Modèles de Markov Latents et Modèles de Mélange Finis pour l'Estimation sur Petits Domaines

M. Giovanna Ranalli[1]

[1]Université de Pérouse – ITALIE

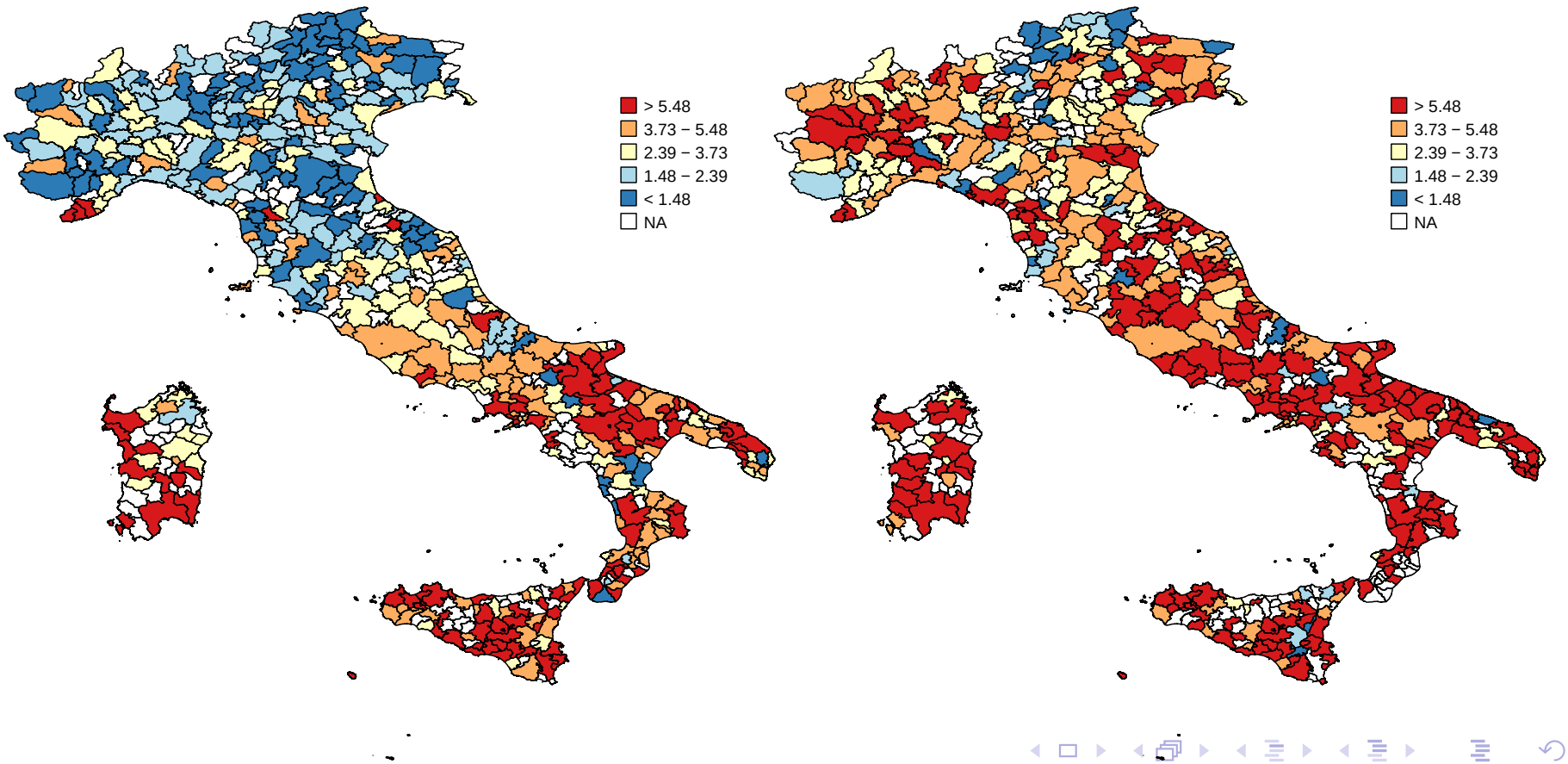
9e Colloque francophone sur le sondages
À l'Université du Québec en Outaouais
Gatineau | Quebéc | Canada, 11 – 14 Octobre, 2016

Motivation – Enquête Emploi Italienne (LFS) (1)

- L'Enquête Emploi (Labor Force Survey, LFS) est une enquête trimestrielle à deux degrés – commune/ménage.
- Pour chaque échantillon trimestriel environ 1350 communes et 100,000 individus sont concernés.
- Depuis 2000, ISTAT produit des estimations annuelles issues de LFS de nombres/taux d'actifs et d'inactifs relatifs aux Local Labour Market Areas (LLMAs).
- Les LLMAs sont des domaines non plannifiés obtenus à partir de grappes de communes intersectées avec des provinces (LAU1) qui constituent les domaines plannifiés de LFS les plus fins.

Motivation – Enquête Emploi Italienne (LFS) (1)

Taux de chômage. Estimateurs directs (%). 2004.1 et 2014.4



Motivation – Enquête Emploi Italienne (LFS) (2)

- Depuis 2004, après la refonte de la stratégie d'échantillonnage de LFS, on utilise un estimateur EBLUP au niveau unité avec des effets aléatoires par domaines autocorrélés spatialement.
- En 2011, avec le dernier Recensement, les LLMAs ont été redéfinis en termes de flux de déplacement quotidiens liés au travail

Occasion de repenser la stratégie d'estimation sur petits domaines

- Evaluation des méthodologies existantes

$\Rightarrow$ Test de nouvelles méthodologies

Aperçu de la présentation

- Les modèles d'estimation sur petits domaines utilisent des effets aléatoires pour modéliser l'hétérogénéité non capturée par les covariables
- Ces effets aléatoires sont supposés avoir une distribution CONTINUE (en général Normale)
- Nous voulons examiner l'utilisation de distributions DISCRÈTES
  - $\Rightarrow$ pour prendre en compte la non-normalité
  - $\Rightarrow$ pour obtenir une classification des petits domaines
  - $\Rightarrow$ pour réduire le volume de calcul dans certains cas

Deux différents cadres

SAE avec Séries Temporelles pour les taux de chômages

- Apporte de la force dans le temps
- Modèle au niveau domaine
- Approche Bayes hiérarchique
- Réponse Normale
- Introduit les Modèles de Markov Latents en SAE
- Enquête LFS: données trimestrielles 2004-2014

Meilleure Prédiction Empirique SAE avec données binaires

- Transversal
- Modèle au niveau unité
- Approche par Maximum de vraisemblance
- Réponse binaire
- Introduit des Mélanges Finis en SAE
- Enquête LFS: 4ème trimestre 2011 (A FAIRE)

Collaborateurs et remerciements

Je ne pourrais pas tout faire moi-même!

- F. Bartolucci, G. Bertarelli, M.F. Marino (Université de Pérouse)
- N. Salvati (Université de Pise)
- M. D'Aló, F. Solari (ISTAT, Institut National de Statistique)
- M. Alfó (Université Sapienza, Rome)

Projet PRIN-SURWEY

SUrvey Research on households WEalth and Youth unemployment
`http://www.sp.unipg.it/surwey/`

Plan

Notation

Modèle [Rao et Yu, 1994]

$$\begin{aligned} \text{Modèle d'échantillonnage} \quad & \hat{\theta}_{it} = \theta_{it} + e_{it} \\ \text{Modèle de lien} \quad & \theta_{it} = \boldsymbol{x}_{it}^T \boldsymbol{\beta} + v_i + \alpha_{it} \\ & i = 1, \ldots, m, \; t = 1, \ldots, T. \end{aligned}$$

- $\hat{\theta}_{it} \longrightarrow$ estimateur direct pour le petit domaine i au temps t;
- $\theta_{it} = g(\bar{Y}_{it}) \longrightarrow$ fonction de la moyenne sur le petit domaine.
- $e_{it}|\theta_{it}$ sont des erreurs d'échantillonnage normalement distribuées, de moyenne zéro, de matrice de variance-covariance connue $\boldsymbol{\Psi} = \text{blockdiag}\{\boldsymbol{\Psi}_i\}$

Notation

Modèle [Rao et Yu, 1994]

$$\begin{aligned} \text{Modèle d'échantillonnage} \quad & \hat{\theta}_{it} = \theta_{it} + e_{it} \\ \text{Modèle de lien} \quad & \theta_{it} = \boldsymbol{x}_{it}^T \boldsymbol{\beta} + v_i + \alpha_{it} \\ & i = 1, \ldots, m, \; t = 1, \ldots, T. \end{aligned}$$

- $\boldsymbol{x}_{it}$ covariables spécifiques du domaines (pouvant varier avec le temps) + effets saisonniers pour des séries temporelles longues
- $v_i \sim N(0, \sigma_v^2) \rightarrow$ effet domaine
- $\alpha_{it} = \rho \alpha_{i,t-1} + \epsilon_{it}$, $|\rho| < 1$, $\epsilon_{it} \sim N(0, \sigma_\epsilon^2) \rightarrow$ effet domaine-par-temps
- Les e_{it}, v_i et ϵ_{it} sont indépendants entre eux

Autres approches

- Modèle AR(1) pour $\alpha_{it} \rightarrow$ Modèles ARMA plus généraux
- [Datta et al., 1999], [Datta et al., 2002], [You, Rao, Gambino, 2003] utilisent un modèle de marche aléatoire

$$\alpha_{it} = \alpha_{i,t-1} + \epsilon_{it},$$

  i.e. avec $\rho = 1$
- Modèles Dynamiques/d'espace d'état pour des coefficients variant dans le temps $\boldsymbol{\beta}_{it}$ [Pfefferman et Burck, 1990] ou pour les erreurs sur les domaines [Ghosh et al., 1996].

Aperçu des Modèles de Markov Latents (LMM) (1)

Voir le livre de [Bartolucci, Farcomeni, Pennoni, 2013]

- Les LMMs sont une famille particulière de modèles pour les données longitudinales
- Les LMMs supposent l'existence d'un processus latent qui affecte la distribution des variables de réponse
- Le processus latent est supposé généré selon une chaîne de Markov avec un certain nombre d'états (états latents)
- Les LMMs generalisent les Classes Latentes (dynamiques) [Fabrizi et al., 2016]
- Les LMMs étendent les modèles de Chaînes de Markov (erreur de mesure)

Aperçu des Modèles de Markov Latents (LMM) (2)

- Modèle de mesure: distribution de la variable de réponse sachant le processus latent
- Modèle Latent: distribution du processus latent
- Les covariables peuvent intervenir dans le modèle de mesure ou dans le modèle latent.

Hétérogénéité non observée

- Elle est prise en compte par des variables latentes discrètes (vs. continues)
- Elle est modélisée de façon dynamique (chaque domaine peut se déplacer entre plusieurs états latents au cours du temps)

Aperçu des Modèles de Markov Latents (LMM) (2)

- Modèle de mesure: distribution de la variable de réponse sachant le processus latent
- Modèle Latent: distribution du processus latent
- Les covariables peuvent intervenir dans le modèle de mesure ou dans le modèle latent.

Hétérogénéité non observée

- Elle est prise en compte par des variables latentes discrètes (vs. continues)
- Elle est modélisée de façon dynamique (chaque domaine peut se déplacer entre plusieurs états latents au cours du temps)

LMMs Spécification pour SAE

MODÈLE d'ÉCHANTILLONNAGE: $\widehat{\boldsymbol{\theta}}_i|\boldsymbol{\theta}_i \sim N_T(\boldsymbol{\theta}_i, \boldsymbol{\Psi}_i)$

MODÈLE DE LIEN: Modèle de mesure

Modèle Latent

Modèle de Lien

Modèle de mesure

$$\theta_{it}|U_{it} = u, \boldsymbol{x}_{it} \sim N(\boldsymbol{x}_{it}^T \boldsymbol{\beta}_{u(it)}, \sigma^2_{u(it)})$$

$U_{it} \in \{1, \ldots, k\}$ état latent, $k \ll m$

$u(it)$ est l'état latent du domaine i au temps t,

$i = 1, \ldots, m,\ t = 1, \ldots, T.$

Modèle Latent

Probabilités Initiales $P(U_{i1} = u) = \pi_{(u)}$

$t = 1;\ u = 1, \ldots, k$

Probabilités de Transition $P(U_{it} = u|U_{i,t-1} = \bar{u}) = \pi_{(u|\bar{u})}$

$t = 2, \ldots, T;\ u, \bar{u} = 1, \ldots, k.$

Modèle de Lien

Modèle de mesure

$$\theta_{it}|U_{it} = u, \boldsymbol{x}_{it} \sim N(\boldsymbol{x}_{it}^T \boldsymbol{\beta}_{u(it)}, \sigma^2_{u(it)})$$

$U_{it} \in \{1, \ldots, k\}$ état latent, $k \ll m$

$u(it)$ est l'état latent du domaine i au temps t,

$i = 1, \ldots, m,\ t = 1, \ldots, T.$

Modèle Latent

Probabilités Initiales $P(U_{i1} = u) = \pi_{(u)}$

$t = 1;\ u = 1, \ldots, k$

Probabilités de Transition $P(U_{it} = u|U_{i,t-1} = \bar{u}) = \pi_{(u|\bar{u})}$

$t = 2, \ldots, T;\ u, \bar{u} = 1, \ldots, k.$

Une comparison avec [Datta et al., 1999] et [You, Rao, Gambino, 2003]

Modèles de Lien

$$\text{You-Rao-Gambino:} \quad \theta_{it} = \boldsymbol{x}_{it}^T \boldsymbol{\beta} + \alpha_{it} + v_i$$

$$\text{Datta et al.:} \quad \theta_{it} = \boldsymbol{x}_{it}^T \boldsymbol{\beta}_i + \alpha_{it} + v_i$$

$$\text{LMMs:} \quad \theta_{it} = \boldsymbol{x}_{it}^T \boldsymbol{\beta}_{u(it)}$$

$u(it)$ est l'état latent du domaine i au temps t, $u = 1, \ldots, k$.
Noter que $k \ll m$.

Estimation – approche HB

- Paramètres du modèle $\rightarrow \boldsymbol{\lambda} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k, \sigma_1^2, \ldots, \sigma_k^2, \boldsymbol{\pi}, \boldsymbol{\Pi})$
- Paramètres d'intérêt sur petits domaines $\rightarrow$ $\boldsymbol{\mu} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)^T$
- Echantillonneur de Gibbs avec Augmentation de Données [Tanner et Wong, 1987, Van Dyk et Meng, 2001]

Gibbs Conditionnel

$$[\ldots] \sim [\ldots]$$

$$[\theta_{it}|\boldsymbol{U}, \boldsymbol{\pi}, \boldsymbol{\Pi}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \widehat{\boldsymbol{\theta}}] \sim \mathsf{N}(\hat{\theta}^B_{it}(\boldsymbol{U}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2), \gamma_{it}\psi_{it})$$

- $\hat{\theta}^B_{it}(\boldsymbol{U}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \gamma_{it}\hat{\theta}_{it} + (1-\gamma_{it})\boldsymbol{x}^T_{it}\boldsymbol{\beta}_{u(it)}$
- $\gamma_{it} = \sigma^2_{u(it)}/(\sigma^2_{u(it)} + \psi_{it})$

En dehors des domaines de l'échantillon

$$[\hat{\theta}_{it}|\boldsymbol{U}, \boldsymbol{\pi}, \boldsymbol{\Pi}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}] \sim \mathsf{N}(\hat{\theta}^D_{it}(\boldsymbol{U}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2), \gamma^*_{it}\psi_{it})$$

- $\hat{\theta}^D_{it}(\boldsymbol{U}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \gamma^*_{it}\hat{\theta}_{i,t-1} + (1-\gamma^*_{it})\boldsymbol{x}^T_{it}\boldsymbol{\beta}_{u(i,t-1)}$
- $\gamma^*_{it} = \sigma^2_{u(i,t-1)}/(\sigma^2_{u(i,t-1)} + \psi_{it})$

Gibbs Conditionnel

$$[\ldots] \sim [\ldots]$$

$$[\theta_{it}|\boldsymbol{U}, \boldsymbol{\pi}, \boldsymbol{\Pi}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \widehat{\boldsymbol{\theta}}] \sim \mathsf{N}(\hat{\theta}^B_{it}(\boldsymbol{U}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2), \gamma_{it}\psi_{it})$$

- $\hat{\theta}^B_{it}(\boldsymbol{U}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \gamma_{it}\hat{\theta}_{it} + (1-\gamma_{it})\boldsymbol{x}^T_{it}\boldsymbol{\beta}_{u(it)}$
- $\gamma_{it} = \sigma^2_{u(it)}/(\sigma^2_{u(it)} + \psi_{it})$

En dehors des domaines de l'échantillon

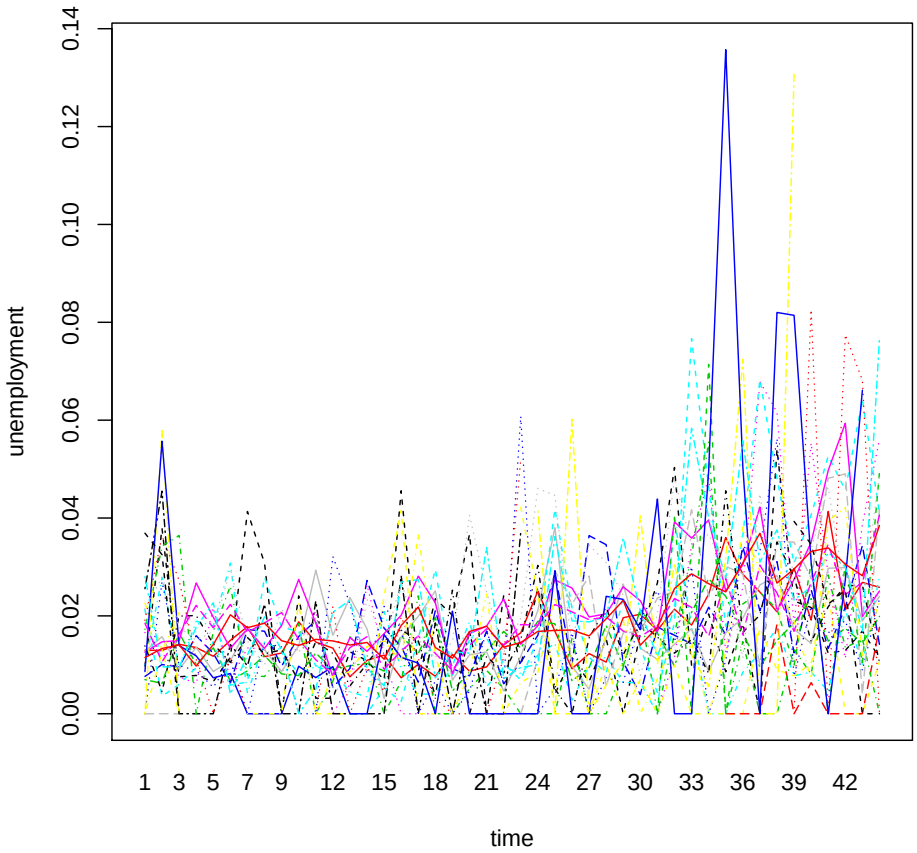
$$[\hat{\theta}_{it}|\boldsymbol{U}, \boldsymbol{\pi}, \boldsymbol{\Pi}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}] \sim \mathsf{N}(\hat{\theta}^D_{it}(\boldsymbol{U}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2), \gamma^*_{it}\psi_{it})$$

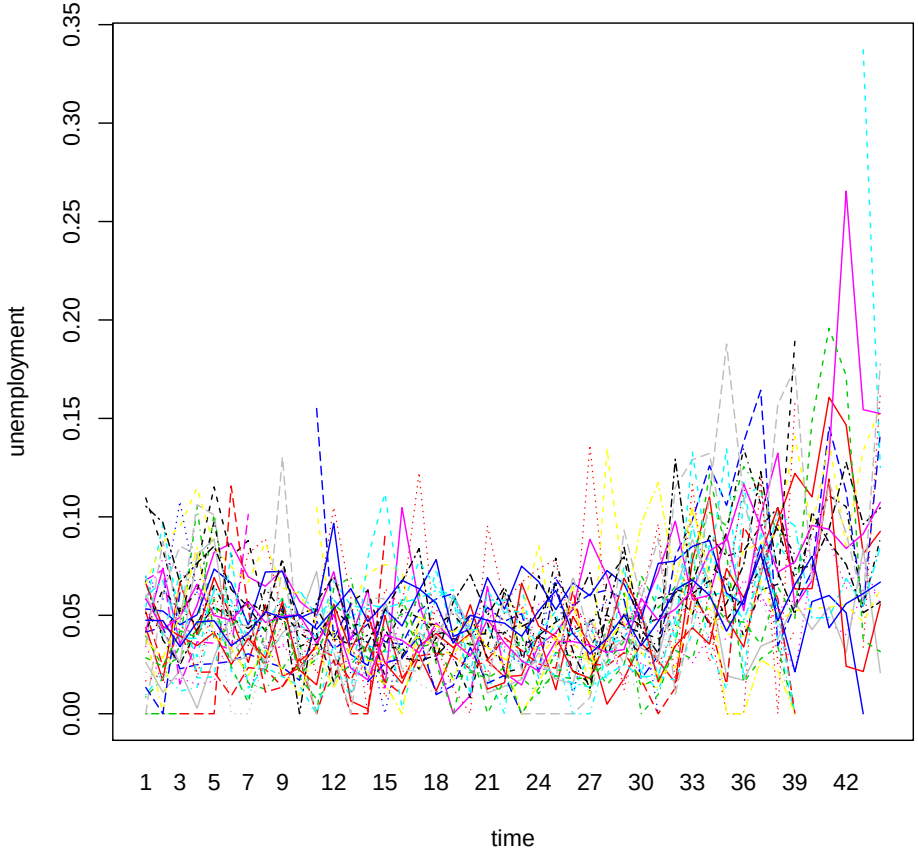
- $\hat{\theta}^D_{it}(\boldsymbol{U}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \gamma^*_{it}\hat{\theta}_{i,t-1} + (1-\gamma^*_{it})\boldsymbol{x}^T_{it}\boldsymbol{\beta}_{u(i,t-1)}$
- $\gamma^*_{it} = \sigma^2_{u(i,t-1)}/(\sigma^2_{u(i,t-1)} + \psi_{it})$

Données LFS

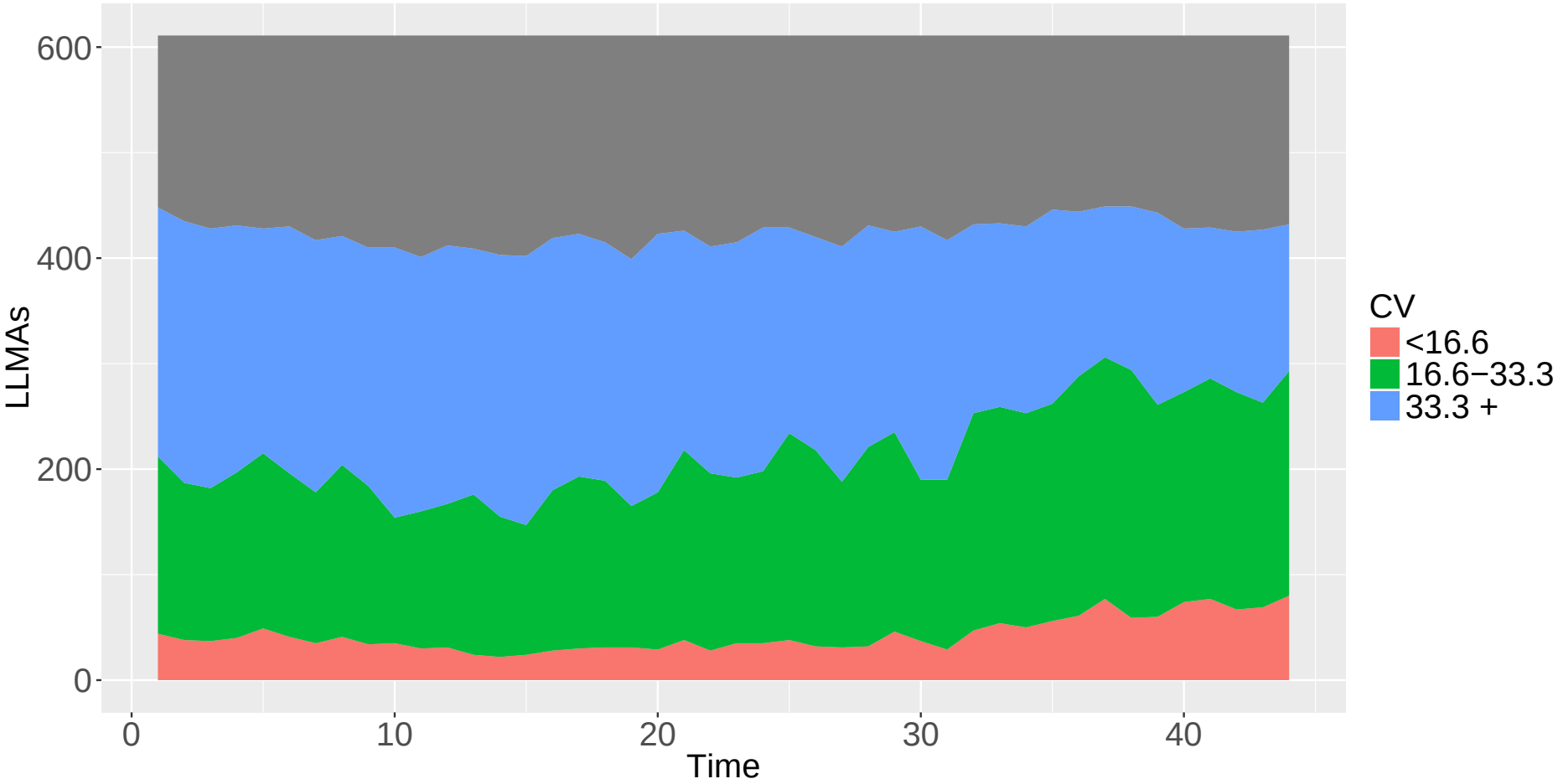
- Estimateurs directs des taux de chômage sur les petits domaines
- $m = 611$ LLMAs (Local Labour Market Areas)
- $T = 44$ données trimestrielles de 2004 à 2014
- en tout $611 \times 44 = 26884$ points dans l'échantillon

Séries temporelles – Regions 4 (Nord) et 17 (Sud)





Estimateurs directs – % CVs



Comparaison des Estimateurs

Information auxiliaire

- population par `sexe` $\times$ 7 classes d'âge
- occupation culturelle (4 catégories)
- spécialisation dominante (5 catégories)
- année et effet du trimestre (variables muettes)

DIR Estimateur Direct

FH Estimateur Fay-Harriot – transversal

YRG Estimateur You-Rao-Gambino – $\rho = 1$

Datta Datta et al. – trop de domaines et de données

LMM Estimateur basé sur le Modèle de Markov Latent

- 30,000 chanes MCMC
- 15,000 burn-in

Comparaison des Estimateurs

Information auxiliaire

- population par `sexe` $\times$ 7 classes d'âge
- occupation culturelle (4 catégories)
- spécialisation dominante (5 catégories)
- année et effet du trimestre (variables muettes)

DIR Estimateur Direct

FH Estimateur Fay-Harriot – transversal

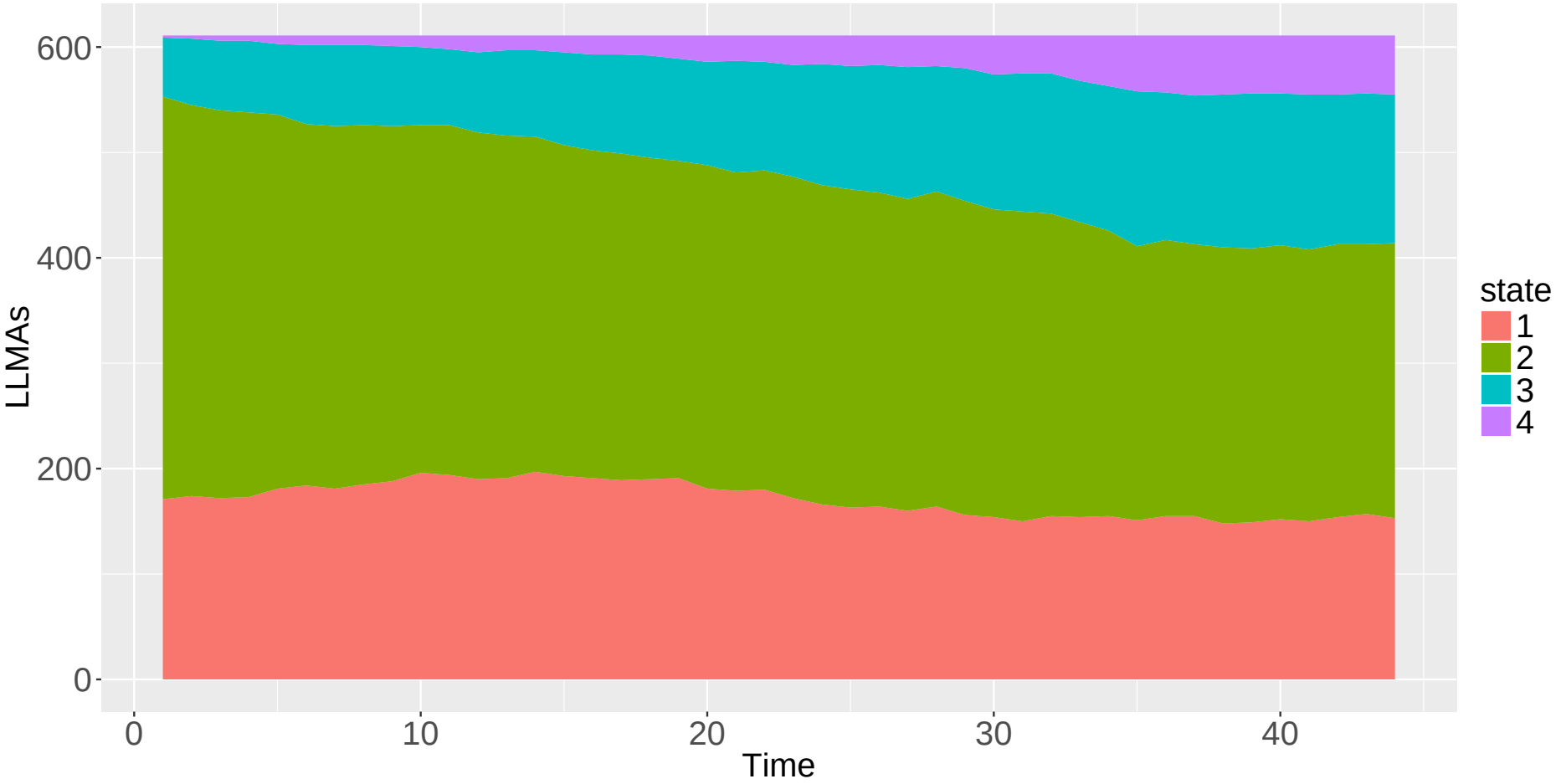
YRG Estimateur You-Rao-Gambino – $\rho = 1$

Datta Datta et al. – trop de domaines et de données

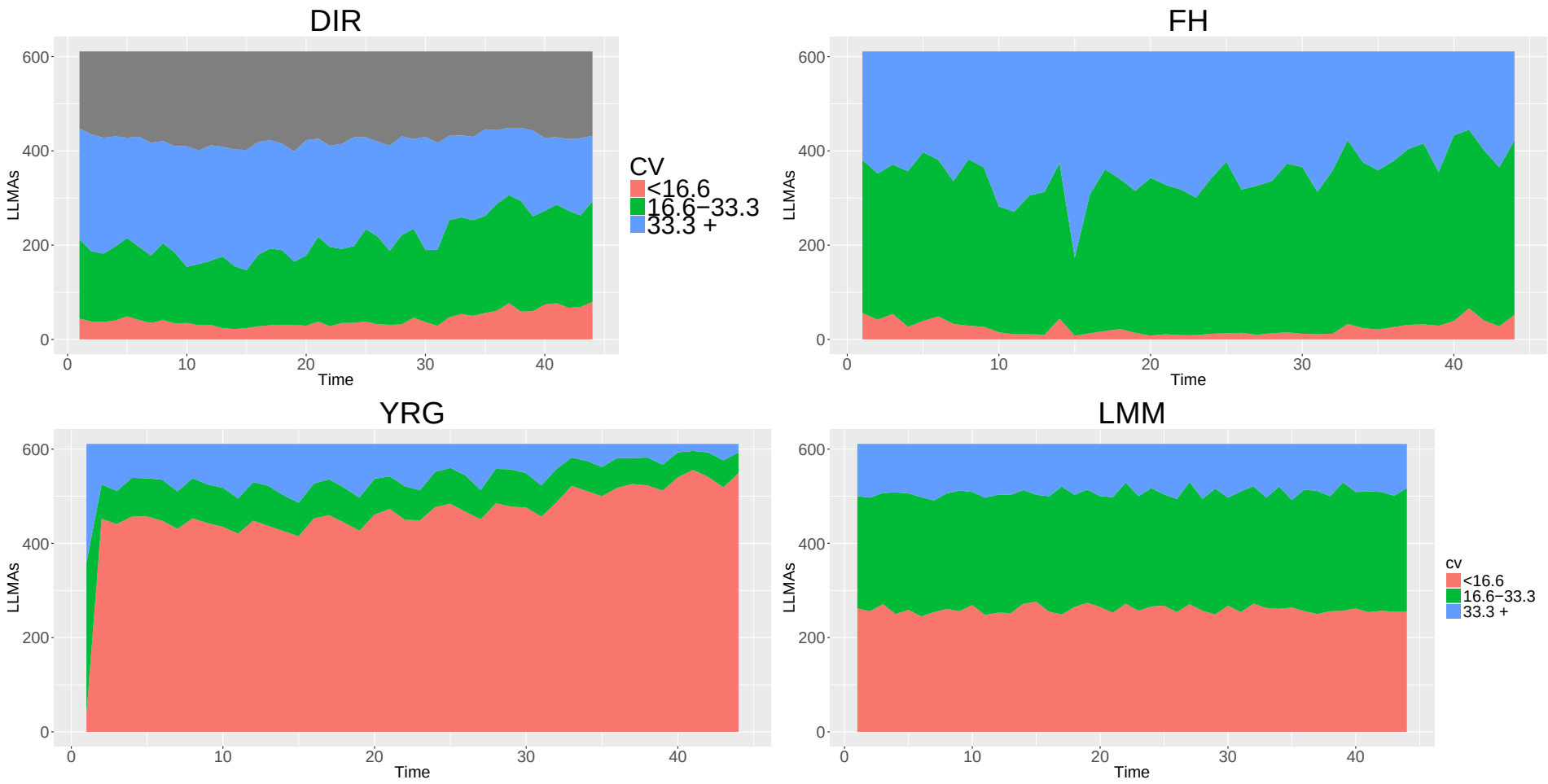
LMM Estimateur basé sur le Modèle de Markov Latent

- 30,000 chanes MCMC
- 15,000 burn-in

Distribution sur les 4 états latents

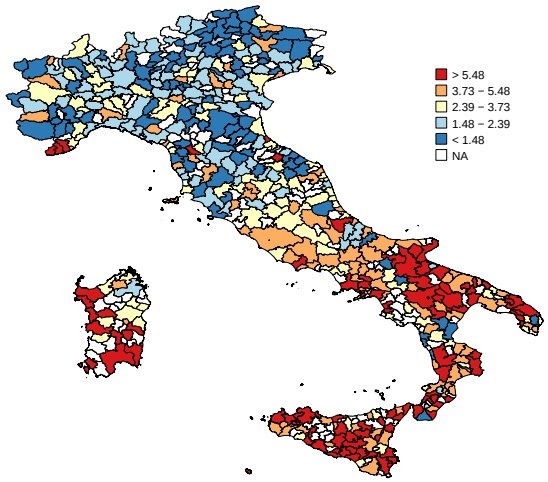


Comparaison de CV

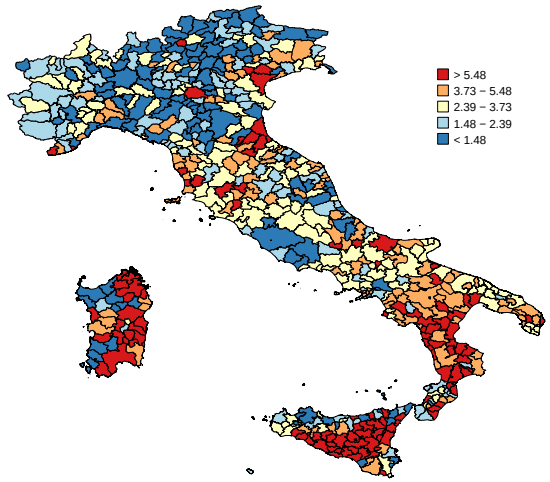


Estimations 2004.1

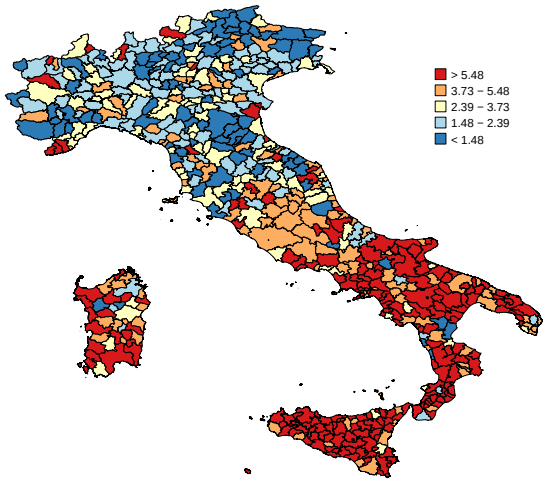
DIR



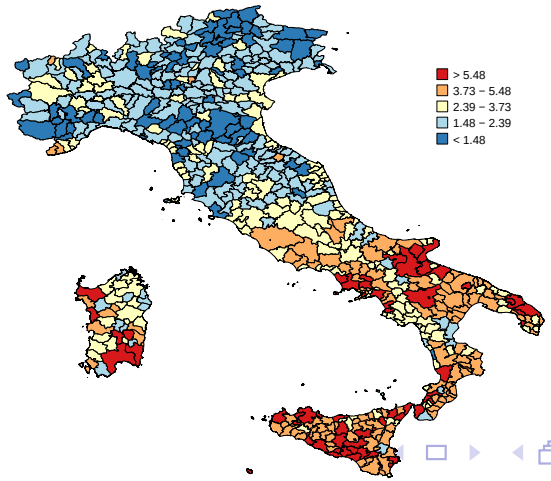
FH



YRG

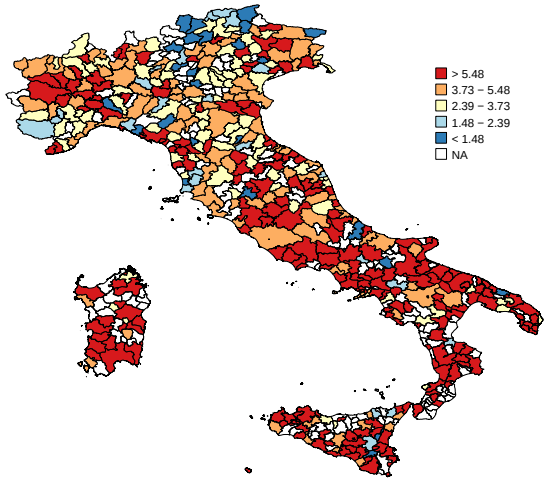


LMM

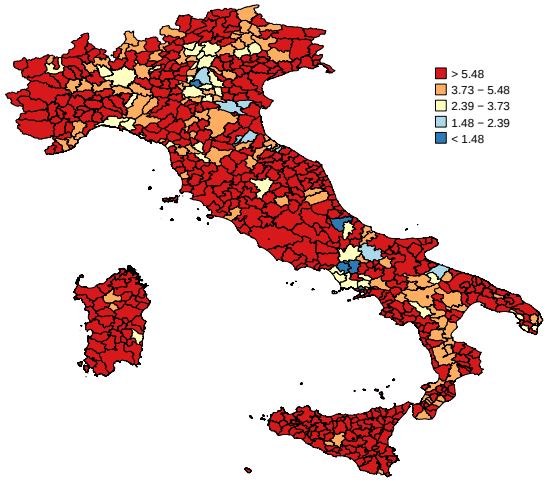


Estimations 2014.4

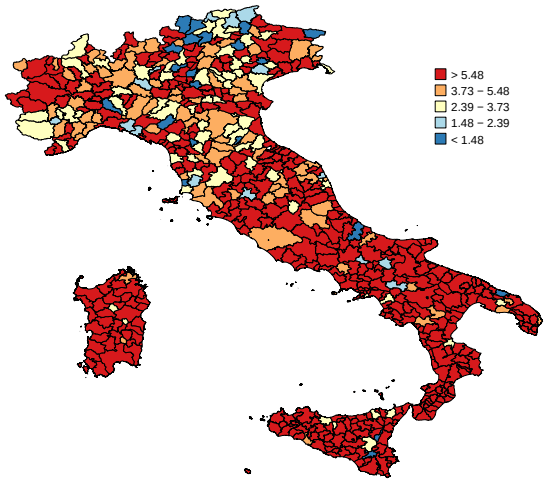
DIR



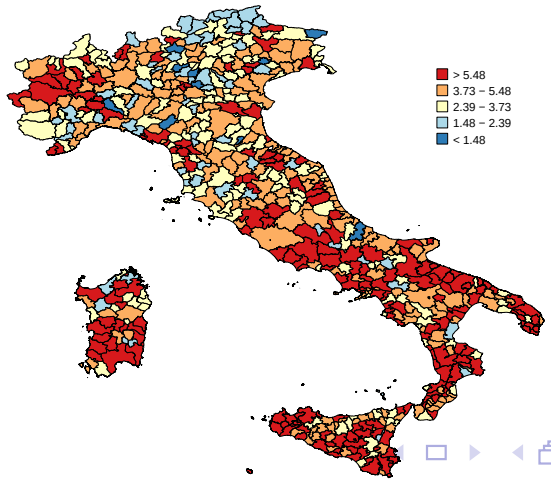
FH



YRG



LMM



Comparaison avec les valeurs du Recensement 2011

Erreur Relative Absolue (ARE) des estimations de 2011_4

$$\text{ARE} = |\hat{\theta}_i - \text{Cens}_i|/\text{Cens}_i$$

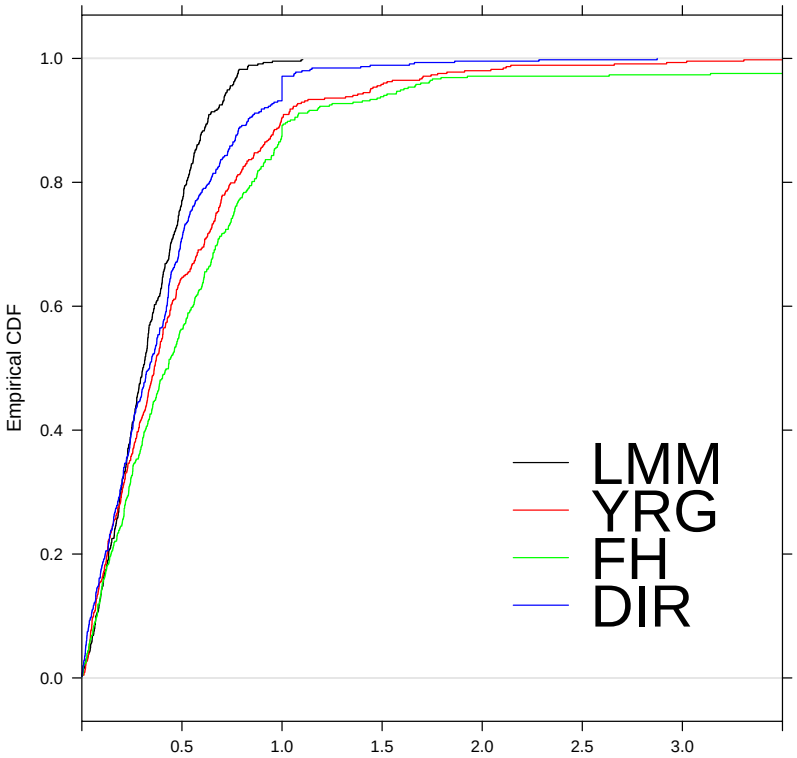
Table: LLMAs observées

	DIR	LMM	YRG	FH
Min.	0.001	0.000	0.000	0.000
1st Qu.	0.157	0.172	0.159	0.202
Median	0.338	0.304	0.362	0.432
Mean	0.399	0.336	0.494	0.625
3rd Qu.	0.541	0.485	0.669	0.759
Max.	2.874	1.103	4.386	10.745

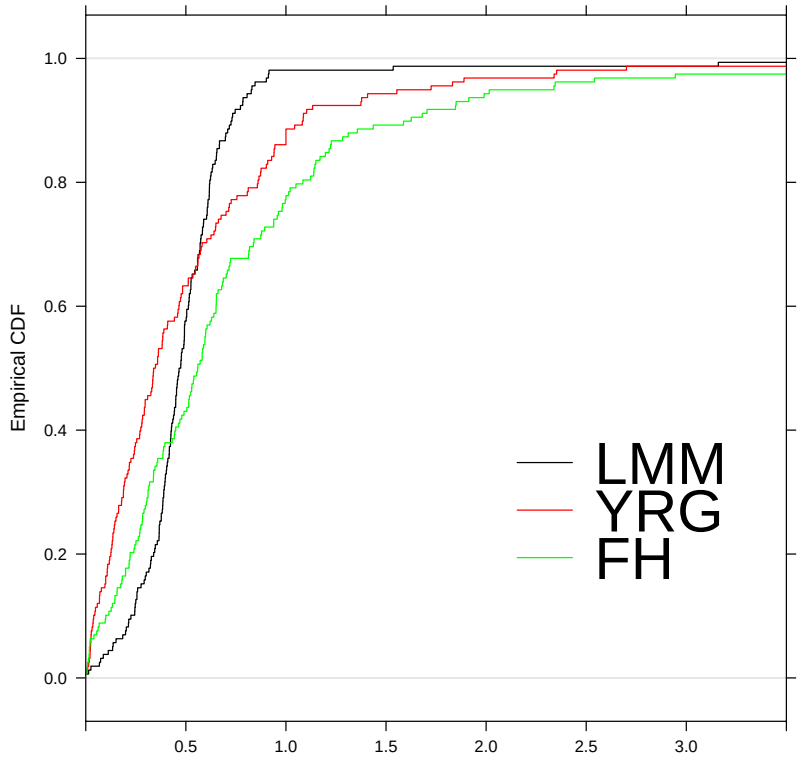
Table: LLMAs non observées

LMM	YRG	FH
0.000	0.005	0.000
0.372	0.147	0.283
0.470	0.343	0.559
0.517	0.550	0.804
0.606	0.695	0.961
4.412	7.323	10.773

CDF empirique des ARE



(a) observed LLMAs



(b) non observed LLMAs

Plan

Deux différents cadres

SAE avec Séries Temporelles pour les taux de chômages

- Apporte de la force dans le temps
- Modèle au niveau domaine
- Approche Bayes hiérarchique
- Réponse Normale
- Introduit les Modèles de Markov Latents en SAE
- Enquête LFS: données trimestrielles 2004-2014

Meilleure Prédiction Empirique SAE avec données binaires

- Transversal
- Modèle au niveau unité
- Approche par Maximum de vraisemblance
- Réponse binaire
- Introduit des Mélanges Finis en SAE
- Enquête LFS: 4ème trimestre 2011 (A FAIRE)

Notation mise à jour

- Soit y_{ij}, une variable de réponse binaire (=1 inactif; =0 sinon) pour l'unité $j \in U_i$ dans le domaine $i = 1, \ldots, m$
- Soit $\boldsymbol{x}_{ij}$ un vecteur de dimension p de covariables individuelles
- Nous nous intéressons à la prédiction de proportions sur les petits domaines

$$\theta_i = \frac{1}{N_i} \sum_{j \in U_i} y_{ij}$$

Modèle et hypothèses

- Soit α_i un coefficient aléatoire spécifique du domaine distribué selon une loi Gaussienne $\Rightarrow \alpha_i \sim N(0, \sigma^2)$
- Conditionnellement à α_i, les réponses y_{ij} sont des variables aléatoires de Bernoulli indépendantes avec une probabilité de succès

$$p_{ij} = \Pr(y_{ij} = 1) = \frac{e^{\alpha_i + \boldsymbol{x}_{ij}^T \boldsymbol{\beta}}}{1 + e^{\alpha_i + \boldsymbol{x}_{ij}^T \boldsymbol{\beta}}}$$

Les approches possibles

- Approche naïve $\Rightarrow$ Prediction par substitution avec estimation PQL

$$\hat{\theta}_i = \left[\frac{1}{N_i} \sum_{j \in U_i} \frac{e^{\alpha_i + \boldsymbol{x}_{ij}^T \boldsymbol{\beta}}}{1 + e^{\alpha_i + \boldsymbol{x}_{ij}^T \boldsymbol{\beta}}} \right]_{(\alpha_i, \boldsymbol{\beta}) = (\hat{\alpha}_i, \hat{\boldsymbol{\beta}})}$$

- Approche EBP $\Rightarrow$ empirical best prediction [Jiang et Lahiri, 2001, Jiang, 1998] avec estimation par la méthode des moments

$$\hat{\theta}_i = \left[\int \left(\frac{1}{N_i} \sum_{j \in U_i} \frac{e^{\alpha_i + \boldsymbol{x}_{ij}^T \boldsymbol{\beta}}}{1 + e^{\alpha_i + \boldsymbol{x}_{ij}^T \boldsymbol{\beta}}} \right) f_\alpha(\alpha_i \mid \boldsymbol{y}_i) \, d\alpha_i \right]_{(\alpha_i, \boldsymbol{\beta}) = (\hat{\alpha}_i, \hat{\boldsymbol{\beta}})}$$

EBP par le Maximum de Vraisemblance

La vraisemblance sur les données observées pour l'ensemble des paramètres du modèle $\mathbf{\Phi}$ est

$$L(\mathbf{\Phi}) = \prod_{i=1}^{m} \int f_{y|\alpha}(\boldsymbol{y}_i \mid \alpha_i) f_{\alpha}(\alpha_i) d\alpha_i \qquad (1)$$

avec $f_{y|\alpha}(\boldsymbol{y}_i \mid \alpha_i)$ le produit de densités de Bernoulli $f(y_{ij} \mid \alpha_i)$

- Pour obtenir les estimations, nous devons traiter des intégrales en grande dimension qui n'ont pas de solution explicite
- On utilise typiquement des méthodes d'approximation pour maximiser $L(\mathbf{\Phi})$ et cela peut être délicat informatiquement

Problème avec l'EBP

- Calculer l'EBP est lent informatiquement à cause de la solution des intégrales qui n'ont pas de forme explicite (pour la méthode des moments et le Maximum de Vraisemblance)
- L'estimation du MSE peut être prohibitif informatiquement, même avec un nombre limité de domaines
- Quand on utilise l'EBP avec des modèles mixtes logistiques (distribution normale avec effets aléatoires), le Bootstrap est moins gourmand informatiquement que les estimations analytiques du MSE.
- Dans le LFS nous avons $\approx 100,000$ observations pour chaque trimestre et $m = 611 \rightarrow$ infaisable

Une approche par le Maximum de Vraisemblance Non Paramétrique [Aitkin, 1996]

Nous utilisons l'approche par le Maximum de Vraisemblance Non Paramétrique (NPML) pour estimer les paramètres du modèle, ce qui nous permet

- de calculer l'EBP et l'approximation analytique de son MSE
- d'éviter des hypothèses invérifiables sur $f_\alpha(\alpha_i)$
- de réduire considérablement le volume de calcul

Distribution non-paramétrique pour les coefficients aléatoires

- On laisse $f_\alpha(\cdot)$ non spécifiée, et on l'estime à partir des données
- On l'approxime par une distribution discrète sur $G < m$ positions $\{\alpha_1, \ldots, \alpha_G\}$, avec des probabilités associées définies par $\pi_g = \Pr(\alpha_i = \alpha_g)$, $i = 1, \ldots, m$ et $g = 1, \ldots, G$.

$$\alpha_i \sim \sum_{g=1}^{G} \pi_g \delta_{\alpha_g}$$

où δ_θ est une distribution ponctuelle de masse unité sur θ.

La vraisemblance

- Pour estimer $\mathbf{\Phi} = \{\alpha_1, \ldots, \alpha_G, \pi_1, \ldots, \pi_G, \boldsymbol{\beta}\}$, nous maximisons

$$L(\mathbf{\Phi}) = \prod_{i=1}^{m} \sum_{g=1}^{G} f_{y|\alpha}\left(\boldsymbol{y}_i \mid \alpha_g\right) \pi_g$$

- $L(\mathbf{\Phi})$ ressemble à la vraisemblance d'un mélange fini

Estimation du modèle

- Via l'algorithme EM [Dempster et al., 1977]
- Le nombre de composantes du mélange G est traité comme connu, et estimé à l'aide de techniques de sélection de modèle (e.g. AIC)
- Il est également possible d'estimer analytiquement la matrice de variance-covariance de $\hat{\boldsymbol{\Phi}}$ [Oakes, 1999]

(Empirical) Best Prediction

- Le meilleur predicteur (BP) de θ_i est donné par

$$\tilde{\theta}_i = E_\alpha(\theta_i \mid \boldsymbol{y}_i) = \sum_{g=1}^{G} \left(\frac{1}{N_i} \sum_{j \in U_i} \frac{e^{\alpha_g + \boldsymbol{x}_{ij}^T \boldsymbol{\beta}}}{1 + e^{\alpha_g + \boldsymbol{x}_{ij}^T \boldsymbol{\beta}}} \right) w_{ig} = \sum_{g=1}^{G} \theta_i(g)\, w_{ig}$$

où w_{ig} est la probabilité a posteriori que le petit domaine i appartienne à la composante g du mélange fini

- Le meilleur prédicteur empirique (EBP-np) de θ_i est obtenu en remplaçant les vrais paramètres du modèle par leurs estimateurs

$$\hat{\theta}_i = \sum_{g=1}^{G} \left(\frac{1}{N_i} \sum_{j \in U_i} \frac{e^{\hat{\alpha}_g + \boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}}}}{1 + e^{\hat{\alpha}_g + \boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}}}} \right) \hat{w}_{ig} = \sum_{g=1}^{G} \hat{\theta}_i(g)\, \hat{w}_{ig}$$

(Empirical) Best Prediction

- Le meilleur predicteur (BP) de θ_i est donné par

$$\tilde{\theta}_i = E_\alpha(\theta_i \mid \boldsymbol{y}_i) = \sum_{g=1}^{G} \left(\frac{1}{N_i} \sum_{j \in U_i} \frac{e^{\alpha_g + \boldsymbol{x}_{ij}^T \boldsymbol{\beta}}}{1 + e^{\alpha_g + \boldsymbol{x}_{ij}^T \boldsymbol{\beta}}} \right) w_{ig} = \sum_{g=1}^{G} \theta_i(g)\, w_{ig}$$

où w_{ig} est la probabilité a posteriori que le petit domaine i appartienne à la composante g du mélange fini

- Le meilleur prédicteur empirique (EBP-np) de θ_i est obtenu en remplaçant les vrais paramètres du modèle par leurs estimateurs

$$\hat{\theta}_i = \sum_{g=1}^{G} \left(\frac{1}{N_i} \sum_{j \in U_i} \frac{e^{\hat{\alpha}_g + \boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}}}}{1 + e^{\hat{\alpha}_g + \boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}}}} \right) \hat{w}_{ig} = \sum_{g=1}^{G} \hat{\theta}_i(g)\, \hat{w}_{ig}$$

Le MSE de EBP-np

- Il est possible d'obtenir l'expression analytique du MSE et de l'estimer comme dans [Jiang et Lahiri, 2001] (non traité ici, très technique!)
- Quand on utilise l'EBP avec des modèles logistiques mixtes, le Bootstrap est moins gourmand informatiquement que les estimations analytiques du MSE.
- Avec cette distribution discrète, le MSE analytique est beaucoup moins gourmand informatiquement que le Bootstrap (particulièrement quand m augmente) $\longrightarrow$ faisable

Etude par simulations - basées sur un modèle

- On simule $\alpha_i \sim N(0, 0.25)$ (pire des scénarios pour NPML)
- $y_{ij} \sim$ Bernoulli(p_{ij}) avec

$$\text{logit}(p_{ij}) = \alpha_i + x_{ij} \quad \text{et} \quad x_{ij} \sim \text{Unif}(-1, j/a)$$

  Cadre de [González-Manteiga et al., 2007].
- Nous comparons EBP-np avec
  - EBP (Intégration MC)
  - Approche Naïve (prédiction par substitution)
- Nous comparons l'estimateur analytique du MSE pour l'EBP-np avec
  - [Jiang et Lahiri, 2001] (Intégration MC - pour l'EBP)
  - [González-Manteiga et al., 2007] (Approximation linéaire - pour l'approche naïve)

Etude par simulations - basées sur un modèle

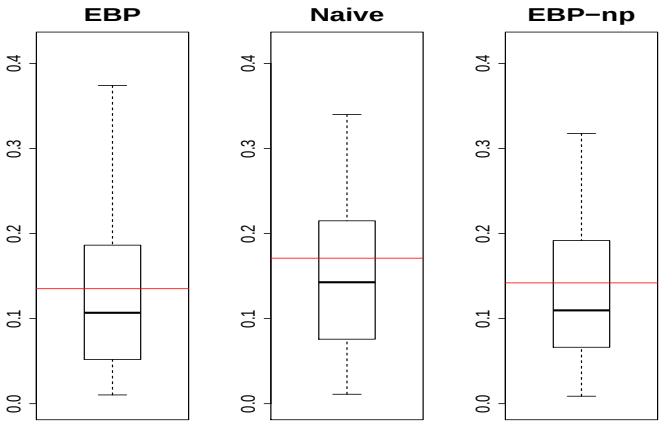
- On simule $\alpha_i \sim N(0, 0.25)$ (pire des scénarios pour NPML)
- $y_{ij} \sim$ Bernoulli(p_{ij}) avec

$$\text{logit}(p_{ij}) = \alpha_i + x_{ij} \quad \text{et} \quad x_{ij} \sim \text{Unif}(-1, j/a)$$

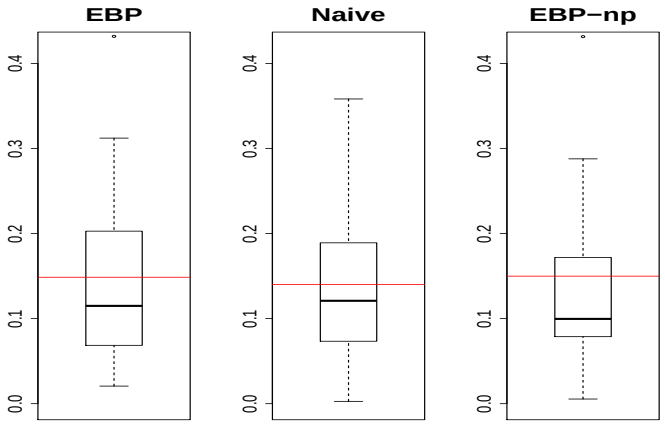
  Cadre de [González-Manteiga et al., 2007].
- Nous comparons EBP-np avec
  - EBP (Intégration MC)
  - Approche Naïve (prédiction par substitution)
- Nous comparons l'estimateur analytique du MSE pour l'EBP-np avec
  - [Jiang et Lahiri, 2001] (Intégration MC - pour l'EBP)
  - [González-Manteiga et al., 2007] (Approximation linéaire - pour l'approche naïve)

Biais absolu de $\hat{\theta}_i$

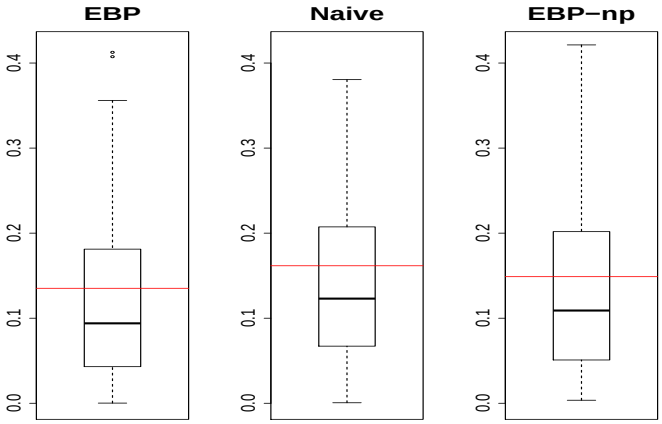
$m = 30, N_i = 100, n_i = 10$



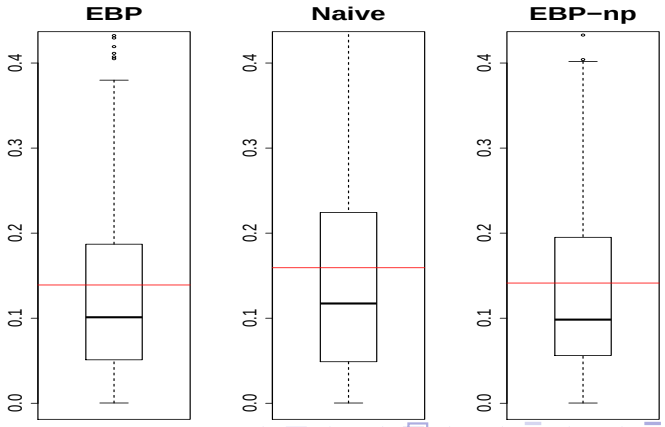
$m = 30, N_i = 1000, n_i = 10$



$m = 100, N_i = 100, n_i = 10$

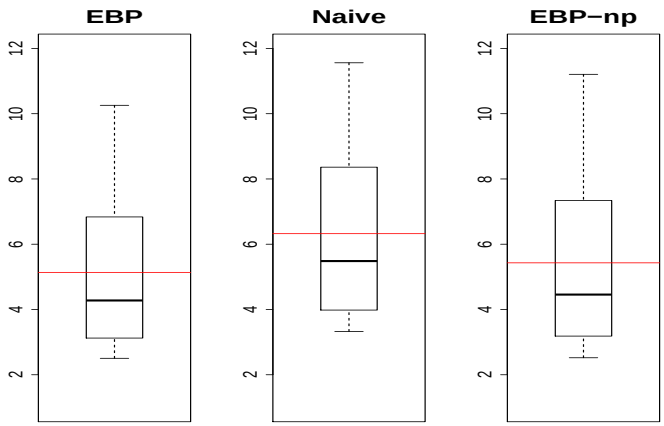


$m = 200, N_i = 100, n_i = 10$

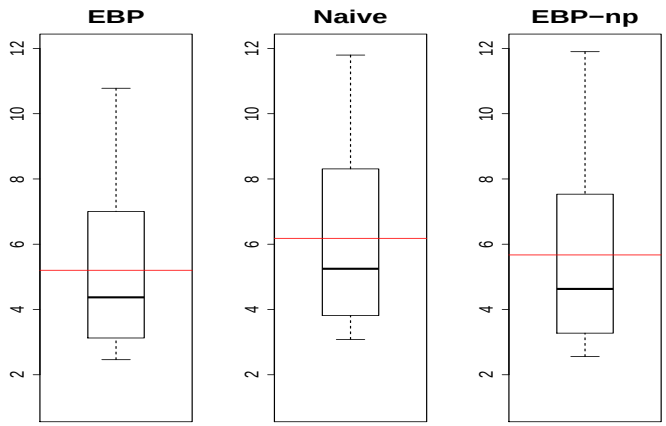


RMSE de $\hat{\theta}_i$

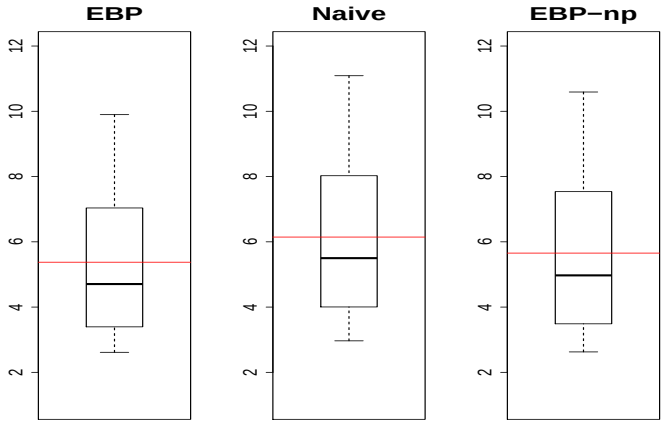
$m = 30, N_i = 100, n_i = 10$



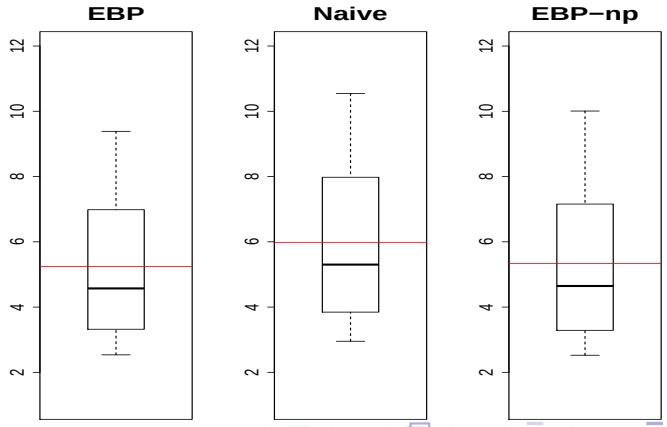
$m = 30, N_i = 1000, n_i = 10$



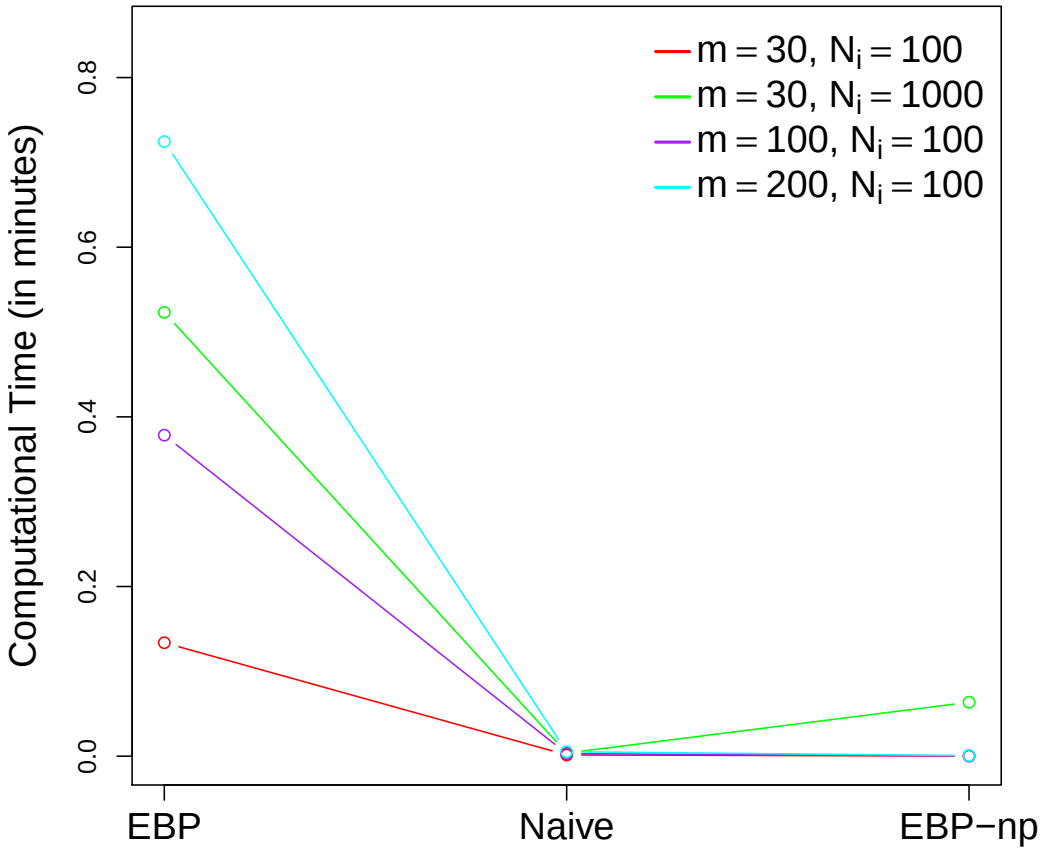
$m = 100, N_i = 100, n_i = 10$



$m = 200, N_i = 100, n_i = 10$



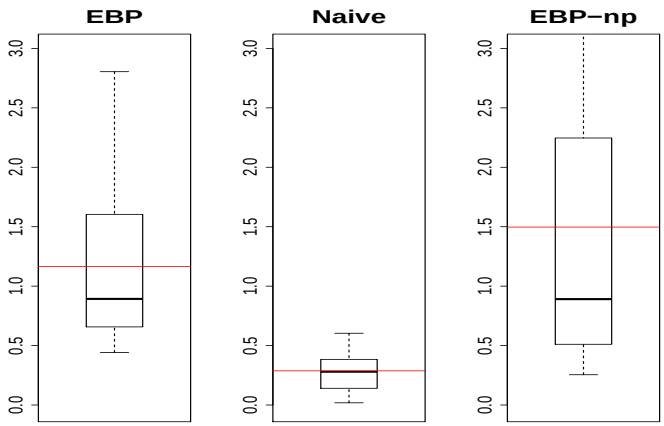
Temps de calcul pour obtenir $\hat{\theta}_i$ $(n_i = 10)$



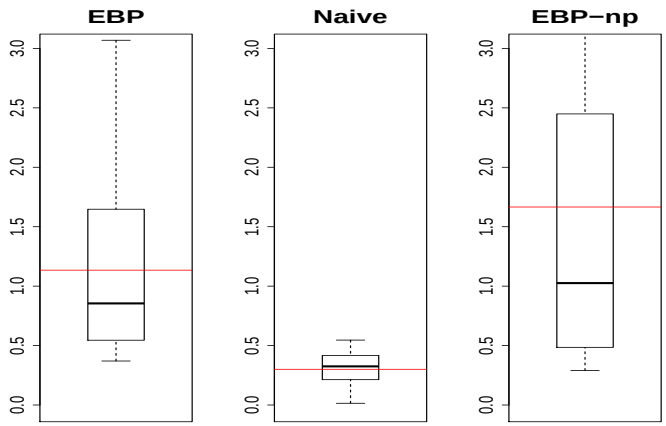
*Architecture Intel Core I5 2.40 GHz

Biais de $\widehat{RMSE}(\hat{\theta}_i)$

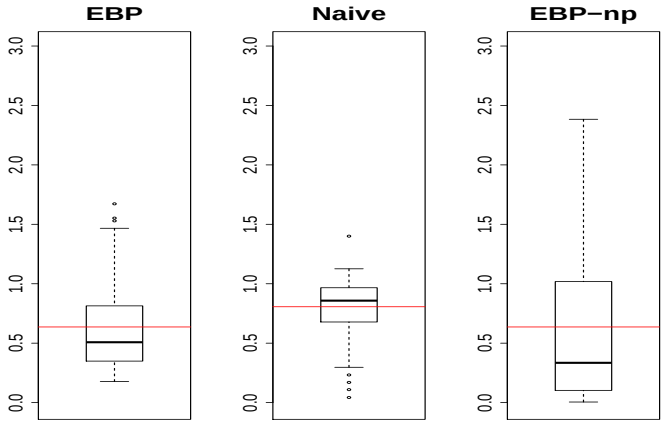
$m = 30, N_i = 100, n_i = 10$



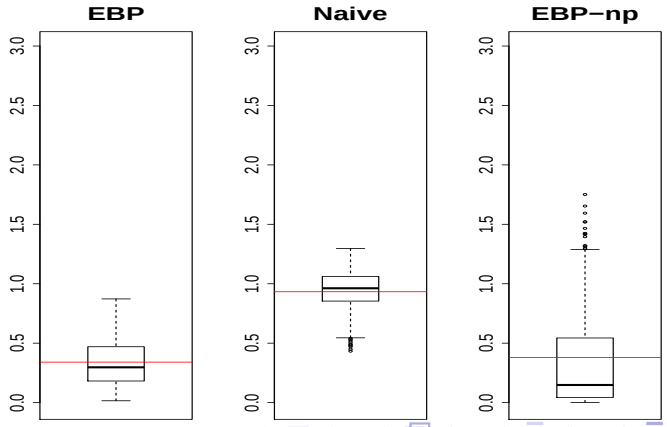
$m = 30, N_i = 1000, n_i = 10$



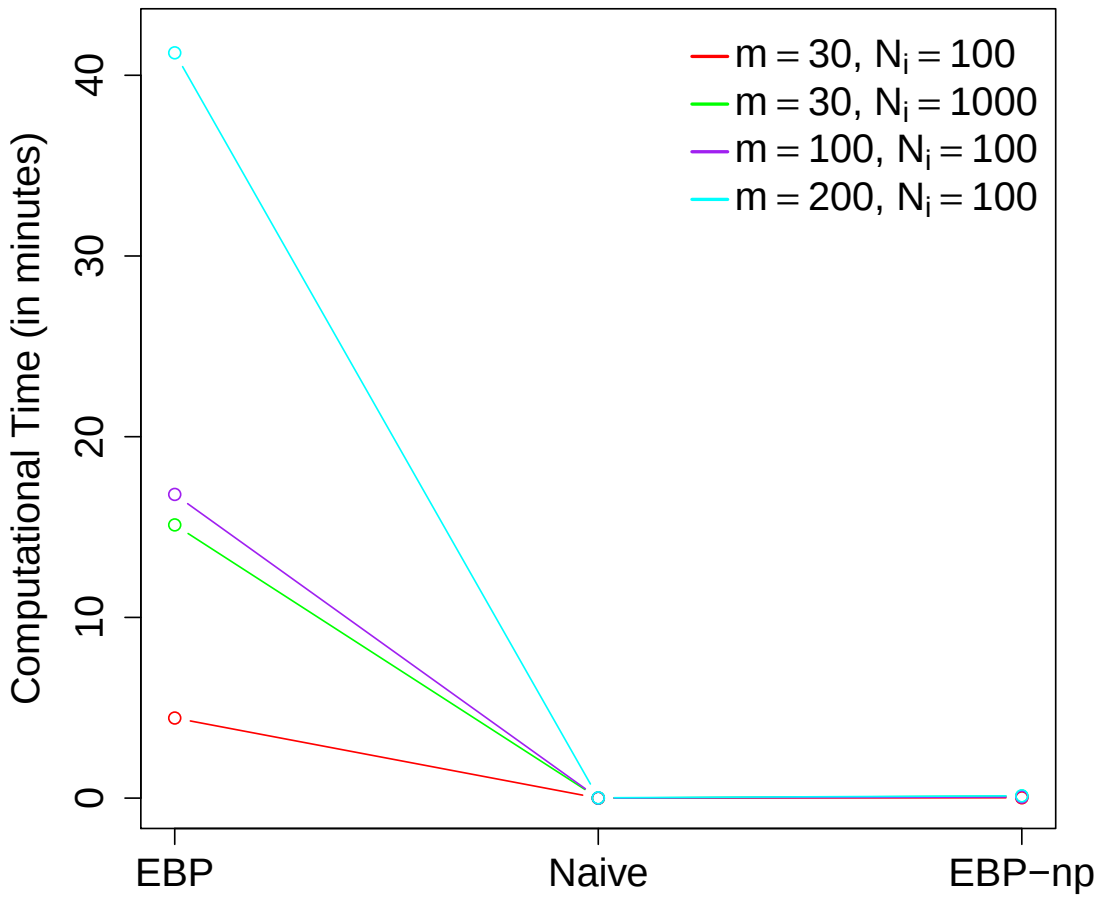
$m = 100, N_i = 100, n_i = 10$



$m = 200, N_i = 100, n_i = 10$



Temps de calcul pour obtenir $\widehat{RMSE}(\hat{\theta}_i)$



Résumé

- Pour un petit m, l'estimateur du MSE obtenu avec l'approche NPML est moins précis que ses adversaires
- Cependant, quand m augmente, notre proposition est meilleure que l'estimateur de González-Manteiga et devient plus proche de celui de Jiang et Lahiri
- Si on considère également le volume de calcul, l'approche NPML semble à privilégier

Lumières ... et ombres

- Les Modèles de Markov Latents apportent un cadre de modélisation très flexible
  - apportent un aperçu sur l'évolution des estimateurs dans le temps
  - traitent des réponses multivariées
  - et/ou des erreurs de mesure
  - fournissent une classification des petits domaines
  - compromis entre des coefficients de régression spécifiques au domaine et un coefficient de régression commun
- Choix de modèle
- Echange de Labels

Lumières ... et ombres

- Les Modèles de Markov Latents apportent un cadre de modélisation très flexible
  - apportent un aperçu sur l'évolution des estimateurs dans le temps
  - traitent des réponses multivariées
  - et/ou des erreurs de mesure
  - fournissent une classification des petits domaines
  - compromis entre des coefficients de régression spécifiques au domaine et un coefficient de régression commun
- Choix de modèle
- Echange de Labels

Développements futurs

- Cadre multivarié (emploi-chômeur-inactif)
- Inclut la corrélation spatiale – dans le modèle latent
- A voir - réponses Poisson/Binomiale/Multinomiale (MH)

Lumières ... et ombres

- Le Maximum de Vraisemblance Non Paramétrique apporte un outil très utile pour
  - développer un EBP pour des réponses binaires (famille exponentielle en général)
  - estimer la distribution du paramètre aléatoire à partir des données (prise en compte de la non-normalité)
  - obtenir un estimateur analytique du MSE calculable sur le plan informatique
  - classer les petits domaines
- Choix de modèle

Lumières ... et ombres

- Le Maximum de Vraisemblance Non Paramétrique apporte un outil très utile pour
  - développer un EBP pour des réponses binaires (famille exponentielle en général)
  - estimer la distribution du paramètre aléatoire à partir des données (prise en compte de la non-normalité)
  - obtenir un estimateur analytique du MSE calculable sur le plan informatique
  - classer les petits domaines
- Choix de modèle

Développements futurs

- Théorie
  - obtenir un terme de correction du biais pour l'estimateur du MSE afin d'améliorer la qualité des résultats
- Simulations
  - évaluer notre proposition pour différentes spécifications de $f_\alpha(\alpha_i)$ et comparer les résultats avec d'autres méthodes
  - comparer les résultats pour le MSE analytique proposé avec ceux par une approche Bootstrap
- Application
  - appliquer l'approche basée sur le NPML à l'enquête LFS qui consiste en 611 petits domaines avec $\simeq 100,000$ observations

Références I


Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, 6(3):251–262.


Bartolucci, F., Farcomeni, A., et Pennoni, F. (2013). Latent Markov models forlongitudinal data, Chapman & Hall/CRC Taylor and Francis Group.


Chib, S. (1995). Marginal likelihood from the Gibbs output. JASA 90, 1313-1321.


Datta, G.S., Lahiri, P., Maiti, T., et Lu, K.L. (1999), Hierarchical Bayes Estimation of Unemployment Rates for the U.S. States, JASA 94, 1074-1082.


Datta, G.S., Lahiri, P., et Maiti, T. (2002). Empirical Bayes Estimation of Median Income of Four-Person Families by State Using Time Series and Cross-Sectional Data, JSPI 102, 83-97


Dempster, A. P., Laird, N. M., et Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. JRSS-B 39, 1–38.

Références II


Fabrizi, E., Montanari, G.E., Ranalli, M.G. (2016). A hierarchical latent class model for predicting disability small area counts from survey data, JRSS-A 179-1, 103-131.


Ghosh, M., Nangia, N., et Kim, D. (1996), Estimation of Median Income of Four-Person Families: A Bayesian Time Series Approach, JASA 91, 1423-1431.


González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., et Santamaría, L. (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. CSDA, 51(5):2720–2733.


Jiang, J. et Lahiri, P. (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics*, 53(2):217–243.


Jiang, J. (1998). Consistent estimators in generalized linear mixed models. JASA, 93(442):720–729.

Références III


Oakes, D. (1999). Direct calculation of the information matrix via the EM. JRSS-B, 61(2):479–482.


Pfeffermann, D., et Burck, L. (1990). Robust Small Area Estimation Combining Time Series and Cross-Sectional Data, Survey Methodology 16, 217-237.


Rao, J.N.K. et Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. Canadian Journal of Statistics 22, 511-528.


Tanner, M. A., et Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. JASA 82, 528-540.


Van Dyk, D. A., Meng, X. L. (2001). The art of data augmentation. JCGS 10-1.


You, Y., Rao, J.N.K., Gambino, J. (2003). Model-based unemployment rate estimation for the Canadian Labour Force Survey: a hierarchical Bayes approach. Survey Methodology, 29-1, 25–32.