Non-parametric estimation for aggregated functional data for electric load monitoring

Ronaldo Dias*, Nancy L. Garcia and Angelo Martarelli

June 1, 2007

Abstract

In this work we address the problem of estimating mean and covariance curves when the available sample consists on aggregated functional data. Consider a population divided into sub-populations for which one wants to estimate the mean (typology) and covariance curves for each sub-population. However, it is not possible (or too expensive) to obtain sample curves for single individuals. The available data are collective curves, sum of curves of different subsets of individuals belonging to the sub-populations. We propose an estimation method based on B-splines expansion. This method is consistent and simulation studies suggest that the proposed mean estimator is suitable even with very few replications. This problem was motivated by a real problem concerning the efficient distribution of electric energy in Southeast Brazil.

Key words: functional data, B-splines, aggregated data, electric load monitoring

AMS Classification: Primary: 62G08; Secondary:62P30

1 Introduction

The efficient distribution of energy is a problem of vital importance to the electric companies around the world. In Brazil, the concession contract between the government and the electricity

*Corresponding author: IMECC/UNICAMP - Caixa Postal 6065 - 13.087-970 - Campinas - SP - BRAZIL – dias@ime.unicamp.br and nancy@ime.unicamp.br

distribution companies states that the companies have to apply annually a minimum amount of 0.5% of its net profit into research to prevent the waste of electric energy. One of these projects consists in optimizing the existing resources avoiding construction of new plants and consequently preserving the environment. A perfect system of energy distribution would be that every electric power plant as well as transformers would have a constant load during the entire day, every day of the year. Unfortunately, this is an impossible scenario, there are peak hours during the day due to electric showers, air-conditioners, power-machines, fluorescent lights and so on. To prevent overload, all distribution network has been designed to deal with the maximum demand without considering the different types of consumers profiles. One easy way to maximize the use of existing plants and equipments is to redistribute the consumers in such way that the plants and transformers have a homogeneous demand during the whole period avoiding energy peaks. In order to achieve this goal, it is necessary to identify individual consumer profiles.

There are several different types of consumers, residential, commercial, industrial among others. Each type of consumer has a different typical curve, called *typology*. For example, it is known that, in Brazil, residential consumers have a peak on energy consumption around 6–8pm (due partially to the use of electric showers), while commercial and industrial consumers have their peak between 8am–6pm. These insights, however, are very empirical and not studied very deeply. One way to study these curves would be to take a large sample, maybe stratified, composed with different types of consumers and use functional data analysis to estimate the mean and covariance curves for each type of consumer. However, to obtain such a sample containing the individual observations is not only very costly but also extremely variable. In fact, we can say that the noise masks the signal.

On the other hand, the distribution of electric energy is done in several stages: first substations provide energy for regions in the city. This energy arrives at power transformers (*trafo*, an usual acronym for transformer) that redistributes it to micro-regions. A sample of one week data of the electric load of transformers constitutes our data set. Moreover, the market (the number of consumers of each type) of each trafo is known.

Based on this data set, the companies need to determine the effect of the type of consumers on the energy load in trafos during the whole day for all days of the week (mainly from 5am–10pm, weekdays). Their goal is to redistribute the consumers in such way as to prevent over

dimensionalization of the trafos to coup with the peak periods.

The proposed framework is to assume that each typology is a smooth curve that can be well approximated by a function belonging to a finite dimensional space $\mathcal{H}_K$ which is spanned by K (fixed) basis functions, such as Fourier expansion, wavelets, B-splines, natural splines. See, for example, Silverman (1986), Kooperberg and Stone (1991), Vidakovic (1999), Dias (1998) and Dias (2000). Although this fact might lead one to think that the nonparametric problem becomes a parametric problem, one notices that the number of coefficients can be as large as the number of observations, and there may be difficulties in estimating the curves (Silverman and Green, 1994). Moreover, if the number of observations is large, the system of equations for exact solution is too expensive to solve. This is an inheritance from the approximation theory of functions.

This paper is organized as follows: Section 2 presents the model, the notation and the proposed estimators. The performance of the estimators were studied through several simulation studies and are presented in Section 3. The analysis of the real data that motivated this work is presented in Section 4.

2 The Model

The problem we address is the estimation of sub-population mean curves when we have aggregated data. That is, each individual in a sub-population presents a distinct mean curve plus an error, but it is not possible to obtain individual measurements. The sample comes as sums of curves of a known quantity of individuals belonging to different sub-populations. Although this problem can be stated in a general setup, to make the understanding easier we will use the electric engineering jargon (motivated by our real data set). Thus each sub-population consists of different kind of consumers (residential, commercial, industrial, etc) and we observe a sample of the electric load of trafos in a discrete period of time. Each trafo has a known number of consumers of each type, known as market.

Consider that we have a sample of the electric load of M trafos. Although the load for each trafo is a continuous curve, it is observed only at T points $(t_1 < t_2 < \ldots < t_T)$. Moreover, we know that the electric load of trafo i is composed by the sum of $N_i = N_{1,i} + \ldots + N_{C,i}$ curves

where $N_{i,c}$ is the number of consumers type c, and $(N_{1,i}, N_{2,i}, \ldots, N_{C,i})$ is called the market of trafo i. Therefore, the total load $Y_{i,j}(t)$ for the ith trafo at time t of the jth day can be written as

$$Y_{i,j}(t) = \sum_{c=1}^{C} \sum_{n_c=1}^{N_{c,i}} W_{c,j,n_c,i}(t), \quad t \in [0,24], \quad i = 1, \ldots, M, \quad j = 1, \ldots, J \tag{2.1}$$

where $W_{c,j,n_c,i}(t)$ represents the load of the n_cth consumer of type c, at day j for trafo i.

We are going to assume a nonparametric regression model for the load of each individual consumer, that is, for a type c consumer we assume that there exists a *typology* (mean curve) $\alpha_c(t)$ such that

$$W_{c,j,n_c,i}(t) = \alpha_c(t) + \varepsilon_{c,j,n_c,i}(t),$$

where $\varepsilon_{c,j,n_c,i}(t)$ is a mean zero Gaussian random process. Moreover, we assume that all the processes $\varepsilon_{c,j,n_c,i}(t)$ are independent and they are identically distributed for fixed c. Therefore,

$$Y_{i,j}(t) = \sum_{c=1}^{C} N_{c,i}\alpha_c(t) + \boldsymbol{\varepsilon}_{i,j}(t), \tag{2.2}$$

where

$$\boldsymbol{\varepsilon}_{i,j}(t) = \sum_{c=1}^{C} \sum_{n_c=1}^{N_{c,i}} \varepsilon_{c,j,n_c,i}(t). \tag{2.3}$$

2.1 Estimating the typologies

In this work, we shall restrict ourselves to expand the mean curves in the well-known B-splines basis. That is, there exist a positive integer K and a knot sequence ξ such that

$$\alpha_c(t) \; = \sum_{k=1}^{K} \beta_{c,k} B_k(t), \tag{2.4}$$

where $B_i(t)$, $i = 1, \ldots, K$ are cubic B-splines. More precisely, the i-th B-spline of order m for the knot sequence ξ is defined by

$$B_i(t) = (\xi_{m+i} - x_i)[\xi_i, \ldots, \xi_{m+i}](\xi_i - t)_+^{m-1} \quad \text{for all} \quad t \in \mathbb{R},$$

where $[\xi_i, \ldots, \xi_{m+i}](\xi_i - t)_+^{m-1}$ is mth divided difference of the function $(\xi_j - t)_+^{m-1}$ evaluated at points $\xi_i, \ldots, \xi_{m+i}$, for more details see de Boor (1978).

Moreover, B-splines have an important computational property, they are splines with smallest possible support. In other words, B-splines are zero on a large set. Furthermore, a stable evaluation of B-splines with the aid of a recurrence relation is possible.

Rewriting (2.4) in a matricial form we can see the linear relationship

$$\begin{pmatrix} \alpha_c(t_1) \\ \vdots \\ \alpha_c(t_T) \end{pmatrix}_T = \begin{pmatrix} B_1(t_1) & \cdots & B_K(t_1) \\ \vdots & & \vdots \\ B_1(t_T) & \cdots & B_K(t_T) \end{pmatrix}_{T\times K} \begin{pmatrix} \beta_{c,1} \\ \vdots \\ \beta_{c,K} \end{pmatrix}_K \tag{2.5}$$

for $c = 1, 2, \ldots, C$.

Notice that the design matrix in Expression (2.5) does not depend on the consumer type c since we are using the same number of basis and same knot allocation for all types of consumers. Moreover, in this model the coefficients do not depend on the sampled points and all T points of all aggregated curves can be used to estimate the same $C \times K$ coefficients. The expression

$$Y_{i,j}(t) = \sum_{c=1}^{C}\sum_{k=1}^{K} N_{c,i}\beta_{c,k}B_k(t) + \boldsymbol{\varepsilon}_{i,j}(t), \tag{2.6}$$

is going to be used to estimate the coefficients $\beta_{c,k}$, $c = 1, \ldots, C$ and $k = 1, \ldots, K$.

Expression (2.6) can be written as a usual linear system as

$$\begin{pmatrix} Y_{1,1}(t_1) \\ \vdots \\ Y_{1,1}(t_T) \\ \vdots \\ Y_{1,J}(t_1) \\ \vdots \\ Y_{1,J}(t_T) \\ \vdots \\ Y_{I,1}(t_1) \\ \vdots \\ Y_{I,1}(t_T) \\ \vdots \\ Y_{I,J}(t_1) \\ \vdots \\ Y_{I,J}(t_T) \end{pmatrix} = \begin{pmatrix} N_{1,1}B_1(t_1)\cdots N_{1,1}B_K(t_1)\cdots N_{C,1}B_1(t_1)\cdots N_{C,1}B_K(t_1) \\ \vdots \\ N_{1,1}B_1(t_T)\cdots N_{1,1}B_K(t_T)\cdots N_{C,1}B_1(t_T)\cdots N_{C,1}B_K(t_T) \\ \vdots \\ N_{1,1}B_1(t_1)\cdots N_{1,1}B_K(t_1)\cdots N_{C,1}B_1(t_1)\cdots N_{C,1}B_K(t_1) \\ \vdots \\ N_{1,1}B_1(t_T)\cdots N_{1,1}B_K(t_T)\cdots N_{C,1}B_1(t_T)\cdots N_{C,1}B_K(t_T) \\ \vdots \\ N_{1,I}B_1(t_1)\cdots N_{1,I}B_K(t_1)\cdots N_{C,I}B_1(t_1)\cdots N_{C,I}B_K(t_1) \\ \vdots \\ N_{1,I}B_1(t_T)\cdots N_{1,I}B_K(t_T)\cdots N_{C,I}B_1(t_T)\cdots N_{C,I}B_K(t_T) \\ \vdots \\ N_{1,I}B_1(t_1)\cdots N_{1,I}B_K(t_1)\cdots N_{C,I}B_1(t_1)\cdots N_{C,I}B_K(t_1) \\ \vdots \\ N_{1,I}B_1(t_T)\cdots N_{1,I}B_K(t_T)\cdots N_{C,I}B_1(t_T)\cdots N_{C,I}B_K(t_T) \end{pmatrix} \begin{pmatrix} \beta_{1,1} \\ \vdots \\ \beta_{1,K} \\ \vdots \\ \beta_{C,1} \\ \vdots \\ \beta_{C,K} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_{1,1}(1) \\ \vdots \\ \boldsymbol{\varepsilon}_{1,1}(t_T) \\ \vdots \\ \boldsymbol{\varepsilon}_{1,J}(t_1) \\ \vdots \\ \boldsymbol{\varepsilon}_{1,J}(t_T) \\ \vdots \\ \boldsymbol{\varepsilon}_{I,1}(t_1) \\ \vdots \\ \boldsymbol{\varepsilon}_{I,1}(t_T) \\ \vdots \\ \boldsymbol{\varepsilon}_{I,J}(t_1) \\ \vdots \\ \boldsymbol{\varepsilon}_{I,J}(t_T) \end{pmatrix}. \tag{2.7}$$

That is,

$$Y = \mathbf{X}\beta + \varepsilon. \tag{2.8}$$

However, we cannot use ordinary least squares because the vector ε is not homocedastic. In fact, the variance of Y is the sum of the variance for all consumers and since each trafo has a different market, they will have distinct variance-covariance structures. In this case, we have to use generalized least square. Let $\mathbf{\Sigma}$ denote the variance-covariance matrix of $\boldsymbol{\varepsilon}_{1,1}(t_i), i = 1, \ldots, T$ and $\hat{\mathbf{\Sigma}}$ is a consistent estimator of $\mathbf{\Sigma}$. The generalized least square estimator of the parameter vector β is denoted by $\hat{\beta}$ and can be found by minimizing

$$SQR(\Theta) = (Y - \mathbf{X}\beta)^T \mathbf{\Sigma}^{-1} (Y - \mathbf{X}\beta).$$

If $\hat{\mathbf{\Sigma}}$ is block-diagonal, it is easy to find the solution of the system

$$\mathbf{X}^T \hat{\mathbf{\Sigma}}^{-1} Y = (\mathbf{X}^T \hat{\mathbf{\Sigma}}^{-1} \mathbf{X}) \hat{\beta}. \tag{2.9}$$

The independence of the random variables $Y_{i,j}(t)$, for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$, and the independence among observations from different trafos lead us to the following covariance structure for the model,

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_1 & 0 & \ldots & 0 \\ 0 & \mathbf{\Sigma}_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathbf{\Sigma}_I \end{pmatrix}_{IJT \times IJT} \qquad \mathbf{\Sigma}_i = \begin{pmatrix} Z_i & 0 & \ldots & 0 \\ 0 & Z_i & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & Z_i \end{pmatrix}_{JT \times JT} . \tag{2.10}$$

where

$$Z_i = \begin{pmatrix} Z_i(t_1, t_1) & Z_i(t_1, t_2) & \ldots & Z_i(t_1, t_T) \\ Z_i(t_2, t_1) & Z_i(t_2, t_2) & \ldots & Z_i(t_2, t_T) \\ \vdots & \vdots & \ddots & \vdots \\ Z_i(t_T, t_1) & Z_i(t_T, t_2) & \ldots & Z_i(t_T, t_T) \end{pmatrix}_{T \times T} , \tag{2.11}$$

with $Z_i(t, s) = \mathrm{Cov}\,(Y_i(t), Y_i(s))$.

The estimates of the matrix Z_i are given by

$$\hat{Z}_i(t_u, t_v) = \sum_{j=1}^{J} \frac{(Y_{i,j}(t_u) - \hat{Y}_i(t_u))(Y_{i,j}(t_v) - \hat{Y}_i(t_v))}{J - 1}, \tag{2.12}$$

where $u, v = 1, 2, \ldots, T$.

In order to find the estimates $\hat{Z}_i(t_u, t_v)$ we use a iterative algorithm. Fix $\gamma > 0$.

1. Let $\hat{Z}_i^{(0)}$, $i = 1, 2, \ldots, I$ be given by

$$\hat{Z}_i^{(0)}(t_u, t_v) = \sum_{j=1}^{J} \frac{(Y_{i,j}(t_u) - \bar{Y}_i(t_u))(Y_{i,j}(t_v) - \bar{Y}_i(t_v))}{J-1}. \tag{2.13}$$

2. $m \leftarrow 0$

3. Estimate $\hat{Y}_i^{(m+1)}(t_u)$, $u = 1, 2, \ldots, T$, using $\hat{Z}_i^{(m)}$, $i = 1, 2, \ldots, I$ in (2.9);

4. Estimate $\hat{Z}_i^{(m+1)}$ plugging $\hat{Y}_i^{(m+1)}(t_u)$ in (2.12), for $i = 1, 2, \ldots, I$;

5. If $\hat{Z}_i^{(m)}(t_u, t_v) - \hat{Z}_i^{(m+1)}(t_u, t_v) < \gamma$, for $i = 1, 2, \ldots, I$, $u, v = 1, 2, \ldots, T$;
   Then, $\hat{Z}_i \leftarrow \hat{Z}_i^{(m+1)}$, end.
   Else, $m \leftarrow m + 1$, go back to step 3;

Consistency Notice that representing the typologies as linear combination of basis functions, put our problem in the framework of regression and in this case we obtain consistency.

2.2 Estimating the covariance matrices

The main objective of this section is to propose a consistent estimator for the covariance function for each consumer type. Notice that, although there are many proposals for estimating the covariance function in the literature, see for example Hall, Fisher and Hoffmann (1994), Beder (1988), Antoniadis and Beder (1989), Rice and Silverman (1991), they are not suitable for this case. Due to the aggregated nature of the data, obtaining an estimate for the covariance function for each trafo does not automatically provide us with estimates for the consumer type. More specifically, denote

$$\sigma_c(s, t) := \text{Cov } (\epsilon_{c,j,n_c,i}(s), \epsilon_{c,j,n_c,i}(t)) \tag{2.14}$$

the covariance function for a single individual of type c belonging to the market of the ith trafo. Notice that we are assuming that this function depends only on the consumer type not on the trafo it belongs to. Therefore, if we denote

$$\Sigma_i(s, t) := \text{Cov } (\epsilon_{i,j}(s), \epsilon_{i,j}(t)) \tag{2.15}$$

the covariance function for the i trafo, we obtain the following relationship

$$\Sigma_i(s,t) := \sum_{c=1}^{C} N_{c,i}\sigma_c(s,t). \tag{2.16}$$

The crucial point here is that our interest lies into the estimation of the covariance functions $\sigma_c(s,t)$ but we do not have the individual observations. In this work, we propose a non-parametric estimator for the covariance function based on tensor product of B-splines. The proposed estimator is given by

$$\hat{\sigma}_c(t,s) = \sum_{k_1=1}^{K}\sum_{k_2=1}^{K} \hat{b}_{c,k_1,k_2} B_{k_1}(t)B_{k_2}(s) = \sum_{k_1=1}^{K}\sum_{k_2=1}^{K} \hat{b}_{c,k_1,k_2} B_{k_1,k_2}(t,s), \tag{2.17}$$

for $c = 1, 2, \ldots, C$, where $\hat{b}_{c,k_1,k_2}$ is the solution of the least square problem

$$\hat{b}_{c,k_1,k_2} = \arg\min_{b_{c,k_1,k_2}} \sum_{i=1}^{M} \left(\hat{Z}_i(t,s) - \sum_{c=1}^{C}\sum_{k_1=1}^{K}\sum_{k_2=1}^{K} N_{c,i} b_{c,k_1,k_2} B_{k_1,k_2}(t,s) \right)^2. \tag{2.18}$$

Of course, $\hat{\sigma}_c$ might not be a covariance function because it lacks the semi-definiteness property,

$$\int\int \sigma_c(s,t)h(s)h(t)\,ds\,dt \geq 0 \qquad \text{for all integrable test functions } h.$$

To overcome this difficult, we propose a two-step procedure which is similar to the one proposed by Hall et al. (1994) for stationary processes. It is well-known, Yaglom (1987), that the covariance function of a non-stationary process has spectral representation given by

$$\sigma_c(s,t) = \int_{\mathbb{R}}\int_{\mathbb{R}} \exp\{i(s\omega_1 - t\omega_2)\} d^2F_c(\omega_1,\omega_2), \tag{2.19}$$

where $F_c(\omega_1,\omega_2)$ is the spectral function. If F has a density with respect to the Lebesgue measure, this density is the spectral density f_c which is again the Fourier transform of the covariance function, that is

$$f_c(\omega_1,\omega_2) = \frac{1}{(2\pi)^2}\int_0^T\int_0^T \exp\{-i(s\omega_1 - t\omega_2)\}\sigma_c(s,t)\,ds\,dt. \tag{2.20}$$

The proposed algorithm is:

Step 1: Compute the estimate $\hat{\sigma}_c$ using (2.17).

Step 2: Compute the Fourier transform $\hat{f}_c$, of $\hat{\sigma}_c$, using (2.20).

Step 3: Obtain a new function $\tilde{f}_c$ by truncating $\hat{f}_c$ to be non-negative and if necessary smooth it.

Step 4: Invert $\tilde{f}_c$ using (2.19) to obtain the final estimate $\tilde{\sigma}_c$.

Our claim, is that, for large samples, steps 3–4 do not have to be performed since the estimate $\hat{\sigma}_c$ is consistent and consequently, is already positive-definite.

Consistency From now on, we will assume that the covariance functions σ_c are smooth enough that can be well approximate by linear combination of tensor product of B-splines, Schumaker (1981). That is, for any $\delta > 0$ there exists K, a vector of interior knots $\xi = (\xi_1, \ldots, \xi_K)$ and coefficients $b_{k_1,k_2,c}$, $c = 1, \ldots, C$, $k_1, k_2 = 1, \ldots, K$, such that

$$\|\sigma_c(s,t) - \sum_{k_1}^{K}\sum_{k_2}^{K} b_{k_1,k_2,c} B_{k_1}(s) B_{k_2}(t)\|_2^2 < \delta. \tag{2.21}$$

Therefore, for each trafo $i = 1, \ldots, M$ we have

$$\|\Sigma_i(s,t) - \sum_{c=1}^{C} N_{i,c} \sum_{k_1}^{K}\sum_{k_2}^{K} b_{k_1,k_2,c} B_{k_1}(s) B_{k_2}(t)\|_2^2 < (\sum_{c=1}^{C} N_{i,c})^2 \delta. \tag{2.22}$$

Since the observed data at points $t_1, \ldots, t_T$ follows a multivariate normal distribution we have that as the number of replications $J \to \infty$,

$$|\hat{Z}_i(t_\ell, t_{\ell'}) - \Sigma_i(t_\ell, t_{\ell'})| \to 0. \tag{2.23}$$

From the definition of $\hat{b}_{k_1,k_2,c}$ for $c = 1, \ldots, C$, $k_1, k_2 = 1, \ldots, K$ through least square minimization given by (2.18),

$$\hat{b}_{c,k_1,k_2} = \arg\min_{b_{c,k_1,k_2}} \sum_{i=1}^{M} \left(\hat{Z}_i(t,s) - \sum_{c=1}^{C}\sum_{k_1=1}^{K}\sum_{k_2=1}^{K} N_{c,i} b_{c,k_1,k_2} B_{k_1,k_2}(t,s) \right)^2$$

consistency follows from standard arguments.

3 Simulation Studies

In the simulation study, we will use the same jargon and notation as the previous section. We will use Equation (2.2) to generate the data according to fictitious typologies α_1, α_2 and α_3

given by

$$\begin{aligned}
\alpha_1(x) &= 0.1(0.4 + \exp(-(x-6)^2/3) + 0.2\exp(-(x-12)^2/25) \\
&\quad +0.5\exp(-(x-19)^2/4)) \qquad (3.1)\\
\alpha_2(x) &= 0.1(0.2 + \exp(-(x-5)^2/4) + 0.25\exp(-(x-18)^2/5)) \qquad (3.2)\\
\alpha_3(x) &= 0.35 + 0.6\exp(-(x-10)^2/8) + 0.5\exp(-(x-16)^2/8) \qquad (3.3)
\end{aligned}$$

for $x \in [0, 24]$ and presented in Figure 3.1. These functions mimic the intuitive behavior of the energy load as perceived by the electricity companies for Brazilian residential and commercial consumers.

Moreover, a reasonable assumption is that the variance function is proportional to the mean

$$\sigma_c(t,t) = 0.30\alpha_c(t), \text{ for } c = 1, 2, 3.$$

The covariance function was arbitrarily chosen to be

$$\begin{aligned}
\sigma_c(t,s) &= (0.75 - 0.25|t-s|)\sqrt{\sigma_c(t,t)\sigma_c(s,s)}, \text{for } |s-t| \leq .50\\
&= 0, \text{ for } |s-t| \geq .75\\
&= \text{ smoothly decaying for } .50 < |s-t| < .75.
\end{aligned}$$

For each curve we will generate 96 points (one observation every 15 minutes) and J replications.

3.1 Two types of consumers

In this scenario we will consider only the fictitious residential typologies α_1 and α_2 given by (3.1) and (3.2). First, we will study the consistency of the estimator, that is, what happens as J increases. Therefore, we fix the number of consumers in each trafo to be 50, see Table 3.1.1.

In the simulation, we tried to build the market in such way each consumer type is favored by one of the trafos and one trafo is balanced. Of course, if we have a sample which favors one of the consumers, this one will get better estimates than the others. The fit was made using 14 B-splines basis (10 internal knots) and 5, 50 and 300 replications. From Figures 3.2 and 3.3 we can see that the estimation for the typologies is very good, even with 5 replications (that would correspond to sample the energy load during only one week). For 300 replications we

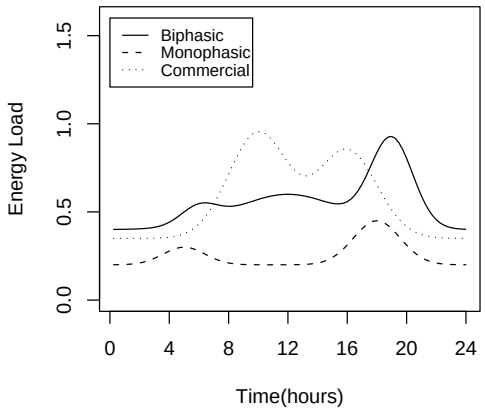

Figure 3.1: Fictitious typologies used in the simulation study.

Table 3.1: Market for trafos 1 – 3 for single phase and two phase consumers.

	Single Phase	Two Phase	Total
Trafo 1	10	40	50
Trafo 2	40	10	50
Trafo 3	25	25	50

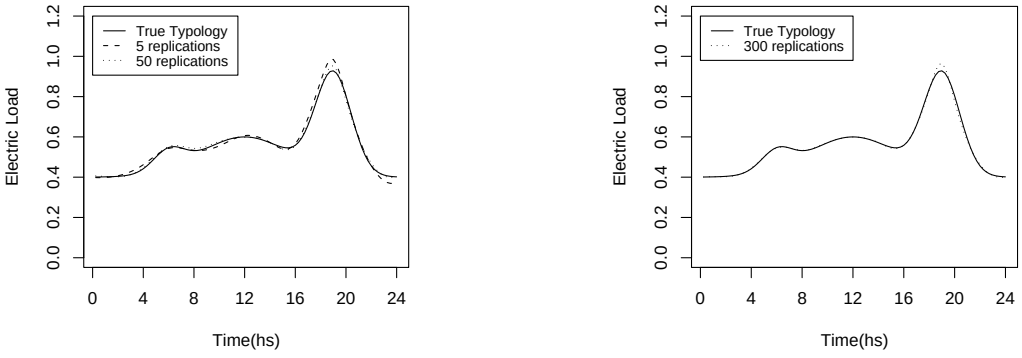

Figure 3.2: Estimated typologies for single phase consumers

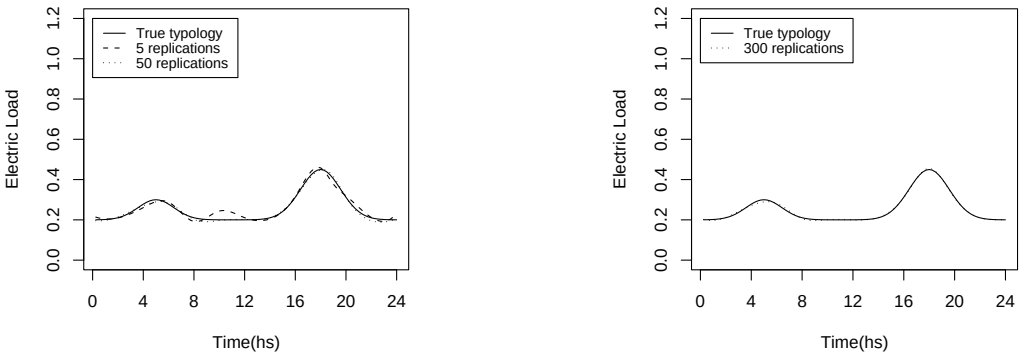

Figure 3.3: Estimated typologies for two phase consumers

have an almost perfect fit. Figures 3.4 and 3.5 show the estimated variances and covariances. As expected, we need more replications to get a good fit, however, we can see that the estimation procedure is consistent. In all cases, the estimated curve was already semi-positive definite and we did not have to invert the Fourier transform.

Table 3.2: Market for trafos 1.1 – 3.1 for single phase and two phase consumers.

	Single Phase	two phase	Total
Trafo 1.1	20	80	100
Trafo 2.1	80	20	100
Trafo 3.1	50	50	100

In order to study the effect of the market on the estimates, we changed the market to 100 consumers for each trafo maintaining the proportions of Table 3.1. Since the estimation gets better and better as we increase the number of replications, the comparison was made using only 5 replications. We defined three new trafos shown in Table 3.1 with twice many consumers. To be fair in the comparison, trafos 1.1 – 1.3 were created by adding consumers to trafos 1 – 3. Figure 3.6 shows that increasing the number of consumers in each trafo does not improve significantly the estimation if we get enough consumers in the first place.

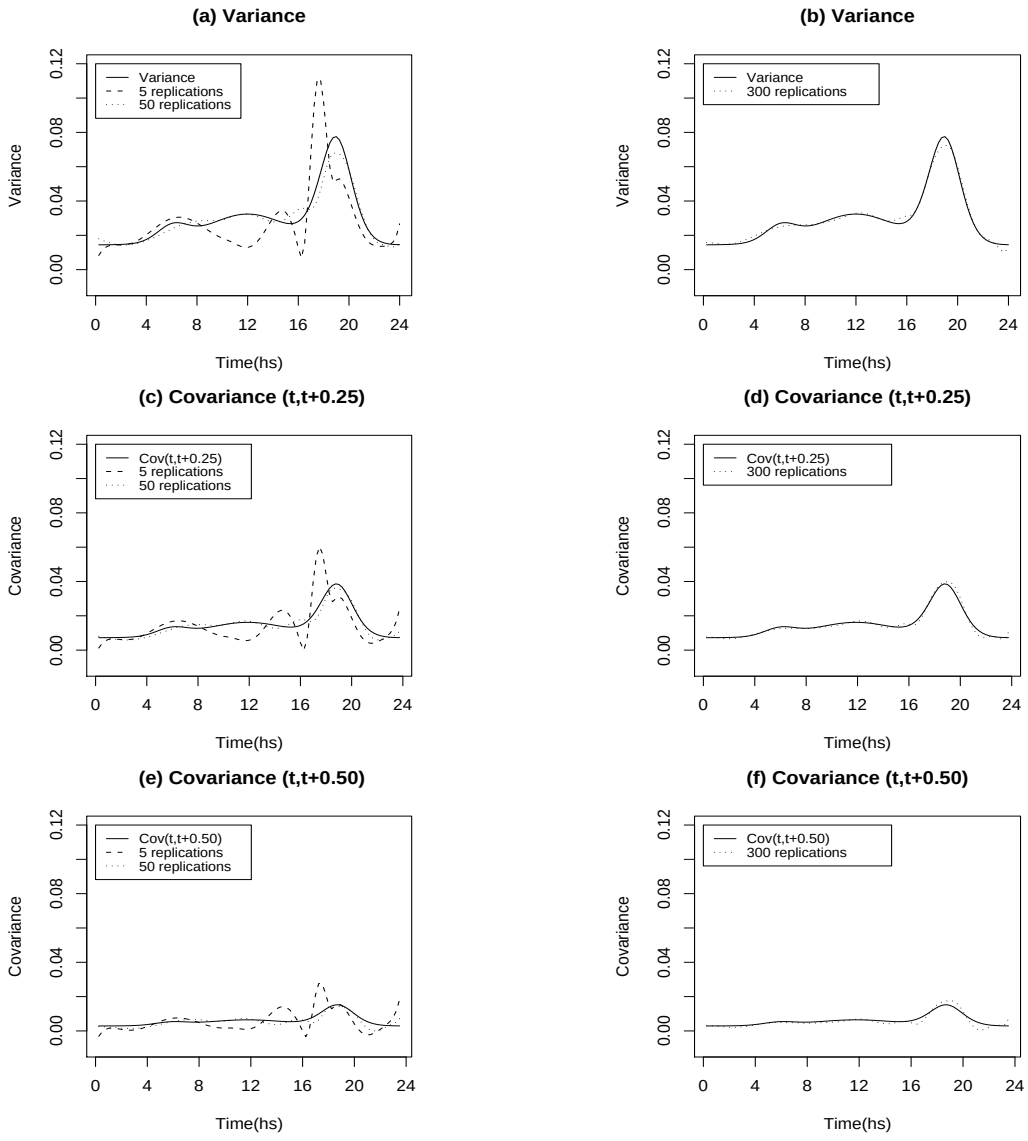

Figure 3.4: Estimated variance and covariance curves for single phase consumers.

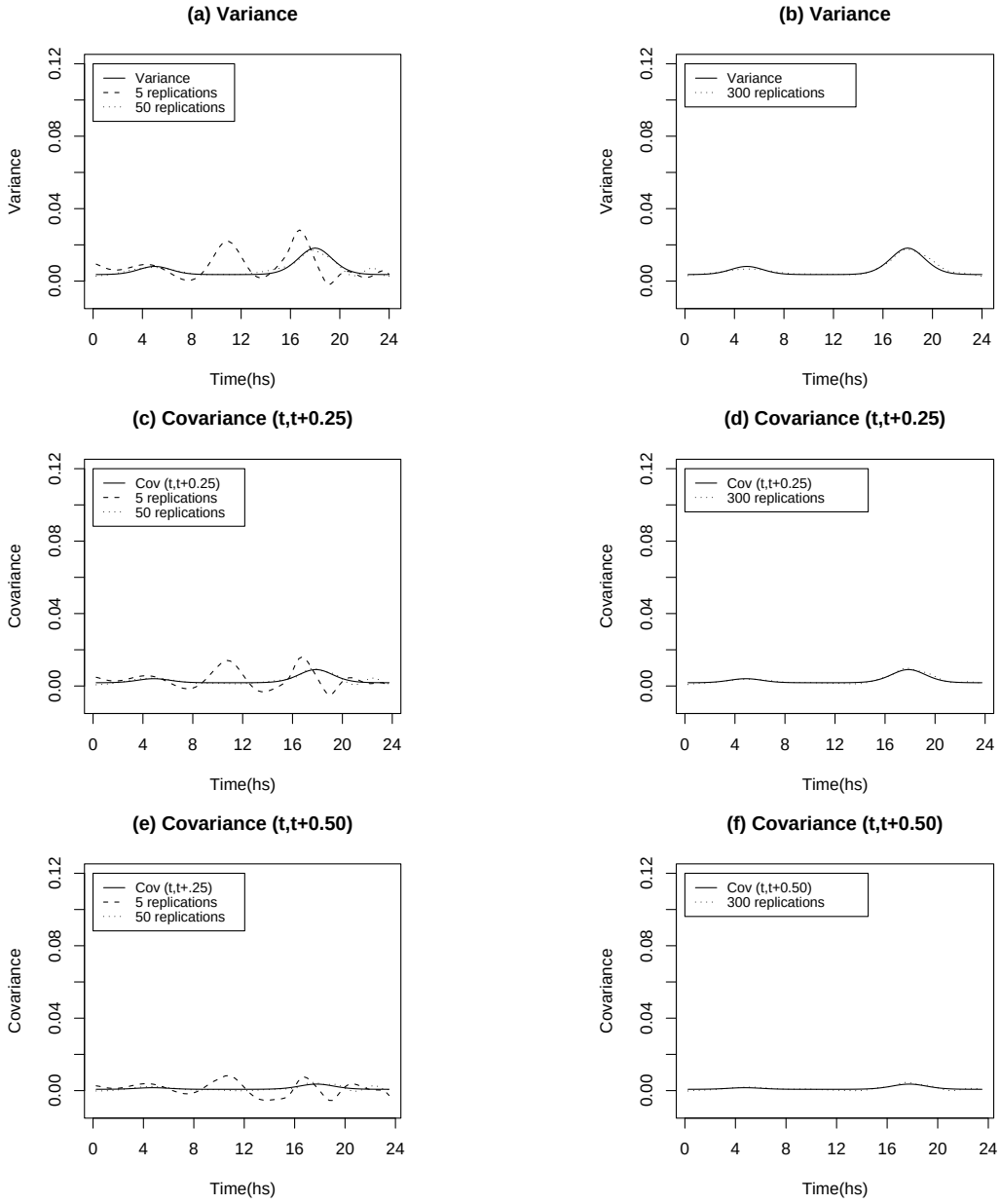

Figure 3.5: Estimated variance and covariance curves for two phase consumers.

3.2 Three types of consumers

In this scenario we will consider all the fictitious typologies presented in Figure 3.1. Again, we will study the consistency of the estimator, that is, what happens as J increases. Therefore, we fix the number of consumers in each trafo to be 50, see Table 3.2. In the simulation, trafos 1 – 7 were constructed in such way that we get one market with same frequency for each

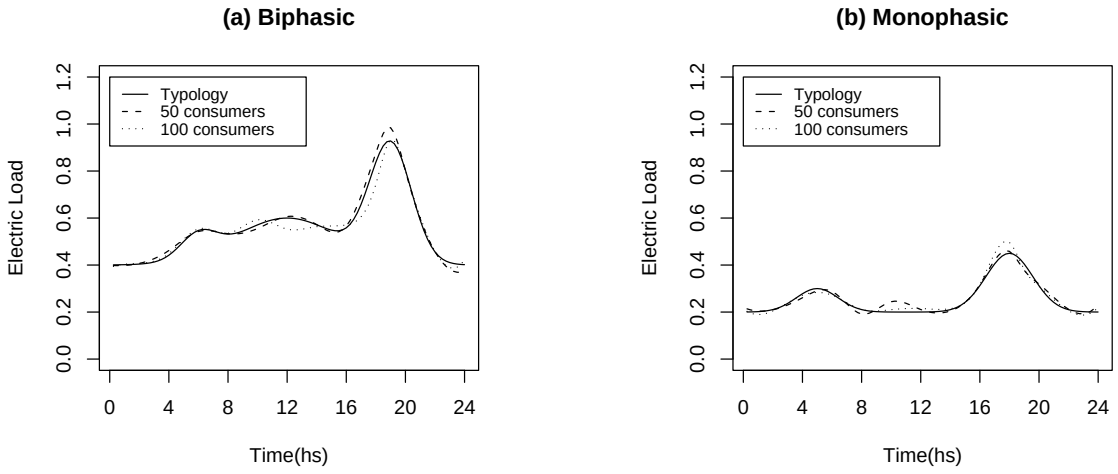

Figure 3.6: Estimated typologies for two phase and single phase consumers for Trafos 1.1–1.3

type of consumer and two other favoring one type of consumer. Since we assume that there is independence among the consumers, if necessary, we could add up trafos in order to get a more suitable market for estimation. This is the idea behind constructing trafos 8 – 10 by adding trafos 2 and 5, 3 and 6, 4 and 7 respectively.

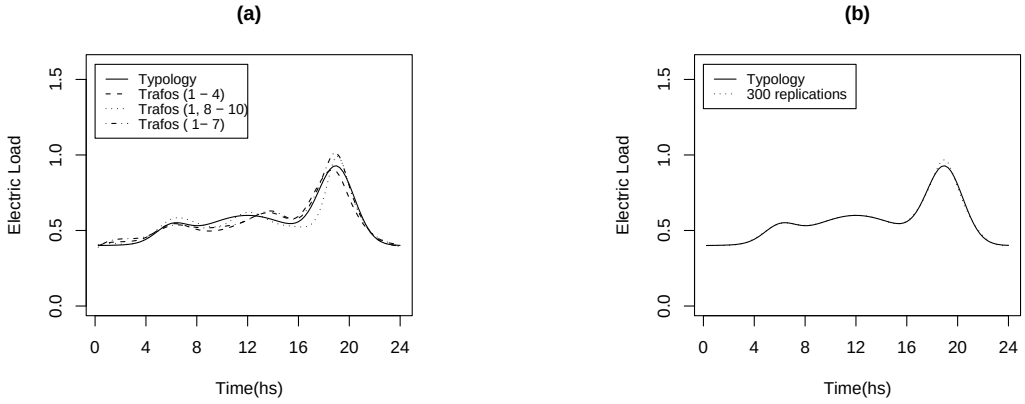

Figure 3.7: Comparison among the estimates for two phase consumers using (a) 10 replications and trafos (1 – 4), (1, 8 – 10) e (1 – 7), (b) 300 replications and trafos (1 – 4).

The typologies were estimated using 14 B-splines basis (10 internal knots). To help visualization we used only 10 replications to compare between different markets. Also, we can see

Table 3.3: Market for trafos 1 – 10 for 3 types of consumers.

	Single Phase	Two Phase	Commercial	Total
Trafo 1	17	16	17	50
Trafo 2	40	5	5	50
Trafo 3	5	40	5	50
Trafo 4	5	5	40	50
Trafo 5	30	10	10	50
Trafo 6	10	30	10	50
Trafo 7	10	10	30	50
Trafo 8	70	15	15	100
Trafo 9	15	70	15	100
Trafo 10	15	15	70	100

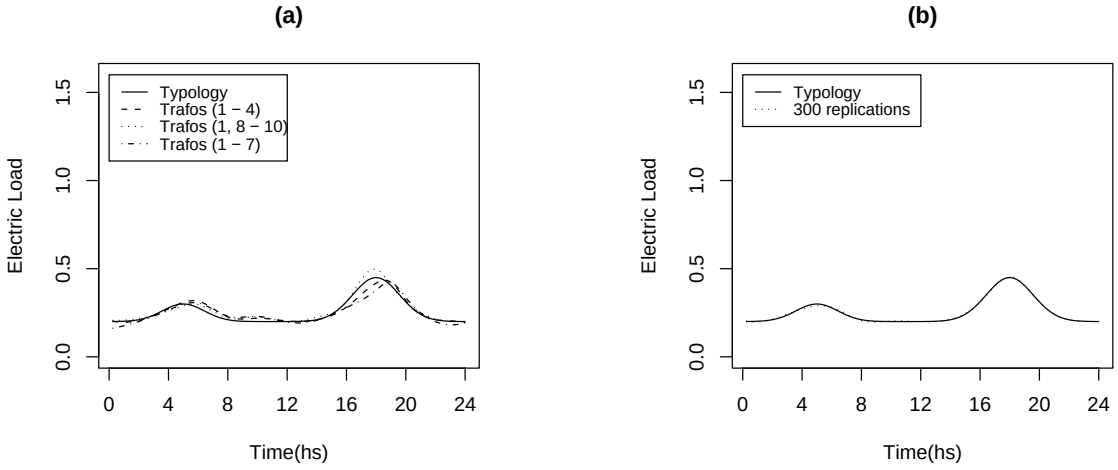

Figure 3.8: Comparison among the estimates for single phase consumers using (a) 10 replications and trafos (1 – 4), (1, 8 – 10) e (1 – 7), (b) 300 replications and trafos (1 – 4).

that the estimator appears to be consistent when the number of replication increases to 300. Figures 3.7 – 3.9 present the estimated functions. We can see that there is not really much difference among the estimates even for 10 replications. Notice that increasing the number of

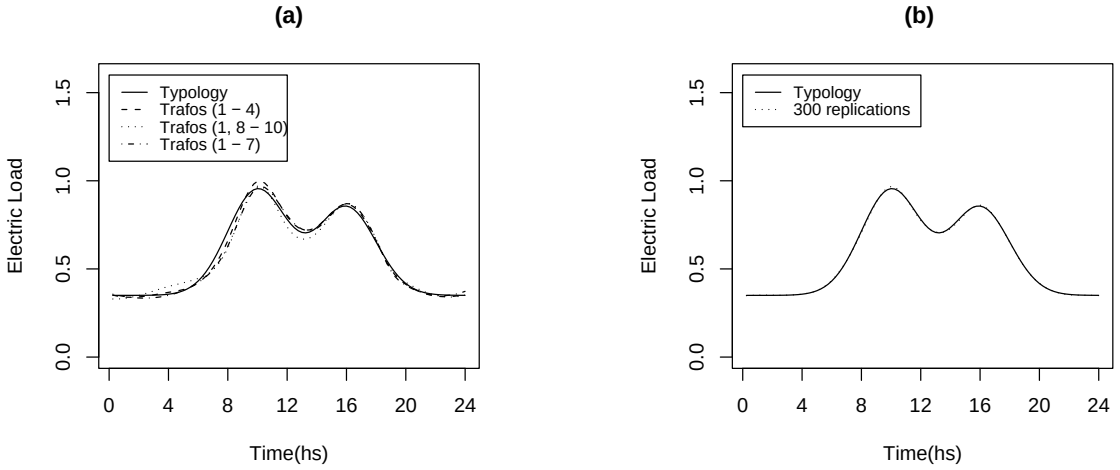

Figure 3.9: Comparison among the estimates for commercial consumers using (a) 10 replications and trafos (1 – 4), (1, 8 – 10) e (1 – 7), (b) 300 replications and trafos (1 – 4).

consumers in the trafos do not help to improve the fit. Also, the fit is better when we have linear independent markets. In fact, if we have one trafo with a market that is just a factor from other, we cannot use both in the equation, we should either choose one of them or add them up. This is the reason why the fit using Trafos 1 –4 presented better results than when we used Trafos 1, 8 – 10, since the proportion of consumers two phase and commercial are bigger in trafos 2 – 4.

4 Real data set

The data used in this section were provided to us by CPFL, a company that distributes electric energy in Southeast Brazil. The identification of typologies is part of a bigger project which has as final goal a rational use of the available resources in order to save energy, equipments and the need to build more and more hydroelectric, thermal or nuclear power plants which may cause serious damage to the environment.

The electric load of trafos were observed during one week every 15 minutes. The market for each trafo is small and variable, consisting of single phase and two phase residential, commercial and industrial consumers among others. We just studied the most prolific types of consumers

(single phase, two phase residential and commercial ones). Industries, in general, are big consumers of energy and each one of them require one or more trafos. Therefore, our analysis is not appropriate for this type of situation.

4.1 Example 1 - Residential consumers

For security measures in Brazil, houses are loaded only with energy tension either 127V or 220V. Therefore, the houses are classified as monophasic (single phase/ 127V) or biphasic (two phases/220V). In general, single phase residencies are more modest due to smaller cost. This reinforces the idea that they have different typologies. In this example we will use two trafos called TR07 and TR09. Figure 4.1 shows the observed curves.

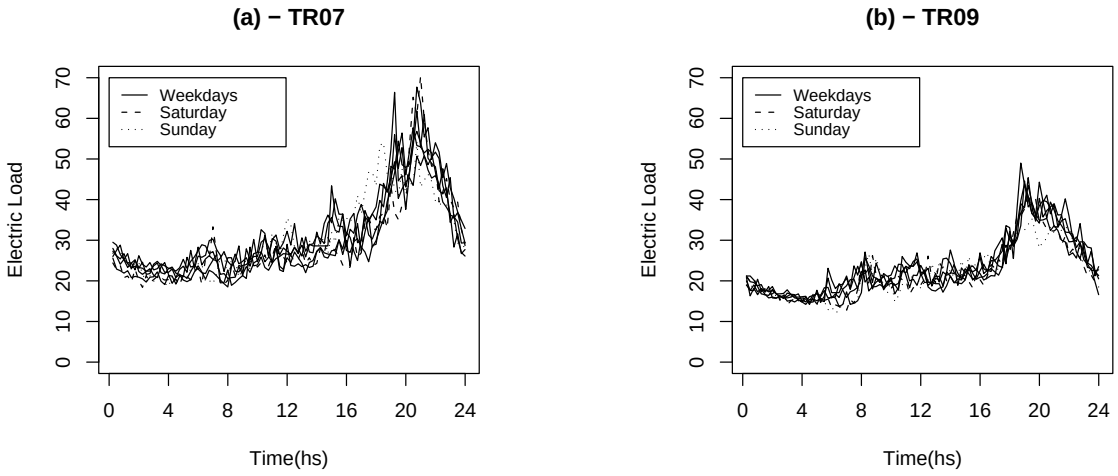

Figure 4.1: Observed curves for TR07 e TR09.

As explained earlier, to improve the performance of the estimator it is desirable to have trafos representing each type of consumer, that is, with a market favoring each type of consumer. However, in the data set we could not find a trafo with majority of two phase consumer, therefore we chose to use a trafo with a equally distributed market. Moreover, we have a sample from just one week and from Figure 4.1 it seems that probably there is a distinct behavior between weekday and weekend curves. Since there are no visible outliers, our sample will consist of 5 weekday curves for each trafo. Table 4.1 presents the market for this example.

Table 4.1: Market for trafos TR07 e TR09.

Trafo	Single Phase	Two Phase	Total
TR07	87	5	92
TR09	25	25	50

4.2 Results

We used 13 B-spline basis with 9 internal knots equally distributed in the interval $[0, 24]$ to estimate the typologies that can be seen in Figure 4.2.

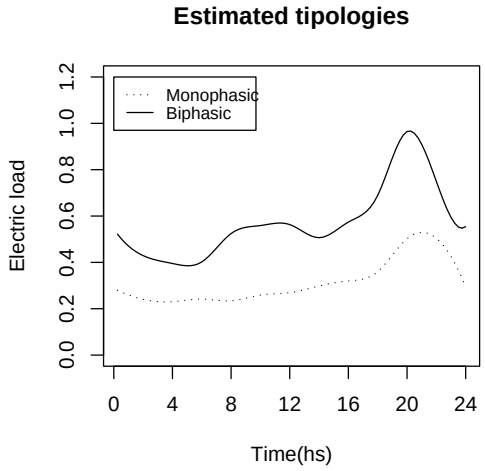

Figure 4.2: Estimated typologies for single phase and two phase consumers based on trafos TR07 and TR09.

As expected the single phase houses have a smaller load than the two phase residencies. Both types have a peak around 8pm, as this coincides with coming home, taking showers, etc. The basic difference is that for two phase residencies there is an increase in the electric load from 8am to 12pm which is not observed for the single phase consumers. To check the goodness of fit, we estimate the electric load for each trafo through the estimated typologies and obtained a very good fit as can be seen from Figure 4.3. There is a slight bias in the estimation for TR09 between 5 and 8pm. This is caused probably by the fact that the number of single phase consumers for this trafo is significantly bigger than the two phase consumers.

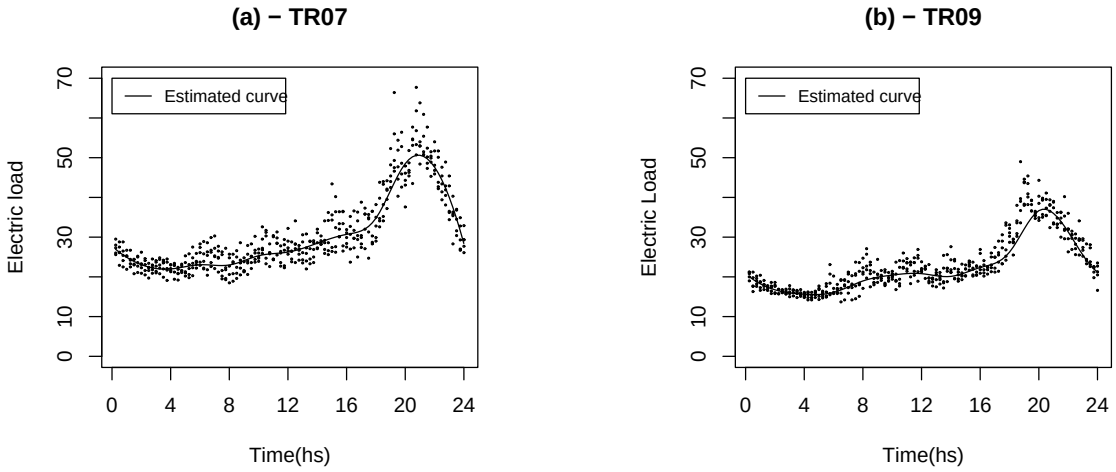

Figure 4.3: Observed and estimated curves for trafos TR07 e TR09.

4.3 Example 2 - Residential and commercial consumers

In this example we will deal with 3 types of consumers: commercial, single phase and two phase residential represented by trafos TR04, TR03 and TR02. Figure 4.4 shows the observed curves and Table 4.3 presents the market for these trafos. Again, we will use only weekdays curves for this analysis.

Table 4.2: Market for trafos TR04, TR03 e TR02.

Trafo	Commercial	Single Phase	Two Phase	Total
TR04	3	48	26	77
TR03	3	4	43	50
TR02	7	5	29	41

Notice that the relative frequency of commercial consumers is very low compare with residential ones. Moreover, two phase consumers are more prevalent in this scenario which will give us a better estimate of their typology.

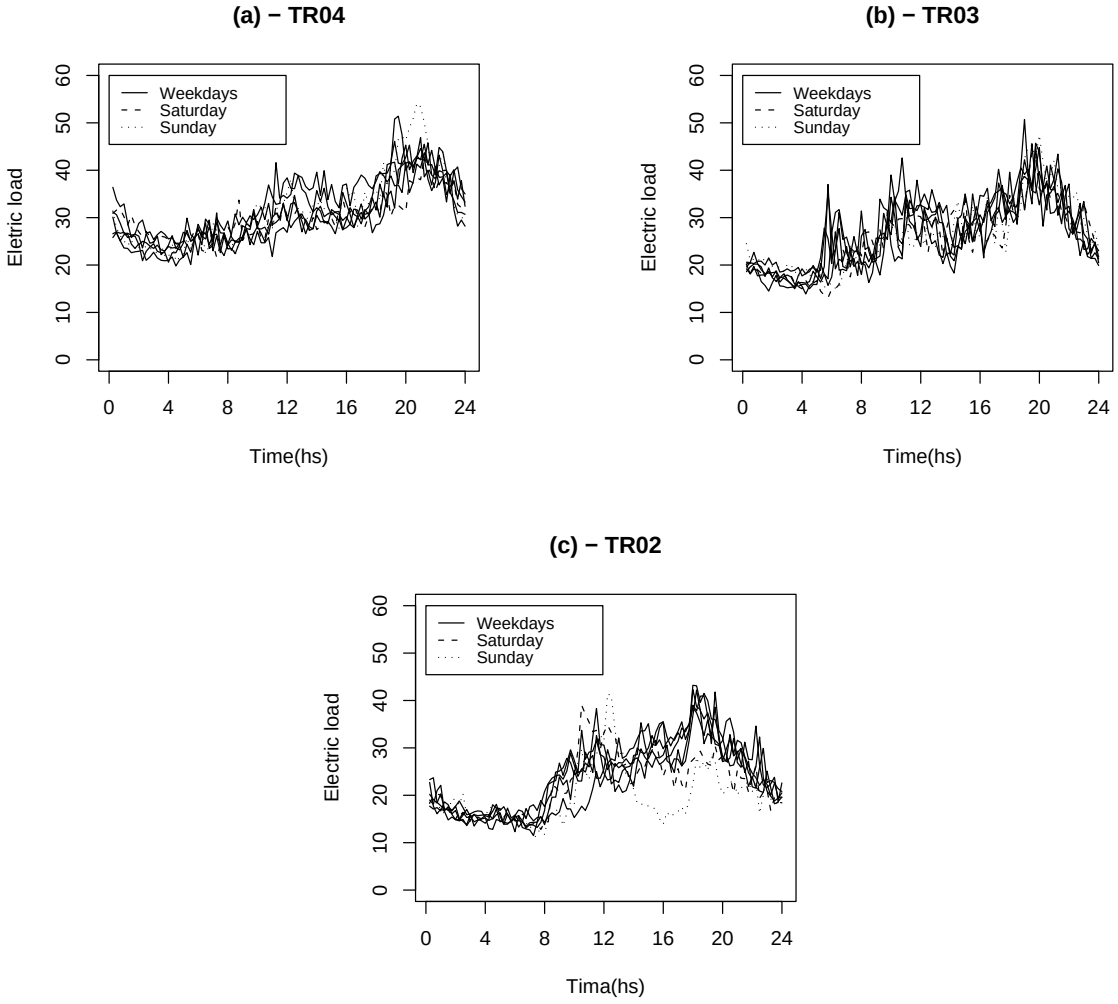

Figure 4.4: Observed curves for TR04, TR03 e TR02.

4.4 Results

We used 11 B-spline basis with 7 internal knots equally distributed in the interval $[0, 24]$ to estimate the typologies that can be seen in Figure 4.5. Figure 4.7 presents the estimated aggregated curves for trafos TR04, TR03 and TR02 obtained through the individual typologies.

Typologies for single phase and two phase consumers were estimated using different methods and trafos. Figure 4.6 compares these estimates and we can see that they basically agree in the main features. Due to the distribution of markets single phase typology is better estimated

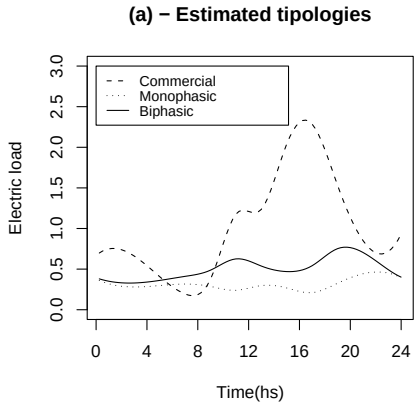

Figure 4.5: Estimated typologies for commercial, single phase and two phase consumers based on trafos TR04, TR03 and TR02.

using Example 1 and two phase typology is better estimated using trafos from Example 2.

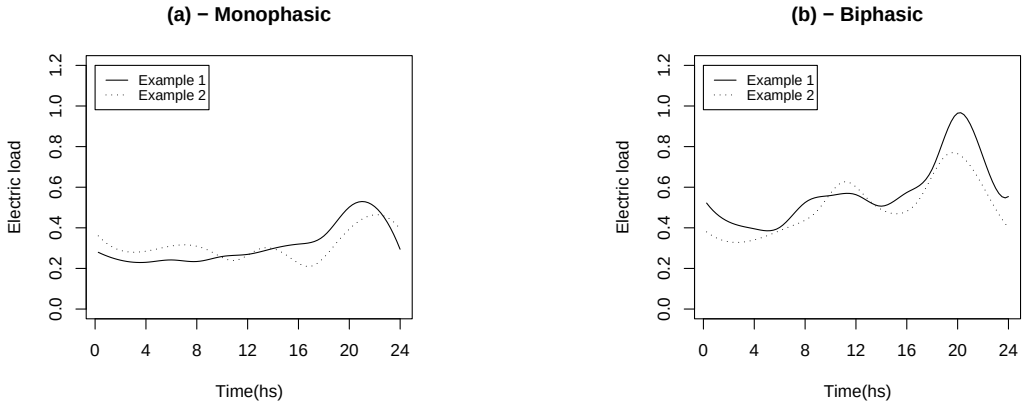

Figure 4.6: Comparing the estimated typologies for residential consumers in Examples 1 and 2

As in Example 1, the single phase houses presented smaller load than the two phase residencies. Both types have a peak around 8pm. Commercial consumers presented a higher load from 8am to 8pm with a peak between 1pm and 7pm. To check the goodness of fit, we estimate the electric load for each trafo through the weighted sum of the estimated typologies and obtained a very good fit as can be seen from Figure 4.7.

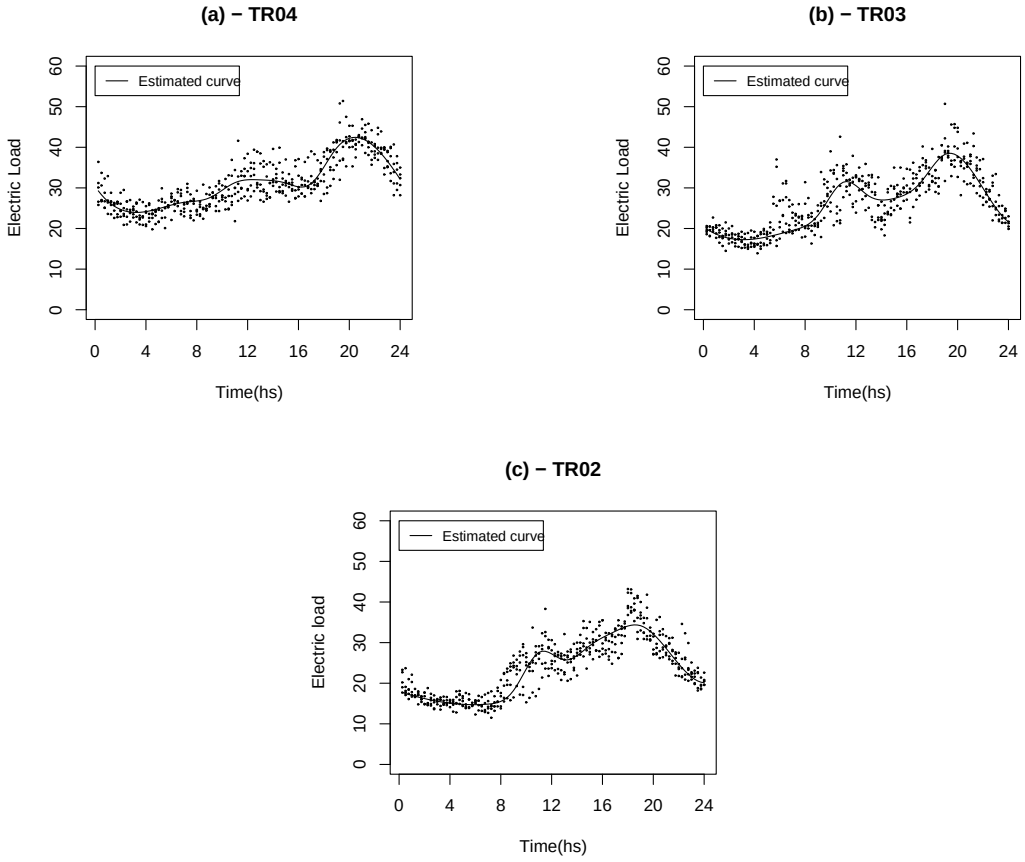

Figure 4.7: Observed and estimated curves for trafos TR04, TR03 e TR02.

4.5 Final considerations

The problem of aggregated functional data analysis, where curves cannot be individually observed, is not restricted to environmental problems, it can appear in other contexts such as finance and medicine. Despite this fact, it has not received much attention in the literature, particularly for the estimation of the covariance function. The usual methodology and transformations cannot be applied directly to this type of data and obtaining a good estimator for the sum of the curves does not give a good estimator for the individual covariances. This is a fascinating topic for current and future research.

One striking feature of our modeling is that the estimates for the typologies were very good even with very few replications. Another important remark is to notice the great influence of the market distribution in terms of improving the estimates. When collecting data, it is crucial

to choose the trafos in order to have a balance between each type of consumer. The good news is that, if there is independence among the consumers, it is always possible to add the energy load from several trafos and obtain a better sample. One surprise is that, the number of consumers in each trafo did not affect the goodness of fit if we have a minimum number of them to begin with.

Acknowledgments: We would like to thank John Rice for many fruitful discussions and insights. We also thank Jane-Ling Wang for discussions on covariance estimates. CPFL presented us with the problem and the data set. This work was partially funded by FAPESP grant 02/01554-5 and CNPq grants 301054/1993-2, 300644/1994-9 and 475763/2003-3.

References

Antoniadis, A. and Beder, J. H. (1989). Joint estimation of the mean and the covariance of a Banach valued Gaussian vector, *Statistics* **20**(1): 77–93.

Beder, J. H. (1988). A sieve estimator for the covariance of a Gaussian process, *Ann. Statist.* **16**(2): 648–660.

de Boor, C. (1978). *A Practical Guide to Splines*, Springer Verlag, New York.

Dias, R. (1998). Density estimation via hybrid splines, *Journal of Statistical Computation and Simulation* **60**: 277–294.

Dias, R. (2000). A note on density estimation using a proxy of the Kullback-Leibler distance, *Brazilian Journal of Probability and Statistics* **13**(2): 181–192.

Hall, P., Fisher, N. I. and Hoffmann, B. (1994). On the nonparametric estimation of covariance functions, *Ann. Statist.* **22**(4): 2115–2134.

Kooperberg, C. and Stone, C. J. (1991). A study of logspline density estimation, *Computational Statistics and Data Analyis* **12**: 327–347.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically whe the data are curves, *Journal of the Royal Statistical Society. Series B.* **53**(1): 233–243.

Schumaker, L. L. (1981). *Spline Functions: Basic Theory*, WileyISci:NJ.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall (London).

Silverman, B. W. and Green, P. J. (1994). *Nonparametric Regression and Generalized Linear Models*, Chapman and Hall (London).

Vidakovic, B. (1999). *Statistical modeling by wavelets*, Wiley Series in Probability and Statistics: Applied Probability and Statistics, John Wiley & Sons Inc., New York. A Wiley-Interscience Publication.

Yaglom, A. M. (1987). *Correlation theory of stationary and related random functions. Vol. I*, Springer Series in Statistics, Springer-Verlag, New York. Basic results.