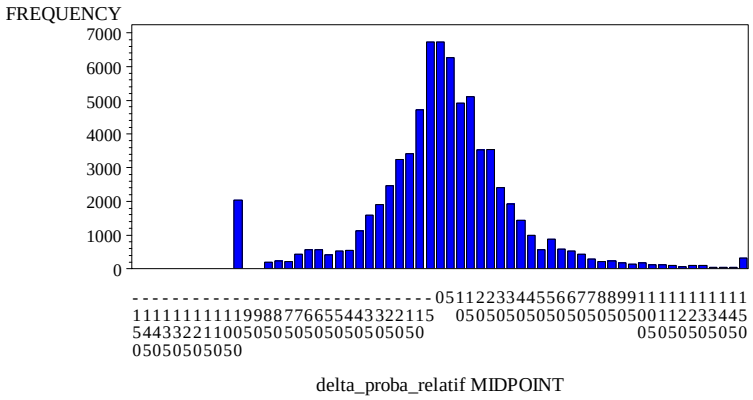
Absolute delta : weighted / non weighted

X-axis in percent



Relative delta : weighted / non weighted

X-axis in percent



Un modèle "sain" n'a pas vocation à produire de forts effets aléatoires - normalement ce sont les effets fixes qui devraient expliquer la plus grande partie de la probabilité d'être chômeur. On pourrait donc imaginer d'annuler l'effet aléatoire d'une ZE dès lors qu'il apparaît prendre une valeur excessive : concrètement, on peut déclarer comme valeur excessive toute valeur ce qui se trouve dans le sous-groupe des ZE situées "à gauche" lorsqu'on représente la distribution des effets locaux (cf graphique supra). Il s'agit donc des ZE où $\hat{v}_d < -3.5$. On ajuste le modèle sur 338 ZE, mais on impose $\hat{v}_d = 0$ pour ces ZE. Cela conduit à une estimation de 2 444 000 chômeurs BIT France entière. Au niveau ZEAT, on obtient avant benchmarking :

Elimination des ZE avec $\hat{v}_d < -3.5$

ZEAT	est_direct	classique_zeat	mixte_zeat	est_Insee	Err classique	Err mixte	Err Insee
1	514275	555801	519982	459889	+ 8.1 %	+ 1.1 %	- 10.6 %
2	394334	392664	405701	403680	- 0.4 %	+ 2.9 %	+ 2.4 %
3	218159	154755	217117	200330	- 29.1 %	- 0.5 %	- 8.2 %
4	181870	203284	178764	201225	+ 11.8 %	- 1.7 %	+ 10.6 %
5	257199	279330	263693	277946	+ 8.6 %	+ 2.5 %	+ 8.1 %
7	247656	241240	237982	248665	- 2.6 %	- 3.9 %	+ 0.4 %
8	241831	281922	254249	250347	+ 16.6 %	+ 5.1 %	+ 3.5 %
9	361093	300394	366486	365490	- 16.8 %	+ 1.5 %	+ 1.2 %
TOTAL	**2 416 417**	**2 409 390**	**2 443 974**	**2 407 572**	**Sigma = 94,0**	**Sigma = 19.2**	**Sigma = 45,0**

On constate, en examinant les estimations par ZE, que les valeurs totalement aberrantes ont disparu. En revanche, on conserve, sur certaines ZE, des valeurs peu crédibles - parce qu'elles apparaissent malgré tout très (trop) différentes des autres estimations, en particulier il demeure un écart considérable avec l'estimation Insee et/ou l'estimation effectuée à partir du modèle standard (non mixte).

Une seconde piste consiste à agir encore plus en amont et à ajuster le modèle en supprimant d'emblée toutes les ZE pour lesquelles on a (par exemple) une proportion de chômeurs inférieure à un certain seuil, mettons 1.5%. On trouve une estimation nationale "petits domaines" de 2 561 000 chômeurs France entière, ce qui n'est pas satisfaisant, d'autant plus que certaines ZE conservent des estimations fantaisistes. L'augmentation de l'estimation nationale était prévisible puisque les effets aléatoires des ZE concernées sont fortement négatifs et que leur suppression conduit donc à augmenter les effectifs de chômeurs des petites ZE.

7. Le modèle EBLUP-B

7.1. Éléments de théorie

On note toujours d l'identifiant de la ZE. On désigne par $\bar{y}_d$ la moyenne arithmétique simple des valeurs individuelles Y_i calculée sur les n_d individus répondants à l'enquête dans la ZE d où, comme dans la partie précédente, Y_i vaut 1 si i est chômeur BIT et 0 sinon. Ainsi, $\bar{y}_d$ est la proportion empirique des chômeurs dans l'échantillon recoupant la ZE. Cet estimateur ne fait pas intervenir les poids de sondage, il est donc (largement) biaisé du point de vue de l'aléa de sondage. L'approche de cette partie est donc uniquement modèle-dépendante puisqu'elle ne tient pas compte de la manière dont les individus sont échantillonnés. On est donc dans l'esprit d'une approche entièrement économétrique. Le modèle théorique est posé au niveau agrégé de la ZE selon :

$$\bar{y}_d = B^t \cdot \bar{X}_d + v_d + e_d$$

où B est un paramètre vectoriel inconnu, avec les hypothèses : $Ev_d = Ee_d = 0$, $Var(v_d) = \sigma_v^2$ et $Var(e_d) = \frac{\sigma_e^2}{n_d}$. Les aléas v_d et e_d sont mutuellement indépendants. Les grandeurs σ_v^2 et σ_e^2 sont des paramètres à estimer.

L'application d'un tel modèle, en particulier la forme de la variance des e_d, trouve une justification naturelle dans un modèle préalable décliné au niveau individuel, à savoir

$$Y_{d,i} = B^t \cdot X_{d,i} + v_d + e_{d,i}$$

où $Ev_d = Ee_{d,i} = 0$, $Var(v_d) = \sigma_v^2$ et $Var(e_{d,i}) = \sigma_e^2$. Il suffit d'effectuer la moyenne des différents termes sur chaque ZE, puis à remplacer $\bar{x}_d$ par $\bar{X}_d$ pour retrouver formellement le modèle agrégé.

Dans notre contexte, il s'agit clairement d'un détournement de finalité : en effet, le modèle défini au niveau individuel ne peut pas avoir de sens lorsque $Y_{d,i}$ est une variable qualitative (ici à deux modalités). En revanche, au stade de l'agrégation, il est tout autant recevable que le modèle de Fay et Herriot.

La théorie de la prédiction linéaire sans biais optimale conduit à retenir l'estimateur suivant, dit EBLUP-B[26] :

$$\hat{\bar{Y}}_d^H = \hat{\gamma}_d \cdot \bar{y}_d + (1-\hat{\gamma}_d) \cdot \hat{B}^t \cdot \bar{X}_d$$

où $\hat{B}$ est l'estimateur des moindres carrés généralisés. Selon l'humeur, on peut opter pour un calcul fondé sur le modèle individuel, soit (V désignant la matrice bloc-diagonale adéquate)

$$\hat{B}_1 = \left(X^t \cdot V^{-1} \cdot X\right)^{-1} \cdot \left(X^t \cdot V^{-1} \cdot Y\right)$$

ou sur le modèle agrégé, soit

$$\hat{B}_2 = \left(\sum_d \frac{\bar{X}_d \cdot \bar{X}_d^t}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_d}} \right)^{-1} \cdot \left(\sum_d \frac{\bar{X}_d \cdot \bar{y}_d}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_d}} \right)$$

On peut aussi remplacer, dans cette dernière expression, les vraies moyennes $\bar{X}_d$ par les moyennes définies sur l'échantillon $\bar{x}_d$ (coefficient de régression $\hat{B}_3$).

Dans tous les cas : $$\hat{\gamma}_d = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \frac{\hat{\sigma}_e^2}{n_d}}$$

Les estimations respectives des paramètres de variance σ_e^2 et σ_v^2 relèvent d'approches dissymétriques. Pour le premier, on choisit (dans cette étude) l'estimateur défini selon :

$$\hat{\sigma}_e^2 = \frac{1}{n-m} \sum_d \sum_{i \in d} \left(Y_{d,i} - \bar{y}_d\right)^2$$

où n est la taille totale de l'échantillon d'individus répondants et m le nombre total de domaines (ici 348 si on n'exclut aucune ZE). Avec une variable qualitative, bien que le modèle individuel n'ait pas de sens, on peut malgré tout effectuer ce calcul. Quand on ajoute l'hypothèse de normalité des aléas - qui de fait s'avère nécessaire - l'estimateur $\hat{\sigma}_v^2$ du paramètre σ_v^2 peut être obtenu par une technique de maximum de vraisemblance (donc non analytique). On est alors en mesure de fournir une estimation de l'erreur de $\hat{\bar{Y}}_d^H$ définie comme $E\left(\hat{\bar{Y}}_d^H - \bar{Y}_d\right)^2$. On montre que l'expression suivante est satisfaisante en toute circonstance :

$$\hat{E}\left(\hat{\bar{Y}}_d^H - \bar{Y}_d\right)^2 = \hat{\gamma}_d \cdot \frac{\hat{\sigma}_e^2}{n_d} + (1-\hat{\gamma}_d)^2 \left(\bar{X}_d^t \cdot \hat{V}(\hat{B}) \cdot \bar{X}_d\right) + 2 \left(\frac{\hat{\sigma}_e^2}{n_d}\right)^2 \cdot \left(\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_d}\right)^{-3} \cdot Var\left(\hat{\sigma}_u^2\right)$$

[26] Empirical Best Linear Unbiased Predictor. On lui attribue un suffixe B, le suffixe A étant réservé au modèle individuel sous-jacent lorsque Y est quantitative.

$\hat{V}(\hat{B})$ varie en $\frac{1}{m \cdot \overline{n}}$ où m désigne le nombre total de ZE participant à l'estimation du modèle et $\overline{n}$ est le nombre moyen d'individus répondants par domaine. Le troisième terme est encore plus petit. Ainsi, lorsque m est grand, le gain de précision estimé par rapport à l'estimateur $\overline{y}_d$ est de l'ordre de grandeur de $\hat{\gamma}_d$ - comme dans le modèle de Fay et Herriot.

Ce modèle est techniquement intéressant parce qu'il s'agit d'un modèle explicité au niveau ZE mais qui nécessite un retour, pour l'estimation des paramètres, aux informations individuelles. C'est donc, en quelque sorte, une troisième forme de modélisation, de niveau intermédiaire.

7.2. Les principaux résultats

On reprend exactement les variables auxiliaires $\overline{X}_d$ définitives du modèle de Fay et Herriot (partie 4). On utilise un programme SAS produit par l'ONS, qui calcule l'estimateur EBLUP_B et son erreur estimée. Ce programme est conçu pour calculer le coefficient $\hat{B}_3$ à partir du modèle individuel initial, donc il effectue à ce stade une régression sur les données de l'échantillon Emploi. Or, dans l'ensemble de l'échantillon Emploi répondant, il y a un unique individu homme de moins de 20 ans diplômé et demandeur d'emploi dans les catégories adéquates. La variable *t_age15_19HdiplBIT* étant (presque) identiquement nulle, on doit la supprimer. Il reste quatre variables, plus la constante.

- Recherche d'un emploi (variable *t_rech_oui*) ;
- Ne pas vivre en couple (variable *t_couple_2*) ;
- Inscription DEFM, catégories 1,2,3 et HAR et appartenir à la catégorie des hommes de 30 à 49 ans, peu diplômés (variable *t_age30_49HnondiplBIT*)[27];
- Inscription DEFM, catégories 1,2,3 et HAR et appartenir à la catégorie des hommes de 50 à 64 ans, peu diplômés (variable *t_age50_64HnondiplBIT*);

La modélisation s'applique aux ZE ayant 50 répondants au moins - c'est un seuil qui empiriquement apparaît satisfaisant. Pour produire les estimations relatives aux ZE ayant moins de 50 répondants, c'est l'estimateur synthétique $\hat{B}^t \cdot \overline{X}_d$ qui a été utilisé.

Le vecteur des coefficients $\hat{\beta}$ associé à la constante et aux quatre variables conservées vaut $(-0{,}035 \quad 0{,}790 \quad 0{,}085 \quad -0{,}212 \quad 0{,}310)$, coefficients respectivement associés aux variables explicatives *constante*, *t_rech_oui, t_couple_2, t_age30_49HnondiplBIT* et *t_age50_64HnondiplBIT* . Ces valeurs sont cohérentes avec celles que donne l'ajustement du modèle de Fay et Herriot, en particulier on retrouve les ordres de grandeur et les signes qui conviennent pour les deux variables de DEFM.

Avec ce modèle à quatre variables, on estime $\hat{\sigma}_e^2 = 0.0479$ et $\hat{\sigma}_v^2 = 0.0001$. On remarque que pour une ZE "médiane", qui a $n_d = 130$ répondants, on a une variance d'échantillonnage mettons $\psi_d = 4 \cdot 10^{-4}$ (voir 4.2) alors que par ailleurs $\frac{\hat{\sigma}_e^2}{130} = 3.7 \cdot 10^{-4}$, ce qui est vraiment proche en ordre de grandeur de la variance d'échantillonnage. Par ailleurs, $\hat{\sigma}_v^2$ était de l'ordre de 1 dans le modèle de Fay et Herriot présenté en partie 4, mais les proportions de chômeurs avaient été multipliées préalablement par 100, ce qui fait qu'avec des taux exprimés en valeurs décimales, $\hat{\sigma}_v^2$ est de l'ordre de 10^{-4} - exactement ce que l'on retrouve ici. Ce fort parallélisme des estimations entre les deux modèles ne constitue évidemment pas une "preuve" à proprement parler, mais il est néanmoins très rassurant. Nous considérons qu'il valide les aspects algorithmiques et plus généralement les

[27] En fait, c'est la variable des inscriptions DEFM toutes catégories qui ressort le mieux, mais la différence de qualité obtenue avec la notion la plus naturelle, à savoir les catégories 1,2,3 HAR est quasi négligeable. Nous avons donc *in fine* retenu le concept qui se limite aux catégories proches de la définition BIT.

traitements informatiques. Il valide également l'utilisation de l'Eblup_B lorsque la variable sous-jacente est qualitative, ce qui était loin d'être évident (on rappelle que le modèle sous-jacent est en fait un modèle de données individuelles fondamentalement réservé à l'estimation de moyennes de variables quantitatives).

Pour ce modèle EBLUP-B, le calcul de l'écart-type $\sqrt{\hat{E}\left(\hat{\bar{Y}}_d^H - \bar{Y}_d\right)^2}$ donne lieu à la distribution suivante :

Quantile		Ecart type
100%	**Max**	0.0103
99%		0.0101
95%		0.0099
90%		0.0097
75%	**Q3**	0.0095
50%	**Median**	**0.0090**
25%	**Q1**	0.0082
10%		0.0069
5%		0.0064
1%		0.0047
0%	**Min**	0.0043

Si on considère une ZE "médiane", deux fois l'écart-type représente près de 2 points de pourcentage : c'est très proche de ce que l'on trouvait avec la modélisation de Fay et Herriot. On rappelle néanmoins que l'aléa est ici un pur aléa de modèle - sans intervention aucune de l'aléa de sondage. L'écart-type (estimé) constitue, dans le cas présent, un bon indicateur pour effectuer un choix de modèle - partant évidemment du principe que l'on compare des modèles qui sont tous "exacts", c'est-à-dire dans lesquels les variables aléatoires résiduelles ont systématiquement des espérances nulles. Hélas, on ne dispose pas d'outils informatiques de type Stepwise pour optimiser le choix des variables explicatives en utilisant cet indicateur comme critère de qualité. En revanche, parmi un ensemble de modèles possibles, si cet ensemble est de taille modeste, on peut retenir le modèle qui fournit le plus petit écart-type.

Il est intéressant d'estimer aussi les erreurs des seules ZE pour lesquelles on a utilisé un estimateur synthétique (qui n'est donc pas l'EBLUP-B) : on constate que leurs incertitudes apparaissent très peu dispersées, et numériquement comparables à celles du haut de la distribution des EBLUP-B.

Quantile		Standard Error (synthetic)
100%	**Max**	0.0117
99%		0.0111
95%		0.0108
90%		0.0107
75%	**Q3**	0.0105
50%	**Median**	**0.0104**
25%	**Q1**	0.0103
10%		0.0102
5%		0.0102
1%		0.0102
0%	**Min**	0.0102

Les tableaux ci-dessous fournissent les distributions d'erreurs EBLUP-B obtenues avec quatre modèles respectifs : en sus de la constante, toujours présente, on utilise les deux variables explicatives "Recherche d'emploi" et "Ne pas vivre en couple" (M1), ou la variable "Recherche d'emploi" (M2), ou la variable "Inscription à Pole Emploi" (M3), ou enfin la variable "Déclaration spontanée de chômage au recensement" (M4). Ces variables proviennent de la base de données par ZE présentée au 1.2.4. De ce point de vue, les possibilités de choix de variables sont considérables. Néanmoins, l'outil informatique offert par l'ONS estime les coefficients de régression à partir des données de l'échantillon (expression $\hat{B}_3$ présentée en 7.1), ce qui impose la présence des variables auxiliaires dans le fichier échantillon. C'est une assez forte contrainte, d'origine purement pratique, que l'on aurait certes pu contourner avec une autre programmation. Cela fait resurgir en substance les risques dus à l'hétérogénéité de variables entre sources mentionnés en partie 6.2. Néanmoins, l'impact numérique de cette hétérogénéité apparait très limité, pour ne pas dire insignifiant. Preuve en est l'analogie assez généralisée avec les résultats de la partie 4 et la valeur très acceptable de l'estimation nationale. Dans ces conditions, il n'y a pas à s'inquiéter particulièrement d'un risque de dérive dans l'estimation des erreurs.

Globalement, on constate que les quatre tableaux sont très proches. On voit aussi que l'étendue des erreurs apparait relativement faible : entre la plus petite erreur et la plus grande erreur, il n'y a jamais qu'un rapport de 1 à 2.5 environ. Par comparaison, on va trouver des rapports considérablement plus grands avec les estimateurs directs !

Modèle M1		
Quantile		Standard Error
100%	**Max**	0.0099
99%		0.0098
95%		0.0097
90%		0.0096
75%	**Q3**	0.0093
50%	**Median**	**0.0089**
25%	**Q1**	0.0081
10%		0.0069
5%		0.0063
1%		0.0047
0%	**Min**	0.0042

Modèle M2		
Quantile		Standard Error
100%	**Max**	0.0098
99%		0.0097
95%		0.0096
90%		0.0095
75%	**Q3**	0.0093
50%	**Median**	**0.0088**
25%	**Q1**	0.0081
10%		0.0069
5%		0.0063
1%		0.0047
0%	**Min**	0.0042

Modèle M3		
Quantile		Standard Error
100%	**Max**	0.0112
99%		0.0111
95%		0.0110
90%		0.0108
75%	**Q3**	0.0105
50%	**Median**	**0.0098**
25%	**Q1**	0.0089
10%		0.0073
5%		0.0066
1%		0.0048
0%	**Min**	0.0043

Modèle M4		
Quantile		Standard Error
100%	**Max**	0.0100
99%		0.0100
95%		0.0098
90%		0.0097
75%	**Q3**	0.0095
50%	**Median**	**0.0090**
25%	**Q1**	0.0082
10%		0.0069
5%		0.0064
1%		0.0047
0%	**Min**	0.0042

Pour les cinq modèles alternatifs évalués, il n'y a aucune évolution perceptible de l'estimateur $\hat{\sigma}_e^2$. En revanche, il y a un peu plus de sensibilité (qui reste néanmoins bien modeste…) pour l'estimation $\hat{\sigma}_v^2$.

	Modèle complet	M1	M2	M3	M4
$\hat{\sigma}_e^2$	0.0479	0.0479	0.0479	0.0479	0.0479
$\hat{\sigma}_v^2$	$1{,}025 \cdot 10^{-4}$	$1{,}022 \cdot 10^{-4}$	**$1{,}012 \cdot 10^{-4}$**	$1{,}036 \cdot 10^{-4}$	$1{,}07 \cdot 10^{-4}$

On retient désormais le modèle M2. Dans environ 80 % des ZE, l'incertitude - définie comme deux fois l'écart-type - est de l'ordre de 1,5 à 2 points de pourcentage. C'est une marge d'erreur qui peut sembler en soi assez conséquente, mais c'est à comparer avec l'estimateur direct - dont l'écart-type médian est de l'ordre de 2 points de pourcentage (cf. partie 4). L'incertitude associée au modèle EBLUP-B est tout à fait cohérente avec celle que l'on trouve en utilisant le modèle de Fay et Herriot (cf. partie 4). La meilleure précision est celle de la ZE de Paris (la plus grosse ZE), avec un écart-type de 0,4 points de pourcentage - à rapprocher de l'écart-type estimé à 0.6 points pour l'estimateur direct.

Lorsqu'on calcule le coefficient γ_d , propre à chaque ZE (cf. 7.1), on obtient la distribution suivante, laquelle s'avère comparable à celle du coefficient similaire calculé avec le modèle de Fay et Herriot.

Quantile		Gamma
100%	**Max**	0.83
99%		0.78
95%		0.61
90%		0.54
75%	**Q3**	0.36
50%	**Median**	**0.24**
25%	**Q1**	0.17
10%		0.12
5%		0.11
1%		0.10
0%	**Min**	0.10

A noter que la distribution de la variable gamma est quasiment la même si on ajuste le modèle complet (4 variables).

Lorsqu'on considère l'intégralité des 348 ZE de France métropolitaine, et que l'on s'intéresse à l'estimation nationale obtenue à partir de différentes méthodes concurrentes, on trouve :

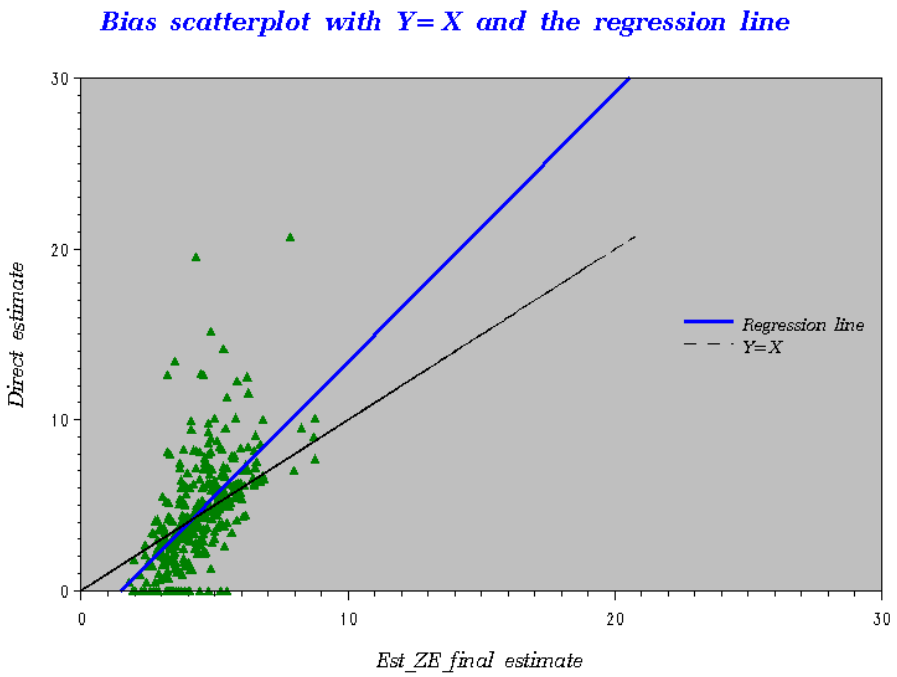
Estimateur national	Estimation totale
Eblup_B (ou synthétique)	2 441 000
Fay et Herriot (ou synthétique)	2 432 000
Enquête Emploi	2 436 000
Méthodologie actuelle Insee	2 408 000

L'effectif national Eblup_B est donc satisfaisant, en tout cas il l'est tout autant que ceux qui relèvent des deux approches concurrentes. Si on compare à la méthode actuellement utilisée par l'Insee, cette méthode Eblup_B modifie sensiblement certaines estimations. Voici la distribution de l'écart relatif :

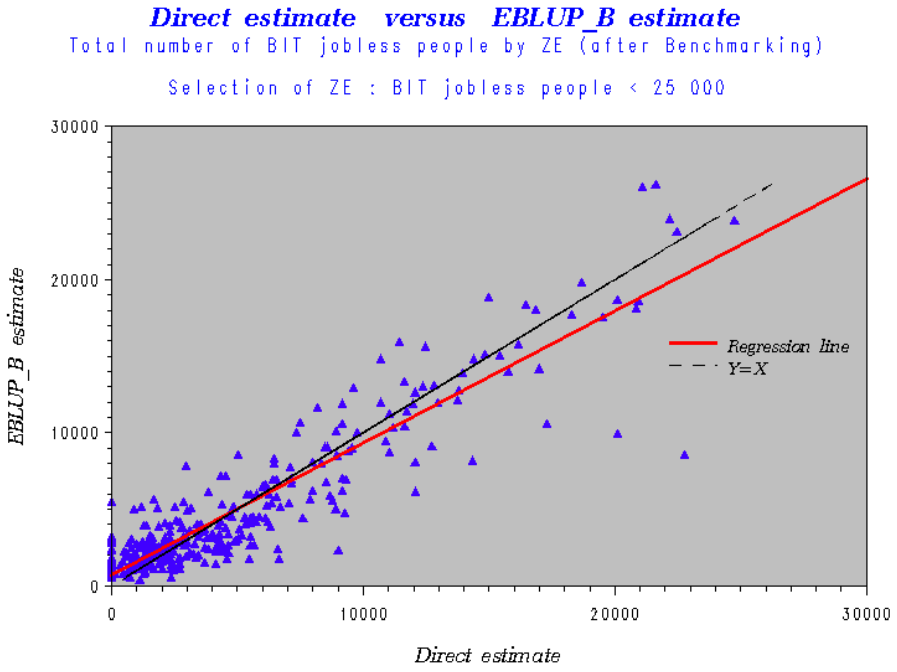
Quantile		Ecart relatif
100%	**Max**	54.4
99%		31.0
95%		20.4
90%		14.9
75%	**Q3**	3.0
50%	**Median**	**-6.5**
25%	**Q1**	-14.2
10%		-21.6
5%		-26.2
1%		-32.1
0%	**Min**	-43.8

L'écart relatif est défini par $100 \cdot \frac{\hat{\bar{Y}}_{ZE,EBLUP\_B} - \hat{\bar{Y}}_{ZE,INSEE}}{\hat{\bar{Y}}_{ZE,INSEE}}$.

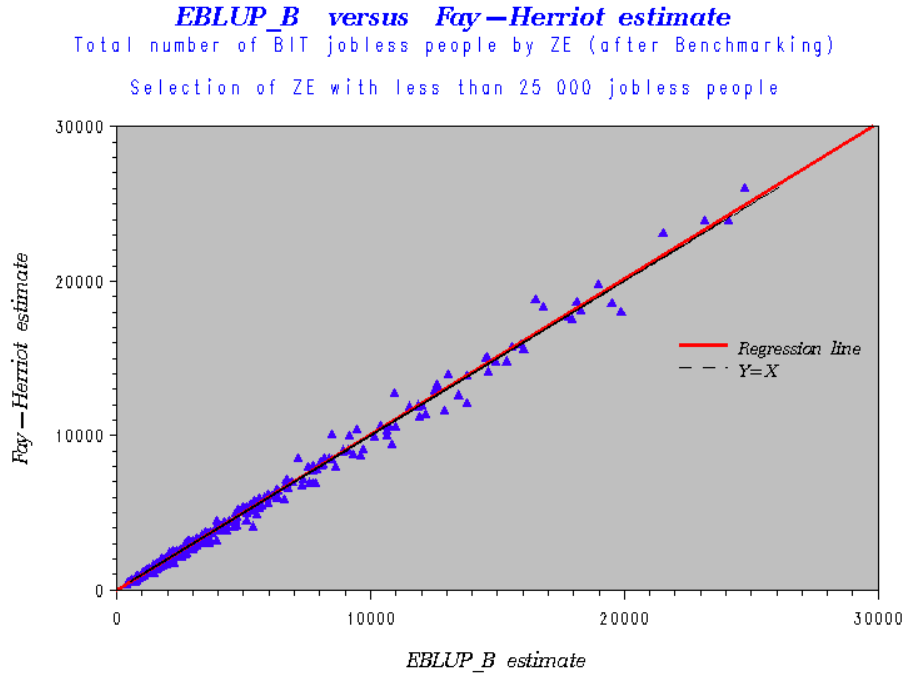
Du point de vue de l'aléa de sondage seulement, on a une présomption de biais si on examine le graphique suivant. L'effet shrinkage s'y distingue assez clairement. Cette intuition est confirmée par un test classique qui montre que la droite $Y = X$ est significativement différente de la droite de régression.



Nous avons *in fine* effectué un benchmarking pour nous caler sur l'estimation directe issue de l'enquête Emploi, à savoir 2 416 000 chômeurs. Le graphique suivant, limité aux ZE ayant moins de 25 000 chômeurs estimés, permet d'apprécier le risque de biais, après benchmarking.



Par curiosité, nous avons voulu rapprocher les estimations locales obtenues par la méthode de Fay et Herriot et les estimations locales obtenues par Eblup-B. Le graphique suivant montre qu'il y a une très forte corrélation entre les deux estimations.



Conclusion

Le premier point de conclusion consiste à rappeler que l'objectif originel et primordial consiste à faire (sensiblement) "mieux" que l'estimateur direct, qui dans nombre de cas ne produit pas d'estimations acceptables. Il y a donc une satisfaction quasi mécanique à appliquer des méthodes plus robustes à l'échantillonnage que l'approche classique, qui doit nécessairement trouver des alternatives, même si elles ne sont pas parfaites, loin de là : la formule est certes provocatrice, mais il faut faire "mieux" avant de faire "bien". Comme on doit s'appuyer de toute façon sur des hypothèses (des "modèles", en terme savant), il faut limiter ses ambitions et se contenter de l'à peu près. La meilleure illustration de ce conseil en matière d'humilité consiste à regarder les normes que d'autres Instituts nationaux ont édictées en matière de diffusion de résultats locaux :

- l'ONS se donne un CV limite de 20% au-dessus duquel il n'y a pas publication locale;
- ISTAT considère qu'au niveau province (107 provinces, donc l'équivalent de nos départements), il faut un CV inférieur à 17,5% pour diffuser les moyennes annuelles et inférieures à 25% pour les moyennes trimestrielles;
- Statistique Canada distingue trois niveaux de CV : inférieur à 16,5%, aucun avertissement, entre 16,5% et 33,3%, mise en garde de l'utilisateur, supérieur à 33,3% diffusion déconseillée (signature d'un "avis de non-responsabilité" si l'utilisateur insiste et passe outre)[28].

Il s'agit là de normes auxquelles nous ne sommes pas habitués (cela étant, à l'Insee, actuellement il n'en existe pas de ce type…) et qui, dans notre culture tournée vers l'estimation nationale, apparaitront probablement laxistes dans bien des esprits.

Nous ne nous sommes certainement pas suffisamment préoccupés de la mesure des erreurs totales associées à chaque méthode, surtout pour les modèles non linéaires où, c'est un fait, aucune

[28] Avec ces normes, environ 75% de nos ZE se trouveraient en diffusion "directe" déconseillée.

estimation d'erreur n'a été produite. Concernant ces derniers, une première raison est la grande complexité de la technique et donc l'absence d'outils logiciels adaptés. Mais cela pourrait se contourner, avec du temps. Une seconde raison, moins matérielle, tient au fait que, finalement, la grandeur "clé" est le biais, plus que la variance. Cela découle de notre premier point : on sait pertinemment, en appliquant l'une quelconque de ces méthodes, que l'on va gagner considérablement en variance, et on parle bien ici de variance d'échantillonnage. Cela est intrinsèque au principe constituant le cœur de toutes ces méthodes, à savoir s'appuyer judicieusement sur de l'information disponible en dehors du domaine considéré[29], ce qui stabilise immanquablement toutes les estimations. L'enjeu essentiel se situe donc du coté du biais. Au demeurant, il faut absolument faire son deuil de l'estimation non-biaisée : de fait, il n'existe pas d'estimation "petits domaines" sans biais si on s'en tient à l'aléa d'échantillonnage classique. Cela signifie, et c'est la clé de l'affaire, que l'on doit en fait se satisfaire de toute méthode qui produit des estimateurs "peu biaisés".

Nous avons le sentiment qu'il y a, dans tous ces processus d'estimation, un mélange savant de théorie très compliquée - mais évidemment très rigoureuse sur le plan mathématique - et d'empirisme assez fort pour prendre des décisions, avec des zones d'ombre intrinsèques à la modélisation non linéaire. D'ailleurs, nous ne nous sommes pas risquer à "conseiller" ou "déconseiller" telle ou telle méthode, encore moins à les classer. Le départage est subtil, et il devrait se faire probablement sur la base d'un sentiment global qui associerait de multiples considérations - d'erreur globale certes quant on peut la mesurer, mais aussi de pertinence des hypothèses initiales, de disponibilité et de fiabilité des informations auxiliaires, de facilité de mise en œuvre - en particulier informatique - de communication même (car il faut ensuite expliquer tout cela à des utilisateurs non-spécialistes). Le choix d'une méthode "optimale" nécessiterait également que l'on s'assure de la stabilité des conclusions dans le temps (il faudrait donc reprendre toute cette étude sur plusieurs trimestres). Au demeurant, la question de la comparaison des méthodes est en soi assez troublante car sa légitimité ne nous paraît pas si claire : d'une part les différentes estimations ne s'appuient pas sur des aléas de même nature (dans ces conditions, compare-t-on des choses comparables ?), d'autre part elles dépendant toutes d'hypothèses initiales fortes (les "modèles") qui ne sont pas vraiment remises en cause dès lors que les résultats apparaissent acceptables. Juger de la pertinence préalable de ces hypothèses - si on le peut … - devrait donc faite partie intégrante de l'exercice de comparaison des méthodes.

La modélisation non linéaire paraît sensiblement plus complexe que la modélisation linéaire et notre sentiment personnel est que l'on y perd presque toute aptitude à utiliser l'intuition : il faut se laisser porter par les développements formels qui construisent des cathédrales et il faut avoir la foi pour y pénétrer même si l'ensemble apparaît très esthétique. Mais se sont là les limites habituelles de la modélisation sophistiquée …

Et puis, restons optimistes quant aux pistes d'amélioration de ces techniques "petits domaines". Elles sont presque innombrables, tant l'univers des possibilités est grand : il existe des théories qui font intervenir des corrélations spatiales (en la circonstance, cela aurait été très intéressant car en matière de chômage, l'effet de grappe géographique est évident), d'autres qui font intervenir des composantes temporelles en empilant différents trimestres pour renforcer encore la stabilité des estimations, d'autres utilisent des modèles optimisant la distribution des estimations, d'autre encore assurent des benchmarking à différents niveaux. Et il y a tout l'univers de l'estimation bayésienne …

Bibliographie

[1] Rao J.N.K., « Small Area Estimation», *Wiley*, 2003.
[2] McCulloch C.E., Searle S.R, Neuhaus J.M., « Generalized, Linear, and Mixed Models », *Wiley*, 2008

[29] Les anglo-saxons utilisent l'expression très imagée "to borrow strenght".