POLITECNICO
MILANO 1863

POLITECNICO DI MILANO
SCHOOL OF INDUSTRIAL AND INFORMATION ENGINEERING
Department of Mathematics
M. Sc. in Mathematical Engineering

Non-parametric mixed effect models for the estimation and analysis of the effect of temperature on householders' monthly electricity demand.

Supervisor: Prof. Simone Vantini

Co-supervisor: Prof. Massimo Tavoni
Dr. Matteo Fontana

Student: Isabella Carioni
Matr. 898937

Academic Year 2018-2019

Contents

List of Figures

Abstract

Climate change is one of the biggest challenges of our time because it directly impacts our economy and society. Temperature increase, caused by the growth in the last decades of greenhouse gases, will cause a major demand of electricity for the use of air conditioning systems that will require more energy potential in the grid with an average predicted growth of 2%. Many studies were performed in the field using linear models and investigating separately the sensitivity of electricity consumption and clients behaviour. This work aims at developing a non parametric model to deal with both mean city monthly temperature electricity response and the identification of common clients behaviour patterns in the population. We propose a functional mixed effect models for the city of Milan, where the fixed effect represents the mean behaviour of the population and the random effects account for the clients' and years' variability. We succeed, studying the mean client's curve, to identify two subgroups in the population that differ for their quadratic or linear trend. After fitting two different models for each subgroup we finally highlight a small number of behaviour's patterns. Given the nature of the data analysed we are able to uniquely identify functions using vertex position and concavity for parables and intercept and slope for lines. This allows us to cluster them using classical statistical tools and reduce complexity.

Sommario

Il cambiamento climatico è una delle più grandi sfide del nostro tempo perchè avrà un impatto diretto sulla nostra economia e società. L'aumeto della temperatura, causato dalla crescita negli ultimi decenni dei gas serra, causerà una maggior domanda di elettricità per l'uso di sistemi di aria condizionata che richiederanno maggior potenziale nella rete con una crescita media prevista del 2%. Molti studi sono stati condotti nel settore utilizzando modelli lineari e analizzando separatamente la sensibilità del consumo elettrico e il comportamento dei clienti. Questo lavoro ha come obiettivo lo sviluppo di un modello non parametrico per studiare sia la risposta media mensile di energia elettrica della città alla temperatura sia l'identificazione di comportamenti comuni dei clienti nella popolazione. Proponiamo un modello funzionale ad effetti misti per la città di Milano. dove l'effetto fisso rappresenta il comportamento medio della popolazione e gli effetti randomici tengono conto della variabilità dei clienti e degli anni. Siamo riusciti, studiando la curva media del cliente, ad identificare due sottogruppi nella popolazione che differiscono per il loro andamento quadratico o lineare. Dopo aver utilizzato due differenti modelli statistici per ogni sottogruppo abbiamo evidenziato un gruppo ristretto di comportamenti comuni. Data la natura dei dati analizzati siamo stati, in grado di identificare univocamente le funzioni tramite la posizione del vertice e la concavità per quanto riguarda le parabole e l'intercetta e il coefficiente angolare per le rette. Questo ci ha permesso di classificarli ugualmente usando i metodi della statistica classsica, riducendo la complessità computazionale.

1. Introduction

Climate change is one of the biggest challenges of our times and we are facing its consequences directly on our lives. Global warming causes more extreme climate events: drought and wild fires started to occur more frequently, storm and hurricane become more destructive, glacier are melting and global mean sea level is rising. It is well known that the human activity, with the increase of greenhouse gasses emissions in the atmosphere, is responsible for the rise of surface temperature that is driven by a myriad of societal factors. It is clear how complicated is the problem and how the economic, social-demographic, energy and climate fields are correlated together for the study of the phenomena. To understand how the climate will impacts on our economy and which policies governments can undertake to reduce global warming, Integrated Assessment Models (IAMs) have been developed, coupling detailed energy system technologies models with simplified economic and climate ones. They use different narratives describing alternative socio-economic advancement, including sustainability, regional rivalry, inequality, fossil-fuelled and middle-of-the-road development to better analyse plausible major global improvement that together would lead in the future to different challenges for mitigation and adaptation to climate change. In this context many studies were elaborated to improve the equations of the mathematical model to better predict future emission and costs. From the analysis performed using scenarios of Integrated Assessment Models is clear that one of the most important change will address the energy system. It has been proven that over the 21th century the surface temperature is projected to rise under all assessed scenarios, causing more frequent hot and fewer cold temperatures extremes over most land areas, and heat waves will occur with a higher frequency and longer duration [9]. This will have a direct impact on the energy system that will have to face the major challenges related to climate change, especially because energy consumption is projected to grow on overage of 2% per year and the 80% of it is still originated by fossil fuels. It is important to note that one third of energy produced is used in the electricity system and is consumed for the 60% in residential and commercial buildings [13]. Electricity consumption will grow with the increasing installation of new air conditioning systems and therefore the power generation grid will require additional resources.

In this context the literature presents different studies on finding the relation between electricity consumption and climatic variables using different approaches. The empirical assessment of the response electricity temperature curve is important for understanding what will be the real impact and cost of the changes that will be necessary to increase the grid potential. Moreover, using the founded relation in Integrated

Assessment Model, it is possible to provide more precise and realistic results in policy evaluations.

- Mukherjee and Roshanak [19] developed a predictive model for residential and commercial electricity usage to understand the relationship between weather, climate and electric power consumption, analysing the state of Florida and testing different non-linear models. They conclude that mean dew point temperature is a more suitable predictor instead degree-days variable.
- Auffhammer et al. [1] used comprehensive high-frequency data at level of load balancing authorities to parametrize the relationship between average or peak electricity demand and temperature across the United State. Their study suggests a significant increase in intensity and frequency of peak events.
- Franco and Sanstad [12] estimated the relationship between temperature and both electricity consumption and peak demand at sample location in California and combined them with global projections to understand the impacts of future temperature change on electricity consumption and peak demand.
- Chen et al. [7] focused on the study of the residential sector analysing the penetration of Air Conditioning systems developing a new classification method with unprecedented spatio-temporal resolution in Los Angeles.

Other studies estimate region-specific predictive models. Climate change has geographically distinct impacts. Doing regional analysis will facilitate assessing the end-use electricity consumption sensitivity because energy consumption are recorded at regional level.

- Christenson et al. [8] investigated the impact of global warming in Switzerland, by means of the degree day methods. They conclude that there will be an increasing in the cooling potential.
- Mirasgedis et al. [18] focused on the potential upcoming impact of climate change on electricity demand at regional/national level for regions where topography and location results in large differences in local climate, to model the sensitivity of electricity demand in Greek power system. The result confirmed an increase of the annual electricity demand in particular during summer, that will lead to the need for increases of the installed capacity.

A study regarding Italy was performed by Bianco et al. [4]. Their objective was to analyse the influence of economic and demographic variables on the annual electricity consumption in order to develop a long term consumption forecasting not considering climate change impacts. The relation between climate change and electricity consumption was studied by Pagliarini et al. [20] who analysed in a five-year period the correlation between daily average outdoor dry bulb temperature and daily electricity consumption. They used a five-parameter estimation approach in order to highlight the effect of both user behaviour and the physical characteristics of building stock.

The aim of this thesis work is to model the overall effect of temperature on electricity consumption in the residential sector and also to identify the different behaviour of single clients to understand common patterns in the consumption. We use a statistical model able to handle the complex hierarchical structure of the data and to consider as a statistical unit the monthly electricity temperature response function for every clients in every year. Consequently we can predict the total increase demand of electricity consumption and simultaneously classify clients with the aim of analysing the presence of air conditioning systems. Furthermore our estimation could be used in the context of Integrated Assessment models to better understand the implication of an increase demand on the electricity grid. We performed our analysis focusing on the city of Milan using a monthly based dataset considering a time interval of 5 years (2015-2019).

This thesis is organized as follows. Chapter 2 contains the fundamental statistical theory used, Chapter 3 explains the procedure that we have followed to construct the final dataset. We present in Chapter 4 the analysis conducted to choose the parameters of the models and the classification of the behaviours of the clients, while in Chapter 5 we draw conclusions summarizing the obtained results.

2. Statistical Methodology

2.1. Functional Data Analysis

Functional Data Analysis or FDA is the branch of statistics that studies complex and high-dimensional data having functional nature (i.e curves, surfaces and images). The basic idea of functional data analysis is to think of observed data as single entities, rather than as a sequence of individual observation. Our functions live in a continuous domain and lie in a functional space. The most common choice is using the L^2 Hilbert space for its good geometric properties. Indeed Hilbert space is a generalization of the concept of Euclidean space to spaces of any dimension, even infinite. In this context we can think our functional data as a point in the space of functions and using the notion of inner product and norm to extend many methods belonging to multivariate statistics.

2.1.1. Basis Function

Functional space are infinite dimensional, therefore we need a strategy for constructing functions with parameters that are easy to estimate and on the other hand we do not want to use more parameters than we need. The solution normally presented in the literature (such as Ramsay and Silverman [22]) is to use a basis function system $\{\phi_1, ..., \phi_k, ...\}$. The functions composing the basis are mathematically independent of each other and have the property that we can approximate arbitrarily well any function by taking a weighted sum or linear combination of a sufficiently large number K of this functions. For example the most familiar basis function system is the collection of monomials used to construct the power series,

$$1, t, t^2, t^3, ..., t^k, ... \tag{2.1}$$

followed by the Fourier expansion

$$1, \sin(\omega t), \cos(\omega t), \sin(2\omega t), \cos(2\omega t)..., \sin(k\omega t), \cos(k\omega t)... \tag{2.2}$$

A function $x(t)$ is represented by the linear expansion

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t) = \mathbf{c}' \boldsymbol{\phi}(t) \tag{2.3}$$

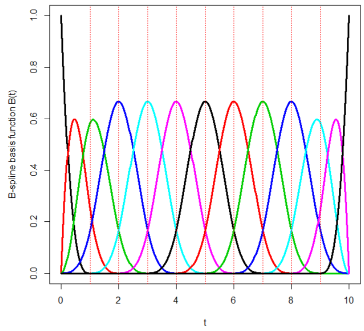

Figure 2.1.: B-spline basis with 13 spline functions of order 4 defined over the interval [0,10] by nine interior boundaries or knots.

The parameters $c_1, c_2, ..., c_k$ are the coefficients of the expansion and K determines the degree to witch the data are smoothed as opposed to interpolated. Basis expansion methods represent the potentially infinite dimensional world of functions within the finite dimensional framework of the vectors c [22]. In Functional Data Analysis is therefore very important to choose the correct basis system.

The most common choice for non-periodic functional data is the *Spline Basis system*. It combines the fast computation of polynomials and greater flexibility achieved with a modest number of basis functions. Splines are piecewise polynomials constructed by dividing the interval of definition T into L subintervals, with boundaries at points called *breakpoints* or *knots*. Over each interval the spline is defined as a polynomial of order m. At each *breakpoint*, neighbouring polynomials are constrained to join smoothly and derivatives up order $m-2$ must also match. Summarizing a spline function is determined by the order of polynomial sequence and the knot sequence τ.

There are several different basis system for constructing spline functions. One of the most popular is the B-spline basis system. Their essential properties are:

- Each $\phi_k(t)$ is itself a spline function of order m and a knot sequence τ
- A linear combination of these basis functions is a spline function
- Any spline function can be expressed as a linear combination of these basis functions

We call $B_k(t, \tau)$ a B-spline basis function in t with sequence of *breakpoints* τ. A spline function is then defined as

$$S(t) = \sum_{k=1}^{m+L-1} c_k B_k(t, \tau) \tag{2.4}$$

and we can see an example in Figure 2.1.

2.1.2. Principal Component Analysis

Once we have our functional data the first method that we turn to after descriptive statistic and plots is Principal Component Analysis. In functional PCA, there is an *eigenfunction* associated with each eigenvalue, rather than an eigenvector. The basic ideas of this procedure were discovered independently by Karnhunen and Loeve [14, 15]. These eigenfunctions describe major variational components.

In multivariate statistics, variation is usually summarized by either a covariance or a correlation matrix. Instead when dealing with functional observations, $x_i(s)$ and $x_i(t)$ have the same origin and scale. Consequently, the estimated *covariance function*

$$v(s,t) = \frac{1}{N-1} \sum_i [x_i(s) - \bar{x}(s)][x_i(t) - \bar{x}(t)] \tag{2.5}$$

or the *cross-product function*

$$c(s,t) = \frac{1}{N} \sum_i x_i(s) x_i(t) \tag{2.6}$$

will tend to be more useful than the correlation function

$$r(s,t) = \frac{v(s,t)}{\sqrt{[v(s,s)v(t,t)]}}. \tag{2.7}$$

We can define PCA as the search of a *probe* ξ that provide a *prob score*, defined as

$$\rho_\xi(x_i) = \int \xi(t) x_i(t) dt, \tag{2.8}$$

with the largest possible variation imposing $\int \xi^2(t)dt = 1$. The probe score variance $Var\left[\int \xi(t)(x_i(t) - \bar{x}(t))^2 dt\right]$ associated with a probe weight ξ is the value of

$$\mu = \max_\xi \left\{ \sum_i \rho_\xi^2(x_i) \right\} \quad \text{subject to} \quad \int \xi^2(t)dt = 1. \tag{2.9}$$

μ and ξ are referred to as the largest *eigenvalue* and *eigenfunction* respectively.

As in multivariate PCA a non increasing sequence of eigenvalues $\mu_1 \geq \mu_2 \geq ... \mu_k$ can be constructed stepwise by requiring each new eigenfunction, computed in step *l*, to be orthogonal to those computed on previous steps,

$$\int \xi_j(t)\xi_l(t)dt = 0 \quad \forall j < l \quad \text{and} \int \xi_l^2(t)dt = 1. \tag{2.10}$$

We can compute eigenfunction ξ_j of the bivariate covariance function $\nu(s,t)$ as solution of the functional eigenequation

$$\int \nu(s,t)\xi_j(t)dt = \mu_j \xi_j(s). \tag{2.11}$$

After computing the pairs (μ_j, ξ_j), we choose $1 \leq l \leq N-1$,using visual inspection of μ_j, to define a basis system for approximating our sample functions x_i. These basis function are referred to *orthonormal* basis and are the most efficient in the sense that the total error sum of squares

$$\mathrm{PCASSE} = \sum_i^N \int [x_i(t) - \bar{x}(t) - c_i' \xi(t)]^2 dt \tag{2.12}$$

is the minimum achievable with only *l* basis functions.

The coefficient vectors c_i i=1,..N contain the *principal component scores* c_{ij} that define the optimal fit to each function x_i:

$$c_{ij} = \rho_\xi(x_i - \bar{x}) = \int \xi_j(t)[x_i(t) - \bar{x}]dt \tag{2.13}$$

They can be useful in interpreting the nature of the variation identified by the PCA and is common practice to use these scores as "data" to be subjected to a more conventional multivariate analysis (Ramsay and Silverman [22]).

2.1.3. Data Alignment and Clustering

One of the major problems that we can encountered in functional data analysis is the misalignment of the data. Functions can vary in both phase and amplitude, as illustrated in Figure 2.2 . Phase variation is illustrated in the bottom panel, opposed to amplitude variation, shown in the top panel. Curve registration is useful if we want to correctly estimate the mean curve. In the bottom panel of the figure we can see that the dashed curves, does not resemble any other curve. The need to register the curves by transforming their argument is motivated, in Ramsay and Silverman [22], by the fact that physical time may not be directly relevant to the dynamic of many real-life systems and there could be a sort of biological time scale that can vary from case to case. The problem of curve misalignment become more important if we want also to perform clustering.

A solution to this problem is *K-mean alignment* proposed by Sangalli et al. [25]. The authors describe a procedure that is able to efficiently cluster and align a set of curves in *k* groups. If the number of clusters is set equal to 1, the algorithm implements

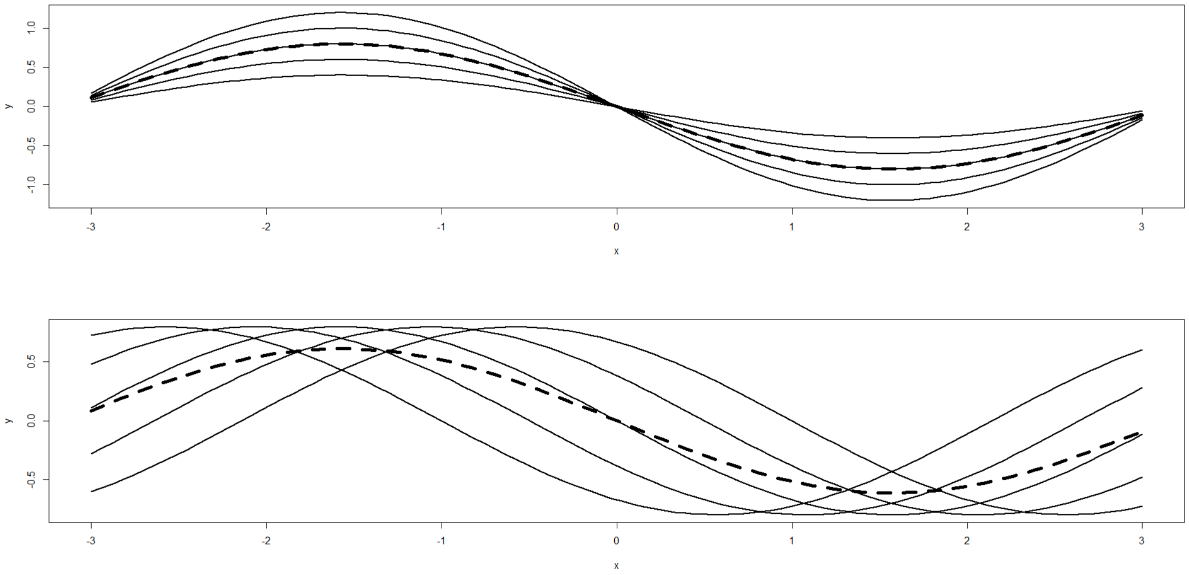

Figure 2.2.: The top panel shows five curves varying only in the amplitude. The bottom panel shows five curves varying only in phase. The dashed line in each panel indicates the mean of the five curves.

the *Procrustes aligning procedure*, whereas if no alignment is allowed, it implements a functional k-mean clustering of curves.

Let us consider a set C of curves $\mathbf{c}(s)$. Aligning $\mathbf{c}_1, \mathbf{c}_2 \in C$ means finding a *warping function* $h(s)$, such that the two curves $\mathbf{c}_1 \circ h$ and $\mathbf{c}_2$ are most similar. We need to specify a similarity index $\rho(.,.) : C \times C \to \mathbb{R}$ and a class W of *warping functions* h, such that $\mathbf{c} \circ h \in C, \forall \mathbf{c} \in C$ and $\forall h \in W$. To align $\mathbf{c}_1$ to $\mathbf{c}_2$, according to (ρ, W), means finding $h^\star \in W$ that maximizes $\rho(\mathbf{c}_1 \circ h, \mathbf{c}_2)$. The choice of (ρ, W) will define what is meant by phase and amplitude variability. One possible option is to use

$$\rho(\mathbf{c}_1, \mathbf{c}_2) = \frac{1}{d} \sum_{p=1}^{d} \frac{\int_{\mathbb{R}} c'_{1p}(s) c'_{2p}(s) ds}{\sqrt{\int_{\mathbb{R}} c'_1 p(s)^2 ds} \sqrt{\int_{\mathbb{R}} c'_{2p}(s)^2}}, \tag{2.14}$$

$$W = \{h : h(s) = ms + q \quad with \quad m \in \mathbb{R}^+, q \in \mathbb{R}\} \tag{2.15}$$

as in **K-mean**.

The couple defined in Equations (2.14) and (2.15) satisfies the following properties:

1. ρ is bounded,with maximum value equal to 1. Moreover ρ is :
   - reflexive: $\rho(\mathbf{c}, \mathbf{c}) = 1 \quad \forall \mathbf{c} \in C$;
   - symmetric:$\rho(\mathbf{c}_1, \mathbf{c}_2) = \rho(\mathbf{c}_2, \mathbf{c}_1) \quad \forall \mathbf{c}_1, \mathbf{c}_2 \in C$;

- transitive: $[\rho(\mathbf{c}_1, \mathbf{c}_2) = 1, \rho(\mathbf{c}_2, \mathbf{c}_3) = 1] \Rightarrow \rho(\mathbf{c}_1, \mathbf{c}_3) = 1, \quad \forall \mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3 \in C$.

2. W is a convex vector space and has a group structures with respect to the operator of function composition $\circ$.

3. ρ and W are consistent: $\rho(c_1, c_2) = \rho(\mathbf{c}_1 \circ h, \mathbf{c}_2 \circ h) \forall h \in W$.
This implies that is not possible to obtain fictitious increment of similarity between two curves simply by warping them simultaneously.

4. Similarity that can be obtained by align $\mathbf{c}_1$ to $\mathbf{c}_2$ is the same as the one that can be obtained by aligning $\mathbf{c}_2$ to $\mathbf{c}_1$:
$\rho(\mathbf{c}_1 \circ h_1, \mathbf{c}_2 \circ h_2) = \rho(\mathbf{c}_1 \circ h_1 \circ h_2^{-1}, \mathbf{c}_2) = \rho(\mathbf{c}_1, \mathbf{c}_2 \circ h_2 \circ h_1^{-1}) \; \forall h_1, h_2 \in W$

5. The similarity index between two curves is unaffected by strictly increasing affine transformations of one or more components of the curves:
Let W^d be the set of transformation $\mathbf{r} : \mathbb{R}^d \to \mathbb{R}^d$ such that: $\mathbf{x} \in \mathbb{R}^d \to \mathbf{r}(\mathbf{x}) \in \mathbb{R}^d$:
$\rho(\mathbf{r}_1(\mathbf{c}_1), \mathbf{r}_2(\mathbf{c}_2)) = \rho(\mathbf{c}_1, \mathbf{c}_2) \; \forall \mathbf{r}_1, \mathbf{r}_2 \in W^d$

Once (ρ, W) are defined we can proceed with Procrustes aligning procedure described in Sangalli et al. [24]. The algorithm perform the following steps:

1. *Expectation step*:
The reference curve is estimated using all the curves obtained at the previous iteration. A new reference curve is obtained.

2. *Maximization step*:
Each curve is shifted and dilated in order to maximize its similarity with the estimated reference curve. New curves are obtained.

The warping functions h_i are given by the comparison of the optimal warping function found at each iteration: $h_i = h_{iterK} \circ ... \circ h_{iter2} \circ h_{iter1}$. The registered centerline is then defined as $\tilde{\mathbf{c}}_i = \mathbf{c}_i \circ h_i^{-1}$

Now we consider the problem of clustering a set of N curves $\{\mathbf{c}_1, ..., \mathbf{c}_N\}$ with respect of k unknown templates $\underline{\boldsymbol{\varphi}} = \{\boldsymbol{\varphi}_1, ..., \boldsymbol{\varphi}_k\}$. What we have to do is to solve the following optimization problem:

(i) find$\underline{\boldsymbol{\varphi}} = \{\boldsymbol{\varphi}_1, ..., \boldsymbol{\varphi}_k\} \subset C$ and $\underline{\mathbf{h}} = \{h_1, .., h_N\} \subset W$ such that

$$\frac{1}{N} \sum_{i=1}^{N} \rho(\boldsymbol{\varphi}_{\lambda(\underline{\boldsymbol{\varphi}}, \mathbf{c}_i)}, \mathbf{c}_i \circ h_i) \geq \frac{1}{N} \sum_{i=1}^{N} \rho(\boldsymbol{\psi}_{\lambda(\underline{\boldsymbol{\psi}}, \mathbf{c}_i)}, \mathbf{c}_i \circ g_i) \qquad (2.16)$$
$$\forall \underline{\boldsymbol{\psi}} = \{\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_k\} \neq \underline{\boldsymbol{\varphi}}, \quad \forall \underline{\mathbf{g}} = \{g_1, .., g_N\} \neq \underline{\mathbf{h}}$$

where

- $\lambda(\underline{\boldsymbol{\varphi}}, \mathbf{c}) = \min\{r : \mathbf{c} \in \delta_r(\underline{\boldsymbol{\varphi}})\}$ is a labelling function,

- $\delta_r(\underline{\boldsymbol{\varphi}}) = \{\mathbf{c} \in C : \sup_{h \in W} \rho(\boldsymbol{\varphi}_j, \mathbf{c} \circ h) \geq \sup_{h \in W} \rho(\boldsymbol{\varphi}_r, \mathbf{c} \circ h), \forall r \neq j\} \quad j = 1, ..k$

is the domain of attraction.

(ii) assign to $\mathbf{c}_i$ to the cluster $\lambda(\underline{\varphi}, \mathbf{c}_i)$ and align it to the corresponding template $\varphi_{\lambda(\underline{\psi}, \mathbf{c}_i)}$ using the warping function h_i

Unfortunately (i) cannot be solved analytically.
K-mean propose to simultaneously deal with (i) and (ii) via k-mean alignment algorithm that iteratively alternates the following step:

- *templates identification step*:
  estimation of the set of k templates associated to the k clusters.
  Ideally the template $\varphi_{j[q]}$ at the iteration q should be estimated as the curve $\varphi \in C$ that maximize the total similarity:

$$\sum_{i:\lambda(\underline{\varphi}_{[q-1]}, \mathbf{c}_{i[q-1]})=1} \rho(\varphi, \mathbf{c}_{i[q-1]}) \tag{2.17}$$

- *assignment and alignment step*:
  The N curves $\{\mathbf{c}_{1[q-1]}, ..., \mathbf{c}_{N[q-1]}\}$ are clustered and assigned to the set of the k templates obtained in the previous step. More precisely, the i-th curve $\mathbf{c}_{i[q-1]}$ is aligned to $\varphi_{\lambda(\underline{\psi}[q], \mathbf{c}_{i[q-1]]})}$ and $\tilde{\mathbf{c}}_{i[q]} = \mathbf{c}_{i[q-1]} \circ h_{i[q]}$ is assigned to the cluster $\lambda(\underline{\psi}[q], \tilde{\mathbf{c}}_{i[q]]})$

- *Normalization step*:
  For $j = 1, ..k$, all the curves $\tilde{\mathbf{c}}_{i[q]]}$ assigned to cluster j are wrapped along $(\bar{h}_{j[q]})^{-}1$, where

$$\bar{h}_{j[q]} = \frac{1}{N_{j[q]}} \sum_{i:\lambda(\underline{\varphi}_{[q]}, \tilde{\mathbf{c}}_{i[q]})=1} h_{i[q]}. \tag{2.18}$$

  The normalization step is used to select the solution to the optimization problem that leaves the average location of the cluster unchanged.

2.1.4. Functional Linear Regression

After building our functional data object, we can use them to model predictive relationship. In classical linear regression,predictive models are often in the form

$$y_i = \sum_{j=0}^{p} x_{ij}\beta_j + \epsilon_i, \quad i = 1, ..N \tag{2.19}$$

where y_i is the response variable, x_{ij} the covariates and ϵ_i measurement error.

If the vector of covariate observation $x_i = (xi1, ..x_{ip})$ is replaced by a function $x_i(t)$ the first idea is to discretize them by choosing a set of times $t_1, ...t_q$ and fitting the model

$$y_i = \alpha_0 + \sum_{j=0}^{q} x_i(t_j)\beta_j + \epsilon_i, \tag{2.20}$$

that choosing a finer mesh of times will approximate the integral equation

$$y_i = \alpha_0 + \int x_i(t)\beta(t) + \epsilon_i. \tag{2.21}$$

To determine the infinite-dimensional $\beta(t)$ we can redefine the problem using a basis coefficient expansion for β, and for $x_i(t)$:

$$\begin{aligned} \beta(t) &= \sum_{k}^{K_\beta} b_k \phi_k(t) = \mathbf{b}'\boldsymbol{\phi}(t), \\ x_i(t) &= \sum_{k}^{K_x} c_{ik}\psi_k(t), \quad x(t) = \mathbf{C}'\boldsymbol{\psi}\mathbf{(t)}. \end{aligned} \tag{2.22}$$

The model can be expressed as

$$\hat{y}_i = \int \mathbf{C}\psi(t)\phi(t)'\mathbf{b} \tag{2.23}$$

We can further simplify notation by defining $(K_\beta + 1)$-vector $\zeta = (\alpha, b_1, ..b_k)'$ and the coefficient matrix ζ to be $Nx(K_\beta + 1)$. Then the model become simply:

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{Z}\hat{\zeta} \\ \mathbf{Z}'\mathbf{Z}\hat{\zeta} &= \mathbf{Z}'\mathbf{y} \end{aligned} \tag{2.24}$$

There are also cases where the interest lies in the prediction of functional response

$$y_i(t) = \beta_o(t) + \sum_{j=1}^{K} x_{ij}\beta_j(t) + \epsilon_i(t) \tag{2.25}$$

where $xi1, ..., x1K$ are known scalar covariates. To estimate $\hat{\boldsymbol{\beta}}$ we need to minimize

$$\sum_{i=1}^{N} \int \left(y_i(t) - \beta_o(t) + \sum_{j=1}^{K} x_{ij}\beta_j(t) \right)^2 dt. \tag{2.26}$$

If there are no particular restriction on the way $\beta(t)$ varies we can minimize Equation (2.26) individually for each t. We calculate $\beta(t)$ for a suitable grid of values of t using ordinary regression analysis, and then interpolate between these values.

2.2. Linear Mixed Effect Model

Mixed-effects models are an extension of linear models and are particularly useful when there is no independence in the observations caused by the hierarchical structure of the data. Linear models hypothesis is that data are characterized by independent observations with an homogeneous variance. In a linear model the distribution of $\mathcal{Y}$ is multivariate normal,

$$(\mathcal{Y}|\mathcal{B} = \mathbf{b}) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{o}, \sigma^2\mathbf{W}^{-1}) \tag{2.27}$$

where

- n is the dimension of the response vector,
- $\mathbf{W}$ is a diagonal matrix of known prior weights,
- $\boldsymbol{\beta}$ is a p-dimensional coefficient vector,
- $\mathbf{X}$ is an $n \times p$ model matrix,
- $\mathbf{o}$ is a vector of prior offset therm.

The parameters of the model are coefficients $\boldsymbol{\beta}$ and the scale parameter σ.

To deal with hierarchical data, a simple approach is to aggregate, so rather than using a single observation that is not independent we study the mean for each level of a factor. In this case data are simply averaged and we run a model with a reduced number of observations. Another approach is to divide the data by factor and analysing one unite at a time, running for each a linear model. However, in doing so we don't take advantage of the information in data from other levels. This can lead also to a poor prediction caused by small amount of data.

In this context linear mixed models are in between the previous approaches. They are able to describe relationships between a response variable and some covariates in data that are grouped according one or more classification factors (i.e. longitudinal data, repeated measured data, multilevel data). By associating common random effects to observation sharing the same level of classification factor, mixed-effects models flexibility represents the covariance structure induced by the grouping of the data.

Linear mixed effect models are defined by the distribution of two vector-valued random variables: $\mathcal{Y}$, the response and $\mathcal{B}$, the vector of random samples [2]. The conditional distribution of $\mathcal{Y}$ given $\mathcal{B} = \mathbf{b}$ has the following form:

$$(\mathcal{Y}|\mathcal{B} = \mathbf{b}) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{o}, \sigma^2\mathbf{W}^{-1}) \tag{2.28}$$

where $\mathbf{Z}$ is the $n \times q$ model matrix for the q-dimensional vector-valued random-effect variable, $\mathcal{B}$, whose value we are fixing at $\mathbf{b}$.

The unconditional distribution of $\mathcal{B}$ is also multivariate normal with mean 0 and a parametrized $q \times q$ variance covariance matrix $\mathbf{\Sigma}$

$$\mathcal{B} \sim \mathcal{N}(0, \mathbf{\Sigma}). \tag{2.29}$$

It is convenient to express the model in terms of *relative covariance factor*, $\mathbf{\Lambda}_\theta$, which is a $q \times q$ matrix, depending on the *variance-component parameter*, θ, and generating the symmetric $q \times q$ variance covariance matrix, $\mathbf{\Sigma}$, according to

$$\mathbf{\Sigma}_\theta = \sigma^2 \mathbf{\Lambda}_\theta \mathbf{\Lambda}_\theta^T. \tag{2.30}$$

The parameters in these models are typically estimated by maximum likelihood or restricted maximum likelihood. In general there is no closed-form solution and they must be determined by iterative algorithm, for example involving a repeated applications of the Penalized Least Squared method proposed by Batesa and DebRoyb [3].

2.2.1. Nonparametric Mixed Effect Models for Functional Data

In the current days with the increasing interest in functional analysis many works develop methodologies and applications to extend linear effect models to this framework, like Rice and Wu [23], Edwards et al. [11], and LoMauro et al. [16].

A typical parametric mixed effects analysis of this type represents each subject's repeated measures as the sum of a population mean function depending on time and other covariates, a low-degree polynomial with random coefficients, and measurement error.

Rice and Wu [23] propose a methodology that is applicable when the curves are sampled at variable and irregularly spaced points. Let there be m subjects, n_i observations at times $0 \leq t_{ij} \leq T$ on the i-th subject, and $n = \sum_{i=1}^{m} n_i$ observations overall. Let $Y_{ij} = Y_i(t_{ij})$ be the outcome measured on the i-th subject at time t_{ij}. The mean function and the random function are approximated non-parametrically with splines

$$\mathbb{E}(Y_i(t)) = \mu(t) = \sum_{K=1}^{p} \beta_k \bar{\phi}_k(t) \tag{2.31}$$

where $\{\bar{\phi}_k(\cdot)\}$ is a basis for spline function on $[0, T]$. The random effect curve for the i-th subject is similarly modelled as $\sum_{k=1}^{q} b_{ik} \phi_k(t_{ij})$. In this case $\{\phi_k(\cdot)\}$ is a basis for a possibly different space of spline functions on $[0, T]$ and b_{ij} are random coefficient with mean zero and covariance matrix $\mathbf{\Sigma}$. Incorporating also the uncorrelated measurement error ϵ_{ij} with mean zero and variance σ^2, we finally obtain the following model

$$Y_{ij} = \mu(t) = \sum_{K=1}^{p} \beta_k \bar{\phi}_k(t) + \sum_{k=1}^{q} b_{ik} \phi_k(t_{ij}) + \epsilon_{ij}. \tag{2.32}$$

The covariance structure is modelled through the b_{ik} and the covariance kernel for a random curve $Y(t)$ is computed as

$$cov(Y(s), Y(t)) = \sum_{k=1}^{q} \sum_{l=1}^{q} \Sigma_{kl} \phi_l(t) + \sigma^2 \delta(s - t), \tag{2.33}$$

where $\delta(\cdot)$ is the Dirac delta function.

Conditioning Equation (2.32) on p and q we obtain the classical mixed effect model, and the vector of observations can be expressed as

$$Y_i = X_i \beta + Z_i b_i + \epsilon_i. \tag{2.34}$$

In this way the estimate of β and σ^2 can be computed using the method developed for mixed effect models.

3. Data

In this chapter we present the set of data we used to performed our analysis. With the aim of studying the monthly relationship between average electricity consumption and climatic variable in cities of Italy we used two datasets one regarding electricity consumption and one on meteorological data.

The consumption dataset is part of a behavioural energy efficiency campaign conducted by Bonan et al. [5] and Bonan et al. [6], that provides customers from a European electric utility with information on their energy use. The dataset we used do not consider only the subsample receiving by e-mail the Home Energy Report (eHER) as in their study, but a 5% random sample of the whole dataset relative to 8048 municipalities of Italy.

Several variables that we take in consideration characterize the householders, namely: id of the electricity supply contract, consumer region and municipalities uniquely identifiable by the ISTAT code. Regarding instead the field of the experiment we considered the average daily electricity consumption in the month [kWh/day], calculated by taking into account the specific monthly duration in days and an aggregated variables identifying the relative month and year of the observation. Specifically we analyse data from January 2015 to December 2019. To join correctly the two dataset we disaggregate the month variable in two separate ones: Month and Year.

We can already understand the complex structure of the data. We have repeated observation of the clients in the different years and month that highlights the hierarchical structure of the data and their natural functional behaviour.

The meteorological data come from E-obs dataset from the EU-FP6 project UERRA (http://www.uerra.eu) and the Copernicus Climate Change Service, and the data providers in the ECA&D project (https://www.ecad.eu) [10].

E-Obs comes as an ensemble dataset available on a 0.1 and 0.25 degree regular grid starting from 01-01-1950 to nowadays on daily basis regarding the following variables: daily mean temperature [°C], daily minimum temperature [°C],daily maximum temperature [°C], daily precipitation sum [mm], daily averaged sea level pressure [hPa] and daily mean global radiation [W/m2].

Each variable was stored separately in NetCDF-4 format and they cover the area: 25N-71.5N x 25W-45E.

The ensemble datasets is based on surface in-situ observations, collected by ground-based observation networks, owned and operated by the National Meteorological Services and is constructed through a conditional simulation procedure. As explained in Cornes et al. [10], they produced for each of the members a spatially correlated random field using a pre-calculated spatial correlation function. They calculated the mean across the members and provided it as the "best-guess" fields, using for global radiation 10-member ensemble, while for the other elements a 100-member ensemble.

We choose E-obs dataset because we could have an homogeneous data on over the Italian peninsula for all the considered interval of time. Moreover E-obs contains the principal predictors for estimating the variation of electricity consumption, first of all temperature, the most important variable to explain climate change. This can be also confirmed in the previous literature [8, 18, 17, 19, 20, 12]. Mukherjee and Roshanak [19] found that the most important predictor was mean dew point temperature followed by precipitation, in Pagliarini et al. [20] dry bulb temperature was the most correlated weather variable to electricity use, followed by solar irradiance.

In this work are considered data from 1 January 2015 to 31 December 2019 using the 0.1 degree regular grid and we consider only the ensemble mean.

To join the weather dataset with the one of electricity consumption we have followed these steps:

1. for every elements of the ensembled data:
   - change the form from 3-dimensional array to list of dataframe(long, lat, variable)
   - extract grid point related to Italy
2. downscale data to every city of Italy present in the consumption dataset
   - Starting from the shape file downloaded by Istat(Istituto Nazionale di Statistica - www.istat.it), we computed the centroids for each city,
   - Using KNN algorithm we calculate the value relative to the city, as the mean of the four nearest gridded points,
3. Finally we aggregate the values monthly using the same variable as Mukherjee and Roshanak [19].

All this procedure has been done using R software, in particular we used *ncdf4* library to read NetCDF format.

In the first step data for each climatic variable were stored in a 3-dimensional array, one for each that represent longitude, latitude and time (days). In order to improve the accessibility of data, we decided to transform the original 3-dimensional array structure into a list of Dataframes. The slicing was performed with respect to the time

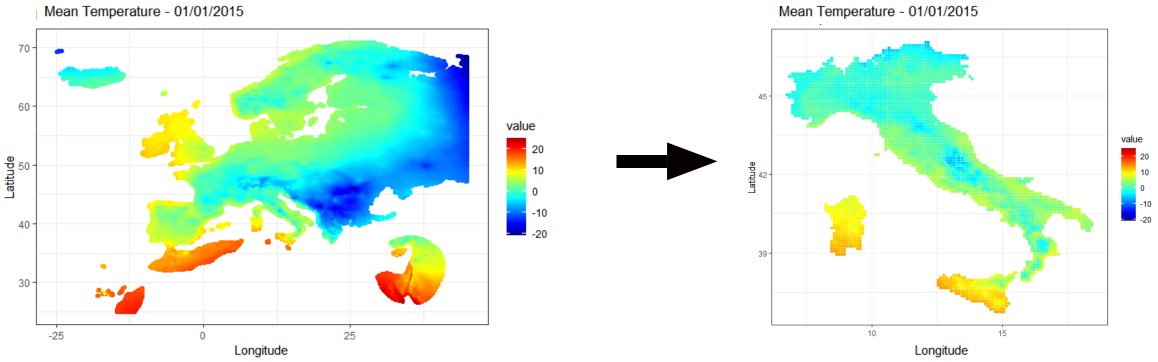

Figure 3.1.: Selection of grid point relative to Italy, tg variable 01/01/2015. On the left Europe data, on the right extracted Italian data.

component. Each observation was characterized uniquely by the couple longitude-latitude, so we were able to construct the Dataframes using the previous variables and another one containing the corresponding value. Subsequently, the shapefile of Italy was used to select the grid point of interest, the one that fall inside the boundaries, to perform the subsequent computations. In this way we obtained for every variable a list of dataframes, one for each day containing the grid point related to Italy,as we can see in Figure 3.1.

The goal of the second step was obtaining the values for each weather variable relatives to the municipalities contained in the consumption dataset. We took the shapefile for every municipalities and we computed the centroid. For each day and for each weather variable we used K-Nearest-Neighbors algorithm. We used the euclidean distance to determine the K nearest neighbours:

$$d(p_i, p_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{3.1}$$

where p_i, p_j are two observation at coordinates (x_i, y_i) and (x_j, y_j). The values of a new observation was estimated as the mean of the k nearest observations.

$$\hat{t}_i = \frac{1}{K}\sum_{j=1}^{K} \tau_j, \quad i = 1, ..N \tag{3.2}$$

where $\hat{t}_i$ is the estimated value of a point in the space where we do not have any observation and $\{\tau_1, .., \tau_K\}$ are the values of the K nearest observations.

After computing leave-one-out cross validation using tg data relative to 1 January 2015 and testing different values for k, we obtained the minimum Predicted Sum of Square (PRESS) using $k = 4$. We can see the result in Figure 3.2. For every day we predicted values for all weather variable, we joined them and add a column relative to Date, obtaining a single dataset.

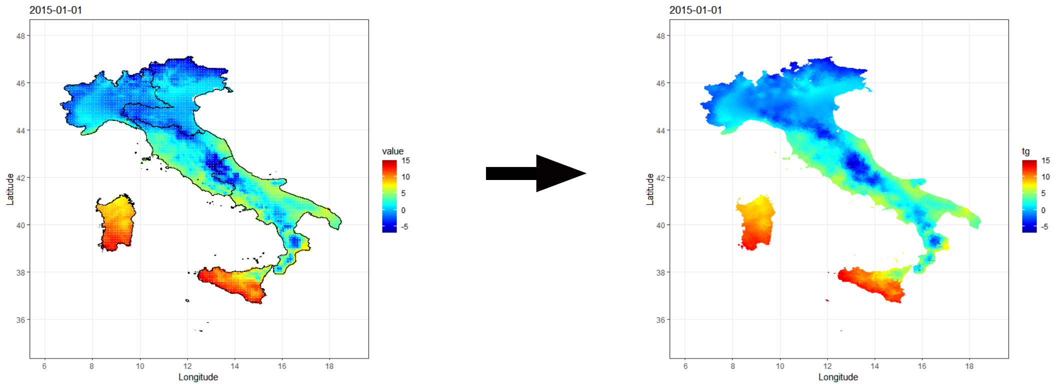

Figure 3.2.: Comparison between grid dataset (left) and municipalities prediction using knn (right), tg variable 01/01/2015.

In the third step we aggregated the meteorological data monthly, computing the same variable as Mukherjee and Roshanak [19]. The most important weather variable computed are: monthly mean temperature [°C], monthly mean maximum temperature [°C],monthly mean minimum temperature [°C], total precipitation in a month [mm], monthly mean pressure [hPa], monthly mean radiation [W/m2] and monthly degree days [°C] computed as $\sum \max(0, 20 - T_e)$ following the Italian normative). For observation identification the data also include the name of the municipality, uniquely identified by ISTAT code, and relative Year and month of the observation.

Finally, we joined the two datasets using the ISTAT code, Year and Month as binding variables.

The variable names of all variables and their description can be found in Appendix A.

4. Model

In this chapter we present our analysis on the previously created dataset with the goal of estimating and analyse the effect of temperature on householders monthly electricity demand. We concentrate our study on 5% of the Italian dataset and we construct our model on the city of Milan.

We started our analysis with a data cleaning process of the Italian dataset. What we wanted to analyse was a population that can represent the real behaviour of electricity consumption in the residential sector. For this reason we removed all the observations below 1 kWh, in this way we considered only clients that effectively were in their house during that month. 1kWh was chosen as a threshold because, as we can found in Raj et al. [21], it approximates the consumption of a refrigerator or standby lights. We also did not consider higher values of average day consumption, because they do not represent a typical residential behaviour. Unfortunately we did not have any information about the family unit, so we decided to rank the observations and to erase the 3000 higher ones, leaving the consumption relative to January, February, July and August. These ones are respectively the coldest and hottest months, with extreme temperature that justify an higher consumption. Finally we kept only the clients with at least four observations. We can see in Figure 4.1 the results of our clean up process. The 3000 higher observations, red in the left image, are a smaller cloud with respect to the other observations. In the right image instead are plotted the final data. We can note that the consumptions relative to January, February, July and August that are effectively the one with higher consumptions.

The original dataset contained 100069 clients and 6332 municipalities and after the cleanup process we ended up with 89942 clients and 5919 municipalities. In Figure 4.2 we can visualize the before and after relative to the city of Milan, the one that we analysed in this work. We started with 1215 clients and end up with 1136.

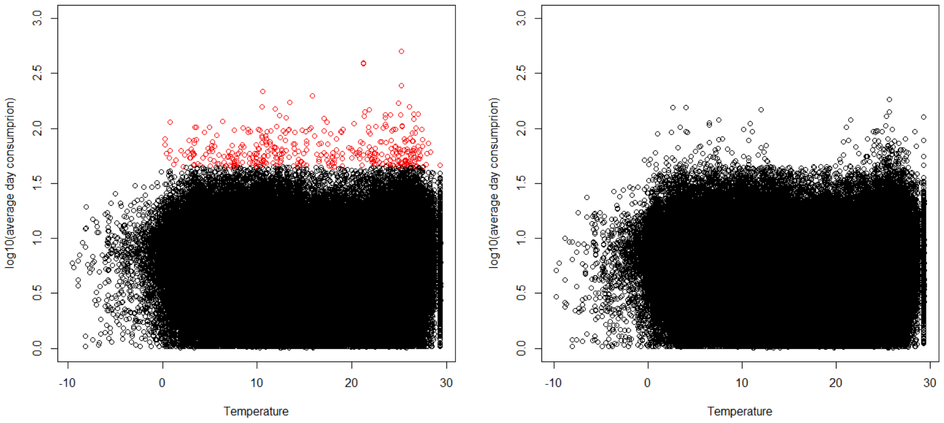

Figure 4.1.: Comparison of Italian dataset before (left) and after (right) cleanup procedure.

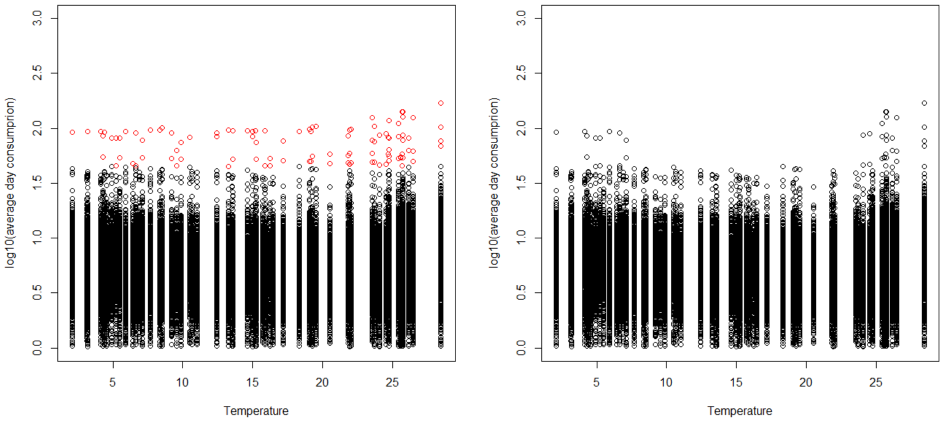

Figure 4.2.: Comparison of Milan dataset before (left) and after (right) cleanup procedure.

4.1. First Approach

The aim of this work was to study the mean monthly electricity temperature response curve relative to the city of Milan and also identifying the different behaviour of single clients to discover common consumption trend. The complex hierarchical structure of the dataset led us to choose a non-parametric mixed effect model, used LoMauro et al. [16], considering the curve as a statistical unit.

The equation of the model is the following:

$$y_{YC}(t) = \sum_{i=1}^{k} \beta_i \phi_i(t) + \sum_{i=1}^{k} B_{Yi} \phi_i(t) + \sum_{i=1}^{k} b_{Ci} \phi_i(t) + \epsilon_{YC}(t) \tag{4.1}$$

where:

- $y_{YC}(t)$ is the datum that one would have recorded if the client C in the year $Y \in (2015, 2019)$ were measured at temperature $t \in (2.127, 28.435)$,
- $\{\phi_1(t), ..., \phi_i(t)\}$ is a basis of spline functions,
- $\sum_{i=1}^{k} \beta_i \phi_i(t)$ indicates the Milan mean curve,
- $\sum_{i=1}^{k} B_{Yi} \phi_i(t)$ is the correction relative to the specific year,
- $\sum_{i=1}^{k} b_{Ci} \phi_i(t)$ is the correction for the specific client,
- $\epsilon_{YC}(t)$ indicates the specific observation measurement error,
- $b_{Ci} \sim \mathcal{N}(0, \sigma_i^2)$, allowing a different variance σ_i^2 for each natural cubic spline,
- $\epsilon_{YC}(t) \sim \mathcal{N}(0, \sigma^2)$ for each client C in the year Y for every time t.

The proposed model was implemented in R with the package *lme4* [2]. lme4 package was developed to compute linear mixed effect models, to use it to compute our non parametric mixed effect model we followed this procedure:

i. We computed the basis spline using temperature data and setting the parameter k, the number of basis functions and d, the degree of the polynomial and evaluated them in each given observation.

ii. We constructed a new dataset composed by: the evaluation of the splines, the logarithmic transformation (base 10) of the average monthly consumption variable to meet the hypothesis on the model coefficients,Year and Clients ID.

iii. We used the new dataset with the lme4 package, using the following formula: $\log_{10}$(avg_day_month_consum) $\sim$ $0{+}splines{+}(0{+}splines|YEAR){+}(0{+}splines|ID)$

The best k and d were chosen computing the Akaike Information Criterion (AIC), for

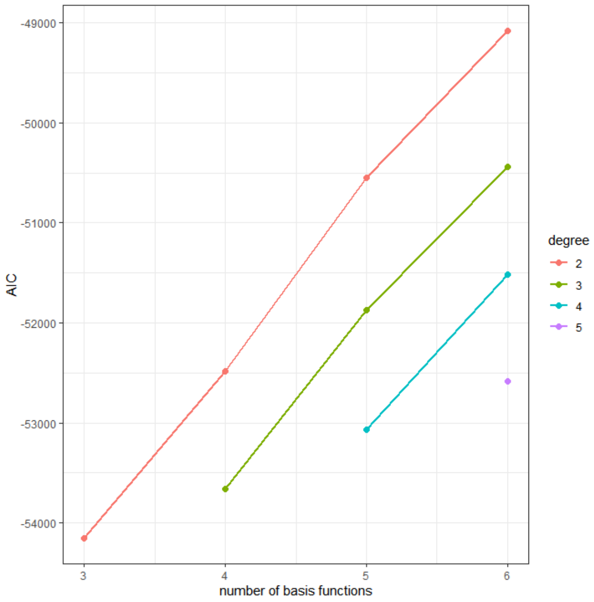

Figure 4.3.: Comparison of AIC for different basis of spline.

each of the following combinations:

- $d = 2, \ k = 3$
- $d = 2, \ k = 4$
- $d = 2, \ k = 5$
- $d = 2, \ k = 6$
- $d = 3, \ k = 4$
- $d = 3, \ k = 5$
- $d = 3, \ k = 6$
- $d = 4, \ k = 5$
- $d = 4, \ k = 6$
- $d = 5, \ k = 6$

We can see in Figure 4.3 a graphic comparison of the models.

The parameters we chose for the model ($d = 2$, $k = 3$) were the ones that minimized the AIC, moreover they confirmed the climatic theory that identifies a quadratic relation between electricity consumption and temperature.

We can see in Table 4.1 the summary of the model. We note that the variability of the annual random effect is less than the variability of the clients groups.

To test the significance of the random effect we performed Likelihood Ratio Test on all the possible combinations of the following model:

- Linear Model: $\log_{10}$(avg_day_month_consum) $\sim 0 + splines$

- Fixed + YEAR: $\log_{10}$(avg_day_month_consum) $\sim 0 + splines + (0 + splines|\text{YEAR})$

- Fixed + ID: $\log_{10}$(avg_day_month_consum) $\sim 0 + splines + (0 + splines|\text{ID})$

Table 4.1.: Model 1: Summary model.

Random effects

Groups	**Name**	**Variance**	**Std. Dev.**
ID	**X3**	0.0826314	0.28746
ID.1	**X2**	0.0872961	0.29546
ID.2	**X1**	0.0637489	0.25249
YEAR	**X3**	0.0011025	0.03320
YEAR.1	**X2**	0.0010647	0.03263
YEAR.2	**X1**	0.0008337	0.02887
Residuals		0.0134927	0.11616

Fixed effects

	Estimate	**Std. Error**	**t value**
X3	0.74637	0.01510	49.42
X2	0.62786	0.01734	36.20
X1	0.69386	0.01727	40.18

Table 4.2.: Model 1: Likelihood Ratio test for random effect significance.

	Linear Model	**Fixed + YEAR**	**Fixed + ID**	**Complete Model**
LogLikelihood	-3260.02	-3175.9	26829	27088

	LM vs YEAR	**LM vs ID**	**LM vs Compl**	**YEAR vs Compl**	**ID vs Compl**
LR	168.2789	60177.46	60696.08	60528	518.63
p-val	1.488066e-36	0	0	0	0

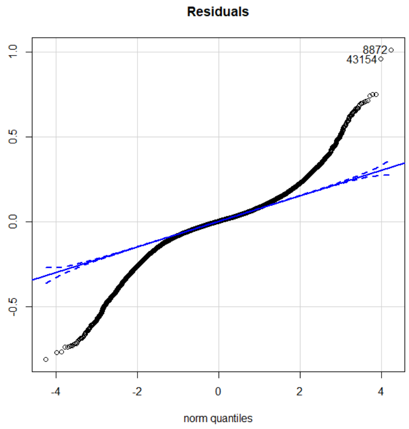

Figure 4.4.: Model 1: Normality plot for the model's residuals.

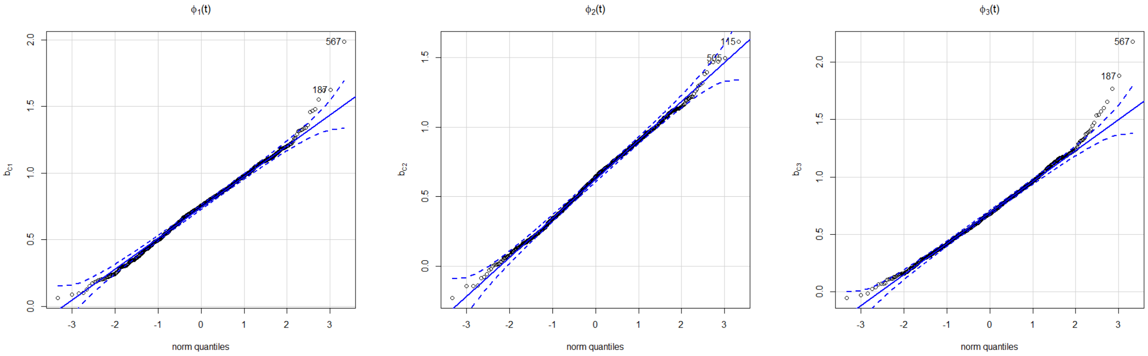

Figure 4.5.: Model 1: Normality plot for coefficients relative to Clients.

- Complete Model: $\log_{10}$(avg_day_month_consum) $\sim 0+\mathtt{splines}+(0+\mathtt{splines}|\mathrm{YEAR})+(0+\mathtt{splines}|\mathrm{ID})$

Results summarized in Table 4.2 confirmed that the more complex model is always better.

Than we verified the hypotesis on the coefficients and residual of the model. The residuals [Figure 4.4] do not properly follow a Gaussian distribution but their distribution is symmetric and with lighter tails, so we can be reasonably satisfied. b_{Ci} [Figure 4.5] for every spline, even if they present a right lighter tail, follow a Gaussian distribution and therefore satisfy the hypothesis.

Finally we examined the the model's results. The fixed effect that represent the mean monthly electricity temperature response curve relative to the city of Milan has a positive concavity with the minimum at 19.1 °C relative to 4.6 kWh. The extremes instead are at 5.5 kWh and at 5kWh. Studying analytically the curves we observed that an increase of temperature of 1°C will lead to an increase in the average day monthly consumption of 1.2%.

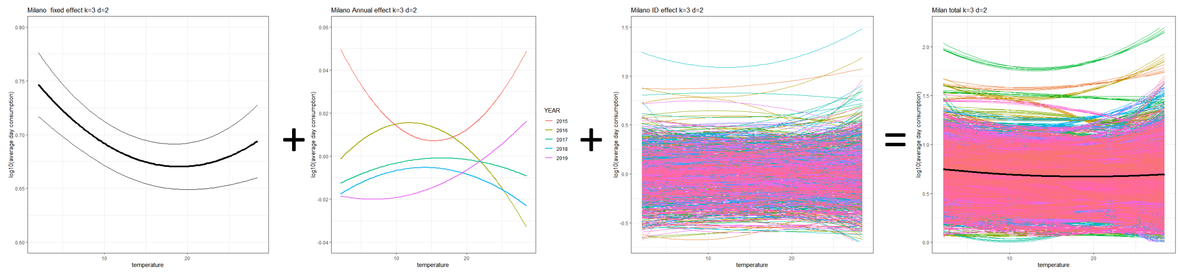

Figure 4.6.: Model 1: Fixed effect with 95% point-wise confident interval, year and clients random effects and total estimated curves of the model.

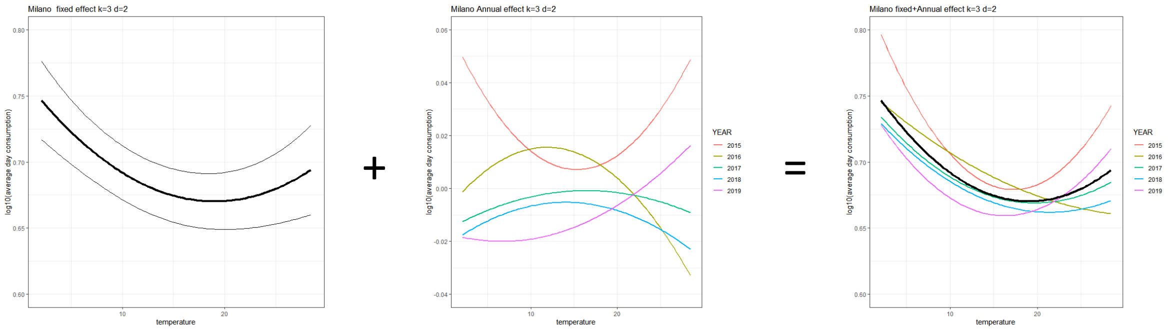

Figure 4.7.: Model 1: Fixed effect with 95% point-wise confident interval, Annual random effect and Annual mean curves.

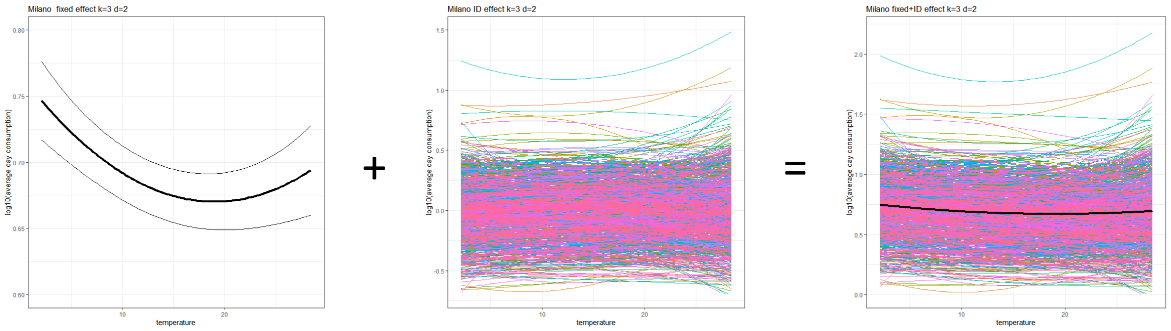

Figure 4.8.: Model 1: Fixed effect with 95% point-wise confident interval, Client random effect and Clients mean curves.

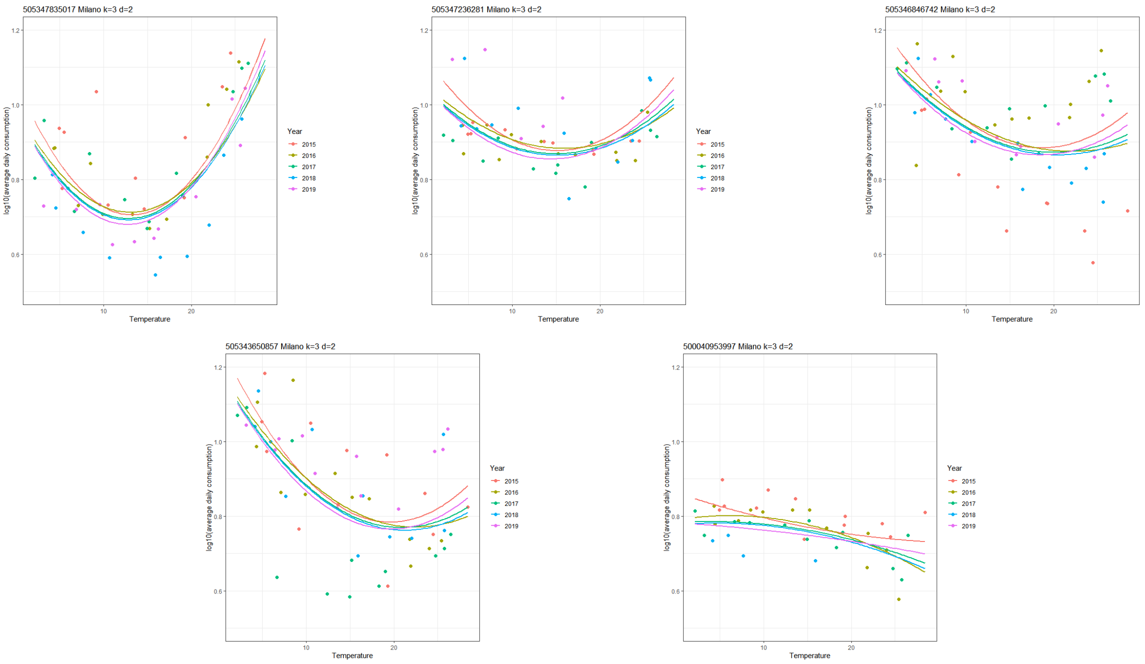

Figure 4.9.: Model 1: Estimated curves for 5 clients.

In the annual mean curves we can see a difference in the minimum of the parabola [Figure 4.7]. In particular, we have seen that 2017 and 2018 had a similar behaviour that can be explained by their similar monthly temperature, with 2017 slightly warmer. Regarding 2016, it was the coolest year with a maximum temperature if 25.4 °C that can explain the decreasing curve. Finally 2015 was the hottest year with a pick at 28.4 °C that justifies the highest consumption. Instead 2019 had a similar behaviour of 2015 with the difference that was a colder year, this can justifies the reduced consumption of electricity.

To understand better this differences and if it could be explained with the heating regulation that impose for the city of Milan the switching on and of respectively the 15 October and the 15 April, we ran the model including a categorical variable in the fixed effect that represents the heating switching. Finally we performed a Likelihood Ratio Test to test the significativity of the heating categorical variable ending up with a p-value of 0.03 . For the goal of our analysis we decided to not consider it in the model because it did not changes the results.

Finally, as we can see in Figure 4.9 and Figure 4.10, we were able to handle the general client behaviour, instead we have very little variability in the years. This confirmed the choice of studying only the overall average curve of the customer to understand the behaviour of the population.

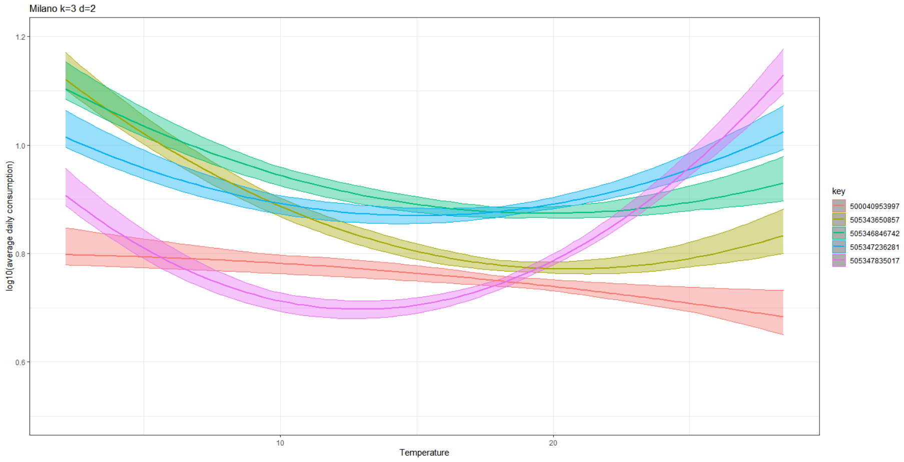

Figure 4.10.: Model 1: mean clients curve, the boundaries represent the minimum and maximum for each point of the functions of each customer.

4.2. Model 1: Analysis

In the following analysis we concentrated on studying only the mean client effect (Fixed effect + Clients Random Effect), to understand common behaviour in the population.

The first analysis we performed was the Functional Principal Component Analysis (PCA). We started from predicted data of previous model on the whole interval (2.127, 28.435) and we interpolated them with a basis of 10 quadratic spline, using *fda* package.

The first two components explained the 98.8% of the variability, as we can see in Figure 4.11. The first component represent the mean behaviour instead the second component the curvature of the function. We can note that the changing point of the second component is 18.3 °C which is the reference temperature of degree days used in model to differentiate between cold and hot season. The minimum of the mean curve instead is at 19.1 °C

We tried to use scores of first and second component for doing clustering but, as we can see in Figure 4.12 there is no a clear division of the data. So we decided to compute functional clustering using the k-mean-alignment explained in Sangalli et al. [25].

We tried both Pearson similarity and L2 distances, using both no-alignment method and only shifting. Pearson performed better because clustered together clients with the same function shape. We can see in Figure 4.13 that we have a knee at k=3. Cluster 1 represent clients with a constant consumption, Cluster 2 clients that reduce their consumption at high temperatures and Cluster 3 clients that have higher consumption at higher temperatures. L2 similarity instead, as we can see in Figure 4.14, captured the changing in the mean consumption clustering high, medium and low consumption clients separately.

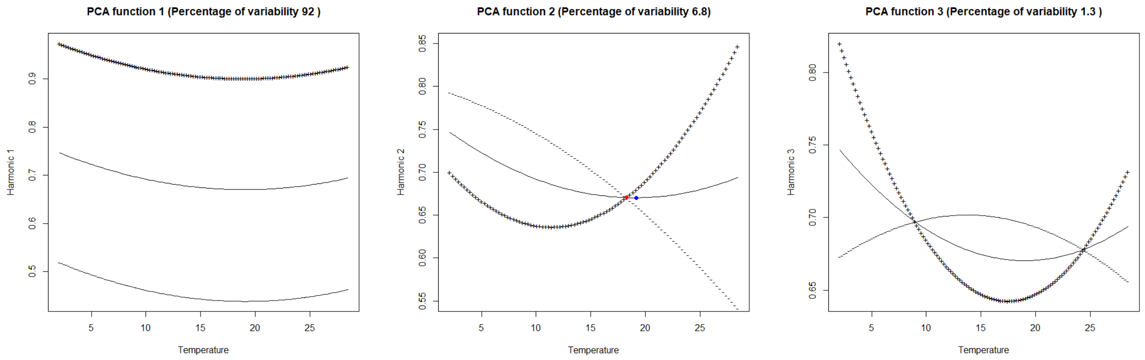

Figure 4.11.: Model 1: Functional PCA. In the second component (centre) are also identified with blue and red points, respectively the minimum of the mean curve and the changing point.

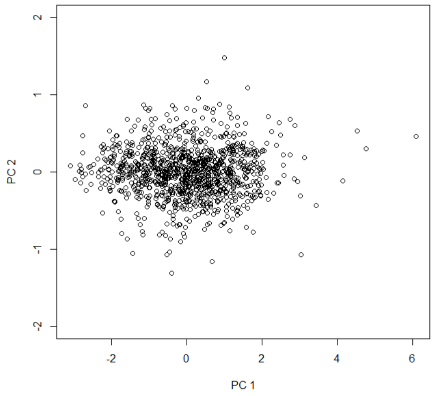

Figure 4.12.: Model 1: Scores first and second principal component.

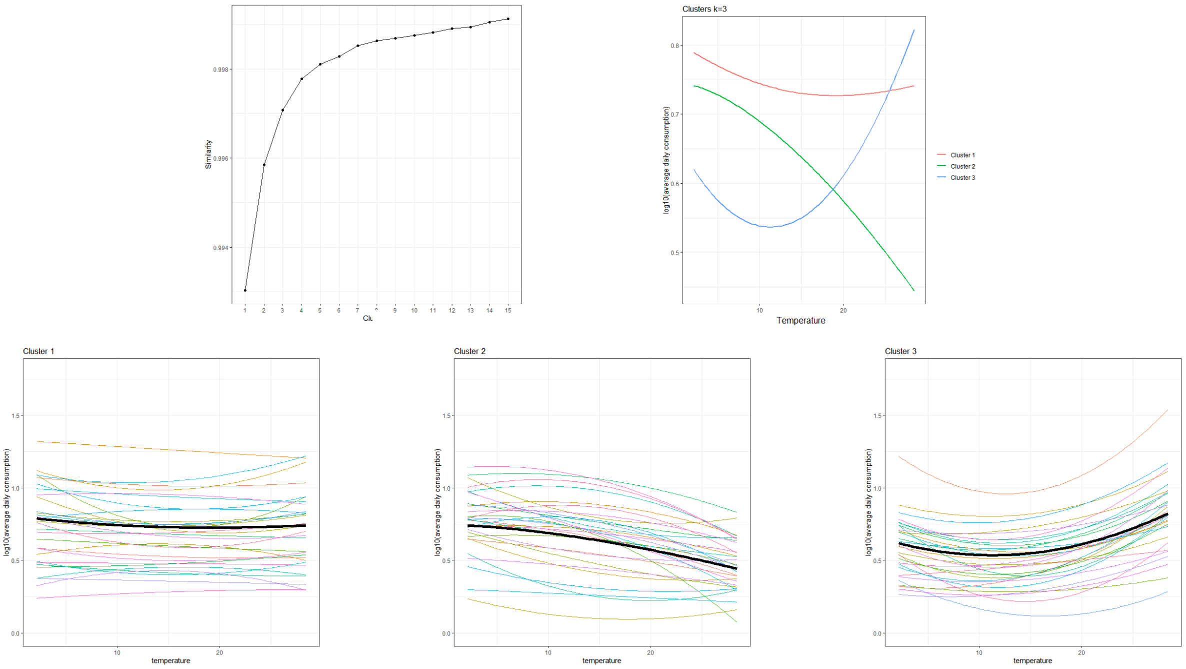

Figure 4.13.: Functional Clustering Pearson: In the top panels are presented the total within similarity for different number of cluster and the cluster's mean curves for k=3. The bottom panels represent the relative classified curves.

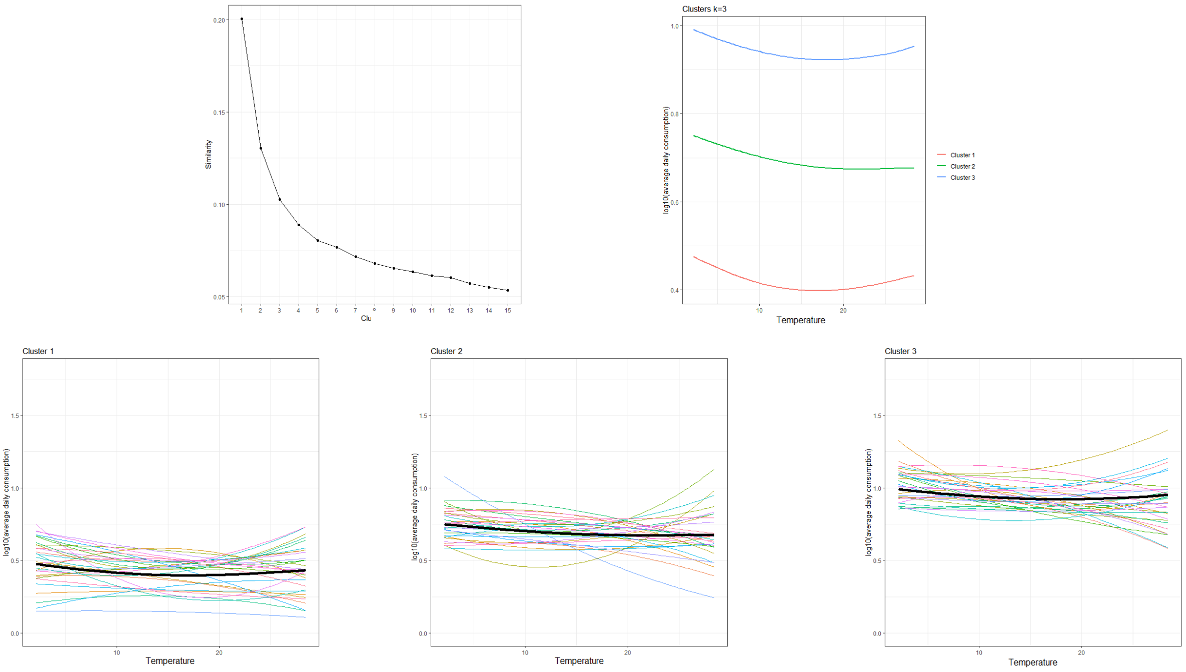

Figure 4.14.: Functional Clustering L2: In the top panels are presented the total within distance for different number of cluster and the cluster's mean curves for k=3. The bottom panels represent the relative classified curves.

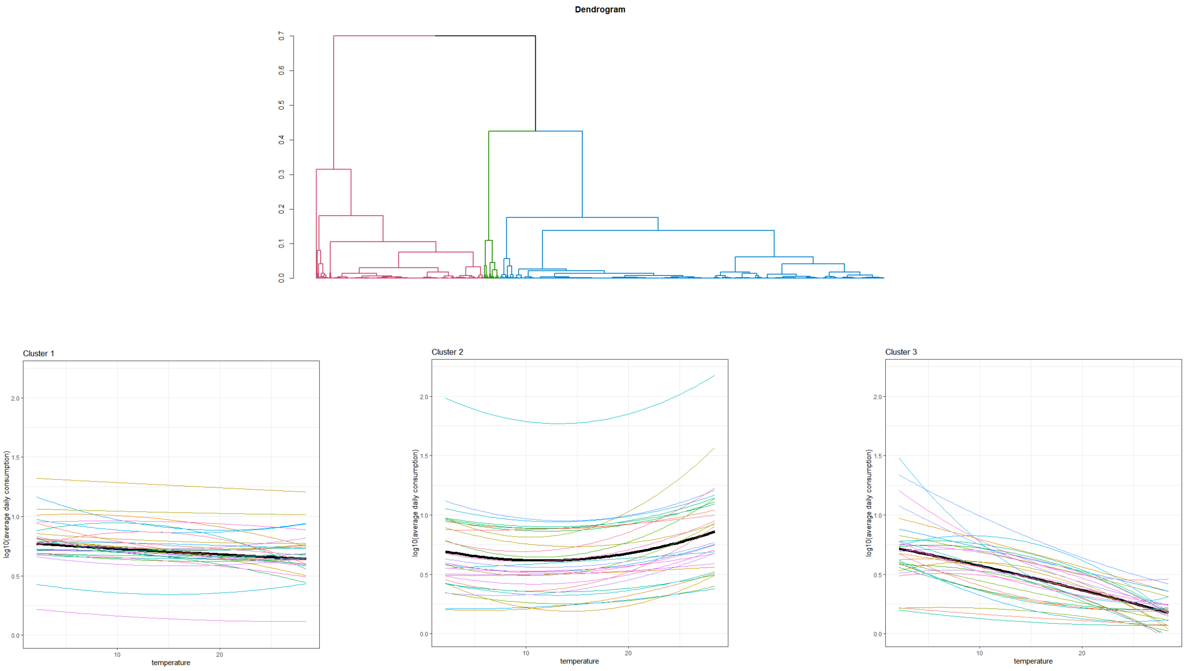

Figure 4.15.: Model 1: The top panel represent the hierarchical dendrogram computed with Pearson similarity and Ward linkage. The bottom panels represent the relative classified curves for each cluster.

We were not fully satisfied from the results so we tried also hierarchical clustering using Pearson similarity to compute the distance matrix with both complete and Ward linkage. Ward linkage, that minimizes the total within-cluster variance, better clustered the different groups: Cluster 1 and 2 are the major part of the population with respectively 770 and 335 clients and represent respectively, constant behaviour and increased consumption at high temperature (see Figure 4.15).

Considering all of the previous clustering results we can note that in each there are some function that were not clustered correctly. Moreover some clients present a negative curvature that goes against climatic theory.

To better analyse this behaviour and tried to study the functions reducing complexity, we decided to compute for each clients the position of the vertex (x, y) and the curvature (d2) analytically, that uniquely identify each parabola.

As we can see in Figure 4.16, we have very high and low value of x outside the interval (0,30). We decided to compute the distribution of x and to consider only the 95% of it, erasing 58 clients that have value outside the 2.5% and 97.5% boundaries [Table 4.3].

From Figure 4.17 we can clearly see parabolas with negative concavity, that did not respect physical behaviour, and degenerate parabolas with zero concavity, that can be represented linearly. At this point we decided to divide the dataset in two different groups using as a threshold d2=0.005:

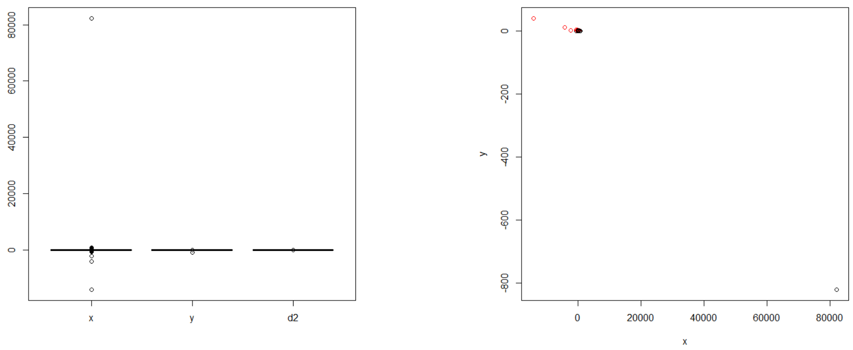

Figure 4.16.: Model 1: On the left Boxplot of the variables relative to vertex position(x,y) and concavity (d2) and Scatterplot of x and y on the right.

Table 4.3.: Distribution of x

Quantiles:

Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
-14028.21	7.80	12.73	66.76	17.39	82191.27

Cut percentiles:

2.5%	**50%**	**97.5%**
-41.90591	12.72733	61.95722

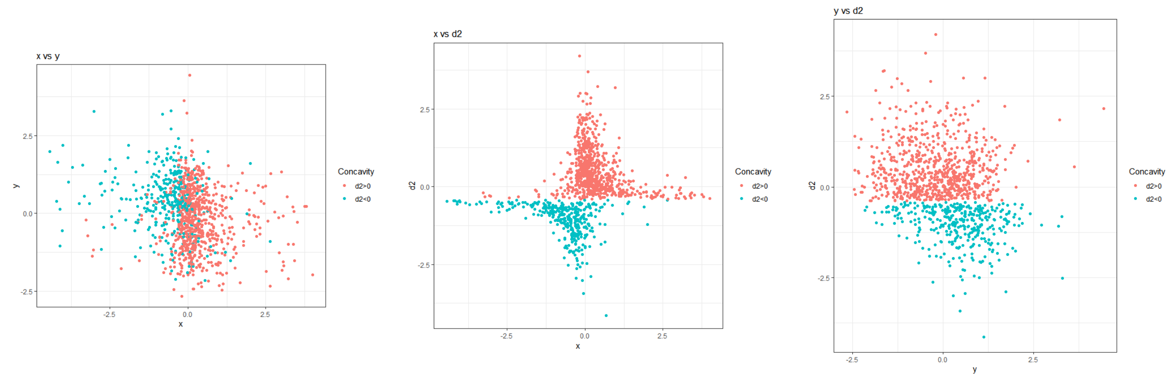

Figure 4.17.: Model 1: Parabolas Standardize Scatterplot. (x,y) on the left, (x,d2) in the centre and (y,d2) on the right.

- d2 $>$ 0.005 : Clients with paraboloid behaviour,
- d2 $\leqslant$ 0.005 : Clients with linear behaviour.

4.3. Second Approach: Paraboloid and Linear Behaviour

In this section we analysed the two groups of clients separately, fitting for each a different model.

4.3.1. Paraboloid Behaviour

We classified as clients with paraboloid behaviour the 48% of the dataset and we refitted on them the same model as before following the Equation (4.1). We also maintained the same value for the parameters (k=3, d=2).

Analysing the clients coefficients relative to each spline (see Figure 4.18), we noted 3 outliers that did not follow properly the Gaussian distribution. We can see from Figures 4.19 and 4.20, that they are the clients with higher mean average electricity consumption reaching 100 kWh and they depart from the main group for about 20 kWh. For these reasons we decided not to include them in the analysis.

The model without outliers led to a greater satisfaction of the model's assumptions. If we look at Figure 4.22, they completely fulfill the hypothesis of Gaussianity, except for the second spline that presents a heavier right tail but definitely better than before. The residual instead, in Figure 4.21, presented the same configuration of the first model. They did not follow completely the Gaussian distribution but at least were symmetric, thus we still were reasonably satisfied.

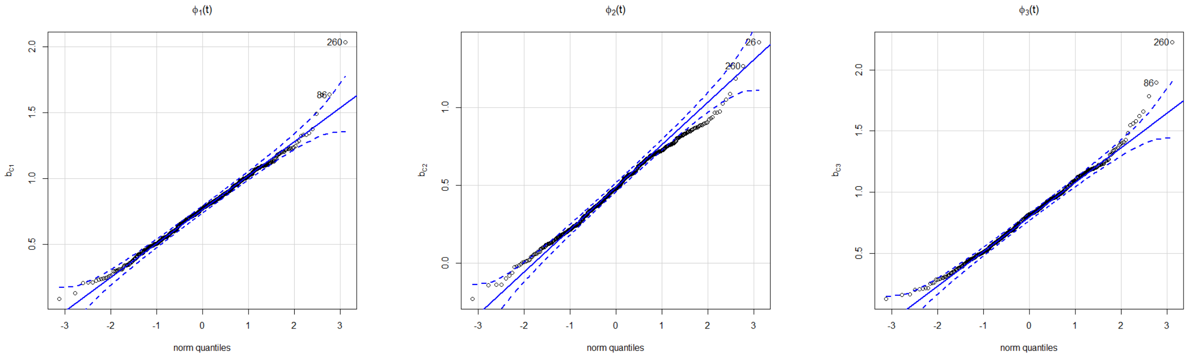

Figure 4.18.: Parabolas: Normality plot for coefficients relative to Clients.

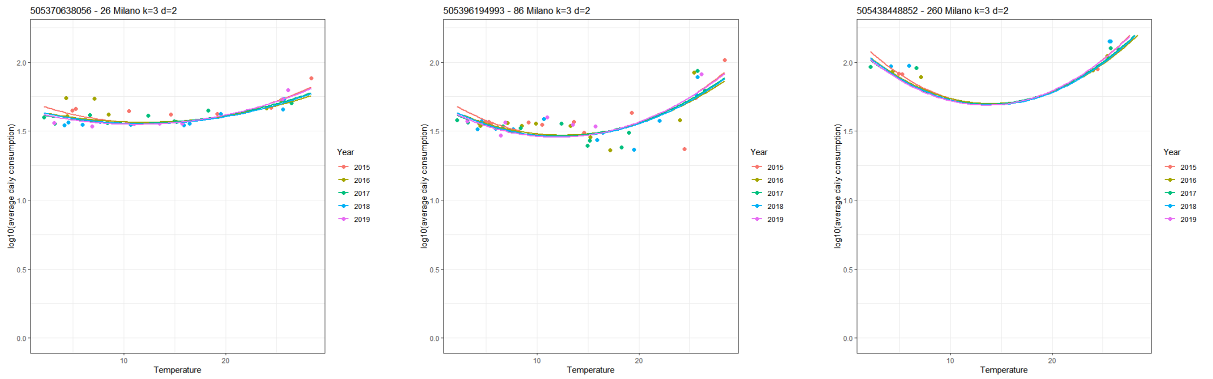

Figure 4.19.: Parabolas: estimated curves for clients with coefficients that not follow Gaussian hypothesis.

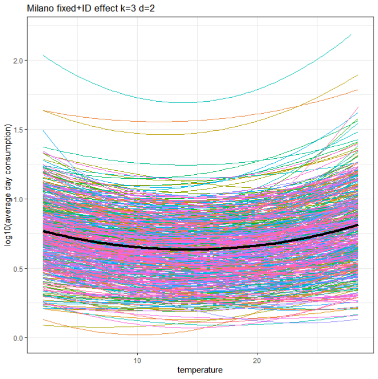

Figure 4.20.: Parabolas: Plot of fixed effect and ID random effects.

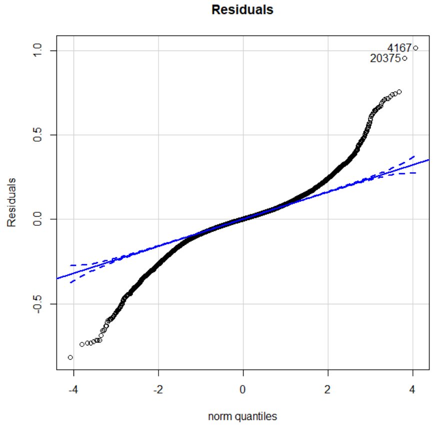

Figure 4.21.: Parabolas: Normality plot for the model's residuals.

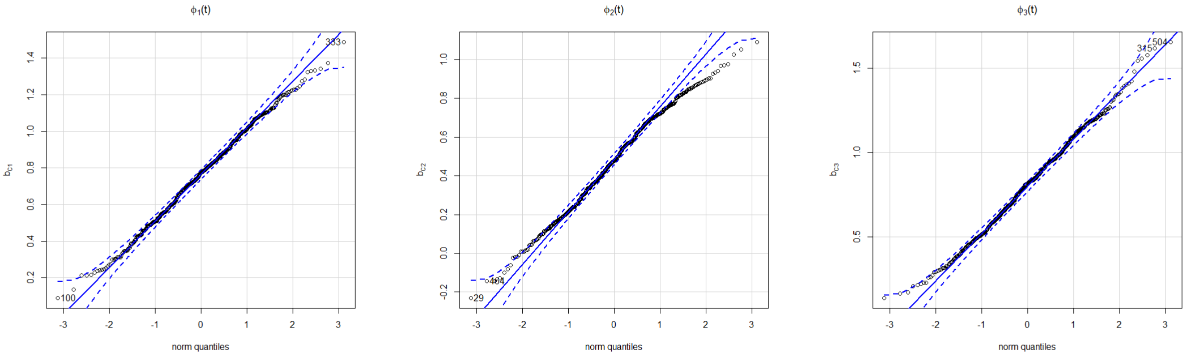

Figure 4.22.: Parabolas: Normality plot for coefficients relative to Clients.

Table 4.4.: Parabolas: Summary model.

Random effects:

Groups	**Name**	**Variance**	**Std. Dev.**
ID	**X3**	0.0800699	0.28297
ID.1	**X2**	0.0670472	0.25893
ID.2	**X1**	0.0640910	0.25316
ANNI	**X3**	0.0007065	0.02658
ANNI.1	**X2**	0.0007810	0.02795
ANNI.2	**X1**	0.0006477	0.02545
Residual		0.0144698	0.12029

Fixed effects:

	Estimate	**Std. Error**	**t value**
X1	0.76300	0.01613	47.30
X2	0.47411	0.01744	27.19
X3	0.80612	0.01732	46.54

Table 4.5.: Parabolas: Likelihood Ratio test for random effect significance.

	Linear Model	**Fixed + YEAR**	**Fixed + ID**	**Complete Model**
LogLikelihood	-1476.13	-1458.58	12095.28	12149.06

	LM vs YEAR	**LM vs ID**	**LM vs Compl**	**YEAR vs Compl**	**ID vs Compl**
LR	35.0928	27142.83	27250.38	27215	107.56
p-val	5.665178e-08	0	0	0	1.821286e-23

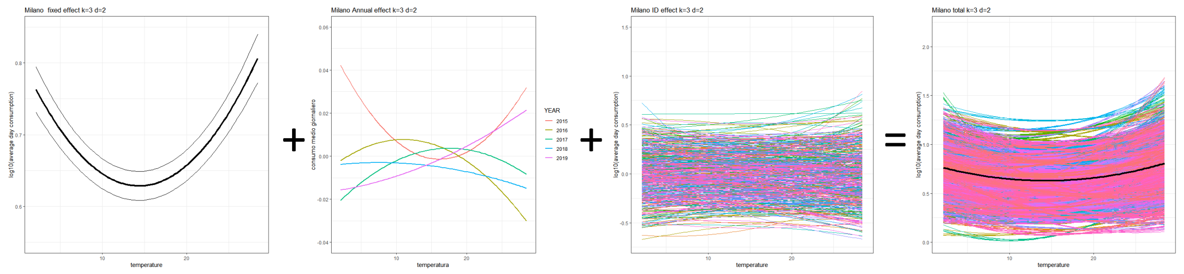

Figure 4.23.: Parabolas: Fixed effect with 95% point-wise confident interval,year and clients random effects and total estimated curves of the model.

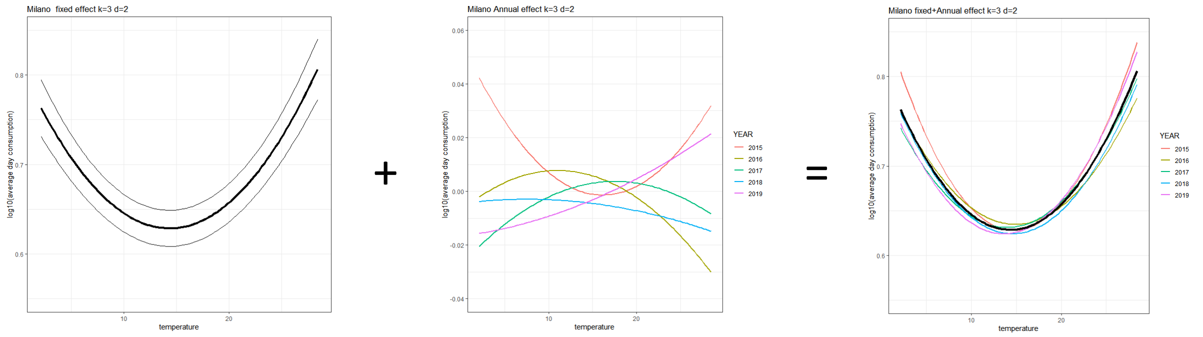

Figure 4.24.: Parabolas: Fixed effect with 95% point-wise confident interval, Annual random effect and Annual mean curves.

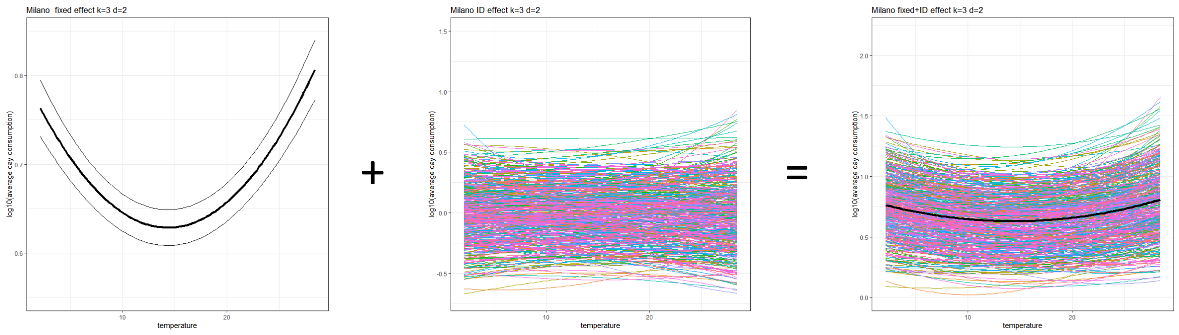

Figure 4.25.: Parabolas: Fixed effect with 95% point-wise confident interval, Client random effect and Clients mean curves.

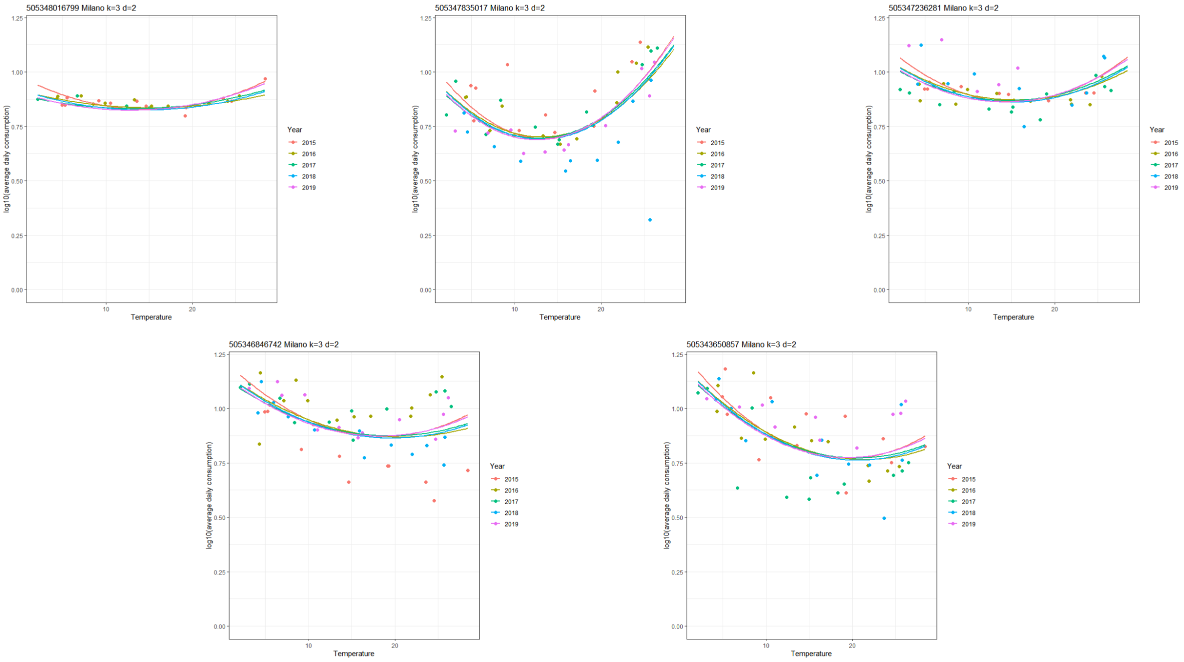

Figure 4.26.: Parabolas: Estimated curves for 5 clients.

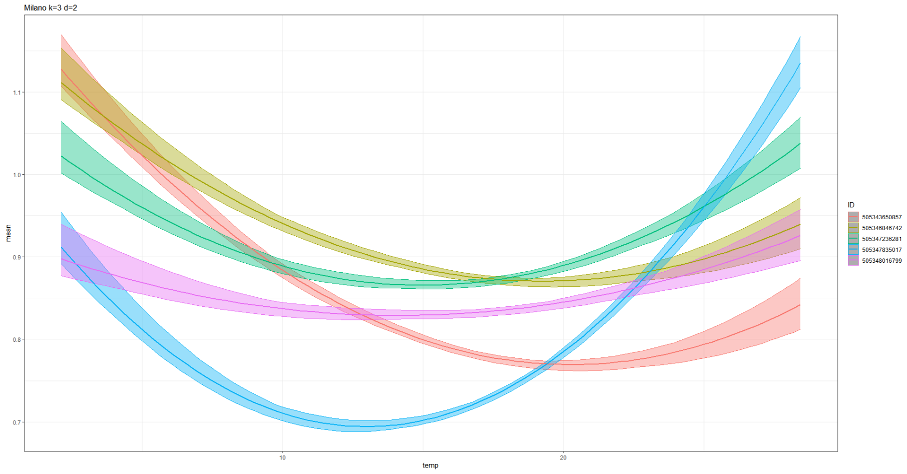

Figure 4.27.: Parabolas: mean clients curve, the boundaries represent the minimum and maximum for each point of the functions of each customer.

As we can see in Table 4.4, we obtained very similar results as the first model and also in this case we found major variability in the clients random effect. Equally we performed the Likelihood Ratio Test to check the significance of the random effects. Once again we had the confirmation of the importance of all the Random effects, included the "YEAR" component. We can see the results in Table 4.5.

Analysing the model results relative to Clients with paraboloid behaviour (see Figure 4.23), we found the mean function of the population had the minimum at 14.37 and that an increase of 1°C of temperature will lead to an increasing in the consumption of the 6.3%. In this context we had also some differences in the annual mean effects similar as the first model but affect less the final result [4.24].

The little variance of the annual effect can be seen also in Figures 4.26 and 4.27, therefore we can use only the mean effect of the clients to study the population behaviour.

4.3.2. Linear Behaviour

In this section we analyse the remaining 52% of clients that present a linear behaviour. In this case we use a classic linear mixed effect model that we can be described with the following equation:

$$y_{YC}(t) = \beta_0 + \beta * t + B_{Y0} + B_Y * t + b_{C0} + b_C * t + \epsilon_{YC}(t) \tag{4.2}$$

where:

- $y_{YC}(t)$ is the datum that one would have recorded if the client C in the year $Y \in (2015, 2019)$ were measured at temperature $t \in (2.127, 28.435)$,
- $\beta_0 + \beta * t$ indicates the Milan mean curve,
- $B_{Y0} + B_Y * t$ is the correction relative to the specific year,
- $b_{C0} + b_C * t$ is the correction for the specific client,
- $\epsilon_{YC}(t)$ indicates the specific observation measurement error,
- $\mathbf{b_C} \sim \mathcal{N}(0, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is the variance covariance matrix of the clients random effect,
- $\epsilon_{YC}(t) \sim \mathcal{N}(0, \sigma^2)$ for each client C in the year Y for every time t.

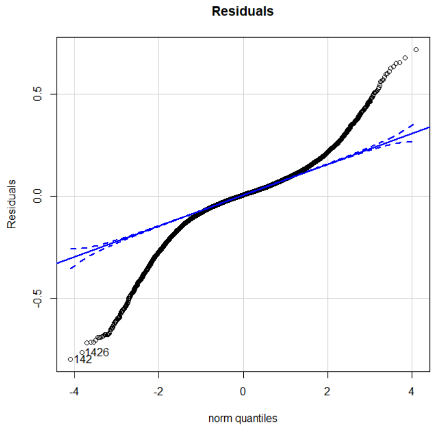

Figure 4.28.: Linear: Normality plot for the model's residuals.

The residuals of the model not follow properly the Gaussian distribution, even so like the previous model is a symmetric distribution with lighter tail. The random coefficients relative to clients (see Figure 4.29) meet the assumption of Gaussianity, instead the ones relative to the temperature have a lighter left tail. If we analyse more specifically we can see that two clients in particular do not follow the quantiles of the Gaussian distribution (60 and 375). Looking at the predicted curves in Figure 4.30, we can note a decreasing paraboloid behaviour estimated by the first model.

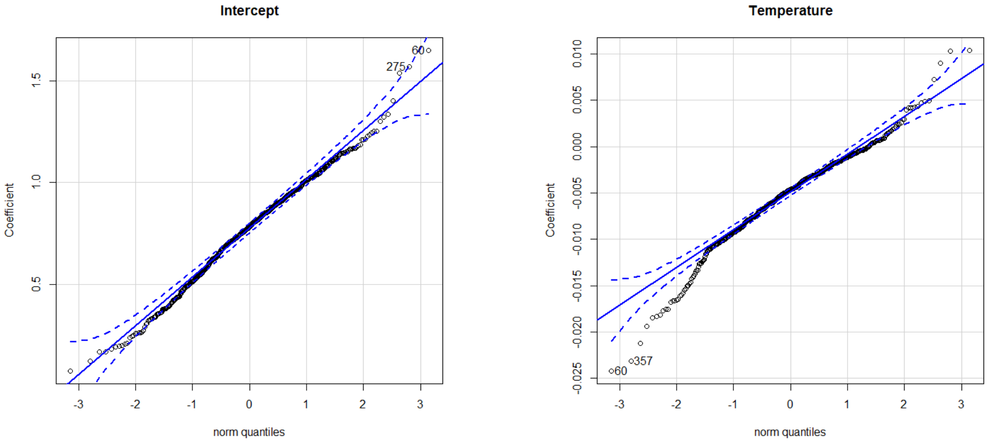

Figure 4.29.: Linear: Normality plot for coefficients relative to Clients.

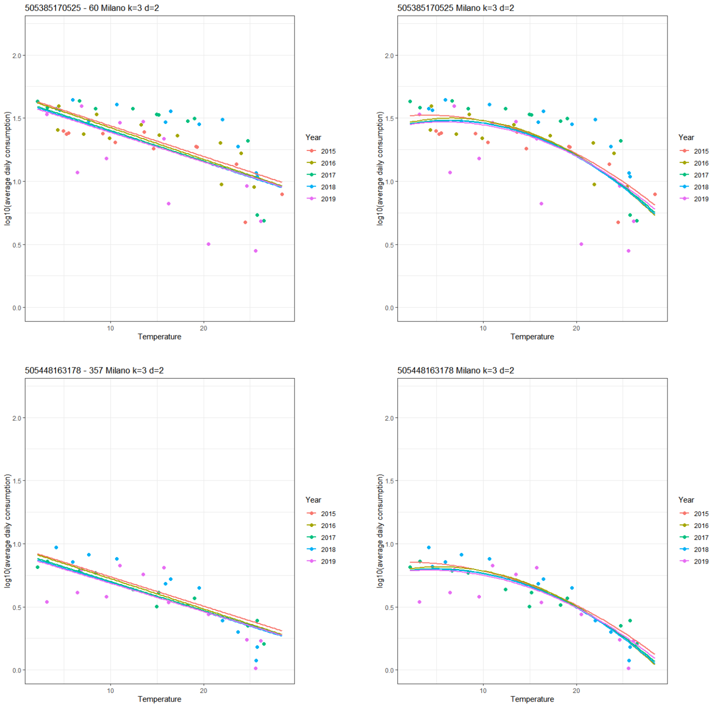

Figure 4.30.: Linear: Comparison between estimated curves using linear model (left) and model 1 (right) for clients with coefficients that not follow Gaussian hypothesis.

Table 4.6.: Linear: Summary model.

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
ID	**(Intercept)**	6.202e-02	0.249037	
	temp	2.661e-05	0.005159	-0.46
ANNI	**(Intercept)**	6.886e-04	0.026242	
	temp	3.708e-07	0.000609	-0.78
Residual		1.305e-02	0.114258	

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	0.769507	0.015701	49.01
temp	-0.005108	0.000364	-14.03

Table 4.7.: Linear : Likelihood Ratio test for random effect significance.

	Linear Model	Fixed + YEAR	Fixed + ID	Complete Model
LogLikelihood	-581.443	-516.5887	15355.14	15562.27

	LM vs YEAR	LM vs ID	LM vs Compl	YEAR vs Compl	ID vs Compl
LR	129.7088	31873.16	32287.42	32158	414.26
p-val	5.665178e-08	0	0	0	1.821286e-23

Unfortunately it is not allowed by theory and our linear model succeeded to estimate the consumption trend considerably. For these motivations we didn't erase the observation from the model.

The summary (see Table 4.6) highlights also in this case a higher variance of the clients random effect, that can also be seen in Figures 4.34 and 4.35.

The result of Likelihood Ratio Test, computed to test the significance of the random effect Table 4.2, confirmed the importance of all.

The model underline a decreasing mean effect for the city of Milan with an intercept of 5.88 kWh and slope of -0.005. a higher consumption for 2019 at higher temperatures. Only 2018 and 2019 intersect each other with a higher consumption for 2019 at higher temperatures.

Finally we can justify the study of the mean consumption curves of the clients for identifying common pattern in the population, because, as we can see in Figures 4.34 and 4.35 we succeed to estimate the general trend of the customers, that have higher variance respect to the annual effect.

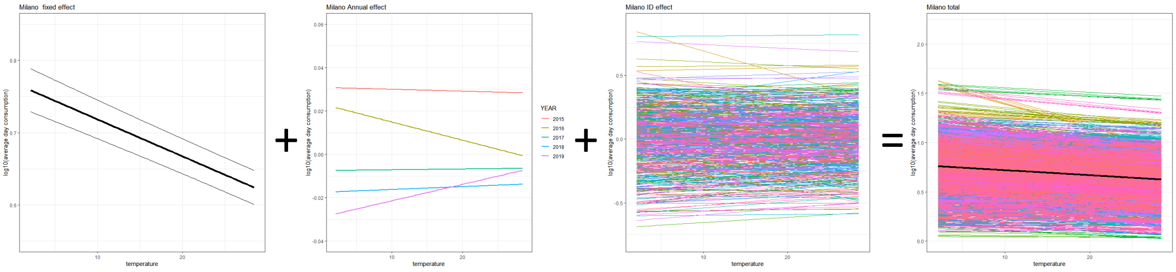

Figure 4.31.: Linear: Fixed effect with 95% point-wise confident interval,year and clients random effects and total estimated curves of the model.

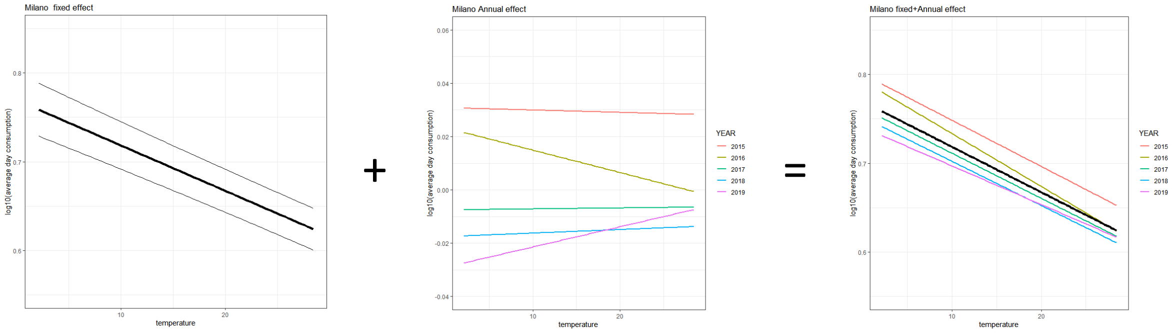

Figure 4.32.: Linear: Fixed effect with 95% point-wise confident interval, Annual random effect and Annual mean curves.

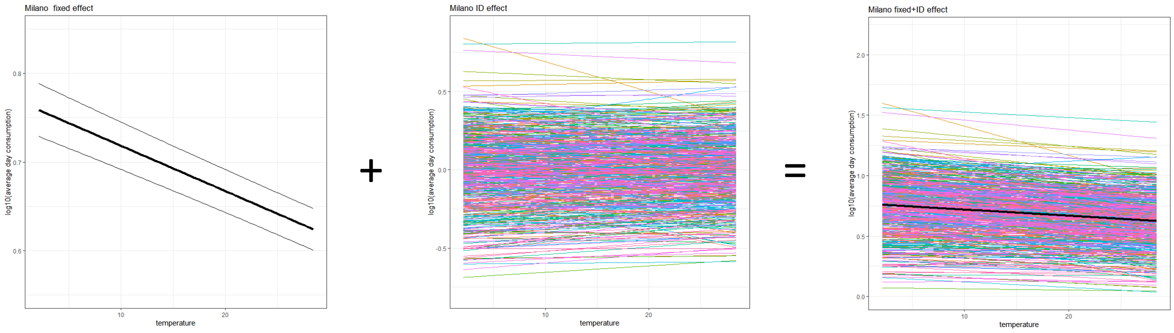

Figure 4.33.: Linear: Fixed effect with 95% point-wise confident interval, Client random effect and Clients mean curves.

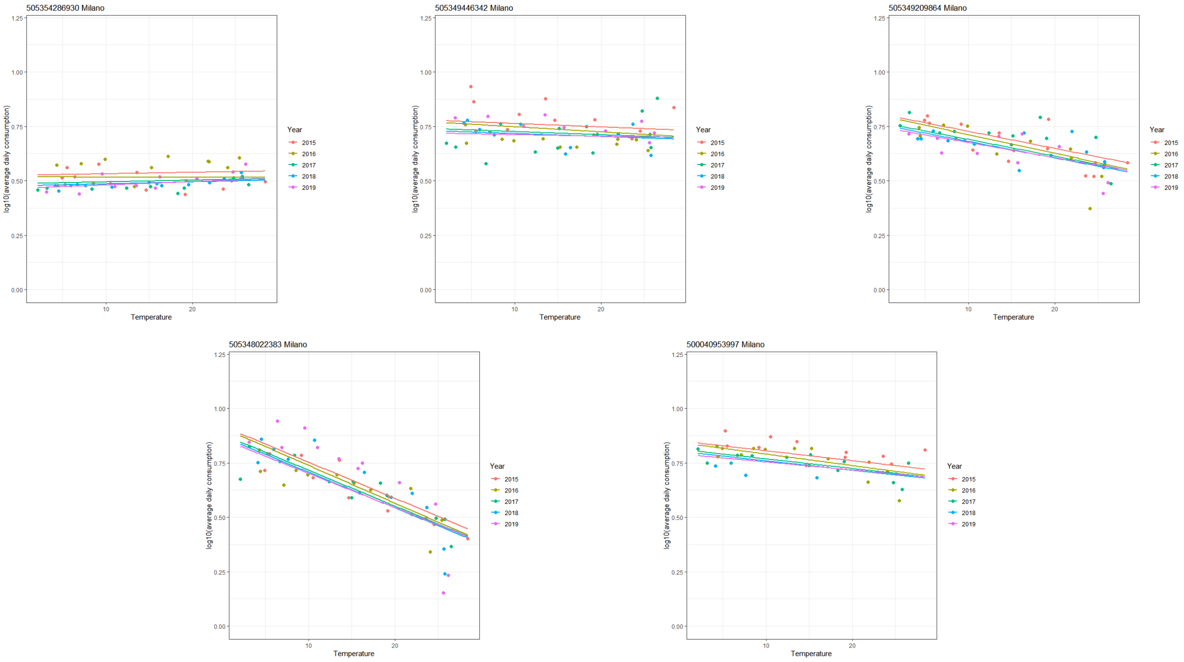

Figure 4.34.: Linear: Estimated curves for 5 clients.

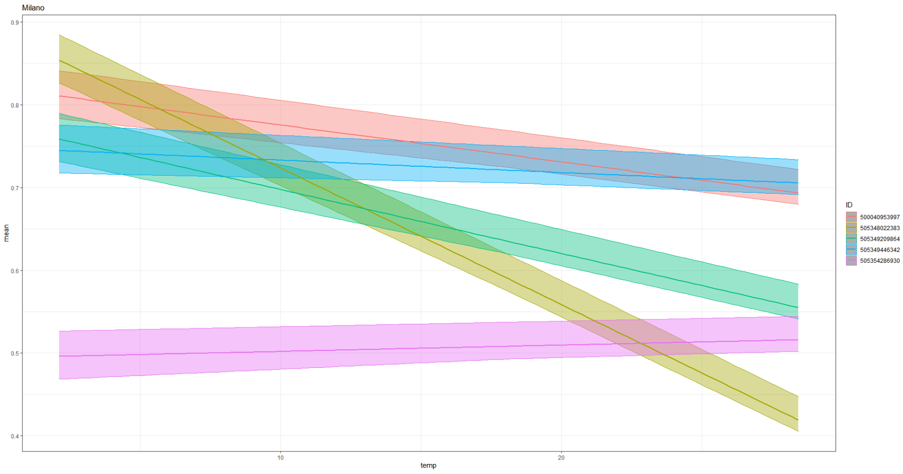

Figure 4.35.: Linear: mean clients curve, the boundaries represent the minimum and maximum for each point of the functions of each customer.

4.4. Second Approach: Analysis

Once obtained the mean function relative to clients of both model we analysed them separately to cluster them and study the population behaviour.

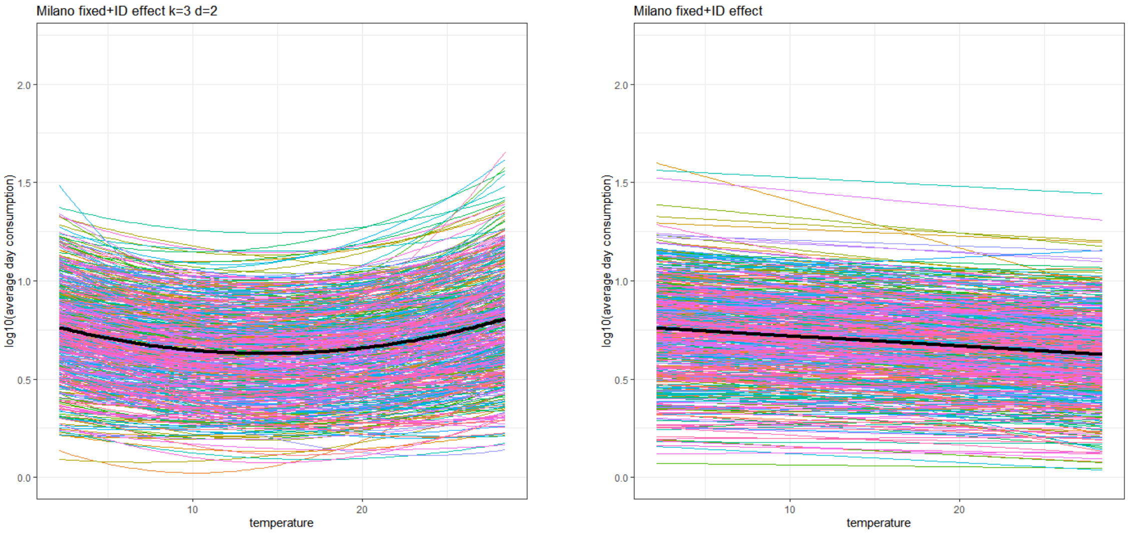

Figure 4.36.: Fixed effect and ID random effect of clients with paraboloid behaviour (left) and linear behaviour (right).

4.4.1. Paraboloid Behaviour

Following the same procedure of the first model, we proceeded with a functional principal component analysis. As we can se in Figure 4.37 the first two principal component explained the 99% of the variability. As before, the first component represents the mean consumption levels and the second one the concavity of the curve. The minimum now is at 14.3 °C and the changing point is at 16.6 °C. If we compare the second components of the parabolas to the one of the first model, we can note that in this case we obtained more concave curves with high consumption of electricity at extreme temperatures. We plotted the scores of the first two principal component to try to use them for clustering, but also this time they didn't present any grouping pattern.

To cluster clients electricity response curve we tried different approaches. The first one was the classical functional clustering without alignment and Pearson similarity, using the procedure of Sangalli et al. [25]. To choose the number of cluster we plotted the mean of the similarities of the curves respect to the cluster mean. As we can see from Figure 4.39 we have an elbow at k=3. The first cluster, composed by the 28% of the clients, represent an increase of consumption relative to higher temperature, probably

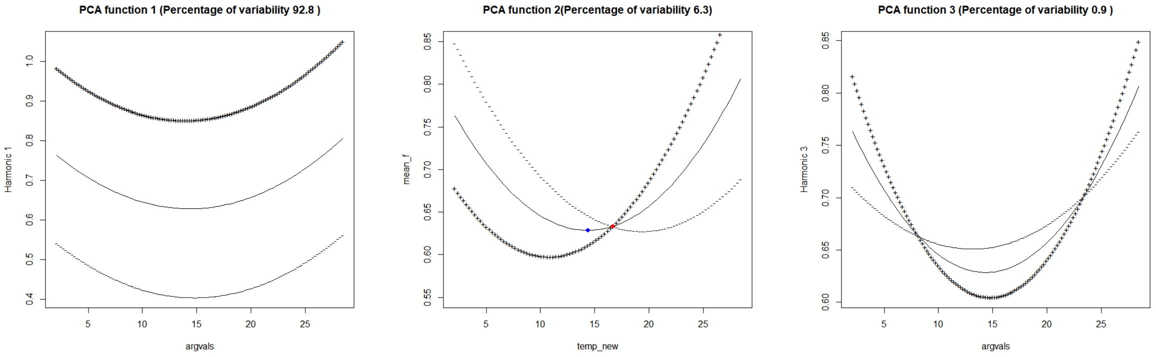

Figure 4.37.: Parabolas: Functional PCA. In the second component (centre) are also identified with blue and red points, respectively the minimum of the mean curve and the changing point.

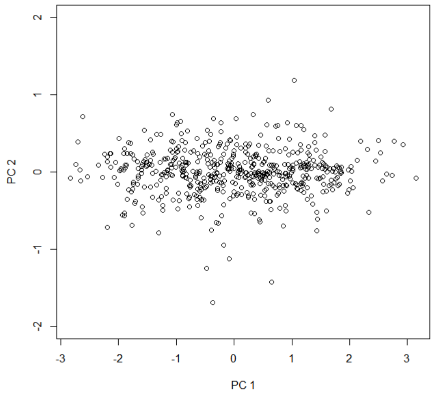

Figure 4.38.: Parabolas: scores of first and second principal component.

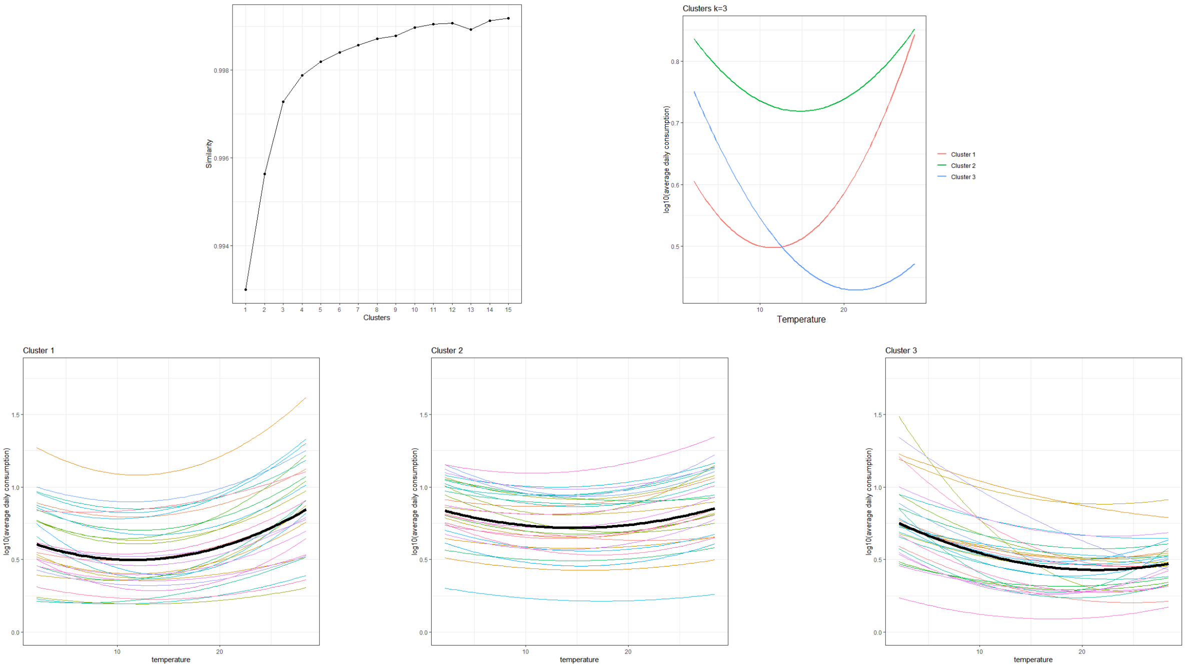

Figure 4.39.: Parabolas: Functional Clustering Pearson. In the top panels are presented the total within similarity for different number of cluster and the cluster's mean curves for k=3. The bottom panels represent the relative classified curves.

caused by the presence of air conditioning. The second cluster (58% of clients) presents what can be considered as a constant behaviour with slightly increasing consumption at temperatures extremes. Finally the remaining 14% presents a decreasing behaviour. It is interesting to note that all the clients will experience an increasing in the consumption of electricity for future higher temperatures.

The second approach we attempted was hierarchical clustering using both complete and Ward linkage. We computed the distance matrix using Pearson similarity. Complete linkage (see Figure 4.40) identifies a major cluster composed by the 90% of the clients, characterized by low concavity and a major consumption of electricity at extreme temperature. Instead the second and the third cluster (5% of clients each), represent respectively clients with higher consumption at high temperature and low temperature. Ward linkage (see Figure 4.41) assigned more clients (20%) to the decreasing pattern in the first cluster, 4% of clients to the third cluster and the 76% to the second one. Concluding, we observed that Ward method better clustered the population, if we look close we can see that first cluster of complete linkage include some clients with decreasing curve that are assigned correctly in Ward to Cluster 1.

Finally, knowing that parabolas can be described uniquely by 3 parameters we decided to compute the position of vertex (x, y) and concavity (d2) of the curves and cluster clients using this new dataset, reducing the complexity of computations.

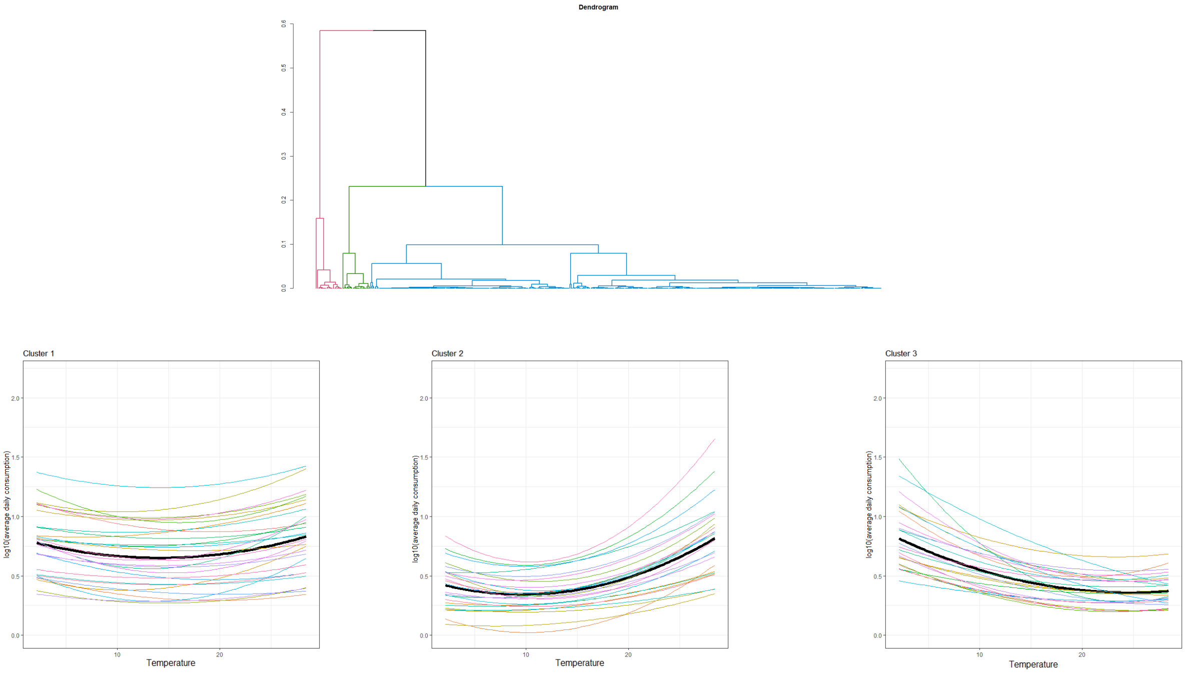

Figure 4.40.: Parabolas: The top panel represent the hierarchical dendrogram computed with Pearson similarity and complete linkage, k=3. The bottom panels represent the relative classified curves for each cluster.

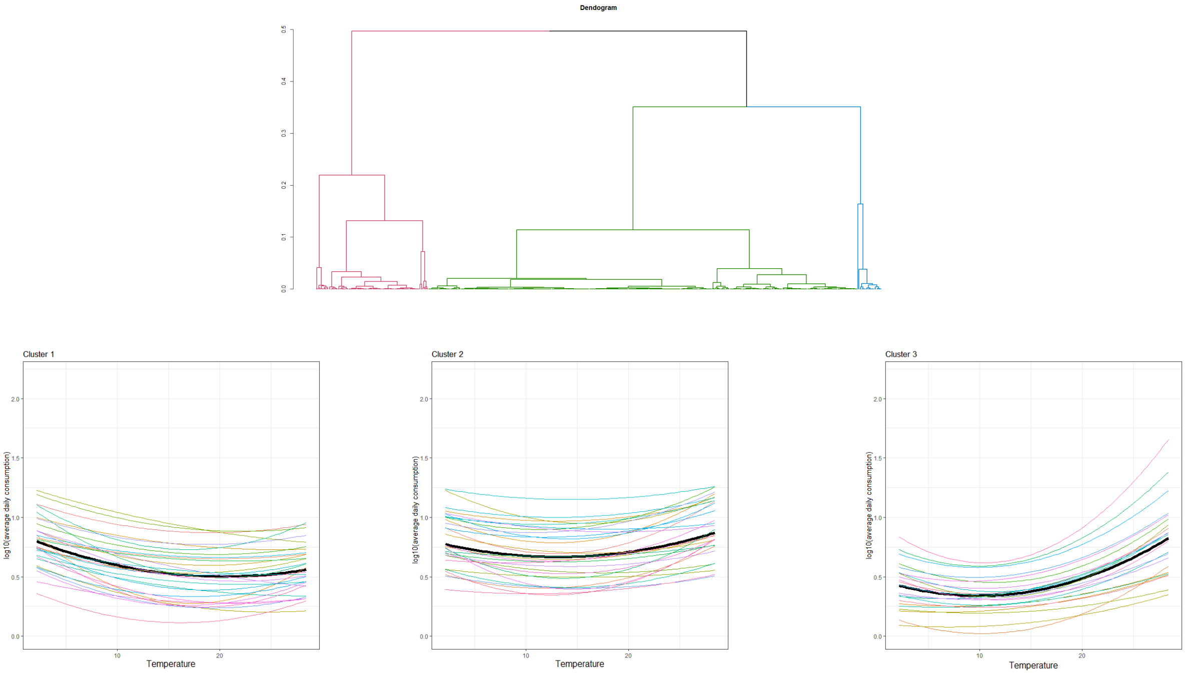

Figure 4.41.: Parabolas: The top panel represent the hierarchical dendrogram computed with Pearson similarity and Ward linkage, k=3. The bottom panels represent the relative classified curves for each cluster.

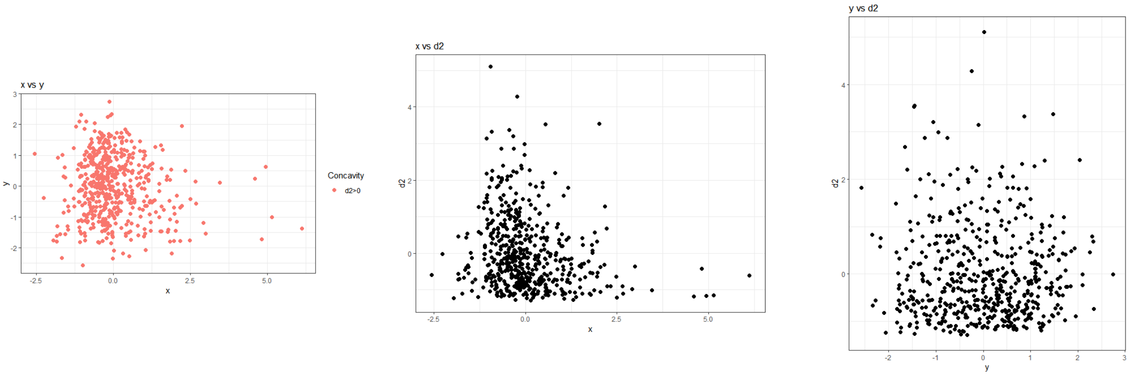

Figure 4.42.: Parabolas: clients vertex position and concavity standardized scatter plot. (x,y) on the left, (x,d2) in the centre and (y,d2) on the right.

As we can see in the scatterplot in Figure 4.42, this time we have only positive concavity and there isn't any pattern in the data.

Applying k-means to the constructed standardized dataset, we choose k=4. As we can see from Figure 4.43, after repeating the algorithm for different seed and different number of cluster, it was the most stable. We can identify a constant behaviour in clusters 1 (38%) and 3 (31%)that differ only for the mean consumption, respectively middle-high and low. Cluster 2 (13%) is characterized by a decreasing behaviour and Cluster 4 (18%) has higher concavity with an increase of consumptions for high temperature. This new constructed dataset is useful also because we can study the marginal density functions relative to each cluster (see Figure 4.44) and use them for tuning the climatic model to have a more precise projection on the future consumptions.

Moreover, we analysed the cluster generated using only all the possible couples of variables. In our opinion the most significative is the one using the abscissa and the concavity. In this way we are highlighting differences regarding the shape of the curve but not on the mean consumption. In Figure 4.45, we can see the results. As before we found 3 cluster representing a constant consumption with a little increase in the consumption at high temperature (60%), clients with a decreasing trend (16%) and a cluster with higher concavity and high consumption at high temperature (24%). Also in this case, we can study the marginal and joint density functions the abscissa and concavity and use them for tuning the parameters of climatic models (see Figure 4.46).

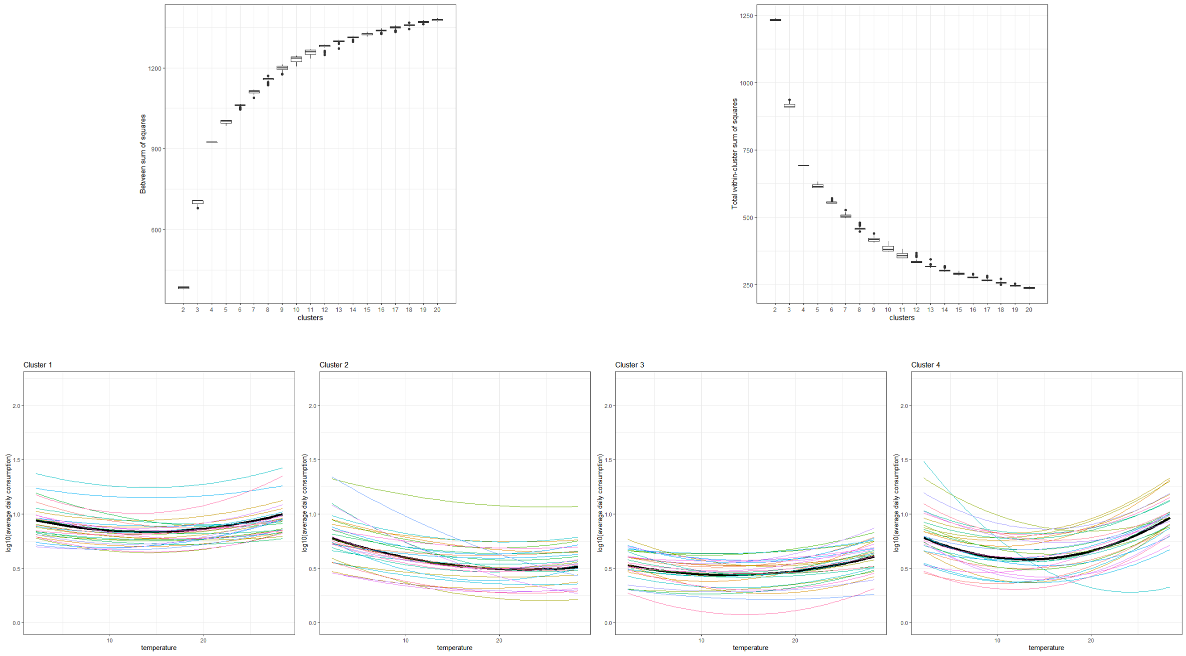

Figure 4.43.: Parabolas: In the top panel are represented the between (left) and within (right) total sum of squares of Kmean clustering using vertex position and concavity. The bottom panels represent the relative classified curves for k=4.

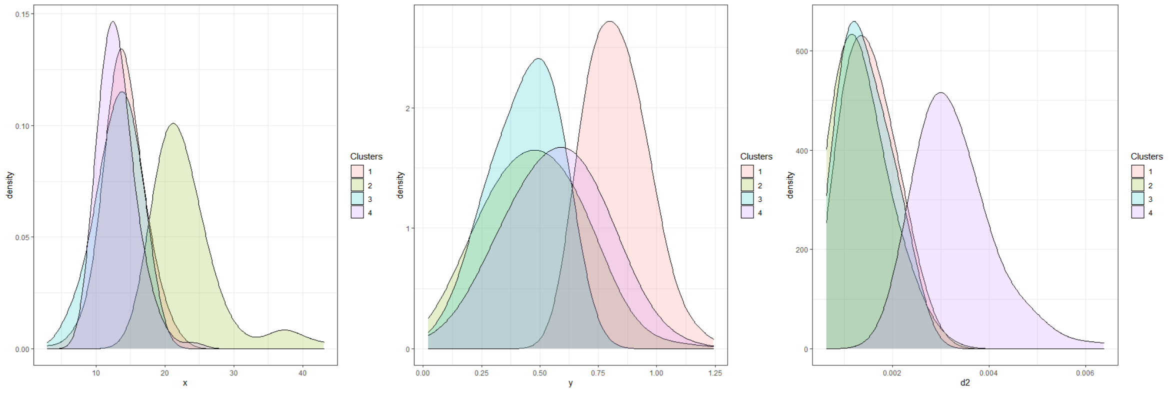

Figure 4.44.: Parabolas: marginal pdfs for x, y, d2 relative to k=4 clusters identified by kmean.

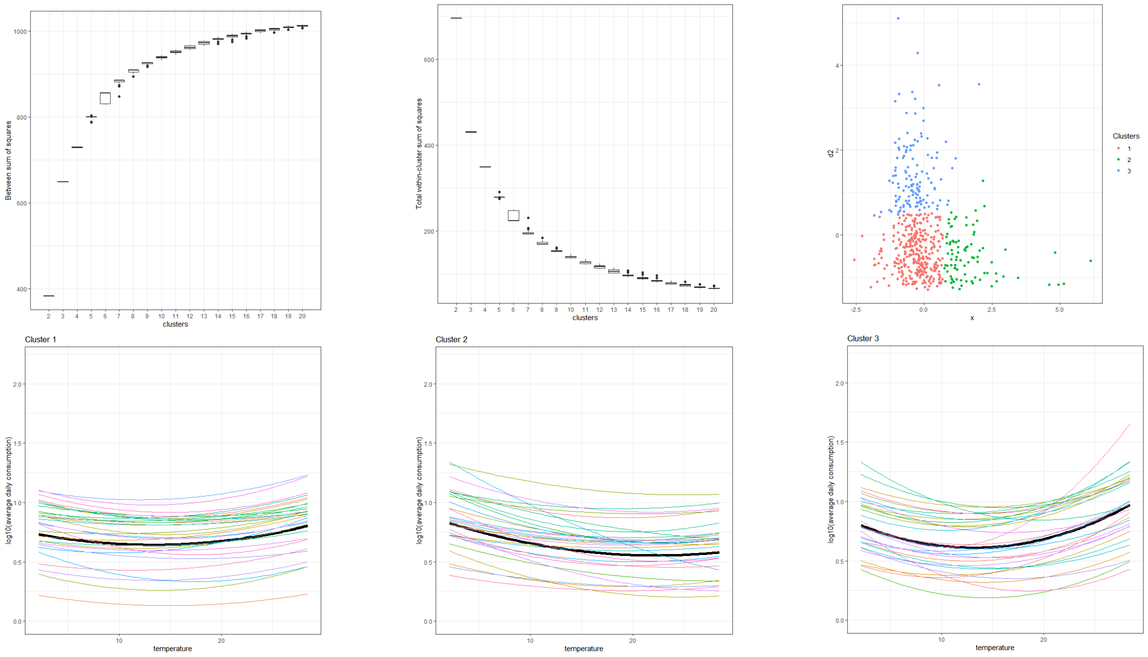

Figure 4.45.: Parabolas: In the top panel are represented the between (left), within (centre) total sum of squares of Kmean clustering using vertex abscissa and concavity and the cluster scatterplot relative to k=3. The bottom panels represent the relative classified curves.

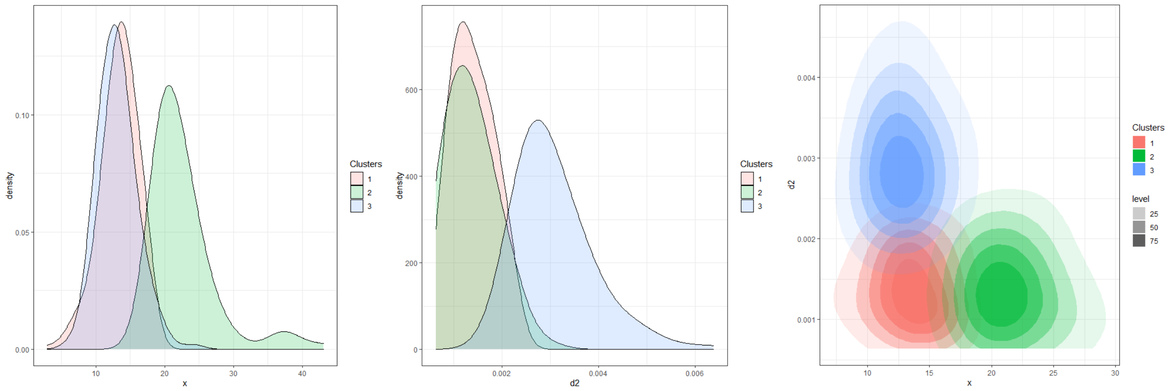

Figure 4.46.: Parabolas: marginal and joint pdfs for x, d2 relative to k=3 clusters identified by kmean.

4.4.2. Linear Behaviour

In the following subsection we present the result of clustering of clients with linear behaviour.

We started with functional kmeans using Sangalli et al. [25] algorithm and Pearson similarity. In Figures 4.47 and 4.48 we consider 2 cases k=2 and k=3. If we consider 3 cluster we obtain two very similar cluster containing decreasing lines and we can note some misclassified clients. For this reason we preferred the case k=2. The first cluster, composed by the 73% of clients, represents a constant consumption while the second one (27%) a decreasing trend. We can justify this behaviour by supposing that these clients didn't have the air conditioning and the consumption of electricity at higher temperatures can be caused by a less use of lighting in the hottest month that correspond to the one with more daily sun hours.

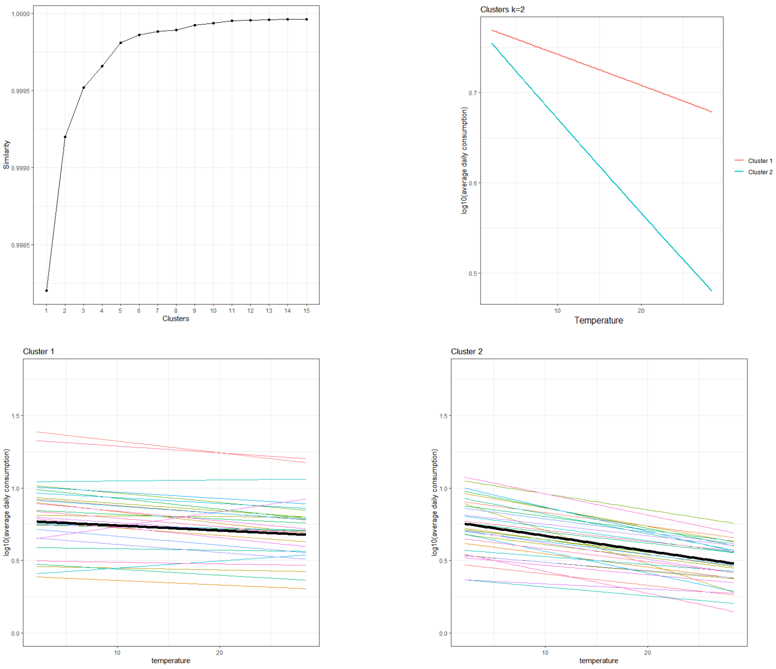

Figure 4.47.: Linear: Functional Clustering Pearson. In the top panels are presented the total within similarity for different number of cluster and the cluster's mean curves for k=2. The bottom panels represent the relative classified curves.

The second approach tested was functional hierarchical clustering. Also in this case we constructed the distance matrix using Pearson similarity. The complete linkage case (see Figure 4.49) identified the same two cluster but in this case clustered as constant the majority of the clients (94%), and only 6% with decreasing trend.

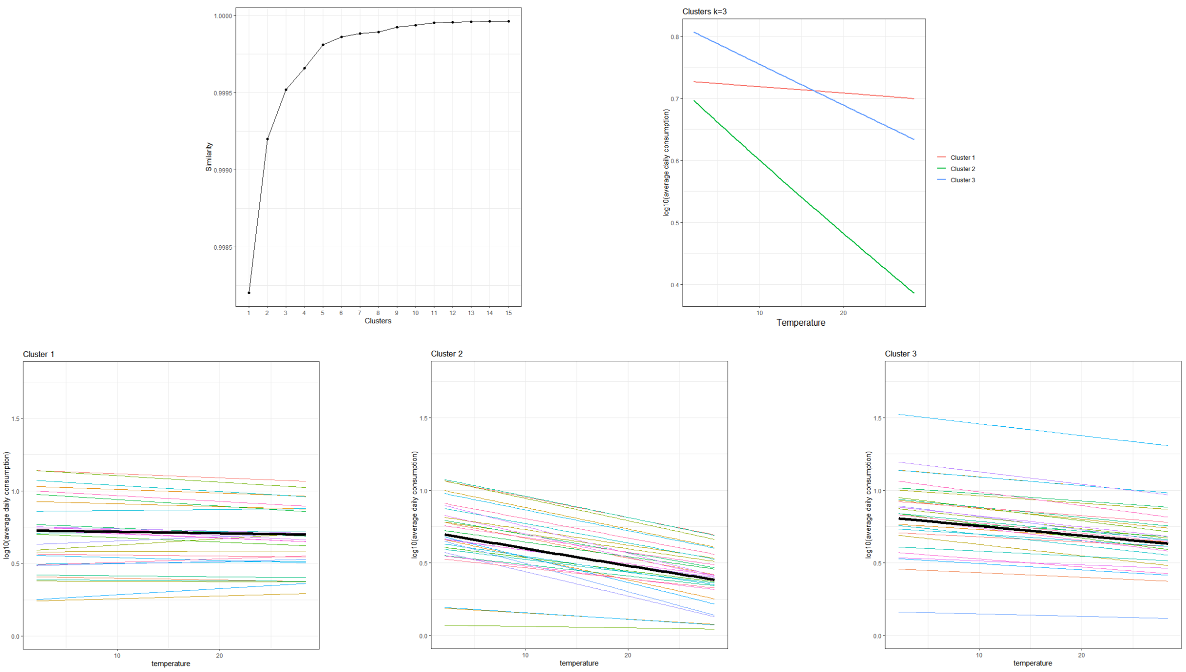

Figure 4.48.: Linear: Functional Clustering Pearson. In the top panels are presented the total within similarity for different number of cluster and the cluster's mean curves for k=3. The bottom panels represent the relative classified curves.

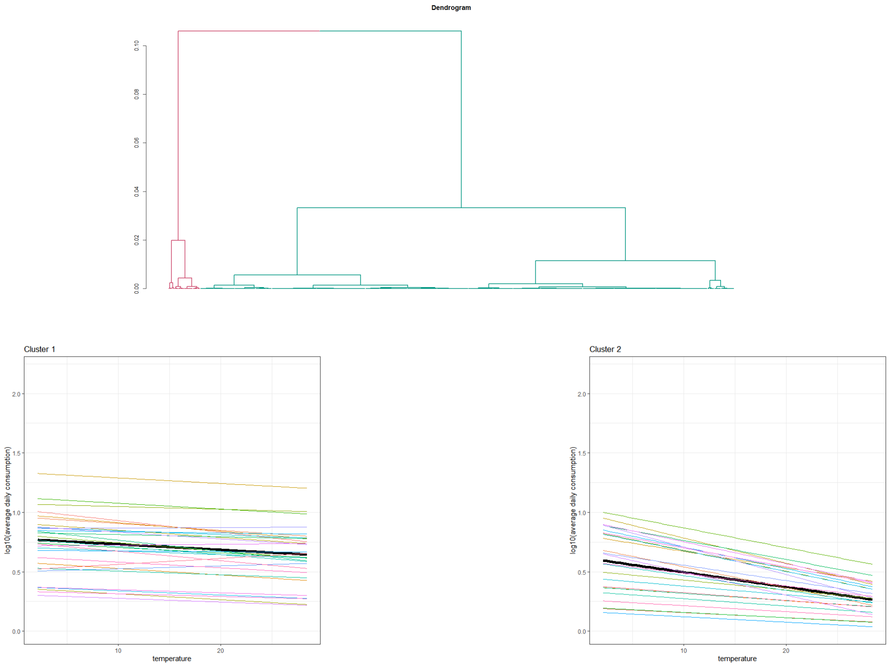

Figure 4.49.: Linear: The top panel represent the hierarchical dendrogram computed with Pearson similarity and complete linkage, k=2. The bottom panels represent the relative classified curves for each cluster.

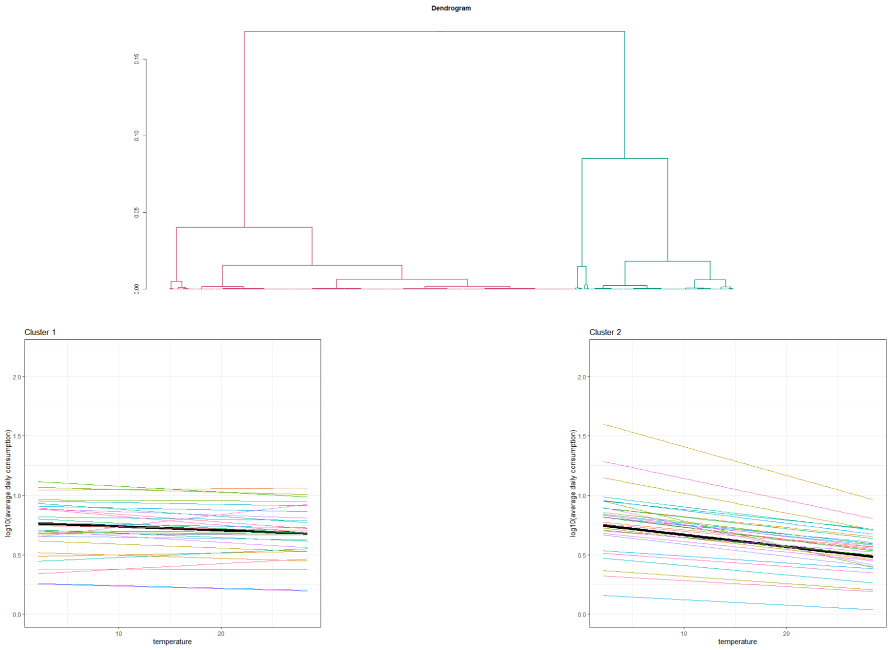

Figure 4.50.: Linear: The top panel represent the hierarchical dendrogram computed with Pearson similarity and Ward linkage, k=2. The bottom panels represent the relative classified curves for each cluster.

The ward linkage instead using k=2 (see Figure 4.50) divided the two behaviours more similar to the functional kmean assigning to the constant one 70% of the clients and 30% to the one with decreasing trend. If we look closely to the constant cluster we can note that there are lines with positive slope. These lines are correctly classified using k=4 (see Figure 4.51). They represent only the 5% of the clients and have a small slope. This explains why they were classified in the constant cluster before, that now represent the 67%. These reasons have led us to prefer the division in two clusters.

To reduce the complexity of the computation we decided to calculate for all the lines the intercept and the slope and use this new dataset to cluster them. From Figure 4.52 we can see a decreasing pattern, reasonable since the lines are in a fixed range and the increase of m corresponds to a decrease of the intercept.

We decide to maintain the two cluster division to perform clustering using the intercept and the slope standardized, Figure 4.53. k=2 is stable, and we can see also the division between the two groups in the scatterplot. In this case we obtained a constant cluster with a low mean consumption, 44% of the clients, and a decreasing cluster with an higher intercept (56%). As we have done with the parabolas we can use the density functions of the intercept and slope for tuning the parameters of climatic models (see Figure 4.46).

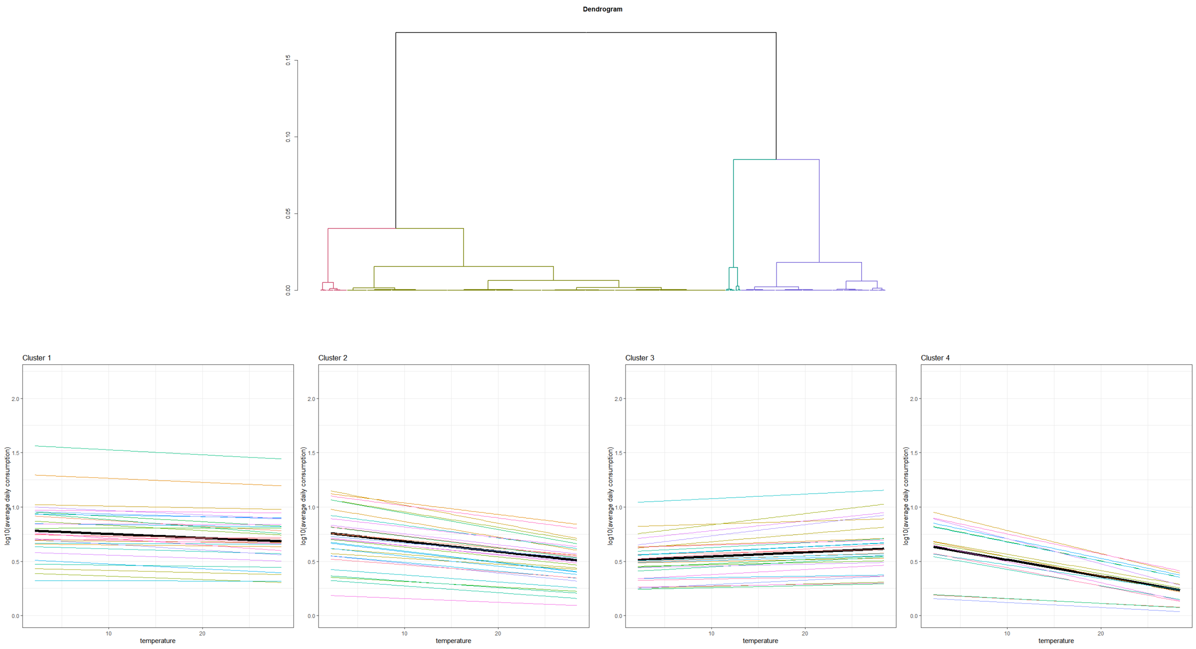

Figure 4.51.: Linear: The top panel represent the hierarchical dendrogram computed with Pearson similarity and Ward linkage, k=4. The bottom panels represent the relative classified curves for each cluster.

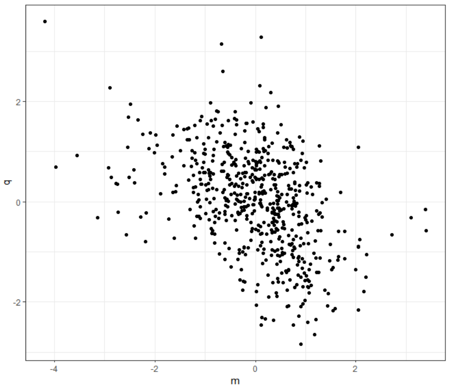

Figure 4.52.: Linear: clients intercept and slope standardized scatter plot.

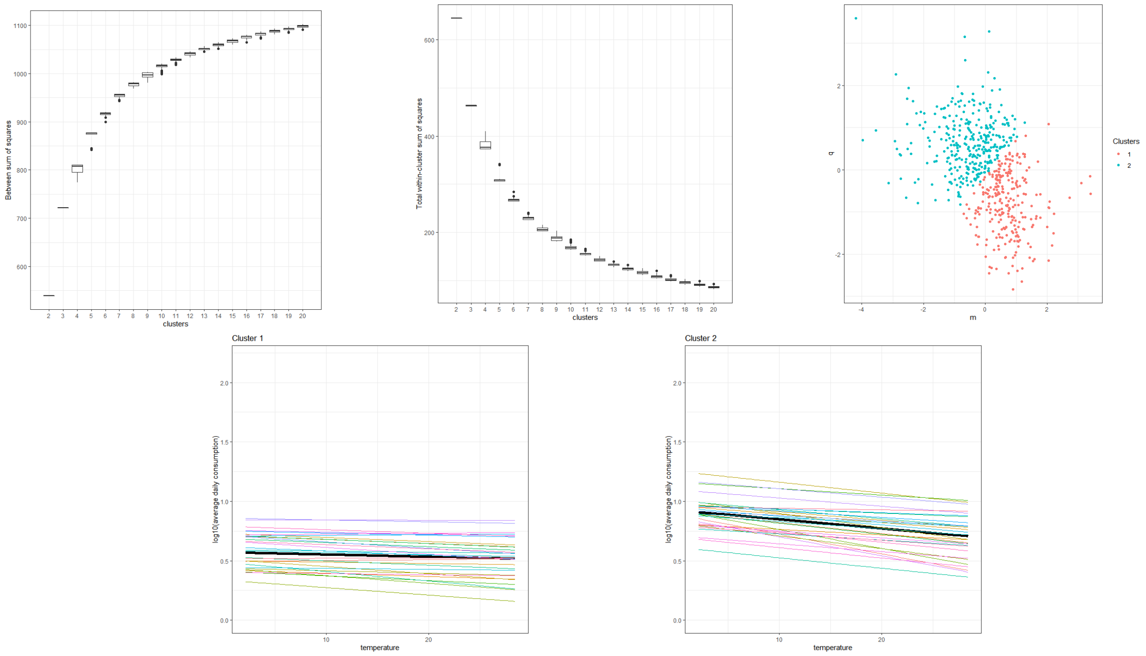

Figure 4.53.: Linear: In the top panel are represented the between (left), within (centre) total sum of squares of Kmean clustering using intercept and slope and the cluster scatterplot relative to k=2. The bottom panels represent the relative classified curves.

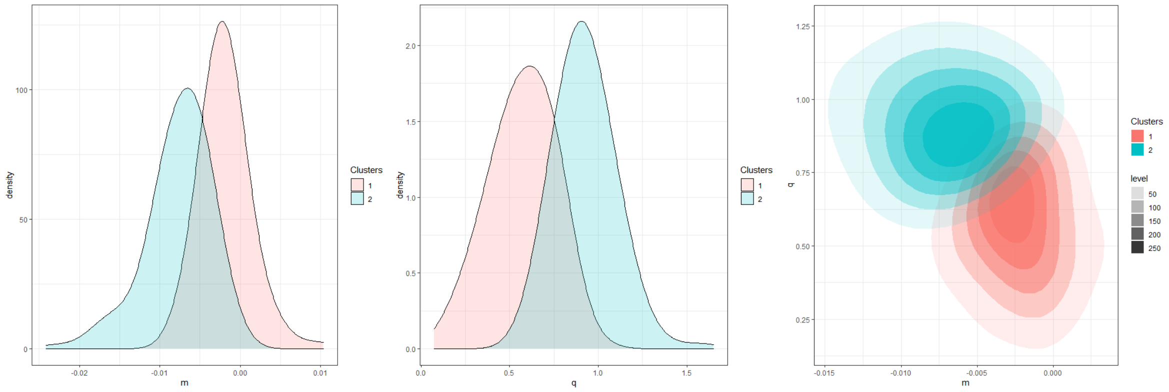

Figure 4.54.: Linear: marginal and joint pdfs for m, q relative to k=2 clusters identified by kmean.

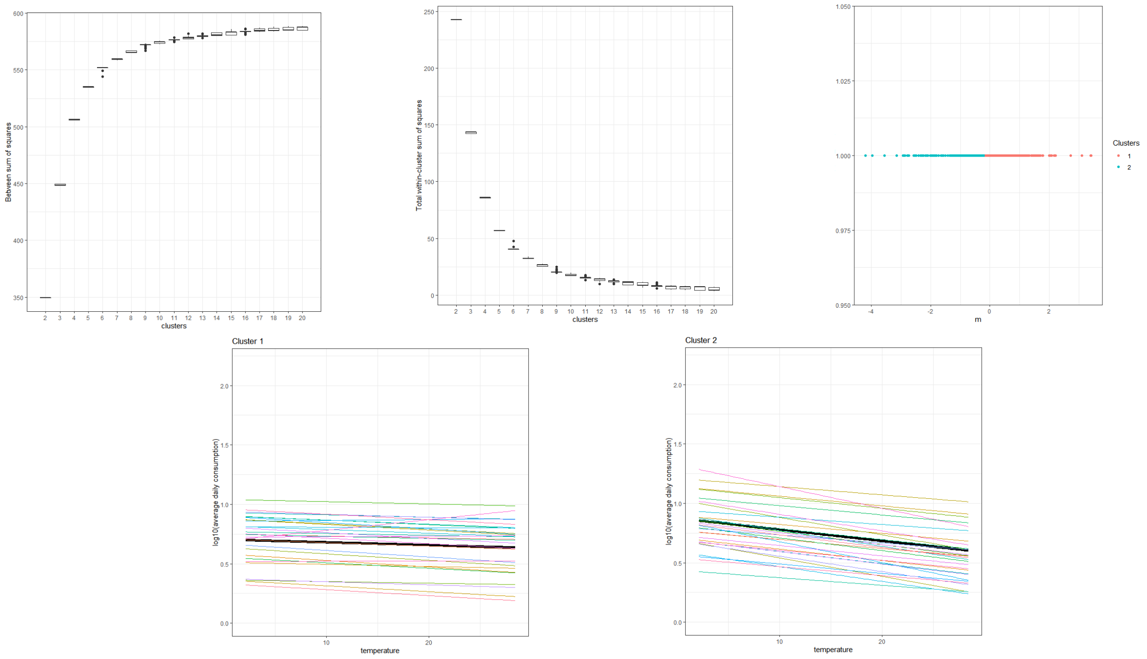

Figure 4.55.: Linear: In the top panel are represented the between (left), within (centre) total sum of squares of Kmean clustering using slope and the cluster scatterplot relative to k=2. The bottom panels represent the relative classified curves.

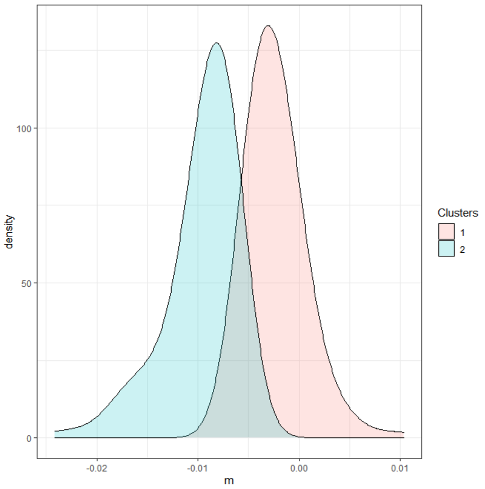

Figure 4.56.: Linear: density function of m relative to k=2 clusters identified by kmean.

Finally if we want to consider only the trend of the clients and not their mean consumption is useful to cluster using only the slope (see Figure 4.55). Also in this case we decided to maintain the two cluster division obtaining the constant cluster composed by 62% of clients and the decreasing one with 38%. We can see the density function of m relative to clusters in Figure 4.56.

5. Conclusions and results

The aim of this work was to propose a model able to estimate the mean monthly electricity consumption response to climate change in the residential sector and also identify the clients electricity temperature response curve to highlights common patterns in the population. Nowadays we are experiencing on our lives the importance of climate change and the impact that it has on our lives. Temperature increasing, accelerated by the growing use of fossil fuel in the last decades, is the climate phenomena that most influences the electricity sector and our economy. The presence of higher temperature will tend to increase the installation and the usage of air conditioning systems and will increase the need of electricity in the grid, produced for the majority part by fossil fuels, principal cause of climate change. To study this complex problem and predict future mitigation policies and their costs scientist developed Integrated Assessment Models that couple detailed models of energy system technologies with simplified economic and climate science models. The empirical assessment of the response electricity temperature curve is useful to obtain more precise impacts and evaluations of policies in specific regions.

In this context many study were performed in the field to analyse the sensitivity of consumption and the penetration of air conditioning. Moreover it is demonstrated that climate change has geographically distinct impacts base on regional level, so the major of study regard a particular state. Italian study on electricity sector and climate change were performed at different time scale analysing the impacts of demographic and climatic variables using regression model, mostly at global level or for specific cities, but not analysing the single client behaviour.

We performed our analysis focusing on the city of Milan using a monthly based dataset considering the time interval of 5 years (2015-2019). To handle the complex hierarchical structure of the data we used a non-parametric mixed effect model, developed by Rice and Wu [23] and used in LoMauro et al. [16], using a functional approach considering as statistical unit the monthly electricity temperature response curve. The fixed effect represented the mean behaviour of the city and random effect accounted the consumer and years effects. To filter out the effect of the different length of months we decided to use the average day monthly energy consumption, and for the reason explained before as a climatic variable the mean monthly temperature.

In the first model we tested, we succeeded to underline the general behaviour of the city of Milan, studying the difference between the years and the mean clients curves. The fixed effect confirmed the quadratic behaviour of consumption and temperature, that we found in the theory. Specifically we will have an increase in consumption of 1.2% for a temperature rise of 1°C. Year effect turn to be statistically significative but with very small variance allowing us to study only the mean clients response function. Analysing the client response function analytically we discover the presence of two macro-groups of clients: the first with quadratic positive behaviour composed by the 48% of clients and the other one with linear behaviour composed by 52% of clients. Consequently we divided the dataset and fitted two different model, fitting the same non parametric model for the first one and a linear mixed effect model for the second one. We obtained a major concavity for the mean response function of the first group that will cause an increasing of 6% of the consumption relative to an increment of °C, instead we observed a decreasing trend for the second group. Studying the mean clients curve we used different clustering methods to to identify the principal trend in the population. We were able to identify for the quadratic group three clusters of clients representing decreasing, constant and increasing trend. We noted that in this group all the clients will experience an increase in consumption. Instead for the linear group we identify 2 principal clusters,representing constant and decreasing behaviour. The method that performed better in functional analysis was kmeans using Pearson similarity. In both cases we we managed in reducing the complexity of functional clustering, obtaining the same one, standard statistics methods analysing variables that uniquely identify the functions: vertex position and concavity for parabolas and intercept and slope for the straight lines.

Unfortunately, we succeeded to perform our model only on the 5% of the dataset of Milan for computational issues. A direct future development would be to extend the model to the whole Italian dataset to have a complete understanding of the national behaviour and to analyse the difference response in the consumption studying the compositions of clusters in each municipalities. In the current work we decided, after finding the two subgroups of the population, to analyse them separately. To avoid a priori identification of the group to which an observation belongs, we propose to define a functional mixture model. Moreover, our analysis focused majorly on the clients behaviours, while it is also necessary to understand the differences between the annual mean consumptions using for example functional non parametric permutation tests. Finally, the electricity temperature response and composition of the cluster of clients found could be plugged in the Integrated Assessment Models to better analyse the cost of possible government policies in the energy field to cope with the increasing demand of electricity.

A. Appendix

A.1. Dataset Variable Specification

Consumption Dataset:

Variable	**Explanation**
utility_ customer_id	Customer number identification
comune_fornitura	Municipality of the clients
istcom	Istat code that uniquely identify a municipality
regione_fornitura	Region
avg_ day_month_consum	Average day monthly energy consumption [kWh/day]
month	Aggregated variable identifying month and year

Meteorological Dataset:

Variable	**Explanation**
tg	Daily mean temperature [°C]
tn	Daily minimum temperature [°C]
tx	Daily maximum temperature [°C]
rr	Daily precipitation sum [mm]
pp	Daily averaged sea level pressure [hPa]
qq	Daily mean global radiation [W/m2]

Monthly Weather Dataset :

Variable	**Explanation**
comune_fornitura	Municipality of the clients
istcom	Istat code that uniquely identify a municipality
Year	
Month	
MMXT	Monthly mean maximum temperature [°C]
MNTM	Monthly mean temperature [°C]
MMNT	Monthly mean minimum temperature [°C]
EMXT	Extreme maximum daily temperature observed in a month [°C]
EMNT	Extreme minimum daily temperature observed in a month [°C]
DT90	Number days in a month with maximum temperature >= 32.2 °C
DT32	Number days in a month with minimum temperature <= 0 °C
DT00	Number days in a month with minimum temperature <= -17.8 °C
DX32	Number days in a month with maximum temperature <= 0 °C
EMXP	Extreme maximum daily precipitation observed in a month [mm]
TPCP	Total precipitation in a month [mm]
DP10	Number of days with >= 25.4 mm of precipitation
DP01	Number of days with >= 2.54 mm of precipitation
DP05	Number of days with >= 12.7 mm of precipitation
MMPR	Monthly mean pressure [hPa]
MMRD	Monthly mean radiation [W/m]
GG	Monthly degree days [°C] ($\sum \max(0, 20 - T_e)$)

Bibliography

[1] M. Auffhammer, P. Baylis, and C. H. Hausman. "Climate change is projected to have severe impacts on the frequency and intensity of peak electricity demand across the United States". In: *Proceedings of the National Academy of Sciences* 114.8 (2017), pp. 1886–1891. DOI: `10.1073/pnas.1613193114`.

[2] D. Bates, M. Mächler, B. Bolker, and S. Walker. "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software, Articles* 67.1 (2015), pp. 1–48. ISSN: 1548-7660. DOI: `10.18637/jss.v067.i01`.

[3] D. M. Batesa and S. DebRoyb. "Linear mixed models and penalized least squares". In: *Journal of Multivariate Analysis* 91 (1 2004), pp. 1–17. DOI: `10.1016/j.jmva.2004.04.013`.

[4] V. Bianco, O. Manca, and S. Nardini. "Electricity consumption forecasting in Italy using linear regression models". In: *Energy* 34.9 (Sept. 2009), pp. 1413–1421. DOI: `10.1016/j.energy.2009.06.034`.

[5] J. Bonan, C. Cattaneo, G. D'Adda, and M. Tavoni. "Can we make social information programs more effective? The role of identity and values". In: (Nov. 2019), pp. 19–21. URL: `https://media.rff.org/documents/Social_information_programs_sep19.pdf`.

[6] J. Bonan, C. Cattaneo, G. d'Adda, and M. Tavoni. "Descriptive and injunctive norms complement eachother in promoting energy conservation through asocial information programme". In: *mimeo* (2020).

[7] M. Chen, K. T. Sanders, and G. A. Ban-Weiss. "A new method utilizing smart meter data for identifying the existence of air conditioning in residential homes". In: *Environmental Research Letters* 14.9 (Aug. 2019), p. 094004. DOI: `10.1088/1748-9326/ab35a8`.

[8] M. Christenson, H. Manz, and D. Gyalistras. "Climate warming impact on degree-days and building energy demand in Switzerland". In: *Energy Conversion and Management* 47.6 (Apr. 2006), pp. 671–686. DOI: `10.1016/j.enconman.2005.06.009`.

[9] Contribution of Working Groups I, II and III to the Fifth Assessment Report of theIntergovernmental Panel on Climate Change [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]. *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of theIntergovernmental Panel on Climate Change*. Tech. rep. IPCC, 2014.

[10] R. C. Cornes, G. van der Schrier, E. J. M. van den Besselaar, and P. D. Jones. "An Ensemble Version of the E-OBS Temperature and Precipitation Data Sets". In: *Journal of Geophysical Research: Atmospheres* 123.17 (2018), pp. 9391–9409. DOI: 10.1029/2017JD028200.

[11] L. Edwards, P. Stewart, J. MacDougall, and R. Helms. "A method for fitting regression splines with varying polynomial order in the linear mixed model". In: *Statistics in medicine* 25 (Feb. 2006), pp. 513–527. DOI: 10.1002/sim.2232.

[12] G. Franco and A. H. Sanstad. "Climate change and electricity demand in California". In: *Climatic Change* 87.S1 (Dec. 2007), pp. 139–151. DOI: 10.1007/s10584-007-9364-y.

[13] T. B. Johansson, N. Nakicenovic, A. Patwardhan, and L. Gomez-Echeverri, eds. *Global Energy Assessment (GEA)*. Cambridge University Press, 2009. DOI: 10.1017/cbo9780511793677.

[14] K. Karhunen. "Über linear Methoden in der Warscheinlichkeitsrechnung". In: *Annales Academiae Scientiorum Fennicae* 37 (1947), pp. 1–79.

[15] M. Loève. "Fonctions aléatoires de second ordre". In: *Comptes Rendus de l'Académie des Sciences, Série I: Mathématique* 220 (1945), p. 469.

[16] A. LoMauro, M. Romei, and S. G. et al. "Evolution of respiratory function in Duchenne muscular dystrophy from childhood to adulthood". In: *European Respiratory Journal* 51:1701418 (2018). DOI: 10.1183/13993003.01418-2017.

[17] N. L. Miller, K. Hayhoe, J. Jin, and M. Auffhammer. "Climate, Extreme Heat, and Electricity Demand in California". In: *Journal of Applied Meteorology and Climatology* 47.6 (June 2008), pp. 1834–1844. DOI: 10.1175/2007jamc1480.1.

[18] S. Mirasgedis, Y. Sarafidis, E. Georgopoulou, V. Kotroni, K. Lagouvardos, and D. Lalas. "Modeling framework for estimating impacts of climate change on electricity demand at regional level: Case of Greece". In: *Energy Conversion and Management* 48.5 (May 2007), pp. 1737–1750. DOI: 10.1016/j.enconman.2006.10.022.

[19] S. Mukherjee and N. Roshanak. "Climate sensitivity of end-use electricity consumption in the built environment: An application to the state of Florida, United States". In: *Energy* 128 (Apr. 2017), pp. 688–700. DOI: 10.1016/j.energy.2017.04.034.

[20] G. Pagliarini, C. Bonfiglio, and P. Vocale. "Outdoor temperature sensitivity of electricity consumption for space heating and cooling: An application to the city of Milan, North of Italy". In: *Energy and Buildings* 204 (Dec. 2019), p. 109512. DOI: 10.1016/j.enbuild.2019.109512.

[21] P. A. .-D. V. Raj, M. Sudhakaran, and P. P.-D.-A. Raj. "Estimation of Standby Power Consumption for Typical Appliances". In: *Journal of Engineering Science and Technology Review* 2.1 (June 2009), pp. 141–144. DOI: 10.25103/jestr.021.26.

[22] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. 2nd Ed. Springer, 2005.

[23] J. A. Rice and C. O. Wu. "Nonparametric Mixed Effects Models for Unequally Sampled Noisy Curves". In: *Biometrics* 57.1 (2001), pp. 253–259. DOI: 10.1111/j.0006-341x.2001.00253.x.

[24] L. M. Sangalli, P. Secchi, S. Vantini, and A. Veneziani. "A Case Study in Exploratory Functional Data Analysis: Geometrical Features of the Internal Carotid Artery". In: *Journal of the American Statistical Association* 104.485 (2009), pp. 37–48. DOI: 10.1198/jasa.2009.0002.

[25] L. M. Sangalli, P. Secchi, S. Vantini, and V. Vitelli. "K-mean Alignment for Curve Clustering". In: *Computational Statistics and Data Analysis* 54 (5 2010), pp. 1219–1233. DOI: 10.1016/j.csda.2009.12.008.