Estimation de courbes moyennes de consommation électrique par sondage pour des petits domaines

Anne De Moliner [1,2] Hervé Cardot [2] Camelia Goga [2]

[1] EDF R&D

[2] Université de Bourgogne

Colloque Francophone sur les sondages, octobre 2016

Sommaire

Sondages pour des courbes, estimations sur petits domaines

- Beaucoup d'études d'EDF sont basées sur l'analyse de courbes de charges moyennes de groupes de clients
- A partir d'échantillons de quelques milliers de courbes de charges, collectées selon un plan de sondage, mesurées toutes les demi-heures pendant une longue période.
- Besoin grandissant d'estimation pour des zones géographiques de plus en plus fines, et plus seulement au niveau national

Petits domaines en sondages

- Lorsqu'il y a trop peu d'individus du domaine dans l'échantillon les estimations directes (= basées uniquement sur les individus appartenant au domaine) deviennent instables.
- Petites sous populations = Petits domaines : Problème usuel en sondages, voir Rao et Molina (2015)
- Besoin d' "emprunter" de l'information au reste de l'échantillon, par une modélisation explicite ou implicite du lien entre variables explicatives et variable d'intérêt

- Objectif : étendre les méthodes d'estimation sur petits domaines en sondages au contexte des données fonctionnelles, proposer de nouvelles méthodes

Petits domaines en sondages

- Lorsqu'il y a trop peu d'individus du domaine dans l'échantillon les estimations directes (= basées uniquement sur les individus appartenant au domaine) deviennent instables.
- Petites sous populations = Petits domaines : Problème usuel en sondages, voir Rao et Molina (2015)
- Besoin d' "emprunter" de l'information au reste de l'échantillon, par une modélisation explicite ou implicite du lien entre variables explicatives et variable d'intérêt
- Objectif : étendre les méthodes d'estimation sur petits domaines en sondages au contexte des données fonctionnelles, proposer de nouvelles méthodes

Notations et cadre de travail (1/2)

Paramètres d intérêt

Population U composée de N individus. A chaque unité i on associe une courbe (de charge) $Y_i(t)$ sur $[0, T]$ mesurée pour p instants équidistants.
U peut être décomposée en D domaines disjoints U_d de taille N_d. On cherche à estimer la courbe moyenne μ_d de chaque domaine :

$$\mu_d(t) = \frac{1}{N_d} \sum_{i \in U_d} y_i(t), t \in [0, T].$$

Echantillon

Echantillon s de taille n tiré aléatoirement selon un plan de sondage conditionnellement non informatif. On notera s_d (de taille n_d) l'intersection du domaine U_d et de l'échantillon s.

Notations et cadre de travail (1/2)

Paramètres d intérêt

Population U composée de N individus. A chaque unité i on associe une courbe (de charge) $Y_i(t)$ sur $[0, T]$ mesurée pour p instants équidistants.
U peut être décomposée en D domaines disjoints U_d de taille N_d. On cherche à estimer la courbe moyenne μ_d de chaque domaine :

$$\mu_d(t) = \frac{1}{N_d} \sum_{i \in U_d} y_i(t), t \in [0, T].$$

Echantillon

Echantillon s de taille n tiré aléatoirement selon un plan de sondage conditionnellement non informatif. On notera s_d (de taille n_d) l'intersection du domaine U_d et de l'échantillon s.

Notations et cadre de travail (2/2)

Information auxiliaire

Au niveau de l'unité : X_i (de moyenne $\bar{X}_d$ sur le domaine d)
Variables réelles et non fonctionnelles

Modèle de superpopulation (forme générale)

$$y_i(t) = f_{d(i)}(X_i, t) + \epsilon_i(t), \quad i \in U_d, \quad t \in [0, T] \tag{1}$$

avec ϵ_i un bruit blanc de moyenne nulle.

Benchmark : moyenne simple

$$y_i(t) = \mu_d(t) + \epsilon_i(t), \quad i \in U_d$$

estimé par la moyenne du domaine

$$\hat{\mu}_d^0(t) = \frac{\sum_{i \in s_d} y_i(t)}{n_d}$$

Notations et cadre de travail (2/2)

Information auxiliaire

Au niveau de l'unité : X_i (de moyenne $\bar{X}_d$ sur le domaine d)
Variables réelles et non fonctionnelles

Modèle de superpopulation (forme générale)

$$y_i(t) = f_{d(i)}(X_i, t) + \epsilon_i(t), \quad i \in U_d, \quad t \in [0, T] \tag{1}$$

avec ϵ_i un bruit blanc de moyenne nulle.

Benchmark : moyenne simple

$$y_i(t) = \mu_d(t) + \epsilon_i(t), \quad i \in U_d$$

estimé par la moyenne du domaine

$$\hat{\mu}_d^0(t) = \frac{\sum_{i \in s_d} y_i(t)}{n_d}$$

Notations et cadre de travail (2/2)

Information auxiliaire

Au niveau de l'unité : X_i (de moyenne $\bar{X}_d$ sur le domaine d)
Variables réelles et non fonctionnelles

Modèle de superpopulation (forme générale)

$$y_i(t) = f_{d(i)}(X_i, t) + \epsilon_i(t), \quad i \in U_d, \quad t \in [0, T] \tag{1}$$

avec ϵ_i un bruit blanc de moyenne nulle.

Benchmark : moyenne simple

$$y_i(t) = \mu_d(t) + \epsilon_i(t), \quad i \in U_d$$

estimé par la moyenne du domaine

$$\hat{\mu}_d^0(t) = \frac{\sum_{i \in s_d} y_i(t)}{n_d}$$

Outline

Modèles linéaires mixtes

Modèles linéaires mixtes niveau unité (Battese et al (1988)) utilisé fréquemment pour des petits domaines :

$$y_i = \mu + \beta X_i + u_{d(i)} + \epsilon_i \tag{2}$$

avec βX_i les effets fixes, $u_d \sim N(0, \sigma_u)$les effets aléatoires et $\epsilon_i \sim N(0, \sigma_\epsilon)$ les résidus

Comment estimer le modèle ?

- Estimation instant par instant ?
- **Pour exploiter les corrélations temporelles : Analyse en Composantes Principales Fonctionnelle**

Modèles linéaires mixtes

Utilisés fréquemment pour des petits domaines ... mais pas sur des données fonctionnelles !

$$y_i(t) = \mu(t) + \beta(t)X_i + u_{d(i)}(t) + \epsilon_i(t) \tag{3}$$

les effets fixes $\beta(t)X_i$, les effets aléatoires $u_d(t)$ et les résidus $\epsilon_i(t)$ (bruits blancs de moyenne nulle) sont fonctionnels.

Comment estimer le modèle ?

- Estimation instant par instant ?
- **Pour exploiter les corrélations temporelles : Analyse en Composantes Principales Fonctionnelle**

Modèles linéaires mixtes

Utilisés fréquemment pour des petits domaines ... mais pas sur des données fonctionnelles !

$$y_i(t) = \mu(t) + \beta(t)X_i + u_{d(i)}(t) + \epsilon_i(t) \tag{3}$$

les effets fixes $\beta(t)X_i$, les effets aléatoires $u_d(t)$ et les résidus $\epsilon_i(t)$ (bruits blancs de moyenne nulle) sont fonctionnels.

Comment estimer le modèle ?

- Estimation instant par instant ?
- **Pour exploiter les corrélations temporelles : Analyse en Composantes Principales Fonctionnelle**

Modèles linéaires mixtes

Utilisés fréquemment pour des petits domaines ... mais pas sur des données fonctionnelles !

$$y_i(t) = \mu(t) + \beta(t)X_i + u_{d(i)}(t) + \epsilon_i(t) \tag{3}$$

les effets fixes $\beta(t)X_i$, les effets aléatoires $u_d(t)$ et les résidus $\epsilon_i(t)$ (bruits blancs de moyenne nulle) sont fonctionnels.

Comment estimer le modèle ?

- Estimation instant par instant ?
- **Pour exploiter les corrélations temporelles : Analyse en Composantes Principales Fonctionnelle**

Des courbes aux scores

L'expansion de Karhunen-Loeve sur les K premières composantes s'écrit :

$$y_i(t) = \mu(t) + \sum_{k=1}^{K} f_{k,i}\zeta_k(t) + \nu_i(t),$$

avec ζ_k les composantes principales, $f_{k,i}$ le score de y_i sur la composante ζ_k et $\nu_i(t)$ le résidu.

Somme par domaine :

$$\mu_d(t) = \frac{1}{N_d}\sum_{i\in U_d} y_i(t) = \mu(t) + \sum_{k=1}^{K} \underbrace{\frac{1}{N_d}\sum_{i\in U_d} f_{k,i}}_{\bar{f}_{k,d}}\,\zeta_k(t) + \underbrace{\frac{1}{N_d}\sum_{i\in U_d}\nu_i(t)}_{residu} \qquad (4)$$

$\bar{f}_{k,d}$ sera estimé par Unit Level Model indépendamment sur chaque composante k. Le résidu sera négligé.

Modèles linéaires mixtes sur les scores des composantes

En suivant l'approche de Battese et al (1988), $\overline{f}_{k,d} = \frac{\sum_{i \in U_d} f_{k,i}}{N_d}$ peut être estimé par :

$$\widehat{\overline{f}}_{k,d} = \gamma_k \, \hat{\theta} \; + (1 - \gamma_k) \, \hat{\psi} \tag{5}$$

avec $\hat{\theta}$ l'estimateur direct, $\hat{\psi}$ l'estimateur model-based et $\gamma_k = \frac{\sigma_u}{\sigma_\epsilon^2 + \sigma_u^2}$: Le poids donné aux estimations directes augmente à mesure que la variance des effets aléatoires augmente.

Des scores moyens estimés aux courbes moyennes estimées

Les courbes moyennes estimées des domaines peuvent être déduites des scores moyens estimés par :

$$\hat{\mu}_d(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \widehat{\overline{f}}_{k,d} \hat{\zeta}_k(t), \tag{6}$$

avec $\hat{\mu}(t)$ la courbe moyenne de l'échantillon et $\hat{\zeta}_k(t)$ les composantes principales estimées.

Outline

Approche prédictive pour des domaines

Approche prédictive (Valliant et al 2000) (pour des domaines) :

$$\hat{\mu}_d(t) = \frac{1}{N_d}\left(\sum_{i\in s_d} y_i(t) + \sum_{i\in U_d - s_d} \hat{y}_i(t)\right). \tag{7}$$

Prédiction de $\widehat{y}_i(t)$ par des arbres de régression ou des forêts aléatoires (modèles non paramétriques) adaptés aux données fonctionnelles.

$$y_i(t) = f(X_i) + \epsilon_i(t)$$

Approche prédictive pour des domaines

Approche prédictive (Valliant et al 2000) (pour des domaines) :

$$\hat{\mu}_d(t) = \frac{1}{N_d}\left(\sum_{i\in s_d} y_i(t) + \sum_{i\in U_d - s_d} \hat{y}_i(t)\right). \tag{8}$$

Prédiction de $\widehat{y}_i(t)$ par des arbres de régression ou des forêts aléatoires (modèles non paramétriques) adaptés aux données fonctionnelles.

$$y_i(t) = f(X_i) + \epsilon_i(t)$$

Arbres de régression pour des courbes (1/3)

Arbres de régression binaires (CART, Breiman (1984))

Partition itérative de l'espace en divisant en deux récursivement le jeu de données selon un seuil sur une variable explicative de manière à miminiser un critère d'hétérogénéité dans les feuilles.
Pour une variable cible réelle Y, le critère d'hétérogénéité est souvent la variance empirique de Y.

Arbres de régression pour des courbes (2/3)

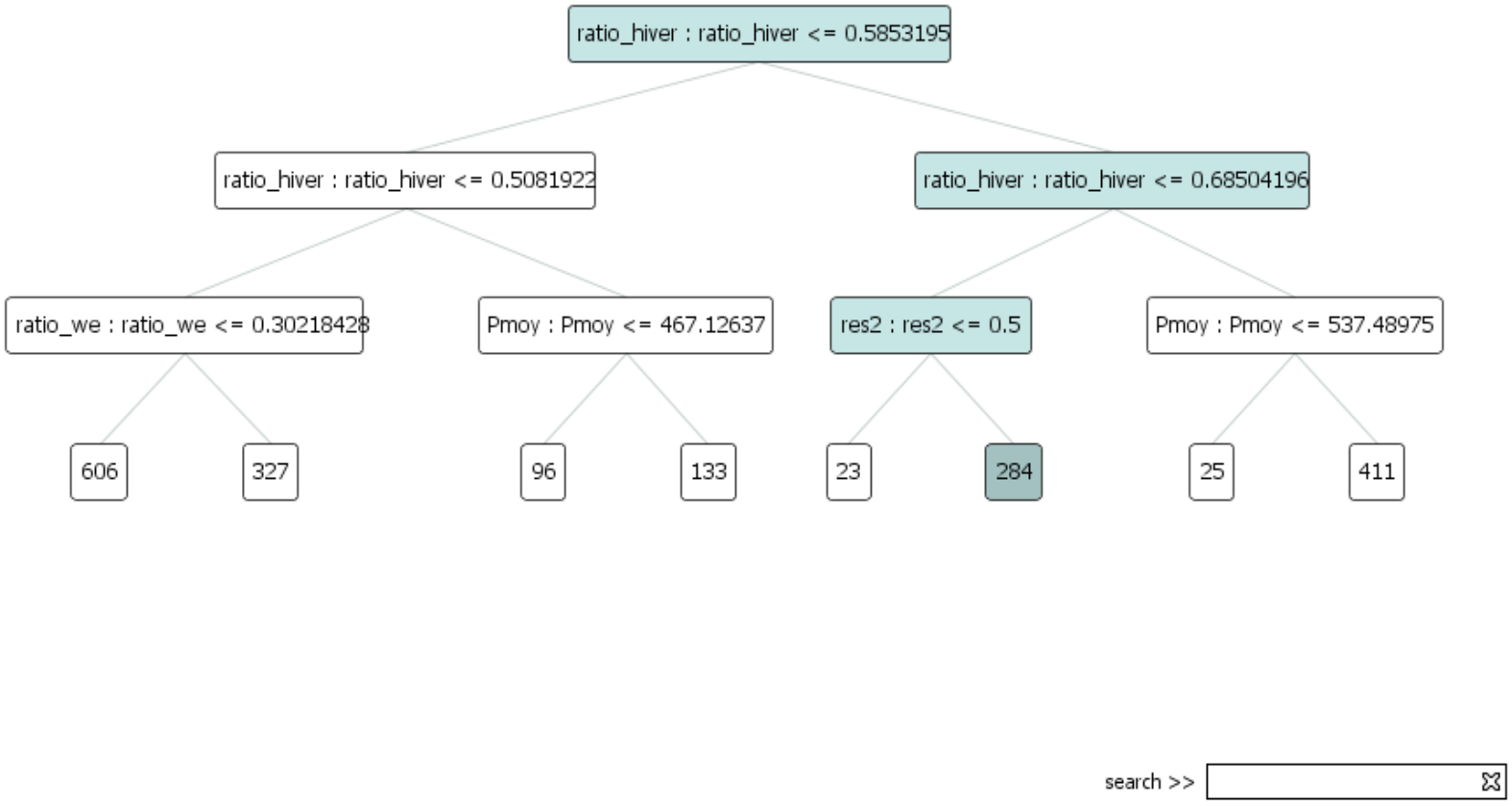


FIGURE – L'outil Courbotree

Arbres de régression pour des courbes (3/3)

Arbres de régression pour des courbes : l'approche Courbotree (Stephan et al 2009)

Critère d'hétérogénéité : distance euclidienne moyenne à la courbe moyenne de la feuille.

La courbe estimée pour l'unité i est la courbe moyenne de la feuille c_l à laquelle i est affectée en fonction de ses variables explicatives

Variantes possibles

- Estimation séparée du niveau (par random forest) et de la forme/ normalisation par un niveau issu des variables auxiliaires
- Modèle linéaire fonctionnel sur les variables au niveau domaine dans chaque feuille, ou médiane de la feuille au lieu de la moyenne.

Forêts aléatoires pour des courbes

Les arbres de régression ont tendance à être instables et très dépendants de l'échantillon :

- Utiliser cette faiblesse pour la transformer en force : estimer un grand nombre d'arbres aléatoires (profonds) puis les agréger en forêt en prenant la moyenne des prédictions de chaque arbre : Random Forest (Breiman (2001))
- Pour chaque arbre de la forêt, rééchantillonner les unités (n unités tirées avec remise) et à chaque split de l'arbre, sélectionner aléatoirement un sous ensemble des variables explicatives.
- Adaptation à des courbes : reprise de l'algorithme de Breiman (2001), en utilisant le critère Courbotree comme critère d hétérogénéité.

Outline

Protocole de test

Le jeu de données

- 1904 courbes de charge de clients résidentiels, d'octobre 2011 à mars 2012 (177 points), répartis en 8 domaines (zones géographiques).
- Information auxiliaire :
  - Niveau individu : puissance souscrite, tarif, consommation de l'année précédente
  - Niveau domaine : taux de chauffage et d'eau chaude sanitaire électrique et surface moyenne des logements.

Protocole de test

- Tirer un grand nombre d'échantillons ($B = 2000$) de taille $n = 200$ (sondage aléatoire simple) et, à partir de chaque échantillon, estimer la moyenne de chaque domaine par les différents estimateurs. .
- Le 8^{th} sera toujours vide.
- Comparer ces estimations avec les moyennes des domaines pour évaluer leurs performances.

Indicateurs de qualité

Pour un domaine d et un instant t, on définit

Indicateur de Biais Relatif

$$RB(\widehat{Y}_d(t)) = 100\frac{|E_{MC}[\hat{Y}_d(t)] - Y_d(t)|}{Y_d(t)}$$

avec $\widehat{Y}_d{}^b(t)$ l'estimateur de la courbe moyenne pour la simulation b et $E_{MC}[\widehat{Y}_d(t)] = \frac{1}{B}\sum_{b=1}^{B}\widehat{Y}_d{}^b(t)$ son espérance Monte Carlo.

Indicateur d'Erreur Quadratique Moyenne

$$MSE_{MC}(\hat{Y}_d(t)) = \frac{1}{B}\sum_{b=1}^{B}(\hat{Y}_d{}^b(t) - Y_d(t))^2$$

Efficacité Relative

$$RE(\hat{Y}_d(t)) = 100\frac{\overline{MSE}_{MC}(t)(\hat{Y}_d)(t)}{\overline{MSE}_{MC}(\hat{Y}_d(t)^{REF})}$$

Indicateurs de qualité

Pour un domaine d et un instant t, on définit

Indicateur de Biais Relatif

$$RB(\widehat{Y}_d(t)) = 100\frac{|E_{MC}[\hat{Y}_d(t)] - Y_d(t)|}{Y_d(t)}$$

avec $\widehat{Y}_d{}^b(t)$ l'estimateur de la courbe moyenne pour la simulation b et $E_{MC}[\widehat{Y}_d(t)] = \frac{1}{B}\sum_{b=1}^{B}\widehat{Y}_d{}^b(t)$ son espérance Monte Carlo.

Indicateur d'Erreur Quadratique Moyenne

$$MSE_{MC}(\hat{Y}_d(t)) = \frac{1}{B}\sum_{b=1}^{B}(\hat{Y}_d{}^b(t) - Y_d(t))^2$$

Efficacité Relative

$$RE(\hat{Y}_d(t)) = 100\frac{\overline{MSE}_{MC}(t)(\hat{Y}_d)(t)}{\overline{MSE}_{MC}(\hat{Y}_d(t)^{REF})}$$

Indicateurs de qualité

Pour un domaine d et un instant t, on définit

Indicateur de Biais Relatif

$$RB(\widehat{Y}_d(t)) = 100\frac{|E_{MC}[\hat{Y}_d(t)] - Y_d(t)|}{Y_d(t)}$$

avec $\widehat{Y}_d^{\,b}(t)$ l'estimateur de la courbe moyenne pour la simulation b et $E_{MC}[\widehat{Y}_d(t)] = \frac{1}{B}\sum_{b=1}^{B}\widehat{Y}_d^{\,b}(t)$ son espérance Monte Carlo.

Indicateur d'Erreur Quadratique Moyenne

$$MSE_{MC}(\hat{Y}_d(t)) = \frac{1}{B}\sum_{b=1}^{B}(\hat{Y}_d^{\,b}(t) - Y_d(t))^2$$

Efficacité Relative

$$RE(\hat{Y}_d(t)) = 100\frac{\overline{MSE}_{MC}(t)(\hat{Y}_d)(t)}{\overline{MSE}_{MC}(\hat{Y}_d(t)^{REF})}$$

Indicateurs de qualité

Pour un domaine d et un instant t, on définit

Indicateur de Biais Relatif

$$RB(\widehat{Y}_d(t)) = 100\frac{|E_{MC}[\hat{Y}_d(t)] - Y_d(t)|}{Y_d(t)}$$

avec $\widehat{Y}_d{}^b(t)$ l'estimateur de la courbe moyenne pour la simulation b et $E_{MC}[\widehat{Y}_d(t)] = \frac{1}{B}\sum_{b=1}^{B}\widehat{Y}_d{}^b(t)$ son espérance Monte Carlo.

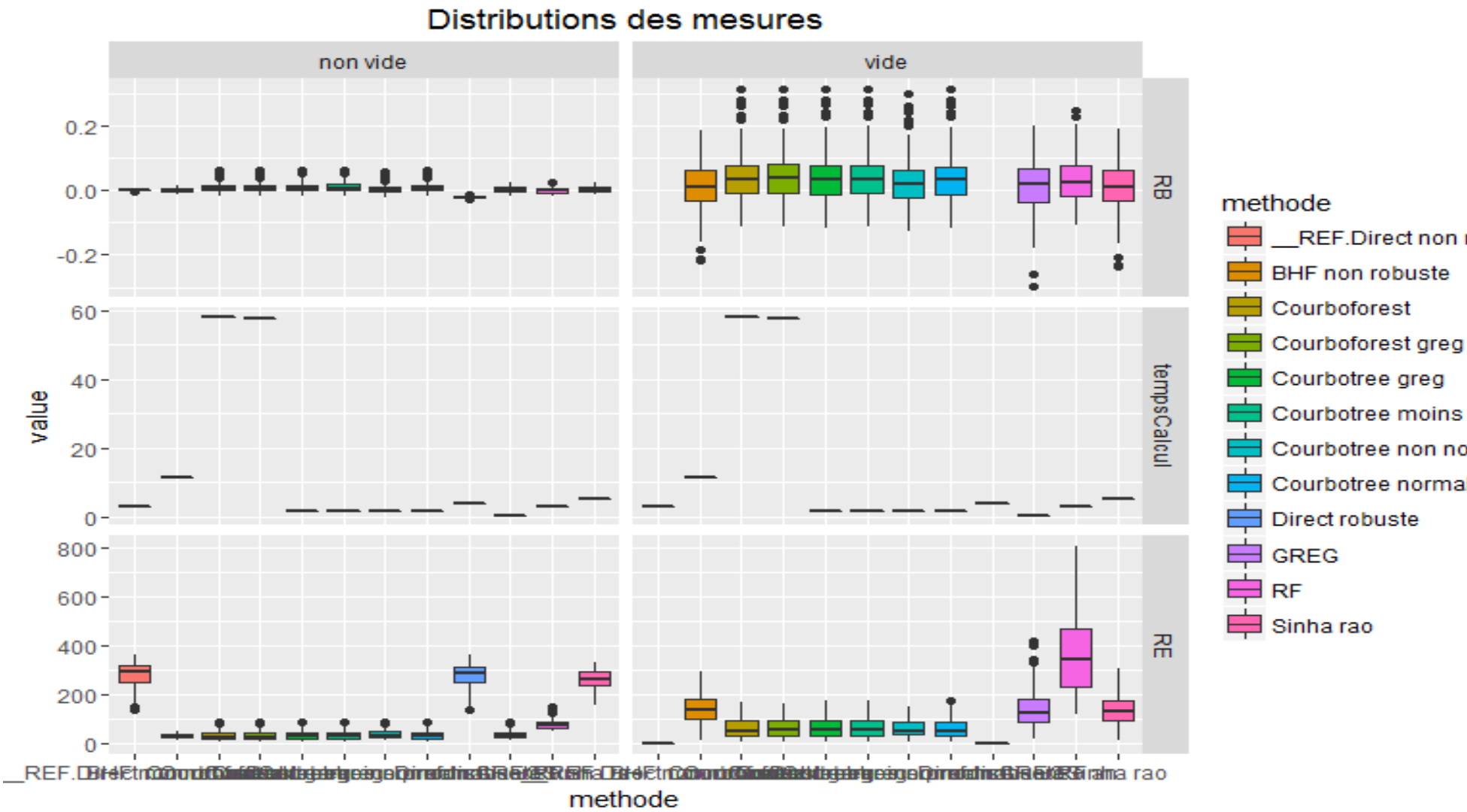
Indicateur d'Erreur Quadratique Moyenne

$$MSE_{MC}(\hat{Y}_d(t)) = \frac{1}{B}\sum_{b=1}^{B}(\hat{Y}_d{}^b(t) - Y_d(t))^2$$

Efficacité Relative

$$RE(\hat{Y}_d(t)) = 100\frac{\overline{MSE}_{MC}(t)(\hat{Y}_d)(t)}{\overline{MSE}_{MC}(\hat{Y}_d(t)^{REF})}$$

Outline


Distributions des mesures
non vide
vide
RB
tempsCalcul
RE
value
0.2
0.0
-0.2
60
40
20
0
800
600
400
200
0
methode
__REF.Direct non
BHF non robuste
Courboforest
Courboforest greg
Courbotree greg
Courbotree moins
Courbotree non no
Courbotree norma
Direct robuste
GREG
RF
Sinha rao

	Echantillonnés		Non échantillonnés		
Méthode	RB (%)	RE (%)	RB (%)	RE (%)	temps (s)
Benchmark	0.00	100	NA	NA	3.31
ACP + LMM	0.00	13.48	0.01	139.45	11.58
Courbotree	0.01	18.94	0.04	76.4	1.99
Courboforest	0.01	19.76	0.04	73.84	57.95

TABLE – Comparaison des modèles, LMM = modèles linéaires mixtes

- Forte amélioration par rapport au benchmark, erreurs acceptables sur les domaines vides. Biais faibles pour l'ensemble des méthodes.
- Meileure méthode ici : CourboTree/CourboForest sur les domaines vides, modèles linéaires mixtes sur Composantes Principales sur domaines non vides
- Courbotree beaucoup plus rapide que CourboForest pour des performances similaires

Conclusions et perspectives

- Adaptation des modèles linéaires mixtes, des arbres de régression et des forêts aléatoires dans le contexte des données fonctionnelles pour l'estimation sur petits domaines.
- Améliore fortement la précision par rapport à la moyenne simple (MSE divisé par 4 dans les tests)
- Meilleures méthodes : CourboTree : arbres de régression pour des courbes et modèles linéaires mixtes sur composantes principales.
- Travaux futurs : Robustesse aux individus atypiques

Bibliographie I


Breiman, L. and Friedman, J. and Stone, C. and Olshen, R. (1984)
Classification and regression trees
CRC press.


Rao, J.N. and Molina I. (2015)
Small area estimation,2nd edition.
Wiley, 2nd edition


Valliant, R. and Dorfman, A. and Royall, R. (2000)
Finite population sampling and inference : a prediction approach
Wiley


Battese, G.E. and Harter, R. and Fuller, W. (1988)
An error-components model for prediction of county crop areas using survey and satellite data
Journal of the American Statistical Association ,**83**,28–36.

Bibliographie II


Stephan V. and Cordogan F.(2009)
Courbotree : application des arbres de regression multivaries pour la classification de courbes
Revue MODULAD ,**33**,129–138