Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Sample-based estimation of mean electricity consumption curves for small domains

by Anne De Moliner and Camelia Goga

Release date: December 20, 2018

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at STATCAN.infostats-infostats.STATCAN@canada.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

. not available for any reference period
.. not available for a specific reference period
... not applicable
0 true zero or a value rounded to zero
0^s value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p} preliminary
^{r} revised
x suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E} use with caution
F too unreliable to be published
* significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version **is also available.**

Cette publication est aussi disponible en français.

Survey Methodology, December 2018
Vol. 44, No. 2, pp. 193-214
Statistics Canada, Catalogue No. 12-001-X

Sample-based estimation of mean electricity consumption curves for small domains

Anne De Moliner and Camelia Goga[1]

Abstract

Many studies conducted by various electric utilities around the world are based on the analysis of mean electricity consumption curves for various subpopulations, particularly geographic in nature. Those mean curves are estimated from samples of thousands of curves measured at very short intervals over long periods. Estimation for small subpopulations, also called small domains, is a very timely topic in sampling theory.

In this article, we will examine this problem based on functional data and we will try to estimate the mean curves for small domains. For this, we propose four methods: functional linear regression; modelling the scores of a principal component analysis by unit-level linear mixed models; and two non-parametric estimators, with one based on regression trees and the other on random forests, adapted to the curves. All these methods have been tested and compared using real electricity consumption data for households in France.

Key Words: Regression trees; functional data; random forests; linear mixed models; robustness.

1 Introduction and context

Many studies conducted by the French electric company EDF are based on the analysis of the mean curves of electricity consumption by groups of customers who share common characteristics (e.g., similar electrical equipment or a common rate). In this text, these groups will be called domains. These mean consumption curves, also called demand curves, are estimated using a sample of several thousand curves measured at half-hourly intervals over long periods (often years).

In the literature, estimation of a total or mean demand curve for various sampling plans and the construction of confidence intervals has been examined in the recent work of Cardot, Dessertaine, Goga, Josserand and Lardin (2013), Cardot, Degras and Josserand (2013), and Cardot, Goga and Lardin (2013). The estimation of totals or means for functional data raises specific problems regarding the sample estimate of the finite population, as the strong time dependencies of the data must be exploited and preserved.

Here, we will focus on the problem of estimating mean curves for small domains, i.e., cases where we look simultaneously at several subpopulations, which may be small in size. With the advent of smart meters, it will become increasingly easy and less and less costly to create and maintain large samples of demand curves. It will therefore be possible to produce estimates of mean curves not only throughout France, but also for small geographic areas such as regions, departments and even cities. For example, these estimates could be used to propose services based on an analysis of consumption curves in territorial communities or for publication as part of an open data process.

This issue of small domains is frequently addressed in sampling theory outside the framework of functional data. The recent book by Rao and Molina (2015) proposes a state-of-the-art report on existing

1. Anne De Moliner, IMB, Université de Bourgogne Franche-Comté / EDF R&D Paris-Saclay. E-mail: anne.de-moliner@enedis.fr; Camelia Goga, LMB, Université de Bourgogne Franche-Comté. E-mail: camelia.goga@univ-fcomte.fr.

methods. When the domain is small, direct estimators (i.e., constructed solely from individuals in the sample within the domain) are not very effective. To improve the quality of estimates, auxiliary information is used and estimators are constructed based on implicit or explicit modelling of the link between quantity of interest and auxiliary information, common to all domains. In the context of EDF, this auxiliary information can, for example, be from known billing data (rate, contract power, total consumption in the previous year in particular) for each individual in the population, but also from open data proposed by the INSEE for small geographic aggregates (IRIS).

In the literature, there are estimation methods for small domains specific to temporal series. For example, Pfefferman and Burck (1990) and Rao and Yu (1994) superimpose temporal series models on series of variables and/or coefficients of the various instants to take into consideration temporal dependencies. However, those space-state-type models were developed for relatively short temporal series (a few dozen points). They are estimated using Kalman filters, which require a lot of calculation time, which would present a problem in our context, in which the number of domains studied can vary widely.

To our knowledge, the estimation of small domains in surveys for functional data has not yet been examined in the literature. To address this problem, we propose two types of methods. First, we apply parametric methods such as linear mixed models and functional linear regressions to the coordinates of the projected curves in a finite base, e.g., the base of principal components of a principal components analysis. We also propose two non-parametric methods based on regression trees and random forests adapted to the curves, respectively. All these methods are part of the model-based survey approach.

In Section 2, we formalize the problem and introduce a few notations. In Section 3, we present two direct estimators (the Horvitz-Thompson estimator and the calibration estimator for functional data) that will be the references to which we will compare ourselves to evaluate the performance of our methods. In Section 4, we propose two parametric methods based on unit-level linear functional and mixed models, adapted to the context of the functional data, and two non-parametric methods based on regression trees and random forests. For each method, we also propose a procedure for approximating the bootstrap variance. Finally, in Section 5, all estimation methods proposed in this article are tested and compared to a data set from actual electricity consumption curves for households in France. The conclusions and perspectives are presented in Section 6. In particular, the respective benefits and drawbacks of the various methods are compared in Subsection 5.4.

2 Notations and framework

Let a population of interest U of size N. A (demand) curve $Y_i(t)$ measured for each instant t belonging to an interval of time $[0, T]$ is associated with each unit i of the population. The population U can be decomposed in D disjoint domains U_d of known sizes N_d, $d = 1, \ldots, D$. Our goal is to estimate the mean curve of Y in each domain:

$$\mu_d(t) = \frac{1}{N_d} \sum_{i \in U_d} Y_i(t), \quad t \in [0, T], \quad d = 1, \ldots, D. \tag{2.1}$$

In the population U, we select a sample s of size n, based on a random sampling design $p(\cdot)$. Let $\pi_i = \Pr(i \in s)$ the probability of inclusion of unit i in sample s and assumed to be positive for all units $i \in U$. Let $s_d = s \cap U_d$ the portion of s belonging to domain U_d of random size n_d, which can be equal to 0 for one or more domains.

We assume that we also have a dimensional vector p of auxiliary variables (non-dependent on time) $\mathbf{X}_i$ that will be assumed to be known for each individual i in the population and with a known average $\overline{\mathbf{X}}_d = \sum_{i \in U_d} \mathbf{X}_i / N_d$ on the domain $d = 1, \ldots, D$.

In practice, the curves are not observed continuously, but only for a series of measurement instants $0 = t_1 < t_2 < \ldots < t_L = T$ that are also assumed to be equidistant and identical for all individuals. It is also assumed that there are no missing values.

3 Direct estimation methods in the design-based approach

In this section, we adopt the sampling design approach. This means that the variable interest values Y_i for each population unit are considered to be deterministic and the only variable present is that of the construction of the sample. The statistical inference then only describes the randomness created by the sampling design.

We will present two classic estimators, the Horvitz-Thompson estimator and the calibration estimator, which will be the references to which we will compare our methods to evaluate performances. These are direct estimators, i.e., estimators constructed by using, for the estimation of the mean for each domain, only units and auxiliary information related to the domain in question.

The functional Horvitz-Thompson estimator (Horvitz and Thompson, 1952; Cardo, Chaouch, Goga and Labruère, 2010) of μ_d is given by:

$$\hat{\mu}_d^{\mathrm{HT}}(t) = \frac{1}{N_d} \sum_{i \in s_d} d_i Y_i(t), \quad d = 1, \ldots, D, \quad t \in [0, T], \tag{3.1}$$

with $d_i = 1/\pi_i$ the sampling weight of unit i, also called the Horvitz-Thompson weight. It obviously cannot be calculated for the unsampled domains (i.e., domains d such that s_d is empty) and it is extremely unstable for small domains. Moreover, it in no way uses the predictor variables available to us.

To take advantage of the auxiliary information, again in a sampling design approach, we can use the calibration estimator proposed by Deville and Särndal (1992).

The calibration estimator for the mean μ_d is given by:

$$\hat{\mu}_d^{\mathrm{cal}}(t) = \frac{1}{N_d} \sum_{i \in s_d} w_{id}^{\mathrm{cal}} Y_i(t) \quad d = 1, \ldots, D, \quad t \in [0, T], \tag{3.2}$$

where the calibration weights w_{id}^{cal}, $i \in s_d$ are as close as possible to the sampling weights d_i units of s_d within the meaning of a certain distance or pseudo-distance $G(w, d)$ defined by the statistician:

$$\min_{w_{id}} \sum_{i \in s_d} d_i G(w_{id}, d_i) \quad \text{subject to} \quad \sum_{i \in s_d} w_{id} \mathbf{X}_i = \sum_{i \in U_d} \mathbf{X}_i. \tag{3.3}$$

For the distance of chi-square $G(w_{id}, d_i) = \sum_{i \in s_d} (w_{id} - d_i)^2 / d_i$, the weights are given by

$$w_{id}^{\text{cal}} = d_i + d_i \left(\sum_{i \in U_d} \mathbf{X}_i - \sum_{i \in s_d} d_i \mathbf{X}_i \right)' \left(\sum_{i \in s_d} d_i \mathbf{X}_i \mathbf{X}_i' \right)^{-1} \mathbf{X}_i, \quad i \in s_d$$

and the estimator becomes

$$\hat{\mu}_d^{\text{cal}}(t) = \frac{1}{N_d} \sum_{i \in s_d} d_i y_i - \frac{1}{N_d} \left(\sum_{i \in s_d} d_i \mathbf{X}_i - \sum_{i \in U_d} \mathbf{X}_i \right)' \hat{\boldsymbol{\beta}}_d(t),$$

where $\hat{\boldsymbol{\beta}}_d(t) = \left(\sum_{i \in s_d} d_i \mathbf{X}_i \mathbf{X}_i' \right)^{-1} \sum_{i \in s_d} d_i \mathbf{X}_i Y_i(t)$. The calibration weights are not dependent on time t, but they are dependent in this case on the domain d, therefore, the estimator $\hat{\mu}_d^{\text{cal}}(t)$ does not satisfy the additivity property, i.e., $\sum_{d=1}^{D} \hat{\mu}_d^{\text{cal}}(t) = \hat{\mu}^{\text{cal}}(t)$ where $\hat{\mu}^{\text{cal}}(t)$ is the calibration estimator of $\mu = \sum_{i \in U} Y_i / N$. Where the vector $\mathbf{1} = (1, 1, \ldots, 1)'$ is in the model, thus,

$$\hat{\mu}_d^{\text{cal}}(t) = \frac{1}{N_d} \sum_{i \in U_d} \mathbf{X}_i' \hat{\boldsymbol{\beta}}_d(t) = \overline{\mathbf{X}}_d \hat{\boldsymbol{\beta}}_d(t), \quad t \in [0, T].$$

If size n_d is large, this estimator is approximately bias-free regarding the sampling plan. We can consider the modified estimator:

$$\hat{\mu}_d^{\text{mod}}(t) = \frac{1}{N_d} \sum_{i \in s_d} d_i Y_i(t) - \frac{1}{N_d} \left(\sum_{i \in s_d} d_i \mathbf{X}_i - \sum_{i \in U_d} \mathbf{X}_i \right)' \hat{\boldsymbol{\beta}}(t), \quad t \in [0, T], \tag{3.4}$$

where

$$\hat{\boldsymbol{\beta}}(t) = \left(\sum_{i \in s} d_i \mathbf{X}_i \mathbf{X}_i' \right)^{-1} \sum_{i \in s} d_i \mathbf{X}_i Y_i(t), \quad t \in [0, T], \tag{3.5}$$

does not depend on domain d and, therefore, the estimator $\hat{\mu}_d^{\text{mod}}$ satisfies the additivity property, i.e., $\sum_{d=1}^{D} \hat{\mu}_d^{\text{mod}}(t) = \hat{\mu}^{\text{cal}}(t)$ where $\hat{\mu}^{\text{cal}}(t)$ is the calibration estimator of $\mu = \sum_{i \in U} Y_i / N$. As well, if n is large, it has no asymptotic bias even if size n_d is not large. The asymptotic variance functions of $\hat{\mu}_d^{\text{cal}}(t)$ and $\hat{\mu}_d^{\text{mod}}(t)$ are equal to the Horvitz-Thompson variances of residuals $Y_i(t) - \mathbf{X}_i' \hat{\boldsymbol{\beta}}_d(t)$ and $Y_i(t) - \mathbf{X}_i' \hat{\boldsymbol{\beta}}(t)$ (see Rao and Molina, 2015).

Nonetheless, for each domain, these estimates are based only on data from the domain in question (curves and explanatory variables) without considering the rest of the sample. Like the Horvitz-Thompson estimator, they are therefore inaccurate for small domains and cannot be calculated for unsampled domains.

The methods that we present in the following section will allow us, by presenting a model common to all units of the population that describes the link between variables of interest and auxiliary information, to jointly use all data from the sample to perform the estimate for each domain, and thus increase the accuracy for each one. It will also make it possible to even provide estimates for unsampled domains.

4 Model-based estimation methods

In this section, we use the model from Valliant, Dorfman and Royall (2000), in which curves Y_i are considered to be random and we propose four innovative approaches for responding to our problem estimating curves with a mean demand for small domains. Assuming that $Y_i(t)$ and the auxiliary information vector $\mathbf{X}_i$ are available for each individual i in the domain d and that the mean $\overline{\mathbf{X}}_d = \sum_{i \in U_d} \mathbf{X}_i / N_d$ is also known.

We assume that the auxiliary variables are related to the demand curves according to a functional model of superpopulation on all of the population that is generally expressed as:

$$\xi: \quad Y_i(t) = f_d(\mathbf{X}_i, t) + \epsilon_i(t), \quad i \in U_d, \quad t \in [0, T], \tag{4.1}$$

with f_d, an unknown regression function to be estimated, which can vary from one domain to another, and ϵ_i a process of zero expectation noise, zero covariance for different individuals and non-null regarding time.

If the size of domain N_d is large, then the mean μ_d will be estimated by

$$\hat{\mu}_d(t) = \frac{1}{N_d} \sum_{i \in U_d} \hat{Y}_i(t), \quad t \in [0, T],$$

where $\hat{Y}_i(t) = \hat{f}_d(\mathbf{X}_i, t)$ is the prediction of $Y_i(t)$. Otherwise, the mean μ_d is estimated by (see Valliant et al., 2000):

$$\hat{\mu}_d(t) = \frac{1}{N_d}\left(\sum_{i \in s_d} Y_i(t) + \sum_{i \in U_d - s_d} \hat{Y}_i(t)\right), \quad t \in [0, T]. \tag{4.2}$$

The quality of our estimates thus depends on the quality of our model: if the model is false, that may lead to biases in the estimates.

4.1 Functional linear model

The simplest model of form (4.1) is the functional linear regression model from Faraway (1997):

$$Y_i(t) = \mathbf{X}_i' \boldsymbol{\beta}(t) + \varepsilon_i(t), \quad t \in [0, T], \quad i \in U_d. \tag{4.3}$$

where the residuals $\varepsilon_i(t)$ are independent for $i \neq j$, distributed based on a law of means of 0 and of variance of $\sigma_i^2(t)$. If the size of domain N_d is large, then the mean of Y in domain d is estimated by

$$\hat{\mu}_d^{\text{blu}}(t) = \overline{\mathbf{X}}_d' \hat{\boldsymbol{\beta}}_{\text{BLU}}(t), \quad t \in [0, T],$$

where $\hat{\boldsymbol{\beta}}_{\text{BLU}}(t) = \left(\sum_{i \in s} \mathbf{X}_i \mathbf{X}_i' / \sigma_i^2(t)\right)^{-1} \sum_{i \in s} \mathbf{X}_i Y_i(t) / \sigma_i^2(t)$ is the best linear unbiased (BLU) estimator of $\boldsymbol{\beta}$ that does not depend on domain d. Estimator $\hat{\mu}_d^{\text{blu}}(t)$ can be expressed as a weighted sum of $Y_i(t)$:

$$\hat{\mu}_d^{\text{blu}}(t) = \frac{1}{N_d} \sum_{i \in s} w_{id}^{\text{blu}}(t) Y_i(t), \quad t \in [0, T],$$

where the weight $w_{id}^{\text{blu}}(t) = \left(\sum_{j \in U_d} \mathbf{X}_j'\right)\left(\sum_{j \in s} \mathbf{X}_j \mathbf{X}_j' / \sigma_j^2(t)\right)^{-1} \mathbf{X}_i / \sigma_i^2(t)$ now dependant on time t. If n_d / N_d is not insignificant, then the mean μ_d is estimated using (4.2) by:

$$\hat{\mu}_d^{\text{blu}}(t) = \frac{1}{N_d}\sum_{i\in s_d}\left(Y_i(t) - \mathbf{X}_i^T\hat{\boldsymbol{\beta}}_{\text{BLU}}(t)\right) + \overline{\mathbf{X}}_d'\hat{\boldsymbol{\beta}}_{\text{BLU}}(t), \quad t \in [0, T].$$

This estimator can again be expressed as a weighted sum of $Y_i(t)$ with weights that will still depend on time t. The variance function (based on the model) of $\hat{\mu}_d^{\text{blu}}(t)$ can be derived using Rao and Molina (2015), Chapter 7. The variance function $\sigma_i^2(t)$ is unknown and can be estimated by following Rao and Molina (2015). By replacing $\sigma_i^2(t)$ with $\hat{\sigma}_i^2(t)$, we will obtain the empirical best linear unbiased predictor (EBLUP) of μ_d and its variance can be obtained using the method set out by Rao and Molina (2015). This EBLUP estimator does not use the sample weight d_i and therefore is not consistent in terms of the sampling plan (unless the sample weights are constant for units in the same domain d). A modified estimator, also referred to as a pseudo-EBLUP, can be constructed using the new approach described by Rao and Molina (2015, Chapter 7), equal in this case to the estimator set out in (3.4).

If $Y_i(t)$ is unknown for the units in domain d, the following indirect estimator can be used:

$$\hat{\mu}_d^{\text{ind}}(t) = \overline{\mathbf{X}}_d\hat{\boldsymbol{\beta}}(t) = \frac{1}{N_d}\sum_{i\in s}\tilde{w}_{id}^{\text{ind}}Y_i(t), \quad t \in [0, T], \tag{4.4}$$

with $\hat{\boldsymbol{\beta}}(t)$ given in (3.5) and the weights $\tilde{w}_{id}^{\text{ind}} = \left(\sum_{j\in U_d}\mathbf{X}_j'\right)\left(\sum_{i\in s}d_i\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\sum_{i\in s}d_i\mathbf{X}_i$ are not dependant on time t, unlike w_{id}^{blu}. Thus, the estimators proposed in this section have the benefit of being able to be used for unsampled domains.

4.2 Unit-level linear mixed models for functional data

The unit-level linear mixed models proposed by Battese, Harter and Fuller (1988) are very useful in estimating total actual variables for domains. As we will see later in more detail, they can translate both the effect of auxiliary information on the interest variable (by fixed effects), and the specifics of the domains (by random effects).

In this section, we thus attempt to adapt those models to the context of functional data. To that end, we will project curves in a space of defined dimensions and in that way, transform our functional problem into several problems in estimating total or mean real uncorrelated variables for small domains, which we will then resolve using the usual methods. The use of projection bases thus makes it possible to preserve the temporal correlation structure of our data while arriving at several unrelated subproblems in estimating real variables, which we treat independently using the usual methods.

4.2.1 Estimation of curves using unit-level linear mixed models applied to PCA scores

Like PCA in finite dimensions, functional PCA is a dimension-reduction method that makes it possible to summarize information contained in a data set. It was proposed by Deville (1974), its theoretical properties were studied in Dauxois, Pousse and Romain (1982) or Hall, Mülller and Wang (2006) and, finally, it was adapted to surveys by Cardot et al. (2010).

More formally, the curves $Y_i = (Y_i(t))_{t\in[0,\,T]}$ are functions of time t and we assume that they belong to $L^2[0,\,T]$, the space of square-integrable functions in the interval $[0,\,T]$. That space is equipped with the usual scalar product $< f,\, g >= \int_0^T f(t)\, g(t)\, dt$ and the standard $\| f \| = \left(\int_0^T f^2(t)dt\right)^{1/2}$. The variance covariance function $v(s,\,t)$ defined by:

$$v(s,\,t) = \frac{1}{N}\sum_{i=1}^{N}(Y_i(s) - \mu(s))(Y_i(t) - \mu(t)), \quad s,\,t \in [0,\,T], \tag{4.5}$$

with $\mu = \sum_{i=1}^{N} Y_i \big/ N$ the mean curve of Y on the population U.

Let $(\lambda_k)_{k=1}^{N}$ the eigen values of v with $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_N \geq 0$ and $(\xi_k)_{k=1}^{N}$ the related orthonormal eigen vectors, $v(s,\,t) = \sum_{k=1}^{N} \lambda_k \xi_k(s)\, \xi_k(t)$.

The best approximation of Y in a dimensional space K smaller than N is given by the projection of Y in the space created by the first eigen vectors ξ_k, $k = 1, \ldots, q$ (Ramsay and Silverman, 2005):

$$Y_i(t) = \mu(t) + \sum_{k=1}^{K} f_{ik}\xi_k(t) + R_i(t), \quad i \in U, \quad t \in [0,\,T], \tag{4.6}$$

where f_{ik} is the projection (or score) of Y_i on the component ξ_k and $R_i(t)$, the rest representing the difference between curve i and its projection. The score f_{ik} is independent of the domain and can be calculated as the scalar product between ξ_k and $Y_i - \mu$, $f_{ik} =< Y_i - \mu,\, \xi_k >= \int_0^T (Y_i - \mu)(t)\, \xi_k(t)\, dt$. The decomposition given in (4.6) is also known as Karhunen-Loève.

Using (4.6), the mean μ_d on the domain d can be approximated by

$$\mu_d(t) \simeq \mu(t) + \sum_{k=1}^{K}\left(\frac{1}{N_d}\sum_{i\in U_d} f_{ik}\right)\xi_k(t), \quad d = 1, \ldots, D, \quad t \in [0,\,T]. \tag{4.7}$$

The unknown mean μ is estimated using the Horvitz-Thompson estimator

$$\hat{\mu}(t) = \frac{1}{N}\sum_{i\in s} d_i Y_i(t), \quad t \in [0,\,T] \tag{4.8}$$

and the ξ_k, $k = 1, \ldots, K$ are estimated by $\hat{\xi}_k$, the eigen vectors in the estimated variance-covariance function $\hat{v}(s,\,t) = \sum_{i\in s} d_i (Y_i(s) - \hat{\mu}(s))(Y_i(t) - \hat{\mu}(t)) \big/ N$ (Cardot et al., 2010).

Thus, to estimate μ_d, we must estimate the mean scores on the principal components for the domain d, i.e., $\overline{f}_{dk} = \sum_{i\in U_d} f_{ik} \big/ N_d$. To that end, for each component f_{ik}, $k = 1, \ldots, K$, we consider a unit-level linear mixed model, also known as a nested error regression model (Battese et al., 1988):

$$f_{ik} = \boldsymbol{\beta}_k' \mathbf{X}_i + \nu_{dk} + \epsilon_{ik}, \quad i \in U_d, \quad k = 1, \ldots, K, \tag{4.9}$$

with $\boldsymbol{\beta}_k' \mathbf{X}_i$ the fixed effect of auxiliary information, ν_{dk} the random effect of the domain d and ϵ_{ik} the residual of unit i. We assume that the random effects of the domains are independent and follow a common law of means of 0 and of variance of $\sigma_{\nu k}^2$. The residuals are also independent, distributed based on a law of means of 0 and of variance of $\sigma_{\epsilon k}^2$. In addition, the random effects and residuals are also assumed to be independent. The parameter $\boldsymbol{\beta}$ in the model can be estimated by $\tilde{\boldsymbol{\beta}}_k$, the best linear unbiased estimator

(BLUP) (Rao and Molina, 2015, Chapter 4.7) and the BLUP estimator $\overline{f}_{dk}$ is thus expressed as a composite estimator (see Rao and Molina, 2015):

$$\widetilde{\overline{f}}_{dk} = \gamma_k \left(\overline{f}_{dk,s} - \left(\overline{\mathbf{X}}_{d,s} - \overline{\mathbf{X}}_d \right)' \tilde{\boldsymbol{\beta}}_k \right) + (1 - \gamma_k) \overline{\mathbf{X}}_d' \tilde{\boldsymbol{\beta}}_k, \quad k = 1, \ldots, K \tag{4.10}$$

with $\gamma_k = \sigma_{\nu k}^2 / (\sigma_{\nu k}^2 + \sigma_{\epsilon k}^2 / n_d)$ and $\overline{\mathbf{X}}_{d,s} = \sum_{i \in s_d} \mathbf{X}_i / n_d$, $\overline{f}_{dk,s} = \sum_{i \in s_d} f_{ik} / n_d$ the respective means of the vectors $\mathbf{X}_i$ and the scores $\hat{f}_{ik}$ on s_d. Finally, the mean μ_d is estimated by

$$\hat{\mu}_d^{\mathrm{BHF}}(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \widehat{\overline{f}}_{dk} \hat{\xi}_k(t), \quad t \in [0, T], \tag{4.11}$$

with $\hat{\mu}$ and $\hat{\xi}_k$ the estimates of μ and k^{th} principal component ξ_k given previously.

The variances $\sigma_{\nu k}^2$ and $\sigma_{\epsilon k}^2$ for $k = 1, \ldots, K$ are unknown and are estimated by $\hat{\sigma}_{\nu k}^2$ and $\hat{\sigma}_{\epsilon k}^2$ obtained, for example, by restricted maximum likelihood (Rao and Molina, 2015). The estimator for $\overline{f}_{dk}$ is obtained by replacing γ_k in (4.10) with $\hat{\gamma}_k = \hat{\sigma}_{\nu k}^2 / (\hat{\sigma}_{\nu k}^2 + \hat{\sigma}_{\epsilon k}^2 / n_d)$ and is known as empirical best linear unbiased prediction (EBLUP). Nonetheless, the calculation of the variance function (based on the model) of $\hat{\mu}_d^{\mathrm{BHF}}(t)$ is more complicated in this case because of the estimators $\hat{\xi}_k$ of the principal components ξ_k and will be examined elsewhere.

We note that a simpler model, without the random effects, could have been considered for the scores f_{ik}:

$$f_{ik} = \boldsymbol{\beta}_k' \mathbf{X}_i + \epsilon_{ik}, \quad i \in U, \quad k = 1, \ldots, K, \tag{4.12}$$

with ϵ_{ik} a null mean residual σ_k^2. In this case, the parameter $\boldsymbol{\beta}_k$ is estimated by $\hat{\boldsymbol{\beta}}_k$, the BLUP estimator and the mean score on the domain d by $\widehat{\overline{f}}_{dk} = \hat{\boldsymbol{\beta}}_k' \overline{\mathbf{X}}_d$, $k = 1, \ldots, K$.

If the rate of n_d / N_d is not insignificant, then $\hat{\mu}_d$ is obtained using the procedure described in Section 4.1.

Note 1: Here, the PCA is not used as a dimension-reduction method, but to decompose our problem into several unrelated subproblems in the estimation of total real variables, which we know how to resolve. We thus keep a number K of principle components as high as possible, i.e., equal to the minimum number of instants of discretization and the number of individuals in the sample.

Note 2: When certain explanatory variables in vector $\mathbf{X}_i$ are categorical, our method, defined in the case of real variables, must be adapted: to that end, we propose transforming each categorical variable into a series of one hot encoding indicators $0 - 1$. As well, when the number of explanatory variables p is large, it may also be relevant to introduce penalties, RIDGE-style for example, in the regression problem.

Note 3: Other projection bases can be considered, such as wavelets (see Mallat, 1999), as they are particularly suited to irregular curves. Finally, another solution would be to apply the functional linear mixed models for curve values to the instants of discretization; however, that method would not allow for consideration of temporal correlations in the problem, unlike the previous ones.

4.2.2 Estimating variance by parametric bootstrap

To estimate the accuracy (variance based on the model) of mean curve estimators, we propose declining the parametric bootstrap method proposed by González-Manteiga, Lombarda, Molina, Morales and Santamara (2008) and then reiterated by Molina and Rao (2010). This is a replicate method that consists of generating a large number B of replicates $s^{*(b)}$, $b = 1, \ldots, B$ of size n by simple random sampling with replacement in s and randomly generating the random and fixed effects in the estimated superpopulation model. Note $\hat{R}_i(t) = Y_i(t) - \hat{\mu}(t) + \sum_{k=1}^{K} f_{k,i}\hat{\xi}_k(t)$ the estimated projection residual for the unit $i \in s$ (see also the formula in (4.6)). For $b = 1, \ldots, B$, we proceed as follows for each $t \in [0, T]$:

1. Generate the random bootstrap effects of each domain, for each principal component:

$$\nu_{k,d}^{*(b)} \sim \mathcal{N}\left(0, \hat{\sigma}_{k,\nu}^2\right), \quad d = 1, \ldots, D, \quad k = 1, \ldots, K$$

and, independent of those random effects, generate the individual bootstrap errors for each unit $i = 1, \ldots, n$ and for each principal component:

$$\epsilon_{k,i}^{*(b)} \sim \mathcal{N}\left(0, \hat{\sigma}_{k,\epsilon}^2\right), \quad i = 1, \ldots, n, \quad k = 1, \ldots, K.$$

2. Calculate the n bootstrap replicates of the projection residuals $\hat{R}_i^{*(b)}(t)$, for $i \in s^{*(b)}$ (this means selection with replacement of n projection residuals among the n residuals $\hat{R}_i(t)$).
3. Calculate the bootstrap replicates $Y_i^{*(b)}(t)$ conditional to the explanatory variables $\mathbf{X}_i$ using the estimated model:

$$Y_i^{*(b)}(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \underbrace{\left(\mathbf{X}_i'\hat{\beta}_k + \nu_{k,d}^{*(b)} + \epsilon_{k,i}^{*(b)}\right)}_{f_{k,i}^{*(b)}} \hat{\xi}_k(t) + \hat{R}_i^{*(b)}(t), \quad \forall i \in s_d^{*(b)} = s^{*(b)} \cap s_d.$$

We see that $f_{k,i}^{*(b)}$, the simulated score for the unit i, is obtained using the same approach as in González-Manteiga et al. (2008).

4. For each domain d, calculate the bootstrap replicate $\hat{\mu}_d^{*(b)}$ on the replicate $s^{*(b)}$ by declining the entire process: PCA and estimation of linear mixed models on principal components by means of EBLUP.
5. Estimate the estimator's variance $\hat{\mu}_d(t)$ by the empirical variance of the B replicates $\hat{\mu}_d^{*(b)}$:

$$\hat{V}\left(\hat{\mu}_d(t)\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\mu}_d^{*(b)}(t) - \frac{1}{B}\sum_{b=1}^{B}\hat{\mu}_d^{*(b)}(t)\right)^2.$$

This approach will also be the one used to approximate the variance of the functional linear regression (omitting step 1 of generating random effects $\nu_{k,d}^{*}$).

4.3 Non-parametric estimation using regression trees and random forests for small curve domains

In this section, to obtain individual predictions $\hat{Y}_i(t)$, we use non-parametric models, regression trees adapted to functional data, and random forests, which no longer require a linear form in the relation between

auxiliary information and interest variable and allow more flexibility in modelling. In fact, regression trees for functional data are frequently used at EDF and are known to give satisfactory results on electricity consumption curves. As well, in literature, regression trees have been adapted to surveys by Toth and Eltinge (2011), but not for estimating totals in small domains.

In this section and the next section, we are therefore seeking to estimate a specific case of the general model (4.1) in which the function f does not depend on the domain of the unit i,

$$Y_i(t) = f(X_i, t) + \epsilon_i(t) \quad \forall i \in U, \quad t \in [0, T]. \tag{4.13}$$

4.3.1 Regression trees for functional data

The classification and regression tree (CART) proposed by Breiman, Friedman, Stone and Olshen (1984) is a very popular non-parametric statistical technique. Its goal is to predict the value of a real or categorical target variable Y based on a vector of real or categorical explanatory variables $\mathbf{X} = (X_1, \ldots, X_j, \ldots, X_p)$. To that end, we determine a partitioning of the space of $\mathbf{X}$ by repeatedly splitting the data set in two, using the decision rule (split criteria) involving a single explanatory variable. Thus, our sample s is the first node λ in a tree (its "root") that we seek to subdivide into two separate nodes λ_l and λ_r such that the values of the real target variable Y_i are as homogenous as possible in each node. The inertia criterion $\kappa(\lambda)$ used to quantify the homogeneity of a node is frequently the sum of the squares of residuals between the values of Y_i for units i in node λ and the mean of those values in the node: $\kappa(\lambda) = \sum_{i \in \lambda} \left(Y_i - \overline{Y}_\lambda\right)^2$ where $\overline{Y}_\lambda$ is the mean of Y_i in node λ.

For the variables X_j which are quantitative, the decision rules are expressed as

$$\begin{cases} i \in \lambda_l & \text{if} \quad X_{ji} < c \\ i \in \lambda_r & \text{otherwise,} \end{cases} \tag{4.14}$$

with c a cut-off point to be optimized among all possible values of X_j. For qualitative variables, they consist of dividing into two separate subsets of modalities. The search for the optimal split criterion is a matter of resolving the optimization problem

$$\arg \max_{\lambda_l, \lambda_r} \left(\kappa(\lambda) - \kappa(\lambda_l) - \kappa(\lambda_r)\right). \tag{4.15}$$

Each of these nodes will then be subdivided in turn into two child nodes and the partitioning process continues until a minimal node size is obtained, until the value of the target variable is the same for all units of the node, or until a given maximum depth is attained. The final partition of the space is then made up of the final nodes in the tree, also called leaves. A summary of each of those leaves (often the mean for a quantitative target variable) then becomes the predicted variable for all units assigned to the leaf. The various parameters (minimum node size and depth) can be selected by cross-validation.

When the variable Y to be predicted is not a real variable, but a dimension vector $m > 1$, the regression tree principle extends very naturally: the tree construction algorithm and the choice of cross-validation parameters remain unchanged, but the inertia criterion is modified. Thus, the minimization problem is still

in the form of (4.15), but this time the criterion is in the form of $\kappa(\lambda) = \sum_{i \in \lambda} \left\| Y_i - \overline{Y}_\lambda \right\|^2$ where $\|\cdot\|$ is, for example, the Euclidean norm or the Mahalanobis distance norm. Multivariate regression trees were used, for example, by De'Ath (2002) in an ecological application.

Finally, when the variable to be predicted Y is a curve, the algorithm for construction of the tree and for choosing the parameters is the same, but this time a functional inertia criterion κ must be used. There are many possible choices. We chose to follow the "Courbotree" approach described in Stéphan and Cogordan (2009) and frequently used at EDF for building segmentations of data sets of electricity consumption curves based on explanatory variables. In that approach, we apply the method presented in the previous paragraph for multivariate Y on vectors $\mathbf{Y}_i = (Y_i(t_1), \ldots, Y_i(t_L))$ of curve values at the instants of discretization, with the Euclidian distance. The Euclidian distance on instants of discretization can thus be seen as an approximation of the norm $L^2[0, T]$. More formally, the functional criterion is thus expressed as

$$\kappa(\lambda) = \sum_{i \in \lambda} \sum_{l=1}^{L} \left(Y_i(t_l) - \overline{Y}_\lambda(t_l) \right)^2, \tag{4.16}$$

with $\overline{Y}_\lambda(t_l) = \sum_{i \in \lambda} Y_i(t_l) \big/ n_\lambda$ where n_λ is the number of units in the sample that belong to the node λ.

In practice, when working on electricity consumption data, the curves considered are at extremely similar levels, and the Courbotree algorithm based on the Euclidian distance may not work well when applied to raw data. Often, the Courbotree algorithm is therefore only used on the curve forms, i.e., the normalized curves $\tilde{Y}_i(t) = Y_i(t) \big/ \overline{Y}_i$ where $\overline{Y}_i = \sum_{\ell=1}^{L} Y_i(t_\ell) \big/ L$ is the mean of $Y_i(t)$ (or the level) on all measurement instants $t_1, \ldots, t_L$ (method also known as *normalized Courbotree*). We then calculate the prediction $\overline{Y}_i$ using a linear regression and finally obtain the prediction of $Y_i(t)$ by obtaining the product between the prediction of $\tilde{Y}_i(t)$ and that of $\overline{Y}_i$.

4.3.2 Variance estimation

To estimate the variance under the model of our estimators for mean curves by domain, we will follow a bootstrap approach very similar to the one proposed for parametric models. Here, our superpopulation model is expressed as

$$Y_i(t) = f(\mathbf{X}_i, t) + \epsilon_i(t), \quad \forall i \in U. \tag{4.17}$$

Let $\hat{f}(\mathbf{X}_i, t)$, for all $i \in s$ the predicted value for the unit i by regression tree, and $\hat{\epsilon}_i(t) = Y_i(t) - \hat{f}(\mathbf{X}_i, t)$, for all $i \in s$ the estimated residual for that unit. The idea of our accuracy approximation method is, as with linear mixed models, to generate a large number B of replicates $s^{*(b)}$, $b = 1, \ldots, B$ of size n by simple random sampling with replacement in s, and calculate the estimator of the mean curve by domain on each replicate and, finally, deduct the variance from the estimator by the variability between replicates. The bootstrap method used here is also known as residual bootstrap in linear model cases.

More specifically, for $b = 1, \ldots, B$, we proceed as follows for each $t \in [0, T]$:

1. Calculate the bootstrap replications of the estimated residuals $\hat{\epsilon}_i^{*(b)}(t)$ for $i \in s^{*(b)}$.
2. Calculate the bootstrap replications for $Y_i(t)$:

$$Y_i^{*(b)}(t) = \hat{f}(\mathbf{X}_i, t) + \hat{\epsilon}_i^{*(b)}(t), \quad \forall i \in s^{*(b)}$$

and recalculate, for each domain d, the mean estimator $\hat{\mu}_d^{*(b)}(t)$ on this replicate.

3. Estimate the variance by the empirical variance of the B bootstrap replicates $\hat{\mu}_d^{*(b)}(t)$, $\hat{V}(\hat{\mu}_d(t)) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\mu}_d^{*(b)}(t) - \frac{1}{B}\sum_{b=1}^{B}\hat{\mu}_d^{*(b)}(t)\right)^2$.

The process is identical if we estimate the function f by random forests rather than regression trees.

4.4 Aggregation of predictions by random forests for curves

The literature often highlights the mediocre predictive performances of regression trees compared to other techniques such as SVMs (see, for example, Cristianini and Shawe-Taylor, 2000). Regression trees can be unstable and highly dependant on the sample on which they were built. To resolve that default, Breiman (2001) proposed the random forest algorithm. This is a set technique that, as its name suggests, consists of aggregating predictions resulting from different regression trees. The fact that the aggregation of unstable predictors leads to a reduction in variance was particularly shown by Breiman (1998). For a quantitative target variable, the aggregation of predictions is performed by taking the mean predictions for each of the trees.

To reduce the variance of the aggregate prediction, the objective is to build trees that are very different from each other. The Breiman algorithm introduces variability in the construction of the trees on the one hand by means of replication (simple random sampling with replacement) of units and, on the other hand, by means of random selection, for each "split" in the tree, of a subset of candidate explanatory variables. For a regression tree, there are therefore two additional parameters to be adjusted for a random forest: the number of trees and the number of candidate explanatory variables in each split.

When the interest variable is multivariate (or functional), the algorithm proposed by Breiman adapts easily, by aggregating the multivariate (or functional) regression trees presented in the previous paragraph. Multivariate random forests have, for example, been studied by Segal and Xiao (2011).

The algorithm that we are proposing here, called "Courboforest," simply consists of aggregating the functional regression trees constructed using the "Courbotree" approach, i.e., the multivariate regression trees applied to the vectors $(Y_i = Y_i(t_1), \ldots, Y_i(t_L))$ of the values of the curves at the instants of discretization, with the split criterion being the inertia based on the Euclidean distance defined by equation (4.16).

5 Application to electricity consumption curves

We will now test the methods that we have just presented to compare their performance on electricity consumption data for French residential clients.

5.1 Presentation of the data set

We worked with a data set belonging to EDF that contains electricity consumption curves for $N = 1{,}905$ French residential clients by daily interval from October 2011 to March 2012, without any missing values ($L = 177$ points). This population is subdivided into $D = 8$ domains corresponding to geographic areas with respective sizes of 573, 195, 304, 121, 228, 219, 45 and 220. For confidentiality purposes, we cannot describe the data set in great detail, or show the mean curves by domain.

By way of illustration, Figure 5.1 shows the appearance of the standardized curves (i.e., each curve is divided by its mean calculated over the period of time studied) for five random individuals, and Figure 5.2 shows the appearance of the first five principal components of the functional PCA created for this data set.

We see that the first component, the overall appearance of which is similar to that of the mean curve, is a "level" effect. Components two and three, which present peaks during the coldest period in February, describe the sensitivity of consumption to outside temperatures. The fourth compares "mid-season" consumption to "wintertime" consumption and, finally, the fifth shows a low at about Christmas (and about February 14).

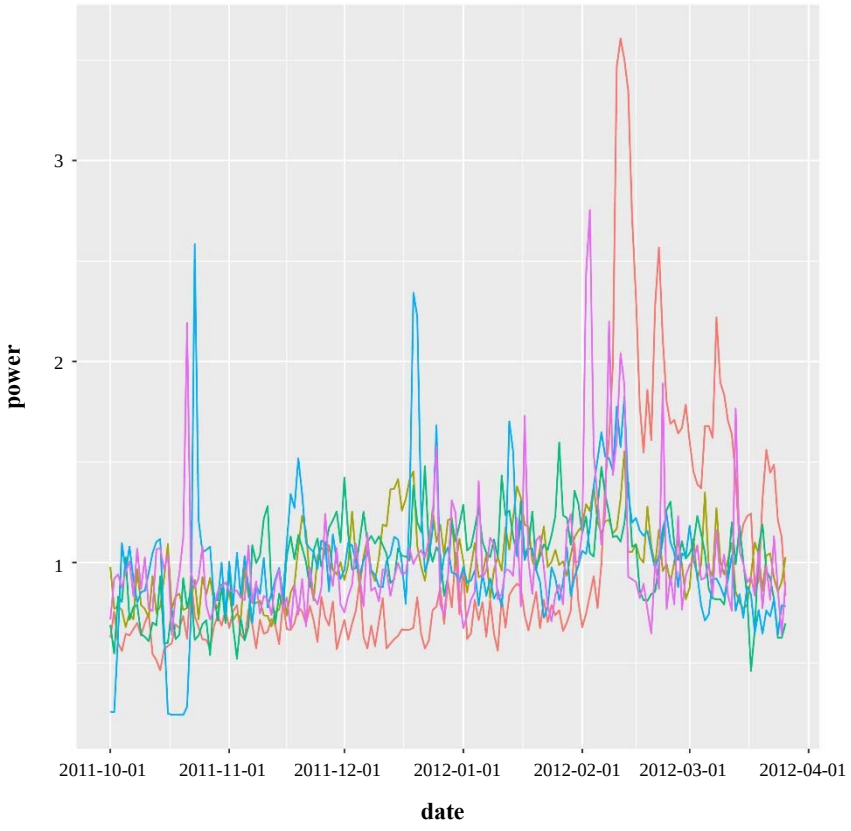

Figure 5.1 Standardized electricity consumption curves (i.e., divided by their mean over the study period) by daily interval for residential clients, winter, 2011/2012.

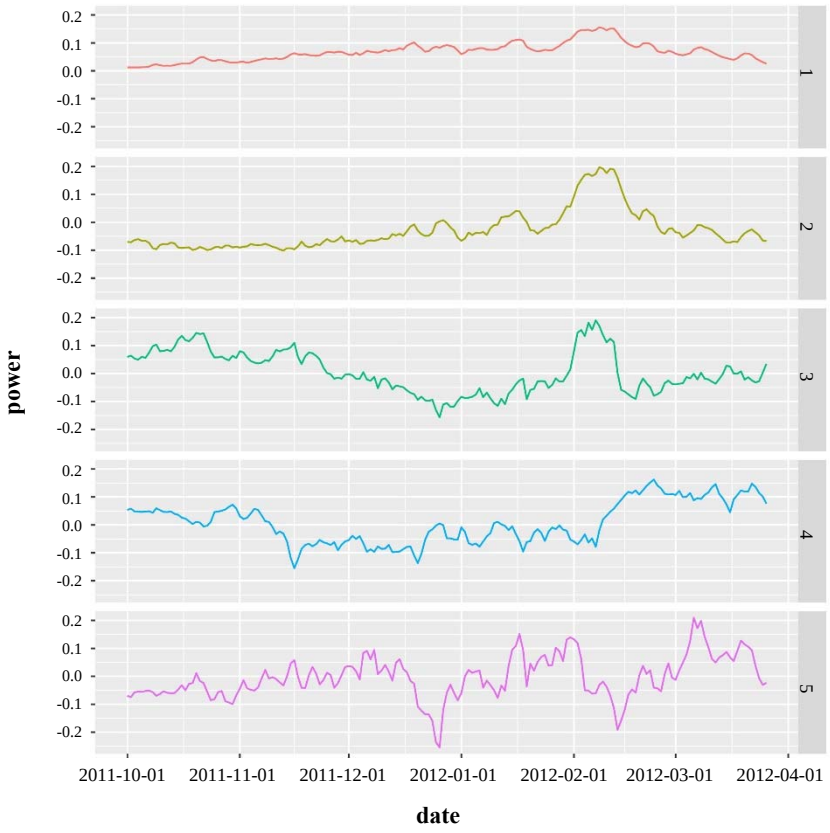

Figure 5.2 First five components of the principal component analysis.

For each individual in our population of study, we have four auxiliary variables at the individual level: contract power (in three classes), rate option (base or off-peak periods) (in the base option, the price per kWh remains constant, while the rate for off-peak periods is reduced for eight hours [referred to as off-peak]. The largest consumers tend to prefer that rate. Off-peak periods can vary from one client to another, but this factor has no impact here, as we are working on a daily interval), the previous year's annual consumption, and the type of dwelling (apartment or single home). These auxiliary variables remain the same for all methods used in order to compare identical auxiliary information. All tests were implemented in R.

5.2 Test protocol

We compare various estimators obtained using the methods set out in this chapter, for various types of modelling (unit-level linear mixed models, linear functional regressions, regression trees, random forests). We test two versions of the unit-level linear mixed model, one by placing linear mixed models on the PCA scores, as suggested in Section 4.2, and the other by applying them directly to the values of the curves of instants of discretization.

For non-parametric methods, the forests and trees have a depth (number of levels) of 5 and a minimum size of 5 leaves. There are 40 trees in the forests. The algorithms can be applied by separating the estimation of the level of the curve and its form (standardization = "yes") or not separating (standardization = "no"). To not multiply the possible combinations, we finally focused on the estimators listed in Table 5.1. The parameters of the regression tree and random forest models are set out in Table 5.2.

Table 5.1
Various estimation method tests

Title	Reference	Projection
Horvitz-Thompson	Equation (3.1)	None
Calibration	Equation (3.2)	None
Linear mixed model	Section (4.2)	None
Linear mixed model on PCA	Equation (4.11)	PCA
Linear regression	Equation (4.4)	None
Courbotree	Section (4.3)	None
Standardized Courbotree	Section (4.3)	None
Courboforest	Section (4.4)	None

Table 5.2
Parameters for trees and random forests

Title	Depth (number of levels)	Number of trees	Standardization
Courbotree	5	1	No
Standardized Courbotree	5	1	Yes
Courboforest	5	40	No

To evaluate the quality of our estimation methods, our test protocol consists of conducting a large number E of sampling simulations from our original population and then estimating the mean curve for each $D = 8$ domain based on each sample gathered by the various proposed methods. In our simulations, the eighth domain $(d = D = 8)$ will always be unsampled in order to measure the performance of our various estimators in this scenario. For each simulation, we select $n = 200$ individuals by simple random sampling from among those in the seven sampled domains $(d = 1, \ldots, 7)$.

Let $\mu_d(t_l)$ the mean curve for the domain d at the instant t_l and $\hat{\mu}_d(t_l)$ its estimator by a given method. We calculate the relative bias of $\hat{\mu}_d(t_l)$:

$$\mathrm{RB}(\hat{\mu}_d(t_l)) = 100\frac{E_{\mathrm{MC}}[\hat{\mu}_d(t_l)] - \mu_d(t_l)}{\mu_d(t_l)}, \quad d = 1, \ldots, D, \quad l = 1, \ldots, L, \tag{5.1}$$

where $E_{\mathrm{MC}}[\hat{\mu}_d(t_l)] = \sum_{e=1}^{E} \hat{\mu}_d^{(e)}(t_l) \big/ E$ is the Monte Carlo expectation of the estimator $\hat{\mu}_d(t_l)$ with $\hat{\mu}_d^{(e)}(t_l)$ the estimator of the mean curve obtained for the e^{th} simulation, for $e = 1, \ldots, E$. A second indicator, known as relative efficiency (RE), is calculated as follows:

$$\mathrm{RE}(\hat{\mu}_d)(t_l) = 100\frac{\mathrm{MSE}_{\mathrm{MC}}(\hat{\mu}_d)(t_l)}{\mathrm{MSE}_{\mathrm{MC}}(\hat{\mu}_d^{\mathrm{HT}})(t_l)}, \quad d = 1, \ldots, D-1, \quad l = 1, \ldots, L. \tag{5.2}$$

where $\mathrm{MSE}_{\mathrm{MC}}(\hat{\mu}_d(t_l)) = \sum_{e=1}^{E}\left(\hat{\mu}_d^{(e)}(t_l) - \mu_d(t_l)\right)^2 \Big/ E$ is the Monte Carlo mean square error, $d = 1, \ldots, D$, $l = 1, \ldots, L$. The lower the RE indicator, the more the estimator will be considered effective. An RE of 100 corresponds to an indicator as effective as the reference estimator.

Here, the reference estimator $\hat{\mu}_d^{\mathrm{HT}}$ is the Horvitz-Thompson estimator (which, for our simple random sampling plan, is the simple mean of the curves in the domain considered); it corresponds to the model described by equation (3.1). This estimator cannot be calculated for the unsampled domain. The RE estimator is then obtained by dividing the MSE of the various estimators by the mean MSE of the Horvitz-Thompson estimator over the seven sampled domains, i.e.

$$\mathrm{RE}(\hat{\mu}_D)(t_l) = 100\,\frac{\mathrm{MSE}_{\mathrm{MC}}(\hat{\mu}_D)(t_l)}{\overline{\mathrm{MSE}}_{\mathrm{MC}}^{\mathrm{HT}}(t_l)}, \qquad l = 1, \ldots, L, \tag{5.3}$$

with $\overline{\mathrm{MSE}}_{\mathrm{MC}}^{\mathrm{HT}}(t_l) = \sum_{d=1}^{D-1}\mathrm{MSE}_{\mathrm{MC}}(\hat{\mu}_d^{\mathrm{HT}})(t_l)$, $l = 1, \ldots, L$.

For each indicator and each instant t_l, the results obtained for the various sampled domains are then aggregated for all domains, $\mathrm{RB}_{\mathrm{ech}}(\hat{\mu})(t_l) = \frac{1}{D-1}\sum_{d=1}^{D-1}\mathrm{RB}(\hat{\mu}_d)(t_l)$ and $\mathrm{RE}_{\mathrm{ech}}(\hat{\mu})(t_l) = \frac{1}{D-1}\sum_{d=1}^{D-1}\mathrm{RE}(\hat{\mu}_d)(t_l)$ for $l = 1, \ldots, L$, while the indicators obtained for the unsampled domain are used as-is.

Finally, to evaluate overall performance, we consider the mean of those indicators for all instants in the test period, while still separating the sampled domains from the unsampled domain. We also look at the calculation times of the various estimators.

5.3 Results and test conclusions

The test results of the methods are presented in Table 5.3 and illustrated in Figures 5.3 to 5.5.

Table 5.3
Mean method performance indicators (RB, RE) for all instants of discretization and domains, separating the unsampled domain from the others

Domain type	Method	RE (%)	RB (%)
Sampled	Horvitz-Thompson	100,00	0,25
Sampled	Calibration	37,13	-0,47
Sampled	Linear mixed model	*14,69*	0,60
Sampled	Linear mixed model PCA	15,40	0,67
Sampled	Linear regression	24,87	1,20
Sampled	Courbotree	20,54	0,80
Sampled	Standardized Courbotree	22,35	1,45
Sampled	Courboforest	24,66	0,62
Unsampled	Horvitz-Thompson		
Unsampled	Calibration		
Unsampled	Linear mixed model	*13,43*	4,66
Unsampled	Linear mixed model PCA	13,49	4,77
Unsampled	Linear regression	14,38	5,09
Unsampled	Courbotree	14,29	3,48
Unsampled	Standardized Courbotree	16,63	5,88
Unsampled	Courboforest	15,97	0,37

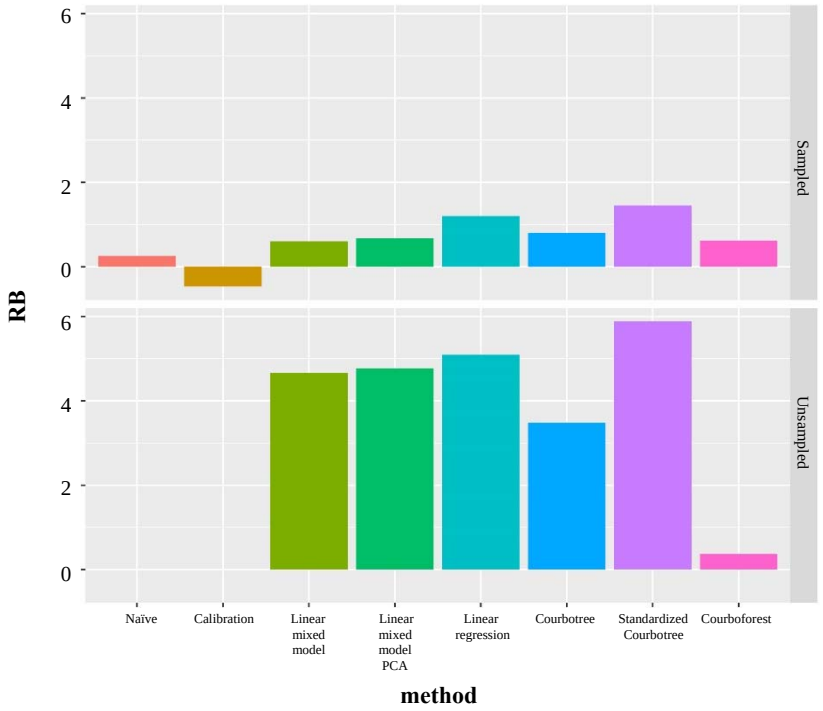

Figure 5.3 Mean relative biases as % (formula (5.1)) of estimation methods, for all instants in the domains, separating unsampled and sampled domains.

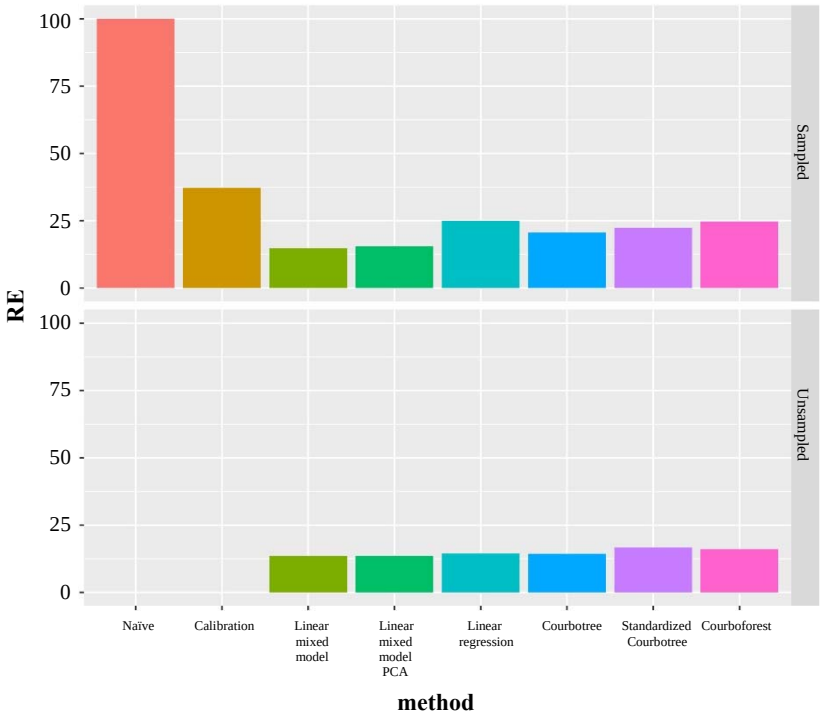

Figure 5.4 Mean relative efficiency (RE) (formula (5.2)) of the estimation methods for all instants and domains, separating unsampled and sampled domains.

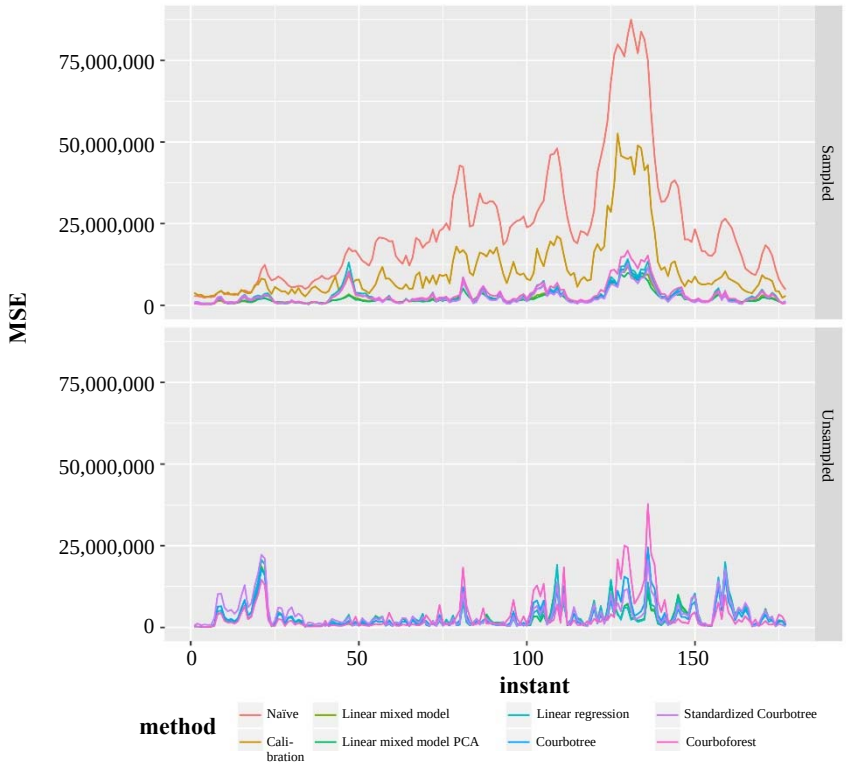

Figure 5.5 Evolution of the mean MSEs for domains over time, for the various estimation methods.

For sampled domains, we see that the integration of explanatory variables in the estimate, regardless of the method used, leads to a net gain in performance: thus, for the least effective method (the estimator by calibration), the error is divided by three when explanatory variables are used.

As well, the use of our various estimators based on superpopulation models leads to an additional gain in accuracy: the RE for our various methods thus range from 15% for linear mixed models to 25% for random forests.

The linear mixed models are the most effective method, so we can assume that there are characteristics of the domains that are unexplainable using only the auxiliary variables that this type of model is able to capture. We therefore go from an RE of 25% for the linear functional regression to an RE of approximately 15% by including these random effects.

The tree and random forest methods capture non-linearities in the relationship between explanatory variables and the interest variable, which explains why these methods give better results than linear functional regressions: the RE of the various non-parametric methods are between 20% and 25%, compared to 25% for linear functional regressions. Very surprisingly, the regression tree gives better results than the random forest. We can put forth the theory that this is because our objective is to best estimate the mean curve of a series of units, not each curve individually. It is therefore possible that the tree is not as good for predicting each curve, but better at the aggregate level. As well, on this particular data set, the method gives the best results when working on raw curves, not when distinguishing between the estimation of form and level.

Projecting curves based on the PCA does not seem to lead to any significant gains in accuracy here.

The Horvitz-Thompson estimator cannot be produced on unsampled domains. The differences between the other methods are much more restricted than on the sampled domains: the random effects cannot be estimated for unsampled domains.

Finally, in Figure 5.5, we trace the mean square error of our estimators for the sampled and unsampled domains. We note that this square error is higher in the winter (January and February). This high variability could be due to a sharp drop in outside temperatures during those months, which increases the variability of heating consumption (difference in behaviour and electrical heating equipment depending on clients). The naive and calibration estimators adapt least well to this situation.

5.4 Comparison of methods and selection criteria

Each model-based method has benefits and drawbacks. Unit-level linear mixed models are the only ones that, due to random effects, make it possible for the modelling to include domain characteristics not reflected in auxiliary information. It thus seems relevant to use them when assuming that the explanatory variables do not make it possible to explain all differences between domains.

The linear functional regression ignores the random effect of the domains, so we expect it to be less effective than linear mixed models due to its construction. Finally, the two non-parametric methods allow for better modelling of the non-linear relationships between the explanatory variables and the interest variable, but on the other hand, does not make it possible to capture the differences between domains that are not reflected in the auxiliary information. They also require the availability of auxiliary information $\mathbf{X}_i$ for each individual in the population when, in the past, we only needed mean values $\overline{\mathbf{X}}_d$ for each domain in the population and $\mathbf{X}_i$ for the sample. The choice between a parametric and non-parametric approach will therefore depend on the nature of the problem, the diversity of domains and the explanatory variables available. Be believe that neither of the two approaches is systematically preferable over the other.

A process for choosing between the two approaches could be to estimate the respective variances in the random effects and the residuals in the linear mixed models and, depending on the relative scope of those effects, moving more toward one or the other type of model. Conversely, cross-validation can be used to quantify the respective performance of the linear mixed models and the non-parametric models for predicting the aggregates of individual curves in order to direct our choice.

Among the non-parametric methods, the choice between regression trees and random forests will depend on the predictive performance of those methods on data, for the mean curves of domains. Generally, we can assume that random forests will give better results than regression trees for individual data (see Breiman et al., 1984); however, it is entirely possible that the best of the two methods for predicting each curve may not be the one that gives the best results to all domains or, at the very least, that the two methods are reduced when we consider the prediction of mean curves of individual aggregates. As well, due to their construction,

random forests require a lot more calculation time than regression trees and that aspect cannot be ignored when the data sets being processed are large in size.

6 Conclusions and outlooks

In this article, we proposed four approaches for estimating mean curves by sampling for small domains. The first two consist of projecting curves in a finite space and using the usual methods for estimating total real variables for each base vector in the projection space. In this case, we use either unit-level linear mixed models or linear regression. The last two approaches consist of predicting each curve of the unsampled units using a non-parametric model and aggregating those predictions to determine the estimated mean curves for each domain. The models used to build the predictions are regression trees adapted to functional data build using the Courbotree approach of Stéphan and Cogordan (2009) or random forests adapted to functional data built by aggregating random Courbotree trees. For each approach, we also proposed a process for approximating the variance of mean curve estimators based on a bootstrap.

Our tests showed that the linear mixed models gave the best results and, for this particular data set, made it possible to divide the error committed by approximately seven in relation to the Horvitz-Thompson estimators. The regression trees come next, followed by the linear functional regressions.

This work can be extended in various ways. In particular, we feel that the approach based on the aggregation of non-parametric estimates of curves using regression trees or random forests is promising. An interesting possibility for improvement could be the use of more relevant distances than the Euclidean distance in the split criteria that builds our regression trees. We could thus use the Mahalanobis distance, the Manhattan distance, or a “dynamic time warping” distance.

Another possibility could be to build this split criterion by applying the Euclidian distance not on the discretized curves, but on a transformation of those curves, by projection in a wavelet base, or on non-linear summaries, such as variational autoencoders from deep learning models (see, for example, LeCun, Bengio and Hinton, 2015).

We can also question the choice of depth of the regression tree, the minimum size of the leaves and the number of trees in the forest. The criteria usually used in non-parametric statistics to answer this question are usually based on the principle of cross-validation. However, our objective here is not to determine the best possible prediction for each population unit, but a prediction that gives the best estimate of the mean curve by domain, which is not necessarily the same thing. It would therefore be best to adapt the cross-validation criteria to reflect our objective.

Finally, we note that the introduction of random effects in the linear models results in improved prediction, which leads us to think that there are characteristics in the domains that are not explained solely

by the auxiliary information. It could therefore be relevant to adapt the functional regression trees to include the random effects. One solution, for example, would be to extend the algorithm from Hajjem, Bellavance and Larocque (2014), based on an EM algorithm as part of the functional data.

Acknowledgements

The authors thank Hervé Cardot for the fruitful discussions and the associate editor and two referees for their remarks and comments, which helped greatly improve this article.

References

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401), 28-36.

Breiman, L. (1998). Arcing classifiers (with a discussion and a response from the author). *The Annals of Statistics*, 26(3), 801-849.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Breiman, L., Friedman, J., Stone, C.J. and Olshen, R.A. (1984). *Classification and Regression Trees*. CRC press.

Cardot, H., Degras, D. and Josserand, E. (2013). Confidence bands for Horvitz-Thompson estimators using sampled noisy functional data. *Bernoulli*, 19(5A), 2067-2097.

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.

Cardot, H., Chaouch, M., Goga, C. and Labruère, C. (2010). Properties of design-based functional principal components analysis. *Journal of Statistical Planning and Inference*, 140(1), 75-91.

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, 2, 283-301. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2013002/article/11888-eng.pdf.

Cristianini, N., and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge University Press Cambridge.

Dauxois, J., Pousse, A. and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis*, 12(1), 136-154.

De'Ath, G. (2002). Multivariate regression trees: A new technique for modeling species-environment relationships. *Ecology*, 83(4), 1105-1117.

Deville, J.-C. (1974). Méthodes statistiques et numériques de l'analyse harmonique. In *Annales de l'INSEE*, JSTOR, 3-101.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418), 376-382.

Faraway, J.J. (1997). Regression analysis for a functional response. *Technometrics*, 39(3), 254-261.

González-Manteiga, W., Lombarda, M.J., Molina, I., Morales, D. and Santamara, L. (2008). Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model. *Computational Statistics & Data Analysis*, 52(12), 5242-5252.

Hajjem, A., Bellavance, F. and Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, 84(6), 1313-1328.

Hall, P., Müller, H.-G. and Wang, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics*, 1493-1517.

Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663-685.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

Mallat, S. (1999). *A Wavelet Tour of Signal Processing*. Academic press.

Molina, I., and Rao, J.N.K. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38(3), 369-385.

Pfeffermann, D., and Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 2, 217-237. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1990002/article/14534-eng.pdf.

Ramsay, J.-O., and Silverman, B.-W. (2005). *Functional Data Analysis*. Springer Series in Statistics, New York, Second Edition.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Rao, J.N.K., and Yu, M. (1994). Small-area estimation by combining time-series and cross-sectional data. *Canadian Journal of Statistics*, 22(4), 511-528.

Segal, M., and Xiao, Y. (2011). Multivariate random forests. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1), 80-87.

Stéphan, V., and Cogordan, F. (2009). CourboTree: Application des arbres de régression multivariés pour la classification de courbes. *La Revue MODULAD*, June.

Toth, D., and Eltinge, J.L. (2011). Building consistent regression trees from complex sample data. *Journal of the American Statistical Association*, 106(496), 1626-1636.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons, Inc.