# Estimation de courbes moyennes de consommation électrique à partir d'échantillons pour des petits domaines

Anne De Moliner *Université De Bourgogne-France Comté / EDF R&D Paris-Saclay, anne.de-moliner@edf.fr*
Hervé Cardot *Institut de Mathématiques de Bourgogne, Université de Bourgogne-Franche-Comté, herve.cardot@u-bourgogne.fr*
Camelia Goga *Institut de Mathématiques de Bourgogne, Université de Bourgogne-Franche-Comté, camelia.goga@u-bourgogne.fr*

**Résumé**

De nombreuses études menées à EDF R&D se basent sur l'analyse de courbes de consommation électrique moyennes pour différentes sous-populations, en particulier de nature géographique. Ces courbes moyennes sont estimées à partir d'échantillons de milliers de courbes mesurées à un pas de temps fin pendant de longues périodes. Or, lorsque les domaines d'intérêt sont petits, on ne dispose souvent que de peu d'unités par domaine et les estimations sont imprécises. Cette problématique de l'estimation sur des petits domaines a été très étudiée en théorie des sondages [9]): les méthodes usuelles consistent à rendre plus précises les estimations sur chaque domaine à l'aide de modélisations implicites ou explicites de la relation entre variable d'intérêt et information auxiliaire. Notre objectif ici est d'adapter ces travaux existants, développés pour l'estimation de totaux de variables réelles, aux données fonctionnelles. Pour cela, nous proposons trois méthodes: la modélisation des scores d'Analyse en Composantes Principales par des modèles linéaires mixtes, un estimateur par régression linéaire fonctionnelle et un arbre de régression adapté à la prévision de courbes [12]). Ces méthodes ont été testées et comparées sur des données réelles de consommations électriques de ménages francais.

**Mots Clés** Données fonctionnelles, estimation sur petits domaines, modèles linéaires mixtes, arbres de régression,...

## 1 Introduction et contexte

De nombreuses études menées à EDF R&D sont basées sur l'analyse simultanée des courbes de consommations électriques moyennes de plusieurs groupes de clients, que l'on appellera par la suite des domaines, et qui partagent des caractéristiques communes. En particulier, il devient nécessaire de réaliser des estimations de courbes de consommations moyennes par zones géographiques (régions, départements, villes, voire quartiers) par exemple afin de proposer des services basés sur l'analyse des courbes de consommation aux collectivités territoriales.

Ces courbes de consommations électriques moyennes, aussi appelées courbes de charge, sont estimées à partir d'échantillons de plusieurs milliers de courbes mesurées toutes les demi-heures pendant de longues périodes (souvent des années). La question de l'estimation d'une courbe de charge totale ou moyenne pour différents plans de sondage et la construction d'intervalles de confiance ont été traitées dans les travaux récents de [5], [6] and [7].

On traitera ici plus précisément le cas où on s'intéresse à des sous-populations de taille réduite. Il s'agit donc d'un problème d'estimation sur petits domaines. Il s'agit d'une question fréquemment abordée en théorie des sondages que de nombreux auteurs traitent, hors du cadre des données fonctionnelles. Le livre très récent de [9] propose un état de l'art des méthodes existantes. Quand le domaine est petit, les estimateurs directs (c'est-à-dire construits uniquement à partir des individus de l'échantillon appartenant au domaine) ne sont pas très performants. Afin d'améliorer la qualité des estimations, des informations auxiliaires sont utilisées et on construit des estimateurs basés sur une modélisation implicite ou explicite du lien entre quantité d'intérêt et information auxiliaire.

A EDF, On dispose en général d'information auxiliaire pertinente, tant au niveau individuel qu'à la maille des domaines étudiés: ainsi, les informations de facturation sont disponibles pour chaque individu de la population, et on peut également exploiter des données en accès libre proposée par l'INSEE à la maille de petits agrégats géographiques (IRIS).

A notre connaissance, l'estimation sur petits domaines en sondages pour des données fonctionnelles n'a pas encore été traitée dans la littérature. Pour répondre à cette problématique, on propose trois méthodes: une estimation par modèles linéaires mixtes appliqués aux scores de l'analyse en composantes principales des courbes de charge, un estimateur par régression linéaire fonctionnelle implémenté en utilisant l'algorithme du calage selon la méthode proposée par [1] ainsi qu'une approche non paramétrique basée sur des arbres de régression adaptés à la prévision de courbes.

Dans le paragraphe 2, nous formaliserons le problème et introduirons quelques notations et hypothèses, puis au paragraphe 3 nous présenterons les différentes méthodes proposées, qui seront ensuite testées et comparées entre elles au paragraphe 4. Les conclusions et perspectives seront exposées au paragraphe 5.

# 2 Notations et hypothèses de travail

Dans cette section, nous introduisons quelques notations sur l'estimation en population finie, la définition des domaines, les données fonctionnelles et l'information auxiliaire.

Soit une population d'intérêt $U$ de taille $N$. A chaque unité $i$ de la population on associe une courbe (de charge) définie sur un intervalle de temps $[0, T]$: pour chaque unité $i$, on a une fonction continue $y_i(t), \quad t \in [0, T]$, où l'index $t$ représente le temps. En pratique, les courbes sont observées non pas de manière continue mais uniquement sur

un ensemble V d'instants de mesure $0 = t_1 < t_2 < ... < t_v = T$ que l'on suppose en outre être équidistants et identiques pour l'ensemble des individus. On suppose de plus qu'il n'y a pas de valeurs manquantes.

La population $U$ peut être décomposée en $D$ domaines disjoints $U_d$ de tailles $N_d$. Notre but est d'estimer la courbe moyenne $\mu_d$ de chaque domaine, i.e.

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

On définit les indicatrices $1_{d,i} = 1_{U_d(i)}$ qui valent 1 si l'unité $i$ appartient au domaine $d$ et zéro sinon. De plus, on note $d(i)$ le domaine auquel appartient l'unité $i$.

On dispose d'un vecteur d'information auxiliaire $X_i$ connu pour chaque individu $i$ de la population ainsi que d'information auxiliaire complémentaire, cette fois au niveau des domaines, $Z_d$. Pour appliquer directement les méthodes linéaires, on regroupe ces informations dans le nouveau vecteur $X_i^* = (X_i, Z_{d(i)})$. On connait également les moyennes $\bar{X}_d^*$ des variables $X_i^*$ pour chaque domaine. Pour simplifier, on considère ici que ces variables explicatives sont réelles et constantes au cours du temps.

On suppose que ces informations auxiliaires sont liées aux courbes de charge selon un modèle, dit de superpopulation, valide sur l'ensemble de la population, et qui s'écrit de manière générale

$$y_i(t) = f_{d(i)}(X_i^*, t) + \epsilon_i(t), \quad i \in U_d, \quad t \in [0, T], \tag{2}$$

avec $f_d$ une fonction de régression inconnue à estimer, qui peut varier d'un domaine à l'autre et $\epsilon_i$ un processus de bruit d'espérance nulle.

Parmi la population $U$ on sélectionne un échantillon de taille $n$. On note $s_d$ l'intersection du domaine $U_d$ et de l'échantillon $s$ et $n_d$ la taille de $s_d$. La taille $n_d$ est aléatoire et peut être égale à 0 pour un ou plusieurs domaines. Les tailles $n_d$ et $N_d$ sont supposées connues.

On suppose que cet échantillon est issu d'une collecte conditionnellement non informative, c'est-à dire-que, pour tout $i$, la distribution des courbes $y_i$ conditionnellement aux variables explicatives est la même dans l'échantillon et dans la population. Du fait des contraintes techniques fortes qui pèsent parfois sur la sélection des échantillons de courbes de charge, on préfère travailler dans un contexte d'inférence basée sur un modèle plutôt que d'inférence basée sur le plan. On suppose cependant que l'on dispose d'information auxiliaire suffisamment riche pour corriger d'éventuels biais de sélection ou défauts de couverture, ce qui rend raisonnable l'hypothèse de plan conditionnellement non informatif.

Un des modèles les plus simples que l'on puisse poser et qui nous sert de référence afin d'évaluer les performances de nos méthodes est

$$y_i(t) = \mu_{d(i)}(t) + \epsilon_i(t) \quad \forall i \in U_d,$$

avec $\epsilon_i$ un processus d'espérance nulle.

L'estimateur naif de la courbe d'un domaine est alors simplement la moyenne des courbes de ce domaine, i.e.

$$\widehat{\mu_d}^0 = \frac{\sum_{i \in s_d} y_i}{n_d}. \tag{3}$$

Cependant, cet estimateur ne peut évidemment pas être calculé pour les domaines non échantillonnés et il est extrêmement instable pour les domaines de petite taille.

# 3 Méthodes d'estimation

Dans cette section, nous présentons trois approches permettant de répondre à notre problématique d'estimation de courbes de charge moyennes par domaine: les modèles linéaires mixtes sur composantes principales, l'estimateur par la régression linéaire fonctionnelle et enfin les arbres de régression pour données fonctionnelles.

## 3.1 Modèles linéaires mixtes sur les scores de composantes principales

Dans ce paragraphe, on cherche à adapter les modèles linéaires mixtes au niveau unité, très usités dans le cadre des petits domaines (on pourra se référer à [2]) au contexte des données fonctionnelles.

Pour cela, notre méthodologie consiste à utiliser une ACP pour transformer notre problème d'estimation de courbes moyennes par domaines en un problème d'estimation de plusieurs moyennes de variables réelles non corrélées, que l'on sait résoudre par les méthodes usuelles. Plus précisément, on réalise une ACP fonctionnelle (voir [10]) puis on décompose notre courbe sur la base des $K$ premières composantes principales définies en suivant l'expansion de Karhunen-Loeve:

$$y_i(t) = \mu(t) + \sum_{k=1}^{K} f_{k,i}\zeta_k(t) + \nu_i(t), \quad i \in U, \tag{4}$$

avec $\zeta_k(t)$ la $k^{eme}$ composante principale, $k = 1, \ldots, K$, $\nu_i(t)$ le résidu, $f_{k,i}$ le score de l'unité $i$ pour la composante $k$.[1] En suivant (4), la moyenne $\mu_d$ du domaine $d$ peut être approximée par

$$\mu_d(t) \simeq \mu(t) + \sum_{k=1}^{K} \left( \frac{1}{N_d} \sum_{i \in U_d} f_{k,i} \right) \zeta_k(t). \tag{5}$$

[1] Ici, l'ACP n'est pas utilisée en temps que méthode de réduction de dimension mais dans le but de décomposer notre problème en plusieurs petits problèmes non corrélés d'estimation de totaux de variables réelles, que l'on sait résoudre. On garde donc un nombre $K$ de composantes principales aussi élevé que possible.

Les composantes principales $\zeta_k(t)$ sont inconnues et peuvent être estimées par $\hat{\zeta}_k(t)$ comme suggéré dans [4]. Donc, afin d'estimer $\mu_d$, il nous faut estimer $\mu$ ainsi que la moyenne des scores sur les composantes principaux pour le domaine $d$, i.e. $\overline{f}_{k,d} = \frac{1}{N_d} \sum_{i \in U_d} f_{k,i}$. Pour cela, on considère le modèle linéaire mixte au niveau unité sur $f_{k,i}$, $k = 1, \ldots, K$ comme dans [9]:

$$f_{k,i} = \beta_k' X_i^* + u_{k,d(i)} + \epsilon_{k,i}, \quad \forall i \in U_d, \tag{6}$$

avec $\beta_k' X_i^*$ l'effet fixe (fonctionnel) de l'information auxiliaire, $u_{k,d(i)}$ l'effet aléatoire (fonctionnel) du domaine $d(i)$ qui suit une loi normale de moyenne 0 et de variance $\sigma_{d,k}^2$ et $\epsilon_{k,i}$ le résidu, distribué selon une loi normale de moyenne 0 et de variance $\sigma_{\epsilon,k}^2$. Ce modèle est une version paramétrique de notre modèle général (2). En suivant la démarche proposée dans [9], Chapter 7, on estime $\beta_k$ par un EBLUP (Empirical Best Linear Unbiased Prediction) et on en déduit l'estimateur $\widehat{\overline{f}}_{k,d}$ de $\overline{f}_{k,d}$. Pour conclure, la moyenne $\mu_d$ est estimée par:

$$\hat{\mu}_d(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \widehat{\overline{f}}_{k,d} \hat{\zeta}_k(t), \tag{7}$$

avec $\hat{\mu}(t)$ la moyenne de l'échantillon et $\hat{\zeta}_k(t)$ les composantes principales estimées, pour $k = 1, \ldots, K$.

## 3.2 L'estimateur par régression linéaire fonctionnelle

Dans cette section, on suppose que l'on a le modèle de superpopulation suivant, valable pour l'ensemble de la population, i.e.

$$y_i(t) = \mu(t) + \beta(t) X_i^* + \epsilon_i(t), \forall i \in U,$$

avec $\epsilon_i$ un processus d'espérance nulle.

Il s'agit du modèle précédent auquel on a enlevé les effets aléatoires: on fait donc l'hypothèse que, conditionnellement aux variables explicatives, la distribution des courbes $y(t)$ est la même sur l'ensemble des domaines et qu'il n'y a pas de spécificités des domaines non prises en compte dans l'information auxiliaire. On est alors dans le contexte usuel de la régression fonctionnelle (régression d'une variable fonctionnelle sur des variables réelles). Ce problème se résout aisément en projetant les courbes sur une base adaptée (par exemple celle de l'ACP comme précédemment, mais on peut aussi utiliser des B-splines ou des ondelettes ou simplement discrétiser la courbe selon les instants de mesure). Si on discrétise selon les instants, on a l'estimateur de la courbe de charge moyenne par domaine:

$$\hat{y_d}(t) = \hat{\mu}(t) + \hat{\beta}(t) \bar{X}_d^*,$$

avec $\hat{\mu}(t)$ et $\hat{\beta}(t)$ les paramètres de la régression linéaire estimés par les moindres carrés ordinaires sur l'ensemble de l'échantillon.

L'estimation de ce modèle fonctionnel peut être lourde en temps de calcul si l'on travaille sur de grosses bases de données (beaucoup de domaines et/ou beaucoup d'instants de mesure), c'est pourquoi on propose de suivre l'approche proposée par [1] pour implémenter notre estimateur. En effet, sous l'approche assistée par un modèle, l'auteur propose d'estimer rapidement des totaux de petits domaines à l'aide d'un estimateur par la régression généralisée (voir [11]) en remarquant que, comme prouvé dans [8], cet estimateur est égal à l'estimateur par calage. On peut donc estimer un grand nombre de totaux en estimant un unique vecteur de poids de calage pour chaque domaine. De plus, notre estimateur par la régression linéaire (pour chaque composante principale, ou chaque vecteur de base), est égal à l'estimateur par la Régression Généralisée pour le Sondage Aléatoire Simple. En suivant cette idée, on peut donc implémenter notre estimateur en calculant un vecteur de poids unique pour chaque domaine et en projetant nos courbes sur la base de notre choix.

## 3.3 Arbres de régression pour courbes (Courbotree)

Dans ce paragraphe, nous proposons d'estimer la courbe de chacune des unités non échantillonnées puis d'en déduire la courbe moyenne de chaque domaine en sommant les prédictions de la manière suivante (voir [13]):

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

Afin d'obtenir la prédiction $\hat{y}_i$ pour chaque unité nous allons chercher à estimer un modèle de la forme:

$$y_i(t) = f(X_i^*, t) + \epsilon_i(t).$$

Il s'agit du modèle (2) dans lequel la fonction $f$ ne dépend plus du domaine. Pour estimer cette fonction $f$, nous allons utiliser une approche par arbre de régression. Cette méthode, suggérée par [3] est une technique d'estimation non paramétrique consistant à séparer en deux, itérativement, le jeu de données suivant l'une des variables explicatives, de façon à maximiser un critère d'homogénéité (ou, de manière équivalente, minimiser un critère d'inertie) sur chacune des partitions ainsi créées.

Ici la variable à prédire est une courbe et non pas une variable réelle comme usuellement. On propose donc d'adapter le critère d'inertie, en suivant l'approche dite du courbotree [12] fréquemment utilisée à EDF. Le critère d'inertie à minimiser au sein d'un noeud contenant les individus $i = 1, ..., I$ est celui basé sur la distance Euclidienne qui est une approximation de la norme $L^2$ pour des instants de mesure équidistants

$$W = \sum_{i=1}^{I}\sum_{t=1}^{T}(y_i(t) - g_I(t))^2,$$

avec $g_I(t)$ la courbe moyenne empirique dans le noeud.

L'estimateur de la courbe d'une unité non échantillonnée est alors la courbe moyenne estimée de la feuille à laquelle il est affecté selon les règles de décision définies par l'arbre: soit une unité $i$, affectée à la feuille $l = l(X_i^*)$, on a :

$$\hat{y}_i = \hat{\mu}_l = \frac{\sum_{k \in s_l} y_k}{n_l},$$

avec $s_l$ l'ensemble des $n_l$ unités de l'échantillon affectées à la feuille $l$.

On remarque que, comme l'arbre de régression est une méthode non linéaire, nous avons besoin de disposer des informations auxiliaires $X_i^*$ pour chaque individu de la population (alors que précédemment nous avions juste besoin de totaux sur la population et des $X_i^*$ sur l'échantillon).

En pratique, comme les courbes de charge ont des niveaux extrêmement hétérogènes, les algorithmes de classification fonctionnent mal, c'est pourquoi on recommande de travailler sur les courbes divisées par le niveau de consommation de l'année précédente (puis de remultiplier les estimations $\hat{y}_i$ par ce niveau au cours de la dernière étape de sommation des estimations). De plus, on recommande également de lisser les courbes de charge (par moyenne mobile d'ordre 5) et éventuellement de réaliser une agrégation temporelle, par exemple en recréant une "semaine type" avant l'étape de classification. Les courbes moyennes $\hat{\mu}_l$ sont cependant calculées sur les courbes non lissées et non agrégées.

# 4 Tests des méthodes sur des données réelles

Nous allons maintenant tester les méthodes que nous venons de présenter sur des données de consommations électriques de clients résidentiels francais afin de comparer leurs performances.

## 4.1 Description du jeu de données

La population de travail est constituée de 1904 courbes, disponibles au pas journalier, d'octobre 2011 à mars 2012 (177 points), sans valeurs manquantes. On considère huit domaines qui correspondent à un découpage de la France en zones géographiques.

Pour chacun des individus de notre population de test, on dispose de plusieurs variables auxiliaires au niveau individuel: puissance souscrite, option tarifaire, consommation de l'année précédente. Au niveau domaine, on dispose des informations suivantes: taux de chauffage électrique et surface moyenne des logements.

## 4.2 Le protocole de test

Afin d'évaluer la qualité de nos méthodes d'estimation, notre protocole de test consiste à tirer aléatoirement un grand nombre $B$ d'échantillons de courbes de charge parmi notre

population de départ et ensuite à estimer la courbe moyenne de chacun des huit domaines à partir de chaque échantillon tiré par les différentes méthodes proposées. On compare ensuite les courbes estimées aux courbes réelles afin de calculer des indicateurs de qualité.

On suppose que le huitième domaine est toujours vide. On effectue $B = 2000$ simulations de tirage d'échantillons. Pour chaque simulation, on effectue un sondage aléatoire simple, et on sélectionne 200 individus parmi ceux appartenant aux 7 domaines échantillonnés par sondage aléatoire simple.

### 4.2.1 Indicateurs de qualité

Lors de l'évaluation de la qualité des estimations, on sépare les résultats obtenus sur le domaine vide de ceux obtenus sur les domaines échantillonnés.

Soit $Y_d(t)$ la courbe moyenne du domaine $d$ à l'instant $t$ et $\hat{Y}_d(t)$ son estimateur par une méthode donnée. On note $E_{MC}[\hat{Y}_d(t)] = \frac{1}{B}\sum_{b=1}^{B} \hat{Y}_d^{\,b}(t)$ l'espérance Monte Carlo de l'estimateur $\hat{Y}_d(t)$ avec $\hat{Y}_d^{\,b}(t)$ l'estimateur de la courbe moyenne obtenu à partir de l'échantillon $b$, pour $b = 1, ..., B$.

Pour un instant $t$ et un domaine $d$ donnés, on construit d'abord un indicateur de biais

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

Pour un instant $t$ et un domaine $d$ donnés, on définit ensuite un indicateur d'erreur globale

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

Cet indicateur englobe à la fois le carré du biais et la variance. Plus il est faible mieux c'est.

Afin de faciliter l'intercomparaison des méthodes, on en déduit un autre indicateur plus facile à lire: l'efficacité relative, obtenue en divisant le MSE par domaine obtenu pour chaque méthode par le MSE par domaine obtenu pour une méthode de référence qui est l'estimateur naif par la moyenne du domaine (voir Eq. (3)).[2] On obtient alors l'estimateur $RE$ qui est une version "normalisée" du $MSE$, i.e.

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

avec $\overline{MSE}_{MC}(\hat{Y}_d)$ le MSE moyen de la méthode sur l'ensemble des instants, pour le domaine $d$. On prend la moyenne de ces indicateurs sur l'ensemble des instants.

[2]Pour le domaine vide, ce MSE n'existe pas, et on divise les MSE des différentes méthodes par le MSE moyen de la moyenne simple sur les domaines échantillonnés.

## 4.3 Les résultats

Comme suggéré, le modèle linéaire mixte a été appliqué sur les composantes principales (ACP + LMM) ou diretement sur les données discrétisées (Discrétisation + LMM). On obtient les résultats suivants:

| | Echantillonnés | | Non échantillonnés | | |
|---|---|---|---|---|---|
| Méthode | RB (%) | RE(%) | RB (%) | RE(%) | temps (s) |
| Estimateur naif | 0.28 | 100 | NA | NA | 0.07 |
| ACP + LMM | 0.52 | 20.98 | 5.1 | 200 | 0.53 |
| Discrétisation + LMM | 0.28 | 25.82 | 5.3 | 409 | 4.16 |
| Regression | 0.49 | 29 | 5.5 | 403 | 0.05 |
| Courbotree | 1.47 | 30.22 | 4.1 | 39 | 0.2 |

Ces tests nous ont permis de montrer que l'intégration de variables explicatives dans les estimations par le biais de modélisations permet d'améliorer grandement les estimations: ainsi sur les domaines échantillonnés, les meilleures performances sont obtenues par les modèles linéaires mixtes sur les composantes principales (RE de 21%), suivis par la régression linéaire fonctionnelle (RE de 29%) presque à égalité avec le CourboTree (RE de 30%). Les biais pour les différentes méthodes sont très modérés. L'utilisation d'une ACP, avant les modèles linéaires mixtes, en prenant en compte l'aspect temporel de la problématique, permet un gain de précision de quelques pourcents sur les domaines échantillonnés. La différence de quelques pourcents entre le modèle linéaire mixte sur la courbe discrétisée et la régression fonctionnelle correspond à l'apport des effets aléatoires dans le modèle. Elle se gagne au prix d'un temps de calcul un peu plus élevé. Les moindres performances du courboTree peuvent s'expliquer par le fait que cette méthode n'intègre pas l'apport des variables explicatives au niveau domaine contrairement aux autres techniques.

Pour les domaines non échantillonnés, la méthode la plus performante est de loin le CourboTree, suivie par les modèles linéaires mixtes sur ACP et enfin, la régression ou les modèles linéaires mixtes sur courbes discrétisées [3]. Les estimations ont alors un biais modéré pour l'ensemble des méthodes.

Sur notre jeu de données, les modèles linéaires mixtes s'estiment en quelques secondes (dix fois moins en cas de projection par ACP), les arbres sont vingt fois plus rapides, et la régression linéaire fonctionnelle cent fois plus rapide, ce qui est un atout en grande dimension.

[3] pour les domaines non échantillonnés, les estimations directes sont impossibles, et la base 100 du RE correspond au MSE moyen sur domaines échantillonnés.

# 5 Conclusions et perspectives

Nous avons proposé des méthodes permettant de réaliser des estimations de courbes de charge moyennes sur des petits domaines en intégrant de l'information auxiliaire disponible à la maille des domaines et des individus. Des tests sur données réelles nous ont permis de montrer que ces méthodes engendraient des gains de précision notables.

Ces travaux pourraient se poursuivre par la construction d'estimateurs de précision (erreur quadratique moyenne) ou par l'adaptation des méthodes de façon à les rendre robustes aux individus atypiques. Enfin, les méthodes par arbre de régression pourraient être améliorées en intégrant les variables explicatives au niveau domaine par des régressions fonctionnelles dans chaque feuille de l'arbre. On pourrait également remplacer les arbres de régression par des forêts aléatoires.

# References


[1] Ardilly, P. (2014). *Estimation régionale de taux de pauvreté utilisant une technique de calage*, Actes du 8ème colloque francophone sur les sondages, Dijon, France.

[2] Battese, G.E., 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.

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

[4] Cardot, H., Chaouch, M., Goga, C. and C. Labruère (2010). *Properties of Design-Based Functional Principal Components Analysis*. Journal of Statistical Planning and Inference, **140**, 75-91.

[5] Cardot, H., Dessertaine, A., Goga, C., Josserand, E. and Lardin, P. (2013). *Comparison of different sample designs and construction of confidence bands to estimate the mean of functional data: An illustration on electricity consumption*, Survey Methodology, **39**, 283–301.

[6] Cardot, H., Degras, D. and Josserand, E. (2013). *Confidence bands for Horvitz–Thompson estimators using sampled noisy functional data*, Bernoulli, **19**, 2067–2097.

[7] Cardot, H., Goga, C. and Lardin, P. (2013). *Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data*, Electronic Journal of Statistics, **7**, 562–596.

[8] Deville, J.-C. and Särndal, C.-E. (1992). *Calibration estimators in survey sampling*, Journal of the American Statistical Association, **87**, 376–382.

[9] Rao, J.N.K. and Molina I. (2015). *Small area estimation*, 2nd edition, Wiley.

[10] Ramsay, J. and Silverman B. (2005). *Functional data analysis*, 2nd edition, Springer.

[11] Särndal C.-E., Swensson, B. and Wretman, J. (2003). *Model assisted survey sampling*, 2nd edition, Springer.

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

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