Estimations régionales et départementales du *Nombre de recours au généraliste* dans l'Enquête Décennale Santé 2002/2003

Marius STEFAN (mastefan@ulb.ac.be)[1]
Jean-Jacques Droesbeke (jjdroesb@ulb.ac.be)[2]
Jean-Claude Deville (Jean-Claude.Deville@ensae.fr)[3]

1 Introduction

Dans cet article nous allons étudier la variable R02AM (*Nombre de recours au généraliste dans les derniers douze mois*) qui se trouve dans l'enquête santé réalisé en 2002/2003. Le but est d'obtenir des estimations régionales et départementales pour le nombre moyen de recours au généraliste. La variable R02AM sera traitée en détail, en montrant la méthodologie que nous avons utilisée. L'estimation «petits domaines» est basée sur un modèle qui en général utilise des variables auxiliaires. Dans la deuxième section nous avons réalisé une analyse exploratoire qui montre quelles sont les variables qui influencent le plus R02AM. Nous avons retenu ces variables que nous utilisons dans la troisième section pour construire un modèle pour R02AM. La construction du modèle est progressive et adaptée aux caractéristiques des données. On obtient comme résultats trois modèles, dont le dernier sera retenu pour faire des estimations. Ce modèle permet seulement l'obtention d'estimations régionales et non pas départementales. Dans la section quatre on présente les principes d'estimation de la statistiques bayésienne et on dérive les formules théoriques des estimations régionales et de leur précisions. On voit que la seule information hors enquête dont on a besoin sont les vrais effectifs des cellules déterminées par les variables auxiliaires retenues. Dans la section cinq on compare les trois modèles du point de vue de leurs ajustements aux données, on sélectionne le troisième modèle et on décide qu'il est suffisamment bien adapté aux données pour pouvoir l'utiliser dans l'estimation. Dans la section six on utilise les formules de la section quatre pour estimer les moyennes régionales ainsi que leurs précisions. On le fait dans le cas des deux échantillons (avec et sans extension) et on compare nos résultats avec les estimations produites par l'INSEE pour les cinq régions à extension et la France Métropolitaine. Comme à l'époque on ne disposait pas des vrais effectifs des cellules, on utilise des effectifs estimés par la somme des poids de sondage. Dans la section sept on modifie le modèle trois en remplaçant l'effet région par un effet département. Le nouveau modèle, appelé modèle quatre, permet le calcul d'estimations départementales, mais aussi régionales et pour la France. Il est aussi bien adapté aux données que le modèle trois, et les estimations régionales et pour la France qu'il fournit sont identiques à celles basées sur le modèle trois. Dans la dernière section on recalcule toutes les estimations mais cette fois-ci en

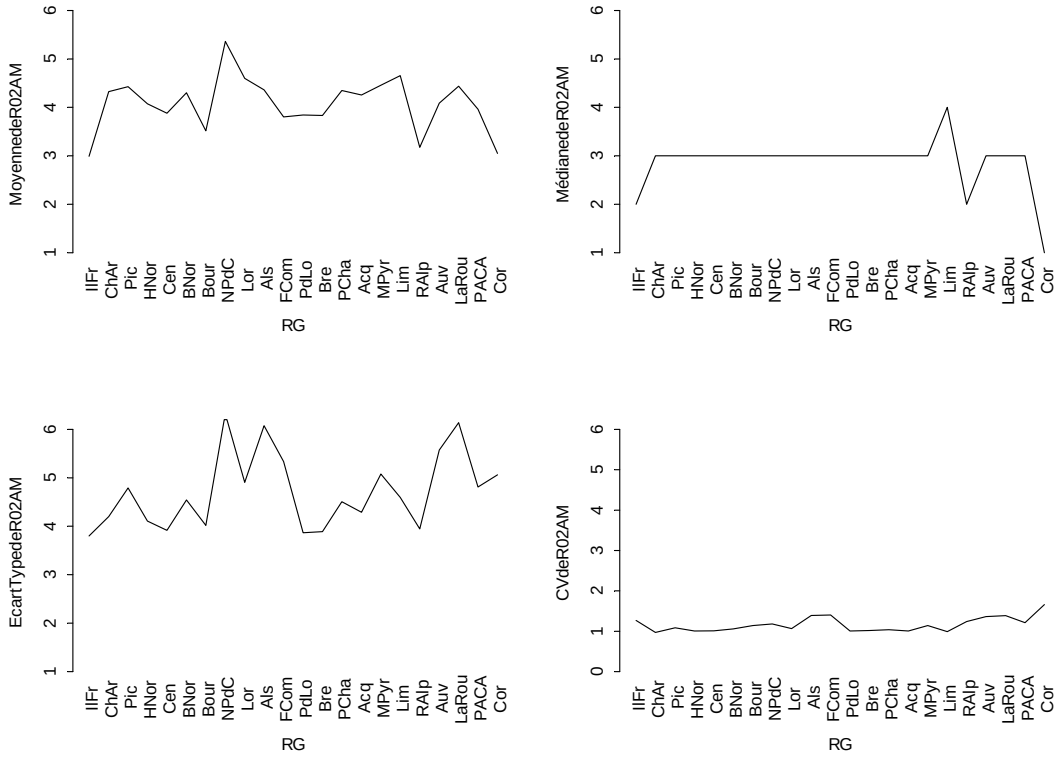
[1] Université Libre de Bruxelles (ULB) et Université Polytechnique de Bucarest (UPB)
[2] Université Libre de Bruxelles (ULB)
[3] Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE) et Institut National de Statistique et d'Etudes Economiques (INSEE)

utilisant les vrais effectifs des cellules fournis par l'INSEE, ensuite on les compare à celles utilisant les effectifs estimés.

2 Analyse exploratoire de la variable R02AM

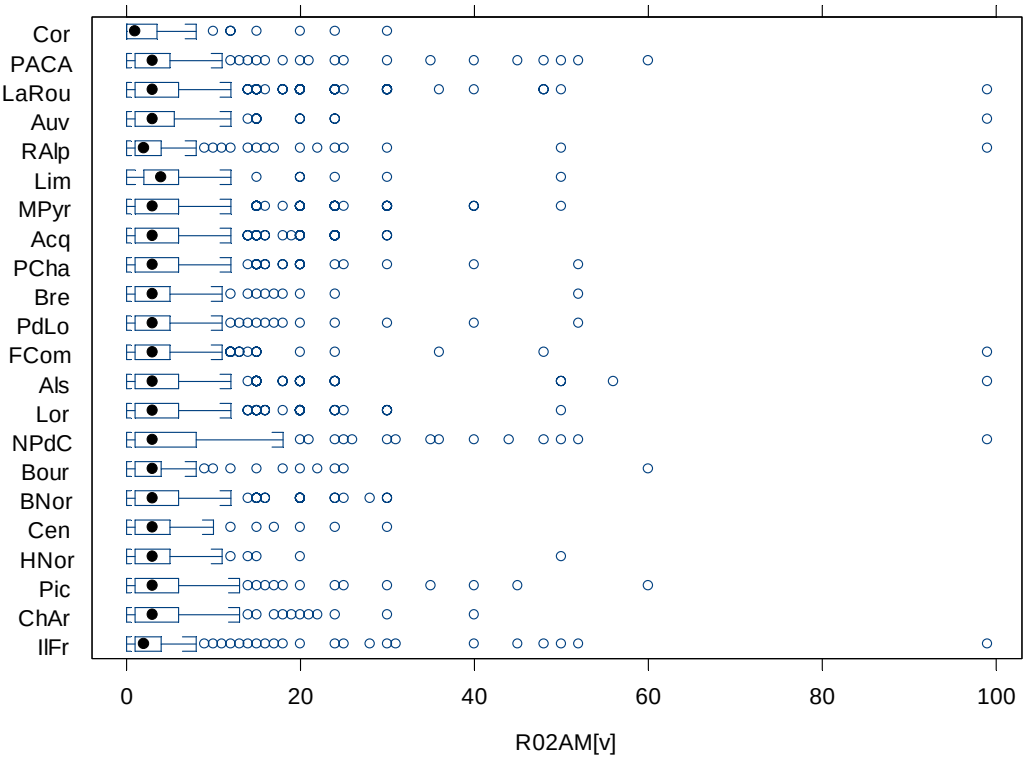
Dans un premier temps, nous avons réalisé une analyse exploratoire à partir de quelques variables qui se trouvent dans la base de données et donc nous pensons que le comportement de R02AM en dépend. Nous avons retenu les variables qui ont l'influence la plus grande sur R02AM: région du ménage (RG), la strate (STRATE), le sexe (SEXE) et l'âge (AGE). Il faut souligner que les variables explicatives qu'on peut introduire dans l'analyse doivent non seulement être liées à la variable d'étude, mais − comme nous allons le voir plus bas − on doit disposer de leurs valeurs pour tous les individus de la population. Etant donnée la façon dont nous avons utilisé les quatre variables ci-dessus dans notre analyse, cette contrainte nécessite de connaître les effectifs de chaque cellule $RG \times STRATE \times SEXE \times AGE$. Pour illustrer l'impact des variables RG, STRATE, SEXE et AGE sur la variable R02AM, on a choisi d'étudier quatre paramètres de R02AM : la moyenne, la médiane, l'écart-type et le coefficient de variation. La figure 1 représente les quatre paramètres calculés par région. On voit l'influence de RG sur tous ces paramètres : les graphiques sont irréguliers avec des hauts et des bas mettant en évidence les régions où on va plus souvent consulter le généraliste et celles où on y va moins.

Figure 1: *Paramètres de R02AM calculés par région*



Dans la figure 2, nous avons représenté les boîtes-à-moustaches de R02AM par région. Elles ne sont pas homogènes, confirmant un effet région dont il faut tenir compte quand on construit le modèle.

Figure 2: *Boîtes à moustaches de R02AM par région*



On a refait la même analyse pour la variable STRATE. STRATE est une variable qualitative avec cinq valeurs selon la taille de la commune. Les résultats sont présentés dans les figures 3 et 4. Les différences sont moins importantes que dans le cas des régions. C'est seulement la strate quatre (unité urbaine de Paris) qui présente une différence importante par rapport aux autres strates.

Figure 3 : *Paramètres de R02AM calculés par strate*

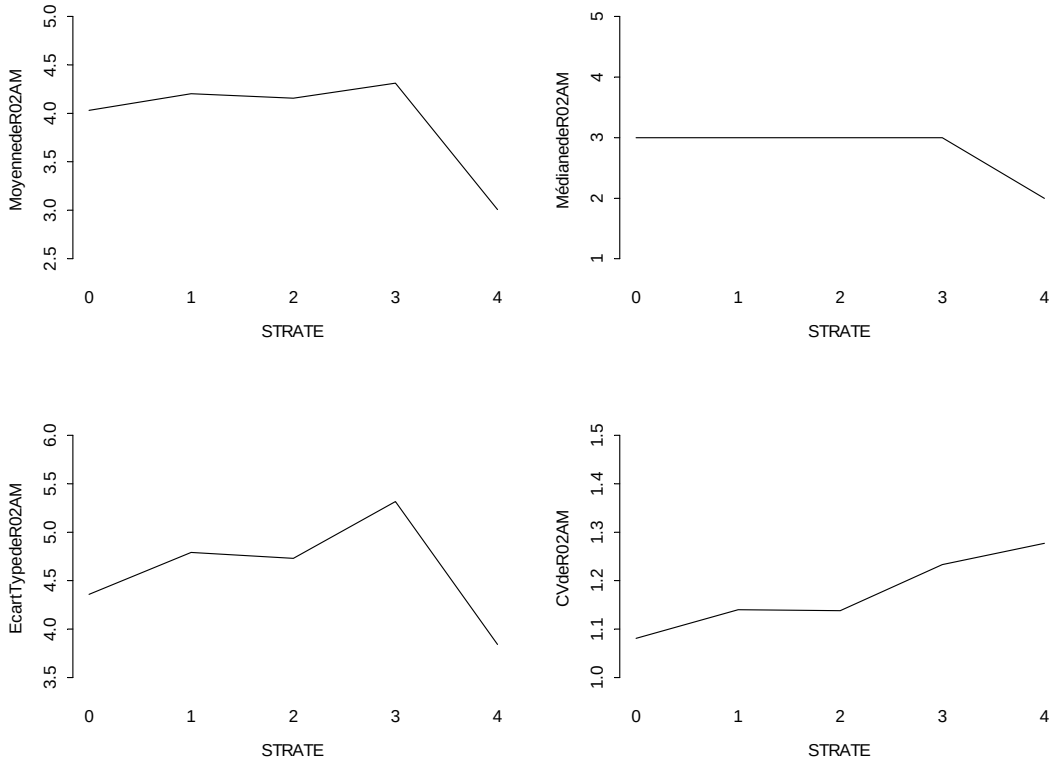
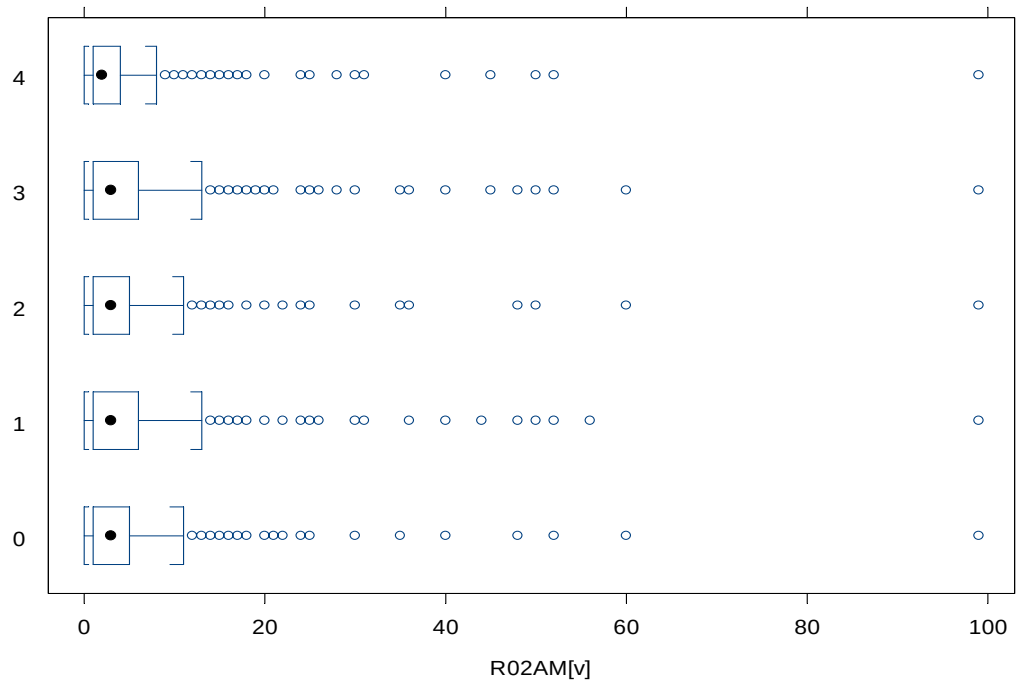
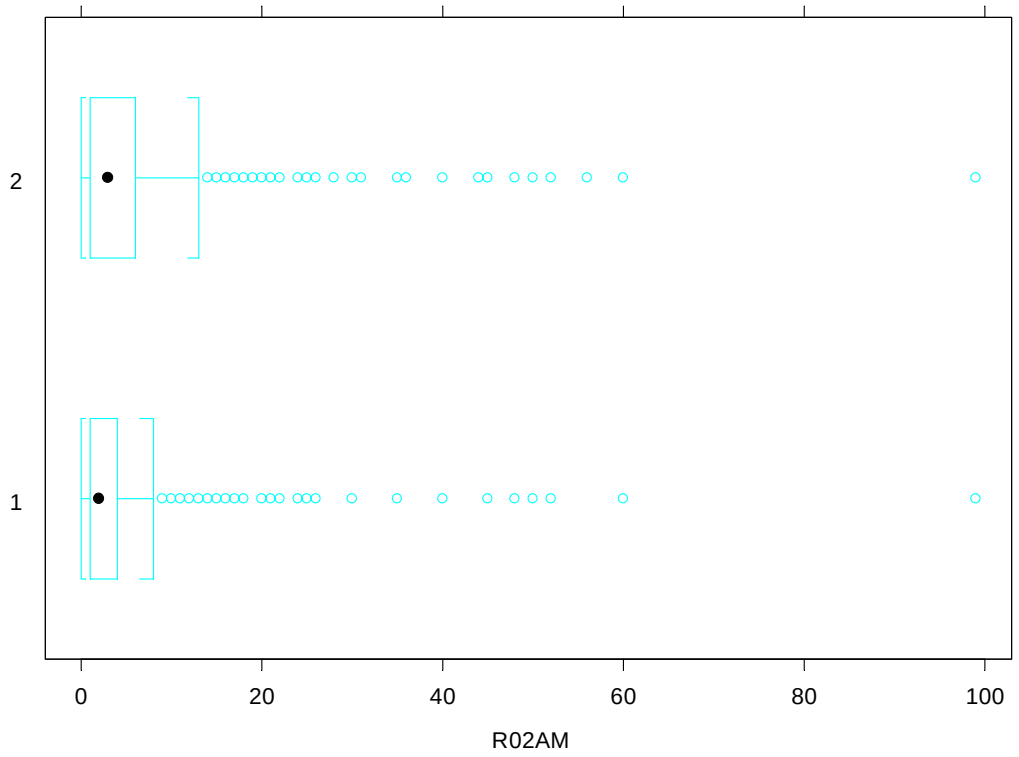


Figure 4: *Boîtes à moustaches de R02AM par strate*



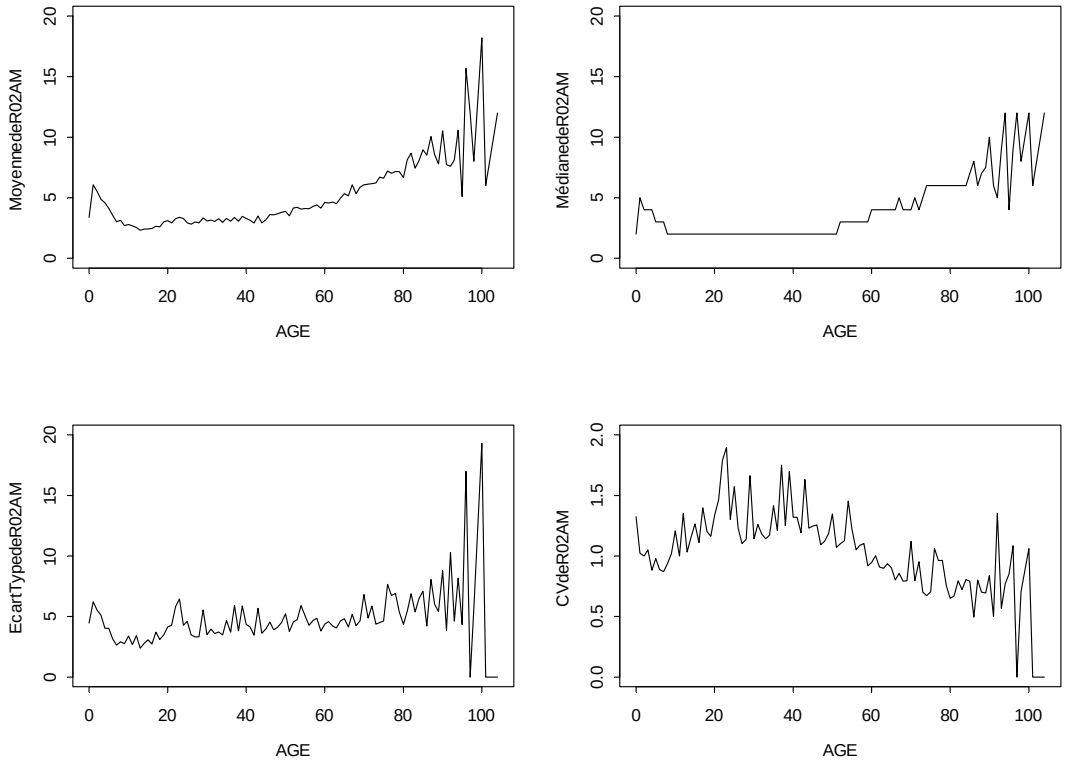
Nous avons également construit les boîtes –à- moustaches de R02AM pour les hommes d'une part (SEXE=1) et pour les femmes d'autre part (SEXE=2) (voir figure 5). Les boîtes sont différentes mettant en évidence un effet SEXE dont il faut tenir compte.

Figure 5: *Boîtes à moustaches de R02AM par SEXE (1=Hommes, 2=Femmes)*



En ce qui concerne la variable AGE, nous avons réalisé les graphiques de la figure 6. Sur l'axe horizontal se trouvent les valeurs de AGE observées dans la base de données. Sur l'axe vertical, nous avons reporté, pour chacune de ces valeurs, la moyenne, l'écart-type, la médiane et le coefficient de variation de R02AM. Les quatre graphiques permettent de visualiser un effet AGE important. Pour ce qui est de la moyenne de R02AM, on observe d'abord un pic correspondant aux âges 0-1, ensuite une diminution jusqu'aux alentours 15 ans, ensuite, jusqu'aux âges les plus élevés, une augmentation permanente. L'écart-type a aussi tendance à augmenter avec l'AGE, mais les fortes fluctuations observables au-delà de 85 ans sont aussi dues au fait que dans cette plage-là de valeurs de AGE, on a de moins en moins d'observations.

Figure 6: *Paramètres de R02AM en fonction de AGE*

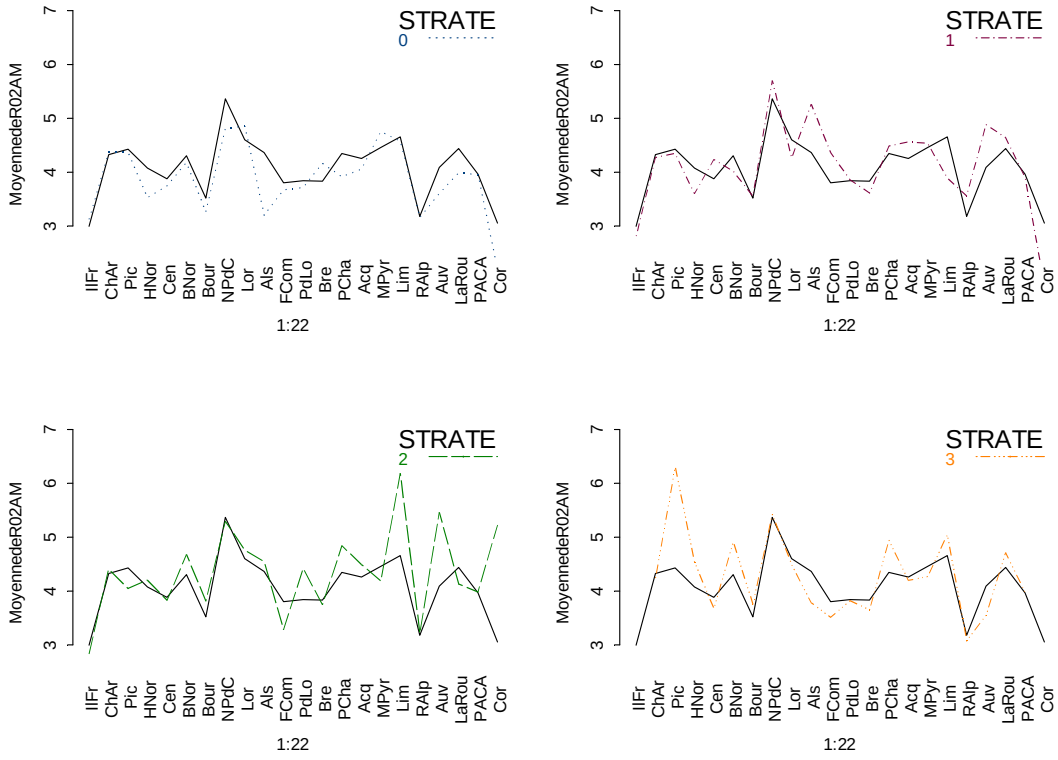


Etant données les observations faites sur base de la figure 6, nous avons transformé la variable AGE en une variable catégorielle en prenant 8 tranches d'âges, la plupart de longueur égale à 10: [0,1], [2,12], [13,23],…[56,67], [68,104]. Nous avons noté cette nouvelle variable AGE10. Le premier intervalle est de longueur égale à 1, les suivants de longueur égale à 10 et le dernier comprend tous les individus d'âge supérieur à 68. De cette manière nous pensons pouvoir capter fidèlement la dépendance entre AGE et R02AM. Une analyse plus fine que nous n'avons pas encore réalisée consisterait à prendre des tranches d'âges de longueur égale à cinq ou plus petite pour voir si ceci aurait un effet au niveau des estimations (à faire).

Nous nous sommes ensuite intéressés aux interactions entre ces variables. Pour voir s'il y a interaction entre RG et STRATE, nous avons réalisé la figure 7 (nous avons inséré seulement les graphiques relatifs à la moyenne de R02AM, mais pour les autres paramètres les conclusions sont similaires). Nous avons calculé et représenté la valeur de la moyenne de R02AM par région et par strate. Pour mieux visualiser une meilleur interaction, nous avons superposé au graphique de chaque strate le premier graphique de la figure 1, à savoir les moyennes de R02AM calculées par région seulement (ligne continue). En absence d'interaction on devrait avoir des lignes identiques ou parallèles. On peut observer que les interactions de STRATE 0 et STRATE 1 avec les régions sont presque nulles. STRATE 2 et STRATE 3 semblent chacune présenter une interaction avec une seule région sur les 22, alors que dans le cas de la STRATE 4, on ne peut pas parler d'interaction étant donné que les

individus de cette strate se trouvent exclusivement dans la région Ile de France. Nous avons d'ailleurs estimé une version du modèle 3 (voir plus bas) avec interaction entre RG et STRATE et nous n'avons pas trouvé de différences importantes au niveau de l'ajustement du modèle par rapport à ce que nous fournit le modèle sans interaction.

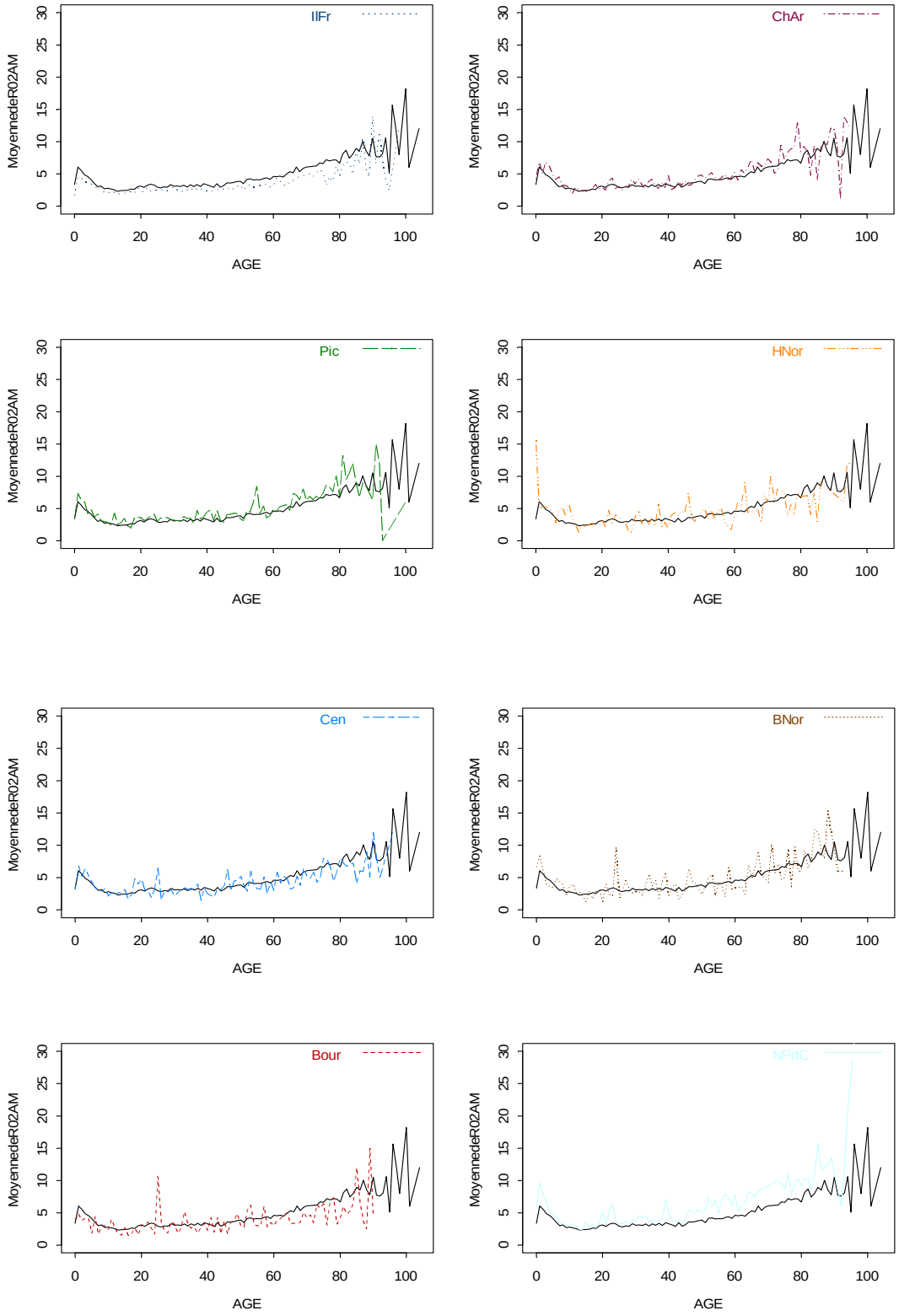
Figure 7:*Interaction région×strate*

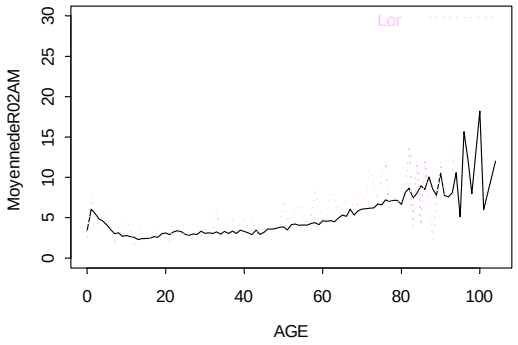


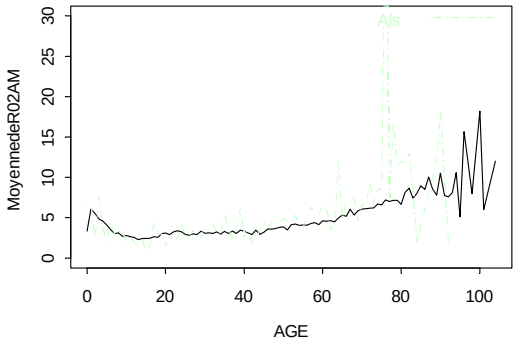
Nous avons ensuite étudié l'interaction entre RG et AGE. De nouveau, on présente ici seulement la moyenne de R02AM, ce qui donne lieu aux graphiques de la figure 8 : pour chaque région et pour chaque valeur observée de AGE, nous avons calculé la moyenne de R02AM de tous les individus correspondants et nous avons superposé (ligne continue) le premier graphique de la figure 6. Les interactions sont faibles et le modèle avec interactions n'a pas fourni d'estimations différentes du modèle sans interaction.

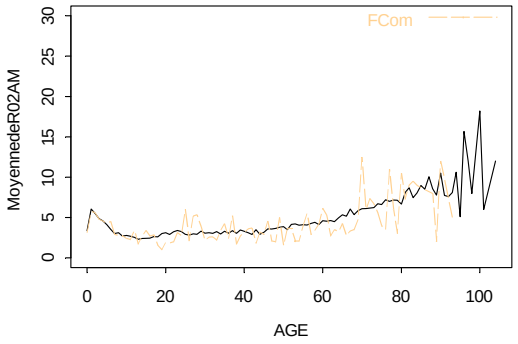
Les graphiques de la figure 9 sont réalisés selon le même principe et sont relatifs à l'interaction entre STRATE et AGE : ils semblent indiquer une absence d'interaction.

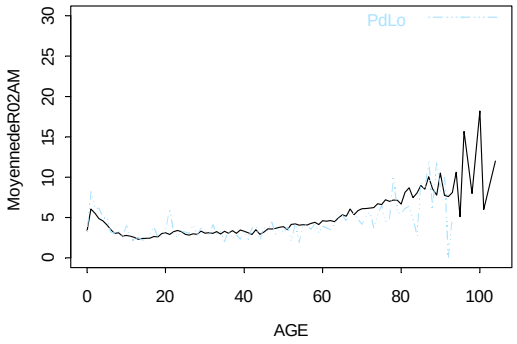
Figure 8: *Interaction région×âge*

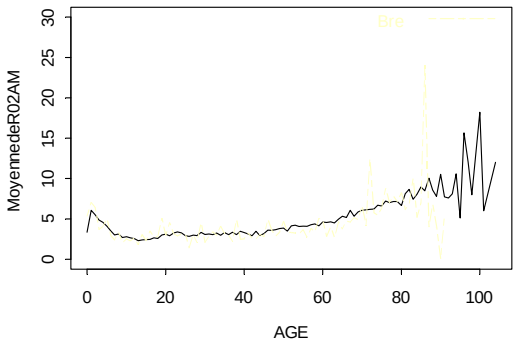


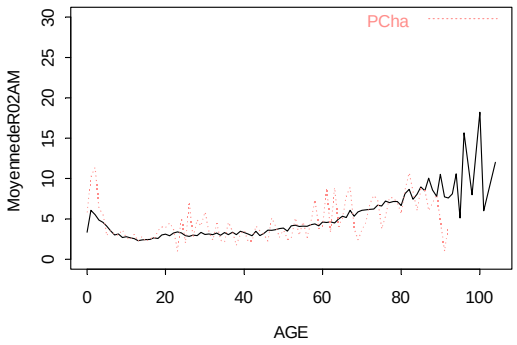

Lor
MoyennedeR02AM
AGE
0
5
10
15
20
25
30
0
20
40
60
80
100

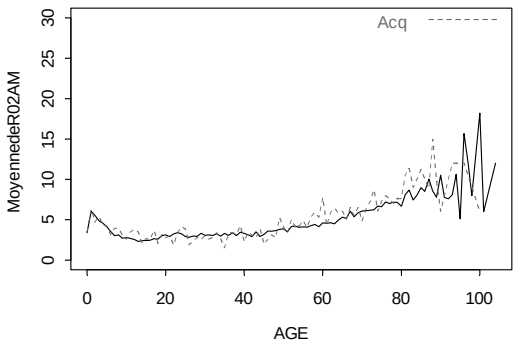

Als
MoyennedeR02AM
AGE
0
5
10
15
20
25
30
0
20
40
60
80
100

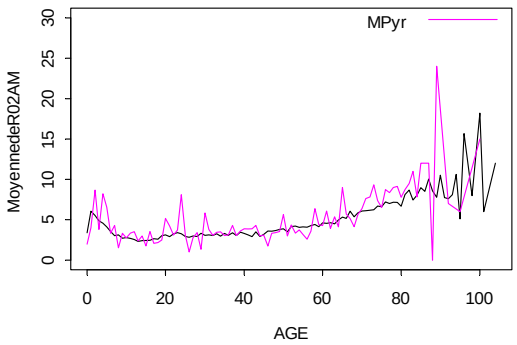

FCom
MoyennedeR02AM
AGE
0
5
10
15
20
25
30
0
20
40
60
80
100

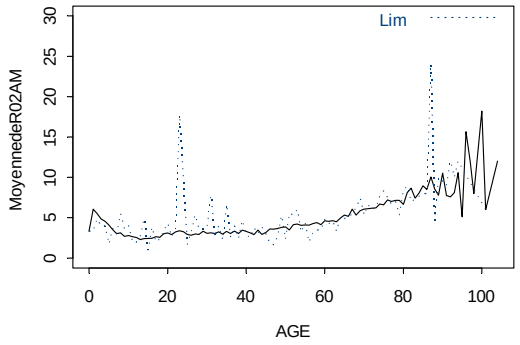

PdLo
MoyennedeR02AM
AGE
0
5
10
15
20
25
30
0
20
40
60
80
100

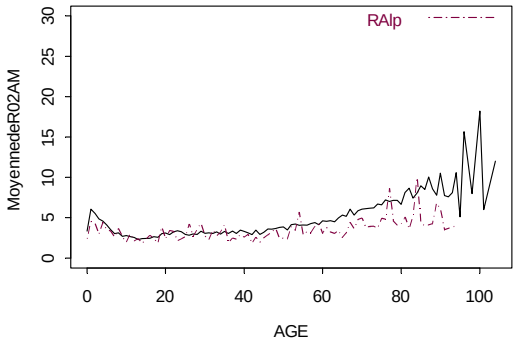

Bre
MoyennedeR02AM
AGE
0
5
10
15
20
25
30
0
20
40
60
80
100

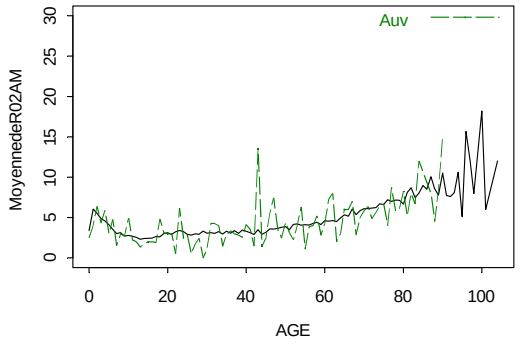

PCha
MoyennedeR02AM
AGE
0
5
10
15
20
25
30
0
20
40
60
80
100

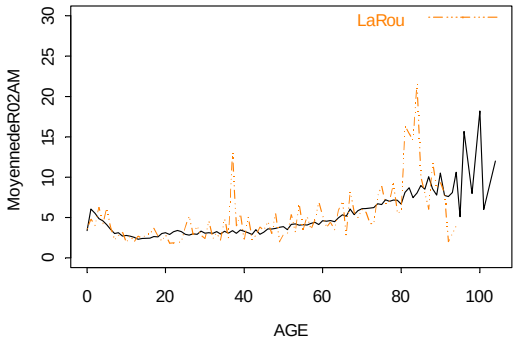

Acq
MoyennedeR02AM
AGE
0
5
10
15
20
25
30
0
20
40
60
80
100

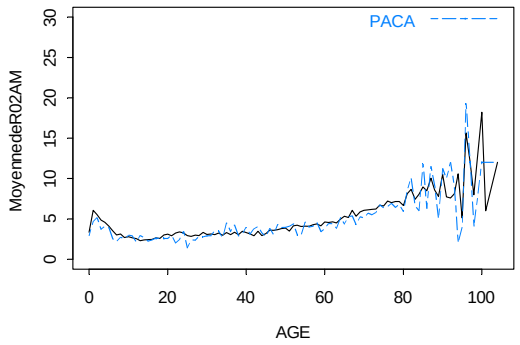

MPyr
MoyennedeR02AM
AGE
0
5
10
15
20
25
30
0
20
40
60
80
100


Lim
MoyennedeR02AM
AGE
0
5
10
15
20
25
30
0
20
40
60
80
100


RAlp
MoyennedeR02AM
AGE
0
5
10
15
20
25
30
0
20
40
60
80
100


Auv
MoyennedeR02AM
AGE
0
5
10
15
20
25
30
0
20
40
60
80
100


LaRou
MoyennedeR02AM
AGE
0
5
10
15
20
25
30
0
20
40
60
80
100


PACA
MoyennedeR02AM
AGE
0
5
10
15
20
25
30
0
20
40
60
80
100

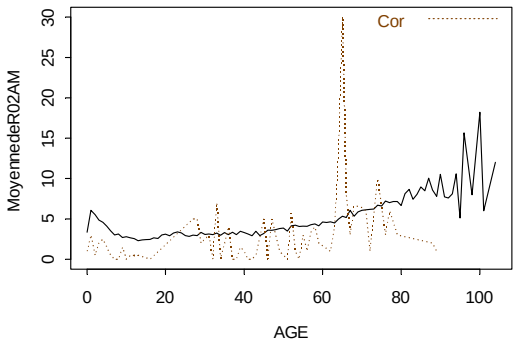
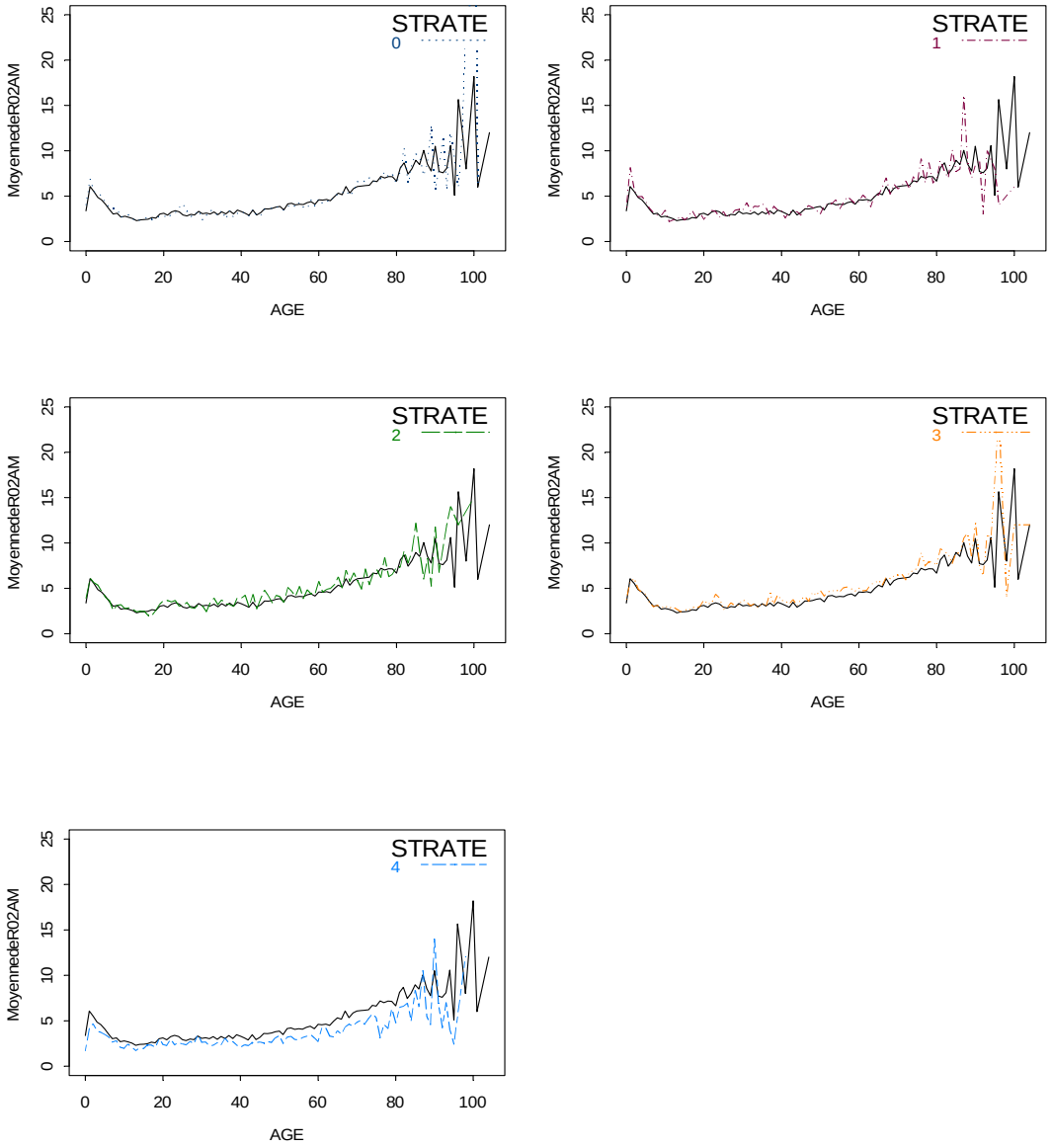

Cor
MoyennedeR02AM
AGE
0
5
10
15
20
25
30
0
20
40
60
80
100

Figure 9: *Interaction strate×âge*



Les graphiques suivants correspondent à la variable SEXE et à ses interactions avec les variables RG, STRATE et AGE. Les graphiques sont parallèles ou identiques: nous n'avons donc pas d'interaction entre SEXE et les trois autres variables.

Figure 10: *Interaction sexe×région*

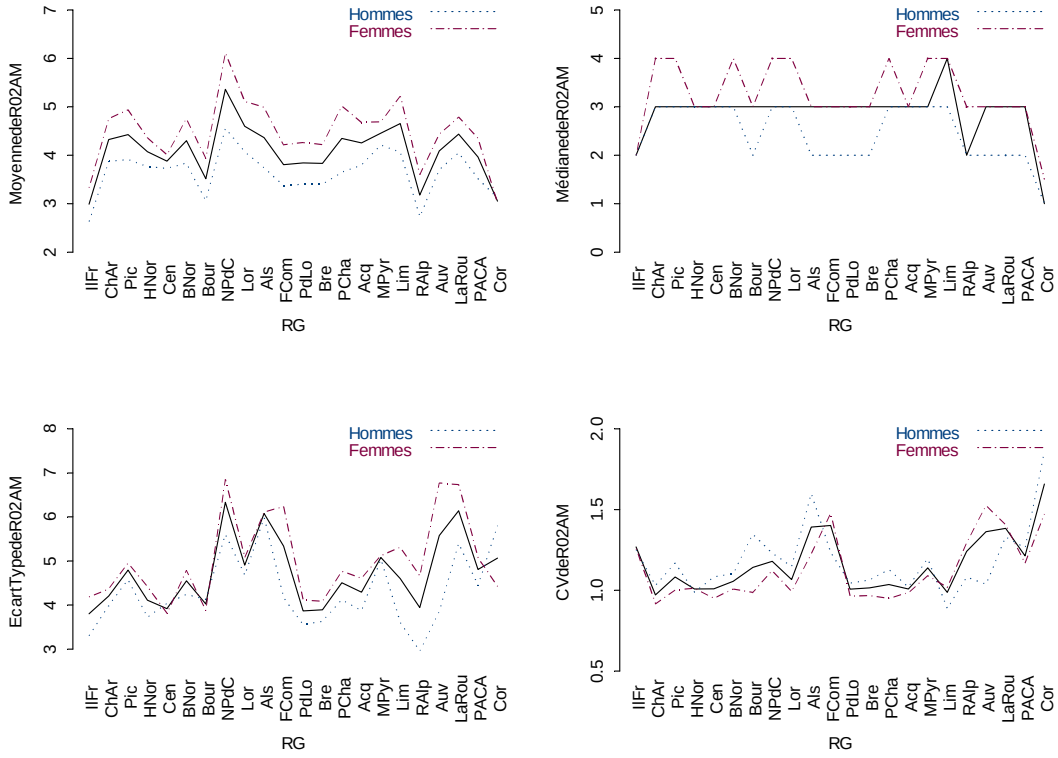


Figure 11: *Interaction sexe×strate*

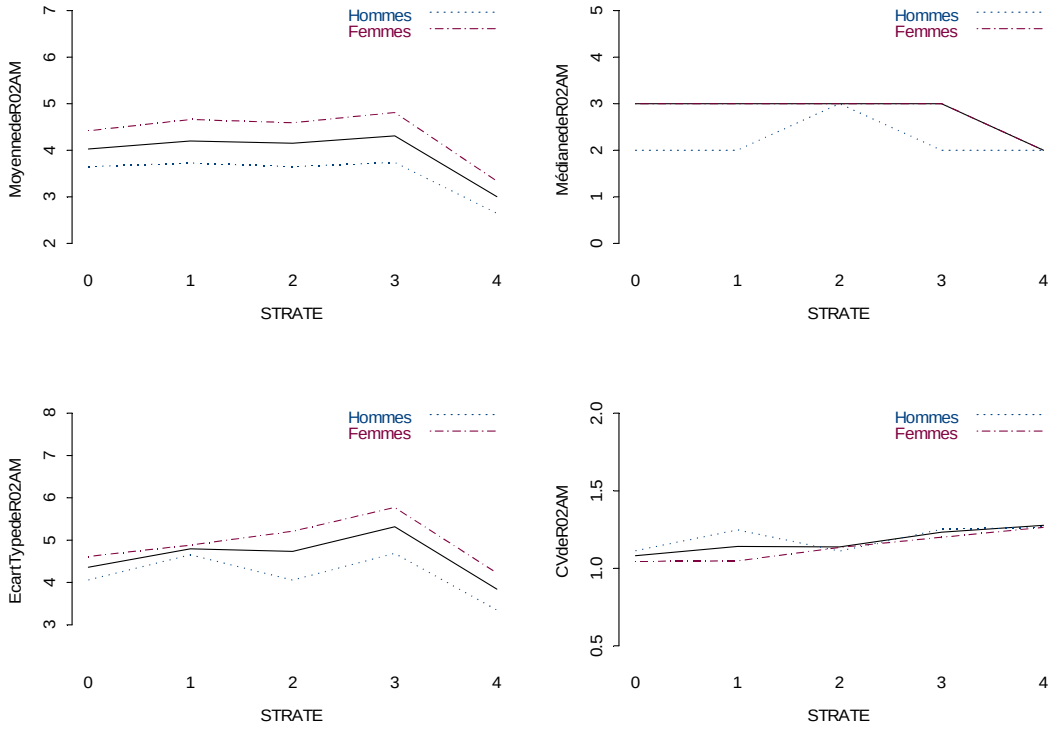
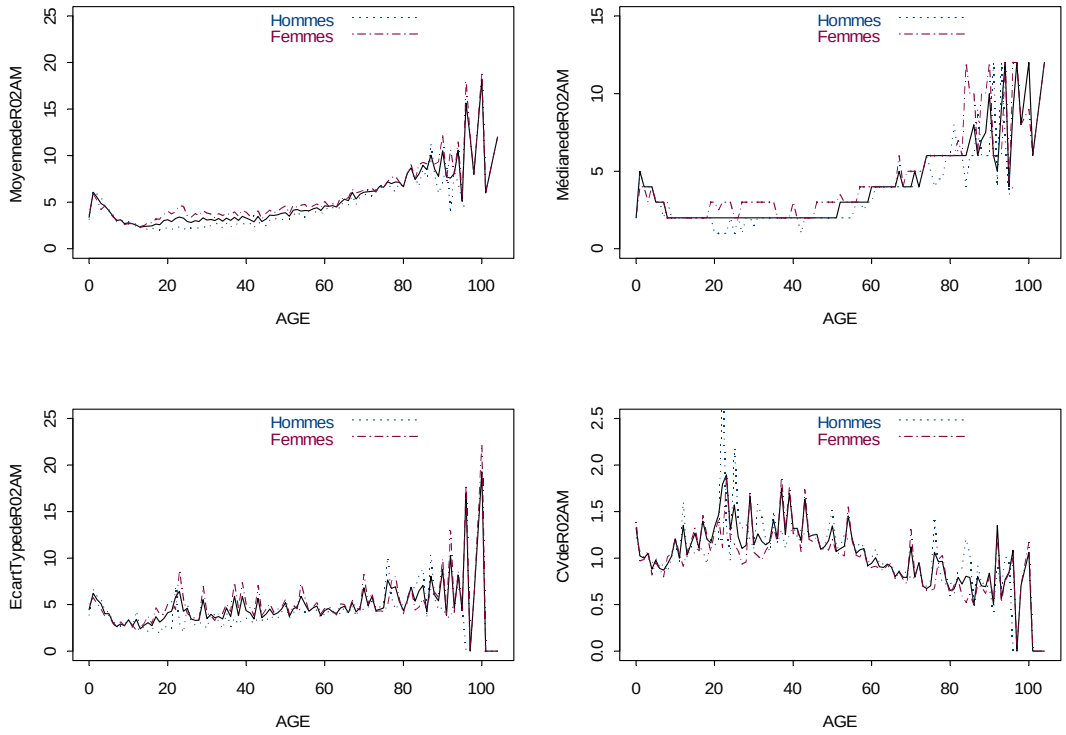


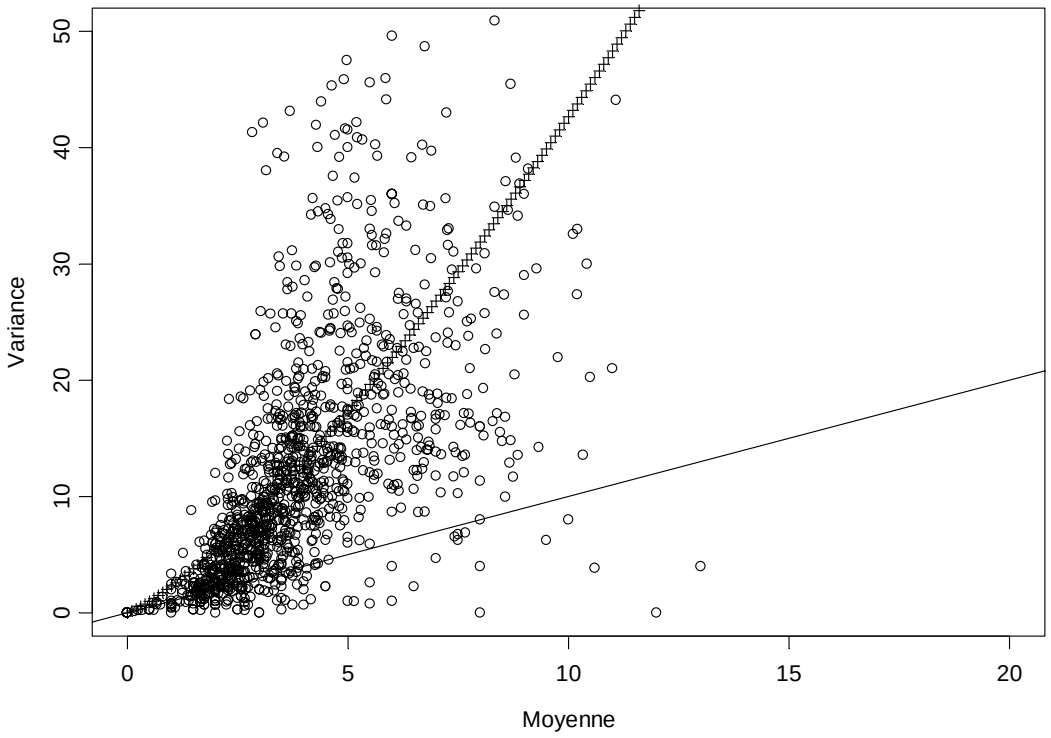
Figure 12 *Interaction sexe×âge*



3 Construction du modèle pour la variable R02AM

Etant donnée l'analyse exploratoire que nous avons réalisée, nous postulons que la valeur de la variable R02AM dépend essentiellement de la région, de la strate, du sexe et de l'âge de l'individu. Soit y_{ijskl} la valeur de R02AM pour l'individu *l* de la tranche d'âges *k*=1,…,8, de sexe *s*=1,2, se trouvant dans la strate *j*=1,…,5 et dans la région *i*=1,…,22. Les valeurs de y_{ijskl} pour tout *l* appartenant à la cellule $i \times j \times s \times k$ devraient tourner autour d'une même moyenne étant donné notre hypothèse. R02AM est une variable de comptage; il semble donc naturel de la modéliser à l'aide d'une distribution de Poisson. Cette distribution suppose que la moyenne est égale à la variance. Nous nous proposons d'investiguer si R02AM vérifie cette propriété. Dans ce but, pour chaque cellule $i \times j \times s \times k$, nous avons calculé la moyenne et la variance des y_{ijskl} et nous avons représenté graphiquement la relation entre les deux dans la figure 13:

Figure 13: *Relation entre moyenne et variance des cellules* $i \times j \times s \times k$



La droite représente la première bissectrice. On voit donc que pour la plupart des points (c'est-à-dire la plupart des cellules $i \times j \times s \times k$) la variance est plus grande que la moyenne : on est en situation de surdispersion. L'autre courbe a pour équation $f(x) = x + 1.3x^{1.5}$, ce qui semble indiquer une relation de ce type entre la variance et la moyenne de la variable R02AM dans les différentes cellules $i \times j \times s \times k$. Il apparaît donc clairement que la loi de Poisson classique n'est pas appropriée pour modéliser R02AM.

Nous avons néanmoins considéré ci-après un modèle fondé sur une loi de Poisson de paramètre ν_{ijsk} non aléatoire (le modèle 1 plus bas) afin de voir à quoi on aboutit lorsque on ignore la surdispersion.

Une des méthodes proposées dans la littérature (voir par exemple Peter Congdon (2005)) pour modéliser une variable de comptage pour laquelle la variance plus grande que la moyenne est de considérer que le paramètre de la loi de Poisson est à son tour aléatoire distribué selon une loi Gamma. Nous adoptons ici cette démarche parce qu'elle nous fournit une relation du type $\text{var} = \text{moy} + alpha \times \text{moy}^{1+kappa}$ comme nous allons le voir ci-dessous. Clairement, la première ligne du modèle est :

$$y_{ijskl} \mid \nu_{ijskl} \underset{ind}{\sim} \text{Poisson}(\nu_{ijskl}) \quad (1)$$

Nous avons pris ultérieurement une version dans laquelle le paramètre aléatoire de la loi de Poisson n'est pas ν_{ijskl} mais ν_{ijsk} (il s'agit du modèle 2 ci-dessous), mais nous avons préféré envisager en premier lieu la situation (1), dans l'idée que nous laissions une liberté plus grande au modèle si l'on associait à chaque observation y_{ijskl} son propre paramètre. Nous modélisons ensuite ν_{ijskl} à l'aide d'une loi Gamma comme suit:

$$\nu_{ijskl} \mid \mu_{ijsk}, alpha, kappa \underset{ind}{\sim} \text{Gamma}\left(\frac{\mu_{ijsk}^{1-kappa}}{alpha}, \frac{\mu_{ijsk}^{-kappa}}{alpha}\right) \quad (2)$$

Le choix des deux paramètres de la loi Gamma est justifié par le raisonnement suivant :

$$E(y_{ijskl}) = E_{\nu_{ijskl}} E(y_{ijskl} \mid \nu_{ijskl}) = E(\nu_{ijskl}) = \mu_{ijsk}$$
$$V(y_{ijskl}) = E_{\nu_{ijskl}} V(y_{ijskl} \mid \nu_{ijskl}) + V_{\nu_{ijskl}} E(y_{ijskl} \mid \nu_{ijskl}) = E(\nu_{ijskl}) + V(\nu_{ijskl}) =$$
$$= \mu_{ijsk} + alpha \times \mu_{ijsk}^{1+kappa}$$

On voit donc que par le choix d'une loi Gamma comme en (2), la relation entre la moyenne marginale et la variance marginale de y_{ijskl} est celle qui semble être indiquée par la figure 13.

Ensuite, nous considérons la fonction-lien habituelle pour la moyenne d'une loi de Poisson : c'est-à-dire l'équation modélisant le log de μ_{ijsk} en fonction de RG, STRATE, SEXE et AGE :

$$\log(\mu_{ijsk}) = \beta_{1i} + \beta_{2j} + \beta_{3s} + \beta_{4k} \quad (3)$$

Pour éviter d'avoir des paramètres redondants nous avons ajouté à (3) les restrictions habituelles (« corner constraints ») $\beta_{21} = \beta_{31} = \beta_{41} = 0$. Notons que nous aurions travaillé avec une relation (3) contenant une constante β_0 mais où les « corner constraints » auraient alors été $\beta_{11} = \beta_{21} = \beta_{31} = \beta_{41} = 0$. Ceci ne change rien au niveau des estimateurs. Cela modifie seulement l'interprétation des β. Nous avons opté pour la spécification (3) car elle permet d'éviter un problème d'identification que nous avons pu remarquer au niveau de β_0 : la chaîne Markov de β_0 converge lentement.

Pour les modèles que nous avons considérés, nous avons pris les paramètres β fixes. Nous avons également mis en œuvre ces modèles avec β aléatoires et nous avons observé une légère diminution de la déviance, mais des β aléatoires impliquent des μ_{ijsk} aléatoires et ceci pose des problèmes au niveau de l'utilisation du modèle pour obtenir la formule des estimateurs et de leur précision.

Si les β sont fixes, il faut encore, pour compléter la hiérarchie, spécifier des lois a priori pour eux ainsi que pour *alpha* et *kappa*. Nous avons travaillé avec des lois uniformes sur des intervalles suffisamment longs, ce qui correspond à une situation où on ne dispose pas d'information à priori sur ces paramètres (plus bas nous allons réaliser une analyse de sensibilité par rapport aux lois a priori en considérant d'autres lois non informatives):

$$\begin{gathered} alpha \sim \text{Unif}(0,100)\,,\ kappa \sim \text{Unif}(-1,100)\,, \\ \beta_{1i} \sim \text{Unif}(-10,10)\,, \beta_{2j} \sim \text{Unif}(-10,10)\,, \beta_{3s} \sim \text{Unif}(-10,10)\,, \beta_{4k} \sim \text{Unif}(-10,10) \quad (4) \end{gathered}$$

Si on met ensemble (1)-(4) on obtient le modèle suivant, que nous avons appelé Modèle 3 :

Modèle 3

$$\begin{gathered} y_{ijskl} \mid \nu_{ijskl} \underset{ind}{\sim} \text{Poisson}(\nu_{ijskl})\,, \\ \nu_{ijskl} \mid \mu_{ijsk}, alpha, kappa \underset{ind}{\sim} \text{Gamma}\left(\frac{\mu_{ijsk}^{1-kappa}}{alpha}, \frac{\mu_{ijsk}^{-kappa}}{alpha}\right), \\ \log(\mu_{ijsk}) = \beta_{1i} + \beta_{2j} + \beta_{3s} + \beta_{4k}\,, \\ alpha \sim \text{Unif}(0,100)\,,\ kappa \sim \text{Unif}(-1,100)\,, \\ \beta_{1i} \sim \text{Unif}(-10,10)\,, \beta_{2j} \sim \text{Unif}(-10,10)\,, \beta_{3s} \sim \text{Unif}(-10,10)\,, \beta_{4k} \sim \text{Unif}(-10,10) \quad (5) \end{gathered}$$

Comme nous l'avons fait remarquer plus haut, les modèles 1 et 2 sont respectivement, celui qui utilise une loi de Poisson de paramètre fixe ne tenant pas compte de la surdispersion, et celui qui utilise ν_{ijsk} au lieu de ν_{ijskl} comme paramètre aléatoire de la loi de Poisson, à savoir:

Modèle 1

$$y_{ijskl} \mid \nu_{ijsk} \underset{ind}{\sim} \text{Poisson}(\nu_{ijsk}),$$

$$\log(\nu_{ijsk}) = \beta_{1i} + \beta_{2j} + \beta_{3s} + \beta_{4k},$$

$$alpha \sim \text{Unif}(0,100),\ kappa \sim \text{Unif}(-1,100),$$

$$\beta_{1i} \sim \text{Unif}(-10,10), \beta_{2j} \sim \text{Unif}(-10,10), \beta_{3s} \sim \text{Unif}(-10,10), \beta_{4k} \sim \text{Unif}(-10,10) \quad (6)$$

et

Modèle 2

$$y_{ijskl} \mid \nu_{ijsk} \underset{ind}{\sim} \text{Poisson}(\nu_{ijsk}),$$

$$\nu_{ijsk} \mid \mu_{ijsk}, alpha, kappa \underset{ind}{\sim} \text{Gamma}\left(\frac{\mu_{ijsk}^{1-kappa}}{alpha}, \frac{\mu_{ijsk}^{-kappa}}{alpha}\right),$$

$$\log(\mu_{ijsk}) = \beta_{1i} + \beta_{2j} + \beta_{3s} + \beta_{4k},$$

$$alpha \sim \text{Unif}(0,100),\ kappa \sim \text{Unif}(-1,100),$$

$$\beta_{1i} \sim \text{Unif}(-10,10), \beta_{2j} \sim \text{Unif}(-10,10), \beta_{3s} \sim \text{Unif}(-10,10), \beta_{4k} \sim \text{Unif}(-10,10) \quad (7)$$

4 Estimation des paramètres du modèle 3 pour la variable R02AM

Nous voulons estimer le nombre moyen de recours au généraliste par individu et par région. Pour chaque région, ce paramètre sera donné par :

$$\mu_i = \frac{1}{N_i} \sum_j \sum_s \sum_k \sum_l y_{ijskl}$$

où N_i représente la taille de la région *i*. L'estimateur de μ_i sera donné par sa moyenne a posteriori et sa précision par sa variance a posteriori, à savoir :

$$\hat{\mu}_i = E(\mu_i \mid \mathbf{y}_{obs}) \text{ et } V(\hat{\mu}_i) = V(\mu_i \mid \mathbf{y}_{obs}) \quad (9)$$

où $\mathbf{y}_{obs}$ représente le vecteur de toutes les observations. Il faudra donc calculer ces deux paramètres de la loi a posteriori de μ_i dans le cas de chaque modèle. Il est impossible de déterminer cette loi ; c'est la raison pour laquelle nous avons utilisé l'échantillonnage de Gibbs.

4.1 L'échantillonnage de Gibbs

L'échantillonnage de Gibbs est une méthode itérative par laquelle sont générées, pour chaque paramètre d'un modèle, des suites Markov convergentes vers la distribution à posteriori du paramètre. On initialise la suite, ensuite on génère pour chaque paramètre de nouvelles valeurs à partir de sa distribution conditionnelle complète : la distribution du paramètre conditionnée par tous les autres paramètres et $\mathbf{y}_{obs}$. Nous allons illustrer pour le modèle 3. Considérons les distributions conditionnelles complètes des paramètres du modèle (pour les μ on n'a pas besoin de leurs distributions conditionnelles étant donnée qu'ils seront updatés automatiquement à partir des valeurs des β):

$$\nu_{ijskl} : f(\nu_{ijskl} \mid \mathbf{y}_{obs}, \overline{\mathbf{v}}, alpha, kappa, \boldsymbol{\beta})$$

($\overline{\mathbf{v}}$ est le vecteur de tous les ν sans le ν_{ijskl} pour lequel on calcule la distribution)

$$alpha : f(alpha \mid \mathbf{y}_{obs}, \mathbf{v}, kappa, \boldsymbol{\beta})$$

$$kappa: f(kappa \mid \mathbf{y}_{obs}, \mathbf{v}, alpha, \boldsymbol{\beta})$$

$$\beta_{1i} : f(\beta_{1i} \mid \mathbf{y}_{obs}, \mathbf{v}, alpha, kappa, \overline{\boldsymbol{\beta}}), i=1,\ldots,22$$

($\overline{\boldsymbol{\beta}}$ est le vecteur de tous les β sans le β_{1i} pour lequel on calcule la distribution)

$$\beta_{2j} : f(\beta_{2j} \mid \mathbf{y}_{obs}, \mathbf{v}, alpha, kappa, \overline{\boldsymbol{\beta}}), j=1,\ldots,5$$

($\overline{\boldsymbol{\beta}}$ est le vecteur de tous les β sans le β_{2j} pour lequel on calcule la distribution)

$$\beta_{3s} : f(\beta_{3s} \mid \mathbf{y}_{obs}, \mathbf{v}, alpha, kappa, \overline{\boldsymbol{\beta}}), s=1,2$$

($\overline{\boldsymbol{\beta}}$ est le vecteur de tous les β sans le β_{3s} pour lequel on calcule la distribution)

$$\beta_{4k} : f(\beta_{4k} \mid \mathbf{y}_{obs}, \mathbf{v}, alpha, kappa, \overline{\boldsymbol{\beta}}), k=1,\ldots,8$$

($\overline{\boldsymbol{\beta}}$ est le vecteur de tous les β sans le β_{4k} pour lequel on calcule la distribution)

Les chaînes Markov des paramètres doivent être initialisées. Soient $\mathbf{v}^0$, $alpha^0$, $kappa^0$ et $\boldsymbol{\beta}^0$ les valeurs initiales et supposons qu'on est arrivés à l'itération g-1 et qu'il faut obtenir les valeurs pour l'itération g. ν_{ijskl}^g s'obtient en échantillonnant une valeur à partir de $f(\nu_{ijskl} \mid \mathbf{y}_{obs}, alpha^{g-1}, kappa^{g-1}, \boldsymbol{\beta}^{g-1})$(on a supprimé $\overline{\mathbf{v}}$, voir plus bas). Une fois qu'on a obtenu toutes les valeurs de $\mathbf{v}^g$ on update la valeur de *alpha* à partir de $f(alpha \mid \mathbf{y}_{obs}, \mathbf{v}^g, kappa^{g-1}, \boldsymbol{\beta}^{g-1})$, celle de *kappa* à partir de $f(kappa \mid \mathbf{y}_{obs}, \mathbf{v}^g, alpha^g, \boldsymbol{\beta}^{g-1})$ et celle de β_{1i} à partir de $f(\beta_{1i} \mid \mathbf{y}_{obs}, \mathbf{v}^g, alpha^g, kappa^g, \overline{\boldsymbol{\beta}}^{g-1})$,etc... On observe qu'à l'intérieur d'une itération on utilise la valeur la plus récente d'un paramètre.

Quelques soient les valeurs initiales, les chaînes Markov convergent vers les distributions à posteriori des paramètres. A savoir, après *d* itérations («burn-in period »), on peut considérer que les valeurs des itérations *d*+1, *d*+2,.. proviennent des distributions à posteriori, donc ces valeurs peuvent être utilisées pour calculer différents paramètres de ces distributions (moyennes, variances, quantiles, fonctions de densités), en prenant la moyenne, la variance, etc… de la chaîne correspondante.

Une question naturelle est combien doit-on attendre jusqu'à la convergence ? Autrement dit qu'elle est la valeur de *d* ? Une deuxième question est combien d'itérations doit-on utiliser pour estimer un paramètre de la loi a posteriori avec suffisamment de précision ? Comme nous pouvons le voir ci-dessus, les valeurs de l'itération g dépendent des valeurs de l'itération g-1, donc il y a de l'autocorrélation à l'intérieur d'une chaîne Markov, ce qui fait que des longueurs plus grandes que dans le cas d'un échantillonnage indépendant seront nécessaires pour arriver à la même précision. Nous allons nous occuper de ces deux questions plus loin.

Pour réaliser l'échantillonnage de Gibbs il faut déterminer les distributions conditionnelles complètes. Nous allons montrer le principe en illustrant pour les ν. De la définition du modèle 3 on obtient la distribution jointe de tous les paramètres et du vecteur d'observations :

$$f(\mathbf{y}_{obs}, \mathbf{v}, alpha, kappa, \boldsymbol{\beta}) \propto \prod_{i,j,s,k,l} \frac{\nu_{ijskl}^{y_{ijskl}+\frac{\mu_{ijsk}^{1-kappa}}{alpha}-1} e^{-\nu_{ijskl}(1+\frac{\mu_{ijsk}^{-kappa}}{alpha})} (\frac{\mu_{ijsk}^{-kappa}}{alpha})^{\frac{\mu_{ijsk}^{1-kappa}}{alpha}}}{y_{ijskl}!\Gamma(\frac{\mu_{ijsk}^{1-kappa}}{alpha})}$$

($\propto$ signifie proportionnelle à, à savoir la densité est connue jusqu'à une constante de proportionnalité ; dans ce cas il s'agit des constantes qui normalisent les lois a priori uniformes du modèle et comme elles vont se simplifier nous les omettons). En intégrant la loi jointe par rapport à un ν_{ijskl} particulier on obtient :

$$f(\mathbf{y}_{obs}, \overline{\mathbf{v}}, alpha, kappa, \boldsymbol{\beta}) \propto q(\mathbf{y}_{obs}, \overline{\mathbf{v}}, alpha, kappa, \boldsymbol{\beta}) \int \nu_{ijskl}^{y_{ijskl}+\frac{\mu_{ijsk}^{1-kappa}}{alpha}-1} e^{-\nu_{ijskl}(1+\frac{\mu_{ijsk}^{-kappa}}{alpha})} d\nu_{ijskl}$$

($q(\mathbf{y}_{obs}, \overline{\mathbf{v}}, alpha, kappa, \boldsymbol{\beta})$ est une fonction qui dépend seulement de $\mathbf{y}_{obs}$, $\overline{\mathbf{v}}$, $alpha$, $kappa$ et $\boldsymbol{\beta}$). Quand on divise la loi jointe par la fonction ci-dessous, q se simplifie et on obtient:

$$f(\nu_{ijskl} \mid \mathbf{y}_{obs}, \overline{\mathbf{v}}, alpha, kappa, \boldsymbol{\beta}) = \frac{f(\mathbf{y}_{obs}, \mathbf{v}, alpha, kappa, \boldsymbol{\beta})}{f(\mathbf{y}_{obs}, \overline{\mathbf{v}}, alpha, kappa, \boldsymbol{\beta})} =$$

$$= \frac{\nu_{ijskl}^{y_{ijskl}+\frac{\mu_{ijsk}^{1-kappa}}{alpha}-1} e^{-\nu_{ijskl}(1+\frac{\mu_{ijsk}^{-kappa}}{alpha})}}{\int \nu_{ijskl}^{y_{ijskl}+\frac{\mu_{ijsk}^{1-kappa}}{alpha}-1} e^{-\nu_{ijskl}(1+\frac{\mu_{ijsk}^{-kappa}}{alpha})} d\nu_{ijskl}}$$

à savoir :

$$\nu_{ijskl} \mid \mathbf{y}_{obs}, alpha, kappa, \boldsymbol{\beta} = \text{Gamma}(y_{ijskl} + \frac{\mu_{ijsk}^{1-kappa}}{alpha}, 1 + \frac{\mu_{ijsk}^{-kappa}}{alpha})$$

(on peut observer que la distribution conditionnelle complète de ν_{ijskl} ne dépend pas des autres ν, ce pourquoi nous les avons supprimé de son écriture).

On voit donc que les distributions conditionnelles complètes des ν sont des lois standard et on peut utiliser un des algorithmes qui existent pour échantillonner à partir de ces lois. Malheureusement ceci n'est pas vrai pour tous les paramètres du modèle 3. Si on refait le même raisonnement pour les autres paramètres, alors on peut constater que leurs distributions conditionnelles peuvent être obtenues jusqu'à des constantes près :

$$f(kappa \mid \mathbf{y}_{obs}, \mathbf{v}, alpha, \boldsymbol{\beta}) \propto \prod_{i,j,s,k} [\frac{(\frac{\mu_{ijsk}^{-kappa}}{alpha})^{\frac{\mu_{ijsk}^{1-kappa}}{alpha}}}{\Gamma(\frac{\mu_{ijsk}^{1-kappa}}{alpha})}]^{n_{ijsk}} \prod_{i,j,s,k,l} v_{ijskl}^{\frac{\mu_{ijsk}^{1-kappa}}{alpha}} e^{-v_{ijskl}\frac{\mu_{ijsk}^{-kappa}}{alpha}},$$

$$f(alpha \mid \mathbf{y}_{obs}, \mathbf{v}, kappa, \boldsymbol{\beta}) \propto \prod_{i,j,s,k} [\frac{(\frac{\mu_{ijsk}^{-kappa}}{alpha})^{\frac{\mu_{ijsk}^{1-kappa}}{alpha}}}{\Gamma(\frac{\mu_{ijsk}^{1-kappa}}{alpha})}]^{n_{ijsk}} \prod_{i,j,s,k,l} v_{ijskl}^{\frac{\mu_{ijsk}^{1-kappa}}{alpha}} e^{-v_{ijskl}\frac{\mu_{ijsk}^{-kappa}}{alpha}},$$

$$f(\beta_{1i} \mid \mathbf{y}_{obs}, \mathbf{v}, kappa, alpha, \bar{\boldsymbol{\beta}}) \propto \prod_{j,s,k} [\frac{(\frac{\mu_{ijsk}^{-kappa}}{alpha})^{\frac{\mu_{ijsk}^{1-kappa}}{alpha}}}{\Gamma(\frac{\mu_{ijsk}^{1-kappa}}{alpha})}]^{n_{ijsk}} \prod_{j,s,k,l} v_{ijskl}^{\frac{\mu_{ijsk}^{1-kappa}}{alpha}} e^{-v_{ijskl}\frac{\mu_{ijsk}^{-kappa}}{alpha}}$$

$$f(\beta_{2j} \mid \mathbf{y}_{obs}, \mathbf{v}, kappa, alpha, \bar{\boldsymbol{\beta}}) \propto \prod_{i,s,k} [\frac{(\frac{\mu_{ijsk}^{-kappa}}{alpha})^{\frac{\mu_{ijsk}^{1-kappa}}{alpha}}}{\Gamma(\frac{\mu_{ijsk}^{1-kappa}}{alpha})}]^{n_{ijsk}} \prod_{i,s,k,l} v_{ijskl}^{\frac{\mu_{ijsk}^{1-kappa}}{alpha}} e^{-v_{ijskl}\frac{\mu_{ijsk}^{-kappa}}{alpha}}$$

$$f(\beta_{3s} \mid \mathbf{y}_{obs}, \mathbf{v}, kappa, alpha, \bar{\boldsymbol{\beta}}) \propto \prod_{i,j,k} [\frac{(\frac{\mu_{ijsk}^{-kappa}}{alpha})^{\frac{\mu_{ijsk}^{1-kappa}}{alpha}}}{\Gamma(\frac{\mu_{ijsk}^{1-kappa}}{alpha})}]^{n_{ijsk}} \prod_{i,j,k,l} v_{ijskl}^{\frac{\mu_{ijsk}^{1-kappa}}{alpha}} e^{-v_{ijskl}\frac{\mu_{ijsk}^{-kappa}}{alpha}}$$

$$f(\beta_{4k} \mid \mathbf{y}_{obs}, \mathbf{v}, kappa, alpha, \bar{\boldsymbol{\beta}}) \propto \prod_{i,j,s} [\frac{(\frac{\mu_{ijsk}^{-kappa}}{alpha})^{\frac{\mu_{ijsk}^{1-kappa}}{alpha}}}{\Gamma(\frac{\mu_{ijsk}^{1-kappa}}{alpha})}]^{n_{ijsk}} \prod_{i,j,s,l} v_{ijskl}^{\frac{\mu_{ijsk}^{1-kappa}}{alpha}} e^{-v_{ijskl}\frac{\mu_{ijsk}^{-kappa}}{alpha}}$$

Pour obtenir les constantes de proportionnalité il faudrait intégrer par rapport à chaque paramètre et on voit que la forme de l'intégrant est telle qu'on ne peut pas obtenir une forme fermée pour la distribution correspondante qui dans ce cas sont des distributions non standard. La solution est de recourir à des algorithmes qui permettent d'échantillonner à partir d'une telle distribution.

Nous avons obtenu les chaînes Markov à l'aide du logiciel WinBugs 4. Celui-ci utilise trois types d'algorithmes pour les distributions non standard : l'échantillonnage réjectif adaptif (Gilks 1992) pour les cas où la densité ou son noyau sont des fonctions log-concaves, le « slice sampling » (Neal 1997) si l'ensemble des valeurs du paramètre est borné, sinon l'algorithme de Metropolis-Hastings. *alpha*, *kappa* et les β sont tous bornés étant données les lois a priori uniformes que nous avons considérées donc dans le cas de ces paramètres et du modèle 3 le logiciel utilise le « slice sampling ». Si on avait considéré par exemple pour les β une loi normale de moyenne zéro et de variance 1000 qui est aussi une distribution a priori non informative étant donné la valeur de sa variance, alors pour ces paramètres l'algorithme

de Metropolis-Hastings aurait été utilisé à la place. Dans ce qui suit nous présentons brièvement les deux méthodes d'échantillonnage à partir d'une distribution non standard.

Le « slice sampling» (Neal 1997)

On veut échantillonner à partir d'une distribution dont la densité est proportionnelle à une fonction $f(x)$, $x \in R^n$. On peut le faire en échantillonnant uniformément dans la surface n+1 dimensionnelle qui se trouve sous le graphe de $f(x)$. Cette idée peut être réalisée en introduisant une variable réelle y et en définissant une distribution uniforme sur la région $U = \{(x, y), 0 < y < f(x)\}$. La fonction de densité de cette distribution sera donnée par :

$$p(x,y) = \begin{cases} \dfrac{1}{\int f(x)dx} & \text{si } 0 < y < f(x) \\ 0 & \text{sinon} \end{cases}$$

La densité marginale de x sera alors la densité à partir de laquelle on veut échantillonner:

$$p(x) = \int_0^{f(x)} \frac{1}{\int f(x)dx} dx = \frac{f(x)}{\int f(x)dx}$$

Donc, si on réussit échantillonner à partir de $p(x, y)$, il faudra tout simplement ignorer y et garder x pour avoir un échantillon de $p(x)$. Echantillonner à partir de $p(x, y)$ ne peut pas se faire directement mais par l'intermédiaire d'une chaîne Markov générée selon l'échantillonnage de Gibbs à partir de distributions $p(y \mid x)$ et $p(x \mid y)$. $p(y \mid x)$ est la distribution uniforme sur l'intervalle $(0, f(x))$, donc son échantillonnage ne pose aucun problème. $p(x \mid y)$ est la distribution uniforme sur l'ensemble $S = \{x \mid y < f(x)\}$ (S est le « slice » défini par y). Echantillonner uniformément sur S est la partie non évidente. Dans cette situation il faudra imaginer un update pour x qui laisse invariante la distribution uniforme sur S. Clairement, si on est à l'itération g pour trouver les valeurs de l'itération g+1 on procède en trois étapes :

1) on sélectionne uniformément une valeur y^{g+1} de $(0, f(x^g))$ et on définit $S^{g+1} = \{x \mid y^{g+1} < f(x)\}$;

2) on trouve un intervalle I^{g+1} autour de x^g qui contient une partie aussi grande que possible de S^{g+1} (pour que la nouvelle valeur de x soit aussi loin que possible de x^g) mais qui ne soit pas trop large (ce qui va rendre la chaîne Markov inefficiente) ; en même temps, pour l'intervalle choisi, on doit pouvoir déterminer l'ensemble A défini comme $A = \{x \mid x \in S \cap I, P(I \mid x) = P(I \mid x^g)\}$; voir Neal (1997) qui montre plusieurs façons de trouver I^{g+1} ;

3) on choisit uniformément x^{g+1} dans I^{g+1} jusqu'au moment où on trouve un point qui soit aussi dans S^{g+1}.

Neal (1997) montrent la correctitude de 1)-3) : la chaîne Markov qui en résulte laisse la distribution cible $f(x)$ invariante.

L'algorithme de Metropolis-Hastings

Pour obtenir une nouvelle valeur d'un paramètre l'algorithme utilise une densité candidat. Clairement, supposons qu'on est arrivé à l'itération *g* et qu'on veut updater la valeur x^g à partir de la distribution de *x* notée $f(x)$. Soit $g(x^* \mid x^g)$ la densité candidat qui virtuellement peut être n'importe quelle densité à partir de laquelle on peut échantillonner facilement et soit x^* une valeur tirée de $g(x^* \mid x^g)$. On accepte x^* comme valeur avec la probabilité $\alpha_{MH}(x^* \mid x^g)$ donnée par :

$$\alpha_{MH}(x^* \mid x^g) = \min(1, \frac{\frac{f(x^*)}{g(x^* \mid x^g)}}{\frac{f(x^g)}{g(x^g \mid x^*)}})$$

Pour résumer, la valeur x^{g+1} s'obtient comme suit :

1) on échantillonne x^* de $g(x^* \mid x^g)$;
2) on échantillonne *u* uniformément sur (0,1) ;
3) si $u \le \alpha_{MH}(x^* \mid x^g)$ alors $x^{g+1} = x^*$; sinon $x^{g+1} = x^g$

De la formule de $\alpha_{MH}(x^* \mid x^g)$ on voit que $f(x)$ peut être connue jusqu'à une constante près puisqu'elle va se simplifier.

WinBugs 4 utilise des fonctions de densité candidat qui dépendent de x^* et de x^g par la différence $|x^* - x^g|$ (par exemple des lois normales de cette moyenne). Dans ce cas la densité candidat est symétrique ($g(x^* \mid x^g) = g(x^g \mid x^*)$) est la probabilité d'acceptation se réduit à :

$$\alpha_{MH}(x^* \mid x^g) = \min(1, \frac{f(x^*)}{f(x^g)})$$

Cette nouvelle formule pour $\alpha_{MH}(x^* \mid x^g)$ est plus intuitive: quelque soit l'état où on se trouve on choisit toujours de visiter des points à plus forte densité que des points à faible densité.

Si la densité candidat est une normale de moyenne $|x^* - x^g|$, alors il nous reste de choisir sa variance. Le choix doit avoir comme résultat une chaîne qui converge vite et dont le paramètre traverse en peu d'itérations une grande partie de la densité $f(x)$. Si la variance est trop grande alors le paramètre essaie de faire de grands sauts dans l'espace de ses valeurs mais une grande partie des valeurs seront rejetés. Si la variance est trop petite alors les valeurs proposées sont presque toutes acceptées mais le paramètre a besoin de beaucoup d'itérations pour couvrir une grande partie de la densité. Pour cette raison, WinBugs 4 a une phase d'adaptation où sur base des probabilités d'acceptation des premières itérations choisit la valeur optimale pour la variance de la densité candidat.

4.2 L'obtention des formules théoriques des estimateurs

μ_i n'est pas un paramètre du modèle 3 étant donné qu'on a modélisé les valeurs individuelles de R02AM, donc WinBugs 4 ne fournit pas une chaîne Markov pour μ_i. Ci-après nous montrons comment on peut estimer (9) à partir des chaînes du modèle. Nous illustrons pour le modèle 3. Pour les modèles 1 et 2 nous ne faisons pas les démonstrations parce que, en raison de leur mauvais ajustement (voir plus bas), nous n'allons pas les utiliser pour faire des estimations.

En partageant les individus entre individus observés et non observés, μ_i peut être réécrit sous la forme:

$$\mu_i = \frac{1}{N_i}[\sum_j \sum_s \sum_k \sum_{l \in obs_i} y_{ijskl} + \sum_j \sum_s \sum_k \sum_{l \in nobs_i} y_{ijskl}]$$

où obs_i et $nobs_i$ représentent les éléments observés, respectivement non observés, de la région *i*. Dès lors:

$$\hat{\mu}_i = E(\mu_i \mid \mathbf{y}_{obs}) = \frac{1}{N_i}[\sum_j \sum_s \sum_k \sum_{l \in obs_i} y_{ijskl} + \sum_j \sum_s \sum_k \sum_{l \in nobs_i} E(y_{ijskl} \mid \mathbf{y}_{obs})] \quad (10)$$

et

$$V(\hat{\mu}_i) = V(\mu_i \mid \mathbf{y}_{obs}) = \frac{1}{N_i^2} V(\sum_j \sum_s \sum_k \sum_{l \in nobs_i} y_{ijskl} \mid \mathbf{y}_{obs}) \quad (11)$$

Pour obtenir le (10), il faut évaluer $E(y_{ijskl} \mid \mathbf{y}_{obs})$ pour $l_0 \in nobs_i$:

$$E(y_{ijskl_0} \mid \mathbf{y}_{obs}) = E_{\nu_{ijskl_0} \mid \mathbf{y}_{obs}} E(y_{ijskl_0} \mid \nu_{ijskl_0}, \mathbf{y}_{obs})$$

Pour $E(y_{ijskl_0} \mid \nu_{ijskl_0}, \mathbf{y}_{obs})$ on a :

$$E(y_{ijskl_0} \mid \nu_{ijskl_0}, \mathbf{y}_{obs}) = \int y_{ijskl_0} f(y_{ijskl_0} \mid \nu_{ijskl_0}, \mathbf{y}_{obs}) dy_{ijskl_0} =$$
$$\int y_{ijskl_0} f(y_{ijskl_0}, \mathbf{y}_{obs} \mid \nu_{ijskl_0}) / f(\mathbf{y}_{obs} \mid \nu_{ijskl_0}) dy_{ijskl_0} =$$

$$\int y_{ijskl_0} f(y_{ijskl_0} \mid \nu_{ijskl_0}) dy_{ijskl_0} = E(y_{ijskl_0} \mid \nu_{ijskl_0}) = \nu_{ijskl_0} \Rightarrow$$
$$E(y_{ijskl_0} \mid \nu_{ijskl_0}, \mathbf{y}_{obs}) = \nu_{ijskl_0} \quad (12)$$

On a utilisé le fait que les *y* sont indépendants quand on conditionne par les ν et donc $f(y_{ijskl_0}, \mathbf{y}_{obs} \mid \nu_{ijskl_0}) / f(\mathbf{y}_{obs} \mid \nu_{ijskl_0}) = f(y_{ijskl_0} \mid \nu_{ijskl_0}) f(\mathbf{y}_{obs} \mid \nu_{ijskl_0}) / f(\mathbf{y}_{obs} \mid \nu_{ijskl_0}) = f(y_{ijskl_0} \mid \nu_{ijskl_0})$
Alors $E(y_{ijskl_0} \mid \mathbf{y}_{obs})$ sera :

$$E(y_{ijskl_0} \mid \mathbf{y}_{obs}) = E_{\nu_{ijskl_0} \mid \mathbf{y}_{obs}}(\nu_{ijskl_0}) = E(\nu_{ijskl_0} \mid \mathbf{y}_{obs})$$

Si on refait le raisonnement avec les ν à la place des *y* et les μ à la place des ν on obtient :

$$E(\nu_{ijskl_0} \mid \mathbf{y}_{obs}) = E_{\mu_{ijsk}, alpha, kappa \mid \mathbf{y}_{obs}} E(\nu_{ijskl_0} \mid \mu_{ijsk}, alpha, kappa, \mathbf{y}_{obs})$$

On voit donc qu'on a besoin de la loi $\nu_{ijskl_0} \mid \mathbf{y}_{obs}, alpha, kappa, \mu_{ijsk}$ pour un individu l_0 non observé (plus haut nous avons déterminé les distributions conditionnelles complètes des ν, à savoir la loi $\nu_{ijskl} \mid \mathbf{y}_{obs}, alpha, kappa, \mu_{ijsk}$ pour un *l* observé). Pour la déterminer on procède de la manière suivante :

- on détermine la loi jointe de tous les *y* (regroupé dans le vecteur $\mathbf{y}$) et de tous les ν regroupés dans le vecteur $\mathbf{v}$, conditionnée par les hyperparamètres du modèle:

$$f(\mathbf{y}, \mathbf{v} \mid alpha, kappa, \boldsymbol{\beta}) \propto \prod_{i,j,s,k,l} \frac{\nu_{ijskl}^{y_{ijskl} + \frac{\mu_{ijsk}^{1-kappa}}{alpha} - 1} e^{-\nu_{ijskl}(1 + \frac{\mu_{ijsk}^{-kappa}}{alpha})}}{y_{ijskl}!}$$

(dans ce cas $\propto$ signifie proportionnelle par rapport à une constante dépendant de *alpha*, *kappa* et $\boldsymbol{\beta}$) ;
- on intègre par rapport à tous les *y* non observés et par rapport à tous les ν sauf ν_{ijskl_0} et on obtient :

$$f(\mathbf{y}_{obs}, \nu_{ijskl_o} \mid alpha, kappa, \boldsymbol{\beta}) \propto \nu_{ijskl_0}^{\frac{\mu_{ijsk}^{1-kappa}}{alpha} - 1} e^{-\nu_{ijskl_0}(1 + \frac{\mu_{ijsk}^{-kappa}}{alpha})} G(\mathbf{y}_{obs}, alpha, kappa, \boldsymbol{\beta})$$

($G(\mathbf{y}_{obs}, alpha, kappa, \boldsymbol{\beta})$ est une fonction de $\mathbf{y}_{obs}, alpha, kappa$ et $\boldsymbol{\beta}$) ;
- on intègre par rapport à ν_{ijskl_0} et on obtient:

$$f(\mathbf{y}_{obs} \mid alpha, kappa, \boldsymbol{\beta}) \propto G(\mathbf{y}_{obs}, alpha, kappa, \boldsymbol{\beta})$$

- utilisant le fait que :

$$f(\nu_{ijskl_o} \mid \mathbf{y}_{obs}, alpha, kappa, \boldsymbol{\beta}) = \frac{f(\mathbf{y}_{obs}, \nu_{ijskl_o} \mid alpha, kappa, \boldsymbol{\beta})}{f(\mathbf{y}_{obs} \mid alpha, kappa, \boldsymbol{\beta})}$$

- on obtient finalement :

$$\nu_{ijskl_o} \mid \mathbf{y}_{obs}, alpha, kappa, \boldsymbol{\beta} = \text{Gamma}(\frac{\mu_{ijsk}^{1-kappa}}{alpha}; \frac{\mu_{ijsk}^{-kappa}}{alpha})$$

Donc :

$$\nu_{ijskl_o} \mid \mathbf{y}_{obs}, alpha, kappa, \boldsymbol{\beta} = \nu_{ijskl_o} \mid alpha, kappa, \boldsymbol{\beta} \quad (13)$$

Alors, utilisant :

$$E_{\mu_{ijsk}, alpha, kappa \mid \mathbf{y}_{obs}}(\mu_{ijsk}) = \int \mu_{ijsk} [\int f(\mu_{ijsk}, alpha, kappa \mid \mathbf{y}_{obs}) d(alpha) d(kappa)] d\mu_{ijsk} =$$
$$= \int \mu_{ijsk} f(\mu_{ijsk} \mid \mathbf{y}_{obs}) d\mu_{ijsk} = E(\mu_{ijsk} \mid \mathbf{y}_{obs}) \Rightarrow$$
$$E(\nu_{ijskl} \mid \mathbf{y}_{obs}) = E(\mu_{ijsk} \mid \mathbf{y}_{obs}) \quad (14)$$

De (10) et (14) on déduit:

$$\hat{\mu}_i = \frac{1}{N_i} [\sum_j \sum_s \sum_k \sum_{l \in obs_i} y_{ijskl} + E(\sum_j \sum_s \sum_k (N_{ijsk} - n_{ijsk}) \mu_{ijsk} \mid \mathbf{y}_{obs})] \quad (15)$$

où N_{ijsk} et n_{ijsk} représentent le nombre total et respectivement le nombre d'individus sélectionnés dans la cellule $i \times j \times s \times k$. Si on veut estimer la moyenne de la France Métropolitaine, alors on peut procéder de la même manière et on obtient une formule similaire à (15), avec une somme supplémentaire, la somme d'après *i* et l'effectif de la région remplacé par la population de la France *N* :

$$\hat{\mu} = \frac{1}{N} [\sum_i \sum_j \sum_s \sum_k \sum_{l \in obs_i} y_{ijskl} + E(\sum_i \sum_j \sum_s \sum_k (N_{ijsk} - n_{ijsk}) \mu_{ijsk} \mid \mathbf{y}_{obs})] \quad (15.1)$$

Pour la variance, l'expression (11) nous permet d'écrire :

$$V(\mu_i \mid \mathbf{y}_{obs}) = \frac{1}{N_i^2} [E_{\nu_{ijskl} \mid \mathbf{y}_{obs}} V(\sum_j \sum_s \sum_k \sum_{l \in nobs_i} y_{ijskl} \mid \nu_{ijskl}, \mathbf{y}_{obs}) +$$
$$+ V_{\nu_{ijskl} \mid \mathbf{y}_{obs}} E(\sum_j \sum_s \sum_k \sum_{l \in nobs_i} y_{ijskl} \mid \nu_{ijskl}, \mathbf{y}_{obs})] \quad (16)$$

Tenant compte de l'indépendance conditionnelle et de (14), le premier terme de (16) devient:

$$E_{\nu_{ijskl}|\mathbf{y}_s} V(\sum_j \sum_s \sum_k \sum_{l \in nobs_i} y_{ijskl} | \nu_{ijskl}, \mathbf{y}_{obs}) = E(\sum_j \sum_s \sum_k \sum_{l \in nobs_i} \nu_{ijskl} | \mathbf{y}_{obs}) =$$
$$= \sum_j \sum_s \sum_k (N_{ijsk} - n_{ijsk}) E(\mu_{ijsk} | \mathbf{y}_{obs}) \quad (17)$$

Pour le deuxième terme de (16), on a:

$$V_{\nu_{ijskl}|\mathbf{y}_s} E(\sum_j \sum_s \sum_k \sum_{l \in nobs_i} y_{ijskl} | \nu_{ijskl}, \mathbf{y}_{obs}) = V(\sum_j \sum_s \sum_k \sum_{l \in nobs_i} \nu_{ijskl} | \mathbf{y}_{obs}) =$$
$$E_{\mu_{ijsk}, alpha, kappa | \mathbf{y}_{obs}} V(\sum_j \sum_s \sum_k \sum_{l \in nobs_i} \nu_{ijskl} | \mu_{ijsk}, alpha, kappa, \mathbf{y}_{obs}) +$$
$$V_{\mu_{ijsk}, alpha, kappa | \mathbf{y}_{obs}} E(\sum_j \sum_s \sum_k \sum_{l \in nobs_i} \nu_{ijskl} | \mu_{ijsk}, alpha, kappa, \mathbf{y}_{obs}) \quad (18)$$

Pour le premier terme de (18), on peut utiliser (13):

$$E_{\mu_{ijsk}, alpha, kappa | \mathbf{y}_{obs}} V(\sum_j \sum_s \sum_k \sum_{l \in nobs_i} \nu_{ijskl} | \mu_{ijsk}, alpha, kappa, \mathbf{y}_{obs}) =$$
$$= E(\sum_j \sum_s \sum_k (N_{ijsk} - n_{ijsk}) alpha \mu_{ijsk}^{1+kappa} | \mathbf{y}_{obs}) \quad (19)$$

Pour le deuxième terme de (18), on aura aussi :

$$V_{\mu_{ijsk}, alpha, kappa | \mathbf{y}_{obs}} E(\sum_j \sum_s \sum_k \sum_{l \in nobs_i} \nu_{ijskl} | \mu_{ijsk}, alpha, kappa, \mathbf{y}_{obs}) =$$
$$= V(\sum_j \sum_s \sum_k (N_{ijsk} - n_{ijsk}) \mu_{ijsk} | \mathbf{y}_{obs}) \quad (20)$$

De (16), (17), (18), (19) et (20), on a :

$$V(\mu_i | \mathbf{y}_{obs}) = \frac{1}{N_i^2} [E(\sum_j \sum_s \sum_k (N_{ijsk} - n_{ijsk}) \mu_{ijsk} | \mathbf{y}_{obs}) +$$
$$+ V(\sum_j \sum_s \sum_k (N_{ijsk} - n_{ijsk}) \mu_{ijsk} | \mathbf{y}_{obs}) + E(\sum_j \sum_s \sum_k (N_{ijsk} - n_{ijsk}) alpha \mu_{ijsk}^{1+kappa} | \mathbf{y}_{obs})] \quad (21)$$

Pour la précision de $\hat{\mu}$ il faut faire les mêmes modifications que ci-dessus et on obtient :

$$V(\mu | \mathbf{y}_{obs}) = \frac{1}{N^2} [E(\sum_i \sum_j \sum_s \sum_k (N_{ijsk} - n_{ijsk}) \mu_{ijsk} | \mathbf{y}_{obs}) +$$
$$+ V(\sum_i \sum_j \sum_s \sum_k (N_{ijsk} - n_{ijsk}) \mu_{ijsk} | \mathbf{y}_{obs}) + E(\sum_i \sum_j \sum_s \sum_k (N_{ijsk} - n_{ijsk}) alpha \mu_{ijsk}^{1+kappa} | \mathbf{y}_{obs})] \quad (21.1)$$

(15)-(15.1) et (21)-(21.1) montrent comment les estimateurs $\hat{\mu}_i$ et $\hat{\mu}$ ainsi que leurs précisions peuvent être écrits en fonction des paramètres du modèle 3. Etant données ces

formules, les valeurs des estimateurs et de leurs précisions utilisent les chaînes Markov des paramètres du modèle 3 selon les formules:

$$\hat{\mu}_i = \frac{1}{N_i}[\sum_j \sum_s \sum_k \sum_{l \in obs_i} y_{ijskl} + \frac{1}{G}\sum_g \sum_j \sum_s \sum_k (N_{ijsk} - n_{ijsk})\mu_{ijsk}^g] \quad (22)$$

$$V(\hat{\mu}_i) = \frac{1}{N_i^2}\{\frac{1}{G}\sum_g \sum_j \sum_s \sum_k (N_{ijsk} - n_{ijsk})\mu_{ijsk}^g + \frac{1}{G}\sum_g \sum_j \sum_s \sum_k (N_{ijsk} - n_{ijsk})alpha^g \mu_{ijsk}^{g\,(1+kappa^g)} + \\ + \frac{1}{G}\sum_g [\sum_j \sum_s \sum_k (N_{ijsk} - n_{ijsk})\mu_{ijsk}^g]^2 - [\frac{1}{G}\sum_g \sum_j \sum_s \sum_k (N_{ijsk} - n_{ijsk})\mu_{ijsk}^g]^2\} \quad (23)$$

et pour la France métropolitaine :

$$\hat{\mu} = \frac{1}{N}[\sum_i \sum_j \sum_s \sum_k \sum_{l \in obs_i} y_{ijskl} + \frac{1}{G}\sum_g \sum_i \sum_j \sum_s \sum_k (N_{ijsk} - n_{ijsk})\mu_{ijsk}^g] \quad (24)$$

$$V(\hat{\mu}) = \frac{1}{N^2}\{\frac{1}{G}\sum_i \sum_g \sum_j \sum_s \sum_k (N_{ijsk} - n_{ijsk})\mu_{ijsk}^g + \\ + \frac{1}{G}\sum_g \sum_i \sum_j \sum_s \sum_k (N_{ijsk} - n_{ijsk})alpha^g \mu_{ijsk}^{g\,(1+kappa^g)} + \\ + \frac{1}{G}\sum_g [\sum_i \sum_j \sum_s \sum_k (N_{ijsk} - n_{ijsk})\mu_{ijsk}^g]^2 - [\frac{1}{G}\sum_g \sum_i \sum_j \sum_s \sum_k (N_{ijsk} - n_{ijsk})\mu_{ijsk}^g]^2\} \quad (25)$$

où μ_{ijsk}^g, $alpha^g$ et $kappa^g$, $g = 1,...,G$ sont les termes des chaînes Markov des paramètres μ_{ijsk}, *alpha*, et respectivement *kappa*. De (22)-(25) on voit que nous avons besoin de connaitre les N_{ijsk}, le nombre total d'individus de la cellule $i \times j \times s \times k$.

5 Choix et ajustement des modèles pour la variable R02AM

Nous avons sélectionné quelques critères qui nous permettent de choisir entre les modèles présentés plus haut, ainsi que des critères qui permettent d'évaluer si un modèle est bien adapté aux données. Etant donné le nombre important d'observations, les ajustements des modèles ont été jugés à partir de l'échantillon sans extension (28259 individus).

Un des principes est de générer à partir du modèle un vecteur de données $\mathbf{y}_{new}$ et d'utiliser une mesure de discrépence entre $\mathbf{y}_{new}$ et $\mathbf{y}_{obs}$. On choisira le modèle ayant la plus petite moyenne a posteriori de la discrépence. Pour les trois modèles considérés nous avons calculé trois mesures de discrépence basées sur $\mathbf{y}_{new}$ et $\mathbf{y}_{obs}$ et appropriées quand on travaille avec une variable de comptage:

$$T(\mathbf{y}_{new}, \mathbf{y}_{obs}) = \sum_i \frac{(y_{new,i} - y_{obs,i})^2}{(y_{new,i} + 0.5)},$$

$$d(\mathbf{y}_{new}, \mathbf{y}_{obs}) = 2\sum_i [(y_{obs,i} + 0.5)\log\frac{(y_{obs,i} + 0.5)}{(y_{new,i} + 0.5)} - (y_{obs,i} - y_{new,i})],$$

$$D(\mathbf{y}_{new}, \mathbf{y}_{obs}) = \sum_i (E(y_{new,i} \mid \mathbf{y}_{obs}) - y_{obs,i})^2 + \sum_i V(y_{new,i} \mid \mathbf{y}_{obs})$$

(ici pour faciliter l'écriture, l'indice *i* désigne un individu et pas une région).

Le nouvelles données $\mathbf{y}_{new}$ sont générées à partir des chaînes Markov de ν_i. Par exemple, pour un individu i on a une chaîne Markov ν_i^g, $g = 1,...,G$. Pour chaque ν_i^g, on génère une valeur $y_{new,i}^g$ à partir de la loi de Poisson(ν_i^g). Pour chaque modèle nous avons généré une chaîne Markov, avec une période « burn-in » de 2000 itérations et un nombre d'itérations utilisées égales à 1000, donc nous avons généré pour chaque individu 1000 nouvelles valeurs. Etant données que l'autocorrélation des chaînes des ν est très faibles, ces chaînes convergent très vite et couvrent bien l'espace des distributions a posteriori correspondantes. Ceci veut dire qu'on aurait pu prendre une période « burn-in » plus courte et moins de 1000 nouvelles valeurs pour chaque individu (d'ailleurs prendre plus de 1000 aurait été difficile étant donné le nombre important d'individus- 28259). Les moyennes et variances a posteriori des mesures de discrépence se calculent comme suit:

$$E(T \mid \mathbf{y}_{obs}) = \frac{1}{G}\sum_g \sum_i \frac{(y_{new,i}^g - y_{obs,i})^2}{(y_{new,i}^g + 0.5)},$$

$$E(d \mid \mathbf{y}_{obs}) = 2\frac{1}{G}\sum_g \sum_i [(y_{obs,i} + 0.5)\log\frac{(y_{obs,i} + 0.5)}{(y_{new,i}^g + 0.5)} - (y_{obs,i} - y_{new,i}^g)],$$

$$E(y_{new,i} \mid \mathbf{y}_{obs}) = \frac{1}{G}\sum_g y_{new,i}^g,\ V(y_{new,i} \mid \mathbf{y}_{obs}) = \frac{1}{G}\sum_g (y_{new,i}^g - E(y_{new,i} \mid \mathbf{y}_{obs}))^2$$

En dehors de ces trois mesures, nous avons aussi calculé une autre mesure de l'ajustement du modèle qui est la Déviance. Celle-ci se définit comme moins deux fois le logarithme de la fonction de vraisemblance de $\mathbf{y}_{obs}$:

$$\text{Déviance}(\mathbf{y}_{obs}, \mathbf{v}) = -2 * \sum_i \log(f(y_{obs,i} \mid \nu_i))$$

donc sa moyenne a posteriori se calcule selon :

$$E(\text{Déviance} \mid \mathbf{y}_{obs}) = -2 * \frac{1}{G}\sum_g \sum_i \log(f(y_{obs,i} \mid \nu_i^g))$$

où $f(y_{obs,i} \mid \nu_i^g)$ est la fonction de densité d'une loi Poisson de moyenne ν_i^g calculée en $y_{obs,i}$.

A partir de l'échantillon sans extension nous avons obtenu les résultats suivants:

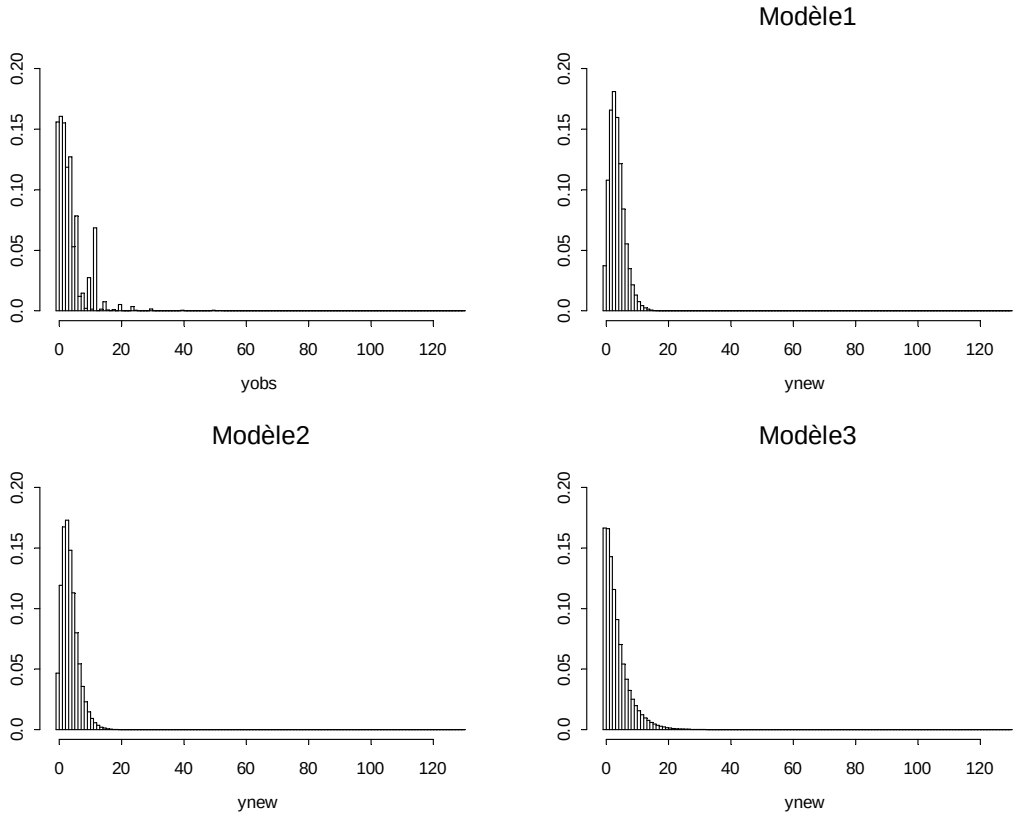
Tableau 1: *Moyennes a posteriori des mesures de discrépence*

Mesure	Modèle 1	Modèle 2	Modèle 3
$E(T(\mathbf{y}_{new}, \mathbf{y}_{obs}) \mid \mathbf{y}_{obs})$	204591.7	180845.9	66212.74
$E(d(\mathbf{y}_{new}, \mathbf{y}_{obs}) \mid \mathbf{y}_{obs})$	114884.9	109110.9	48095.09
$D(\mathbf{y}_{new}, \mathbf{y}_{obs})$	647926.3	619880.3	232111.5
$E(\text{Déviance} \mid \mathbf{y}_{obs})$	174500	168300	103500

Le tableau 1 montre clairement que le modèle 1 utilisant une loi Poisson de paramètre fixe est le pire et que le modèle 2 qui considère un paramètre aléatoire ν_{ijsk} commun pour tous les individus d'une même cellule représente une légère amélioration ; cependant, c'est le modèle 3 qui apporte une amélioration importante.

Une fois générées les nouvelles données on peut visualiser la qualité du modèle en comparant la distribution de $\mathbf{y}_{obs}$ à celle de $\mathbf{y}_{new}$ La distribution de $\mathbf{y}_{new}$ notée $f(\cdot|\mathbf{y}_{obs})$ est appelée distribution postérieure prédictive . Pour les trois modèles nous avons obtenu le graphique suivant:

Figure14:*Distribution de* $\mathbf{y}_{obs}$ *et Distributions Postérieurs Prédictives sous les Modèles 1-3*



On peut de nouveau remarquer le mauvais ajustement des modèles 1 et 2. Dans le cas du modèle 3 on observe une sous estimation de la valeur 12. La tendance naturelle est que des valeurs de plus en plus grandes de R02AM sont de moins en moins observées. C'est ce qui se passe dans la distribution de $\mathbf{y}_{obs}$ (sauf pour la valeur 12 qui est atypique de ce point de vue ; beaucoup d'individus ont tendance à répondre une douzaine de fois même si le nombre réel est 13 ou légèrement plus, ou 10 ou légèrement moins). Pour cette raison le modèle 3 n'estime pas correctement l'effectif de 12.

Toujours en utilisant les nouvelles données, on peut évaluer l'ajustement d'un modèle à l'aide d'une mesure de discrépence, cette fois-ci entre les *y* et les ν. Nous avons déjà vu

plus haut une telle mesure : la Déviance. On calcule $\text{Déviance}(\mathbf{y}_{obs}, \mathbf{v})$ et $\text{Déviance}(\mathbf{y}_{new}, \mathbf{v})$ et on estime la probabilité que $\text{Déviance}(\mathbf{y}_{new}, \mathbf{v})$ soit plus grand que $\text{Déviance}(\mathbf{y}_{obs}, \mathbf{v})$ par :

$$\hat{p} = \frac{1}{G}\sum_{g} I[\text{Déviance}(\mathbf{y}_{new}^{g}, \mathbf{v}^{g}) \geq \text{Déviance}(\mathbf{y}_{obs}, \mathbf{v}^{g})]$$

Une valeur de *p* proche de 0.5 indique un bon ajustement. Des valeurs extrêmes proche de 1 ou de 0 indique un modèle qui n'ajuste pas bien les données Nous avons calculé cette mesures seulement pour le modèles 3. Nous avons obtenu $\hat{p} = 0.40$ qui montre que le modèle 3 est bien adapté aux données.

Une autre mesure appropriée pour une variable de comptage est :

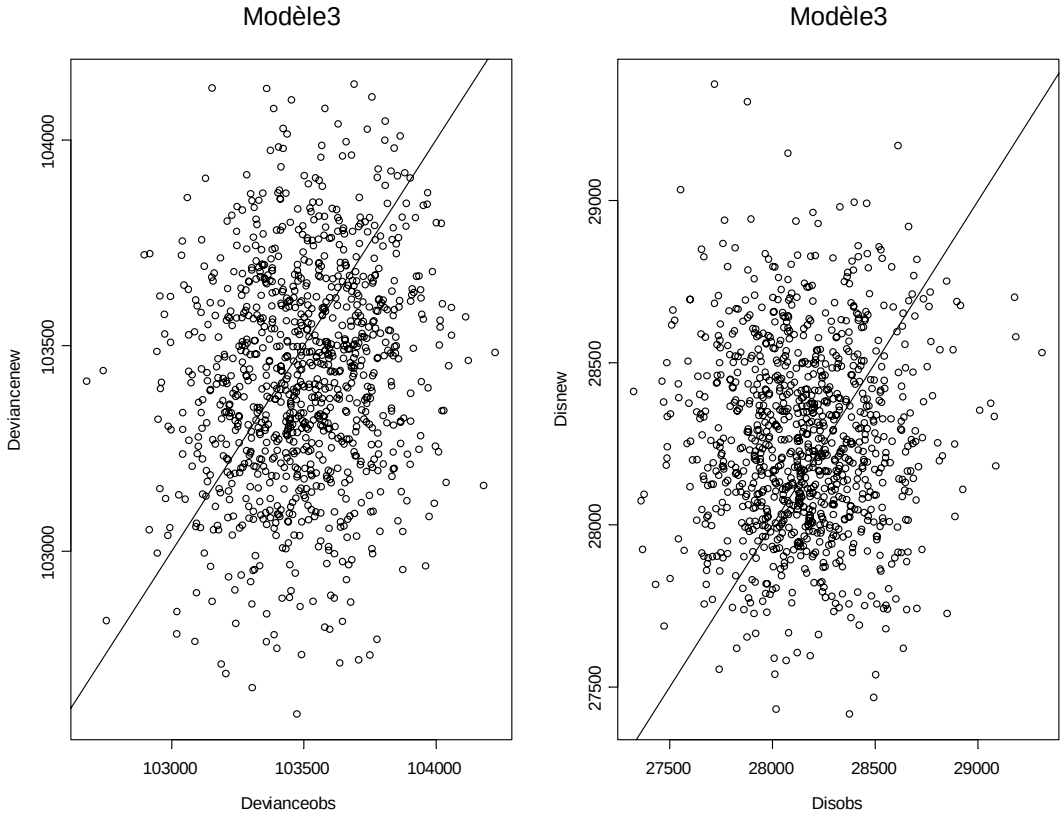
$$\text{Dis}(\mathbf{y}, \mathbf{v}) = \sum_{i} \frac{(y_i - \nu_i)^2}{\nu_i}$$

On calcule $\text{Dis}(\mathbf{y}_{obs}, \mathbf{v})$ et $\text{Dis}(\mathbf{y}_{new}, \mathbf{v})$ et on estime ensuite la probabilité que $\text{Dis}(\mathbf{y}_{new}, \mathbf{v})$ soit plus grand que $\text{Dis}(\mathbf{y}_{obs}, \mathbf{v})$ par :

$$\hat{p} = \frac{1}{G}\sum_{g} I[\text{Dis}(\mathbf{y}_{new}^{g}, \mathbf{v}^{g}) \geq \text{Dis}(\mathbf{y}_{obs}, \mathbf{v}^{g})]$$

Nous avons obtenu $\hat{p} = 0.58$. On peut représenter graphiquement $\text{Déviance}(\mathbf{y}_{obs}, \mathbf{v})$ contre $\text{Déviance}(\mathbf{y}_{new}, \mathbf{v})$ et $\text{Dis}(\mathbf{y}_{obs}, \mathbf{v})$ contre $\text{Dis}(\mathbf{y}_{new}, \mathbf{v})$. Pour un bon modèle, la moitié des points se trouvent au-dessus de la première bissectrice et l'autre moitié en dessous. Nous avons obtenu :

Figure 15 : *Déviance*$(\mathbf{y}_{obs}, \mathbf{v}^g)$ *vs Déviance*$(\mathbf{y}^g_{new}, \mathbf{v}^g)$ *et Dis*$(\mathbf{y}_{obs}, \mathbf{v}^g)$ *vs Dis*$(\mathbf{y}^g_{new}, \mathbf{v}^g)$



Une autre façon de tester le modèle consiste à faire appel à la validation croisée (cross-validation). Ceci consiste à calculer des paramètres de la loi $f(y_i \mid \mathbf{y}_{(i)})$ et à les comparer avec ce qui a été observé, où $\mathbf{y}_{(i)}$ représente tous les individus observés sauf l'individu *i*. Les densités $f(y_i \mid \mathbf{y}_{(i)})$ sont appelées densités prédictives par cross validation. Un premier paramètre est le CPO_i - conditional predictive ordinate de *i* :

$$\text{CPO}_i = f(y_i \mid \mathbf{y}_{(i)})$$

et le modèle à choisir est celui ayant les CPO_i les plus grands. Pour estimer CPO_i (donc $f(y_i \mid \mathbf{y}_{(i)})$) on peut utiliser (voir entre autre Rao (2003)):

$$\text{CPO}_i = \hat{f}(y_i \mid \mathbf{y}_{(i)}) = \frac{1}{\frac{1}{G}\sum_g \frac{1}{f(y_i \mid \nu_i^g)}}$$

Nous n'avons pas représenté les CPO_i étant donné leur nombre important. Pour valider le modèle on peut aussi utiliser des résidus de la forme :

$$r_i = \frac{y_{obs,i} - E(y_i \mid \mathbf{y}_{(i)})}{\sqrt{V(y_i \mid \mathbf{y}_{(i)})}}$$

Afin de calculer les r_i on doit calculer $E(y_i \mid \mathbf{y}_{(i)})$ et $V(y_i \mid \mathbf{y}_{(i)})$. L'espérance conditionnelle d'une fonction $a(y_i)$ peut s'estimer par (voir Rao (2003)):

$$\hat{E}(a(y_i) \mid \mathbf{y}_{(i)}) = \hat{f}(y_i \mid \mathbf{y}_{(i)}) \frac{1}{G} \sum_g \frac{b_i(\nu_i^g)}{f(y_i \mid \nu_i^g)} \quad (26)$$

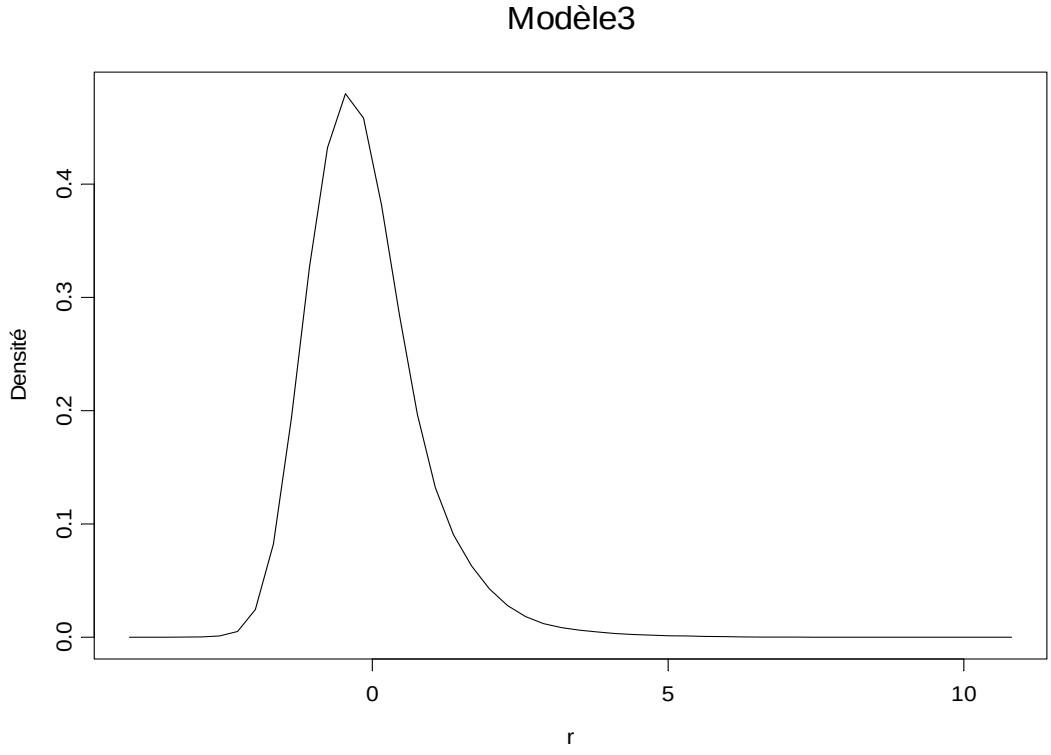
où $b_i(\nu_i^g) = E(a(y_i) \mid \nu_i^g)$. Pour le modèle 3 on aura:

$$\hat{E}(y_i \mid \mathbf{y}_{(i)}) = \hat{f}(y_i \mid \mathbf{y}_{(i)}) \frac{1}{G} \sum_g \frac{\nu_i^g}{f(y_i \mid \nu_i^g)}$$

$$\hat{V}(y_i \mid \mathbf{y}_{(i)}) = \hat{f}(y_i \mid \mathbf{y}_{(i)}) \frac{1}{G} \sum_g \frac{\nu_i^g + \nu_i^{g2}}{f(y_i \mid \nu_i^g)} - [\hat{f}(y_i \mid \mathbf{y}_{(i)}) \frac{1}{G} \sum_g \frac{\nu_i^g}{f(y_i \mid \nu_i^g)}]^2$$

En utilisant les deux formules ci-dessus, on peut maintenant calculer les r_i. Dans le cas d'un modèle qui ajuste bien les données la moyenne des r_i devrait être proche de zéro et l'écart-type proche de 1. Pour le modèle 3, nous avons obtenu une moyenne des r_i égale à -0.10, un écart-type égal à 0.907 et une médiane égale à -0.26, alors que l'intervalle interquartile est [-0.71,0.30]. Le pourcentage des r_i supérieurs en valeur absolu à 2 est égal à 2.94%. Toutes ces valeurs indiquent que le modèle 3 est assez bien adapté aux données. Plus bas il y a la densité des résidus:

Figure 16 : *Distribution des résidus* r_i



(26) peut être appliquée si l'on d'une formule explicite pour $b_i(\nu_i)$. Sinon, alors il faudra échantillonner à partir de $f(y_i \mid \mathbf{y}_{(i)})$. C'est ce qu'il faut faire pour une autre mesure de

l'ajustement représentée par les probabilités $P(y_i \le y_{obs,i} \mid \mathbf{y}_{(i)})$. (26) n'est plus utilisable étant donné que dans ce cas les $a(y_i)$ sont tels qu'on ne peut pas calculer explicitement les $b_i(\nu_i)$.

Gelfand (1996) donne une méthode pour échantillonner $f(y_i \mid \mathbf{y}_{(i)})$ sans devoir relancer les chaînes Markov avec le vecteur $\mathbf{y}_{(i)}$ à la place de $\mathbf{y}_{obs}$ pour chaque individu *i* de l'échantillon (ce qui serait impossible du point de vue temps). Pour notre modèle 3 ceci revient à : dans chaque vecteur $\mathbf{v}_i = (\nu_i^g)$, tirer un échantillon avec remise et avec probabilités proportionnelles à $1/f(y_{obs,i} \mid \nu_i^g)$ (soient $\mathbf{v}_i^* = (\nu_i^{g^*})$ les nouveaux ν); ensuite échantillonner $y_{new,i}^{g^*}$ à partir de Poisson$(\nu_i^{g^*})$. Le vecteur $\mathbf{y}_{new,i}^*$ ainsi obtenu sera composé de *G* observations tirées de $f(y_i \mid \mathbf{y}_{(i)})$. Les probabilités $P(y_i \le y_{obs,i} \mid \mathbf{y}_{(i)})$ seront estimées par:

$$\hat{P}(y_i \le y_{obs,i} \mid \mathbf{y}_{(i)}) = \frac{1}{G}\sum_g I(y_{new,i}^{g^*} - y_{obs,i} \le 0)$$

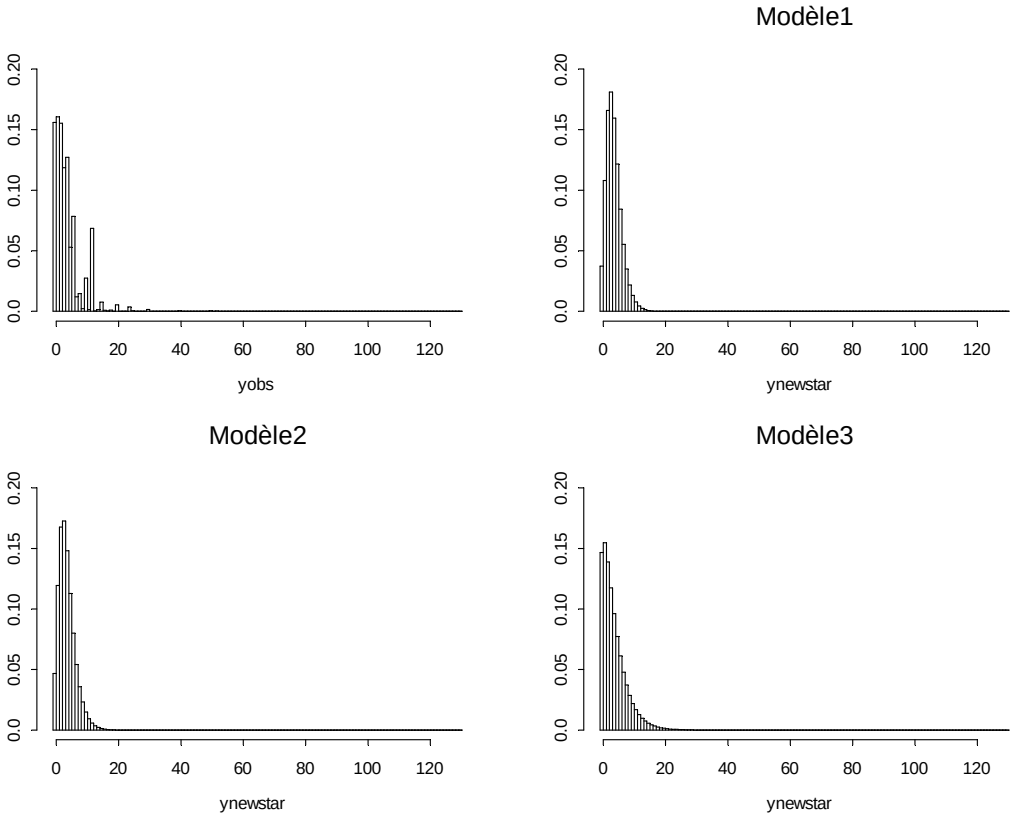
Le résultat est un vecteur contenant pour chaque individu *i* les estimations de ces probabilités. Pour le modèle 3 la moyenne du vecteur est égale à 0.55, la médiane à 0.56 et l'intervalle interquartile est [0.37,0.74], confirmant ainsi à nouveau le bon ajustement du modèle étant donné qu'il faut avoir des probabilités proches de 0.5.

Les valeurs de $\mathbf{y}_{new,i}^*$ peuvent être utilisées pour estimer la moyenne de n'importe quelle fonction de y_i sous la loi $f(y_i \mid \mathbf{y}_{(i)})$:

$$\hat{E}(h(y_i) \mid \mathbf{y}_{(i)}) = \frac{1}{G}\sum_g h(y_{new,i}^{g^*})$$

On peut représenter graphiquement l'histogramme des nouvelles valeurs $\mathbf{y}_{new,i}^*$ et la comparer aux valeurs observées $\mathbf{y}_{obs}$. Les histogrammes pour les trois modèles et l'histogramme de $\mathbf{y}_{obs}$ sont représentés dans la figure 17. On peut faire la même observation que dans le cas de la distribution postérieure prédictive : la valeur 12 est ici aussi sous estimé par le modèle.

Figure 17: *Distribution de* $\mathbf{y}_{obs}$ *et Distribution de* $\mathbf{y}^{*}_{new}$ *sous les Modèles 1-3*



6 Calcul des estimations pour la variable R02AM

6.1 Le cas de l'échantillon sans extension

Utilisant (22)-(25) nous avons obtenu des estimations à partir du modèle 3. Dans ce qui suit nous détaillons le cas de l'échantillon sans extension.

Quand on utilise une chaîne de Markov il faut d'abord s'assurer que la chaîne a convergé vers la distribution a posteriori des paramètres. Il faut faire ceci pour tous les paramètres d'intérêt. De (22)-(25), on peut voir que nous utilisons les paramètres *alpha*, *kappa* et μ_{ijsk}. Pour vérifier la convergence des μ_{ijsk} il suffit de s'assurer que β_1, β_2, β_3 et β_4 convergent étant donnée la relation déterministe (3). Pour visualiser la convergence nous avons roulé trois chaînes Markov pour chaque paramètre du modèle 3. Les chaînes ont été initialisées à partir de valeurs éloignées pour s'assurer que la convergence est réelle. Ci-dessous se trouvent les représentations graphiques des chaînes pour les paramètres d'intérêt jusqu'à l'itération 3000 (pour certains paramètres on a représenté leurs valeurs à partir de l'itération 1000 pour éviter d'avoir la situation de la figure 19 où on ne voit pas grand-chose à cause de la valeur extrême qui a initialisé la chaîne). Comme on peut le constater visuellement, les chaînes des paramètres convergent après 2000 itérations, (pour certains paramètres, même avant).

Figure 18 : *Les chaînes Markov de kappa*

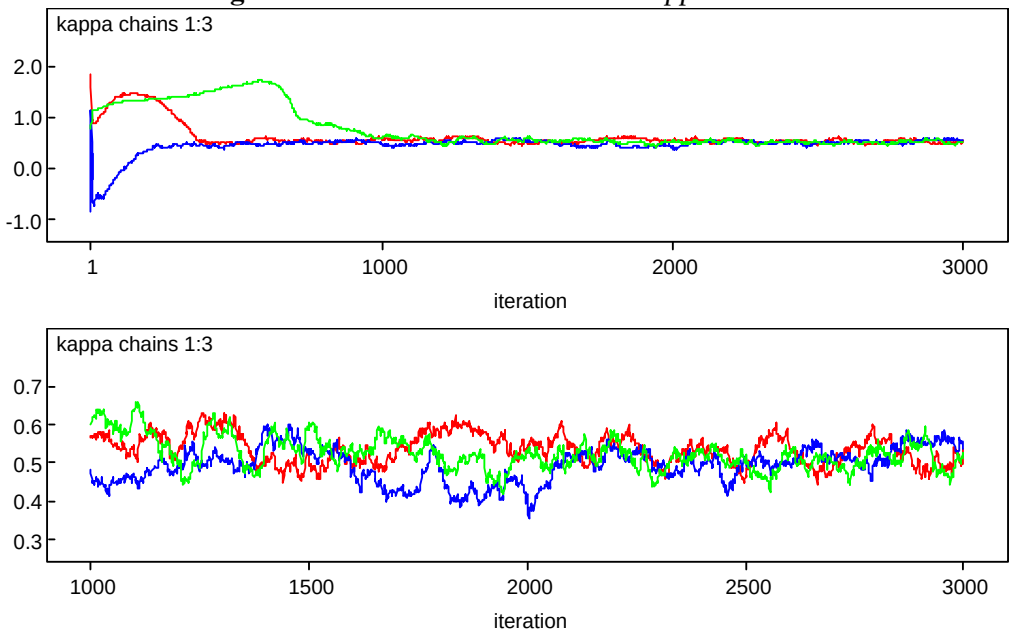


Figure 19 : *Les chaînes Markov de alpha*

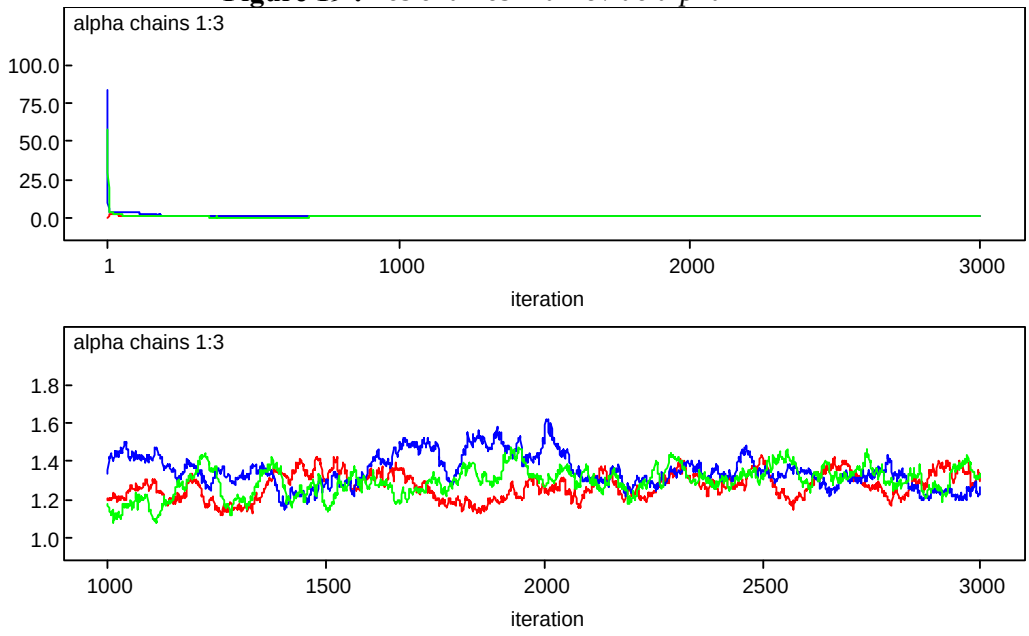
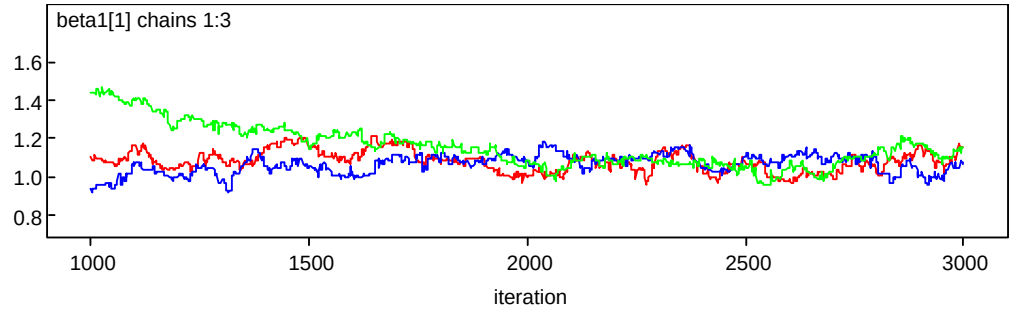
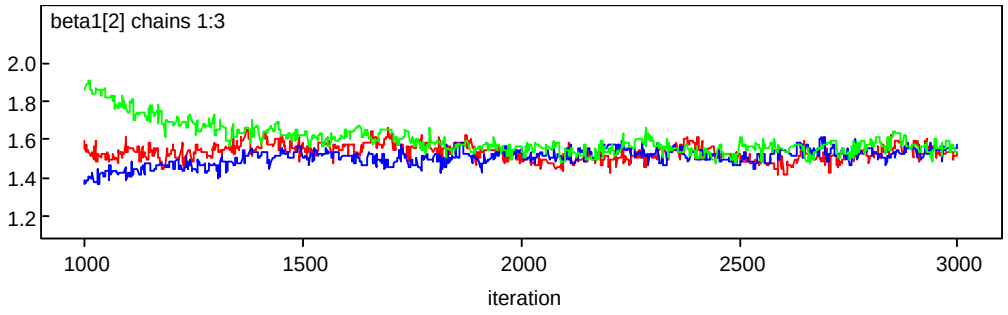
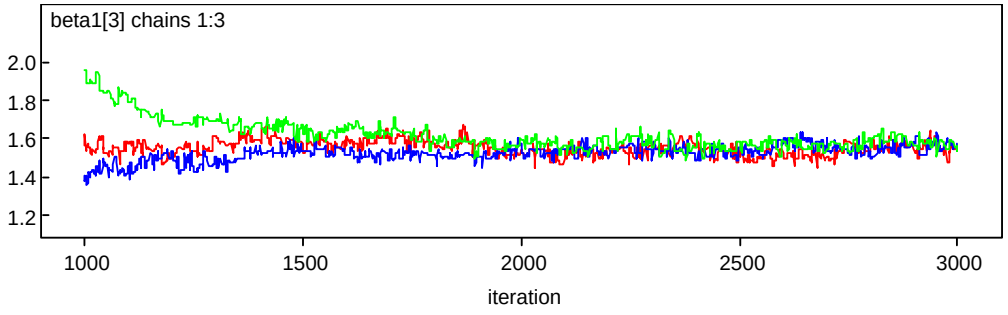
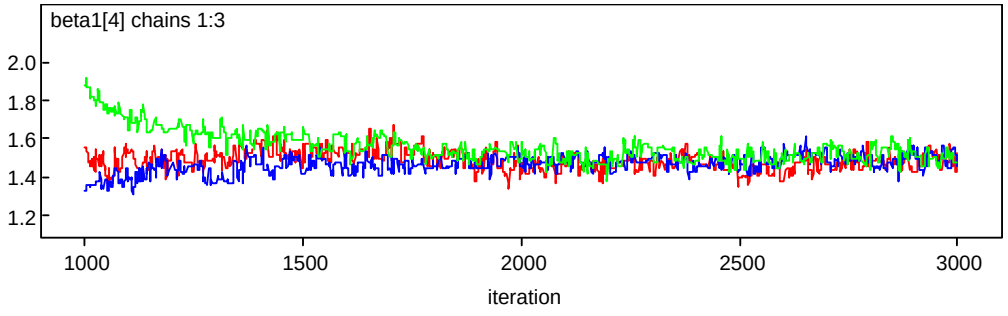


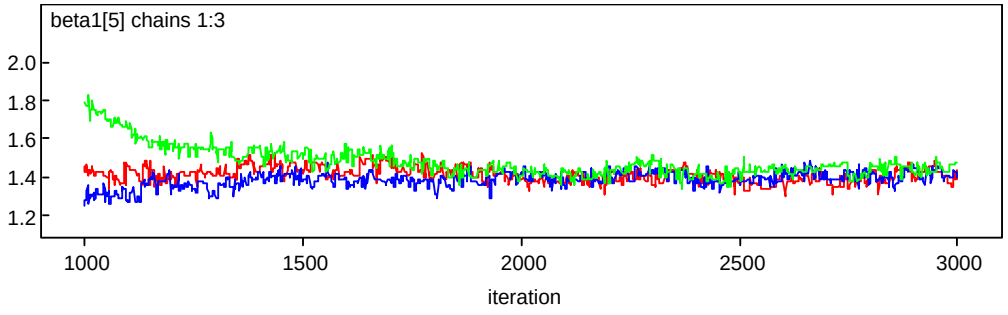
Figure 20 : *Les chaînes Markov de* β_1

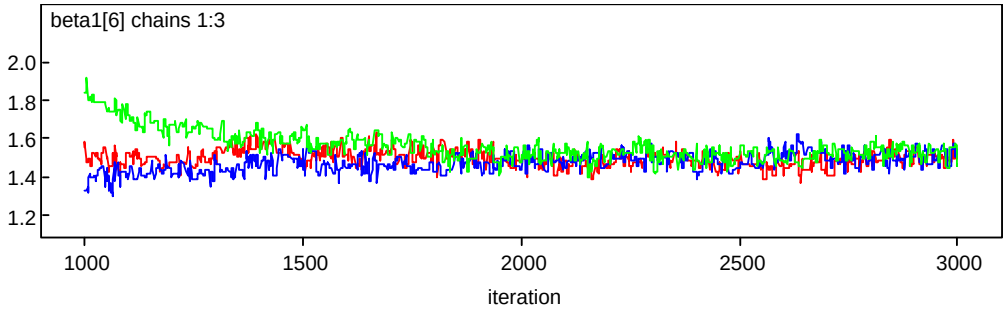


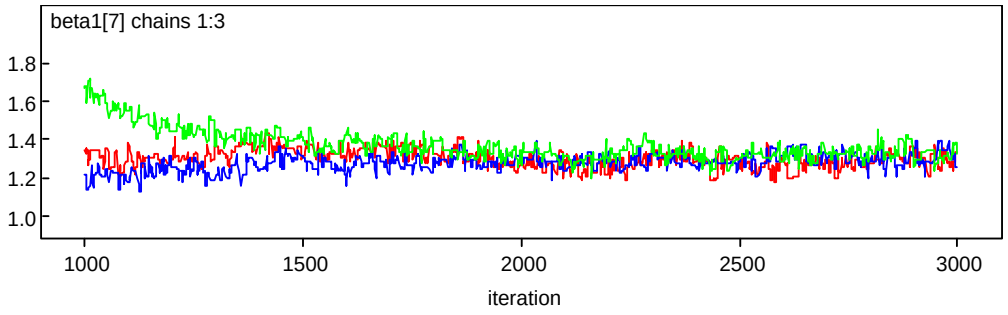

beta1[2] chains 1:3
2.0
1.8
1.6
1.4
1.2
1000
1500
2000
2500
3000
iteration

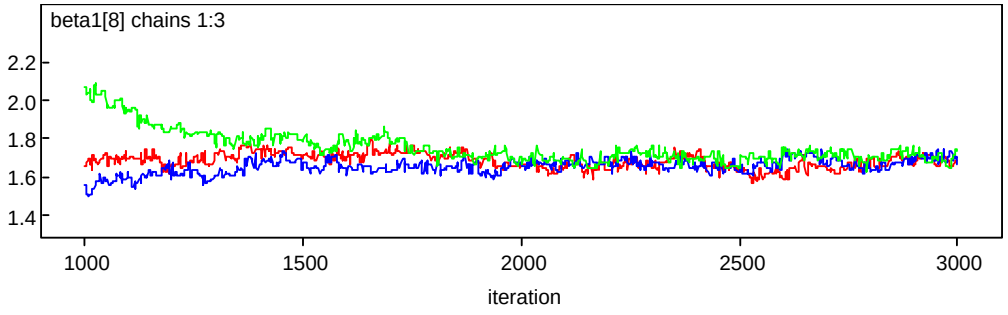

beta1[3] chains 1:3
2.0
1.8
1.6
1.4
1.2
1000
1500
2000
2500
3000
iteration

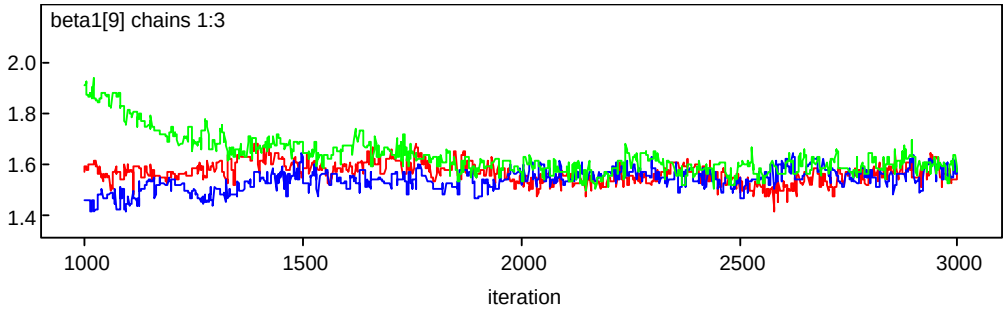

beta1[4] chains 1:3
2.0
1.8
1.6
1.4
1.2
1000
1500
2000
2500
3000
iteration

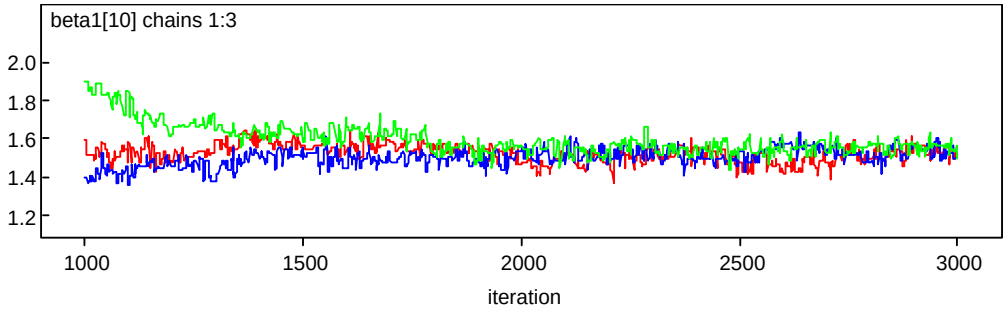

beta1[5] chains 1:3
2.0
1.8
1.6
1.4
1.2
1000
1500
2000
2500
3000
iteration

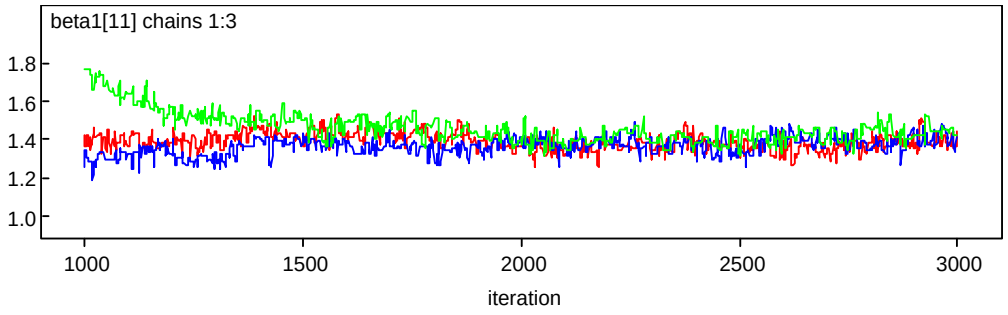

beta1[6] chains 1:3
2.0
1.8
1.6
1.4
1.2
1000
1500
2000
2500
3000
iteration

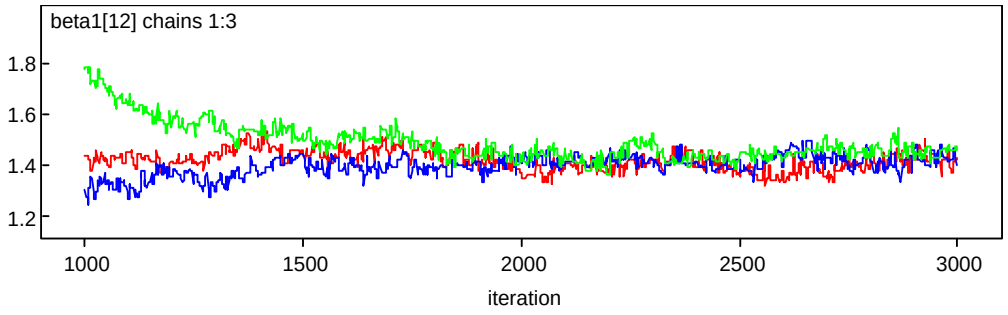

beta1[7] chains 1:3
1.8
1.6
1.4
1.2
1.0
1000
1500
2000
2500
3000
iteration

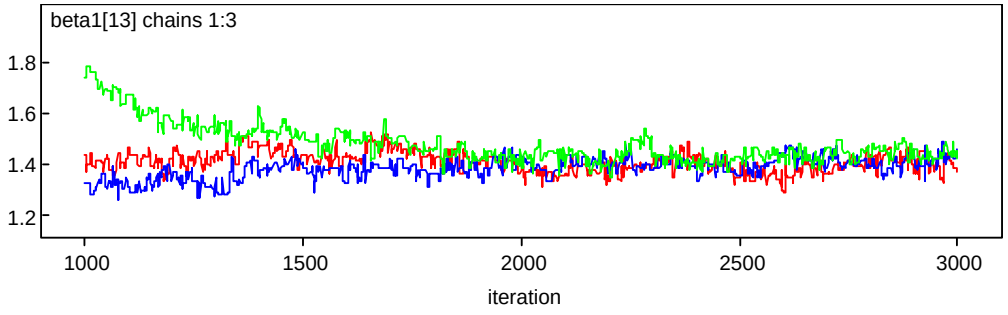

beta1[8] chains 1:3
2.2
2.0
1.8
1.6
1.4
1000
1500
2000
2500
3000
iteration

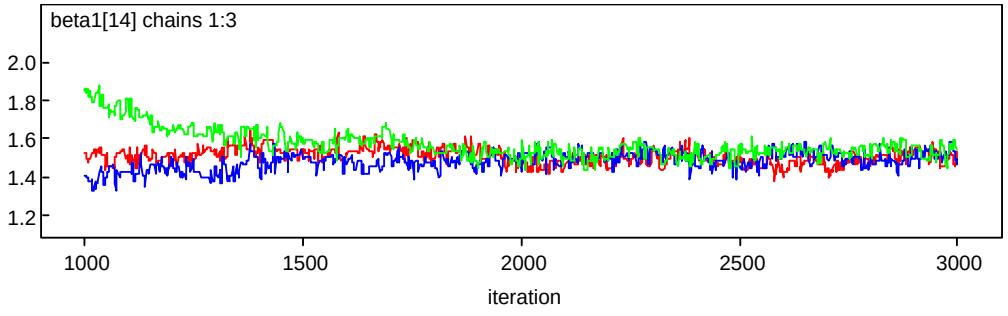

beta1[9] chains 1:3
2.0
1.8
1.6
1.4
1000
1500
2000
2500
3000
iteration

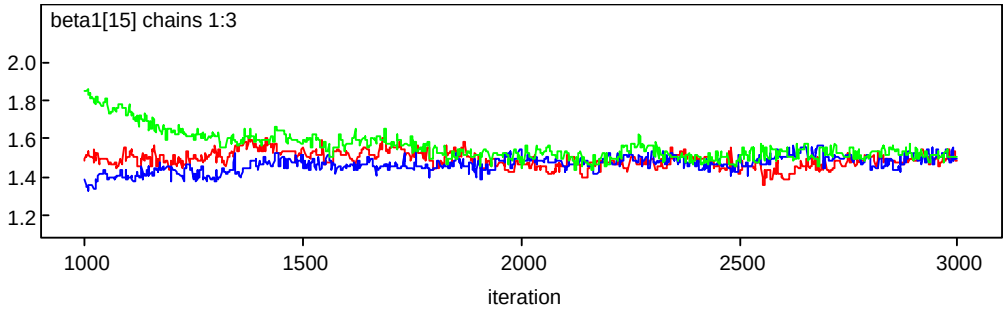

beta1[10] chains 1:3
2.0
1.8
1.6
1.4
1.2
1000
1500
2000
2500
3000
iteration

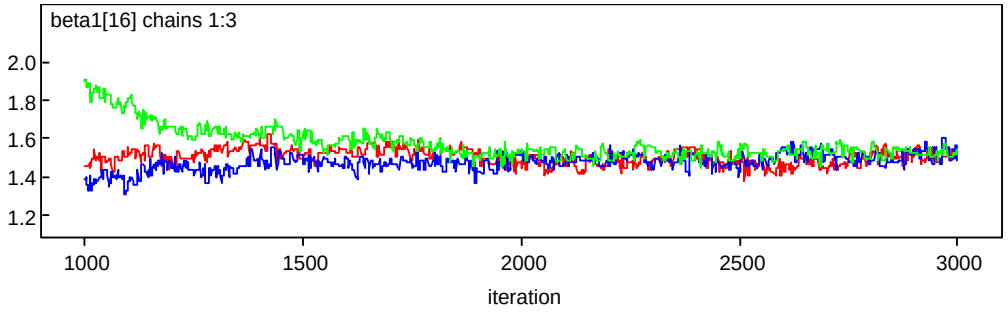

beta1[11] chains 1:3
1.8
1.6
1.4
1.2
1.0
1000
1500
2000
2500
3000
iteration

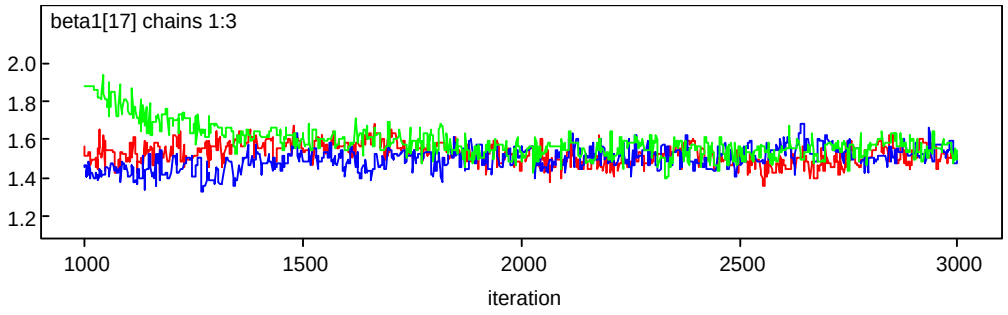

beta1[12] chains 1:3
1.8
1.6
1.4
1.2
1000
1500
2000
2500
3000
iteration

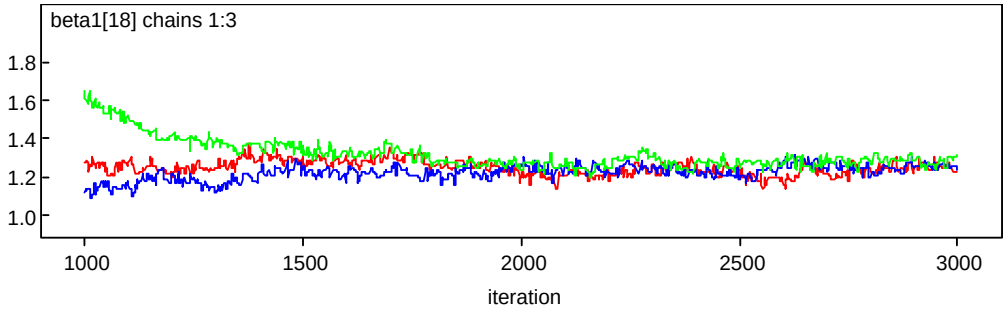

beta1[13] chains 1:3
1.8
1.6
1.4
1.2
1000
1500
2000
2500
3000
iteration

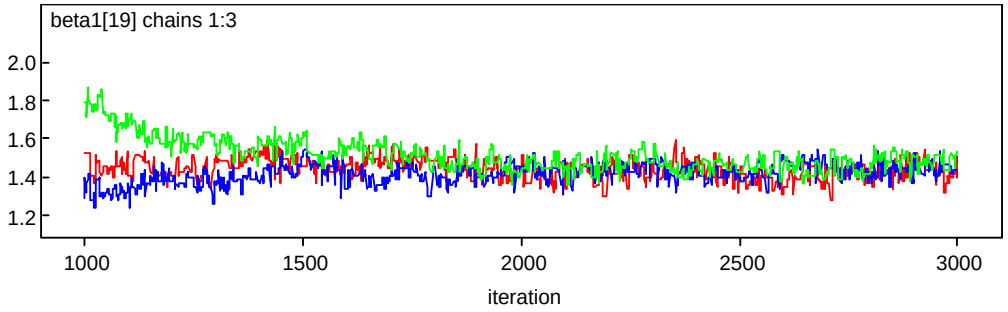

beta1[14] chains 1:3
2.0
1.8
1.6
1.4
1.2
1000
1500
2000
2500
3000
iteration

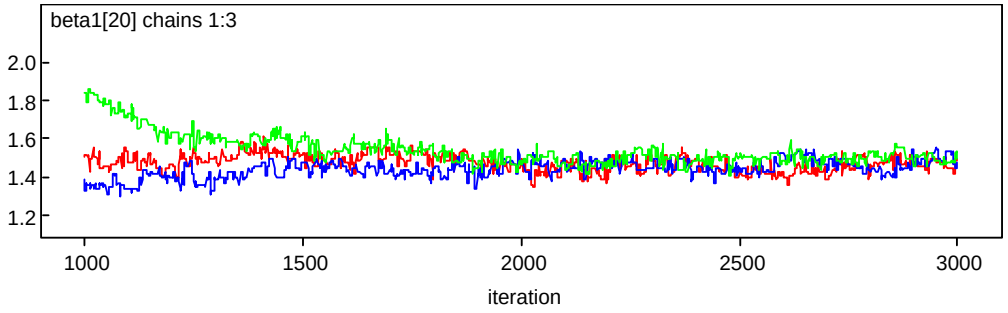

beta1[15] chains 1:3
2.0
1.8
1.6
1.4
1.2
1000
1500
2000
2500
3000
iteration

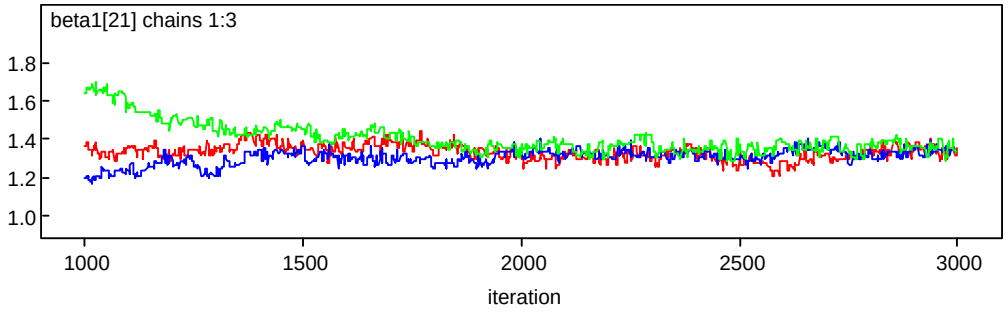

beta1[16] chains 1:3
2.0
1.8
1.6
1.4
1.2
1000
1500
2000
2500
3000
iteration


beta1[17] chains 1:3
2.0
1.8
1.6
1.4
1.2
1000
1500
2000
2500
3000
iteration


beta1[18] chains 1:3
1.8
1.6
1.4
1.2
1.0
1000
1500
2000
2500
3000
iteration


beta1[19] chains 1:3
2.0
1.8
1.6
1.4
1.2
1000
1500
2000
2500
3000
iteration


beta1[20] chains 1:3
2.0
1.8
1.6
1.4
1.2
1000
1500
2000
2500
3000
iteration


beta1[21] chains 1:3
1.8
1.6
1.4
1.2
1.0
1000
1500
2000
2500
3000
iteration

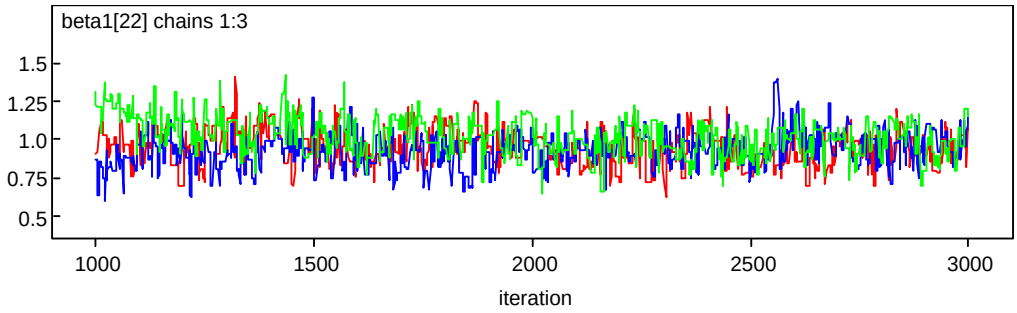
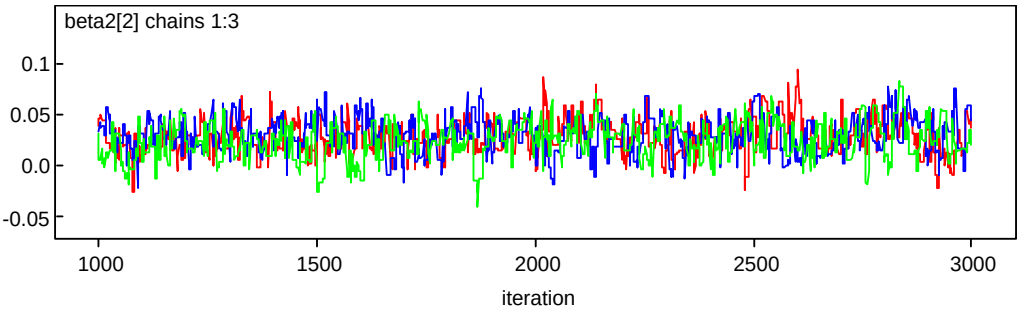
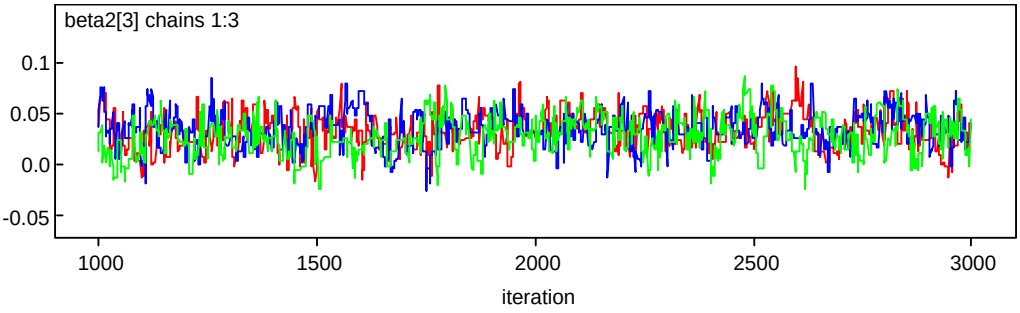
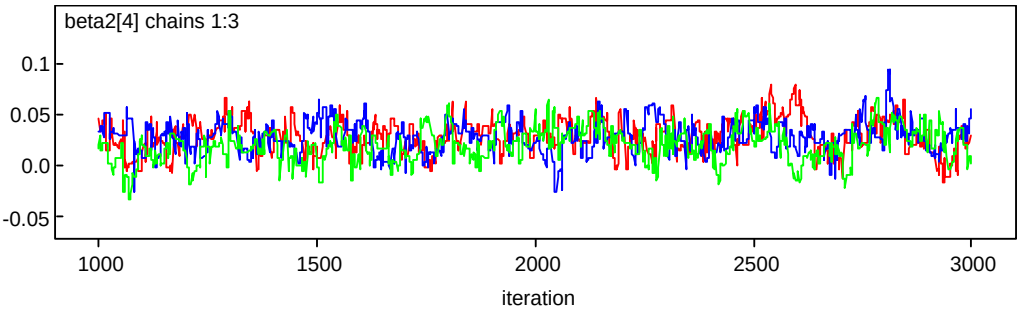


Figure 21 : *Les chaînes Markov de* β_2







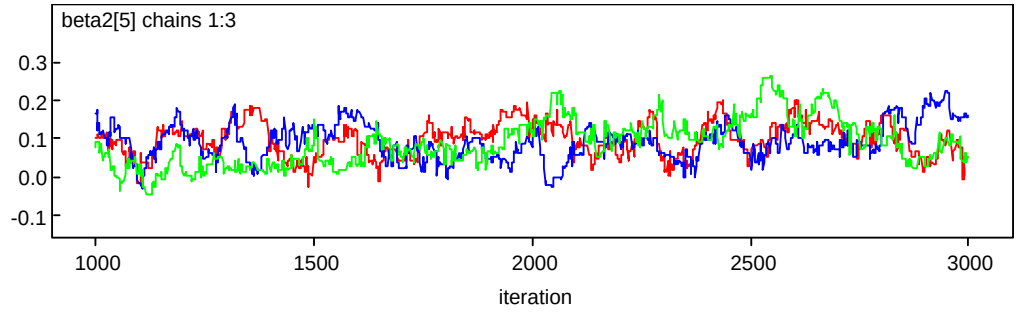


Figure 22 : *Les chaînes Markov de* β_3

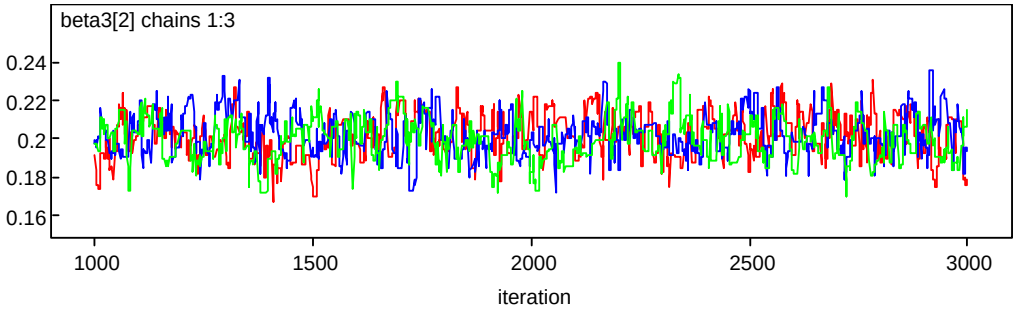
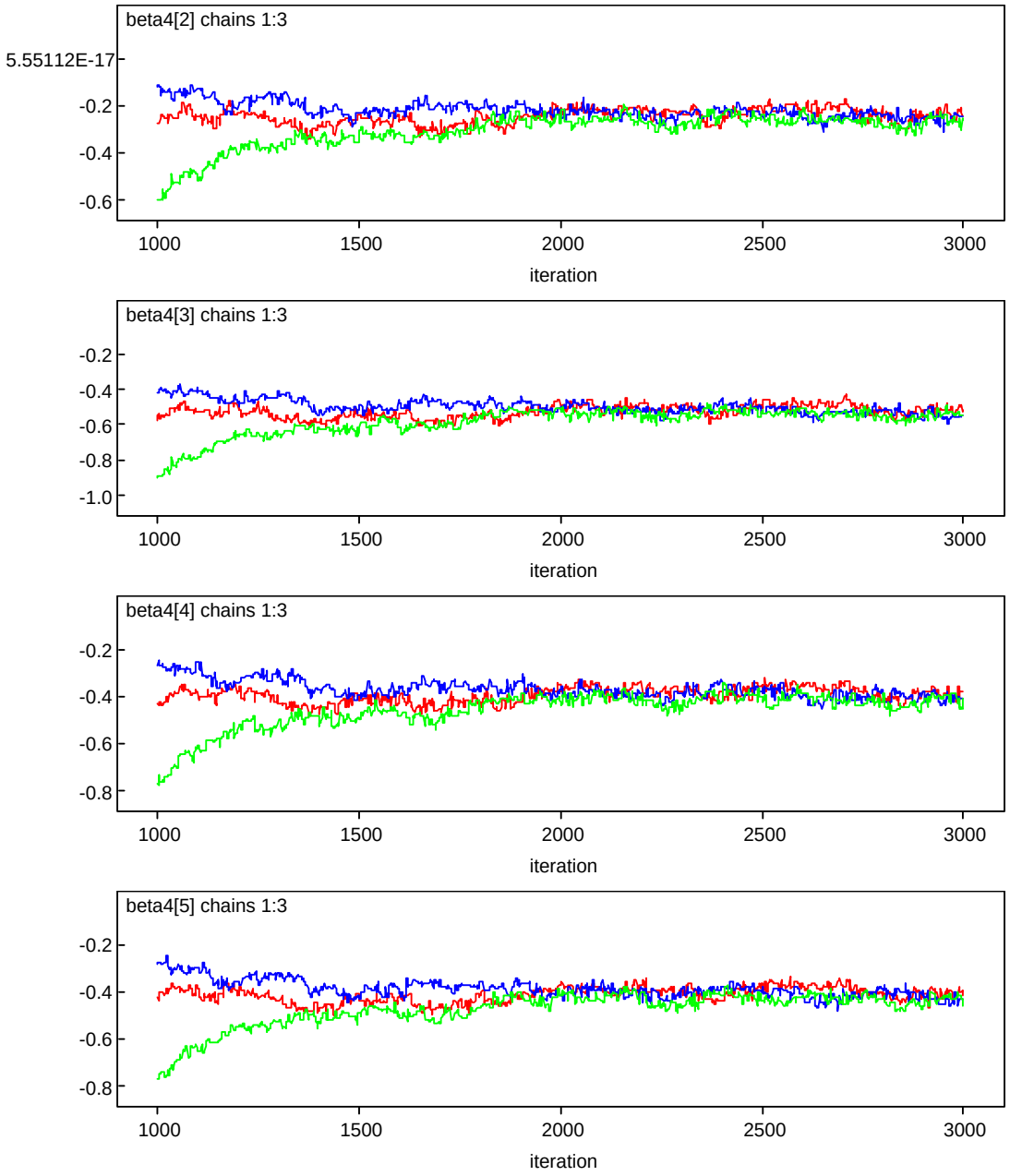
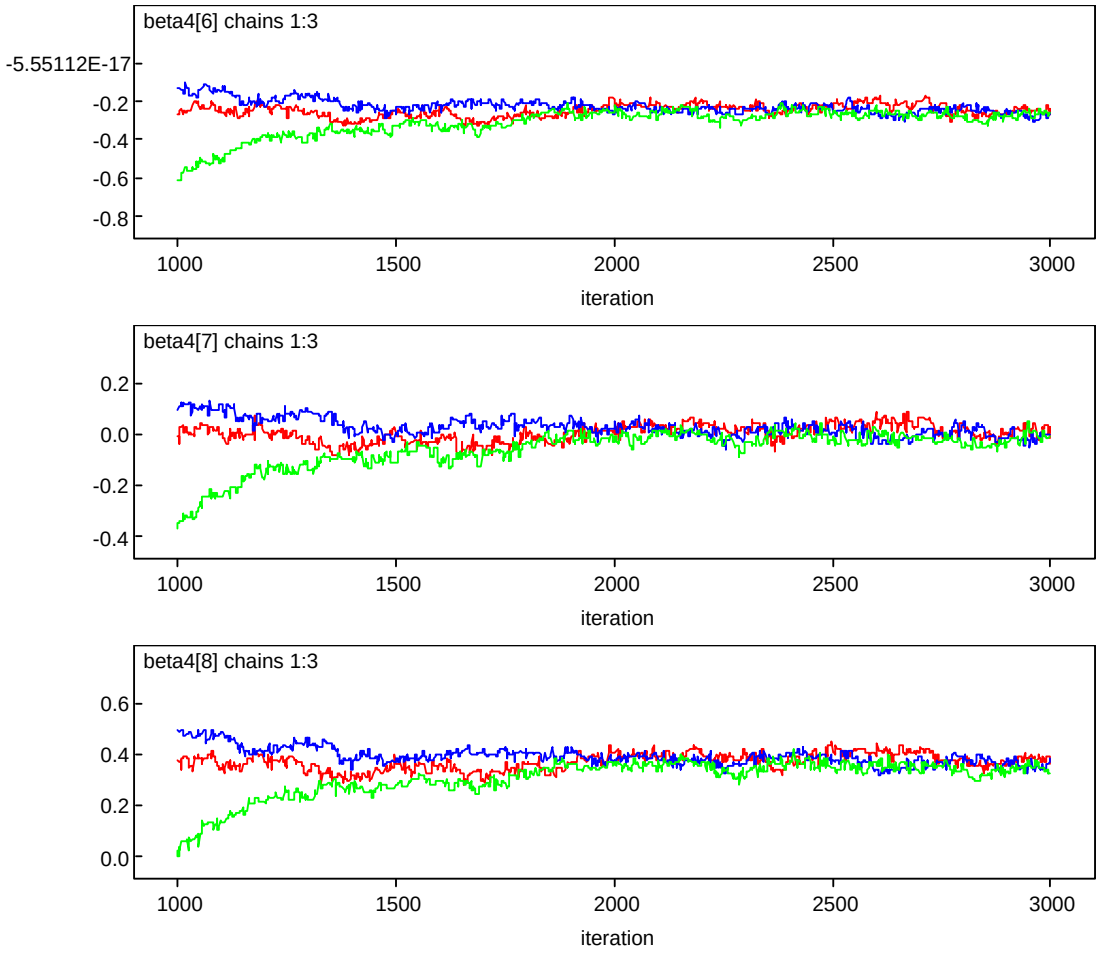


Figure 23 : *Les chaînes de Markov de* β_4





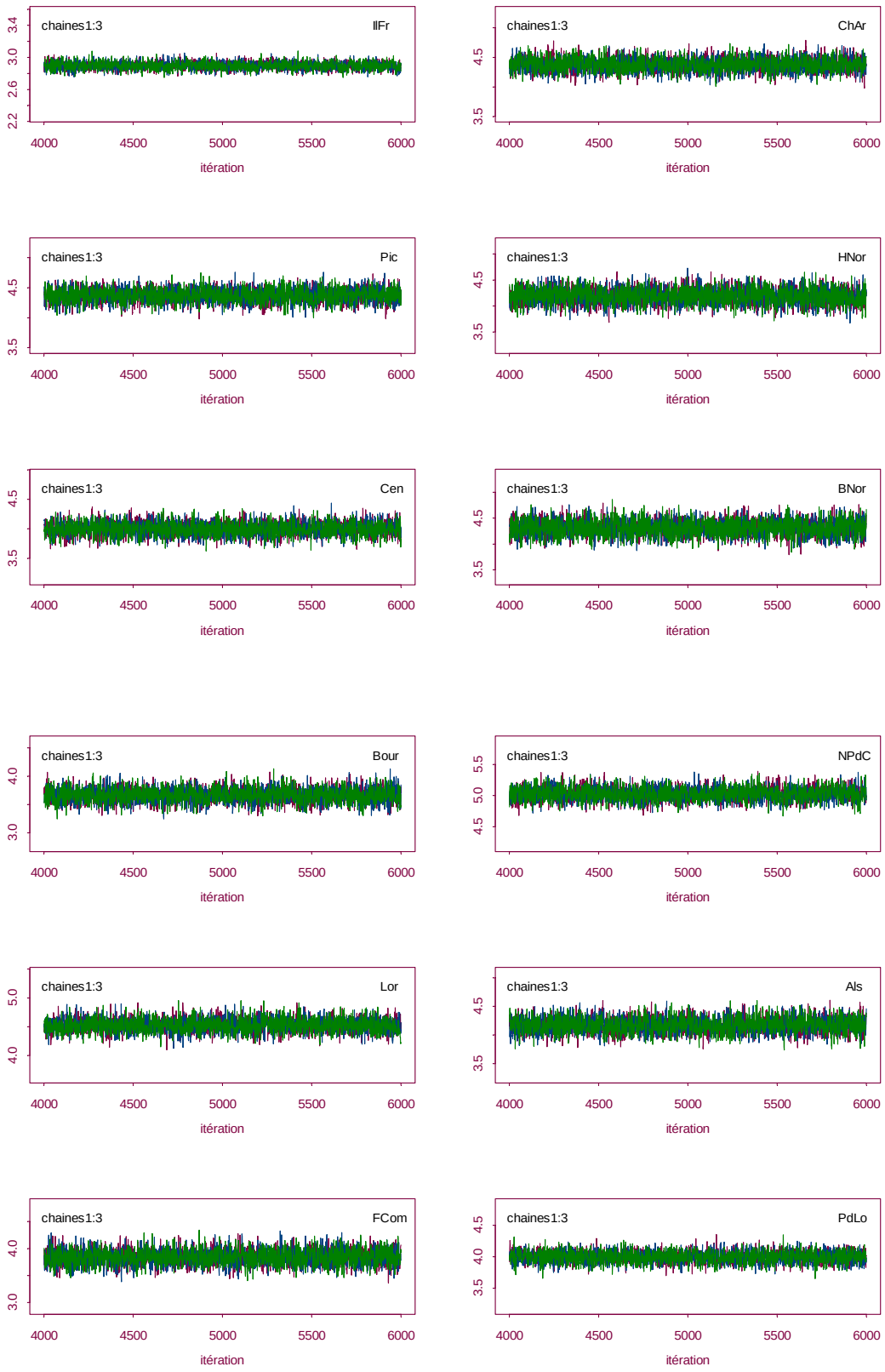
A partir des formules (22) et (24) on peut identifier des chaines Markov pour la moyenne de chaque région $\{\mu_i^g\}$ et pour celle de la France $\{\mu^g\}$:

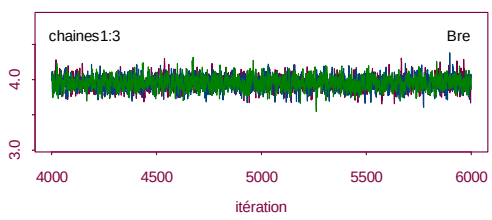
$$\mu_i^g = \frac{1}{N_i}[\sum_j \sum_s \sum_k \sum_{l \in obs_i} y_{ijskl} + \sum_j \sum_s \sum_k (N_{ijsk} - n_{ijsk})\mu_{ijsk}^g]$$

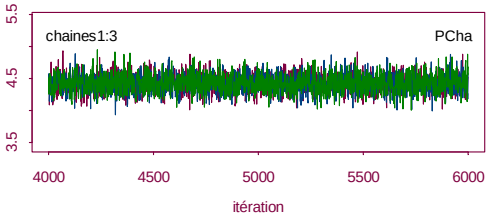
$$\mu^g = \frac{1}{N}[\sum_i \sum_j \sum_s \sum_k \sum_{l \in obs_i} y_{ijskl} + \sum_i \sum_j \sum_s \sum_k (N_{ijsk} - n_{ijsk})\mu_{ijsk}^g]$$

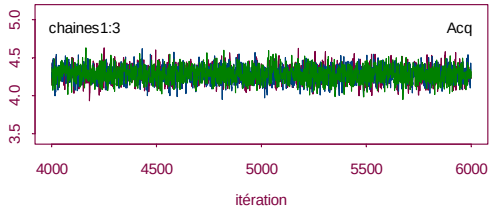
On pourrait vérifier directement que ces chaînes ont convergé en les représentant graphiquement. Dans la figure 24 se trouvent leurs chaînes Markov à partir de l'itération 4000. On peut constater que les valeurs des trois chaînes ne peuvent pas être distinguées, donc elles proviennent d'une même distribution, la distribution a posteriori.

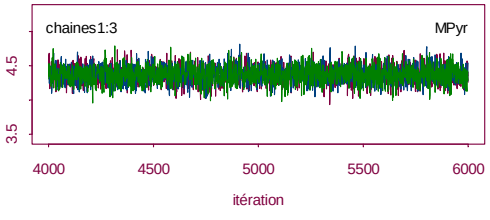
Figure 24: *Les chaînes Markov des moyennes des régions et de la France*

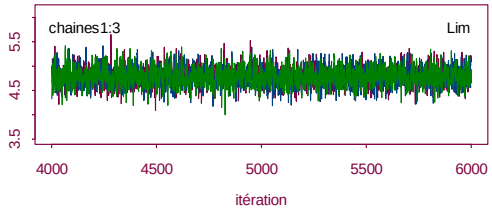


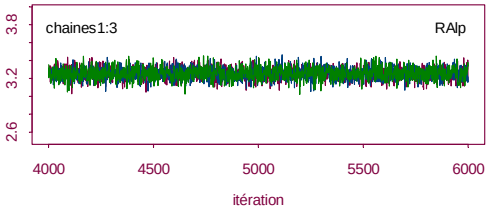

chaines1:3
Bre
4.0
3.0
4000
4500
5000
5500
6000
itération

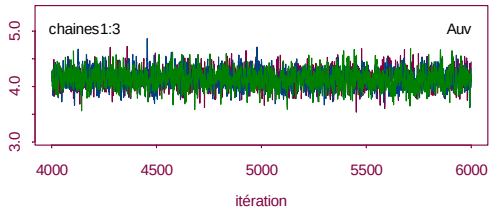

chaines1:3
PCha
5.5
4.5
3.5
4000
4500
5000
5500
6000
itération

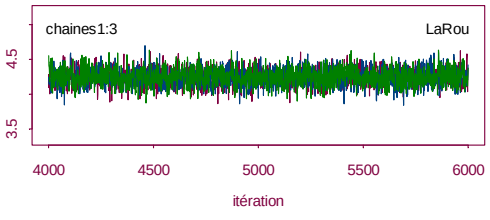

chaines1:3
Acq
5.0
4.5
4.0
3.5
4000
4500
5000
5500
6000
itération

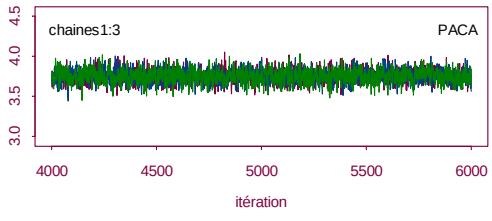

chaines1:3
MPyr
4.5
3.5
4000
4500
5000
5500
6000
itération

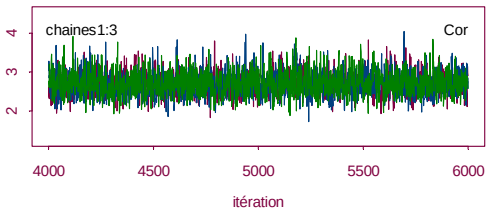

chaines1:3
Lim
5.5
4.5
3.5
4000
4500
5000
5500
6000
itération

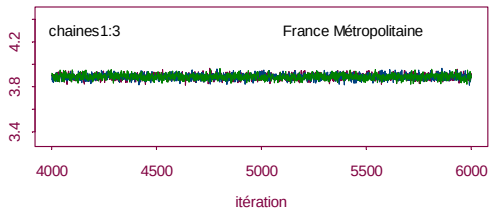

chaines1:3
RAlp
3.8
3.2
2.6
4000
4500
5000
5500
6000
itération


chaines1:3
Auv
5.0
4.0
3.0
4000
4500
5000
5500
6000
itération


chaines1:3
LaRou
4.5
3.5
4000
4500
5000
5500
6000
itération

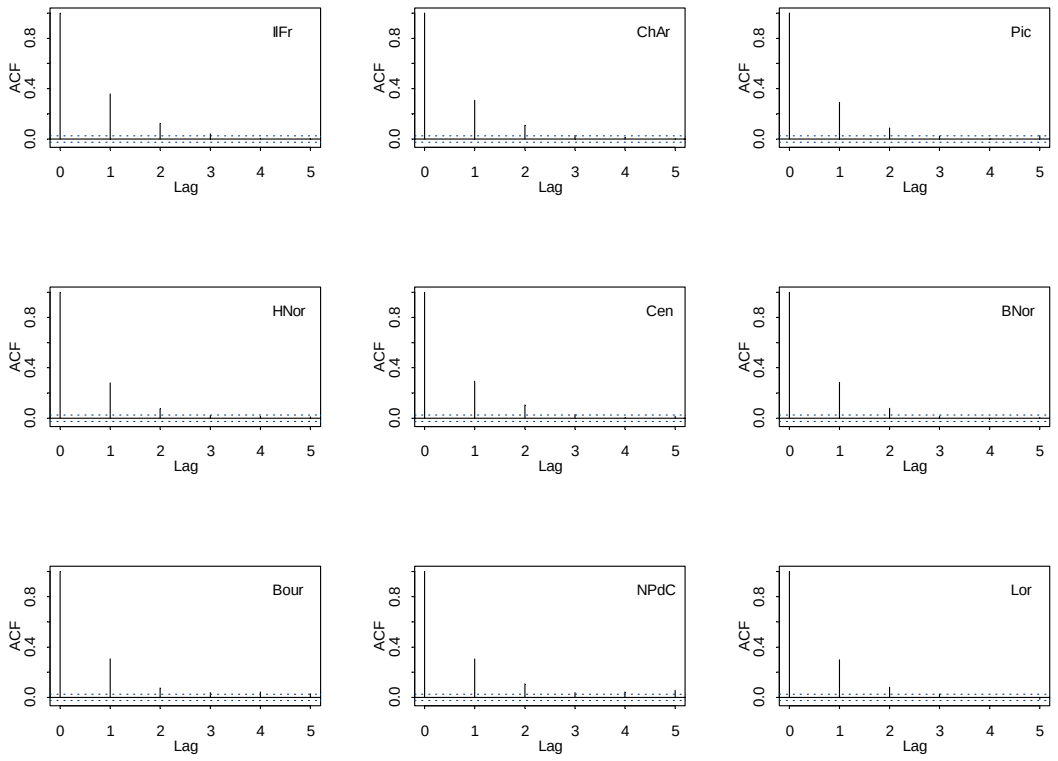

chaines1:3
PACA
4.5
4.0
3.5
3.0
4000
4500
5000
5500
6000
itération

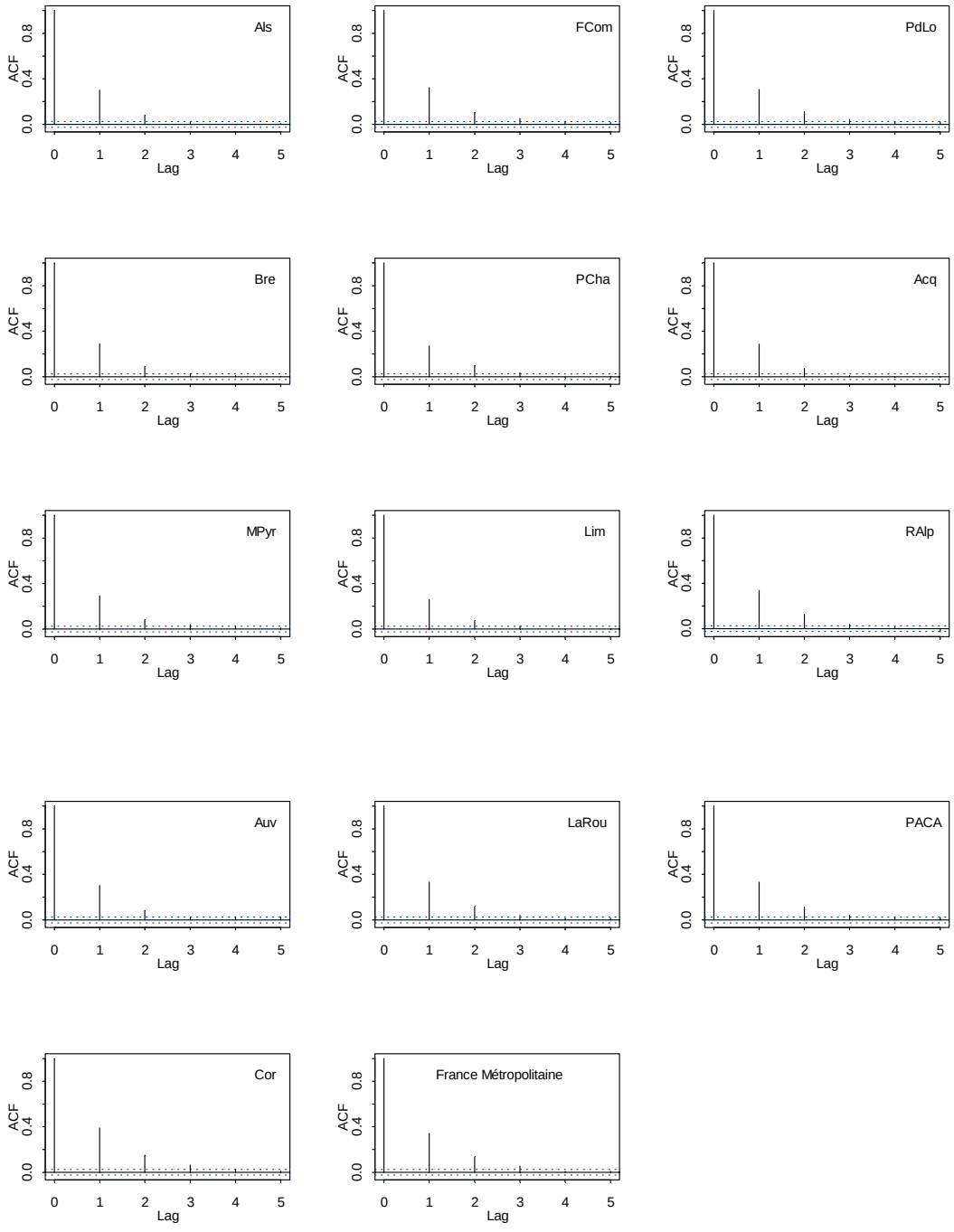

chaines1:3
Cor
4
3
2
4000
4500
5000
5500
6000
itération


chaines1:3
France Métropolitaine
4.2
3.8
3.4
4000
4500
5000
5500
6000
itération

D'ailleurs les chaînes Markov des μ_i et μ convergent plus vite que celles des paramètres utilisés pour calculer $\hat{\mu}_i$ et $\hat{\mu}$. Ceci est dû à l'autocorrélation qui est plus faible dans le cas des moyennes régionales et de la France (à partir du lag 3 elle devient négligeable comme on peut le constater dans la figure 25 ; pour les paramètres *alpha*, *kappa* et les β l'autocorrélation diminue plus lentement, elle étant importante même pour des lag d'ordre 20). Le coefficient d'autocorrélation d'une chaîne Markov détermine directement le nombre d'itérations qu'il faut considérer pour approximer les quantités a posteriori avec une bonne précision. Nous allons voir ceci plus bas.

Figure 25 : *Fonctions d'autocorrélation des chaînes Markov des régions et de la France*




Als
FCom
PdLo
Bre
PCha
Acq
MPyr
Lim
RAlp
Auv
LaRou
PACA
Cor
France Métropolitaine
ACF
0.0
0.4
0.8
0
1
2
3
4
5
Lag

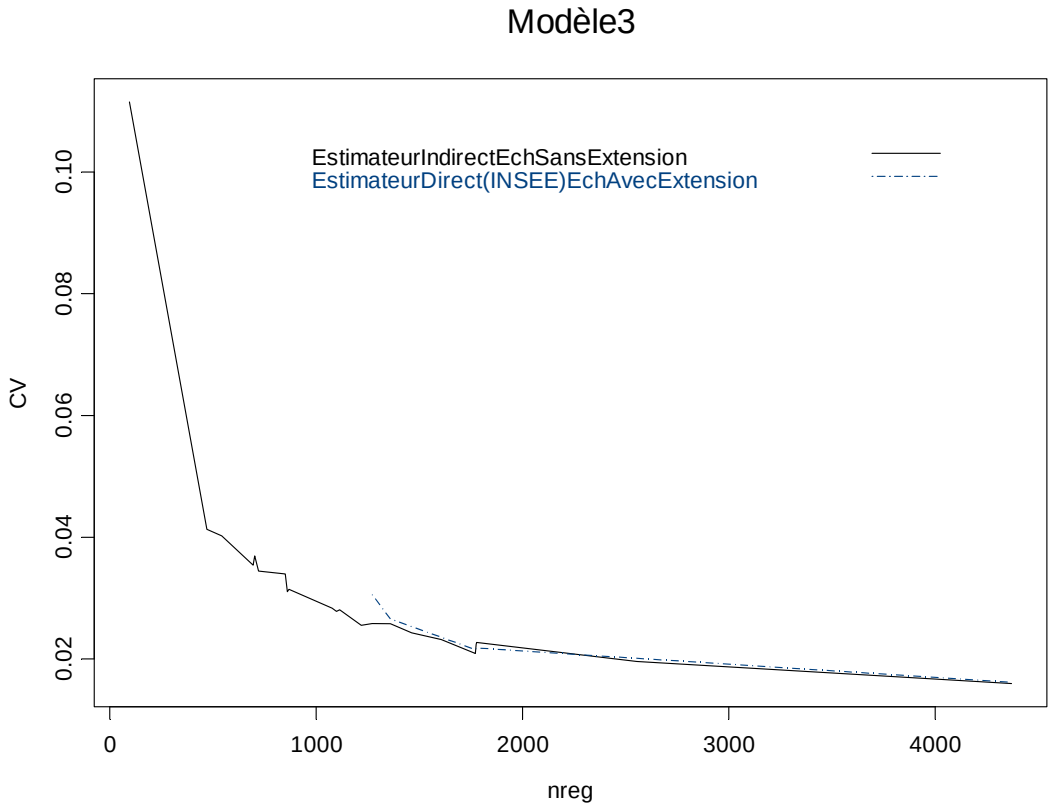
Etant donnés les graphiques ci-dessus, on peut conclure que les trois chaînes Markov ont convergé. On peut donc utiliser leurs valeurs pour calculer différents paramètres des distributions a posteriori des μ_i et μ. Pour chacune des trois chaînes nous avons réalisé 6000 itérations, avec une période de « burn-in » de 4000 itérations. Nous avons donc utilisé pour obtenir les estimations les dernières 2000 itérations de chaque chaîne, à savoir au total 6000 itérations. On rappelle que les estimations utilisent des effectifs N_{ijsk} estimés, non pas les vrais effectifs N_{ijsk}. Dans le tableau 2 se trouvent les moyennes et les écart-types a posteriori calculés en utilisant (22) et (24) pour les moyennes, respectivement (23) et (25) pour les écart-types.

Tableau 2:*Estimations à partir de l'échantillon sans extension (28259 observations)*

Région (*) = région à extension	Estimation	Ecart-Type (formules (23) et (25))
Ile de France(*)	2.89	0.0462
Champagne-Ardenne(*)	4.37	0.1128
Picardie(*)	4.36	0.1128
Haute-Normandie	4.18	0.1481
Centre	4.00	0.1124
Basse-Normandie	4.30	0.1483
Bourgogne	3.66	0.1244
Nord Pas de Calais(*)	5.02	0.1050
Loraine	4.52	0.1154
Alsace	4.16	0.1310
Franche Compté	3.83	0.1417
Pays de la Loire	3.99	0.0896
Bretagne	3.95	0.0959
Poitou Charente	4.42	0.1373
Aquitaine	4.28	0.0993
Midi-Pyrénées	4.37	0.1216
Limousin	4.77	0.1974
Rhône Alpes	3.25	0.0638
Auvergne	4.13	0.1663
Languedoc-Roussillon	4.24	0.1202
PACA(*)	3.74	0.0851
Corse	2.70	0.3011
France Métropolitaine	**3.88**	**0.0227**

La figure 26 représente le coefficient de variation en fonction de la taille de l'échantillon régional. La ligne en pointillé représente les coefficients de variation déterminés à partir des estimations INSEE pour les cinq régions qui ont bénéficié d'extensions, calculés à partir de l'échantillon avec extension. On voit que les estimateurs indirects basés sur le modèle 3 et calculés à partir de l'échantillon sans extension sont tout aussi bons ou meilleurs que les estimateurs directs mais calculés à partir de l'échantillon avec extension. La conclusion est qu'en utilisant les méthodes d'estimation petits domaines, notamment la mise au point d'un modèle, on peut arriver aux mêmes performances que la théorie classique des sondages sans prélever un échantillon supplémentaire.

Figure 26: *Coefficient de Variation vs Taille de l'échantillon régional*



A partir des chaînes Markov on peut calculer d'autres paramètres, non seulement la moyenne. Dans le tableau 3 se trouvent, pour chaque région et la France Métropolitaine, les écart-types des chaînes μ_i^g, les précisions Monte Carlo pour les moyennes et les écart-types a posteriori (voir plus bas les formules utilisées pour leur calcul) et les quantiles d'ordre 0.025, 0.5 et 0.975 pour les moyennes a posteriori (avec les deux quantiles on pourra calculer un intervalle de confiance d'ordre 0.95 pour les moyennes a posteriori).

Tableau 3:*Estimations à partir de l'échantillon sans extension (28259 observations)*

Région (*)=région à extension	Ecart-Type $\hat{\sigma}_i$	MCSE($\hat{\mu}_i$) (MCSE($\hat{\sigma}_i$))	Quantile 0.025	Quantile 0.5	Quantile 0.975
IledeFrance(*)	0.0462	0.0008(0.0004)	2.80	2.89	2.98
Champagne-Ardenne(*)	0.1127	0.0019(0.0011)	4.16	4.37	4.59
Picardie(*)	0.1127	0.0019(0.0011)	4.14	4.36	4.58
Haute-Normandie	0.1480	0.0025(0.0014)	3.89	4.18	4.47
Centre	0.1123	0.0019(0.0011)	3.78	4.00	4.21
Basse-Normandie	0.1483	0.0025(0.0014)	4.02	4.30	4.59
Bourgogne	0.1244	0.0022(0.0012)	3.41	3.66	3.91
NordPasdeCalais(*)	0.1050	0.0018(0.0010)	4.82	5.02	5.23
Loraine	0.1154	0.0020(0.0011)	4.30	4.52	4.75
Alsace	0.1310	0.0023(0.0013)	3.92	4.16	4.43
FrancheCompté	0.1417	0.0025(0.0014)	3.57	3.83	4.12
PaysdelaLoire	0.0895	0.0015(0.0008)	3.81	3.99	4.16
Bretagne	0.0959	0.0016(0.0009)	3.76	3.95	4.13
PoitouCharente	0.1373	0.0023(0.0013)	4.16	4.42	4.70
Aquitaine	0.0993	0.0017(0.0009)	4.09	4.28	4.47
Midi-Pyrénées	0.1216	0.0021(0.0012)	4.13	4.36	4.61
Limousin	0.1973	0.0033(0.0019)	4.40	4.77	5.18
RhôneAlpes	0.0638	0.0011(0.0006)	3.12	3.25	3.37
Auvergne	0.1662	0.0029(0.0016)	3.81	4.13	4.46
Languedoc-Roussillon	0.1201	0.0021(0.0012)	4.01	4.23	4.48
PACA(*)	0.0851	0.0015(0.0008)	3.58	3.74	3.91
Corse	0.3010	0.0058(0.0032)	2.16	2.68	3.33
FranceMétropolitaine	**0.0227**	**0.0004(0.0002)**	**3.84**	**3.88**	**3.93**

Remarquons d'abord que les estimations de l'écart-type a posteriori calculées comme écart-type des chaînes Markov $\{\mu_i^g\}$ et $\{\mu^g\}$ (colonne 2 du tableau 3) et celles utilisant les formules (23) ou (25) (colonne 3 du tableau 2) sont identiques.

Pour estimer un paramètre a posteriori, on utilise une chaîne Markov d'une certaine longueur. La précision Monte Carlo est la valeur qui montre l'erreur de la chaîne Markov dans l'estimation du paramètre. Par exemple, le fait que la précision Monte Carlo de la moyenne a posteriori de l'Ile de France est égale à 0.0008 veut dire que la vraie valeur de la moyenne a posteriori $E(\mu_i \mid \mathbf{y}_{obs})$ (et non pas de la moyenne μ_i de l'Ile de France) se trouvent dans l'intervalle [2.89-1.96*0.0008 ;2.89+1.96*0.0008] avec une probabilité égale à 0.975. Pour les écart-types on a la même interprétation.

Clairement, on a donc ici deux processus d'estimation, ou plutôt pour les différencier, on a une estimation et une approximation. La moyenne μ_i est estimée par sa moyenne a posteriori résultant du modèle 3, et la moyenne a posteriori est approximée par la moyenne de la chaîne Markov $\{\mu_i^g\}$. La précision de l'approximation est l'erreur Monte Carlo, alors que la précision de l'estimation est l'écart-type a posteriori.

Si le paramètre envisagé est la moyenne a posteriori et si la chaîne est un processus autorégressif d'ordre 1, alors la précision Monte Carlo se calcule selon :

$$\mathrm{MCSE}(\hat{\mu}) = \frac{\hat{\sigma}}{\sqrt{G}}\sqrt{\frac{1+\hat{\rho}_1}{1-\hat{\rho}_1}}$$

où $\hat{\sigma}$ est l'écart-type de la chaîne, $\hat{\rho}_1$ est l'autocorrélation d'ordre 1 de la chaîne et G est la longueur de la chaîne. Si le paramètre est l'écart-type a posteriori, dans les mêmes conditions, l'erreur Monte Carlo se calcule selon :

$$\mathrm{MCSE}(\hat{\sigma}) = \frac{\hat{\sigma}}{\sqrt{2G}}\sqrt{\frac{1+\hat{\rho}_1^2}{1-\hat{\rho}_1^2}}$$

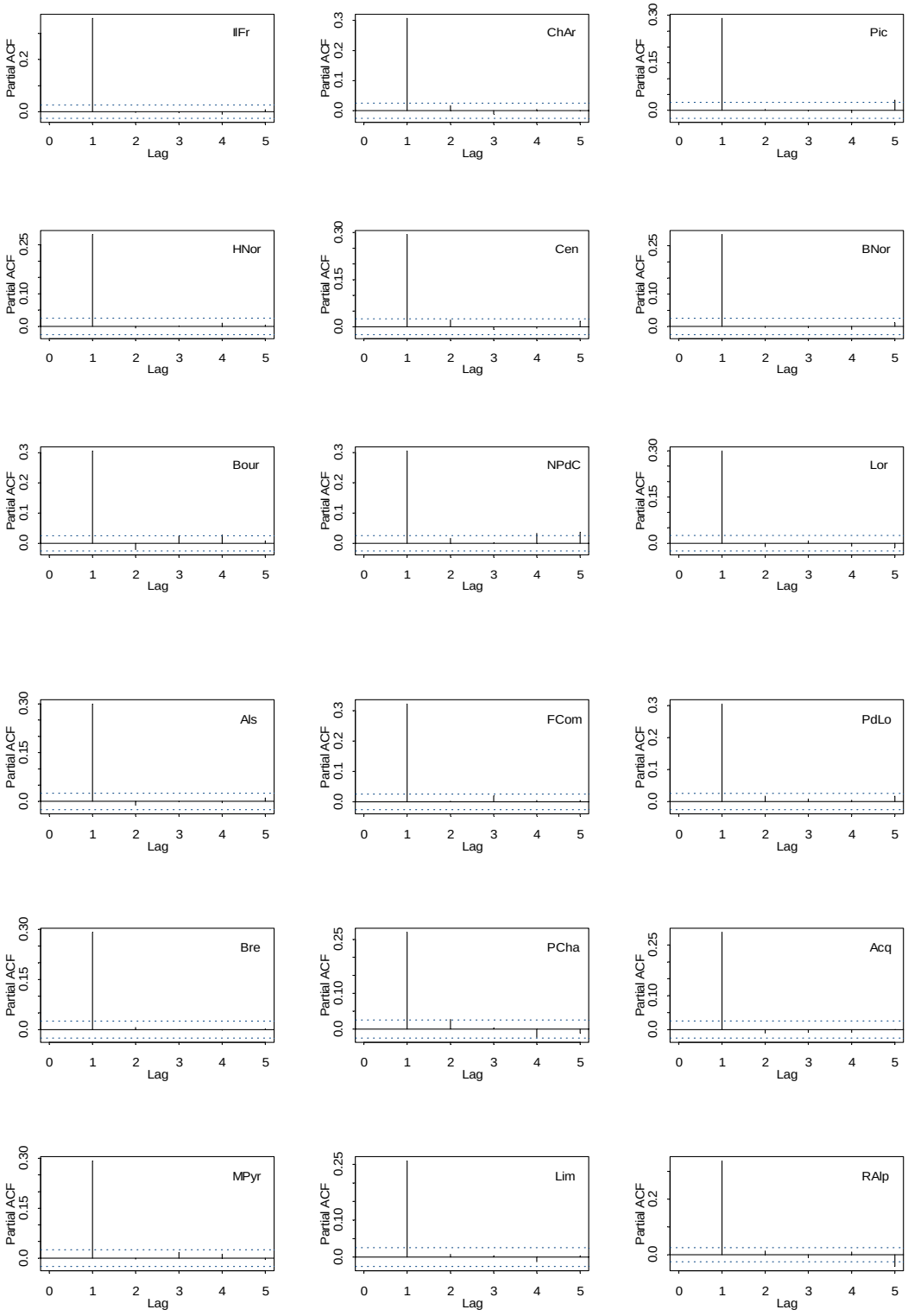
Des deux formules ci-dessus on voit que plus l'autocorrélation est grande plus la longueur G doit être grande pour arriver à la même précision Monte Carlo (si $\rho_1 = 0$ alors on arrive aux formules bien connues d'un échantillonnage iid ; malheureusement ce n'est pas le cas des chaînes Markov qui par définition échantillonnent des valeurs dépendantes).

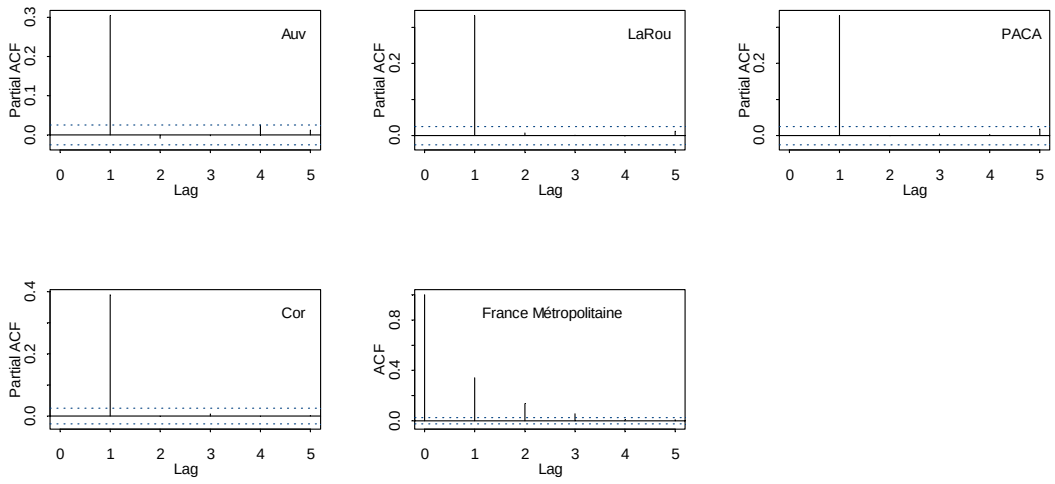
Du tableau 3 on voit que les erreurs Monte Carlo fournissent des approximations à un décimale près pour les moyennes a posteriori (sauf pour la Corse) et en général à deux décimales près pour les écart-types a posteriori. On peut donc remarquer à partir des valeurs calculées mais aussi des deux formules ci-dessus qu'un écart-type a posteriori s'estime avec une précision Monte Carlo plus grande qu'une moyenne a posteriori.

Nous avons vu plus haut que les précisions estimées comme écart-type des chaînes Markov est celle estimées utilisant les formules (23) et (25) sont identiques, donc les précisions Monte Carlo des écart-types calculées dans le tableau 3 peuvent être considérées comme précisions Monte Carlo des estimations utilisant (23) et (25), ceci faute de savoir comment on pourrait dériver de telles précisions pour (23) et (25) étant donnée qu'elles ne sont ni moyennes ni variances d'une chaîne, mais un mix des deux.

Il reste bien sûr vérifier que les chaînes Markov des régions et de la France sont des processus autorégressifs d'ordre 1. Pour voir ça, les fonctions d'autocorrélation partielle sont l'outil idéal. Pour un lag d donné, elles calculent la corrélation entre μ_i^g et μ_i^{g+d} après avoir éliminé l'effet des autocorrélations d'ordre 1, 2,…,d-1. Donc, au cas d'un processus AR(1), c'est seulement le coefficient d'autocorrélation partielle d'ordre 1 qui est significativement différent de zéro. Ces fonctions sont représentées dans la figure 27 :

Figure 27:*Fonctions d'autocorrélation partielle des chaînes régionales et de la France*





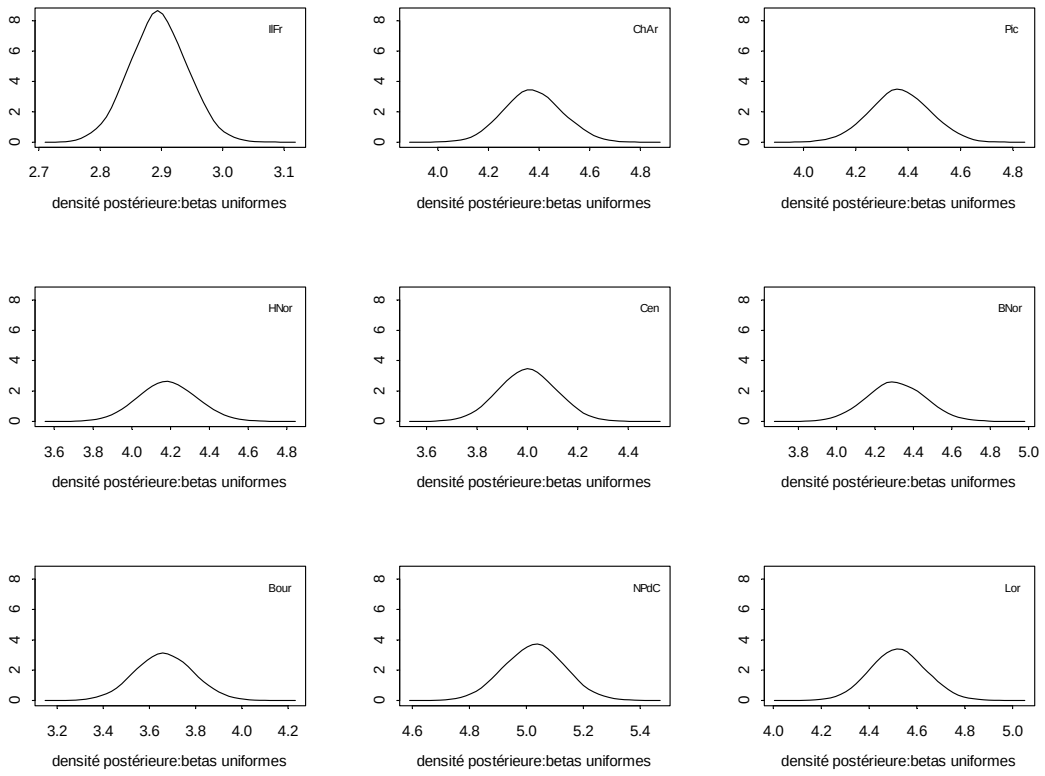
Les lignes en pointillé qui se trouvent en bas de chaque graphique représentent un intervalle de confiance en dehors duquel un coefficient d'autocorrélation partielle est significativement différent de zéro. On voit que tous les régions présentent seulement un seul coefficient d'autocorrélation partielle significatif – celui d'ordre 1. Donc, leurs chaînes Markov sont des AR(1). Pour la France Métropolitaine celui d'ordre 2 semble tomber en dehors de l'intervalle mais il n'est pas trop important.

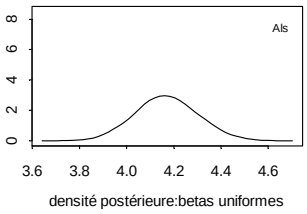
Pour le modèle 3 nous avons choisi comme densités a priori pour les β ainsi que *alpha* et *kappa* des lois uniformes sur des intervalles suffisamment grands pour marquer l'absence d'information a priori sur tous ces paramètres. Toute étude doit être accompagnée d'une analyse de sensibilité par rapport aux lois a priori des hyper paramètres du modèle.

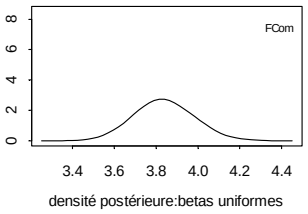
La loi a posteriori d'un paramètre résulte de la combinaison par le théorème de Bayes entre sa fonction de vraisemblance et sa loi a priori. Si les données ne contiennent pas suffisamment d'information sur le paramètre, alors sa loi a posteriori sera déterminée principalement par la loi a priori. Dans le cas contraire, la loi a priori aura peu d'influence sur la loi a posteriori et donc sur les estimations basées sur celle-ci. En général, avec un échantillon de taille importante (comme c'est le cas de l'échantillon sans extension), on est plutôt dans la deuxième situation. On s'attend donc que les lois a priori n'influence pas les estimations.

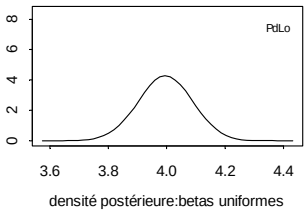
Nous avons réalisé une analyse de sensibilité seulement par rapport aux lois a priori des paramètres β pour lesquels nous avons considéré des lois normales de moyennes zéro et de variance 1000 (la valeur importante de la variance est censée fournir une loi non informative). Comme les estimations sont calculées à partir des chaînes μ_i^g et μ^g, nous avons représenté dans les figures 28 et 29 les densités a posteriori de μ_i et μ estimées à partir de ces chaînes dans les deux cas : les β sont uniformes et les β sont normales:

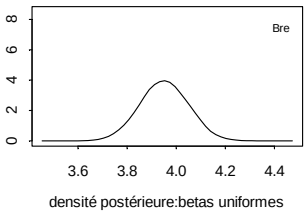
Figure 28 : *Densités postérieures des estimateurs des moyennes des régions et de la France: lois a priori de β uniformes*

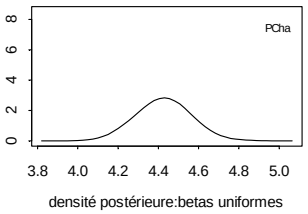


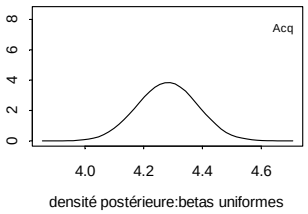

Als
densité postérieure:betas uniformes

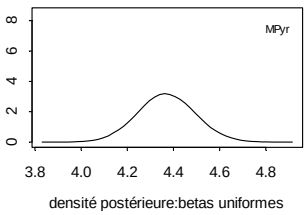

FCom
densité postérieure:betas uniformes

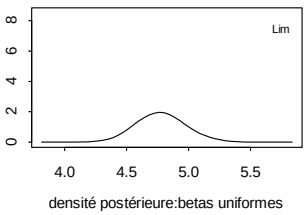

PdLo
densité postérieure:betas uniformes

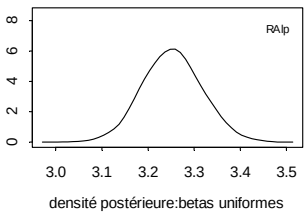

Bre
densité postérieure:betas uniformes

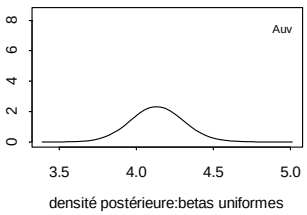

PCha
densité postérieure:betas uniformes

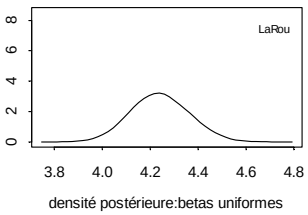

Acq
densité postérieure:betas uniformes

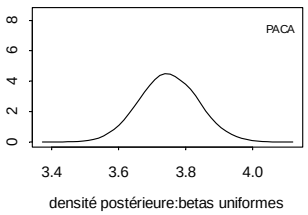

MPyr
densité postérieure:betas uniformes

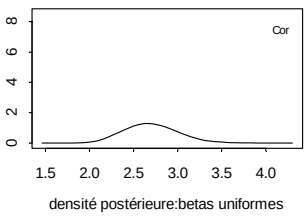

Lim
densité postérieure:betas uniformes


RAlp
densité postérieure:betas uniformes


Auv
densité postérieure:betas uniformes


LaRou
densité postérieure:betas uniformes


PACA
densité postérieure:betas uniformes


Cor
densité postérieure:betas uniformes

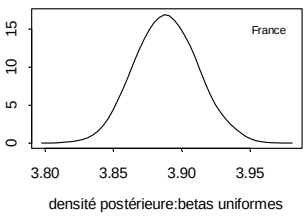
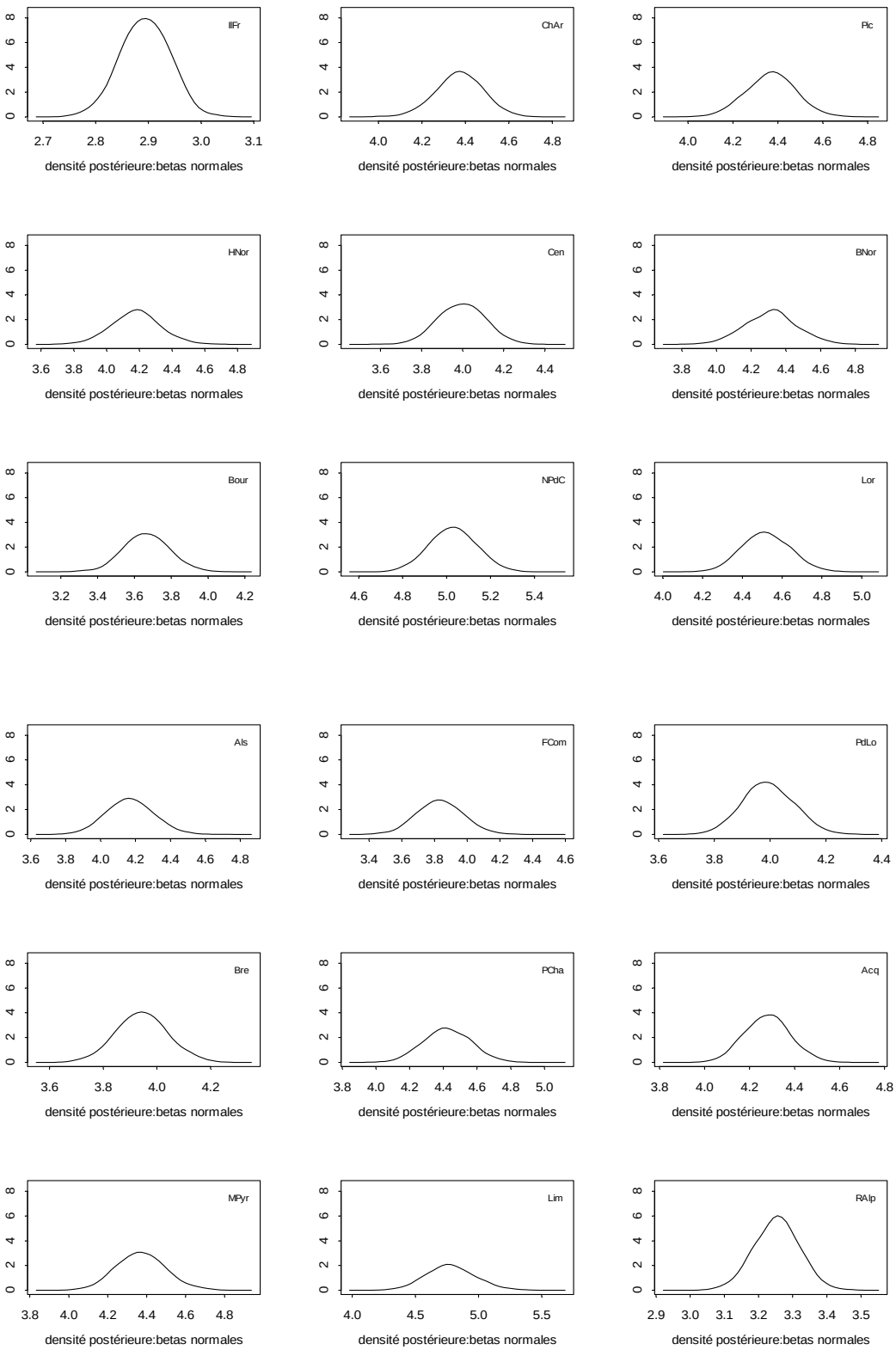
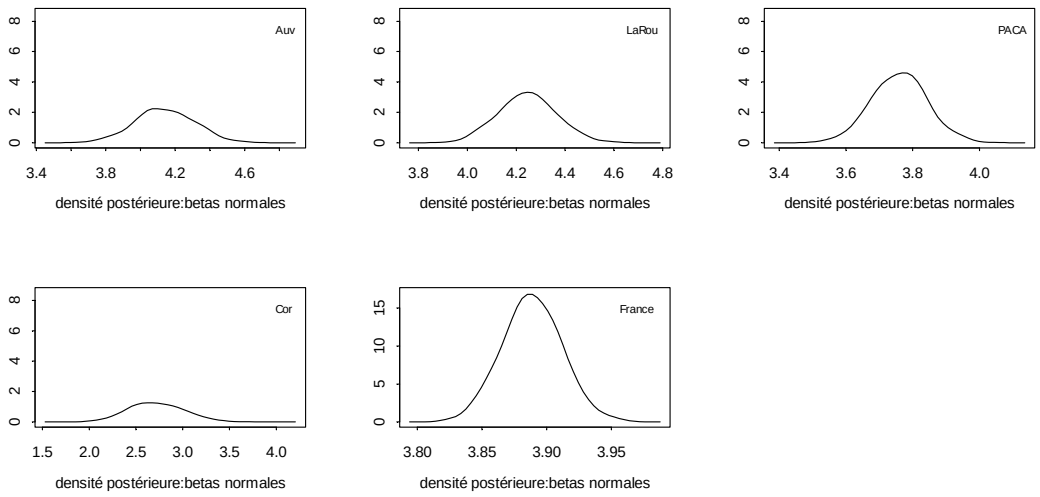

France
densité postérieure:betas uniformes

Figure 29 : *Densités postérieures des estimateurs des moyennes des régions et de la France : lois a priori de β normales*





En comparant on peut voir que les densités sont les mêmes dans les deux cas. D'ailleurs nous avons calculé les estimations dans le deuxième cas (nous n'avons pas inséré les valeurs dans ce document) et nous avons obtenue pratiquement les mêmes valeurs. Cependant, nous avons remarqué que dans le deuxième cas les coefficients d'autocorrélation des chaînes Markov sont le double (de l'ordre de 0.6 pour un lag égal à 1) que dans le cas des β uniformes. Ceci veut dire qu'il faut utiliser des chaînes plus longues pour avoir la même précision Monte Carlo.

6.2 Le cas de l'échantillon avec extension

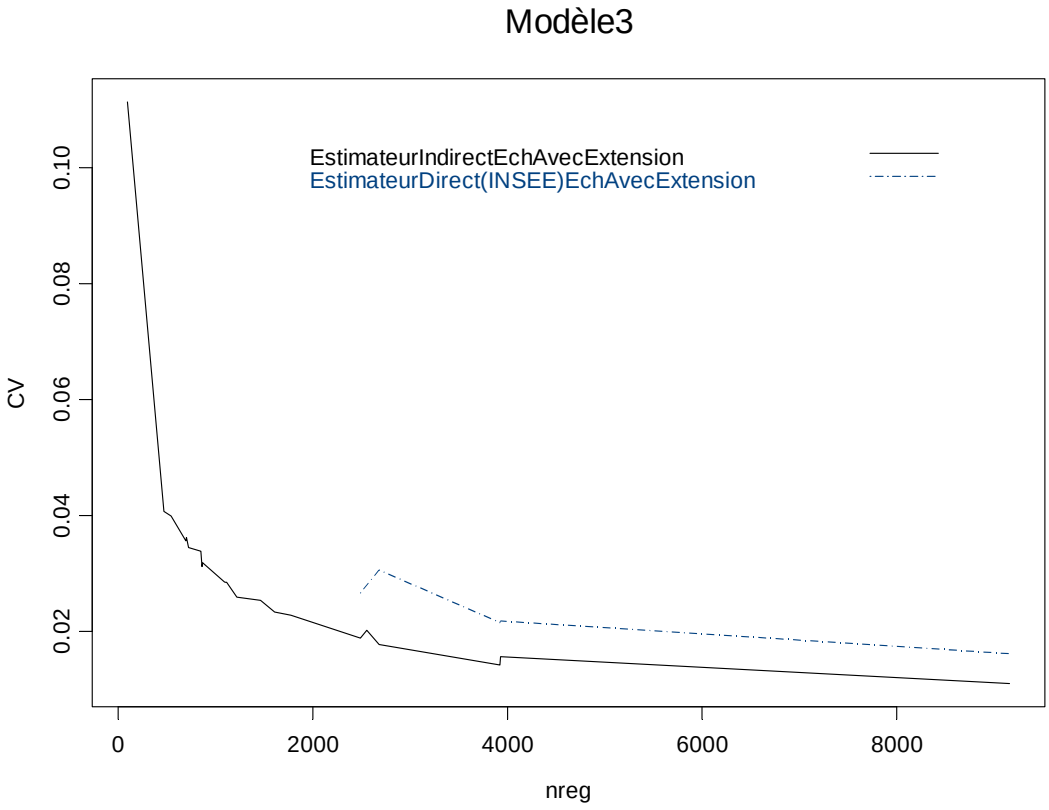
Nous avons suivi une démarche similaire dans le cas de l'échantillon avec extension. Après une période « burn-in » de 4000 itérations, nous avons utilisé les dernières 2000 itérations de chacune des trois chaînes (voir plus bas comment nous avons diagnostiqué la convergence). Le tableau 4 contient les estimations obtenues pour les moyennes des 22 régions et de la France ainsi que les estimations obtenues par l'INSEE pour les cinq régions à extension et la France :

Tableau 4 : *Estimations à partir de l'échantillon avec extension (39900 observations)*

Région (*) = région à extension	Estimation	Ecart-Type (formules (23) et (25))	Estimation INSEE	Ecart-Type INSEE
Ile de France(*)	3.00	0.0330	3	0.0485
Champagne-Ardenne(*)	4.41	0.0832	4.4	0.1169
Picardie(*)	4.36	0.0773	4.5	0.1376
Haute-Normandie	4.20	0.1498		
Centre	4.02	0.1145		
Basse-Normandie	4.31	0.1488		
Bourgogne	3.68	0.1248		
Nord Pas de Calais(*)	5.13	0.0729	5.4	0.1162
Loraine	4.53	0.1176		
Alsace	4.18	0.1331		
Franche Compté	3.85	0.1395		
Pays de la Loire	4.01	0.0916		
Bretagne	3.97	0.1006		
Poitou Charente	4.44	0.1386		
Aquitaine	4.30	0.1003		
Midi-Pyrénées	4.37	0.1247		
Limousin	4.80	0.1957		
Rhône Alpes	3.27	0.0660		
Auvergne	4.15	0.1657		
Languedoc-Roussillon	4.24	0.1212		
PACA(*)	3.88	0.0607	4	0.0872
Corse	2.71	0.3019		
France Métropolitaine	**3.93**	**0.0207**	**4**	**0.0303**

La figure 30 contient la relation entre le coefficient de variation et la taille de l'échantillon régional. La ligne en pointillé représente les coefficients de variation des estimations obtenues par l'INSEE pour les cinq régions qui ont bénéficié d'extensions. On observe qu'en appliquant les méthodes d'estimations pour petits domaines, on arrive à des coefficients de variation plus petits. Il serait intéressant d'appliquer la méthodologie INSEE pour toutes les régions et pour les deux échantillons (avec et sans extension) pour comparer les estimations directes basées sur le plan de sondage aux estimations indirectes basées sur le modèle 3 pour l'ensemble des régions.

Figure 30 : *Coefficient de Variation vs Taille de l'échantillon régional*



Si on compare les valeurs des écart-type des tableaux 2 et 4 on observe que dans les régions à extension le fait d'avoir plus d'individus a augmenté et la précision de l'estimation et la précision des approximations Monte Carlo. Dans les régions sans extension les différences de précision se trouvent au niveau de la troisième décimales et comme pour une telle approximation il faut un nombre d'itérations de l'ordre de quelques dizaines de milliers pour la plupart des régions, nous pouvons conclure que la précision des estimations dans ces régions n'est pas influencée.

Tableau 5:*Estimations à partir de l'échantillon avec extension (39900 observations)*

Région (*)=région à extension	Ecart-Type $\hat{\sigma}_i$	MCSE($\hat{\mu}_i$) (MCSE($\hat{\sigma}_i$))	Quantile 0.025	Quantile 0.5	Quantile 0.975
IledeFrance(*)	0.0330	0.0006(0.0003)	2.93	3.00	3.06
Champagne-Ardenne(*)	0.0831	0.0014(0.0008)	4.25	4.42	4.58
Picardie(*)	0.0773	0.0013(0.0007)	4.21	4.36	4.51
Haute-Normandie	0.1498	0.0025(0.0014)	3.91	4.20	4.49
Centre	0.1144	0.0019(0.0011)	3.80	4.01	4.24
Basse-Normandie	0.1488	0.0026(0.0014)	4.03	4.32	4.61
Bourgogne	0.1247	0.0022(0.0012)	3.45	3.68	3.93
NordPasdeCalais(*)	0.0728	0.0012(0.0007)	4.99	5.13	5.27
Loraine	0.1176	0.0020(0.0011)	4.31	4.53	4.76
Alsace	0.1331	0.0023(0.0013)	3.92	4.17	4.44
FrancheCompté	0.1395	0.0024(0.0013)	3.59	3.84	4.12
PaysdelaLoire	0.0915	0.0016(0.0009)	3.83	4.01	4.19
Bretagne	0.1006	0.0018(0.0010)	3.77	3.96	4.17
PoitouCharente	0.1385	0.0023(0.0013)	4.17	4.44	4.72
Aquitaine	0.1003	0.0017(0.0009)	4.10	4.29	4.50
Midi-Pyrénées	0.1247	0.0022(0.0012)	4.14	4.37	4.62
Limousin	0.1956	0.0034(0.0019)	4.42	4.80	5.18
RhôneAlpes	0.0660	0.0012(0.0006)	3.14	3.27	3.39
Auvergne	0.1657	0.0028(0.0016)	3.83	4.15	4.48
Languedoc-Roussillon	0.1212	0.0021(0.0012)	4.01	4.23	4.48
PACA(*)	0.0607	0.0010(0.0006)	3.76	3.88	4.00
Corse	0.3018	0.0057(0.0031)	2.16	2.69	3.34
FranceMétropolitaine	**0.0206**	**0.0003(0.0002)**	**3.89**	**3.93**	**3.97**

Dans la section 6.1 où nous avons étudié l'échantillon sans extension, nous avons visualisé la convergence des chaînes de Markov pour chaque paramètre d'intérêt en représentant graphiquement les valeurs des chaînes de Markov respectives. Dans la littérature, il existe des diagnostics de convergence qui évaluent à partir de quelle itération les valeurs de la chaîne proviennent de la distribution a posteriori (voir Cowles et Carlin (1996) pour une présentation détaillée de ces diagnostiques).

Pour cette section nous avons choisi le diagnostic de Gelman et Rubin (1992). C'est un diagnostic qui utilise plusieurs chaînes Markov pour calculer le « shrink factor ». Plus précisément, soit *L* le nombre de chaînes Markov qui ont 2*d* valeurs chacune. Soient *B* et *W* la variance entre les chaînes et respectivement à l'intérieur des chaînes calculées à partir des dernières *d* itérations :

$$B = \frac{d}{L-1}\sum_{l=1}^{L}(\bar{h}_{l\cdot} - \bar{h}_{\cdot\cdot})^2 \ , \ W = \frac{1}{L(d-1)}\sum_{l=1}^{L}\sum_{k=d+1}^{2d}(h_{lk} - \bar{h}_{l\cdot})^2$$

où

$$\bar{h}_{l\cdot} = \frac{1}{d}\sum_{k=d+1}^{2d} h_{lk} \text{ et } \bar{h}_{\cdot\cdot} = \frac{1}{L}\sum_{l=1}^{L}\bar{h}_{l\cdot}$$

Le « shrink factor » sera alors donné par :

$$R = \frac{(\frac{d-1}{d}W + \frac{1}{d}B)}{W}$$

Si les chaînes de Markov convergent, alors R doit être proche de 1. Si R est largement supérieur à 1, alors il faudra augmenter la valeur de d. Les figures 31-36 correspondent aux paramètres utilisés dans l'estimation des μ_i dont on veut savoir à partir de quelle itération leurs chaînes Markov respectives ont atteint la stationnarité. Nous avons utilisé L=3 chaînes chacune de longueur égale à 6000. Les chaînes ont été partagées en 50 intervalles de la manière suivante : le premier intervalle contient les itérations 1:50, le deuxième 1:50+120, le troisième 1:50+2*120, le quatrième 1:50+3*120, etc…Pour chaque segment on calcule le R en obtenant ainsi un échantillon de valeurs observées de R. A partir de cet échantillon on calcule la médiane et le quantile d'ordre 97.5 de R et on les représente en fonction de l'itération maximum du segment.

Pour le paramètre *alpha*, on peut observer que R s'est stabilisé autour de 1 à partir de l'itération 1200. Etant donné que les quantiles de R sont calculés à partir de la deuxième moitié de chaque segment, on peut conclure que la convergence a été atteinte autour de l'itération 600. Les conclusions sont similaires pour les autres paramètres.

Figure 31

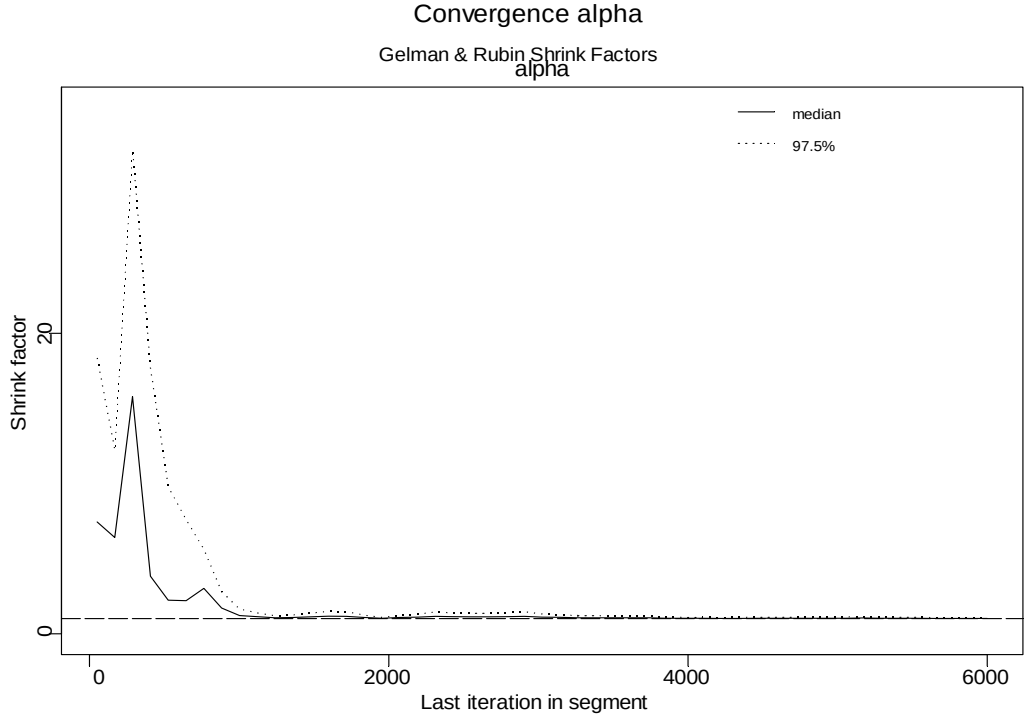


Figure 32

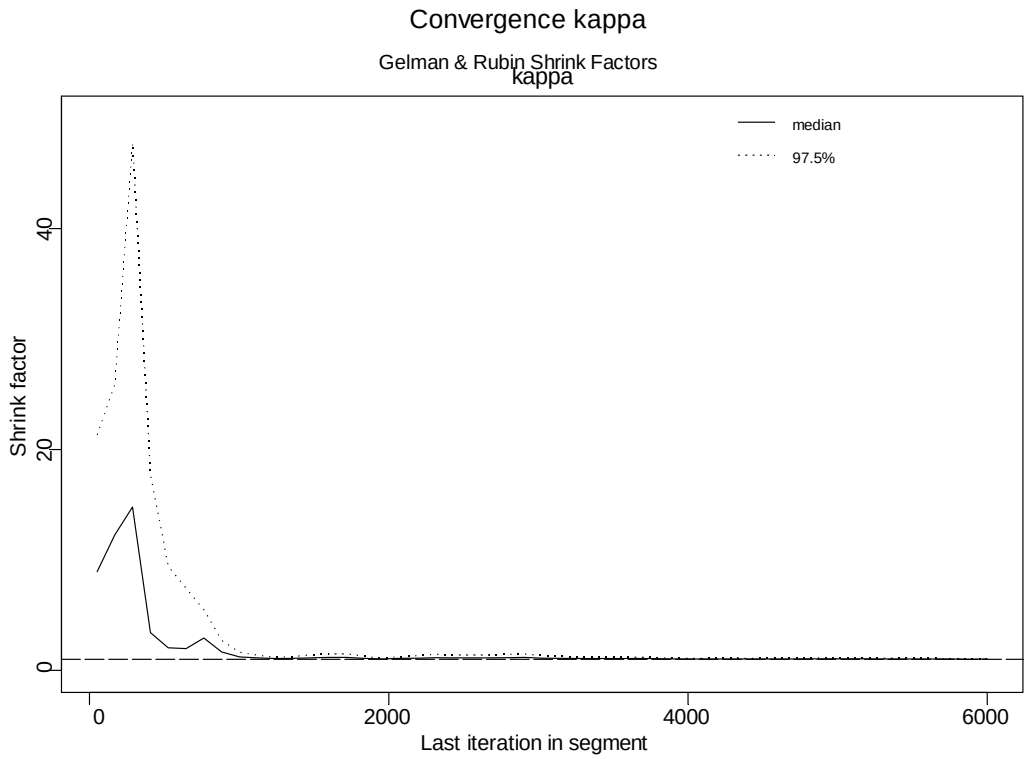
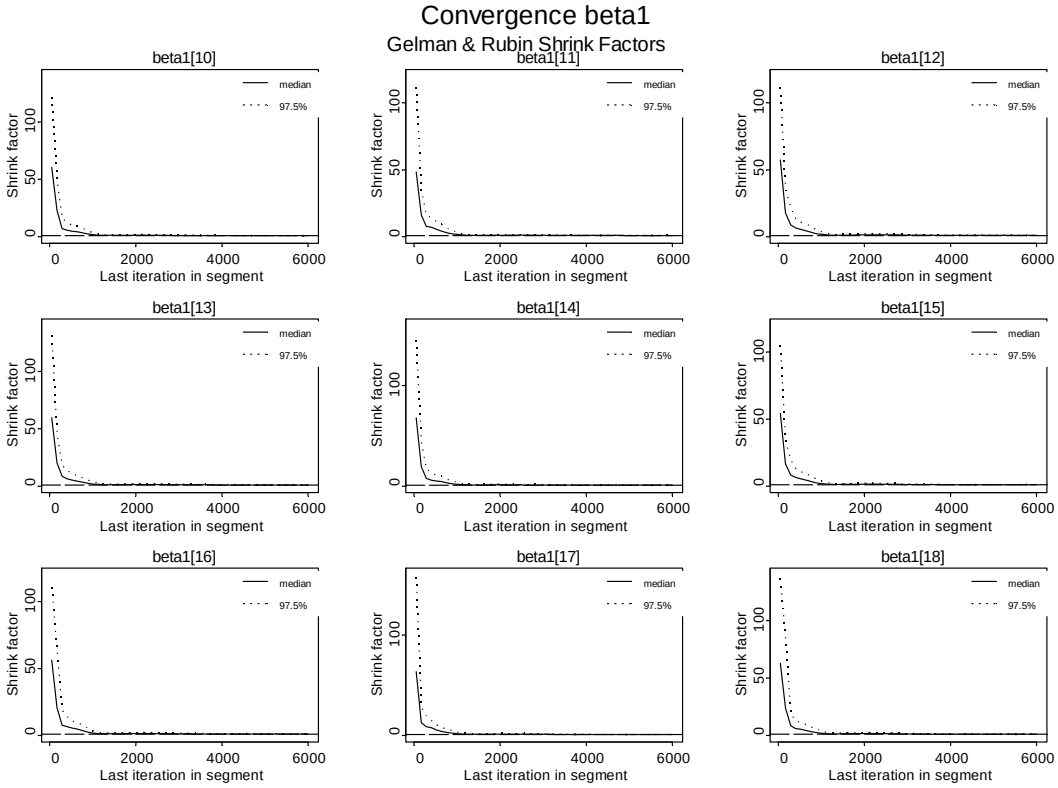
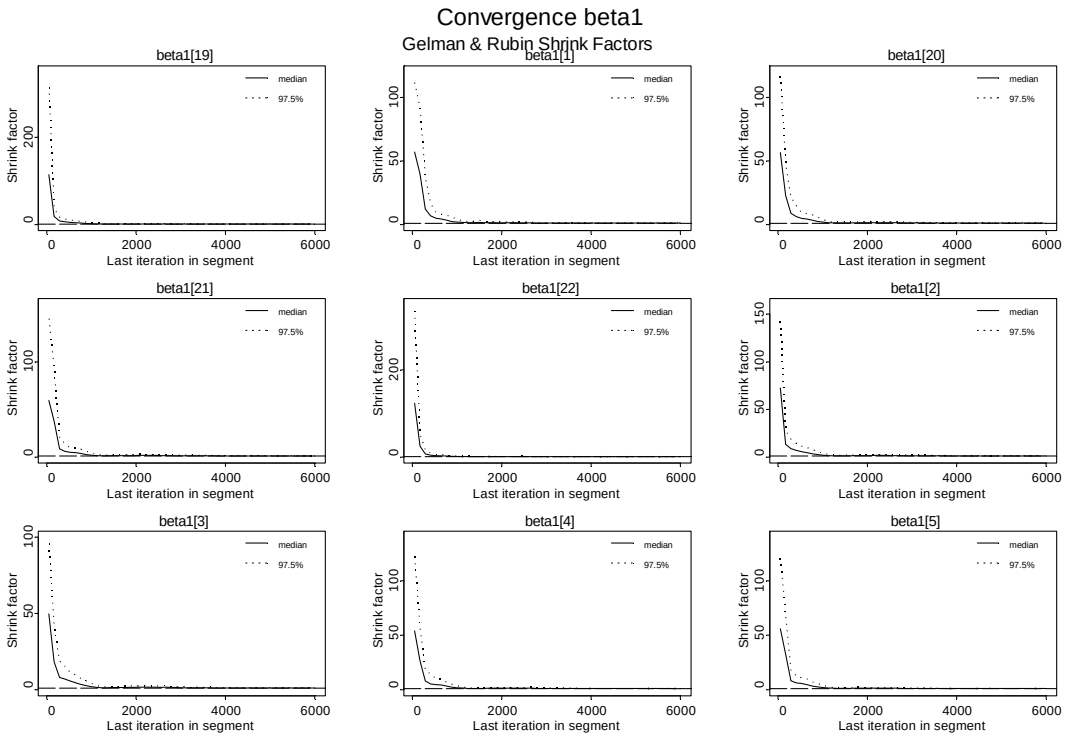


Figure 33




Convergence beta1
Gelman & Rubin Shrink Factors
beta1[19]
beta1[1]
beta1[20]
beta1[21]
beta1[22]
beta1[2]
beta1[3]
beta1[4]
beta1[5]
median
97.5%
Shrink factor
Last iteration in segment

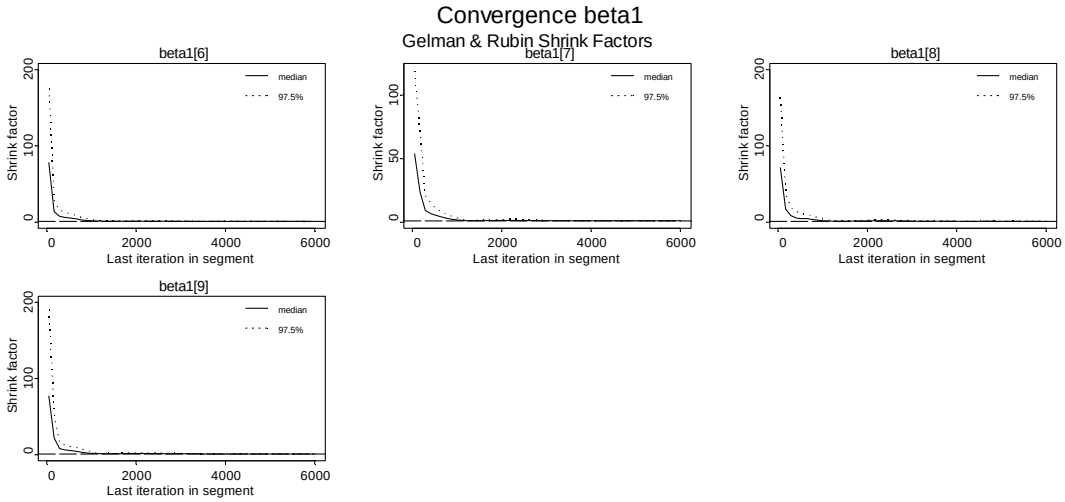

Convergence beta1
Gelman & Rubin Shrink Factors
beta1[6]
beta1[7]
beta1[8]
beta1[9]
median
97.5%
Shrink factor
Last iteration in segment

Figure 34

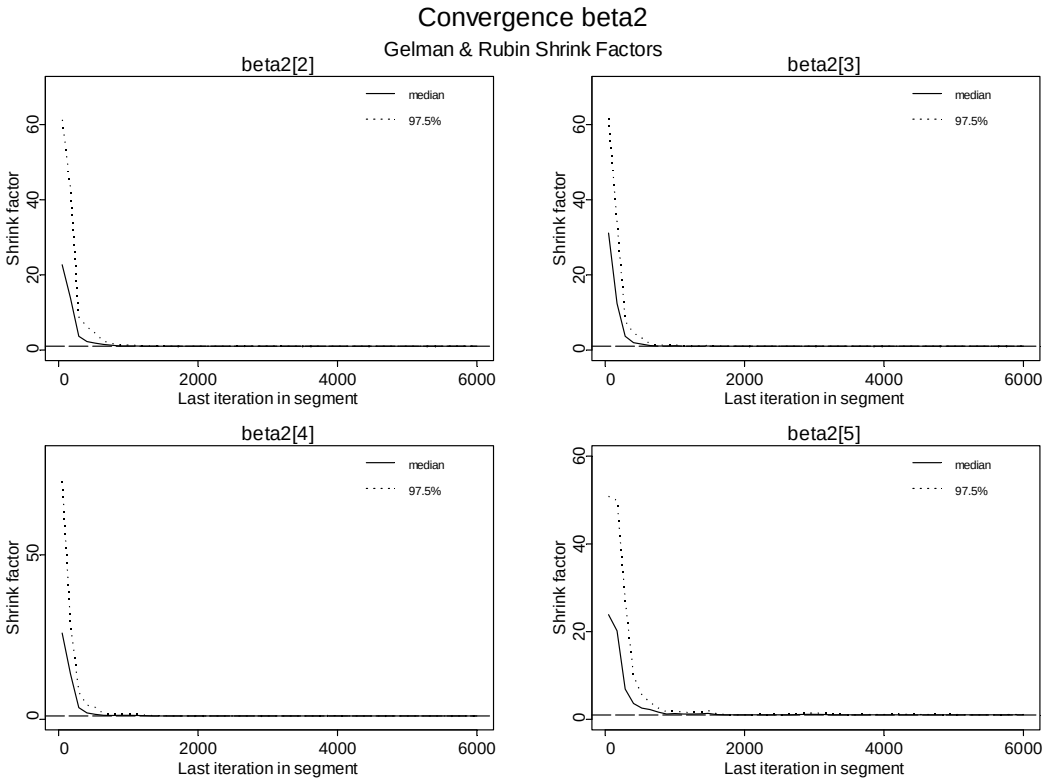


Figure 35

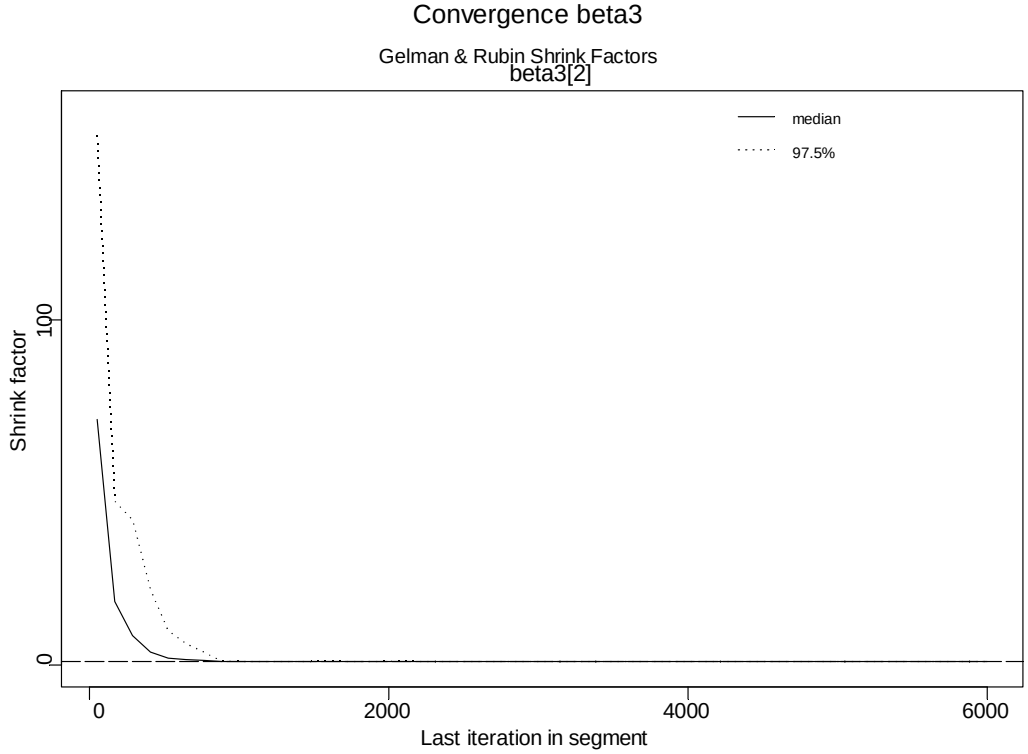
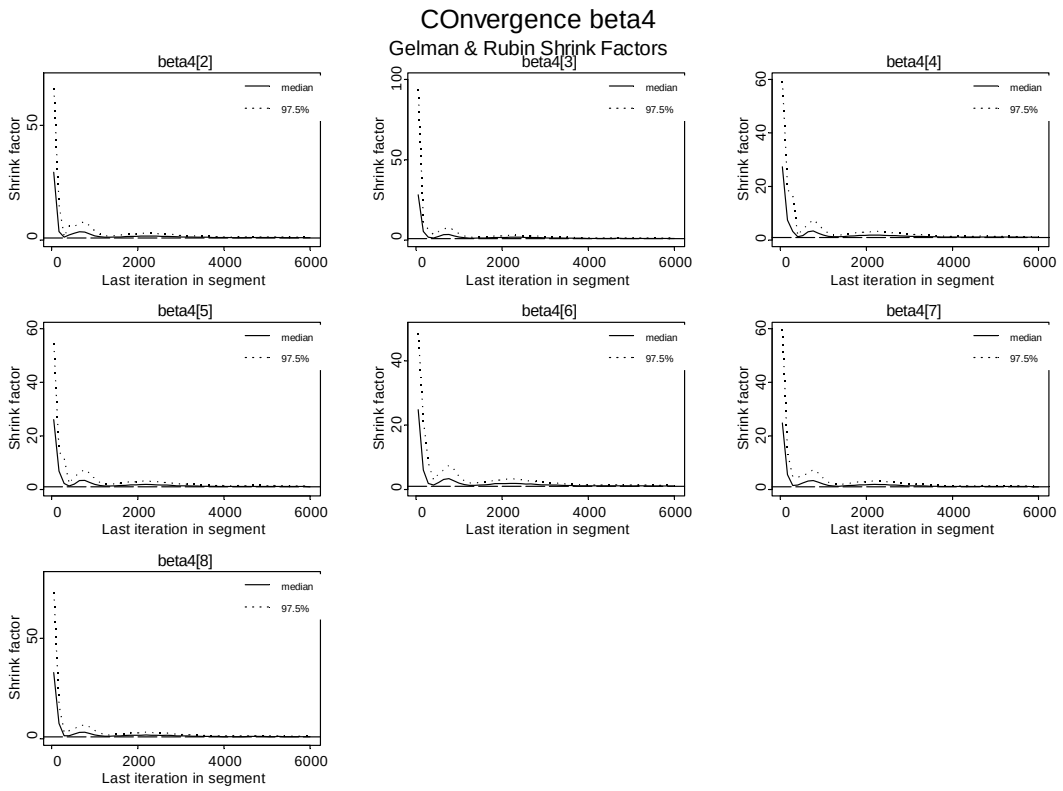


Figure 36



Dans le tableau 6, on a calculé le *R* pour chaque paramètre en utilisant les dernières 3000 itérations de chaque chaîne. On voit que les valeurs obtenues sont presque égales à 1, confirmant de nouveau la convergence.

Comme plus haut, on peut s'intéresser directement aux chaînes Markov des moyennes des régions et de la France plutôt qu'aux paramètres *alpha*, *kappa* et les β. Le tableau 7 contient les valeurs de *R* pour ces chaînes. Elles sont encore plus proches de 1 que celles du tableau 6, reconfirmant ce qu'on a déjà observé : en raison de l'autocorrélation plus faible, les chaînes $\{\mu_i^g\}$ et $\{\mu^g\}$ convergent plus vite et couvrent mieux les valeurs de la loi a posteriori.

Tableau 6 : *« Shrink Factor » des chaines Markov des paramètres du Modèle 3*

		R	
Paramètre	**Itérations utilisées**	**Estimateur**	**Quantile 97.5%**
alpha	3001:6000	1.01	1.04
kappa	3001:6000	1.01	1.04
beta1[1]	3001:6000	1.02	1.07
beta1[2]	3001:6000	1.04	1.14
beta1[3]	3001:6000	1.04	1.13
beta1[4]	3001:6000	1.02	1.08
beta1[5]	3001:6000	1.03	1.09
beta1[6]	3001:6000	1.03	1.10
beta1[7]	3001:6000	1.03	1.09
beta1[8]	3001:6000	1.05	1.16
beta1[9]	3001:6000	1.03	1.10
beta1[10]	3001:6000	1.03	1.09
beta1[11]	3001:6000	1.01	1.02
beta1[12]	3001:6000	1.03	1.09
beta1[13]	3001:6000	1.02	1.08
beta1[14]	3001:6000	1.03	1.09
beta1[15]	3001:6000	1.03	1.09
beta1[16]	3001:6000	1.03	1.11
beta1[17]	3001:6000	1.01	1.04
beta1[18]	3001:6000	1.03	1.10
beta1[19]	3001:6000	1.03	1.09
beta1[20]	3001:6000	1.02	1.07
beta1[21]	3001:6000	1.03	1.11
beta1[22]	3001:6000	1.00	1.00
beta2[2]	3001:6000	1.00	1.01
beta2[3]	3001:6000	1.00	1.01
beta2[4]	3001:6000	1.00	1.00
beta2[5]	3001:6000	1.02	1.08
beta3[2]	3001:6000	1.01	1.02
beta4[2]	3001:6000	1.06	1.18
beta4[3]	3001:6000	1.06	1.20
beta4[4]	3001:6000	1.07	1.21
beta4[5]	3001:6000	1.07	1.22
beta4[6]	3001:6000	1.06	1.19
beta4[7]	3001:6000	1.06	1.18
beta4[8]	3001:6000	1.08	1.25

Tableau 7 : *« Shrink Factor » des chaines Markov des régions et de la France*

		R	
Région (*) = région à extension	**Itérations utilisées**	**Estimateur**	**Quantile 97.5%**
Ile de France(*)	4001:6000	1.00	1.01
Champagne-Ardenne(*)	4001:6000	1.00	1.00
Picardie(*)	4001:6000	1.00	1.01
Haute-Normandie	4001:6000	1.00	1.00
Centre	4001:6000	1.00	1.00
Basse-Normandie	4001:6000	1.00	1.01
Bourgogne	4001:6000	1.00	1.00
Nord Pas de Calais(*)	4001:6000	1.01	1.04
Loraine	4001:6000	1.00	1.00
Alsace	4001:6000	1.00	1.00
Franche Compté	4001:6000	1.00	1.01
Pays de la Loire	4001:6000	1.00	1.00
Bretagne	4001:6000	1.01	1.02
Poitou Charente	4001:6000	1.00	1.00
Aquitaine	4001:6000	1.00	1.02
Midi-Pyrénées	4001:6000	1.00	1.00
Limousin	4001:6000	1.00	1.00
Rhône Alpes	4001:6000	1.01	1.03
Auvergne	4001:6000	1.00	1.00
Languedoc-Roussillon	4001:6000	1.00	1.01
PACA(*)	4001:6000	1.00	1.00
Corse	4001:6000	1.00	1.00
France Métropolitaine	**4001:6000**	**1.00**	**1.02**

7 Estimations départementales pour la variable R02AM

Le modèle 3 ci-dessous a été utilisé pour obtenir des estimations régionales et au niveau de la France mais il ne peut pas être utilisé pour obtenir des estimations au niveau des départements parce que il ne contient pas un effet département. En France il existe une forte demande pour des estimations départementales. Pour cette raison dans cette section nous allons voir comment transformer le modèle 3 dans un nouveau modèle qui permette l'estimation des 96 moyennes départementales. Le nouveau modèle, que nous allons appeler modèle 4, pourra être utilisé pour estimer les moyennes régionales et la moyenne de la France selon le même principe selon lequel le modèle 3 a permis l'estimation de la moyenne nationale. Il sera alors intéressant de comparer les moyennes régionales et la moyenne de la France obtenues sur base des deux modèles. Les outils et les développements statistiques sont identiques à ceux des sections précédentes. Pour cette raison, nous allons seulement présenter les formules et les résultats sans détailler.

L'adaptation la plus naturelle que nous allons effectuer est de remplacer l'effet région β_i, i=1,…,22 par un effet département noté β_d, d=1,…,96. Dans ce cas une cellule $i \times j \times s \times k$ sera remplacée par une cellule $d \times j \times s \times k$ avec l'inconvénient d'augmenter de beaucoup le nombre des cellules. En plus, une grande partie d'entre elles seraient vides et dans une analyse que nous avons réalisées mais qui n'a pas été insérée dans ce rapport, nous avons observé qu'en travaillant sous un modèle 3 sans effet strate, l'ajustement du modèle et les résultats de l'inférence ne sont pas modifiés. Nous aurions donc pu enlever l'effet strate mais nous l'avons gardé pour nous prémunir contre un plan de sondage informatif : la strate est présente dans le plan de sondage, donc pour ne pas risquer que celui-ci soit informatif il faudrait que le modèle contiennent aussi la strate.

Avec un effet département, pour les raisons énoncées ci-dessus, nous avons décidé d'enlever l'indice *j*, la strate, et de travailler avec des cellules du type $d \times s \times k$. Le sujet du plan de sondage informatif/non informatif dans les nouvelles circonstances n'a pas été étudié. Si le plan de sondage est informatif alors il faudrait introduire les poids de sondage dans l'analyse. Des idées de les intégrer sous une approche bayesienne hiérarchique se trouvent dans Gelman et al (1995), mais nous n'avons pas étudié ces méthodes. Par conséquent, notre modèle 4 pour départements sera donné par (les notations sont analogues à celles utilisées dans les sections précédentes):

Modèle 4

$$y_{dskl} \mid \nu_{dskl} \underset{ind}{\sim} \text{Poisson}(\nu_{dskl})$$

$$\nu_{dskl} \mid \mu_{dsk}, alpha, kappa \underset{ind}{\sim} \text{Gamma}\left(\frac{\mu_{dsk}^{1-kappa}}{alpha}, \frac{\mu_{dsk}^{-kappa}}{alpha}\right)$$

$$\log(\mu_{dsk}) = \beta_{1d} + \beta_{3s} + \beta_{4k}$$

$$alpha \sim \text{Unif}(0,100),\ kappa \sim \text{Unif}(-1,100)$$

$$\beta_{1d} \sim \text{Unif}(-10,10),\ \beta_{3s} \sim \text{Unif}(-10,10),\ \beta_{4k} \sim \text{Unif}(-10,10)$$

En utilisant les mêmes outils statistiques que celles de la section 5 nous avons pu constaté que les modèles 3 et 4 assurent des ajustements identiques par rapport aux données. Les mesures du tableau 1 ainsi que les autres paramètres calculés dans la section 5 sont pratiquement les mêmes dans le cas des deux modèles. Pour cette raison les valeurs calculées sous le modèle 4 ne sont pas repris dans cette section. On peut conclure que le fait de remplacer l'effet région par un effet département ne change rien au niveau de l'ajustement.

Les paramètres du modèle 4 s'estiment à partir de leurs distributions a postériori en utilisant les chaines Markov obtenues avec l'échantillonnage de Gibbs. Les moyennes départementales et leurs variances utilisent ces chaines selon des formules analogues à (22) et (23):

$$\hat{\mu}_d = \frac{1}{N_d}[\sum_s \sum_k \sum_{l \in obs_d} y_{dskl} + \frac{1}{G}\sum_g \sum_s \sum_k (N_{dsk} - n_{dsk})\mu_{dsk}^g] \quad (27)$$

et

$$V(\hat{\mu}_d) = \frac{1}{N_d^2}\{\frac{1}{G}\sum_g \sum_s \sum_k (N_{dsk} - n_{dsk})\mu_{dsk}^g + \frac{1}{G}\sum_g \sum_s \sum_k (N_{dsk} - n_{dsk}) alpha^g \mu_{dsk}^{g\,(1+kappa^g)} + \\ + \frac{1}{G}\sum_g [\sum_s \sum_k (N_{dsk} - n_{dsk})\mu_{dsk}^g]^2 - [\frac{1}{G}\sum_g \sum_s \sum_k (N_{dsk} - n_{dsk})\mu_{dsk}^g]^2\} \quad (28)$$

(27) permet d'identifier des chaines Markov $\{\mu_d^g\}$, g=1,…,G pour la moyenne de chaque département:

$$\mu_d^g = \frac{1}{N_d}[\sum_s \sum_k \sum_{l \in obs_d} y_{dskl} + \sum_s \sum_k (N_{dsk} - n_{dsk})\mu_{dsk}^g]$$

Si on fait la «somme» d'après d correspondant aux départements d'une région i, alors le modèle 4 peut être utilisé pour estimer les moyennes régionales μ_i et les précisions de ces estimations. Si on fait la «somme» d'après tous les départements, alors on peut estimer la moyenne de la France μ et la précision de cette estimation. Les formules se trouvent ci-dessous :

$$\hat{\mu}_i = \frac{1}{N_i}[\sum_{d \in i} \sum_s \sum_k \sum_{l \in obs_d} y_{dskl} + \frac{1}{G}\sum_{d \in i} \sum_g \sum_s \sum_k (N_{dsk} - n_{dsk})\mu_{dsk}^g] \quad (29)$$

$$V(\hat{\mu}_i) = \frac{1}{N_i^2}\{\frac{1}{G}\sum_g \sum_{d \in i} \sum_s \sum_k (N_{dsk} - n_{dsk})\mu_{dsk}^g + \frac{1}{G}\sum_g \sum_{d \in i} \sum_s \sum_k (N_{dsk} - n_{dsk}) alpha^g \mu_{dsk}^{g\,(1+kappa^g)} + \\ + \frac{1}{G}\sum_g [\sum_{d \in i} \sum_s \sum_k (N_{dsk} - n_{dsk})\mu_{dsk}^g]^2 - [\frac{1}{G}\sum_g \sum_{d \in i} \sum_s \sum_k (N_{dsk} - n_{dsk})\mu_{dsk}^g]^2\} \quad (30)$$

$$\hat{\mu} = \frac{1}{N}[\sum_d \sum_s \sum_k \sum_{l \in obs_d} y_{dskl} + \frac{1}{G}\sum_d \sum_g \sum_s \sum_k (N_{dsk} - n_{dsk})\mu_{dsk}^g] \quad (31)$$

$$V(\hat{\mu}) = \frac{1}{N^2}\{\frac{1}{G}\sum_g \sum_d \sum_s \sum_k (N_{dsk} - n_{dsk})\mu_{dsk}^g + \frac{1}{G}\sum_g \sum_d \sum_s \sum_k (N_{dsk} - n_{dsk}) alpha^g \mu_{dsk}^{g\,(1+kappa^g)} + \\ + \frac{1}{G}\sum_g [\sum_d \sum_s \sum_k (N_{dsk} - n_{dsk})\mu_{dsk}^g]^2 - [\frac{1}{G}\sum_g \sum_d \sum_s \sum_k (N_{dsk} - n_{dsk})\mu_{dsk}^g]^2\} \quad (32)$$

($d \in i$ signifie qu'on fait la somme d'après les départements *d* se trouvant dans la région *i*; on remarque aussi que (29) et (31) identifient des chaines Markov $\{\mu_i^g\}$ et $\{\mu^g\}$ pour les régions et la France; les formules sont évidentes et on les omet).

Les chaines Markov $\{\mu_d^g\}$, $\{\mu_i^g\}$ et $\{\mu^g\}$ sont importantes parce qu'elles permettent de calculer les précisions des estimateurs d'une autre manière que les formules théoriques (28), (30) et (32) en faisant tout simplement leurs variances. En même temps elles permettent le calcul des précisions Monte Carlo tant pour les moyennes départementales, régionales et nationale que pour leurs écart-types respectifs.

Comme dans les sections précédentes, nous avons roulé trois chaines Markov simultanément. Nous avons observé que pour la plupart des paramètres la convergence est atteinte entre l'itération 1000 et 2000 et que toutes les chaines convergent à partir de l'itération 3000. Pour cette raison nous avons abandonné les premières 3000 itérations et nous avons utilisé les dernières 2000 de chaque chaine, donc un total de 6000 itérations. Utilisant les formules (27)-(32) nous avons calculé des estimations et leurs précisions pour chaque département, régions et pour la France Métropolitaine. Les résultats pour l'échantillon avec extension se trouvent dans le tableau 8.

Les estimations du tableau 8 utilisent des effectifs N_{dsk} estimés comme somme des poids de sondage. Dans l'échantillon avec extension il y un seul département sans aucune observation : Lozère dans la région Languedoc-Roussillon. Pour les N_{dsk} de ce département nous n'avons aucune estimation, donc dans son cas nous n'avons pas pu estimer la moyenne départementale. Si on avait une estimation pour les N_{dsk} ou les vrais effectifs alors des estimations pourrait être calculées mais comme $\hat{\beta}_d$ est très instable (l'échantillon n'apporte aucune information donc sa loi a posteriori est identique à sa loi a priori, une loi uniforme sur [-10,10] dans ce cas), $\hat{\mu}_d$ sera instable, ce qui va rendre $\hat{\mu}_i$ de la région de Languedoc-Roussillon et $\hat{\mu}$ de la moyenne nationale instables. Nous avons remarqué ce phénomène sur l'échantillon sans extension (les résultats n'ont pas été insérés) et dans le cas de deux autres départements sans aucune observation dans l'échantillon sans extension (Alpes de Haute Provence et Hautes Alpes dans la région PACA).

On peut remarquer que les estimations régionales et celle nationales ainsi que leurs précisions sont identiques à celles obtenues sur base du modèle 3. Nous avons aussi calculé les quantiles d'ordre 0.025, 0.5 et 0.975 qui fournissent un intervalle de confiance de niveau 0.95 pour les moyennes départementales.

Tableau 8:*Estimations à partir de l'échantillon avec extension (39900 observations)*

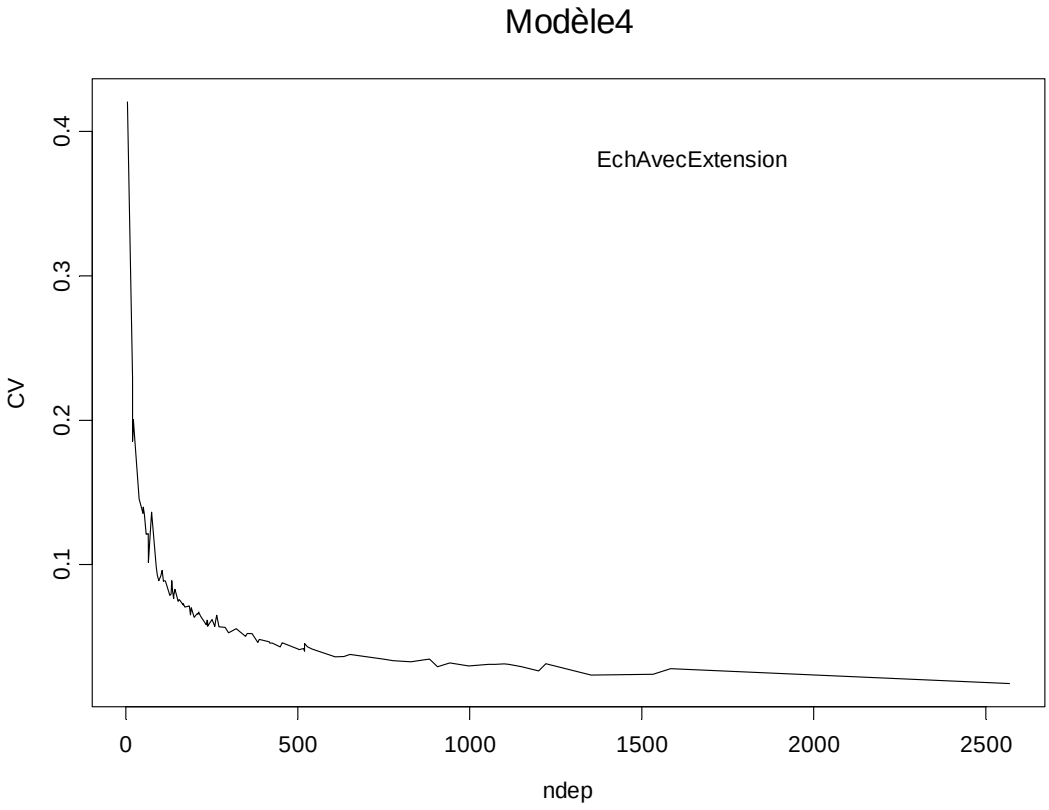
Département **Région**	Estimation	Ecart-Type	MCSE($\hat{\mu}_d$) (MCSE($\hat{\sigma}_d$))	Quantiles 0.025;0.5;0.975
Paris	2.42	0.0676	0.0012(0.0007)	2.28;2.42;2.55
Seine et Marne	3.01	0.0944	0.0017(0.0009)	2.83;3.01;3.20
Yvelines	3.19	0.0986	0.0017(0.0009)	3.00;3.19;3.39
Essonne	3.32	0.1127	0.0019(0.0011)	3.10;3.32;3.54
Hauts de Seine	2.56	0.0806	0.0015(0.0008)	2.40;2.56;2.72
Seine Saint Denis	3.43	0.1008	0.0018(0.0010)	3.23;3.42;3.63
Val de Marne	3.18	0.0983	0.0018(0.0010)	2.99;3.17;3.37
Val d'Oise	3.25	0.1005	0.0017(0..0010)	3.06;3.25;3.45
Ile de France(*)	**2.99**	**0.0330**	**0.0006(0.0003)**	**2.93;2.99;3.06**
Ardennes	4.92	0.2116	0.0035(0.0020)	4.52;4.91;5.35
Aube	3.84	0.1766	0.0031(0.0017)	3.50;3.83;4.20
Marne	4.29	0.1132	0.0019(0.0011)	4.07;4.29;4.51
Haute Marne	4.91	0.2264	0.0038(0.0022)	4.48;4.90;5.36
Champagne-Ardenne(*)	**4.43**	**0.0829**	**0.0014(0.0008)**	**4.27;4.43;4.59**
Aisne	4.27	0.1431	0.0025(0.0014)	3.99;4.27;4.55
Oise	4.04	0.1206	0.0021(0.0012)	3.81;4.04;4.29
Somme	4.79	0.1400	0.0023(0.0013)	4.52;4.79;5.07
Picardie(*)	**4.35**	**0.0771**	**0.0013(0.0007)**	**4.20;4.35;4.50**
Eure	3.67	0.2578	0.0044(0.0025)	3.18;3.66;4.19
Seine Maritime	4.40	0.1807	0.0031(0.0017)	4.06;4.40;4.77
Haute-Normandie	**4.20**	**0.1487**	**0.0025(0.0014)**	**3.92;4.20;4.50**
Cher	4.23	0.3050	0.0053(0.0030)	3.66;4.23;4.83
Eure et Loir	3.49	0.2894	0.0051(0.0029)	2.95;3.47;4.09
Indre	4.32	0.6287	0.0109(0.0062)	3.19;4.29;5.60
Indre et Loire	3.99	0.2461	0.0043(0.0024)	3.53;3.99;4.48
Loir et Cher	4.29	0.3946	0.0070(0.0039)	3.56;4.28;5.10
Loiret	4.02	0.1838	0.0031(0.0018)	3.67;4.02;4.39
Centre	**4.02**	**0.1136**	**0.0019(0.0011)**	**3.79;4.02;4.24**
Calvados	4.15	0.2352	0.0042(0.0023)	3.71;4.14;4.63
Manche	4.35	0.2588	0.0046(0.0026)	3.87;4.33;4.88
Orne	4.54	0.2962	0.0051(0.0029)	3.97;4.54;5.14
Basse-Normandie	**4.32**	**0.1500**	**0.0026(0.0015)**	**4.04;4.32;4.62**
Côte d'Or	3.57	0.1991	0.0035(0.0020)	3.20;3.57;3.97
Nièvre	4.30	0.5212	0.0090(0.0051)	3.35;4.28;5.38
Saône et Loire	3.55	0.1860	0.0032(0.0018)	3.20;3.55;3.93
Yonne	4.04	0.3528	0.0063(0.0035)	3.40;4.03;4.79
Bourgogne	**3.68**	**0.1252**	**0.0022(0.0012)**	**3.44;3.68;3.92**
Nord	4.99	0.0881	0.0015(0.0008)	4.82;4.99;5.16
Pas de Calais	5.42	0.1274	0.0021(0.0012)	5.17;5.42;5.67
Nord Pas de Calais(*)	**5.14**	**0.0729**	**0.0012(0.0007)**	**4.99;5.14;5.28**
Meurthe et Moselle	4.18	0.1948	0.0034(0.0019)	3.82;4.18;4.58
Meuse	3.63	0.4891	0.0088(0.0049)	2.74;3.60;4.65
Moselle	4.82	0.1746	0.0029(0.0017)	4.49;4.82;5.17
Vosges	4.70	0.3596	0.0062(0.0035)	4.03;4.69;5.44
Lorraine	**4.54**	**0.1210**	**0.0021(0.0012)**	**4.30;4.54;4.78**

Bas Rhin	4.28	0.1709	0.0029(0.0016)	3.94;4.27;4.62
Haut Rhin	4.01	0.2020	0.0035(0.0020)	3.63;4.01;4.41
Alsace	**4.17**	**0.1306**	**0.0022(0.0013)**	**3.92;4.17;4.43**
Doubs	3.95	0.2546	0.0044(0.0025)	3.46;3.94;4.46
Jura	3.63	0.2443	0.0043(0.0024)	3.18;3.62;4.14
Haute Saône	4.09	0.2710	0.0047(0.0027)	3.59;4.08;4.66
Territoire de Belfort	3.44	0.4182	0.0076(0.0042)	2.68;3.42;4.31
Franche Comté	**3.85**	**0.1407**	**0.0024(0.0014)**	**3.57;3.85;4.13**
Loire Atlantique	3.69	0.1594	0.0027(0.0015)	3.38;3.69;4.00
Maine et Loire	4.13	0.1422	0.0024(0.0013)	3.85;4.13;4.42
Mayenne	4.48	0.3969	0.0068(0.0039)	3.74;4.46;5.29
Sarthe	4.20	0.2964	0.0050(0.0029)	3.64;4.19;4.81
Vendée	3.94	0.2592	0.0046(0.0026)	3.46;3.93;4.48
Pays de la Loire	**4.01**	**0.0900**	**0.0015(0.0008)**	**3.83;4.01;4.18**
Côtes d'Armor	3.64	0.2225	0.0039(0.0022)	3.22;3.64;4.10
Finistère	3.95	0.1495	0.0026(0.0015)	3.66;3.95;4.24
Ille de Vilaine	4.53	0.2394	0.0040(0.0023)	4.08;4.53;5.01
Morbihan	3.66	0.2266	0.0039(0.0022)	3.22;3.65;4.11
Bretagne	**3.97**	**0.1009**	**0.0018(0.0010)**	**3.78;3.97;4.17**
Charente	5.12	0.4739	0.0081(0.0046)	4.25;5.10;6.09
Charente Maritime	4.40	0.1833	0.0031(0.0018)	4.05;4.40;4.77
Deux Sèvres	4.15	0.3684	0.0062(0.0036)	3.45;4.14;4.92
Vienne	4.40	0.3291	0.0057(0.0032)	3.79;4.40;5.07
Poitou Charente	**4.44**	**0.1411**	**0.0024(0.0013)**	**4.17;4.44;4.73**
Dordogne	4.42	0.4351	0.0076(0.0043)	3.63;4.40;5.33
Gironde	4.43	0.1611	0.0027(0.0015)	4.12;4.43;4.75
Landes	4.13	0.2505	0.0044(0.0025)	3.66;4.12;4.63
Lot et Garonne	4.61	0.3631	0.0062(0.0035)	3.91;4.61;5.36
Pyrénées Atlantiques	4.09	0.1715	0.0030(0.0017)	3.77;4.08;4.44
Aquitaine	**4.29**	**0.0995**	**0.0017(0.0009)**	**4.10;4.29;4.49**
Ariège	4.22	0.3193	0.0056(0.0031)	3.62;4.21;4.86
Aveyron	3.54	0.4954	0.0086(0.0049)	2.65;3.51;4.57
Haute Garonne	4.27	0.1944	0.0032(0.0019)	3.89;4.26;4.65
Gers	5.17	0.5598	0.0095(0.0054)	4.17;5.15;6.36
Lot	3.19	0.7320	0.0131(0.0074)	1.97;3.11;4.82
Hautes Pyrénées	4.88	0.3102	0.0053(0.0030)	4.29;4.87;5.50
Tarn	4.32	0.3149	0.0053(0.0030)	3.73;4.31;4.99
Tarn et Garonne	4.41	0.8841	0.0155(0.0088)	2.86;4.35;6.27
Midi-Pyrénées	**4.38**	**0.1228**	**0.0021(0.0012)**	**4.14;4.38;4.63**
Corrèze	5.01	0.2920	0.0049(0.0028)	4.46;5.00;5.61
Creuse	4.20	0.3321	0.0058(0.0033)	3.58;4.18;4.91
Haute Vienne	5.19	0.4614	0.0077(0.0045)	4.31;5.17;6.13
Limousin	**4.80**	**0.1987**	**0.0034(0.0019)**	**4.42;4.80;5.19**

Ain	3.13	0.1417	0.0025(0.0014)	2.86;3.13;3.41
Ardèche	4.03	1.6968	0.0286(0.0166)	1.55;3.76;7.95
Drôme	4.15	0.3687	0.0062(0.0036)	3.46;4.14;4.88
Isère	3.37	0.1760	0.0031(0.0017)	3.04;3.36;3.73
Loire	3.97	0.2262	0.0039(0.0022)	3.54;3.97;4.43
Rhône	3.13	0.1079	0.0019(0.0010)	2.92;3.12;3.34
Savoie	3.23	0.2876	0.0051(0.0029)	2.69;3.22;3.81
Haute Savoie	2.78	0.1810	0.0033(0.0018)	2.45;2.78;3.15
Rhône Alpes	**3.27**	**0.0651**	**0.0011(0.0006)**	**3.14;3.27;3.40**
Allier	6.04	0.6133	0.0102(0.0059)	4.91;6.01;7.30
Cantal	3.91	0.3200	0.0056(0.0032)	3.33;3.90;4.58
Haute Loire	3.92	0.2971	0.0051(0.0029)	3.36;3.91;4.53
Puy de Dôme	3.89	0.2780	0.0049(0.0027)	3.37;3.88;4.46
Auvergne	**4.15**	**0.1694**	**0.0029(0.0016)**	**3.83;4.15;4.49**
Aude	3.97	0.5388	0.0096(0.0054)	3.01;3.93;5.13
Gard	4.42	0.2528	0.0043(0.0024)	3.94;4.41;4.92
Hérault	3.93	0.1635	0.0028(0.0016)	3.63;3.92;4.26
Lozère	NA	NA	NA(NA)	NA;NA;NA
Pyrénées Orientales	4.84	0.2780	0.0048(0.0027)	4.31;4.83;5.40
Languedoc-Roussillon	**4.24**	**0.1198**	**0.0020(0.0011)**	**4.02;4.24;4.49**
Alpes de Haute Provence	4.38	0.3491	0.0059(0.0034)	3.73;4.37;5.08
Hautes Alpes	3.78	0.3628	0.0062(0.0035)	3.10;3.76;4.53
Alpes Maritimes	3.40	0.1084	0.0019(0.0010)	3.19;3.40;3.63
Bouches du Rhône	4.05	0.0975	0.0017(0.0009)	3.86;4.05;4.24
Var	4.06	0.1326	0.0023(0.0013)	3.80;4.05;4.32
Vaucluse	3.87	0.1866	0.0033(0.0018)	3.51;3.86;4.24
PACA(*)	**3.88**	**0.0603**	**0.0010(0.0006)**	**3.76;3.88;4.00**
Haute Corse	5.95	1.1047	0.0183(0.0107)	4.04;5.87;8.34
Corse du Sud	1.95	0.2656	0.0052(0.0028)	1.48;1.93;2.51
Corse	**2.76**	**0.3043**	**0.0054(0.0030)**	**2.20;2.75;3.39**
France Métropolitaine	**3.93**	**0.0208**	**0.0003(0.0002)**	**3.89;3.93;3.97**

Dans la figure 37 se trouve le coefficient de variation représenté en fonction de la taille de l'échantillon départemental.

Figure 37 : *Coefficient de Variation vs Taille de l'échantillon départemental*



8 Estimations régionales et départementales calculées avec des effectifs de R99

Nous avons vu que les formules des estimations (régionales et départementales) et de leurs précisions dérivées dans les sections précédentes utilisent les effectifs totaux des cellules (N_{ijsk} ou N_{dsk}). Au moment où nous avons calculé ces estimations et rédigé le présent chapitre, nous ne disposions pas de ces effectifs. Pour cette raison nous les avons estimés par la somme des poids de sondage et nous les avons remplacés dans les formules par les valeurs données par ces estimations. En même temps nous avons souligné que les précisions ne tiennent pas compte de cette estimation supplémentaire. Plus tard, INSEE nous a fourni ces effectifs calculés à partir du recensement de 1999. Dans cette section nous présentons les nouvelles estimations qui utilisent les «vraies effectifs ». Nous allons voir que les valeurs des estimations et de leurs précisions sont pratiquement les mêmes.

Dans le tableau 9 se trouvent les estimations régionales pour la variable R02AM obtenues sur base du modèle 3. Si on compare avec les valeurs du tableau 4 on voit que les estimations sont pratiquement identiques.

Tableau 9:*Estimations de R02AM utilisant les effectifs de R99(échantillon avec extension)*

Région (*) = région à extension	Estimation	Ecart-Type	Quantile 0.025	Quantile 0.5	Quantile 0.975
Ile de France(*)	3.01	0.0327	2.94	3.01	3.07
Champagne-Ardenne(*)	4.42	0.0797	4.26	4.42	4.58
Picardie(*)	4.39	0.0784	4.24	4.39	4.54
Haute-Normandie	4.15	0.1449	3.87	4.16	4.44
Centre	3.98	0.1114	3.76	3.98	4.20
Basse-Normandie	4.30	0.1496	4.01	4.30	4.60
Bourgogne	3.60	0.1232	3.36	3.60	3.85
Nord Pas de Calais(*)	5.13	0.0735	4.99	5.13	5.28
Loraine	4.54	0.1190	4.30	4.53	4.77
Alsace	4.31	0.1345	4.07	4.31	4.60
Franche Compté	3.82	0.1341	3.57	3.82	4.10
Pays de la Loire	3.96	0.0901	3.78	3.96	4.14
Bretagne	3.97	0.1033	3.77	3.97	4.18
Poitou Charente	4.44	0.1416	4.17	4.43	4.74
Aquitaine	4.38	0.0993	4.18	4.38	4.57
Midi-Pyrénées	4.41	0.1278	4.17	4.40	4.67
Limousin	4.64	0.1934	4.28	4.63	5.02
Rhône Alpes	3.31	0.0663	3.18	3.31	3.44
Auvergne	4.15	0.1713	3.82	4.15	4.50
Languedoc-Roussillon	4.27	0.1280	4.04	4.27	4.53
PACA(*)	3.84	0.0575	3.73	3.84	3.95
Corse	2.55	0.2887	2.00	2.53	3.13
FranceMétropolitaine	**3.92**	**0.0209**	**3.88**	**3.92**	**3.96**

On veut maintenant utiliser le modèle 4 pour les départements et calculer les estimations et leurs précisions avec les vrais effectifs. Dans le tableau 10 se trouvent les résultats. En comparant les tableaux 8 et 10 on peut observer que les estimations pour les moyennes départementales (estimation de la moyenne et sa précision) sont pratiquement les mêmes. Par contre quand on utilise le modèle 4 pour estimer les moyennes régionales et de la France, alors on peut observer que pour une grande partie des régions les écart types sont légèrement plus élevés dans le cas où on utilise les effectifs calculés à partir du recensement de 1999. Ceci est particulièrement vrai pour la Corse où l'écart type a doublé dans le cas où on utilise les effectifs du recensement. Pour les régions à extension il n'y a pas de différence.

Ceci est dû à la qualité des estimations des effectifs par la somme des poids de sondage. Dans les régions à extension où l'échantillon est suffisamment grand, il n'y pas de différence. Dans le cas des autres régions sauf la Corse, les différences sont assez faibles. Dans le cas de la Corse où l'échantillon régional a la taille la plus petite, on voit une différence nette entre les deux estimations. Il faut considérer et travailler avec les estimations de cette section utilisant les vrais effectifs.

Tableau 10:*Estimations de R02AM utilisant les effectifs de R99(échantillon avec extension)*

Département **Région**	Estimation	Ecart-Type	MCSE($\hat{\mu}_d$) (MCSE($\hat{\sigma}_d$))	Quantiles 0.025;0.5;0.975
Paris	2.42	0.0675	0.0012(0.0007)	2.28 ;2.41 ;2.55
Seine et Marne	3.06	0.0957	0.0017(0.0009)	2.87 ;3.05 ;3.25
Yvelines	3.19	0.0984	0.0017(0.0009)	2.99 ;3.19 ;3.38
Essonne	3.23	0.1098	0.0019(0.0011)	3.02 ;3.23 ;3.45
Hauts de Seine	2.63	0.0827	0.0015(0.0008)	2.47 ;2.62 ;2.79
Seine Saint Denis	3.45	0.1015	0.0018(0.0010)	3.26 ;3.45 ;3.66
Val de Marne	3.18	0.0984	0.0018(0.0010)	2.99 ;3.18 ;3.38
Val d'Oise	3.30	0.1021	0.0018(0.0010)	3.11 ;3.31 ;3.51
Ile de France(*)	**3.00**	**0.0331**	**0.0006(0.0003)**	**2.94 ;3.00 ;3.07**
Ardennes	4.80	0.2064	0.0034(0.0020)	4.41 ;4.80 ;5.22
Aube	3.91	0.1798	0.0032(0.0018)	3.57 ;3.90 ;4.27
Marne	4.29	0.1133	0.0019(0.0011)	4.08 ;4.29 ;4.52
Haute Marne	4.91	0.2262	0.0038(0.0022)	4.47 ;4.90 ;5.36
Champagne-Ardenne(*)	**4.41**	**0.0838**	**0.0014(0.0008)**	**4.25 ;4.41 ;4.58**
Aisne	4.29	0.1438	0.0026(0.0014)	4.01 ;4.29 ;4.58
Oise	4.11	0.1227	0.0021(0.0012)	3.88 ;4.11 ;4.36
Somme	4.76	0.1390	0.0023(0.0013)	4.49 ;4.76 ;5.04
Picardie(*)	**4.36**	**0.0776**	**0.0013(0.0007)**	**4.20 ;4.36 ;4.51**
Eure	3.55	0.2495	0.0043(0.0024)	3.08 ;3.54 ;4.06
Seine Maritime	4.40	0.1807	0.0031(0.0017)	4.06 ;4.40 ;4.77
Haute-Normandie	**4.15**	**0.1467**	**0.0025(0.0014)**	**3.87 ;4.14 ;4.44**
Cher	4.26	0.3070	0.0053(0.0030)	3.68 ;4.26 ;4.86
Eure et Loir	3.52	0.2924	0.0052(0.0029)	2.98 ;3.51 ;4.13
Indre	4.49	0.6542	0.0113(0.0064)	3.31 ;4.46 ;5.83
Indre et Loire	4.19	0.2585	0.0045(0.0025)	3.70 ;4.19 ;4.71
Loir et Cher	3.99	0.3673	0.0065(0.0037)	3.32 ;3.98 ;4.75
Loiret	3.83	0.1750	0.0030(0.0017)	3.49 ;3.83 ;4.18
Centre	**4.00**	**0.1243**	**0.0021(0.0012)**	**3.76 ;4.00 ;4.25**
Calvados	3.96	0.2244	0.0040(0.0022)	3.54 ;3.95 ;4.42
Manche	4.37	0.2602	0.0047(0.0026)	3.89 ;4.36 ;4.91
Orne	4.71	0.3069	0.0052(0.0030)	4.11 ;4.70 ;5.32
Basse-Normandie	**4.25**	**0.1488**	**0.0026(0.0015)**	**3.97 ;4.25 ;4.55**
Côte d'Or	3.53	0.1966	0.0035(0.0019)	3.15 ;3.52 ;3.92
Nièvre	4.52	0.5477	0.0094(0.0054)	3.52 ;4.50 ;5.64
Saône et Loire	3.43	0.1796	0.0031(0.0017)	3.09 ;3.43 ;3.80
Yonne	3.73	0.3252	0.0058(0.0032)	3.13 ;3.71 ;4.41
Bourgogne	**3.68**	**0.1360**	**0.0024(0.0013)**	**3.41 ;3.68 ;3.95**
Nord	4.99	0.0880	0.0015(0.0008)	4.81 ;4.99 ;5.16
Pas de Calais	5.40	0.1270	0.0021(0.0012)	5.15 ;5.40 ;5.65
Nord Pas de Calais(*)	**5.14**	**0.0729**	**0.0012(0.0007)**	**4.99 ;5.14 ;5.28**
Meurthe et Moselle	4.19	0.1954	0.0034(0.0019)	3.83 ;4.19 ;4.60
Meuse	3.78	0.5090	0.0092(0.0051)	2.85 ;3.75 ;4.84
Moselle	4.80	0.1735	0.0029(0.0017)	4.46 ;4.79 ;5.14
Vosges	4.59	0.3510	0.0060(0.0034)	3.93 ;4.58 ;5.32
Lorraine	**4.49**	**0.1227**	**0.0021(0.0012)**	**4.26 ;4.49 ;4.74**
Bas Rhin	4.47	0.1788	0.0031(0.0017)	4.12 ;4.47 ;4.83
Haut Rhin	4.13	0.2078	0.0036(0.0020)	3.73 ;4.12 ;4.54

Alsace	**4.33**	**0.1357**	**0.0023(0.0013)**	**4.07 ;4.33 ;4.60**
Doubs	3.96	0.2551	0.0044(0.0025)	3.47 ;3.95 ;4.47
Jura	3.56	0.2390	0.0042(0.0024)	3.12 ;3.55 ;4.05
Haute Saône	4.09	0.2706	0.0047(0.0027)	3.59 ;4.07 ;4.66
Territoire de Belfort	3.58	0.4356	0.0079(0.0044)	2.79 ;3.56 ;4.49
Franche Comté	**3.85**	**0.1484**	**0.0026(0.0014)**	**3.56 ;3.84 ;4.14**
Loire Atlantique	3.54	0.1528	0.0026(0.0015)	3.24 ;3.53 ;3.84
Maine et Loire	4.18	0.1440	0.0024(0.0014)	3.90 ;4.18 ;4.47
Mayenne	4.64	0.4118	0.0071(0.0040)	3.88 ;4.63 ;5.48
Sarthe	4.21	0.2969	0.0050(0.0029)	3.64 ;4.20 ;4.82
Vendée	3.61	0.2370	0.0042(0.0023)	3.17 ;3.60 ;4.10
Pays de la Loire	**3.90**	**0.0962**	**0.0017(0.0009)**	**3.72 ;3.90 ;4.09**
Côtes d'Armor	3.86	0.2361	0.0041(0.0023)	3.41 ;3.86 ;4.34
Finistère	3.94	0.1490	0.0026(0.0015)	3.65 ;3.94 ;4.23
Ille de Vilaine	4.47	0.2361	0.0040(0.0023)	4.03 ;4.47 ;4.94
Morbihan	3.66	0.2272	0.0039(0.0022)	3.23 ;3.66 ;4.12
Bretagne	**4.02**	**0.1072**	**0.0018(0.0010)**	**3.82 ;4.02 ;4.24**
Charente	4.69	0.4345	0.0075(0.0042)	3.90 ;4.68 ;5.59
Charente Maritime	4.61	0.1922	0.0032(0.0018)	4.25 ;4.61 ;5.00
Deux Sèvres	3.95	0.3508	0.0059(0.0034)	3.29 ;3.94 ;4.68
Vienne	4.25	0.3177	0.0055(0.0031)	3.66 ;4.25 ;4.90
Poitou Charente	**4.40**	**0.1539**	**0.0026(0.0015)**	**4.11 ;4.40 ;4.72**
Dordogne	4.58	0.4502	0.0079(0.0045)	3.76 ;4.56 ;5.51
Gironde	4.57	0.1663	0.0028(0.0016)	4.25 ;4.57 ;4.90
Landes	4.10	0.2486	0.0044(0.0024)	3.63 ;4.09 ;4.60
Lot et Garonne	4.54	0.3573	0.0061(0.0035)	3.85 ;4.53 ;5.27
Pyrénées Atlantiques	4.16	0.1743	0.0031(0.0017)	3.84 ;4.15 ;4.51
Aquitaine	**4.43**	**0.1124**	**0.0019(0.0011)**	**4.21 ;4.43 ;4.66**
Ariège	4.35	0.3294	0.0058(0.0032)	3.74 ;4.34 ;5.02
Aveyron	3.72	0.5218	0.0091(0.0052)	2.79 ;3.70 ;4.81
Haute Garonne	4.31	0.1965	0.0033(0.0019)	3.94 ;4.31 ;4.71
Gers	5.01	0.5421	0.0092(0.0053)	4.04 ;4.99 ;6.15
Lot	2.86	0.6569	0.0118(0.0066)	1.76 ;2.79 ;4.31
Hautes Pyrénées	4.93	0.3137	0.0053(0.0030)	4.35 ;4.93 ;5.56
Tarn	4.37	0.3189	0.0054(0.0031)	3.78 ;4.36 ;5.05
Tarn et Garonne	4.11	0.8254	0.0145(0.0082)	2.66 ;4.06 ;5.86
Midi-Pyrénées	**4.26**	**0.1417**	**0.0024(0.0014)**	**3.99 ;4.25 ;4.55**
Corrèze	4.81	0.2799	0.0047(0.0027)	4.27 ;4.80 ;5.38
Creuse	4.19	0.3314	0.0058(0.0033)	3.57 ;4.18 ;4.90
Haute Vienne	5.04	0.4487	0.0075(0.0043)	4.19 ;5.03 ;5.96
Limousin	**4.82**	**0.2479**	**0.0041(0.0024)**	**4.35 ;4.81 ;5.32**

Ain	3.32	0.1503	0.0026(0.0015)	3.04 ;3.32 ;3.61
Ardèche	3.62	1.5220	0.0257(0.0148)	1.39 ;3.37 ;7.15
Drôme	3.78	0.3357	0.0057(0.0032)	3.15 ;3.77 ;4.44
Isère	3.28	0.1714	0.0031(0.0017)	2.96 ;3.27 ;3.63
Loire	4.02	0.2289	0.0039(0.0022)	3.58 ;4.01 ;4.48
Rhône	3.17	0.1095	0.0019(0.0011)	2.96 ;3.17 ;3.40
Savoie	3.22	0.2863	0.0051(0.0029)	2.68 ;3.21 ;3.80
Haute Savoie	2.80	0.1821	0.0033(0.0018)	2.46 ;2.79 ;3.17
Rhône Alpes	**3.35**	**0.1002**	**0.0017(0.0009)**	**3.17 ;3.34 ;3.56**
Allier	5.34	0.5420	0.0090(0.0052)	4.34 ;5.32 ;6.44
Cantal	3.99	0.3262	0.0057(0.0032)	3.40 ;3.98 ;4.67
Haute Loire	3.87	0.2935	0.0051(0.0029)	3.32 ;3.86 ;4.47
Puy de Dôme	4.09	0.2922	0.0051(0.0029)	3.54 ;4.08 ;4.69
Auvergne	**4.37**	**0.2077**	**0.0035(0.0020)**	**3.98 ;4.37 ;4.79**
Aude	4.34	0.5886	0.0105(0.0059)	3.29 ;4.30 ;5.61
Gard	4.48	0.2562	0.0044(0.0025)	4.00 ;4.47 ;4.98
Hérault	3.97	0.1652	0.0028(0.0016)	3.67 ;3.97 ;4.31
Lozère	NA	NA	NA(NA)	NA ;NA ;NA
Pyrénées Orientales	4.71	0.2704	0.0046(0.0026)	4.20 ;4.70 ;5.25
Languedoc-Roussillon	**4.30**	**0.1354**	**0.0023(0.0013)**	**4.04 ;4.29 ;4.57**
Alpes de Haute Provence	4.26	0.3387	0.0058(0.0033)	3.62 ;4.25 ;4.93
Hautes Alpes	3.63	0.3488	0.0060(0.0034)	2.99 ;3.62 ;4.36
Alpes Maritimes	3.46	0.1104	0.0019(0.0011)	3.25 ;3.46 ;3.69
Bouches du Rhône	3.98	0.0959	0.0017(0.0009)	3.80 ;3.98 ;4.18
Var	3.99	0.1304	0.0022(0.0012)	3.74 ;3.99 ;4.25
Vaucluse	3.82	0.1845	0.0032(0.0018)	3.47 ;3.82 ;4.20
PACA(*)	**3.85**	**0.0599**	**0.0010(0.0006)**	**3.73 ;3.85 ;3.97**
Haute Corse	6.07	1.1258	0.0187(0.0109)	4.12 ;5.97 ;8.50
Corse du Sud	1.81	0.2470	0.0049(0.0026)	1.38 ;1.79 ;2.33
Corse	**4.13**	**0.62**	**0.0103(0.006)**	**3.05 ;4.08 ;5.46**
France Métropolitaine	**3.93**	**0.0233**	**0.0004(0.0002)**	**3.89 ;3.93 ;3.98**

Bibliographie

[1] Congdon P. (2005), *Bayesian models for categorical data*, John Wiley&Sons, Chichester-England

[2] Cowles M.K. et Carlin B.P. (1996), Markov chain Monte Carlo convergence diagnostics:a comparative review, *Journal of the American Statistical Association*,91,883-904

[3] Gelfand A.E. (1996), *Markov Chain Monte Carlo in practice*, édité par Gilks W.R., Richardson S., Spiegelhalter D.J., chapitre 9, 145-158

[4] Gelman A. et Rubin D.B. (1992), Inference from iterative simulation using multiple sequences. *Statistical science*, 7, 457-472

[5] Gelman A., Carlin B.B., Stern H.S. et Rubin D.B., (1995), Bayesian data analysis, Chapman and Hall, New York, chapitre 7.

[6] Gilks W. (1992), Derivative-free adaptive rejection sampling for Gibbs sampling. In *Bayesian statistics 4* édité par Bernardo J.M., Berger J.O., Dawid A.P., Smith A.F.M. Oxford University Press,U.K., 641-665.

[7] Neal R. (1997), Markov chain Monte Carlo methods based on slicing the density function, Technical report 9722, Department of statistics, University of Toronto.

[8] Rao J.N.K. (2003), *Small area estimation*, John Wiley&Sons, Hoboken-New Jersey