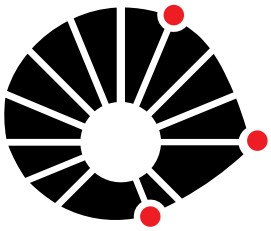


UNICAMP

UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Matemática, Estatística e Computação Científica

GABRIEL FRANCO DE SOUZA

Aggregated functional data model applied on clustering and disaggregation of electrical load profiles

Modelo de dados funcionais agregados aplicado em separação e agrupamento de consumo de energia elétrica

Campinas

2021

Gabriel Franco de Souza

Aggregated functional data model applied on clustering and disaggregation of electrical load profiles

Modelo de dados funcionais agregados aplicado em separação e agrupamento de consumo de energia elétrica

Tese apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutor em Estatística.

Thesis presented to the Institute of Mathematics, Statistics and Scientific Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor in Statistics.

Supervisor: Nancy Lopes Garcia

Co-supervisor: Camila Pedroso Estevam de Souza

Este exemplar corresponde à versão final da Tese defendida pelo aluno Gabriel Franco de Souza e orientada pela Profa. Dra. Nancy Lopes Garcia.

Campinas

2021

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Silvania Renata de Jesus Ribeiro - CRB 8/6592

Franco, Gabriel, 1989-
F848a Aggregated functional data model applied on clustering and disaggregation of electrical load profiles / Gabriel Franco de Souza. – Campinas, SP : [s.n.], 2021.

Orientador: Nancy Lopes Garcia.
Coorientador: Camila Pedroso Estevam de Souza.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Estatística não paramétrica. 2. Teoria do spline. 3. Energia elétrica - Consumo. 4. Processos gaussianos. 5. Análise multivariada. I. Garcia, Nancy Lopes, 1964-. II. Souza, Camila Pedroso Estevam de, 1982-. III. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Modelo de dados funcionais agregados aplicado em separação e agrupamento de consumo de energia elétrica
Palavras-chave em inglês:
Nonparametric statistics
Spline theory
Electric power consumption
Gaussian process
Multivariate analysis
Área de concentração: Estatística
Titulação: Doutor em Estatística
Banca examinadora:
Nancy Lopes Garcia [Orientador]
Aluísio de Souza Pinheiro
Clarice Garcia Borges Demétrio
Alexandra Mello Schmidt
Nancy Elizabeth Heckman
Data de defesa: 11-01-2021
Programa de Pós-Graduação: Estatística

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-9140-1840
- Currículo Lattes do autor: http://lattes.cnpq.br/9730840432900175

Tese de Doutorado defendida em 11 de janeiro de 2021 e aprovada

pela banca examinadora composta pelos Profs. Drs.

Prof(a). Dr(a). NANCY LOPES GARCIA

Prof(a). Dr(a). ALUÍSIO DE SOUZA PINHEIRO

Prof(a). Dr(a). CLARICE GARCIA BORGES DEMÉTRIO

Prof(a). Dr(a). ALEXANDRA MELLO SCHMIDT

Prof(a). Dr(a). NANCY ELIZABETH HECKMAN

A Ata da Defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria de Pós-Graduação do Instituto de Matemática, Estatística e Computação Científica.

Aos meus pais, os que tornaram possível meu sonho de ser cientista

Acknowledgements

A special thanks to professor Gavin Shaddick whom provided the dataset used in this thesis. The following thanks will be addressed in Portuguese.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Agradeço ao excelente trabalho de orientação das professoras Nancy Garcia e Camila de Souza. A paciência, atenção e dedicação dispendidas neste trabalho foram essenciais para guiar essa pesquisa tão interessante e iluminar meus caminhos nesta jornada. Mesmo com muitas outras tarefas administrativas e didáticas, além do cenário de pandemia no ano de 2020, jamais me faltou amparo, apoio e mentoria. São dois exemplos de mentoras que quero guardar para o resto da minha vida e sonho um dia conseguir exercer tal função com tamanha propriedade. Espero que cada pessoa que eu possa vir a orientar no futuro tenha a mesma segurança e confiança que eu tive durante esse período sob a tutela dessas duas excelentes profissionais e pessoas maravilhosas.

Em especial, dedico essa tese e toda minha carreira aos meus pais. Tive o privilégio de contar com o apoio e amor incondicionais que apenas essas duas pessoas são capazes de prover. Agradeço desde o apoio e investimento no período de vestibular, a confiança e apoio ao me ver saindo de casa rumo à Campinas e principalmente o apoio à minha opção de carreira. A todo momento é possível sentir a admiração e o orgulho de ambos em cada conquista e espero que essa seja mais uma para se orgulharem. Newton disse "sobre os ombros de gigantes", eu digo que os gigantes que tornaram possível visualizar e buscar esse horizonte foram meus pais.

Também foi fundamental o apoio dos ciclos de amizade que fizeram parte desta jornada. Amigos de longa data que me apoiam desde a graduação como Diego e Joyce Fujiwara e Renato Reis. Destes mais antigos, tenho um obrigado especial ao Darcy Camargo, meu amigo, meu colega e meu irmão. Além de ser uma ótima fonte de conhecimento para tirar minhas dúvidas em matemática e probabilidade, foi e é companheiro de todas as horas, uma pessoa de confiança e a quem dedico boa parte do meu sucesso. Ainda pelo IMECC, agradeço ao acolhimento de todas as pessoas maravilhosas que circulavam pela sala 1B do prédio anexo da pós-graduação. Era sempre uma alegria subir as escadas para receber e doar alegria ao redor de um café.

Aproveito para dedicar uma parte desses agradecimentos a pessoas que me inspiraram ao longo desses anos a ser um melhor indivíduo e profissional, seja diretamente ou apenas pelo fato de permitir que eu pudesse acompanhar suas carreiras: obrigado a Ligia Steinberg, Charles Almeida, Leonora Cardani, Helen Silva, Kássia Mitally, Victor

Freguglia, Ana Fontenelle e Thays Gomes.

É com alegria que gostaria de também agradecer ao Grupo Ginástico Unicamp, onde tenho amizades que guardo com muito carinho e onde sempre fui muito bem recebido e acolhido. O grupo proporcionou experiências além da grade curricular que vou guardar sempre com muito carinho e me apresentou a pessoas que hoje fazem parte da minha vida além da Unicamp.

Por fim, um imenso agradecimento a todos os professores e professoras que passaram pela minha vida e também a todos e todas do departamento de estatística do Instituto de Matemática, Estatística e Computação Científica, nosso IMECC. Em especial, minhas primeiras orientadoras Mariana Motta e Hildete Pinheiro, meu orientador de mestrado Aluísio Pinheiro, à professora Verónica González-López que me recebeu de braços abertos no curso de estatística. A tese que aqui escrevo é também fruto de um esforço coletivo de todos vocês.

E agradeço a você que está lendo esta tese. Seu interesse nesta pesquisa é minha maior recompensa.

Resumo

O alerta das mudanças climáticas fizeram com que organizações ao redor do mundo dispendessem esforços em programas de energia eficientes. Neste cenário, entender a demanda de energia elétrica dos consumidores finais tem papel fundamental para o planejamento da rede de distruibuição elétrica e avaliar a necessidade de novas usinas de energia. Apesar do crescimento das tecnologias de medidores inteligentes, observar o consumo individual ainda é caro. Por outro lado, curvas de consumo agregadas geralmente são disponíveis em subestações de energia. A metodologia proposta separa o consumo de energia agregado observado nessas subestações em curvas médias de consumo estimadas, chamadas de *curvas típicas*, para cada tipo de consumidor abastecido. Nossa abordagem supõe que cada curva típica de consumidor segue um Processo Gaussiano, cuja média é dada pela curva típica e modelada em termos de covariáveis funcionais e escalares. Além disso, levando em consideração as diferenças entre subestações devido a fatores externos, propomos um modelo de agrupamento de subestações baseado na similaridade de suas curvas típicas e estrutura de covariância. Para verificar o desempenho do modelo, a metodologia é testada em uma série de experimentos sob oito cenários simulados diferentes e aplicada a um conjunto de dados reais de monitoramento de consumo de energia elétrica em subestações no Reino Unido.

Palavras-chave: Separação de sinal, modelo funcional, modelo funcional agregado, processos gaussianos, expansão por funções base.

Abstract

The climate change alert pushed organizations worldwide to devote efforts to energy efficiency programs. In this scenario, understanding electrical demand at the consumer level plays an important role in planning the distribution of electrical networks and evaluating the need for construction of new power plants. Despite the growth of smart meter technology, observing individual consumption loads is still expensive. On the other hand, aggregated load curves are normally available at the substation level. The proposed methodology separates substation aggregated loads into estimated mean consumption curves, called typical curves, one for each supplied customer type. The approach proposed here assumes that each customer load curve follows a Gaussian process with a mean given by the typical curve, which can be modelled in terms of explanatory scalar and functional covariates. In addition, to account for the difference among substations due to external factors, a model-based clustering approach for substations is proposed based on the similarity of their consumers' typical curves and covariance structures. To assess model performance, the methodology is tested in a series of experiments under eight simulated scenarios and applied to a real substation load monitoring dataset from the United Kingdom.

Keywords: Blind source separation, functional aggregated model, Gaussian process, basis function expansion.

List of symbols

i	Replicates (date)
j	Groups (substations)
c	Subject type
m	Subject (customer)
t	Time
b	Cluster
$m_{j\cdot}$	Total number of subjects at group j
m_{jc}	Number of subjects of type c at group j
$\alpha_c(t)$	Mean signal (Typical curve) of subject type c
$W_{ijmc}(t)$	Separated signal (individual consumption) of the i-th replicate of group j for subject m of type c
$Y_{ij}(t)$	Aggregated signal (aggregated consumption) of replicate i for group j

Contents

1 Introduction

1.1 Energy context in Brazil and worldwide

In 2017, 62.5% of the Brazilian energy matrix consisted of hydroelectric power generation, mainly located in the South, Southeast and Northeast regions, where the greatest metropolitan areas are located, followed by thermoelectric power, with 17.1% (EMPRESA DE PESQUISA ENERGÉTICA – EPE, 2018). As an alternative to these energy sources, a solar smart grid in the semiarid Brazilian Northeast is suggested as a strategy to explore the large amount of solar radiation available and improve the economical potential of the region (NOBRE et al., 2019). On the other hand, with smaller urban areas, undersized industrial parks, and consequently lower electrical energy demand, the Northeast region presents a distribution network shortage (NOBRE et al., 2019). However, with Brazilian electrical energy consumption expected to grow exponentially by 2030, a national rationing plan may be implemented if no programs for efficient electrical usage are executed (MINISTÉRIO DE MINAS E ENERGIA, 2007). Therefore, it is of great importance to have statistical and computational tools to understand demand at the individual customer level to analyze and monitor the electrical load profiles of customers and produce metrics for data-driven decisions, such as new power plants and network distribution redesign.

Around the world, the United States (WILLIAMS et al., 2012), Canada (WEAVER et al., 2007), the United Kingdom (BRISTOW et al., 2008) and several other countries have committed to reduce greenhouse gas emissions by at least 80% by no later than 2050, in comparison to 1990. Denmark and Germany are leaders in feed-in tariffs to accelerate investment in renewable energy technologies and are closest to meeting the emission goal (LIPP, 2007). Reusable programs and carbon sequestration are attractive options to reduce industrial carbon emissions (ALLWOOD; CULLEN; MILFORD, 2010). At the commercial and domestic levels, initiatives like building energy modelling (SOUZA; HECKMAN; XU, 2017) and prediction (ZHAO; MAGOULÈS, 2012) are also important tools contributing to efficient energy consumption.

Increasing network distribution and the rise of smart grids have drawn attention to load profile monitoring (WANG et al., 2015). Multiple articles have been published in the literature proposing clustering techniques to segment customers and reduce variability (PRAHASTONO; KING; OZVEREN, 2007; LI et al., 2015a). Efforts are also underway to achieve short-term load forecasting using machine learning (SOUSA; JORGE; NEVES, 2014) and deep learning methods (SHI; XU; LI, 2017). Although load profile modeling is an important task to comprehend electrical demand variability, it does not provide

information on the customer level like smart meters (D'OCA; CORGNATI; BUSO, 2014; GOUVEIA; SEIXAS, 2016), appliances monitoring (HART, 1992; ARGHIRA et al., 2012) and disaggregation methods (SCHIRMER; MPORAS; PARASKEVAS, 2019).

Understanding individual customer consumption behaviour is essential to comprehend electrical energy demand and consequently to take action to reduce substation load, such as educational programs and off-peak tariff policies, or even to consider bigger projects like new power plants and network distribution redesign. Solutions such as the aggregated data model proposed in this thesis provide estimated typical curves for each customer type based only on aggregated data and enhance comprehension of the covariance structure to assess data uncertainty. Regions like the Brazilian Northeast may not have smart meters to provide data on the individual level, and disaggregation methods on substation electrical loads are interesting approaches to help authorities understand electrical demand without major investments in individual load monitoring.

1.2 UK electrical data

The dataset analyzed in this thesis is composed of electrical load profile curves from substations in the United Kingdom serving residential customers of two types. Customers are labeled in two categories: unrestricted (C1) and "Economy 7" (C2) domestic customers, with the latter referring to a program with cheaper electrical tariffs during the off-peak period. Usually, researchers do not have access to individual residential electrical load, but only to aggregated consumption at the substation level. Furthermore, the *market* of each substation, that is, the number of C1 and C2 residences, is also known. The first question to arise in this scenario is the following: "is it possible to estimate the average consumption curve for each of the two types of customers based on the substation aggregated information and its market?" The answer is yes, but it is essential to observe more substations with different numbers of customers to create the variability required to separate the aggregated consumption into customer-type specific curves, which are also known as *typical curves*.

Figure 1 shows an example of aggregated load profiles from four substations in the United Kingdom serving customers of types C1 and C2 as described above. The curves represent the electrical load in KW/h over 61 working days from January 3 to March 30, 2013. Different shapes and scales are observed because substations A to D have distinct customer distributions in their known markets, as shown in Table 1. The divergent peak at substation D probably reflecs its majority of type C2 customers and provides evidence of possibly different typical curves for each type of customer.

It is possible to obtain more information about the dataset, such as temperature and geographical location, to account for the variability of the electrical load profile. For

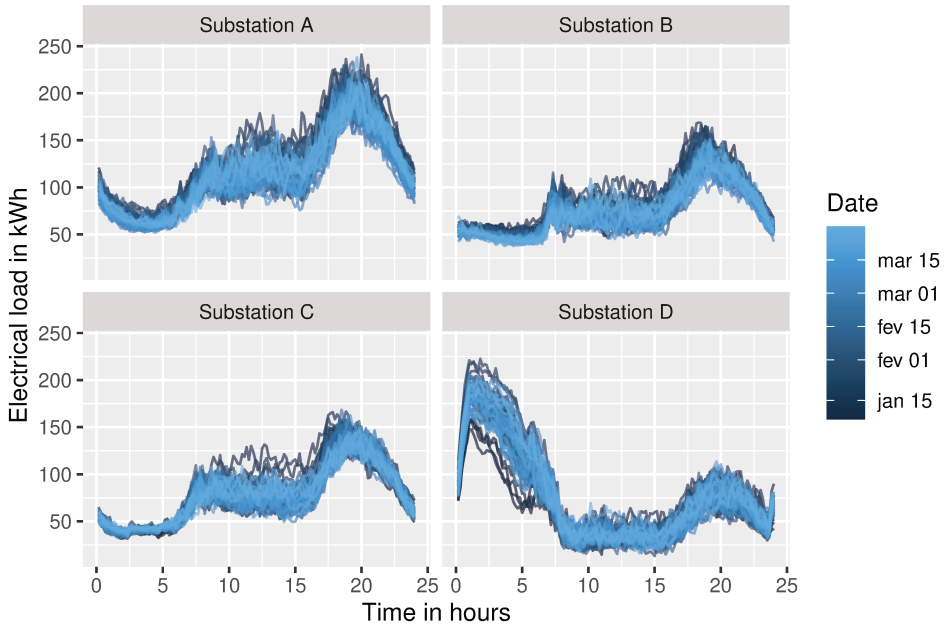


Figure 1 – Daily electrical load profiles at every 10 minutes for four substations where each curve corresponds to one out of 61 winter days in the UK

example, substations A and B show darker curves for higher loads, indicating a potential temperature effect because the dark tones represent the beginning of winter.

Another question arises from observing data from many substations: "Are the typical customer curves the same for every substation?" For example, can it be expected that an unrestricted domestic residence in an urban area like London will have the same load profile as a house located in the countryside? The answer is probably no and this is the motivation to introduce a latent variable to cluster substations based on their disaggregated typical curves and covariance structures.

1.3 Modelling aggregated data

Several widely used machine learning and regression methods were used to study energy disaggregation such as artificial neural networks (LIN; TSAI, 2015; HOSSEINI et al., 2017), random forests (BILSKI; WINIECKI, 2017; SCHIRMER; MPORAS, 2019), Support Vector Machines (BASU et al., 2014; SCHIRMER; MPORAS; SHEIKH-AKBARI,

Table 1 – Market distribution of the four substations presented in Figure 1.

Substation	C1 (Standard)	C2 (Economy 7)
A	228	3
B	146	5
C	151	5
D	21	88

2020), wavelet component analysis (ZHU; LU, 2014) and K-Nearest-Neighbours (KIM et al., 2014). Reviews and comparisons of multiple statistical methods for energy disaggregation are available in literature (SCHIRMER; MPORAS; PARASKEVAS, 2019).

The approach to separate the aggregated load is related to *Blind Source Separation*, where a single channel source is separated into multiple channels using a set of restrictions and conditions (CARDOSO, 1998).Unlike blind source separation, the model proposed in this work provides estimates for covariance structure and accommodates scalar and functional explanatory variables. The proposed model approach offers flexibility and desirable inference properties such as consistency and lack of bias. The estimated covariance structure provides information on load measurement relationships over time and load variability. Finally, the latent variable approach makes it possible to group electrical substations by their common typical curves.

The family of aggregated data models considered in this work was first proposed by (DIAS; GARCIA; MARTARELLI, 2009). Using the observed electrical load from energy transformers and their market information, the authors composed a non-parametric model to estimate the typical consumption curve of customers in the city of Campinas, Brazil, using basis function expansion and the sample covariance matrix as the model covariance structure. More sophisticated structures, under the Bayesian paradigm, were proposed to study the transformer load curves (DIAS; GARCIA; SCHMIDT, 2013; DIAS et al., 2015). Later, considering the market as random, the aggregated data model could identify errors in energy customer classification (LENZI et al., 2017).

In this work, a generalization of the aggregated data model described above is proposed. The novel approach performs the disaggregation task by assuming a Gaussian process with mean functional response as an aggregated linear combination of the market, typical customer curves, and explanatory variables. Additional functional variables are incorporated in the typical curve model to comprehend, for example, the impact of temperature on customer load profile. A model-based clustering approach is also proposed to group energy substations with similar disaggregated curves using a mixture of Gaussian processes (SHI; MURRAY-SMITH; TITTERINGTON, 2005; SHI; CHOI, 2011; TRESP, 2001) estimated by the Expectation-Maximization algorithm (DEMPSTER; LAIRD; RUBIN, 1977; MCLACHLAN; KRISHNAN, 2007). Finally, structured covariance functionals are proposed to model load variability and correlation decay over time.

1.4 Code availability

The methodology proposed in this thesis is implemented as an R package called `aggrmodel`, which is currently available online as an alpha version at the GitHub repository github.com/gabrielfranco89/aggrmodel. The repository contains the functions

used to perform all the analyses conducted in the thesis as well as examples to illustrate package usability, which can be easily explored by the reader.

1.5 Thesis overview

This thesis is organized as follows: Chapter 2 describes the methodology by introducing the simple aggregated data model and expanding it to the full model, where explanatory variables and additional functional components are incorporated, and describes the latent variable modelling to cluster substations based on their typical curves. Chapter 3 is made up of two simulation studies to evaluate the performance of the full-model approach with explanatory variables, a temperature component, and the model-based clustering approach to group substations. Chapter 4 presents an application to real data consisting of electrical load profile data from substations in the United Kingdom.

2 Methods

This section describes the statistical methods used to fit the aggregated data model and is organized as follows: Section 2.1 introduces the simplest aggregated data model followed by Section 2.2 which presents the main features needed to construct the full model with explanatory variables, surface response and different covariance structures. Section 2.3 describes the proposed clustering analysis followed by the estimation procedures in Section 2.4. Section 2.5 concludes this chapter with model diagnosis. Well–established concepts such as B-Spline expansion (RAMSAY; SILVERMAN, 2005; BOOR et al., 1978) and Gaussian Processes (SHI; CHOI, 2011) can be found in the literature and will not be described in detail in this thesis.

2.1 Modelling aggregated data: an introduction

The aggregated data model is here introduced in its simplest form, as in in (DIAS; GARCIA; MARTARELLI, 2009). The observed data consist of aggregated energy consumption curves for J substations observed over I days. Each substation constitutes a distinct market with C types of consumers, – e.g., residential, industrial and business. Each aggregated curve is the sum of all individual consumer curves served by that substation. Suppose that $W_{ijcm}(t)$, the unobserved energy consumption of customer m of type c at time t from substation j on day i, can be represented as

$$W_{ijcm}(t) = \alpha_c(t) + \varepsilon_{ijmc}(t), \tag{2.1}$$

with $\alpha_c(\cdot)$ being the typical curve of a customer of type c and $\varepsilon_{ijmc}(t)$ a Gaussian Process (GP) with zero mean and covariance structure $\Psi_c(\cdot,\cdot)$ to be detailed in Section 2.2.1.

Let $Y_{ij}(t)$ be the observable aggregated energy consumption at substation j, day i and time t. $Y_{ij}(t)$ can then be represented as the sum of individual customer curves, that is,

$$\begin{aligned} Y_{ij}(t) &= \sum_{c=1}^{C}\sum_{m=1}^{m_{jc}} W_{ijcm}(t) \\ &= \sum_{c=1}^{C}\sum_{m=1}^{m_{jc}} \alpha_c(t) + \varepsilon_{ijmc}(t) \\ &= \sum_{c=1}^{C} m_{jc}\alpha_c(t) + \varepsilon_{ij}(t), \end{aligned} \tag{2.2}$$

with m_{jc} being the number of customers of type c in substation j, $\varepsilon_{ij}(\cdot,\cdot) \sim GP\big(0, \Sigma_j(\cdot,\cdot)\big)$, where, assuming independence among individual customers, the covariance structure $\mathbf{\Sigma}_j(\cdot,\cdot)$

can be written as

$$\Sigma_j(s,t) = \sum_{c=1}^{C} m_{jc}\Psi_c(s,t). \tag{2.3}$$

2.1.1 Typical curve basis function expansion

The mean component $\alpha_c(\cdot)$ in Equation (2.1) represents the typical curve of customers of type c and can be modelled using a basis function expansion as

$$\alpha_c(t) = \sum_{k=1}^{K} \phi_k(t)\beta_{ck} = \boldsymbol{\phi}(t)\boldsymbol{\beta}_c, \tag{2.4}$$

where $\boldsymbol{\beta}_c \in \mathbb{R}^K$ is the vector of expansion parameters or coefficients and $\phi_k(\cdot)$ the k-th basis function, which can be B-Splines, Fourier transforms, wavelets or a polynomial basis. In this thesis, it is assumed that the typical curves belong to a Sobolev space and that they can be well approximated by uniform B-splines. For more details, see (REIF, 1997). As in previous studies using aggregated data analysis (DIAS; GARCIA; MARTARELLI, 2009; DIAS; GARCIA; SCHMIDT, 2013; DIAS et al., 2015; LENZI et al., 2017), a cubic B-Spline basis is used with the assumption that the number of basis functions K is known. Several studies have provided methods to select the best number of basis functions, such as (DEVORE; PETROVA; TEMLYAKOV, 2003), which extend the results of (DONOHO, 1993) to L_p, with $p \neq 2$; (DIAS; GARCIA, 2007) which provided a consistent estimate of the optimal number of basis functions by minimizing a penalized proxy of the Kullback–Leibler distance; (KOHN; MARRON; YAU, 2000) for wavelets and fourier basis selection and (LUO; WAHBA, 1997; DIAS, 1998) for adaptive methods. Ideally, K should be large enough to capture function details, but still far from interpolating.

2.2 Full aggregated data model

In the electrical energy consumption example, suppose that the typical curve depends not only on the time t, but also on functional covariates such as the air temperature on day i, or in other words, it can be written as a function $\alpha_{ic} : \mathbb{R}^2 \to \mathbb{R}_+$. It is possible to write, for example, the functionals $u(t) = t$ as the time and $v_i(t)$ as the air temperature on day i at time t. In this case, $\alpha_{ij}(\cdot)$ can be expanded as a tensor product of basis functions as follows:

$$\alpha_{ic}(t) = \alpha_{ic}\big(u(t), v_i(t)\big) = \sum_{k=1}^{K}\sum_{l=1}^{L} \phi_k\big(u(t)\big)\varphi_l\big(v_i(t)\big)\,\beta_{lkc}, \tag{2.5}$$

$$= \boldsymbol{\phi}_i(t)\boldsymbol{\beta}_c \tag{2.6}$$

where $\phi(\cdot)$ and $\varphi(\cdot)$ are basis functions and β_{lkc} are expansion parameters.

In addition, suppose that P explanatory variables related to the substations are known, namely $D_{ij1}, \ldots, D_{ijP}$. Therefore, these variables can be incorporated to write the full aggregated data model as:

$$Y_{ij}(t) = \left(\sum_{c=1}^{C} m_{jc}\alpha_{ic}\big(u(t), v_i(t)\big)\right) + D_{ij1}\gamma_1 + \cdots + D_{ijP}\gamma_P + \varepsilon_{ij}(t) \tag{2.7}$$

$$= \left(\sum_{c=1}^{C}\sum_{k=1}^{K}\sum_{l=1}^{L} m_{jc}\phi_k\big(u(t)\big)\varphi_l\big(v_i(t)\big)\,\beta_{lkc}\right) + \mathbf{D}_{ij}\boldsymbol{\gamma} + \varepsilon_{ij}(t). \tag{2.8}$$

In vector representation,

$$Y_{ij}(t) = \sum_{c=1}^{C} m_{jc}\boldsymbol{\phi}_i(t)\boldsymbol{\beta}_c + \mathbf{D}_{ij}\boldsymbol{\gamma} + \varepsilon_{ij}(t) \tag{2.9}$$

with $\boldsymbol{\gamma} \in \mathbb{R}^P$ being the parameters corresponding to the substation explanatory variables in $\mathbf{D}_{ij}$ and

$$\boldsymbol{\phi}_i(t)^\intercal = \begin{pmatrix} \phi_1\big(u(t)\big)\varphi_1\big(v_i(t)\big) \\ \phi_1\big(u(t)\big)\varphi_2\big(v_i(t)\big) \\ \vdots \\ \phi_1\big(u(t)\big)\varphi_L\big(v_i(t)\big) \\ \phi_2\big(u(t)\big)\varphi_1\big(v_i(t)\big) \\ \vdots \\ \phi_2\big(u(t)\big)\varphi_2\big(v_i(t)\big) \\ \vdots \\ \phi_2\big(u(t)\big)\varphi_L\big(v_i(t)\big) \\ \vdots \\ \phi_K\big(u(t)\big)\varphi_L\big(v_i(t)\big) \end{pmatrix} \quad \text{and } \boldsymbol{\beta}_c = \begin{pmatrix} \beta_{11c} \\ \beta_{12c} \\ \vdots \\ \beta_{1Lc} \\ \beta_{21c} \\ \beta_{22c} \\ \vdots \\ \beta_{2Lc} \\ \vdots \\ \beta_{KLc} \end{pmatrix}. \tag{2.10}$$

Note that the simple model is nested inside the full aggregated data model, because it represents the case when temperature and explanatory variables have no effect on the typical curve.

2.2.1 Covariance structures

Let $\varepsilon_{ijmc}(\cdot)$ be the Gaussian Process introduced in Equation (2.1) with zero mean and covariance structure defined by the functional $\Psi_c(\cdot,\cdot)$. This presentation will use the following decomposition (DIAS; GARCIA; SCHMIDT, 2013):

$$\Psi_c(s,t) = \mathrm{Cov}\Big(\varepsilon_{ijmc}(s), \varepsilon_{ijmc}(t)\Big)$$

$$= \eta_c(s)\,\rho_c(s,t)\,\eta_c(t), \tag{2.11}$$

where $\eta_c(\cdot)$ and $\rho_c(\cdot,\cdot)$ are variance and correlation functionals, respectively. To guarantee the positive definiteness of the Gaussian Process covariance structure, $\rho_c(\cdot,\cdot)$ must be a proper positively defined correlation functional. The following sections describe the different nested forms of $\eta_c(\cdot)$ and $\rho_c(\cdot,\cdot)$.

2.2.1.1 Variance functionals

The variance functional $\eta_c(\cdot)$ describes the variability of customers of type c over time. The identifiability of the model is guaranteed only if $\eta_c(\cdot)$ is positive, otherwise any function multiplied by -1 is also an optimal solution. Hence, the results of (RAMSAY; SILVERMAN, 2005) can be used, and the variance function $\eta_c(\cdot)$ can be written as:

$$\eta_c(\cdot) = \exp\left\{\sum_{k=1}^{K'} \phi_k^{\eta}(\cdot)\beta_{kc}^{\star}\right\}. \tag{2.12}$$

Furthermore, nested functional variances can be created based on a different parametrization of the expansion coefficients of Equation (2.12) (DIAS; GARCIA; SCHMIDT, 2013). If

$$\sigma_c^{\star} = \frac{1}{K'}\sum_{k=1}^{K'} \beta_{kc}^{\star} \tag{2.13}$$

$$\beta_{kc}^{\eta} = \beta_{kc}^{\star} - \sigma_c^{\star}, \tag{2.14}$$

then

$$\eta_c(\cdot) = \exp\left\{\sigma_c^{\star} + \sum_{k=1}^{K'} \phi_k^{\eta}(\cdot)\beta_{kc}^{\eta}\right\}, \tag{2.15}$$

with $\sum_{k=1}^{K'} \beta_{kc}^{\eta} = 0$. Now if $\beta_{kc}^{\eta} = 0, \forall k$, then there is a homogeneous variance $\sigma_c = e^{\sigma_c^{\star}}$ over time for each customer type and if $\sigma_c = \sigma, \forall c$ we have an uniform homogeneity for all types of customer. Hence, the three forms of nested variance functionals are

1. Homogeneous uniform: $\eta_c(t) = \sigma,\ \forall c$;
2. Homogeneous: $\eta_c(t) = \sigma_c$;
3. Complete: $\eta_c(\cdot) = \sigma_c \exp\left\{\sum_{k=1}^{K'} \phi_k^{\eta}(\cdot)\beta_{kc}^{\eta}\right\}$.

2.2.1.2 Correlation functional

The correlation functional $\rho_c(s,t)$ quantifies the relationship between the energy consumption of a customer of type c at two points in time s and t in the time interval

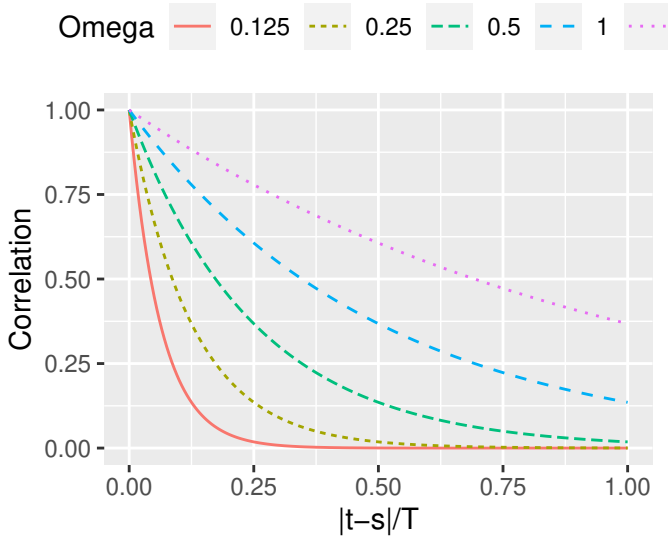


Figure 2 – Exponential correlation structure for different configurations of ω parameter.

$[0, T]$. It is assumed that this relationship is defined by an exponential decay proportional to the absolute difference $|t - s|$ and with parameter $\omega_c > 0, \forall c$, that is,

$$\rho_c(s,t) = \exp\left\{-2\frac{1}{\omega_c}\frac{|t-s|}{T}\right\}. \tag{2.16}$$

2.2.2 Model likelihood

The full aggregated data model in Equation (2.7) includes an error $\varepsilon_{ij}(\cdot)$, which is a Gaussian Process with zero mean and covariance $\mathbf{\Sigma}_j(\cdot,\cdot)$. Therefore, $Y_{ij}(\cdot) \sim GP\Big(\mu_{ij}(\cdot),\, \Sigma_j(\cdot,\cdot)\Big)$ with

$$\mu_{ij}(t) = \sum_{c=1}^{C} m_{jc}\, \boldsymbol{\phi}_i(t)\boldsymbol{\beta}_c + \mathbf{D}_j\boldsymbol{\gamma} \qquad \text{and} \tag{2.17}$$

$$\Sigma_j(s,t) = \sum_{c=1}^{C} m_{jc}\, \eta_c(s)\, \rho_c(s,t)\, \eta_c(t). \tag{2.18}$$

Given a sample $\mathbf{y}$ of N daily observations from J substations over I days, say

$$\mathbf{y} = \Big\{\mathbf{y}_{ij} \,:\, \mathbf{y}_{ij} = \big(y_{ij}(t_1), \ldots, y_{ij}(t_N)\big) \text{ with } i = 1, \ldots, I \text{ and } j = 1, \ldots, J\Big\}. \tag{2.19}$$

Note that $\mathbf{y}$ can be made up of substations observed on different days at different time frequencies. However, to simplify the notation it is assumed that all data are observed on the same days and at the same time frequency. Assuming independence among days and substations and given a set of model parameters $\mathbf{\Theta}$, the likelihood of the aggregated data

model can be written as

$$\mathcal{L}\Big(\boldsymbol{\Theta}|\mathbf{y}\Big) = \prod_{i=1}^{I}\prod_{j=1}^{J} f\Big(\mathbf{y}_{ij};\boldsymbol{\Theta}\Big) \tag{2.20}$$

$$= \prod_{i=1}^{I}\prod_{j=1}^{J} \frac{1}{\sqrt{(2\pi)^n\,|\boldsymbol{\Sigma}_j|}} \exp\left\{-\frac{1}{2}(\boldsymbol{\mu}_{ij}-\mathbf{y}_{ij})^\intercal \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{\mu}_{ij}-\mathbf{y}_{ij})\right\}, \tag{2.21}$$

where

$$\boldsymbol{\mu}_{ij} = \Big\{\mu_{ij}(t_1),\ldots,\mu_{ij}(t_N)\Big\} \qquad \text{and} \tag{2.22}$$

$$\boldsymbol{\Sigma}_j = \Big\{\boldsymbol{\Sigma}_j \in \mathbb{R}^{N\times N} \text{ with elements } \Sigma_j(s,t) : s = t_1,\ldots,t_N;\ t = t_1,\ldots,t_N;\ \Big\}. \tag{2.23}$$

2.3 Model-based clustering analysis

Assume that substations can belong to B distinct clusters depending on the similarity of their consumers typical curves. Let Z_j be the latent variable that identifies to which cluster substation j belongs, with π_b being the probability of substation j belonging to cluster b. In other words, for each substation $j = 1,\ldots,J$, let Z_j be a random multinomial variable such that

$$\mathbb{P}\Big(Z_j = b\Big) = \pi_b, \qquad \text{for } b = 1,2,\ldots,B \tag{2.24}$$

$$\text{and } \sum_{b=1}^{B}\pi_b = 1.$$

It is assumed that given $Z_j = b$, the typical curve of a consumer of type c is given by $\alpha_{cb}(\cdot)$ and the aggregated load is a Gaussian process with mean function $\mu_{jb}(\cdot)$ and covariance function $\Sigma_{jb}(\cdot,\cdot)$, that is,

$$Y_{ij}(\cdot)|Z_j = b \sim GP\Big(\mu_{jb}(\cdot),\, \Sigma_{jb}(\cdot,\cdot)\Big), \tag{2.25}$$

where $\mu_{jb}(t) = \sum_{c=1}^{C} m_{jc}\alpha_{cb}(t)$ and therefore the introduction of the latent variable Z_j leads to a mixture of Gaussian process regression (SHI; MURRAY-SMITH; TITTERINGTON, 2005).

2.3.1 Clustering model likelihood

Let $\mathbf{y}$ be the vector of observed aggregated energy consumption over I days at J substations, as in Equation (2.19), let $\mathbf{z} = \big(z_1,\ldots,z_J\big)$ be the vector of latent variables and

$\boldsymbol{\pi} = \left(\pi_1 \ldots, \pi_B\right)$ its associated parameters. Consider $\mathbf{y}_{\cdot j} = \left(\mathbf{y}_{1j}, \ldots, \mathbf{y}_{Ij}\right)^{\intercal}$, the observed data likelihood can be written as

$$\mathcal{L}\left(\boldsymbol{\Theta}, \boldsymbol{\pi}|\mathbf{y}\right) = \prod_{j=1}^{J} \sum_{b=1}^{B} f\left(\mathbf{y}_{\cdot j}, z_j|\boldsymbol{\Theta}, \boldsymbol{\pi}\right), \tag{2.26}$$

and the corresponding log-likelihood as

$$\begin{aligned} \ell\left(\boldsymbol{\Theta}, \boldsymbol{\pi}; \mathbf{y}, \right) &= \sum_{j=1}^{J} \log \left(\sum_{b=1}^{B} f\left(\mathbf{y}_{\cdot j}, z_j|\boldsymbol{\Theta}, \boldsymbol{\pi}\right) \right) \\ &= \sum_{j=1}^{J} \log \left(\sum_{b=1}^{B} \pi_b f\left(\mathbf{y}_{\cdot j}|z_j, \boldsymbol{\Theta}, \boldsymbol{\pi}\right) \right). \end{aligned} \tag{2.27}$$

The direct optimization of Equation (2.27) is difficult due to the presence of the logarithm of a summation. Section 2.4.2 presents an Expectation-Maximization algorithm (MCLACHLAN; KRISHNAN, 2007; DEMPSTER; LAIRD; RUBIN, 1977), which performs an iterative optimization of Equation (2.27) using the joint distribution of $\mathbf{y}_{ij}$ and $\mathbf{z}_j$, with the so-called complete data likelihood given by:

$$\mathcal{L}\left(\boldsymbol{\Theta}, \boldsymbol{\pi}|\mathbf{y}, \mathbf{z}\right) = \prod_{j=1}^{J} f\left(\mathbf{y}_{\cdot j}, z_j|\boldsymbol{\Theta}, \boldsymbol{\pi}\right) \tag{2.28}$$

$$= \prod_{j=1}^{J} f\left(\mathbf{y}_{\cdot j}|z_j; \boldsymbol{\Theta}\right) \mathbb{P}\left(Z_j = z_j|\boldsymbol{\pi}\right) \tag{2.29}$$

$$= \prod_{j=1}^{J} \prod_{b=1}^{B} \left(f\left(\mathbf{y}_{\cdot j}|z_j; \boldsymbol{\Theta}\right) \mathbb{P}\left(Z_j = z_j|\boldsymbol{\pi}\right) \right)^{\mathcal{I}\left(z_j = b\right)}. \tag{2.30}$$

The components of (2.30) are further detailed in Section 2.4.2.

2.4 Estimation

The estimation procedure assumes that the aggregated load is a Gaussian process and uses the least-squares approach for linear model parameter estimation. The latter provides fast computation and consistent unbiased estimators under a normal distribution (SHI; CHOI, 2011). The clustering analysis is performed by latent variable modeling in the context of a mixture of Gaussian processes using the well-known EM algorithm.

2.4.1 Aggregated data model

The mean function $\mu_{ij}(\cdot)$ in Equation (2.17) can be written as the product of a functional design matrix representation $\mathbf{X}_{ij}(\cdot)$ and the vector of parameters $\boldsymbol{\beta}$:

$$\mu_{ij}(t) = \mathbf{X}_{ij}(t)\boldsymbol{\beta}, \tag{2.31}$$

where $X_{ij}(\cdot)$ is shown in (2.32) as a matrix composed of the basis functions multiplied by its respective market m_{jc} and the covariates $\mathbf{D}_j$, whereas $\boldsymbol{\beta}$ is made up of the parameters of the basis expansion and the coefficients of the explanatory variables shown in (2.33):

$$\mathbf{X}_{ij}(t) = \begin{pmatrix} m_{j1}\boldsymbol{\phi}_i(t) & m_{j2}\boldsymbol{\phi}_i(t) & \cdots & m_{jC}\boldsymbol{\phi}_i(t) & \mathbf{D}_j \end{pmatrix}_{1\times(KLC+P)}, \tag{2.32}$$

$$\boldsymbol{\beta}^\intercal = \begin{pmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \cdots & \boldsymbol{\beta}_C & \boldsymbol{\gamma} \end{pmatrix}_{1\times(KLC+P)}. \tag{2.33}$$

With this vector representation, we can write the aggregated data model as

$$\mathbf{Y}_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{ij}, \tag{2.34}$$

where $\mathbf{X}_{ij}$ and $\boldsymbol{\varepsilon}_{ij}$ are row bindings of evaluations of $\mathbf{X}_{ij}(\cdot)$ and $\boldsymbol{\varepsilon}_{ij}(\cdot)$ at times $t_1, \ldots, t_N$, or in other words

$$\mathbf{X}_{ij} = \begin{pmatrix} \mathbf{X}_{ij}(t_1) \\ \mathbf{X}_{ij}(t_2) \\ \vdots \\ \mathbf{X}_{ij}(t_N) \end{pmatrix}_{N\times(KLC+P)} \quad \text{and} \quad \boldsymbol{\varepsilon}_{ij} = \begin{pmatrix} \varepsilon_{ij}(t_1) \\ \varepsilon_{ij}(t_2) \\ \vdots \\ \varepsilon_{ij}(t_N) \end{pmatrix}_{N\times 1}. \tag{2.35}$$

Furthermore, the model can be represented across all J substations over I days using a single vector $\mathbf{Y}$, that is,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.36}$$

where

$$\mathbf{y} = \begin{pmatrix} \mathbf{Y}_{11} \\ \mathbf{Y}_{21} \\ \vdots \\ \mathbf{Y}_{I1} \\ \mathbf{Y}_{12} \\ \vdots \\ \mathbf{Y}_{IJ} \end{pmatrix}_{NIJ\times 1}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_{11} \\ \mathbf{X}_{12} \\ \vdots \\ \mathbf{X}_{I1} \\ \mathbf{X}_{I2} \\ \vdots \\ \mathbf{X}_{IJ} \end{pmatrix}_{NIJ\times(KLC+P)} \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{\varepsilon}_{11} \\ \boldsymbol{\varepsilon}_{21} \\ \vdots \\ \boldsymbol{\varepsilon}_{I1} \\ \boldsymbol{\varepsilon}_{12} \\ \vdots \\ \boldsymbol{\varepsilon}_{IJ} \end{pmatrix}_{NIJ\times 1}. \tag{2.37}$$

Thus the density $f(\mathbf{y};\boldsymbol{\Theta})$ can be written as a Normal density with mean $\mathbf{X}\boldsymbol{\beta}$ and sparse block diagonal covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{NIJ\times NIJ}$, composed of the matrices $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_J$. Hence, the log-likelihood of the aggregated data model in Equation (2.21) can be written as

$$\begin{aligned} \ell\Big(\boldsymbol{\Theta};\mathbf{y}\Big) &= \log \mathcal{L}\Big(\boldsymbol{\Theta};\mathbf{y}\Big) \\ &\propto -\frac{1}{2}\log|\boldsymbol{\Sigma}| + -\frac{1}{2}\big(\mathbf{X}\boldsymbol{\beta}-\mathbf{y}\big)^\intercal\boldsymbol{\Sigma}^{-1}\big(\mathbf{X}\boldsymbol{\beta}-\mathbf{y}\big). \end{aligned} \tag{2.38}$$

Equation (2.38) configures a Gaussian process regression likelihood (SHI; CHOI, 2011; RAMSAY; SILVERMAN, 2005). The estimator of $\boldsymbol{\beta}$ is obtained using weighted least squares and $\boldsymbol{\Sigma}$ is estimated using the *BFGS* Quasi-Newton numerical optimization method. An alternative to estimate $\boldsymbol{\Sigma}$, possibly avoiding local maxima, but with higher computational cost is *Simulated Annealing* (FLETCHER, 2013; KIRKPATRICK; GELATT; VECCHI, 1983; RUGGIERO; LOPES, 1997). The complete estimation procedure is described below.

2.4.1.1 Estimation algorithm

Given a sample $\mathbf{y}$ and its log-likelihood in (2.38), the parameters in $\boldsymbol{\beta}$ and the covariance parameters in $\boldsymbol{\Theta}_{\boldsymbol{\Sigma}} = \left(\boldsymbol{\sigma}, \boldsymbol{\omega}, \boldsymbol{\beta}^{\eta}\right)$ are estimated using Algorithm 1 as described below.

Algorithm 1. *Fix a precision value $\xi > 0$. Given a sample $\mathbf{y}$, at run $r = 0$ get initial values for $\boldsymbol{\beta}^{(0)}$. At run $r > 0$, do*

1. *Fix $\boldsymbol{\beta}^{(r-1)}$ to obtain $\boldsymbol{\Theta}_{\boldsymbol{\Sigma}}^{(r)}$ by optimizing the log-likelihood in (2.38).*
2. *Fix $\boldsymbol{\Theta}_{\boldsymbol{\Sigma}}^{(r)}$ to obtain $\boldsymbol{\beta}^{(r)}$ via*

$$\boldsymbol{\beta}^{(r)} = \left(\mathbf{X}^{\intercal}\left(\boldsymbol{\Sigma}^{(r)}\right)^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{X}^{\intercal}\left(\boldsymbol{\Sigma}^{(r)}\right)^{-1}\mathbf{y}\right). \tag{2.39}$$

3. *If*

$$\left|\ell\left(\boldsymbol{\Theta}^{(r)}; \mathbf{y}\right) - \ell\left(\boldsymbol{\Theta}^{(r-1)}; \mathbf{y}\right)\right| < \xi,$$

then stop. If not, add one unit to run (r) and repeat.

The precision value $\xi > 0$, also called the convergence criterion, is typically set to 10^{-6}. Because the least squares estimator for $\boldsymbol{\beta}$ is unbiased and its expected value does not depend on $\boldsymbol{\Sigma}$, the initial values for $\boldsymbol{\beta}$ can be obtained by fitting a linear model with no covariance structure for the aggregated data model.

To improve computational performance, (2.39) can be written in terms of the covariance matrices for each substation in the block diagonal matrix $\boldsymbol{\Sigma}$, that is,

$$\boldsymbol{\beta}^{(r)} = \left(\sum_{i=1}^{I}\sum_{j=1}^{J}\mathbf{X}_{ij}^{\intercal}\left(\boldsymbol{\Sigma}_{j}^{(r)}\right)^{-1}\mathbf{X}_{ij}\right)^{-1}\left(\sum_{i=1}^{I}\sum_{j=1}^{J}\mathbf{X}_{ij}^{\intercal}\left(\boldsymbol{\Sigma}_{j}^{(r)}\right)^{-1}\mathbf{y}_{ij}\right). \tag{2.40}$$

2.4.1.2 Conditions for identifiability

To ensure the existence of the inverse of the left-hand size of $\boldsymbol{\beta}^{(r)}$ in (2.39) the number of substations in sample $\mathbf{y}$ must be greater than the number of subject types. In other words, $J > C$. Also, to avoid multicolinearity, markets must be linearly independent, that is, there must be no $M \in \mathbb{R}$ such that $\mathbf{m}_j = M\mathbf{m}_{j'}$, for any $j \neq j'$.

2.4.2 Model-based clustering

The observed data likelihood of the clustering aggregated data model in Section 2.3.1 depends on the latent variable Z_j. Hence, the EM algorithm was used to estimate all parameters of interest using the complete data likelihood. The following subsection present the details of the E and M steps for the proposed model.

2.4.2.1 E-Step

Let $\boldsymbol{\Theta}$ be the set of parameters of the distribution of $\mathbf{Y}$ given $\mathbf{Z}$ and $\boldsymbol{\pi}$ the set of parameters of the distribution of $\mathbf{Z}$, as defined in Section 2.3. Given the observed data $\mathbf{y}$ and the unobserved data $\mathbf{z}$, let ℓ be the complete data log-likelihood of parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$, which can be written as

$$\begin{aligned}
\ell\Big(\boldsymbol{\Theta}, \boldsymbol{\pi}|\mathbf{y}, \mathbf{z}\Big) &= \log \mathcal{L}\Big(\boldsymbol{\Theta}, \boldsymbol{\pi}|\mathbf{y}, \mathbf{z}\Big) \\
&\propto \sum_{j=1}^{J}\sum_{b=1}^{B} \mathcal{I}\big(z_j = b\big) \times \\
&\qquad \left(\log \pi_b - \frac{1}{2}\sum_{i=1}^{I}\Big[\log|\boldsymbol{\Sigma}_{jb}| + (\boldsymbol{\mu}_{jb} - \mathbf{y}_{ij})^{\intercal}\boldsymbol{\Sigma}_{jb}^{-1}(\boldsymbol{\mu}_{jb} - \mathbf{y}_{ij})\Big]\right).
\end{aligned} \tag{2.41}$$

The E-Step of the EM algorithm calculates the expected values of $\ell\Big(\boldsymbol{\Theta}, \boldsymbol{\pi}|\mathbf{y}$ and $\mathbf{z}\Big)$ with respect to the conditional distribution of $\mathbf{Z}$ given the observed data and current parameter estimates $\boldsymbol{\Theta}^{(r)}$ and $\boldsymbol{\pi}^{(r)}$ at run r to obtain

$$\begin{aligned}
Q\Big(\boldsymbol{\Theta}, \boldsymbol{\pi}|\, \boldsymbol{\Theta}^{(r)}, \boldsymbol{\pi}^{(r)}\Big) &\equiv \mathbb{E}_{\mathbf{Z}|\mathbf{y}, \boldsymbol{\Theta}^{(r)}, \boldsymbol{\pi}^{(r)}}\left[\ell\Big(\boldsymbol{\Theta}, \boldsymbol{\pi}|\mathbf{y}, \mathbf{z}\Big)\right] \\
&\propto \sum_{j=1}^{J}\sum_{b=1}^{B} \mathbb{P}\Big(Z_j = b|\mathbf{y}_{1j}, \ldots, \mathbf{y}_{Ij}; \boldsymbol{\Theta}^{(r)}, \boldsymbol{\pi}^{(r)}\Big) \times \\
&\qquad \left(\log \pi_b - \frac{I}{2}\log|\boldsymbol{\Sigma}_{jb}| - \frac{1}{2}\sum_{i=1}^{I}(\boldsymbol{\mu}_{jb} - \mathbf{y}_{ij})^{\intercal}\boldsymbol{\Sigma}_{jb}^{-1}(\boldsymbol{\mu}_{jb} - \mathbf{y}_{ij})\right).
\end{aligned} \tag{2.42}$$

The probability $\mathbb{P}\Big(\big(\big)\cdot\Big)$ in Equation (2.42) can be computed using Bayes Theorem and written as

$$\begin{aligned}
\mathbb{P}\Big(Z_j = b|\mathbf{y}_{\cdot j}; \boldsymbol{\Theta}^{(r)}, \boldsymbol{\pi}^{(r)}\Big) &\coloneqq \frac{f\Big(\mathbf{y}_{\cdot j}|z_j = b; \boldsymbol{\Theta}_b^{(r)}\Big) \times \pi_b^{(r)}}{\sum_{b'=1}^{B} f\Big(\mathbf{y}_{\cdot j}|z_j = b'; \boldsymbol{\Theta}_{b'}^{(r)}\Big) \times \pi_{b'}^{(r)}} \\
&= \frac{\Big[\prod_{i=1}^{I} f\Big(\mathbf{y}_{ij}|z_j = b; \boldsymbol{\Theta}_b^{(r)}\Big)\Big] \times \pi_b^{(r)}}{\sum_{b'=1}^{B}\Big[\prod_{i=1}^{I} f\Big(\mathbf{y}_{ij}|z_j = b'; \boldsymbol{\Theta}_{b'}^{(r)}\Big)\Big] \times \pi_{b'}^{(r)}},
\end{aligned} \tag{2.43}$$

where the product of densities is possible because independence among days $i = 1, 2, \ldots, I$ is assumed.

2.4.2.2 M-Step

Given (2.42) and (2.43), the M-Step maximizes the function $Q(\cdot)$ in terms of the parameters $\boldsymbol{\Theta} = \{\boldsymbol{\beta}, \boldsymbol{\Theta}_{\boldsymbol{\Sigma}}\}$ and $\boldsymbol{\pi}$, where $\boldsymbol{\Theta}_{\boldsymbol{\Sigma}}$ contains the parameters $\boldsymbol{\beta}^{\eta}$ and $\boldsymbol{\omega}$ of the covariance matrix $\boldsymbol{\Sigma}_{jb}$ described in Section 2.2.2. Let

$$p_{jb}^{(r)} = \mathbb{P}\Big(Z_j = b|\mathbf{y}_{\cdot j}; \boldsymbol{\Theta}^{(r)}, \boldsymbol{\pi}^{(r)}\Big) \tag{2.44}$$

and let $Q(\cdot)$ be written as a sum of two terms: one that depends only on $\boldsymbol{\pi}$ and another term that depends only on $\boldsymbol{\Theta}$, that is,

$$Q\Big(\boldsymbol{\Theta}, \boldsymbol{\pi}|\, \boldsymbol{\Theta}^{(r)}, \boldsymbol{\pi}^{(r)}\Big) = Q_1\Big(\boldsymbol{\pi}|\, \boldsymbol{\Theta}^{(r)}, \boldsymbol{\pi}^{(r)}\Big) + Q_2\Big(\boldsymbol{\Theta}|\, \boldsymbol{\Theta}^{(r)}, \boldsymbol{\pi}^{(r)}\Big) \tag{2.45}$$

where

$$Q_1\Big(\boldsymbol{\pi}|\, \boldsymbol{\Theta}^{(r)}, \boldsymbol{\pi}^{(r)}\Big) = \sum_{j=1}^{J}\sum_{b=1}^{B} p_{jb}^{(r)} \log \pi_b \text{ and} \tag{2.46}$$

$$Q_2\Big(\boldsymbol{\Theta}|\, \boldsymbol{\Theta}^{(r)}, \boldsymbol{\pi}^{(r)}\Big) = -\frac{1}{2}\sum_{j=1}^{J}\sum_{b=1}^{B} p_{jb}^{(r)}\Big(\log|\boldsymbol{\Sigma}_{jb}| + \tag{2.47}$$

$$\sum_{i=1}^{I}(\mathbf{X}_j\boldsymbol{\beta}_b - \mathbf{y}_{ij})^{\intercal}\boldsymbol{\Sigma}_{jb}^{-1}(\mathbf{X}_j\boldsymbol{\beta}_b - \mathbf{y}_{ij})\Big). \tag{2.48}$$

Because Q_1 does not depend on $\boldsymbol{\Theta}$, $\pi_b^{(r+1)}$ can be obtained by maximizing Equation (2.46) with respect to π_b, subject to $\sum_{b=1}^{B}\pi_b = 1$. Therefore, using Lagrange multipliers it can be shown that

$$\pi_b^{(r+1)} = \frac{1}{J}\sum_{j=1}^{J} p_{jb}^{(r)}, \tag{2.49}$$

for $b = 1, \ldots, B$.

To obtain $\boldsymbol{\beta}_b^{(r+1)}$ and $\boldsymbol{\Theta}_{\boldsymbol{\Sigma}}^{(r+1)}$, this study uses the so-called Expectation/Conditional Maximization (ECM) algorithm (MENG; RUBIN, 1993; MCLACHLAN; KRISHNAN, 2007), where $\boldsymbol{\Theta}_{\boldsymbol{\Sigma}}$ is set equal to $\boldsymbol{\Theta}_{\boldsymbol{\Sigma}}^{(r)}$ and Q_2 is maximized with respect to $\boldsymbol{\beta}_b$ to obtain

$$\boldsymbol{\beta}_b^{(r+1)} = \Big(I\sum_{j=1}^{J}\mathbf{X}_j^{\intercal}\big(\boldsymbol{\Sigma}_{jb}^{\star}\big)^{-1}\mathbf{X}_j\Big)^{-1}\Big(\sum_{i=1}^{I}\sum_{j=1}^{J}\mathbf{X}_j^{\intercal}\big(\boldsymbol{\Sigma}_{jb}^{\star}\big)^{-1}\mathbf{y}_{ij}\Big), \tag{2.50}$$

for $b = 1, \ldots, B$ and $\boldsymbol{\Sigma}_{jb}^{\star} = p_{jb}^{(r)} \times \boldsymbol{\Sigma}_{jb}^{-1}$.

Next, $\boldsymbol{\beta}_b$ is set to its updated value $\boldsymbol{\beta}_b^{(r+1)}$ and Q_2 is maximized with respect to $\boldsymbol{\Theta}_{\boldsymbol{\Sigma}}$ through numerical optimization algorithms to obtain $\boldsymbol{\Theta}_{\boldsymbol{\Sigma}}^{(r+1)}$.

2.4.2.3 Estimation algorithm

Given the iterative forms of estimation of $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$, we present the following steps based on the ECM algorithm.

Algorithm 2. *Fix a precision value $\xi > 0$. Given a sample $\mathbf{y} = \{\mathbf{y}_{\cdot 1}, \ldots, \mathbf{y}_{\cdot J}\}$, at run $r = 0$, get initial values for $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\Theta}_{\boldsymbol{\Sigma}}^{(0)}$ and $\boldsymbol{\pi}^{(0)}$. For the following runs $r > 0$, do*

1. *E-Step: for $b = 1, \ldots, B$, obtain $p_{jb}^{(r)} = \mathbb{P}\Big(Z_j = b | \mathbf{y}_{\cdot j}; \boldsymbol{\pi}^{(r)}, \boldsymbol{\beta}_b^{(r)}, \boldsymbol{\Sigma}_b^{(r)}\Big)$.*
2. *M-Step: maximize $Q^{\star}\Big(\boldsymbol{\Theta} | \boldsymbol{\Theta}^{(r)}\Big)$ to obtain $\boldsymbol{\Theta}^{(r+1)}$:*
   a) *Obtain $\pi_b^{(r+1)} = \dfrac{1}{J} \displaystyle\sum_{j=1}^{J} p_{jb}^{(r)}$.*
   b) *Set $\boldsymbol{\Theta}_{\boldsymbol{\Sigma}}$ to $\boldsymbol{\Theta}_{\boldsymbol{\Sigma}}^{(r)}$ to obtain*

$$\boldsymbol{\beta}_b^{(r+1)} = \Big(I \sum_{j=1}^{J} \mathbf{X}_j^{\intercal} \big(\boldsymbol{\Sigma}_{jb}^{\star}\big)^{-1} \mathbf{X}_j\Big)^{-1} \Big(\sum_{i=1}^{I} \sum_{j=1}^{J} \mathbf{X}_j^{\intercal} \big(\boldsymbol{\Sigma}_{jb}^{\star}\big)^{-1} \mathbf{y}_{ij}\Big),$$

   c) *Set $\boldsymbol{\beta}_b$ to $\boldsymbol{\beta}_b^{(r+1)}$ in item b) to obtain $\boldsymbol{\Theta}_{\boldsymbol{\Sigma}}^{(r+1)}$ by maximizing Equation (2.48) with respect to $\boldsymbol{\Theta}_{\boldsymbol{\Sigma}}$.*
3. *Let $\ell\big(\boldsymbol{\Theta}, \boldsymbol{\pi}; \mathbf{y}\big)$ be the observed data log-likelihood defined in Equation (2.27). If*

$$\Big|\ell\big(\boldsymbol{\Theta}^{(r)}, \boldsymbol{\pi}^{(r)}; \mathbf{y}\big) - \ell\big(\boldsymbol{\Theta}^{(r-1)}, \boldsymbol{\pi}^{(r-1)}; \mathbf{y}\big)\Big| < \xi,$$

then stop. If not, add one unit to run (r) and repeat.

2.4.2.4 Initial values and number of clusters

Obtaining initial values for all parameters might be a challenge if no previous information is available to guide the initialization. In this work, an approach is proposed to obtain cluster and — initial values, much like the approach in Section 2.4.1.1 for the full aggregated data model.

The first step is to fix the number of clusters B and the number of trials G. For each trial $g \in G$, each substation is randomly assigned to a cluster, where the number of substations in each cluster must be greater than the number of customer types to preserve model identifiability. For each trial g, the clusters are split into datasets with their respective substations, and a simple aggregated data model is fitted to each one. Then the model with the smallest squared error among the G trials is selected to provide an initial $\boldsymbol{\beta}_b$. The initial $\boldsymbol{\pi}$ is the proportion of substations in each cluster and the winning trial can also be used to provide initial covariance parameters.

The total number of clusters B is highly dependent on previous user information. As a first step, one might use the suggested approach of multiple fits with different numbers of clusters to select the configuration with the smallest squared error, which implies in high computing cost, or one might use an approximation of Bayes factors to select the best number of clusters B (SCHWARZ et al., 1978). The latter approach is detailed in Section 2.5 and is the approach selected for this thesis.

It is also possible to assume that B is a random variable and obtain its estimated value through its posterior probability using approaches like the reversible jump algorithm (GREEN, 1995), but with intensive computation.

2.4.2.5 Identifiability condition

As in Section 2.4.1.2, there are necessary conditions for model fitting. Because there are at least B times the number of parameters, the procedure requires $J > CB$, that is, the number of substations must be greater than the number of estimated typical curves. Furthermore, substation markets must not be proportional to ensure full rank matrices in least squares computations.

2.5 Model check

This section will examine how to assess the uncertainty of the estimated mean curves and their covariance parameters. Inferences on the disaggregated mean curves can be performed by taking the closed form of the parameters in (2.39) and (2.50), because they are functions of the Gaussian process $Y_{ij}(\cdot)$ (SHI; CHOI, 2011; TRESP, 2001). In fact, it can be said that

$$\hat{\boldsymbol{\beta}} \sim Normal\Big(\boldsymbol{\beta}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\intercal}\Big), \tag{2.51}$$

with $\boldsymbol{\beta} \in \mathbb{R}^{CK}$ as the true expansion parameters and $\mathbf{A} \in \mathbb{R}^{CK \times NIJ}$ defined as

$$\mathbf{A} = \Big(\mathbf{X}^{\intercal}\boldsymbol{\Sigma}^{-1}\mathbf{X}\Big)^{-1}\mathbf{X}^{\intercal}\boldsymbol{\Sigma}^{-1}. \tag{2.52}$$

Using the distribution of $\hat{\boldsymbol{\beta}}$ given by (2.51), confidence intervals can be determined based on the standard errors in the diagonal of $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\intercal}$. On the other hand, the covariance parameters $\boldsymbol{\Theta}_{\boldsymbol{\Sigma}}$ are obtained by numerical optimization using the Quasi-Newton methods available in the R language (FLETCHER, 2013; R Core Team, 2019), and the parameter standard errors can be obtained from the observed Hessian matrix $\mathbf{H}$, that is,

$$SE\big(\boldsymbol{\Theta}_{cov}\big) = \sqrt{diag\Big(\mathbf{H}^{-1}\Big)}. \tag{2.53}$$

The proposed covariance structures make the aggregated data models a family of nested models, where the uniformly homogeneous one is a particular case of the

homogeneous model which is a particular case of the complete model. Two model fits can be compared using the likelihood ratio test. Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be the two aggregated data models to be compared, with $\mathcal{M}_1$ nested in $\mathcal{M}_2$. Denote by $\ell(\mathcal{M}_1)$ and $\ell(\mathcal{M}_2)$ the log-likelihood of models $\mathcal{M}_1$ and $\mathcal{M}_1$. Then the likelihood ratio statistic L is defined by

$$L = -2\Big(\ell(\mathcal{M}_2) - \ell(\mathcal{M}_1)\Big), \tag{2.54}$$

where the test statistic is asymptotically χ^2 distributed with degrees of freedom equal to the difference in the number of parameters between models.

When comparing clustering models, if they have the same number of clusters but different covariance structures, then the same approach can be used to compare them. However, to compare models with distinct numbers of clusters, the *Bayesian information criterion* (BIC) value comparison as in (SHI; WANG, 2008) is recommended. Let $\ell\big(\boldsymbol{\Theta}, \boldsymbol{\pi}; \mathbf{y}\big)$ be the observed log-likelihood and let $\hat{\boldsymbol{\Theta}}$ and $\hat{\boldsymbol{\pi}}$ be the maximum likelihood estimates, then the BIC is given by

$$BIC = -2\ell\Big(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\pi}}; \mathbf{y}\Big) + H\log(IJN), \tag{2.55}$$

where H is the total number of parameters, I the total number of days, J the number of substationsand N the number of observed point in time. Simulation studies for Gaussian process mixtures have shown that models with the smallest BIC tend to have the correct number of clusters (SHI; WANG, 2008).

Some authors use the right-hand element of the sum in Equation 2.55 as $\log(IJ)$, that is, they consider the number of observations as the number of curves, not the number of observed data points (HULLAIT et al., 2020; LI; WANG; CARROLL, 2013; WEI; ZHOU, 2010). In this thesis BIC is used as in Equation (2.55).

Finally, if a model has a good fit to the observed data, the residual curves can be expected to oscillate randomly around the zero line.

2.5.1 Simulation performance measures

In simulation studies, the true parameters are known and are used to measure the performance of the aggregated data models. To assess the performance of the estimated typical curves, the relative residual curve $R_c(t)$ of the customer of type c is defined as

$$R_c(t) = \frac{\hat{\alpha}_c(t) - \alpha_c(t)}{\alpha_c(t)}. \tag{2.56}$$

Analogously, the relative residual curve of the estimated variance functionals is also defined as

$$R_c(t) = \frac{\hat{\eta}_c(t) - \eta_c(t)}{\eta_c(t)} \tag{2.57}$$

.

Division by the true value in (2.56) and (2.57) is desirable to make the residual curves comparable under different magnitudes.

Let R_{rc} be the relative residual curve of the customer of type c in the r-th simulation run. Define the functional Mean Squared Relative Error (fMSRE) as the mean of the integrals of the squared relative residual curves over time t. That is,

$$fMSRE_c = \frac{1}{R}\sum_{r=1}^{R}\int_0^T R_{rc}^2(t)dt \approx \frac{1}{R}\sum_{r=1}^{R}\left\{\frac{T}{N}\sum_{t=t_1}^{t_N} R_{rc}^2(t)\right\}, \tag{2.58}$$

where N is the number of observed points in time in the data set and T the upper limit of the time domain. Because in this thesis the time frequencies are equally distanced, the fraction T/N is the equally spaced time difference band that approximates the dt of the integral on the left-hand side.

3 Simulation studies

3.1 Overview

This chapter evaluates the proposed aggregated data model in simulated scenarios. This approach provides control over the true model parameters that generate the data and the possibility of assessing the performance of estimated parameters in multiple simulation runs.

Two independent simulation studies were performed: one for the full aggregated data model, and the other for the clustering aggregated data model. Section 3.2 introduces the simulated scenarios for both studies. Section 3.3 presents the first study with its typical surface, explanatory variables and functional variance; focusing on the precision of the estimated parameters under two model fits: one considering a homogeneous covariance structure and the other a complete structure as in data generation. Section 3.4 describes the clustering aggregated data model and the results of its substation allocation under two numbesr of clusters. All parameters used in this chapter are based on the estimated typical curves and estimated covariance parameters obtained in Chapter 4 from the UK electrical energy substation data.

3.2 Simulated scenario setup

The simulated scenarios are different combinations of number of observed days, representing the amount of information available, and market balance, which is detailed below.

In real substation data, it is sometimes observed that a particular customer type may be overrepresented, with more than 95% of the market. If this dominance occurs in all observed substations, this situation is called an unbalanced market scenario, and a balanced market scenario otherwise. To study this phenomenon, the markets were generated as follows:

- Unbalanced: all substations have markets with more customers of Type 1 than Type 2 with percentage varying between 70% and 95%.
- Balanced: six substations have most of their customers of Type 1 and six substations have most of their customers of Type 2, with the majority percentages varying between 70% and 95%.

The percentages are relative to the number of customers for each substation, which is displayed in Table 2. Annex A displays all the simulated markets for each scenario.

Table 2 – Fixed number of customers for each substation in the simulation study.

Substation	1	2	3	4	5	6	7	8	9	10	11	12
Total	231	151	156	109	225	172	206	182	175	160	254	69

The combinations of market balance and number of observed days compose the eight different simulated scenarios presented in Table 3. Scenarios 1 to 4 are related to the full aggregated data model study and Scenarios 5 to 8 to the clustering aggregated data model study. Each scenario is composed of two types of customers observed at 30 minutes time frequency at 12 substations and replicated 15 times. In other words, 15 datasets were generated with these configurations and studied in detail, as described in Sections 3.3 and 3.4.

Table 3 – Covariance structure, number of clusters, number of observed days, market balance and number of generated datasets (replicates). Eight simulated scenarios were proposed: Scenarios 1 to 4 for the full aggregated data models and Scenarios 5 to 8 for the clustering aggregated data model.

Scenario	Covariance	Clusters	Days	Market	Replications
1	Complete	1	5	Unbalanced	15
2	Complete	1	5	Balanced	15
3	Complete	1	30	Unbalanced	15
4	Complete	1	30	Balanced	15
5	Homogeneous	3	5	Unbalanced	15
6	Homogeneous	3	5	Balanced	15
7	Homogeneous	3	30	Unbalanced	15
8	Homogeneous	3	30	Balanced	15

3.3 Full aggregated data model

The full aggregated data model studies the typical surface together with explanatory variables related to substations. In this case, the surface is a function of time and daily air temperature, as presented in Section 2.2. Section 3.3.1 describes the air temperature functional and the explanatory variables used in this simulation, Section 3.3.2 presents the main results and Section 3.3.2 contains a discussion and the conclusions of this study.

3.3.1 True parameters

Recall the typical surface in Equation (2.6) introduced in Section 2.2:

$$\alpha_{ic}(t) = \alpha_{ic}\big(u(t), v_i(t)\big).$$

In this section, $u(t) = t$ and $v_i(t) = T_i(t)$ are used as the temperature at day i. Then, the typical surface is given by

$$\alpha_{ic}\Big(t, T_i(t)\Big) = b_c(t) \times \left(1 - \frac{1}{2}\Phi\big(T_i(t) - 1\big)\right), \tag{3.1}$$

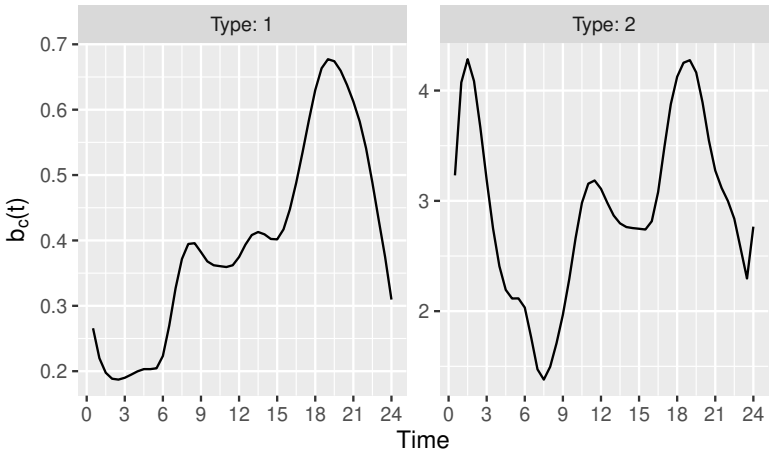
with $b_c(t)$ as the baseline curve for customer of type c and $\Phi(\cdot)$ as the cumulative density function of the standard normal distribution. Hence when the temperature drops below 1°C the typical surface area increases considerably.

Figure 3 shows the baseline curves, the variance functionals and the signal-to-noise ratio (SNR) for each customer type. The SNR is simply the ratio of the typical curve to the variance functional at time t. The baseline curves and variance functionals were based on the estimated typical curves obtained from the real data, as will be described in detail in Chapter 4. The type 1 baseline curve mimics the unrestricted domestic customer with lower consumption in early morning, increasing after 8 AM and reaching its peak at 8 PM. The Type 2 curve mimics the "Economy 7" customer with peaks around 2am and 8pm but with considerably larger electrical load values than Type 1. Customers variance functionals have higher values around the work period between 9am and 5pm, although Type 1 has two peaks that possibly represent when people leave from and arrive at their homes. The typical surfaces are shown in Figure 5.

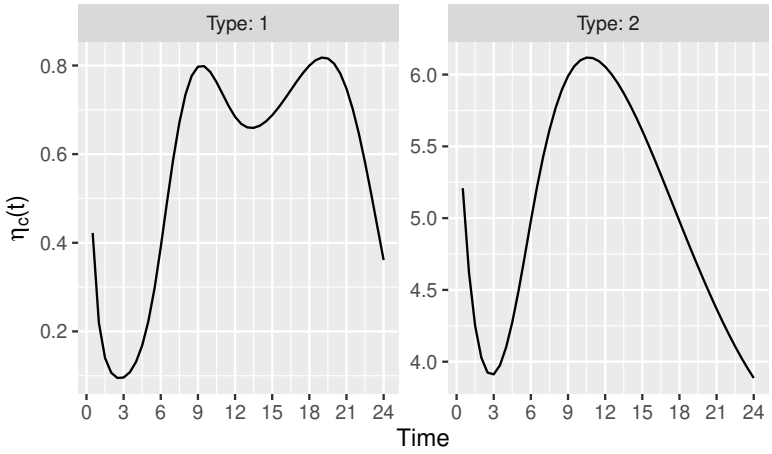
The weather data containing temperature and air humidity were also based on real measurements for winter 2013 in Wales, United Kingdom. For this study, three sets of data were generated, representing three locations labelled T1, T2, and T3. Substations 1 to 4 were assigned to location T1, substations 5 to 8 to location T2, and the remaining substations 9 to 12 to location T3. Figure 4 shows the temperature and air humidity profiles for each location observed over 30 days. In fact, only data for scenarios 3, 4, 7, and 8 were generated in this manner. For scenarios 1, 2, 5, and 6, only the first five days were considered. In this study, temperature was used as the second component of the typical surface, and air humidity was used as an explanatory variable of the full aggregated data model with constant coefficient.

Furthermore, two explanatory variables were considered: air humidity as a functional variable, and a binary variable with value 1 for substations 1 and 2 and 0 otherwise, with associated coefficients $1/90 = 0.0111$ and 13, respectively. Therefore, from Equation (2.7), the full aggregated complete model can be written as

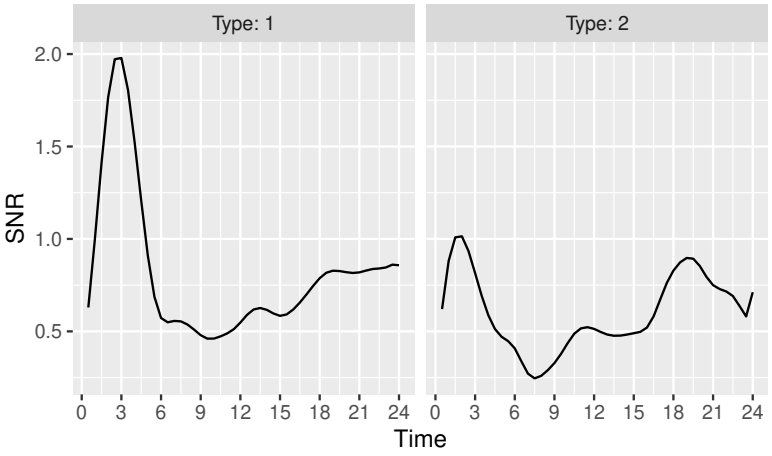
$$Y_{ij}(t) = \left(\sum_{c=1}^{C} m_{jc}\alpha_c\big(t, T_i(t)\big)\right) + 13\, D_{j1} + 0.0111 D_{ij2}(t) + \varepsilon_{ij}(t), \tag{3.2}$$



(a) Baseline curves



(b) Variance functionals



(c) Signal-to-noise ratio (SNR)

Figure 3 – Baseline curves, true variance functionals and signal-to-noise ratio at time t of the simulation study.

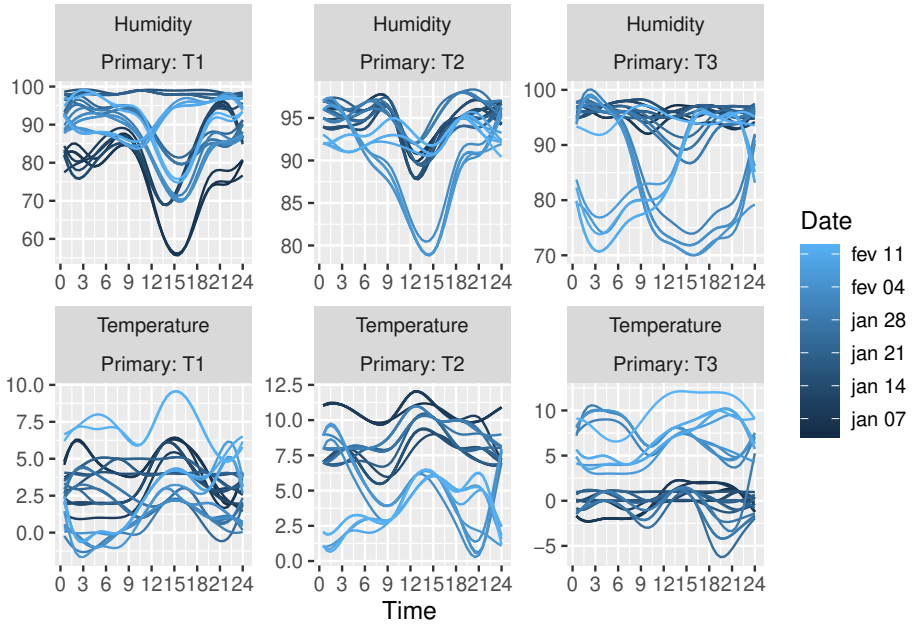


Figure 4 – Air temperature and air humidity for each primary T1, T2 and T3 used in the simulation study.

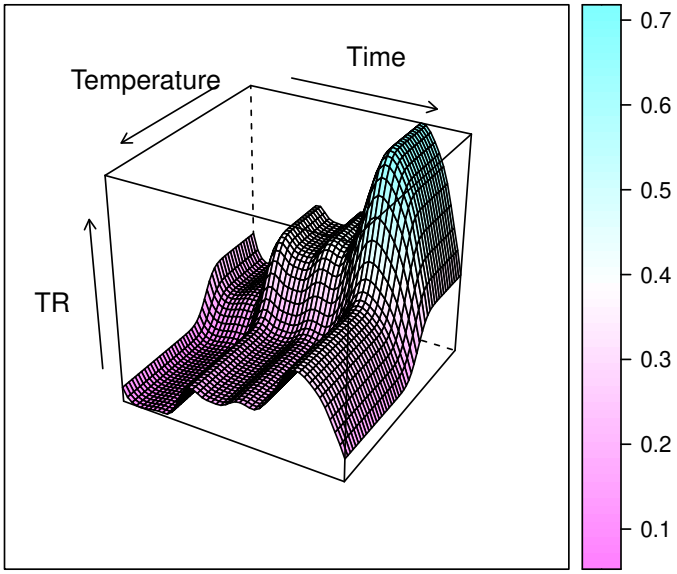
where D_{j1} is the dummy variable for substations 1 and 2 and $D_{j2} = D_{ij2}(t)$ the air humidity of substation j at time t of day i. Finally, the true covariance decay parameters for each customer type were defined as $\omega_1 = 0.03$ and $\omega_2 = 0.7$.

3.3.2 Results

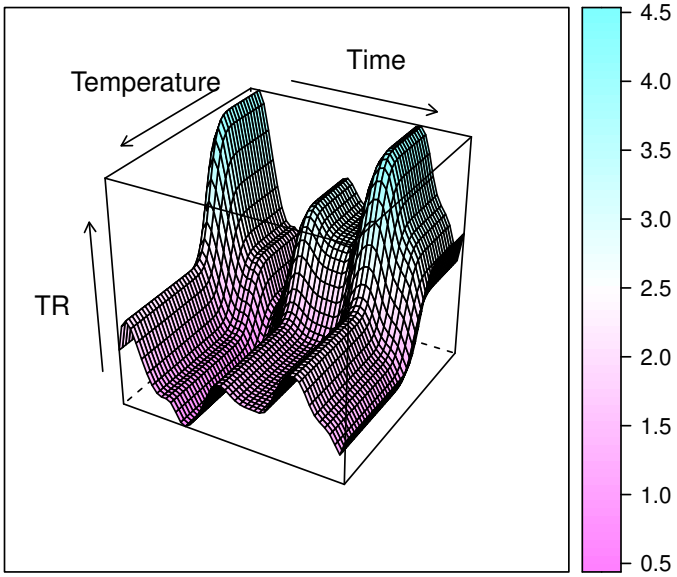
In this study, for each of Scenarios 1 to 4 in Table 3, two models were fitted: one homogeneous and one complete aggregated data model. The homogeneous fit tests the performance of typical surface estimation under an under-parameterized covariance structure and the behaviour of the dispersion parameters by reducing the variance functional to a scalar. On the other hand, the complete model tests check whether, under the correct scenario, the proposed model performs well in terms of typical surface and covariance parameter estimation.

Throughout this chapter, the number of observed days and the market balance are explicitly shown to avoid consulting Table 3 to remember the scenario setup.

Starting with the homogeneous fit study, Figure 6 shows the estimated typical surfaces $\alpha_{ic}(t, T(t))$ for some temperature curves $T(t)$ for every combination of observed days and market balance. The first row in the panels represents a single instance of the temperature $T(t)$ on the first observed day in the simulated data and in its respective primary group T1, T2 or T3. Observe that Figures 6a and 6b show estimated typical surfaces with noticeable variability, where some curves assume negative values. However, the balanced scenario in Figure 6a presents estimated curves for Type 2 with lower



(a) Typical surface for Type 1



(b) Typical surface for Type 2

Figure 5 – Typical surfaces (TR) for each customer type as functions of time and the full range of observed temperatures in the three primaries.

variability than those in Figure 6b. On the lower panels, Figures 6c and 6d show lower variability than the five-day scenarios. Furthermore, observe that the estimated curves for Type 2 in Figure 6c have even lower variability. In general, the median curves in the four scenarios show that the estimated curves are concentrated around their true values. This proximity to the true curve is better visualized in the residual curves shown in Figure 7. As presented in Section 2.5, these curves are standardized so that the scenario performance can be compared. Note that the residual curves for Type 2 in the five-day scenarios have lower variability in Figure 7a than their respective ones in Figure 7b, as mentioned earlier. The same event occurs in the 30-day scenarios, but with lower variability than the five-day scenarios. The four panels of Figure 7 show median curves oscillating around the horizontal zero-reference line, with no major differences among scenarios. To summarize the precision of the estimated typical surfaces shown in Figure 6, Table 4 shows the functional Mean Squared Relative Error for Scenarios 1 to 4 fitted by the homogeneous model. Clearly, the fMSRE for the estimated Type 1 typical curves is considerably higher in the five-day scenarios. It seems that the magnitude of the curves influences the variability of the estimates because the curves with greater magnitude in Type 2 have lower fMSRE than those with lower magnitude in Type 1. Moreover, all fMSRE for the 30-day scenarios are lower than the respective ones in the five-day scenarios.

Figure 8a shows violin plots of the relative error of the estimated coefficients associated with the explanatory variables $\gamma_1 = 13$ and $\gamma_2 = 0.0011$. One run was excluded from the plot in the balanced scenario because it showed an absolute relative error greater than 38. In all scenarios, the estimates with γ_2 have larger violins than those with γ_1. The 30-day scenarios have lower expected variability than the five-day scenario estimates, but their median reference lines above the zero line show visible underestimation of the parameter γ_2. Furthermore, Table 5 shows the mean, median and square root of the Mean Squared Relative Error (srMSRE) of the estimated parameters. Observe that parameter γ_1 has estimates with considerably lower srMSRE than γ_2. The underestimation is notable in the mean and median values of γ_2. Nevertheless, the statistics of parameter γ_2 show slight overestimation of the mean and larger srMSREs in all scenarios, especially the balanced five-day scenario, the one that presented a run with relative error greater than 38.

The estimated covariance parameters for Scenarios 1 to 4 are displayed in Figures 9 and 10 and Table 6. Figure 10 shows the estimated dispersion parameters represented over the true variance functionals, Figure 9 the violin plots of the estimated decay parameters and Table 6 the mean, median and square root of the Mean Squared Relative Error (MSRE)of the estimated decay parameters. Because the homogeneous model estimates a scalar as the dispersion parameter, the estimated values in Figure 10 are represented as constant lines over time. It seems that the horizontal lines are trying to capture an average of the variance functionals over time. In fact, taking the average of the variance functionals in Figure 10 over $t \in T$ yields 0.572 for Type 1 and 5.03 for

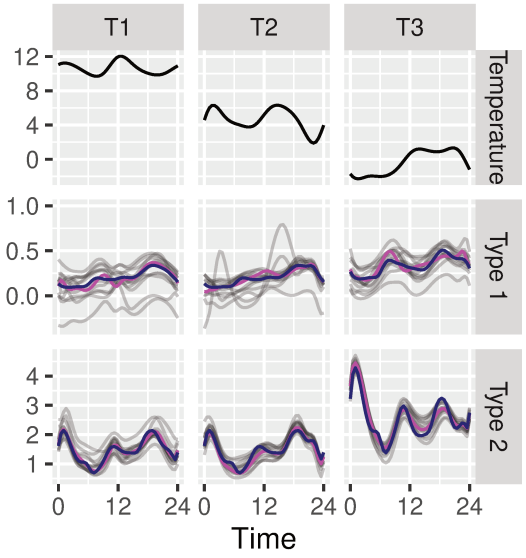
Type 2, which are close to the respective median lines at 0.6324 and 4.6375. Moreover, the visibly overestimated valuefor Type 1 and the underestimated one for Type 2 in the unbalanced five-day scenario belong to the same run. On the other hand, the estimated decay parameters show systematic underestimation for Type 2 in all scenarios, as shown in Figure 9. The reduced estimate variability for the 30-day scenarios is observed only for estimated values of ω_2. Again, the difference in magnitude of the parameters seems to have an influence on their performance, because $\omega_2 > \omega_1$. Furthermore, Table 6 shows the underestimation of ω_2 in the median and mean values and smaller srMSREs in favour of balanced markets in the five-day scenarios.

Figure 11 analogously shows the estimated typical surfaces for Scenarios 1 to 4 under the complete model fit. Again, observe that the estimated curve variability is reduced in the 30-day scenarios, especially for Type 2 under balanced markets. In addition, the advantage of balanced markets under the five-day scenarios can be seen from the lower variability of the residual curves in Figure 12 and the lower fMSRE in Table 4. The complete model does not present clear superiority in terms of fMSRE compared with the homogeneous model study.

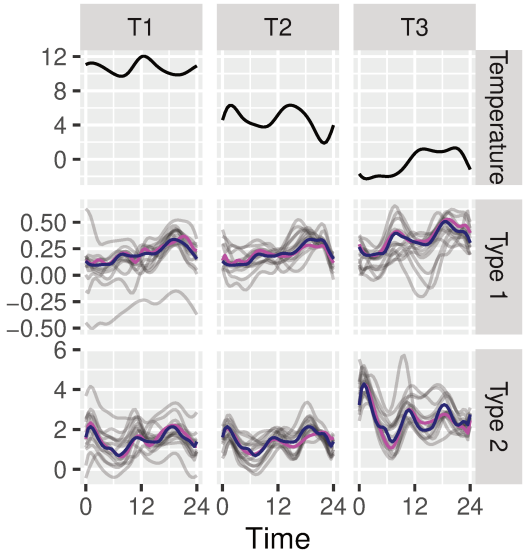
Figure 8b displays the relative errors of the estimated coefficients γ_1 and γ_2 associated with the explanatory variables. The characteristics of the violins are much like the respective ones in the homogeneous model. In fact, note that the srMSREs in Table 5 of both studies have similar values. Consequently, the complete model case shares the aspect of smaller srMSREs for estimates of γ_1, especially in the 30-day scenarios.

Finally, Figure 13 shows the estimated variance functionals of the complete model and Figure 14 their respective residual curves. As observed in the typical curves, the 30-day scenarios present lower estimate variability than the five-day scenarios. In general, the estimates capture the main features of the true curves, such as the prolonged higher values for customers of Type 1 and the decreasing values after 12 AM for Type 2. However, in some regions, the estimated curves present behaviour different from the true curve. In all scenarios, observe that the Type 1 curves begin almost at zero, whereas the true curve has a small peak with rapid decay. Moreover, in the balanced 30-day scenario, the estimated variance functionals for Type 2 customers present a nonexistent local peak at the end of the day. The violin plots of the relative errors of the estimated decay parameters are displayed in Figure 15 and their summary statistics in Table 6. Essentially, the complete model offers estimates with smaller srMSRE compared with the homogeneous model, but the underestimation of ω_2 persists.

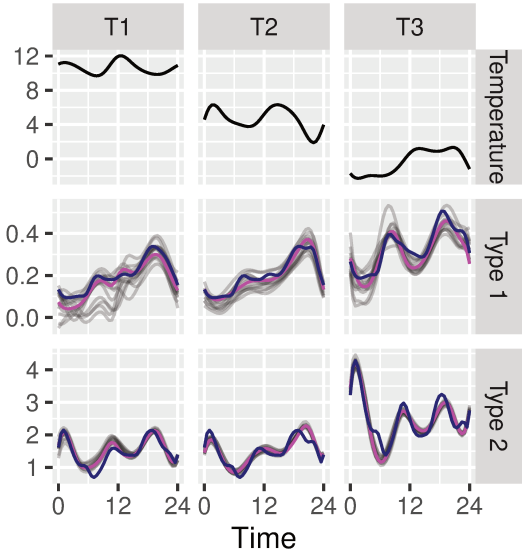
In addition, because the homogeneous model is nested in the complete model, Table 22 in Annex A shows the likelihood ratio test for all runs in every scenario. In all cases the test favours the complete model fit with p-values smaller than 0.0001.



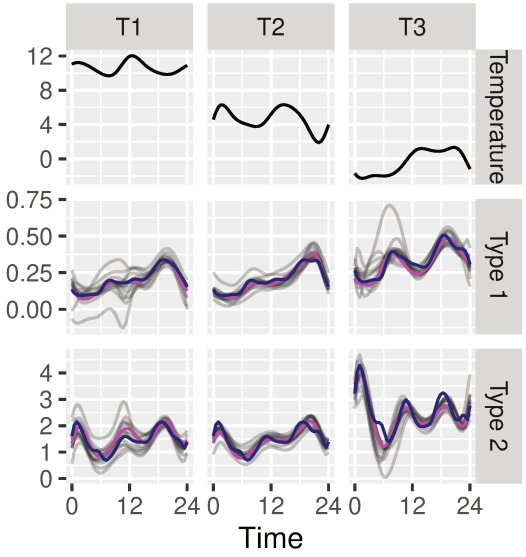
(a) Balanced market with 5-day data



(b) UnBalanced market with 5-day data

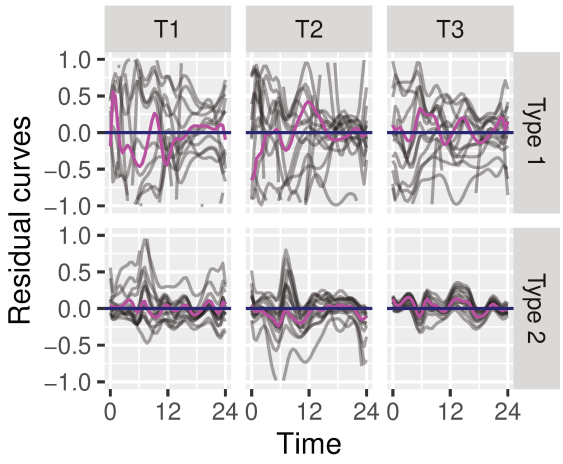


(c) Balanced market with 30-day data

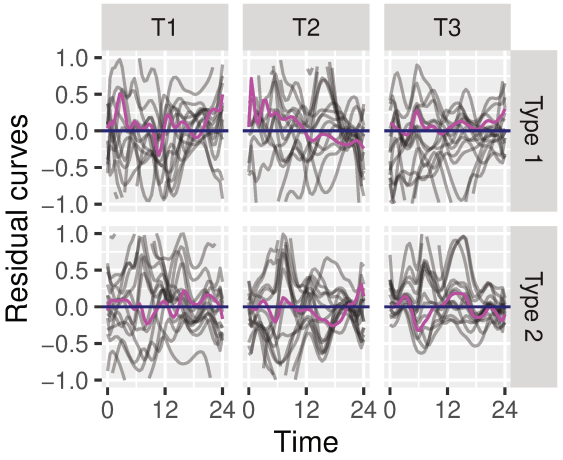


(d) Unbalanced market with 30-day data

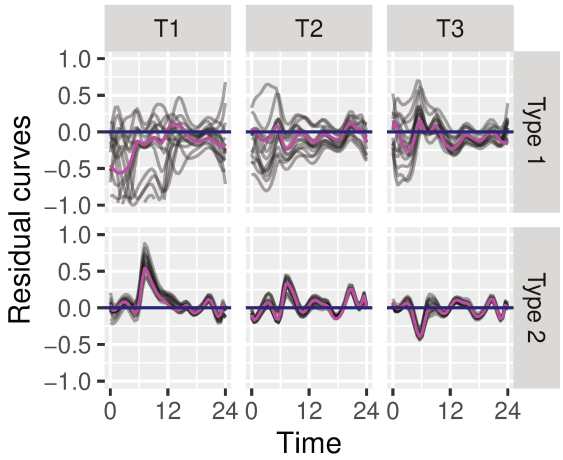
Figure 6 – At every panel, the first row represents the temperatures $T(t)$ for each temperature set T1, T2 and T3; the following rows represent the estimated typical curves of $\alpha_c(t, T(t))$ for customers of Type 1 and 2 in Scenarios 1 to 4 under the homogeneous model fit. Median depth lines are represented in magenta, true typical curves in blue and estimated typical curves for each simulation run in gray.



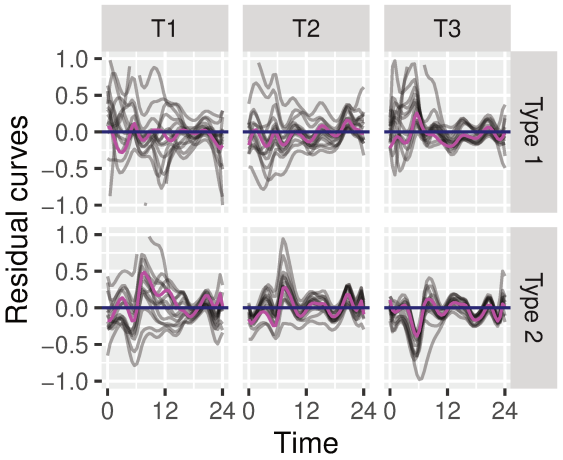
(a) Balanced market with 5-day data



(b) Unbalanced market with 5-day data



(c) Balanced market with 30-day data



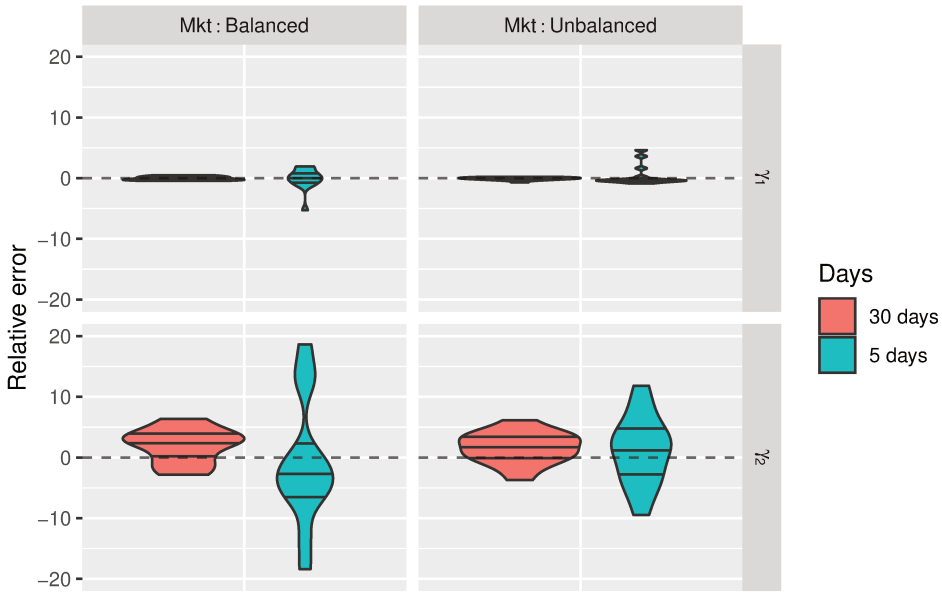
(d) Unbalanced market with 30-day data

Figure 7 – Residual curves of the estimated typical curves for Scenarios 1 to 4 in grey and their median depth in magenta under the homogeneous model fit in Figure 6.

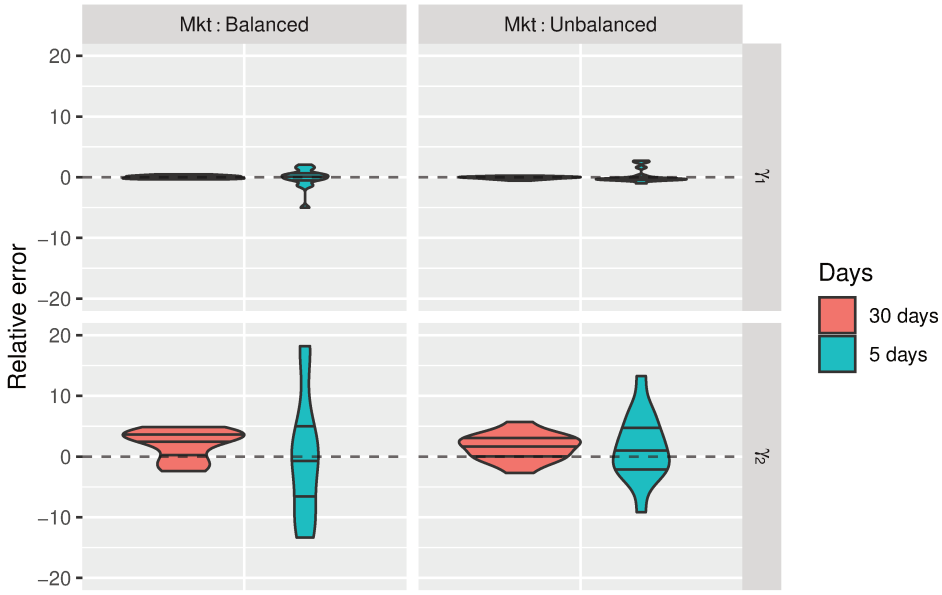
3.3.3 Discussion and conclusions

In all scenarios, the estimated typical surfaces of the homogeneous model are robust to the misspecification of the covariance structure, as shown in Figure 6. Both studies show an expected better performance for the 30-day scenarios in terms of estimation variability and fMSRE, which is also true for balanced scenarios compared to unbalanced ones.

Interestingly, it seems that the magnitude of the parameters may influence the quality of the estimate. The estimated typical curves, for example, show better estimates for customers of Type 2, the ones with higher consumption curves compared to Type 1. The same characteristic is observed in the relative errors of the estimates of γ_1, a parameter much greater than γ_2. Nonetheless, this is not as evident in the decay parameter estimation. The latter seems to be especially difficult to estimate because its performance in terms of



(a) Homogeneous covariance structure



(b) Complete covariance structure

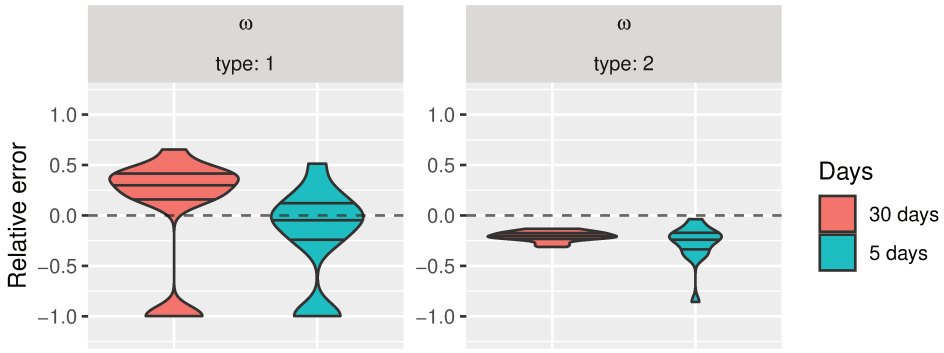
Figure 8 – Relative errors of the estimated explanatory variables coefficients, $\gamma_1 = 13$ and $\gamma_2 = 0.0011$, and their relative error distribution under a) the homogeneous model fit and b) under the complete model fit for Scenarios 1 to 4.

Table 4 – Functional mean squared relative errors of the estimated typical curves under the homogeneous (Figure 6) and complete (Figure 11) model fit for Scenarios 1 to 4.

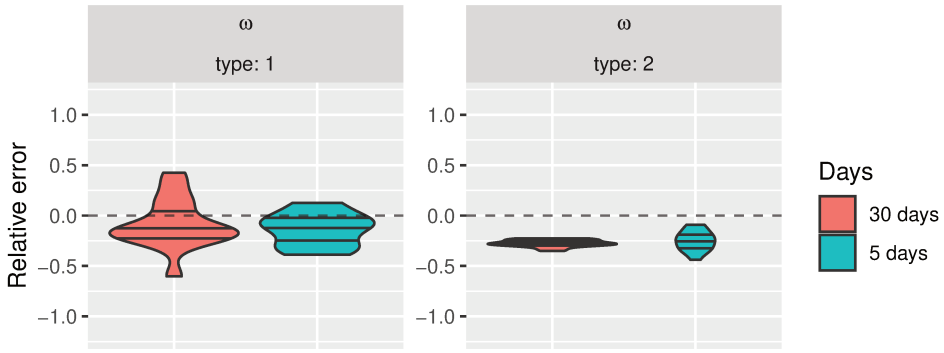
Model	Days	Type	Market balance	fMSRE
Homogeneous	5 days	Type 1	Balanced	17.6706
			Unbalanced	18.4121
		Type 2	Balanced	0.8899
			Unbalanced	4.6557
	30 days	Type 1	Balanced	1.9525
			Unbalanced	2.1867
		Type 2	Balanced	0.5814
			Unbalanced	1.0923
Complete	5 days	Type 1	Balanced	19.5322
			Unbalanced	14.9073
		Type 2	Balanced	0.9747
			Unbalanced	3.9616
	30 days	Type 1	Balanced	1.9122
			Unbalanced	1.9651
		Type 2	Balanced	0.7720
			Unbalanced	1.1850

precision and srMSRE (square root of the mean squared relative error) is not improved under 30-day scenarios or balanced markets. In fact, the estimates present a systematic underestimation of $\omega_2 = 0.70$. Still on the covariance structure, the estimated variance functionals in the complete study can capture the main features of the true ones, despite an unexpected local peak in the 30-day scenario with balanced market.

In general, the advantages of balanced markets and 30-day scenarios is evident. The complete model provides a functional variance structure that can capture different dispersions over time. However, in terms of typical surface estimation, there is no clear difference between the homogeneous and the complete model fit, which could be attributed to the fact that the least-squares estimators are unbiased independently of the covariance structure, as noted in Section 2.5.



(a) Unbalanced market



(b) Balanced market

Figure 9 – Relative errors of the estimated covariance parameters $\omega_1 = 0.03$ and $\omega_2 = 0.70$ under the homogeneous model fit for Scenarios 1 to 4.

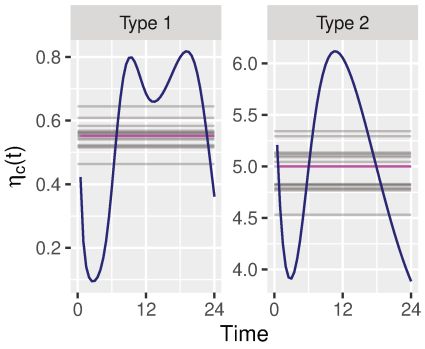
3.4 Clustering the aggregated model

This section studies the clustering approach of the aggregated data model presented in Section 2.3. The method was tested in Scenarios 5 to 8 and is presented in Table 3, where data from three clusters were simulated under the homogeneous covariance structure. In contrast to Section 3.3, typical surfaces were not considered here for the clustering model.

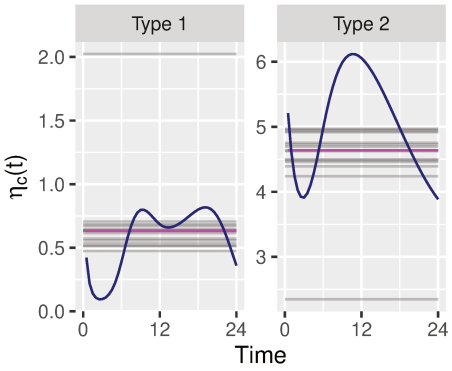
Section 3.4.1 details the clustering configuration and the true parameters. Sections 3.4.2 describes the main results and 3.4.3 contains the discussion and conclusions of this simulation study.

3.4.1 Clustering setup and true parameters

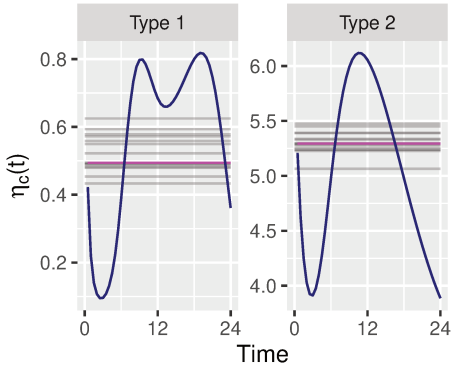
Scenarios 5 to 8 in Table 3 are made up of variations of the number of days (5 and 30) and the market (balanced and unbalanced). The remaining simulation parameters were fixed to three clusters and two types of customers observed in 12 substations every 30 minutes. The true cluster allocation is displayed in Table 7, where substations 1 to



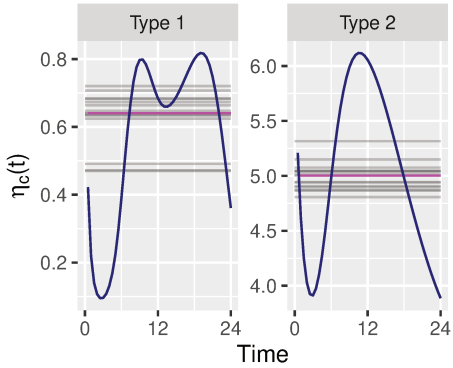
(a) Balanced market with 5-day data



(b) Unbalanced market with 5-day data



(c) Balanced market with 30-day data



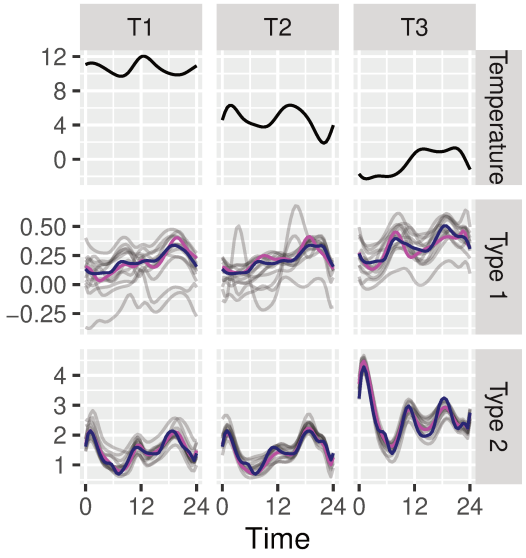
(d) Unbalanced market with 30-day data

Figure 10 – Estimated dispersion parameters for Scenarios 1 to 4 under the homogeneous model fit represented by the horizontal gray lines. Median lines are represented in magenta and the true variance functionals in blue.

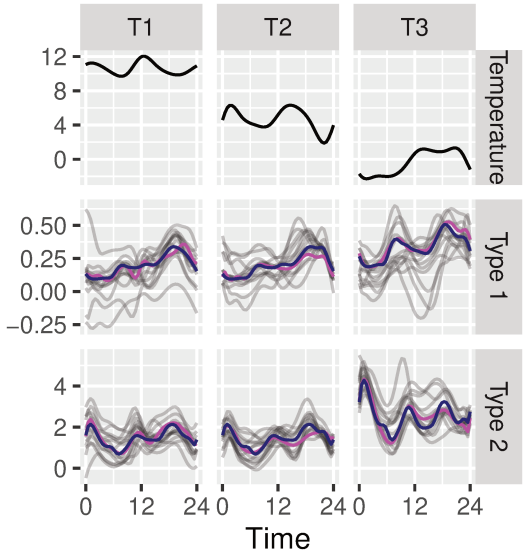
6 are assigned to Cluster 1, substations 7 to 10 to Cluster 2 and finally substations 11 and 12 to Cluster 3. The chosen covariance structure is the homogeneous one, where each customer type has its own dispersion and decay parameters.

Figure 16 shows the six typical curves divided into the two customer types for each of the three clusters. Their shapes were based on the estimated typical curves for the UK energy grid dataset in Section 4.4. Hence, Type 1 mimics the unrestricted domestic customer, with similar shapes among clusters, whereas Type 2 mimics the "Economy 7" customer.

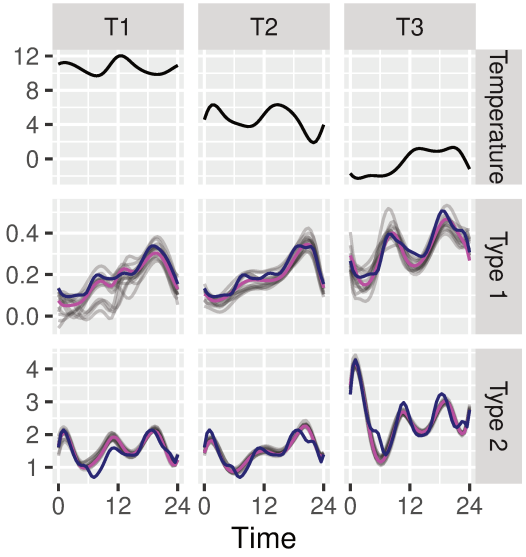
The covariance parameters that compose the homogeneous covariance structure of the simulated scenarios are presented in Table 8 divided by cluster, parameter and customer type, where, again, their values were based on the estimated covariance parameters for the UK energy grid dataset described in Section 4.4.



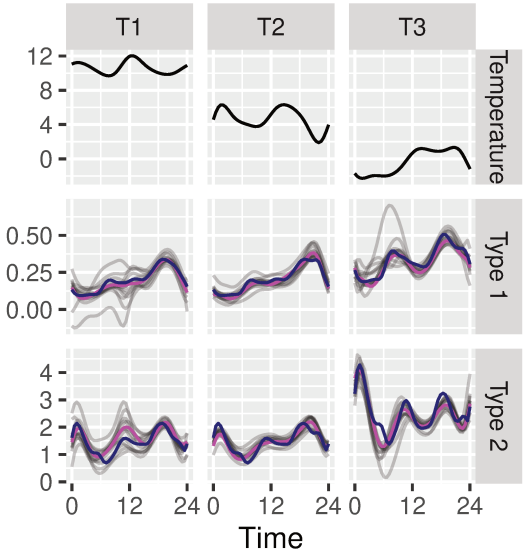
(a) Balanced market with 5-day data



(b) Unbalanced market with 5-day data

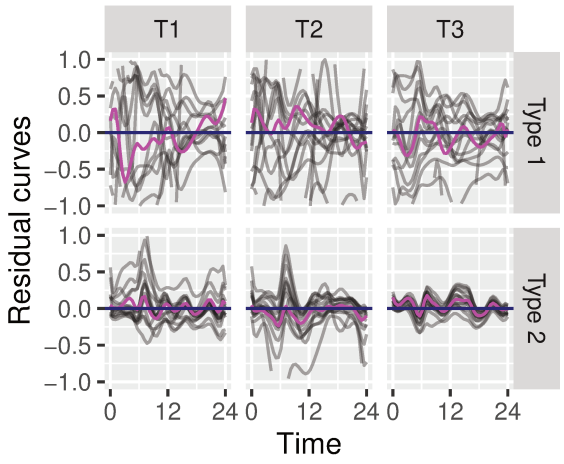


(c) Balanced market with 30-day data

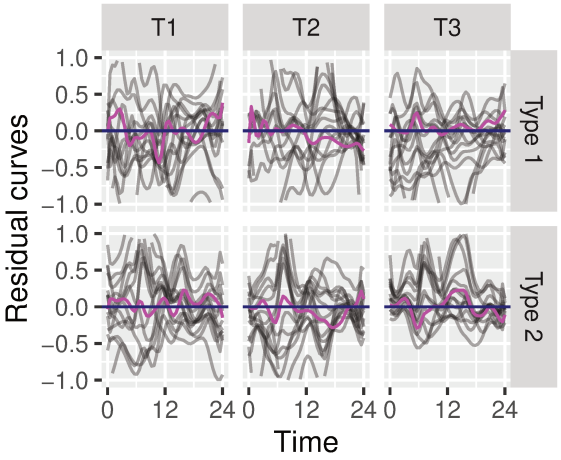


(d) Unbalanced market with 30-day data

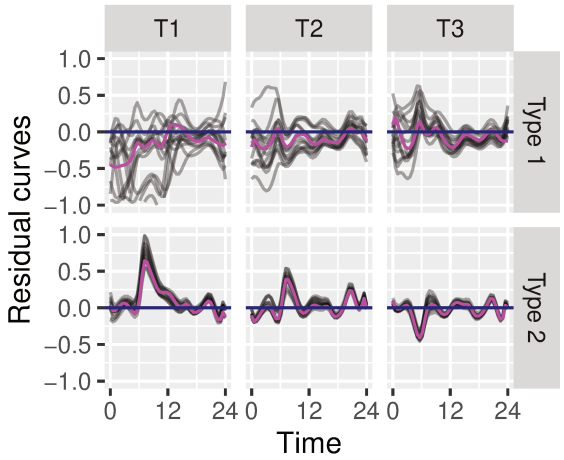
Figure 11 – In every panel, the first row represents the temperatures $T(t)$ for each temperature set T1, T2 and T3; the following rows represent the estimated typical curves of $\alpha_c(t, T(t))$ for customers of Type 1 and 2 in Scenarios 1 to 4 under the complete model fit. Median depth lines are represented in magenta, true typical curves in blue and estimated typical curves for each simulation run in gray.



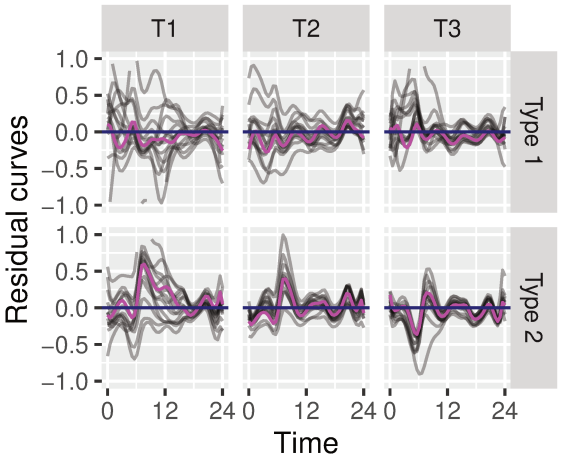
(a) Balanced market with 5-day data



(b) Unbalanced market with 5-day data



(c) Balanced market with 30-day data



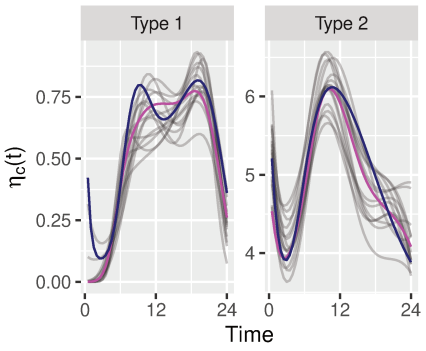
(d) Unbalanced market with 30-day data

Figure 12 – Residual curves of the estimated typical curves in grey and their median depth in magenta under the complete model fit in Figure 11.

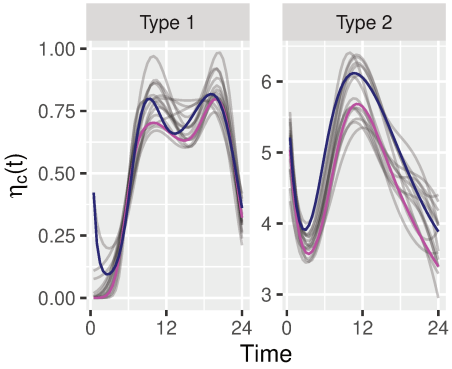
3.4.2 Results

Scenarios 5 to 8 were subjected to two model fits: clustering homogeneous aggregated data models with two and three clusters. The two-cluster fit tested how the model would perform if the number of clusters were underdetermined, that is, how the method groups substations and consequently what are the characteristics of the estimated typical curves and covariance parameters. On the other hand, the three-cluster fit evaluates model performance under correct scenarios.

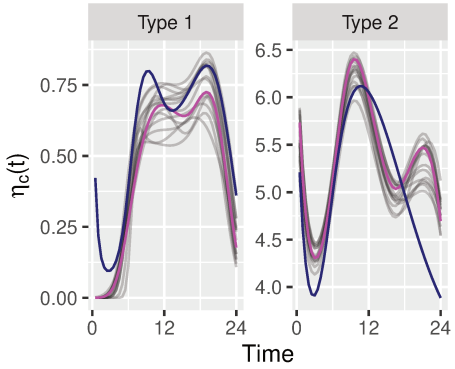
Before presenting the results, it is necessary to note the number of runs that did not converge or converged to a local maximum in this simulation at each model fit. The non-convergent runs in the two-cluster fit were the following: two runs in the unbalanced five-day market scenario, one run in the balanced five-day market scenario, and two runs in the balanced 30-day market scenario. Moreover, the three-cluster fit had



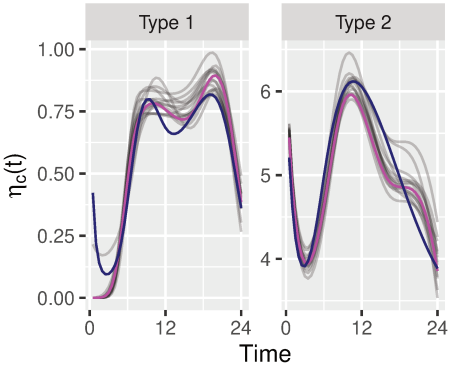
(a) Balanced market with 5-day data



(b) Unbalanced market with 5-day data

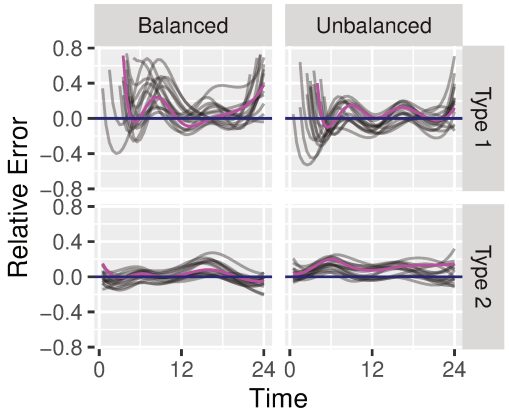


(c) Balanced market with 30-day data

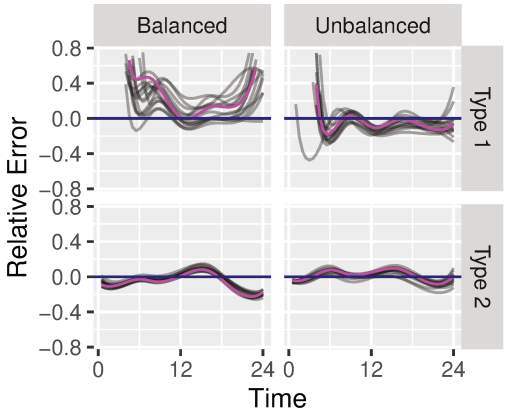


(d) Unbalanced market with 30-day data

Figure 13 – Estimated variance functionals for Scenarios 1 to 4 under the complete model fit. Median depth lines are represented in magenta, true variance functionals in blue and estimated curves in gray.

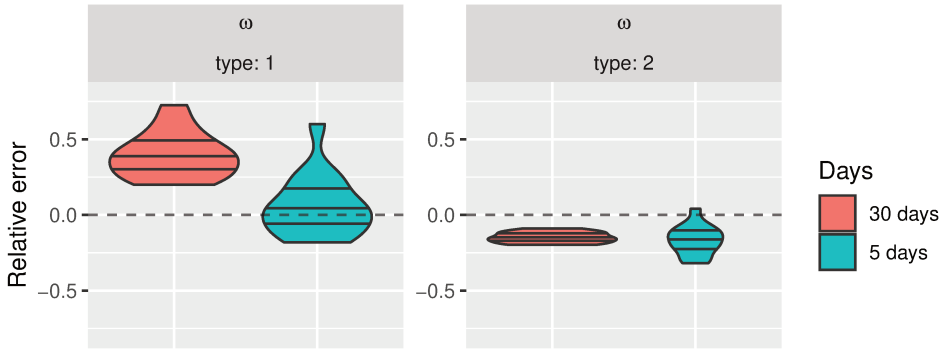


(a) 5 days

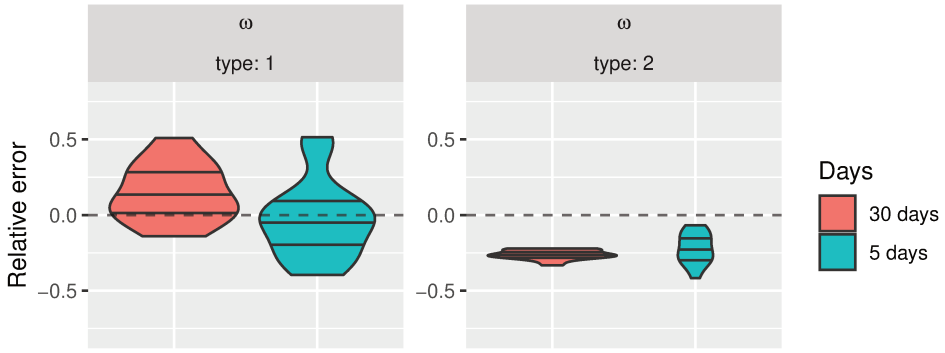


(b) 30 days

Figure 14 – Residual curves of the estimated variance functionals for Scenarios 1 to 4 under the complete model fit. Median lines are represented in magenta and residual curves in gray.



(a) Unbalanced market



(b) Balanced market

Figure 15 – Relative errors of the estimated covariance parameters $\omega_1 = 0.03$ and $\omega_2 = 0.70$ for Scenarios 1 to 4, under the complete aggregated data model fit.

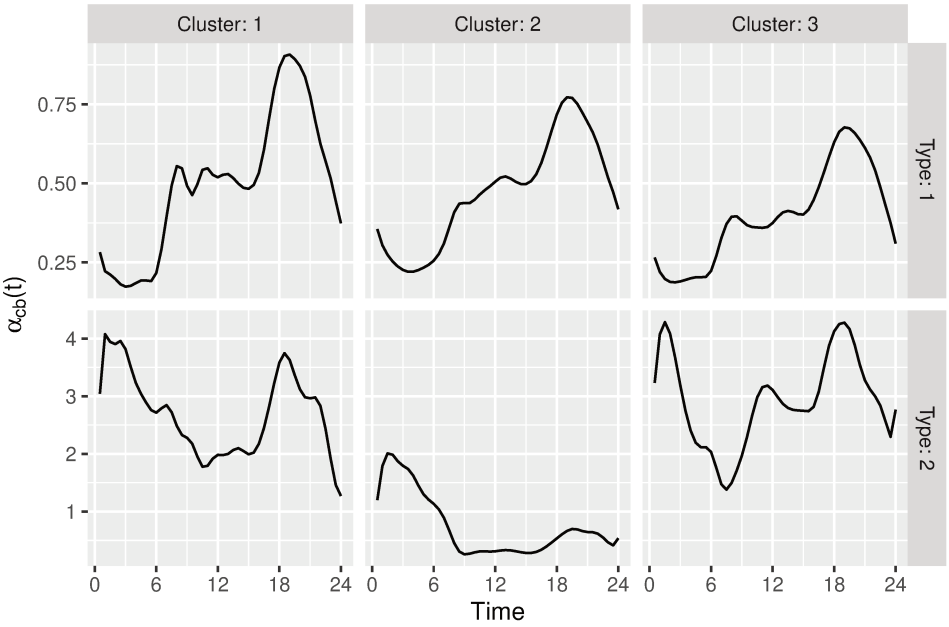


Figure 16 – True typical curves for the clustering simulation considering three clusters and two customers types.

three non-convergent runs in both balanced and unbalanced five-day market scenarios, four in the unbalanced 30-day scenario, and two in the balanced 30-day scenarios. The runs that converged to local maxima presented anomalous estimated typical curves with negative and discrepant values.

Let us begin with the two-cluster fit and its respective substation clustering as shown in Table 9. In all runs, substations are assigned with high probability to the same cluster configuration, and therefore Substations 1 to 6 were assigned to Cluster 1, Substations 7 to 10 to Cluster 2, and Substations 11 and 12 to Cluster 3. Note that the substations of true Cluster 3 were assigned to the larger Cluster 1 in the two-cluster model. Recall also that the true Clusters 1 and 3 in Figure 16 have similar typical curves for Type 1 and Type 2 and that both have approximately the same magnitude and characteristics over time, and hence it is reasonable that they merge into a single cluster in the two-cluster model.

Figure 17 shows the estimated typical curves for Scenarios 5 to 8 under the two-cluster model fit. In general, observe that Cluster 1 curves capture the main characteristics of Clusters 1 and 3: the work period stability, the 8 PM peak of Type 1 curves, and the 2 AM and 8 PM peaks of Type 2 curves. The 30-day scenarios have slightly lower estimate variability than the five-day scenarios, but note that the Type 2 curves in Cluster 1 have runs with different estimated characteristics of the work period, as shown in Figures 17c and 17d. In fact, the main difference between true Clusters 1 and 3 is the work period characterization of Type 2, and therefore it is to be expected that some runs could estimate typical curves in favour of the true Cluster 1 or Cluster 3.

Table 10 shows the summary statistics of the estimated covariance parameters of the two-cluster model fit. Because the estimated Cluster 2 substations coincide with the substations in the true Cluster 2, it is to be expected that their estimated covariance parameters are close to their true values. Observe in Table 10that the median and mean of the estimated parameters for Cluster 2 are close to their true values in the Reference column, especially for 30-day scenarios. Under five-day scenarios, balanced markets have better estimates in terms of precision. On the other hand, Cluster 1 estimates are located between the true values of true Clusters 1 and 3, and therefore the Reference column for Cluster 1 in Table 10 represents the mean of the covariance parameters of the true Clusters 1 and 3. The proximity of estimated and true covariance parameters is greater for customers of Type 2. In contrast, the Type 2 true dispersion parameters of Clusters 1 and 3 present the largest difference in Table 8, 1.54 and 5.18 respectively. Nonetheless, there is no clear evidence that the estimated covariance parameters in Cluster 1 are close to the average of the true parameters of Clusters 1 and 3.

The estimated typical curves of the three-cluster model are displayed in Figure 18 and their associated residual curves in Figure 19. In general, the median curves

show that the estimated curves capture the main characteristics of the true typical curves, although there are visible discrepant examples, more frequently seen in the unbalanced scenarios, as shown in Figures 18b and 18c. Negative values could be avoided by restricting the typical curve estimation, but to avoid overextending the computational burden of this simulation, it was decided to retain the least-squares estimators in exchange for some negative values, and also to show that in general the estimator is robust for different scenarios because the median curves are positive along the entire time axis. Once more, the residual curves show that the 30-day scenarios are more concentrated around the zero-reference line than the five-day scenarios, with even better fits for balanced scenarios. In this clustering approach, the relative residual curves for Cluster 3 in the unbalanced scenario do not concentrate around the zero-reference line, as shown in Figures 19b and 19d. Furthermore, Cluster 1 has residual curves with lower variability than Cluster 3 in all scenarios. In fact, Cluster 1 contains six substations, whereas Cluster 3 contains two substations, the minimum number required for model identifiability. It seems that the larger the number of substations in the cluster, the better is the precision, and consequently this might be the reason for the Cluster 3 overestimation of the typical curves, particularly under unbalanced scenarios.

The estimated covariance parameters of the three-cluster models are represented by their mean, median and srMSRE in Table 11. As observed in previous results throughout this chapter, smallest srMSRE are associated with parameters with larger magnitudes, for example, $\sigma_{22} = 1.28$ and $\sigma_{23} = 5.18$. n contrast, the largest srMSRE are associated with parameters with smaller magnitudes, for example, $\omega_{13} = 0.02$. In the latter case, the unbalanced scenarios presented smaller srMSRE than the balanced markets. In cases like $\omega_{22} = 0.09$, increasing the number of observation days from 5 to 30 improved srMSRE. The same behaviour was observed in most parameters, particularly those with small magnitude. On the other hand, parameters $\omega_{21} = 0.03$ and $\omega_{13} = 0.02$ were the smallest parameters in the simulation, but the srMSREs of ω_{21} were mostly around 1.5, whereas the srMSREs of ω_{13} had three values greater than 10. Recall that Cluster 3 contained only two substations. Therefore, as mentioned earlier for typical curve estimation, both the number of substations and the number of observation days are important to improve parameter estimation in each cluster.

To avoid an overextended table in this section, the comparison of BIC values between the two- and three-cluster models is presented in Table 23 in Annex A. In all cases, the BIC is favourable to the three-cluster model with differences mostly of order 10^2.

Table 11 – Mean, median and square root of the Mean Squared Relative Error (srMSRE) of the estimated covariance parameters for Scenarios 5 to 8, under the three-cluster model fit.

Parameter	Days	Market	Median	Mean	$\sqrt{MSRE}$
$\sigma_{11} = 1.54$	30 days	Balanced	1.0862	1.0114	0.6470
		Unbalanced	1.3929	1.2079	0.4677
	5 days	Balanced	0.9430	0.8281	0.6991
		Unbalanced	1.3859	1.2022	0.4923
$\omega_{11} = 0.16$	30 days	Balanced	0.3220	2.2018	3.5878
		Unbalanced	0.1295	0.3441	1.1971
	5 days	Balanced	3.9317	6.5328	6.3132
		Unbalanced	0.1274	0.9780	2.3491
$\sigma_{21} = 1.53$	30 days	Balanced	2.8842	2.9041	0.9477
		Unbalanced	2.7891	2.8831	0.9404
	5 days	Balanced	2.8640	2.9372	0.9590
		Unbalanced	2.7464	2.7632	0.9386
$\omega_{21} = 0.03$	30 days	Balanced	0.0759	0.0777	1.2608
		Unbalanced	0.0960	0.0985	1.5111
	5 days	Balanced	0.0736	0.0807	1.3000
		Unbalanced	0.1057	0.4043	3.5322
$\sigma_{12} = 1.07$	30 days	Balanced	1.0274	1.0346	0.4578
		Unbalanced	1.0810	1.1472	0.3197
	5 days	Balanced	0.6081	0.6888	0.7774
		Unbalanced	1.2345	1.2259	0.4379
$\omega_{12} = 0.12$	30 days	Balanced	0.1208	0.6763	2.1698
		Unbalanced	0.1226	0.1344	0.4184
	5 days	Balanced	1.4476	2.4465	4.4133
		Unbalanced	0.1083	0.5807	2.0123
$\sigma_{22} = 1.28$	30 days	Balanced	1.5417	1.5858	0.4888
		Unbalanced	1.2823	1.2391	0.4312
	5 days	Balanced	1.5092	1.5814	0.4852
		Unbalanced	0.0363	0.5207	0.9565
		Balanced	0.0960	0.0965	0.2692

Table 11 (continued)

Parameter	Days	Market	Median	Mean	$\sqrt{MSRE}$
$\omega_{22} = 0.09$	30 days	Unbalanced	0.1057	0.4156	1.9330
	5 days	Balanced	0.0877	0.0884	0.2444
		Unbalanced	5.2492	5.6967	7.8929
$\sigma_{13} = 0.43$	30 days	Balanced	0.3390	1.4096	1.7326
		Unbalanced	1.3438	1.4685	1.5540
	5 days	Balanced	0.0282	0.4760	1.1353
		Unbalanced	1.0871	1.0075	1.3569
$\omega_{13} = 0.02$	30 days	Balanced	3.9775	3.2759	12.7591
		Unbalanced	0.1446	0.3754	4.2157
	5 days	Balanced	9.9767	10.3302	22.7049
		Unbalanced	0.9252	2.7126	11.6030
$\sigma_{23} = 5.18$	30 days	Balanced	5.2455	5.0097	0.3249
		Unbalanced	3.9596	3.3274	0.5980
	5 days	Balanced	4.7792	4.6787	0.3502
		Unbalanced	4.0859	3.9616	0.4917
$\omega_{23} = 0.37$	30 days	Balanced	0.2624	0.2492	0.5713
		Unbalanced	0.2968	0.6871	1.1099
	5 days	Balanced	0.1930	0.1960	0.6859
		Unbalanced	0.2759	0.8170	1.3016

3.4.3 Discussion and conclusion

In both fitted models, substations are allocated to the same clusters throughout the series of runs. In the two-cluster model, Substations 11 and 12, which belong to the true Cluster 3, are always assigned to Cluster 1 together with Substations 1 to 6. In this case, with an underdetermined number of clusters, the method groups the clusters with more similarity. Hence the estimated typical curves for Cluster 1 still capture the main features of the true curves for Clusters 1 and 3. Similarly, the estimated covariance parameters for Cluster 1 present values between the true covariance parameters of the true Clusters 1 and 3. In the three-cluster model, substations are assigned to the correct cluster. Consequently, except for some cases under unbalanced scenarios, estimated typical curves for this model are well located around their true curves. In general, 30-day scenarios have less dispersed estimates than five-day scenarios, and balanced markets have less dispersed

estimates than unbalanced ones.

The differences among scenarios have a similar impact on the estimation of covariance parameters. In addition, there is evidence that the number of substations in a cluster is crucial to estimation performance, particularly for small-magnitude parameters. The positive impact of increasing the number of substations on parameter estimation is shown in (LENZI et al., 2017). Two parameters with small values for Clusters 1 and 3 have distinct srMSRE probably because the information available for Cluster 3 estimation is less than for Cluster 1.

When comparing both models, the three-cluster model presents lower BIC than the two-cluster model in all cases.In a real-world problem, where the true number of clusters is unknown, the BIC can be a useful tool to decide between models.

Users of the clustering aggregated data model are encouraged to be careful with the estimated covariance parameters and to try multiple models with different numbers of clusters using these estimated values as input for their initial values. For example, the estimated values of the aggregated two-cluster data model can be used as an input to fit the aggregated three-cluster data model by repeating one of the results. As shown in Section 2.5, after multiple fits, the user can compare the models using the likelihood test ratio to help decide which model is the most adequate to the data.

Table 5 – Mean, median and square root of the Mean Squared Relative Error (MSRE) of the estimated explanatory variables parameters under the homogeneous and complete model fit for Scenarios 1 to 4.

Model	Parameter	Days	Market	Mean	Median	$\sqrt{MSRE}$
Homogeneous	$\gamma_1 = 13$	30 days	Balanced	12.7193	12.0985	0.3054
			Unbalanced	11.8776	12.1702	0.2516
		5 days	Balanced	12.0501	12.1832	1.7045
			Unbalanced	17.2909	8.3669	1.6263
	$\gamma_2 = 0.0011$	30 days	Balanced	0.0330	0.0421	3.4016
			Unbalanced	0.0289	0.0328	2.9574
		5 days	Balanced	0.0156	-0.0159	14.6551
			Unbalanced	0.0231	0.0334	5.6259
Complete	$\gamma_1 = 13$	30 days	Balanced	13.3503	12.9929	0.2578
			Unbalanced	11.8177	12.1773	0.2295
		5 days	Balanced	11.3529	15.1660	1.6096
			Unbalanced	15.3926	9.1294	1.1091
	$\gamma_2 = 0.0011$	30 days	Balanced	0.0330	0.0490	3.0979
			Unbalanced	0.0283	0.0337	2.5957
		5 days	Balanced	0.0254	0.0104	15.3070
			Unbalanced	0.0274	0.0264	5.4715

Table 6 – Mean, median and square root of the Mean Squared Relative Error (MSRE) of the estimated decay parameters for Scenarios 1 to 4, under the homogeneous model fit.

Model	Parameter	Days	Market	Median	Mean	$\sqrt{MSRE}$
Homogeneous	$\omega_1 = 0.03$	30 days	Balanced	0.0257	0.0275	0.2549
			Unbalanced	0.0382	0.0322	0.5544
		5 days	Balanced	0.0270	0.0259	0.2107
			Unbalanced	0.0287	0.0764	6.6934
	$\omega_2 = 0.70$	30 days	Balanced	0.5028	0.5079	0.2764
			Unbalanced	0.5520	0.5562	0.2111
		5 days	Balanced	0.5226	0.5251	0.2675
			Unbalanced	0.5427	0.5073	0.3318
Complete	$\omega_1 = 0.03$	30 days	Balanced	0.0316	0.0298	0.4243
			Unbalanced	0.0419	0.0418	0.4208
		5 days	Balanced	0.0288	0.0291	0.2516
			Unbalanced	0.0312	0.0318	0.2065
	$\omega_2 = 0.70$	30 days	Balanced	0.5169	0.5178	0.2621
			Unbalanced	0.5899	0.5962	0.1521
		5 days	Balanced	0.5323	0.5450	0.2418
			Unbalanced	0.5819	0.5831	0.1917

Table 7 – True cluster assignment for each substation in the simulation study.

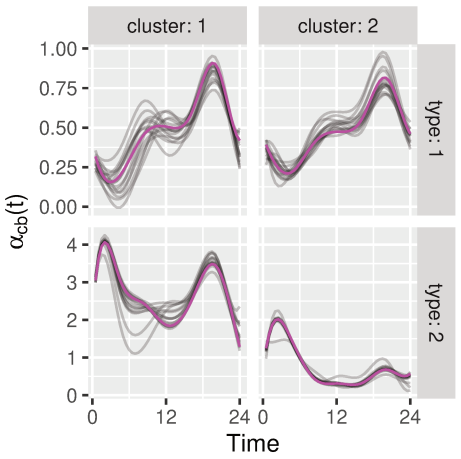
Substation	1	2	3	4	5	6	7	8	9	10	11	12
True cluster	1	1	1	1	1	1	2	2	2	2	3	3

Table 8 – True covariance parameters for clustering simulation considering three clusters and two customer types.

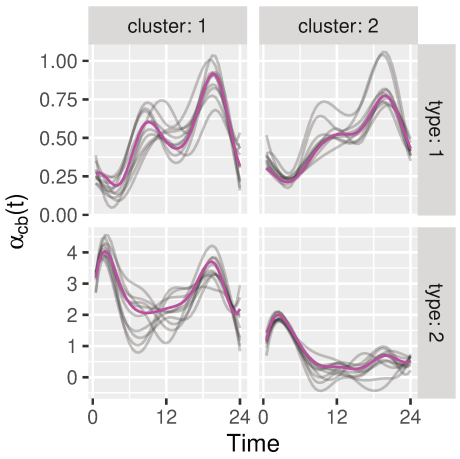
Cluster	Parameter	Type	Value
$b = 1$	σ_{cb}	$c = 1$	1.54
		$c = 2$	1.53
	ω_{cb}	$c = 1$	0.16
		$c = 2$	0.03
$b = 2$	σ_{cb}	$c = 1$	1.07
		$c = 2$	1.28
	ω_{cb}	$c = 1$	0.12
		$c = 2$	0.09
$b = 3$	σ_{cb}	$c = 1$	0.43
		$c = 2$	5.18
	ω_{cb}	$c = 1$	0.02
		$c = 2$	0.37

Table 9 – Cluster allocation of the 12 substations under the clustering models with two and three clusters. The proportion of runs assigned to that cluster are 100% in all runs in both model fit.

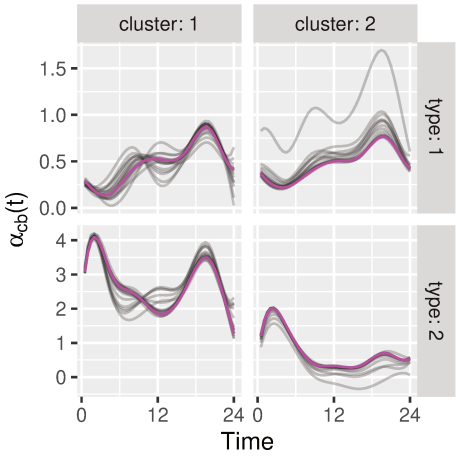
Substation	True	two-cluster fit	three-cluster fit
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	1	1	1
6	1	1	1
7	2	2	2
8	2	2	2
9	2	2	2
10	2	2	2
11	3	1	3
12	3	1	3



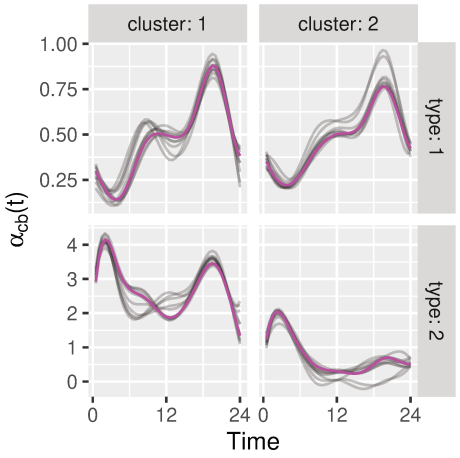
(a) Balanced market with 5-day data



(b) Unbalanced market with 5-day data



(c) Balanced market with 30-day data

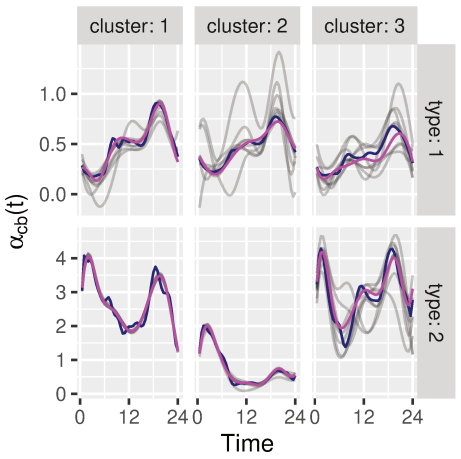


(d) Unbalanced market with 30-day data

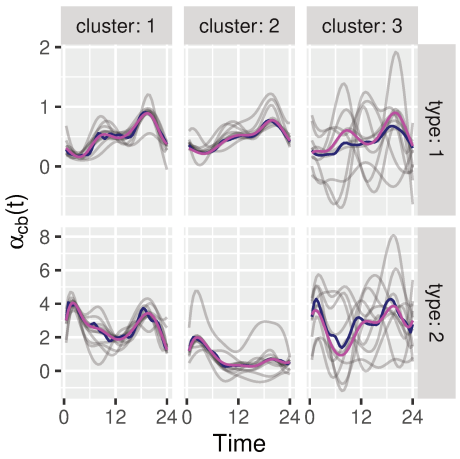
Figure 17 – Estimated typical curves for Scenarios 5 to 8, represented by the combination of market balance and number of days, under the two-cluster model. Median curves are represented in magenta and estimated typical curves in gray.

Table 10 – Summary statistics for the estimated covariance parameters for Scenarios 5 to 8 under the two-cluster model fit. The reference column for Cluster 1 (parameter subindex ending in 1) is the mean value between covariance parameters of the true Clusters 1 and 3 and for Cluster 2 (parameter subindex ending in 2) is the true covariance parameters for the true Cluster 2.

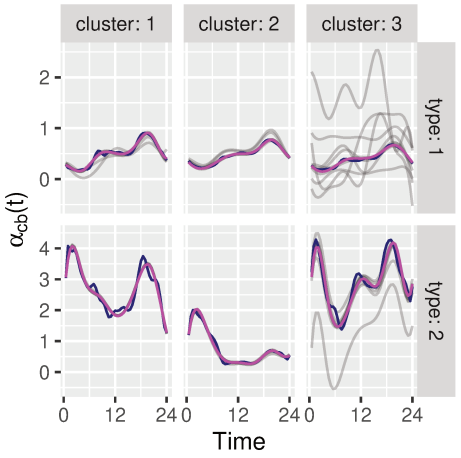
Parameter	Days	Market	Ref	Median	Mean	Std Dev
σ_{11}	30 days	Balanced	0.985	1.3690	1.1717	0.6952
		Unbalanced	0.985	1.4175	1.4324	0.1123
	5 days	Balanced	0.985	1.2360	1.0226	0.5916
		Unbalanced	0.985	1.3903	0.9552	0.7257
ω_{11}	30 days	Balanced	0.090	0.1863	1.7907	3.2170
		Unbalanced	0.090	0.1303	0.1457	0.0323
	5 days	Balanced	0.090	0.2262	2.4892	3.6509
		Unbalanced	0.090	0.4885	1.7235	2.2490
σ_{21}	30 days	Balanced	3.355	3.3149	3.7087	1.0352
		Unbalanced	3.355	3.8129	3.3212	0.7789
	5 days	Balanced	3.355	4.0515	4.0523	0.7751
		Unbalanced	3.355	3.7831	4.1194	0.9537
ω_{21}	30 days	Balanced	0.200	0.1247	0.1559	0.0653
		Unbalanced	0.200	0.1859	0.1728	0.0768
	5 days	Balanced	0.200	0.1983	0.1848	0.0644
		Unbalanced	0.200	0.2595	0.2048	0.0891
σ_{12}	30 days	Balanced	1.070	1.1663	1.2281	0.3476
		Unbalanced	1.070	1.0870	1.1621	0.1524
	5 days	Balanced	1.070	1.0906	1.1274	0.3569
		Unbalanced	1.070	1.2591	1.1776	0.3398
ω_{12}	30 days	Balanced	0.120	0.1290	0.3108	0.7708
		Unbalanced	0.120	0.1143	0.1152	0.0071
	5 days	Balanced	0.120	0.1080	0.3592	1.0368
		Unbalanced	0.120	0.1134	1.0093	2.3378
σ_{22}	30 days	Balanced	1.280	1.4533	1.3147	0.3727
		Unbalanced	1.280	1.3541	1.2999	0.8166
	5 days	Balanced	1.280	1.4583	0.9738	0.7445
		Unbalanced	1.280	0.3949	0.8713	1.0385
ω_{22}	30 days	Balanced	0.090	0.0958	0.3894	0.8949
		Unbalanced	0.090	0.1233	1.4038	2.7188
	5 days	Balanced	0.090	0.0998	2.2790	3.1786
		Unbalanced	0.090	3.6099	4.7933	4.9760



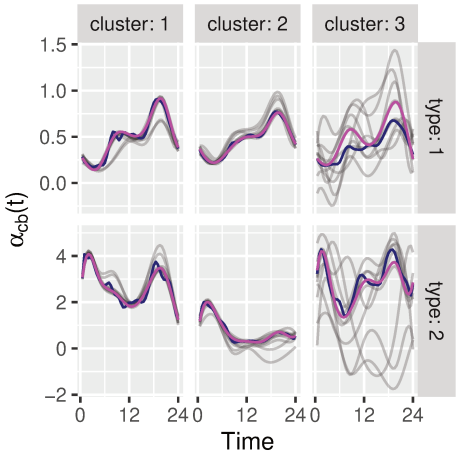
(a) Balanced market with 5-day data



(b) Unbalanced market with 5-day data

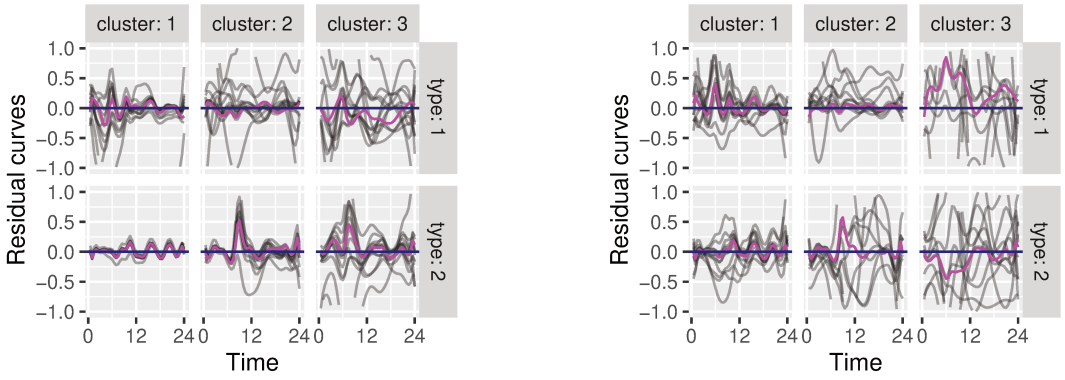


(c) Balanced market with 30-day data



(d) Unbalanced market with 30-day data

Figure 18 – Estimated typical curves for Scenarios 5 to 8, represented by the combination of market balance and number of days, under the three-cluster model. Median curves are represented in magenta and estimated typical curves in gray.



(a) Balanced market with 5-day data

(b) Unbalanced market with 5-day data

(c) Balanced market with 30-day data

(d) Unbalanced market with 30-day data

Figure 19 – Residual curves for Scenarios 5 to 8, represented by the combination of market balance and number of days, under the three-cluster model. Median curves are represented in magenta and estimated typical curves in gray..

4 Analysis of UK electrical substation data

This chapter applies the proposed clustering and full aggregated data models to a real data set containing electrical load profiles from energy substations in the United Kingdom. The first step is to fit the simple aggregated data model described in Section 2.1 to estimate the covariance parameters to be used as initial parameters in the full aggregated data model described in Section 2.2, where air temperature data are used as an additional functional component to create a typical surface to explain customer energy consumption under different weather configurations. Finally, the model-based clustering approach described in Section 2.3 is applied to investigate how substations are grouped according to their estimated typical curves.

Section 4.1 presents the dataset characteristics and an exploratory data analysis. Section 4.2 presents the results of the simple aggregated model fit, followed by the full aggregated model in Section 4.3. Section 4.4 describes the clustering analysis results.

4.1 The dataset

Provided by Professor Gavin Shaddick, the dataset was analysed in two articles co-authored by him: one to cluster and classify substations (LI et al., 2015a) and the other to estimate peak-loads using clusterwise regression (LI et al., 2015b). The data contain information on electrical load profiles observed every 10 minutes across 407 electrical energy substations in the northwest portion of the United Kingdom. Observations were taken from October 28, 2012, to March 30, 2013, for a total of 154 days.

Each substation supplies energy to up to eight types of customers. This eight-customer division dates from the 1990s and is organized as two domestic types, unrestricted and "Economy 7"; two non-domestic types, also unrestricted and "Economy 7"; and four non-domestic classes of maximum-demand customers according to their peak-load factor. This distribution has proven to be inefficient because a small delicatessen and a supermarket can be assigned to the same customer type (WILKS, 2010). The variability of non-domestic groups makes the aggregated data model unsuitable because there is no typical curve that could, for example, represent both a supermarket and a small delicatessen. Hence, to apply the proposed model, only a subset of the data, consisting of substations with only two types of domestic customers, was considered, resulting in a data set with 12 substations and the following two types of customers:

- C1: Non restricted domestic customers;
- C2: Economy 7 domestic customers.

Type C2, the "Economy 7" domestic customers, corresponds to a differential tariff provided by United Kingdom electricity suppliers with cheaper electricity during the off-peak periods.

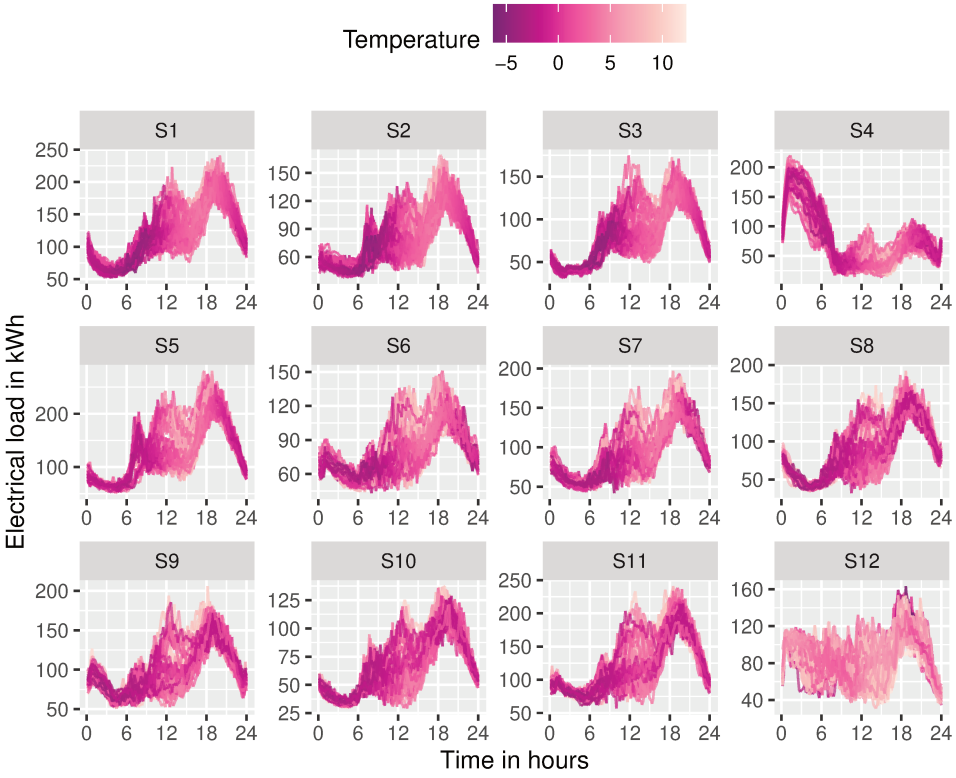
Only working days were considered in the dataset to remove the weekend effect on electrical energy consumption because it is possibly different than the domestic routine between Monday and Friday. Also, to avoid variability during the Christmas and New Year holidays, observations from January 3 were used instead.

Temperature measurements were obtained through the API of the World Weather Online (worldweatheronline.com) Web site, using the substation primary, generally representing a community or a district in Wales, as the location reference, which will be detailed in Section 4.1.1. The downloaded historical weather data, however, contain observations only every three hours. Hence, to achieve the same observation frequency of 10 minutes as in the electrical load dataset, a cubic B-spline interpolation was performed to incorporate temperature as a functional variable in the full aggregated data model.

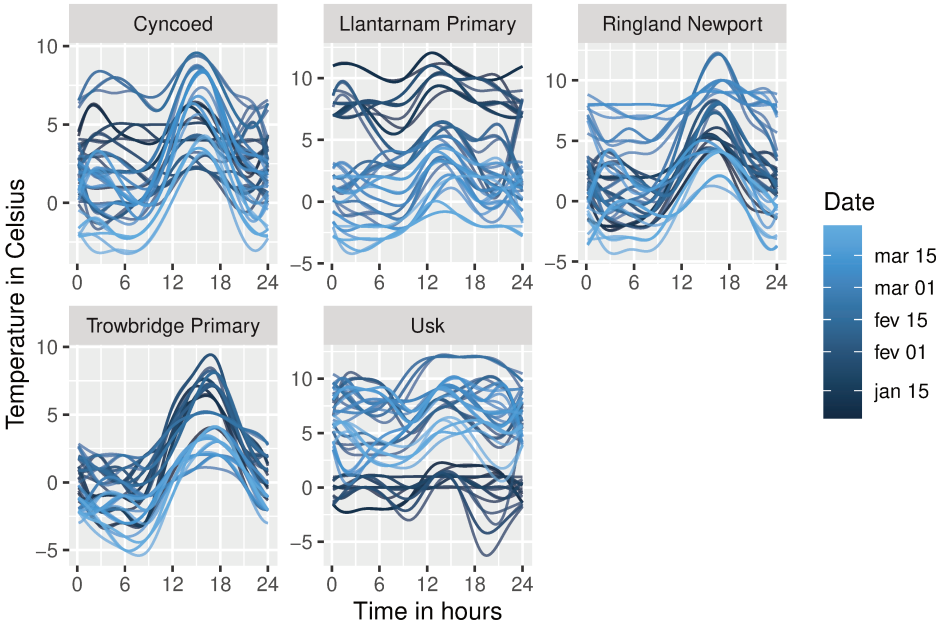
4.1.1 Exploratory data analysis

Figure 20a provides a visualization of the electrical load profiles corresponding to the 61 days from January 3 to March 30, 2013, for each one of the 12 substations, coloured according to the temperature scale located above the panel. Figure 20b shows the observed temperature at the five substation primaries (see Table 12) for the same 61 days as in Figure 20a. The associated market of each substation is displayed in Table 12, which shows that the great majority of the customers are unrestricted domestic customers (C1), dominating more than 90% of the market in 10 of 12 substations, whereas substation S4 is the only one with a majority of "Economy 7" domestic customers (C2), representing 80.73% of its market.

Except for S4 and S12, all substations presented a similar pattern. Early morning showed the lowest energy consumption until approximately 9 AM, with apparently homogeneous variance during this period. The period between 10 AM and 4 PM showed the largest variability, probably because this is the period when people tend to leave their houses to work, but some stay at a home office, for example. From 5 PM onward, the variability apparently stabilized again, even at the load peak at 7 PM. However, substations S4 and S12 not only did not follow this pattern, but also were distinct one from the other. Their peaks, at late night in substation S12 and before sunrise in S4, probably occur because of lower tariffs at night, encouraging energy consumption outside daytime.



(a) Electrical load profile



(b) Temperature in substation primaries

Figure 20 – (a) Electrical load profile data in kWh from 12 substations color-coded by the current air temperature in Celsius and (b) the observed temperature in the five primaries over 61 winter days in the United Kingdom observed every 10 minutes.

Exceptionally, substation S12 has one-third of its market consisting of customers of type C2, which results in relatively higher loads in early morning compared to substations with a great majority of C1 customers, and more variability before 9 AM.

Many other factors such as humidity, precipitation, and temperature may influence load profiles. Figure 20a is color-coded according to the observed air temperatures shown in Figure 20b. There is no clear visual evidence of the impact of temperature on electrical load consumption, although the information might be useful to explain the variability of work periods. Indeed, most substations experience temperature fluctuations during the daytime, whereas night periods tend to be more stable. Nonetheless, substation S12 is an exception for the temperature pattern as well; it is the only one that shows higher temperature values both day and night.

The particularity of substation S12 may be explained by its geographic location within the Usk primary, shown in the Google Maps frame in Figure 21. S12 is in the town of Monmouth, in the countryside of Wales, with a population of less than three thousand, and is the smallest of all the primaries. To the southwest is Llantarnam, a community in the suburb of Cwmbran, with a population size slightly larger than four thousand, where substations S8 to S11 are located, all consisting mostly of customers of type C1. Not far away is Ringland, in the city of Newport, where substations S6 and S7 are located, with populations approximately double that of Llantarnam. Closer to the capital of Wales, there are two primaries: Trowbridge and Cyncoed. Both are in communities with a population greater than ten thousand (16,194 and 11,148, respectively) located in the urban area of Cardiff Central. In fact, Cyncoed, the only substation with a majority of C2 customers, has some of the highest property prices in the country. All cited demographic data are

Table 12 – Primaries, substation names, substation IDs and number of customers of types C1 and C2.

Primary	Substation	Number	C1	C2
Trowbridge Primary	S1	512017	228	3
	S2	512050	146	5
	S3	512051	151	5
Cyncoed	S4	513044	21	88
	S5	513049	218	7
Ringland Newport	S6	531834	155	17
	S7	531835	194	12
Llantarnam Primary	S8	532204	173	9
	S9	532205	163	12
	S10	532206	158	2
	S11	532207	244	10
Usk	S12	535445	46	23

Figure 21 – Geographic location of substation primaries in United Kingdom: (1) Trowbridge, (2) Cyncoed, (3) Ringland, (4) Llantarnam and (5) Usk.

available in the 2011 census of the United Kingdom (STATISTICS, 2016).

4.1.2 Data modelling

The simplest homogeneous aggregated data model provides estimates of the typical curves for unrestricted and "Economy 7" domestic customers supplied by the 12 electrical energy substations and also reasonable estimates of the covariance parameters because the 61 days can be considered as replicates. The results of the simple aggregated data model can be used as initial values in the full model approach to obtain the typical response surface using the observed temperature at substation primaries as the additional functional component.

Furthermore, model-based clustering analysis can be an alternative to model load profile variability and can also be used to investigate how substations can be grouped according to their typical curves. The dataset consists of substations from different locations, each with their own specific weather patterns, population characteristics, and probably different energy consumption habits as well.

The next sections describe the application of the simple homogeneous aggregated data model, followed by the full and clustering models.

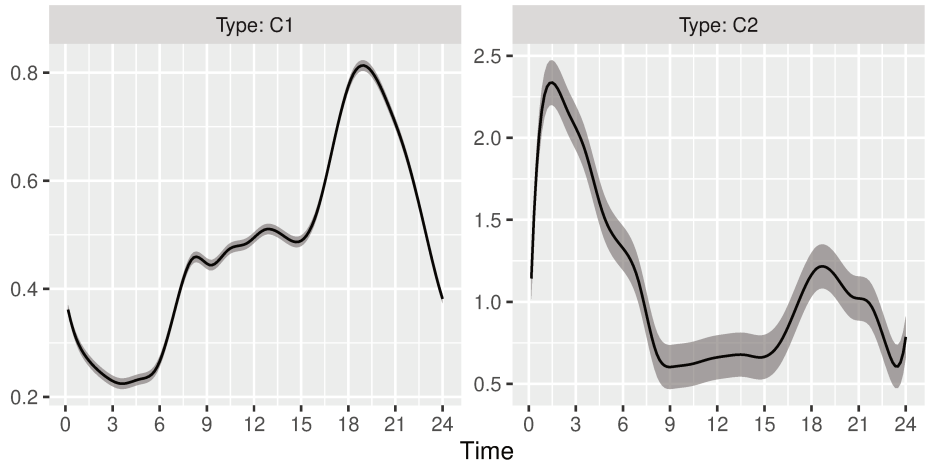


Figure 22 – Estimated typical curves in kWh and their confidence band (in gray) for unrestricted (C1) and "Economy 7" (C2) domestic customers using an homogeneous aggregated model.

4.2 Simple homogeneous aggregated data model

The simple homogeneous aggregated data model described in Section 2.1 assumes the same homogeneous dispersion and decay parameters for all customer types. This might be a naive approach, but its results can be used as initial values for the full model, drastically improving its computational performance.

Figure 22 displays the estimated typical curves considering 24 cubic B-Splines functions expansion. The unrestricted domestic (C1) customers consume less energy than "Economy 7" domestic (C2) customers. The typical consumption curve of C1 customers shows modest values early in the morning, rising to a higher baseline in the traditional work period between 9 AM and 4 PM to finally reach their peak at 7 PM and complete the cycle by slowly returning to low consumption late at night. On the right panel, the typical curve for C2 customers is almost a mirror image of C1: the curve has its peak right after midnight and is constantly decreasing until it reaches its lowest values at 9 AM, when the cheaper tariffs cease. Later, there is a local peak around 7 PM, higher than C1, but still considerably lower than the early morning peak.

Because both these customer types are domestic customers, certain behaviours can be proposed to justify their typical curves. For example, unrestricted customers seem to have the habit of getting up in the morning and turning on electrical appliances that increase their load values, such as microwaves and hairdryers, for example. The work period presents many possibilities: most people leave their houses to go to work, decreasing home energy consumption, but some household members may stay at home to work in a home office. At night, when people arrive from their jobs, the appliances that are now turned on have higher energy consumption, such as washing machines and dryers that were not used in the morning. In contrast, the C2 typical curve has its major peak right after midnight,

Table 13 – Estimated covariance parameters of the simple homogeneous aggregated data model for the UK electrical energy dataset.

Parameter	Type	Value	95% Confidence Interval
σ_c	C1	0.6608	(0.6452, 0.6764)
	C2	5.6094	(5.4494, 5.7693)
ω_c	C1	0.0404	(0.0384, 0.0425)
	C2	0.8205	(0.7721, 0.8689)

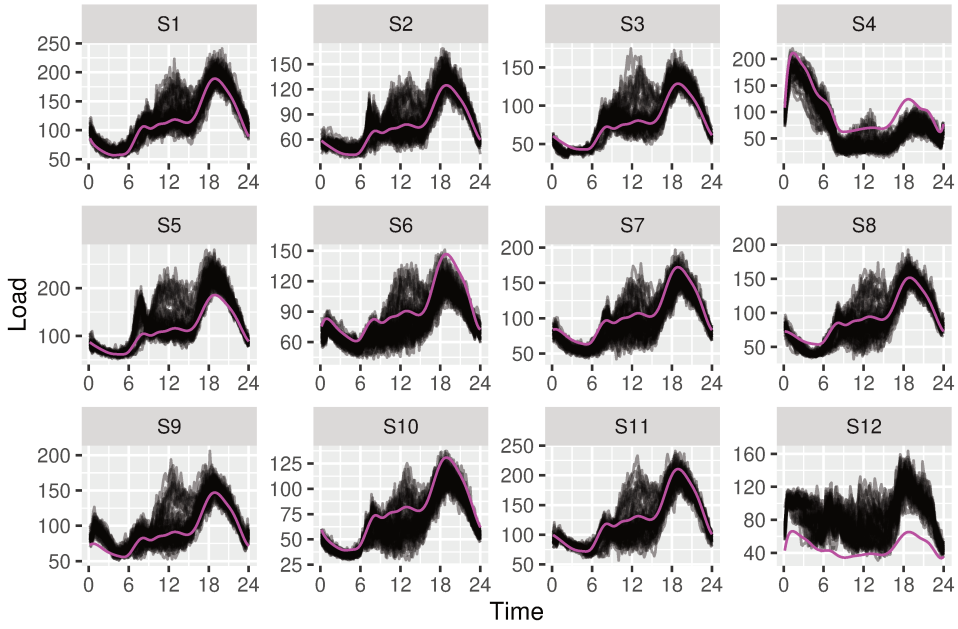
maybe due to appliances with higher energy cost making use of cheaper tariffs. From 9am onward both typical curves increase their loads up to a plateau until approximately 4pm, where the consumption rapidly increase to the local peak. Furthermore, the confidence band for the C2 typical curve is larger than for C1 because the C1 class contains most of the market share (around 90% or more) in most substations (see Table 12), and consequently the amount of information available to estimate the C1 typical curve is greater than for C2.

The estimated covariance parameters for the homogeneous aggregated data model are displayed in Table 13. The dispersion parameter for C2 is considerably greater than for C1, with larger confidence bands in Figure 22. The small decay parameter for C1 indicates that correlation between energy consumption at two distinct points in time decays faster for C1 than for C2. This means that, given the same time window, energy consumption in C2 has a stronger dependence on values in its time neighbourhood than C1. Furthermore, the confidence intervals for the covariance parameters reveal no evidence in favour of the homogeneous uniform model because the intervals for each customer type do not overlap.

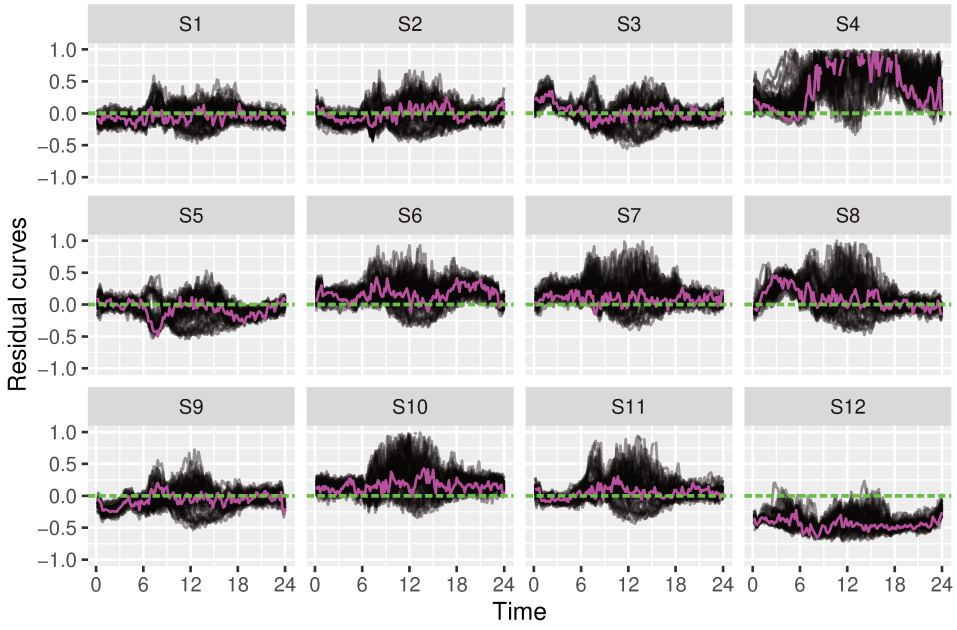
To evaluate whether the model is suitable for the available data, the fitted aggregated curve is plotted along with the observed data in Figure 23a. Apparently, the homogeneous model can capture the main features of the data, but fails to fit the aggregated load in some substations such as S4 and S12: S4 has overestimated fitted curves, whereas S12 has underestimated curves. This suggests that it might be interesting to add dummy variables indicating these two substations in the full model approach because there appears to be a vertical shift of the fitted curves. Other small discrepancies are visible in other substations, but in general they follow the main features of the observed data.

Figure 23b shows the relative residual curves defined in Section 2.5 for each substation with a reference line at zero and their median curve in green. Ideally, the median should almost coincide with the zero-reference line, but note that there are curves positioned above or below the zero reference. Specifically, substations S4 and S12 are clearly under- and overestimated respectively. Furthermore, the homogeneous dispersion hypothesis does not hold for these data because the dispersion of the residual curves varies

over time, which is another piece of evidence in favour of the complete aggregated data model with variance functionals to capture this feature.



(a) Estimated typical curves in kWh.



(b) Relative residual curves.

Figure 23 – Simple homogeneous aggregated data model fit: (a) estimated typical curves in kWh (in gray) and median curves (in magenta) and (b) Relative residual curves (in gray), median residual curve (in magenta) and zero reference line (in green) for each substation.

Important insights can be extracted from the homogeneous model fit before proceeding to the next level of the aggregated model, that is, the full model approach with additional covariates and a complete covariance structure. The bias in the fitted values

for substations S4 and S12 suggests that indicator variables specific to these substations could be used as explanatory variables in the full model. In addition, the air temperature information is incorporated as a functional covariate to build the typical surface and potentially reduce the residual curves dispersion in the work period between 9am and 5pm.

4.3 Full aggregated data model

The full aggregated data model enables a functional covariate to be incorporated to produce typical surface responses for each customer type, as well as explanatory variables to better explain the aggregated data variability. For the tensorial product expansion in Equation (2.5), $K = 24$ and $L = 6$ are used to estimate the typical surface. In addition, two explanatory variables are considered as indicators of substations S4 and S12, as mentioned in Section 4.2. The variance functionals used in the complete covariance structure are expanded as in Equation (2.12) with $K' = 6$.

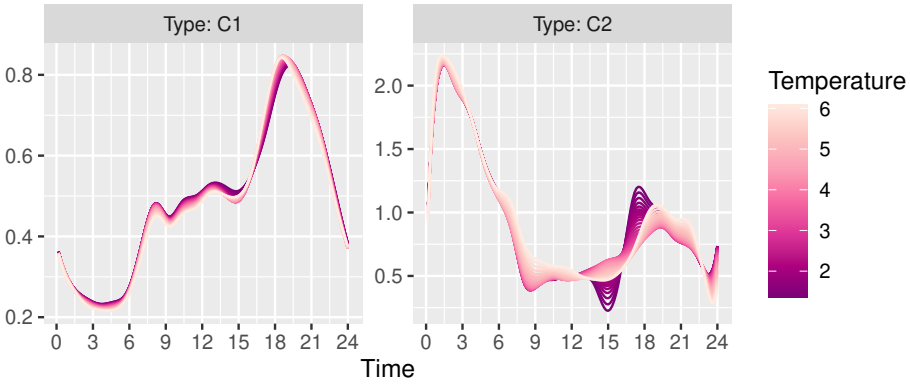
As mentioned in Section 4.1, temperature data were extracted for each primary every three hours, interpolated by cubic B-Splines and displayed in Figure 20b. Table 14 shows the summary statistics of the observed temperatures for each primary.The simulation studies in this project (Section 3.3) showed satisfactory estimated typical surfaces for temperature intervals frequently observed in the data, but higher dispersion in the estimate for rarely observed temperatures. In the case of real data, except for Trowbridge, temperature data are concentrated approximately between 1°C and 4 °C, and therefore the estimated typical surfaces may be well estimated within this range, but present some difficulties outside it.

4.3.1 Full model fit results

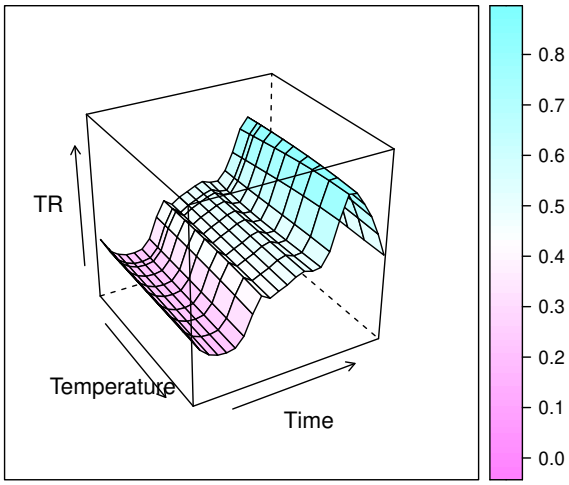
Figure 24a shows the estimated typical curves for certain fixed temperature values, and Figures 24b and 24c show the estimated typical surfaces for C1 and C2 customer types, respectively, for temperatures between 1.21°C and 5.89°C. The selected temperature

Table 14 – Summary statistics of air temperature in degrees Celsius over the 61 observed days in the dataset for each substation primary.

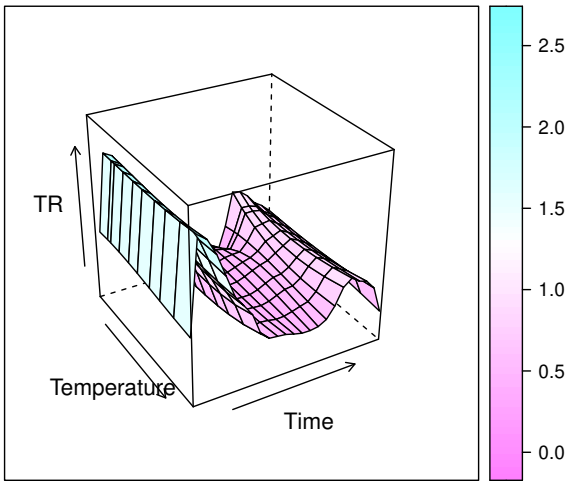
Primary	Minimum	1st Quartile	Median	3rd Quartile	Maximum
Cyncoed	-3.24	0.95	2.54	4.53	9.58
Llantarnam Primary	-4.22	0.89	3.30	7.90	12.06
Ringland Newport	-4.35	0.49	2.57	5.37	12.29
Trowbridge Primary	-5.41	-0.78	1.26	3.29	9.44
Usk	-6.27	1.00	5.74	8.07	12.22



(a) Estimated typical curves for temperatures from 1.21°C to 5.89°C.



(b) Estimated typical surface (TR) for C1 customers.



(c) Estimated typical surface (TR) for C2 customers.

Figure 24 – (a) Estimated typical curves in kWh for customers of type C1 and C2 coloured according to temperatures between 1.21°C and 5.89°, (b) estimated typical surface response in kWh for customers of type C1 between 1.21°C and 5.89°C and (c) estimated typical surface response in kWh for customers of type C2 between 1.21°C and 5.89°C. TR denotes the for typical response in kWh.

range contains 60% of the observed values in the dataset and hence that interval where the typical surfaces are well estimated, which avoids discrepant values that do not contribute to the analysis as shown in Section 3.3. On the time axis, the estimated typical surfaces have similar characteristics to the curves estimated by the simple aggregated model shown in Figure 22. On the temperature axis, unrestricted domestic C1 customers present robust behaviour for different temperatures, but C2 customers are subject to greater variation of energy consumption between 12 PM and 8 PM at different temperatures. In the latter case, extreme temperatures must be considered with caution because for values outside the selected range, the typical curves are unstable and may present negative or extremely high values.

The first two lines of Table 15 show the estimated effect values for the dummy explanatory variables corresponding to substations S4 and S12. Note that the estimated effect of substation S12 is a shift of 37.40, which is a considerable value because the aggregated observations in this location are mainly around 50 kWh and 120 kWh. Substation S4 results in an estimated effect of -8.20, but its 95% confidence interval contains zero, revealing that it may have no effect on the aggregated load data. The remaining lines of Table 15 present the estimated covariance decay parameters for C1 and C2 customers (ω_{C1} and ω_{C2}), which are much like to the ones obtained in Table 13. The correlation of neighbouring observations is stronger in type C2, with a decay parameter estimated at 0.61 versus 0.03 for type C1.

Figure 25 shows the estimated variance functionals for C1 and C2 customers along with their confidence bands built using their standard error as described in Section 2.5. The left panel reveals the higher values of dispersion at 9 AM, when people tend to leave their houses, and at 8 PM, the peak of the estimated typical curve. The lowest values are observed in early morning, a period with few or no activities in residences. On the right panel, note the peak around 10 AM and the lowest values around 10 PM and 3 AM. Interestingly, except for midnight, the lower tariff period has lower dispersion values.

Figure 26a shows the fit for the full aggregated data model along with the observed aggregated data. The over- and underestimation problem for substations S4 and

Table 15 – Estimated coefficients of explanatory variables and estimated covariance parameters followed by their 95% confidence intervals using the full aggregated data model.

Parameter	Value	95% Confidence Interval
S12	37.3968	(33.5605, 41.2331)
S4	-8.1977	(-19.9304, 3.5350)
ω_{C1}	0.0333	(0.0313, 0.0353)
ω_{C2}	0.6127	(0.5803, 0.6450)

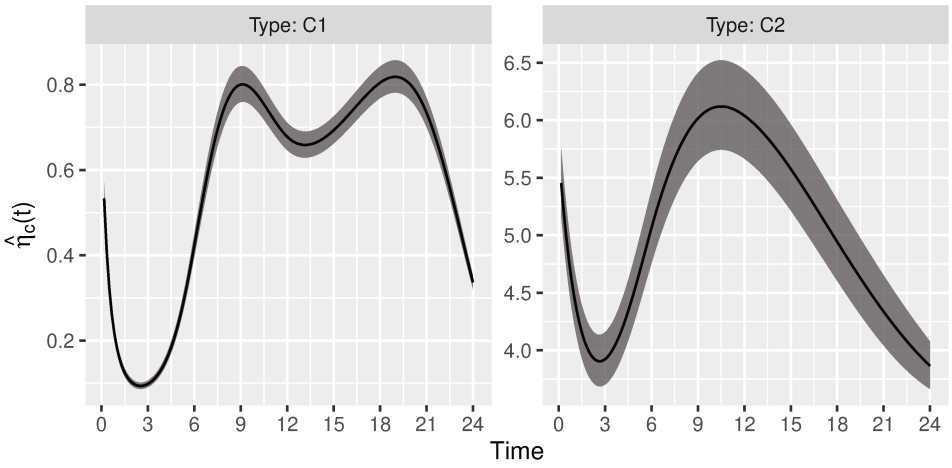


Figure 25 – Estimated variance functionals for C1 and C2 customers along with their confidence bands using the full aggregated data model.

S12 is solved by adding dummy variables, and including temperature captures a portion of data variability, but it is still difficult to explain the work period variability, which may be associated with other factors. Figure 26b presents the associated relative residual curves. Greater variability can be observed between 9 AM and 5 PM, but the residuals for S4 and S12 are closer to the zero-reference line than in Figure 23b.

4.3.2 Comparison with homogeneous aggregated data model

This section compares the proposed full aggregated data model with the homogeneous aggregated data model described in Section 4.2 to assess whether there are advantages to incorporating explanatory variables and temperature in terms of model fitting.

Table 16 shows the functional mean squared relative error (fMSRE) between fitted and observed values for each substation and fitted model. In this case, consider the total number of runs in Equation (2.58) as the number of observed days. Except for substations S8 and S9, the full aggregated data model has smaller fMSRE for most substations. The larger differences are observed in S4 and S12, both substations that were not well fitted by the homogeneous aggregated model, as shown in Figure 23a. Considering all substations, the average fMSRE of the full model is better in terms of fMSRE, with a value of 0.208, whereas the homogeneous model has an average fMSRE of 0.266.

Because the homogeneous aggregated data model is nested inside the full aggregated data model, the likelihood ratio test can be performed to verify whether the model fits are statistically different. Hence, by designating the full model as $\mathcal{M}_1$ and the

Table 16 – Functional mean squared relative errors of estimated versus predicted aggregated data for each substation and fitted model.

Substation	Model	fMSRE
S1	Homogeneous	0.1350
	Full	0.1295
S2	Homogeneous	0.1527
	Full	0.1518
S3	Homogeneous	0.1636
	Full	0.1629
S4	Homogeneous	0.8097
	Full	0.3811
S5	Homogeneous	0.1713
	Full	0.1690
S6	Homogeneous	0.2390
	Full	0.2241
S7	Homogeneous	0.2279
	Full	0.2221
S8	Homogeneous	0.2485
	Full	0.2387
S9	Homogeneous	0.1609
	Full	0.1680
S10	Homogeneous	0.2763
	Full	0.2699
S11	Homogeneous	0.1733
	Full	0.1693
S12	Homogeneous	0.4389
	Full	0.2154

simple model as $\mathcal{M}_2$, the test statistic L can be computed as

$$\begin{aligned} L &= -2\big(\ell(\mathcal{M}_1) - \ell(\mathcal{M}_2)\big) = -2\big(-344,770.3 - (-358,204.1)\big) \\ &= 26,867.67. \end{aligned} \tag{4.1}$$

Under the null hypothesis, the test statistic has a chi-square distribution with 254 degrees of freedom obtained from the difference of the number of model parameters. Hence, the difference between the models is statistically significant with p-value approximately zero.

Therefore, the full aggregated data model is a better fit, improving the explanation of the aggregated load variability by adding the temperature component and dummy variables. The assumption of a complete covariance structure could capture the variability over time by means of the estimated variance functionals. Although the estimated surfaces

might be used with caution in temperature ranges with few observations, in general they are useful to assess electrical energy consumption under different weather conditions. In addition, many other functional variables can be included in the model either as a higher-dimensional surface of additive linear or non-linear terms or as other explanatory variables, scalar or functional, to explain the aggregated data variability.

4.4 Clustering analysis

The clustering aggregated data model groups substations with similar typical curves and covariance structure for domestic customers of type unrestricted and "Economy 7". The model assumes that the aggregated observed data are a mixture of B aggregated models with distinct mean curves, with B as the total number of clusters. This section describes the fitting of a mixture of aggregated models with homogeneous covariance structure considering two and three clusters to obtain the best substation clustering that explains observed aggregated data variability using the Bayesian Information Criterion. There will be no explanatory variables or temperature components in this model.

4.4.1 Two clusters

This first approach may be useful as a tool to explore data with faster computational performance because the number of covariance parameters to be estimated by numerical optimization is $B \times C \times P$; where B is the number of clusters, C the number of customer types and P the number of parameters relative to the covariance structure. In this case with two customer types, two clusters, and a homogeneous covariance structure, eight covariance parameters must be estimated. The estimated typical curves are estimated by least squares.

Figure 27 shows the estimated typical curves for customers of type C1 and C2 in Clusters 1 and 2. Type C1 curves share characteristics in both clusters like the increasing load around 9am, the plateau in the middle of the day, and the highest consumption at 8pm. Customers of type C2 have peaks at 2am in both clusters, but with different magnitudes, with Cluster 1 being the smaller one. The clustering aggregated model reveals new features for C2 customers, such as the different 8 PM peak, that could not be identified with the aggregated data models in Section 4.3.

Table 17 shows the estimated probability $\hat{p}_{jb}$ of substation j being allocated to cluster b. Substations S5 and S12 are grouped in Cluster 2, and Cluster 1 gathers the remaining substations into a large cluster with 10 elements. Interestingly, S12 is the substation located far to the north, as shown in Figure 21, and one of the few substations that does not show an extreme dominance of C1 customer types; on the other hand, substation S5 has one of the markets dominated by C1 customers.

Figure 28a shows the fitted values plotted along with observed aggregated load data. Note that the model can explain most of the aggregated data variability. In contrast to the full homogeneous aggregated data model, the impact of the clustering approach is visible on substation S12. The fact that the model enables this substation to have an estimated typical curve different than most of the others shows that this clustering approach is sufficient to explain the electrical load variability without a dummy explanatory variable to shift estimated fitted values. In substation S5, this impact is not evident because C2 has only 3.11% of market share. Furthermore, the relative residual curves in Figure 28b shows that the clustering aggregated model has median residual curve oscillating around the zero reference line and indicating that it is well adjusted to the observed data. However, the difference in variability over time in the residual curves suggests that the complete covariance structure with variance functionals might be more suitable.

The estimated covariance parameters for both clusters are displayed in Table 18. When compared to the homogeneous model of Section 4.2, the estimated parameters of Cluster 1 are closer to those presented in Table 13. Still in Cluster 1, the results present a large estimated dispersion parameter for C2 customer types, possibly related to the small number of customers in the market and the difficulty of representing the variability of the period between 9 AM and 5 PM by a single typical curve.

In summary, the clustering aggregated data model with two clusters provides satisfactory fitted curves (see Figure 28a) and typical curves that captures different characteristics for each cluster, especially customers of type C2. Even with no explanatory variables or additional temperature component it was possible to explain most of the variability of the load profiles.

4.4.2 Three clusters

The next step was to consider three clusters to fit the clustering aggregated data model assuming homogeneous covariance structure. Figure 29 shows the estimated typical curves for C1 and C2 for the three clusters. The unrestricted customers C1 have

Table 17 – Estimated probability $\hat{p}_{jb}$ of substation j belonging to cluster b, under the two cluster fit.

	Trowbridge			Cyncoed		Ringland		Llantarnam				Usk
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
$\hat{p}_{j1}$	1	1	1	1	0	1	1	1	1	1	1	0
$\hat{p}_{j2}$	0	0	0	0	1	0	0	0	0	0	0	1

Table 18 – Estimated dispersion (σ_{cb}) and decay (ω_{cb}) parameters for customer type c in cluster b of the clustering aggregated model considering two clusters.

Parameter	Value	95% Confidence Interval
σ_{11}	0.7016	(0.6856, 0.7175)
σ_{21}	4.3629	(4.2214, 4.5045)
ω_{11}	0.0491	(0.0466, 0.0515)
ω_{21}	1.0033	(0.9406, 1.0661)
σ_{12}	1.5410	(1.4834, 1.5985)
σ_{22}	1.5375	(1.461, 1.6139)
ω_{12}	0.1588	(0.146, 0.1716)
ω_{22}	0.0277	(0.0244, 0.0310)

once more similar curves in all clusters, with small observable differences during the work period between 9 AM and 5 PM and at the 8 PM peak at night. In contrast, the "Economy 7" customers C2 have distinct estimated load profiles among clusters. The estimated typical curve of Cluster 2 has unmistakably the lowest energy consumption and its only peak in the early morning, whereas Clusters 1 and 3 share some characteristics like the double peak right after midnight and at 8 PM, but minor differences in the morning and during the work period.

The estimated probability of the cluster assignment is shown in Table 19, where each substation is allocated with high probability to its cluster. Again, S5 and S12 form one cluster whereas the large cluster of Section 4.4.1 is divided into a major cluster composed by substations from Llantarnam primary and two from Trowbridge and another cluster with Ringland substations plus S1 and S4. The clustering results show that the big cluster in Section 4.4.1 with 10 of 12 substations was divided into two clusters with estimated typical curves that share some characteristics, but representing different morning and work period behaviors as seen in Figure 29.

The fitted curves over the observed aggregated data are displayed in Figure 30a. There are no apparent differences in fitted values compared to the two-cluster approach in

Table 19 – Estimated probability $\hat{p}_{jb}$ of substation j belonging to cluster b, under the three cluster fit.

	Trowbridge			Cyncoed		Ringland		Llantarnam				Usk
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
$\hat{p}_{j1}$	0	0	0	0	1	0	0	0	0	0	0	1
$\hat{p}_{j2}$	1	0	0	1	0	1	1	0	0	0	0	0
$\hat{p}_{j3}$	0	1	1	0	0	0	0	1	1	1	1	0

Figure 28a. Recall that substation markets are mostly dominated by C1 customers, the ones with similar estimated typical curves in all clusters, with substations with more C2 customers like S4 and S12 remaining in the same cluster. Hence, the impact of different C2 typical curves for Llantarnam, for example, might not be evident in the estimated aggregated load. Therefore, the residual curves in Figure 30b yield the same characteristics as the two clusters residual plot in Figure 28b.he plot suggests a good model fit represented by the median residual curves around the zero-reference line in most substations, except for a slight overestimation in substation S6. Comparisons between the two- and three-cluster models will be detailed in Section 4.4.4.

Table 20 displays the estimated covariance parameters for each combination of cluster and customer type. Cluster 1 probably has a homogeneous dispersion because the two values are close, and their 95% confidence intervals overlap. However, there is high uncertainty in the decay parameters, especially for C2, where its confidence interval is large enough to contain zero, although we know this is not possible due to the parameter positive restriction. Cluster 2, the one with the lowest estimated C2 typical curves, has distinct dispersion and decay parameters for both customers and narrow confidence intervals. Lastly, Cluster 3 presents the largest distinction between dispersion parameters, which is visible in the confidence bands of the estimated typical curve of type C2 in Figure 29.

In summary, the clustering aggregated data model with three clusters divided the large cluster in the two-cluster approach into two groups represented mostly by their primaries. The estimated typical curves for customers of type C1 still show similarities between clusters, but now enable the estimated typical curves of C2 customers to accommodate three different electrical energy consumption profiles.

Table 20 – Estimated dispersion (σ_{cb}) and decay (ω_{cb}) parameters for customer type c in cluster b of the aggregated three-cluster model.

Parameter	Value	95% Confidence Interval
σ_{11}	1.5367	(1.4798, 1.5936)
σ_{21}	1.5269	(1.495, 1.5588)
ω_{11}	0.1584	(0.091, 0.2258)
ω_{21}	0.0272	(-0.1216, 0.1761)
σ_{12}	1.0743	(1.0732, 1.0754)
σ_{22}	1.2794	(1.2762, 1.2825)
ω_{12}	0.1197	(0.1085, 0.1308)
ω_{22}	0.0905	(0.0159, 0.1650)
σ_{13}	0.4278	(0.4151, 0.4405)
σ_{23}	5.1783	(5.1706, 5.1859)
ω_{13}	0.0202	(0.0096, 0.0307)
ω_{23}	0.3743	(0.3487, 0.3998)

4.4.3 More than three clusters

Models with four or more clusters do not meet the condition of identifiability to obtain typical curves and covariance parameter estimates because there are only 12 substations.

4.4.4 Model comparison

The two- and three-cluster aggregated data models resulted in good model fits according to the fitted values (Figures 28a and 30a) and the residual curves (Figures 28b and 30b). The difference between them can be observed in the large cluster with 10 substations for the two-cluster model, which is split into two clusters in the three-cluster model (Tables 17 and 19).To decide which model is best suited to the observed aggregated data, the model comparison tools described in Section 2.5 were used.

Table 21 shows the functional mean squared relative error (fMSRE) of the fitted and observed data under the two- and three-cluster models at each substation. In a comparison of substation fMSREs, S4 and S10 are highlighted because they have the largest differences between models. Observing both the fitted over observed values and the relative residual curves in Figures 28a and 28b under the two-cluster model, it is clear how far their median curves are from the zero-reference line. On the other hand, observing the same substations in Figures 30a and 30b under the three-cluster model, it is apparent that their medians are closer to the zero line. In other words, the three-cluster model improves the model fit for these substations and consequently reduces their fMSRE. The other substations have minor differences between models in terms of fMSRE.

Another complementary tool is to compare the two models by their approximated BIC values. These values are given by:

- Two clusters BIC $= 707,308.3$ and
- Three clusters BIC $= 704,571.4$.

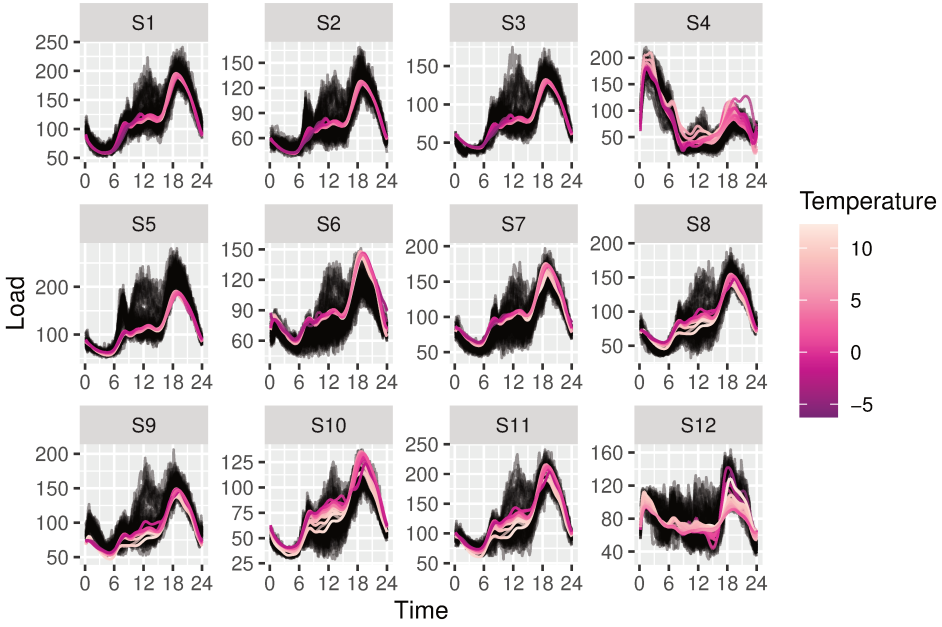
Because the selection is favourable to models with the smallest BIC values, it again favours the aggregated three-cluster data model, although its BIC value is only 0.3% smaller than the BIC for the two-cluster model. Remember that the BIC value approach has worked well as a model selection criterion in simulated mixture Gaussian processes (SHI; WANG, 2008; SHI; CHOI, 2011).

Therefore, the aggregated three-cluster model performed better in terms of fMSRE and BIC. Moreover, substations are grouped in a meaningful way, related to their primaries and avoiding the large cluster in the two-cluster approach. Hence, the three-

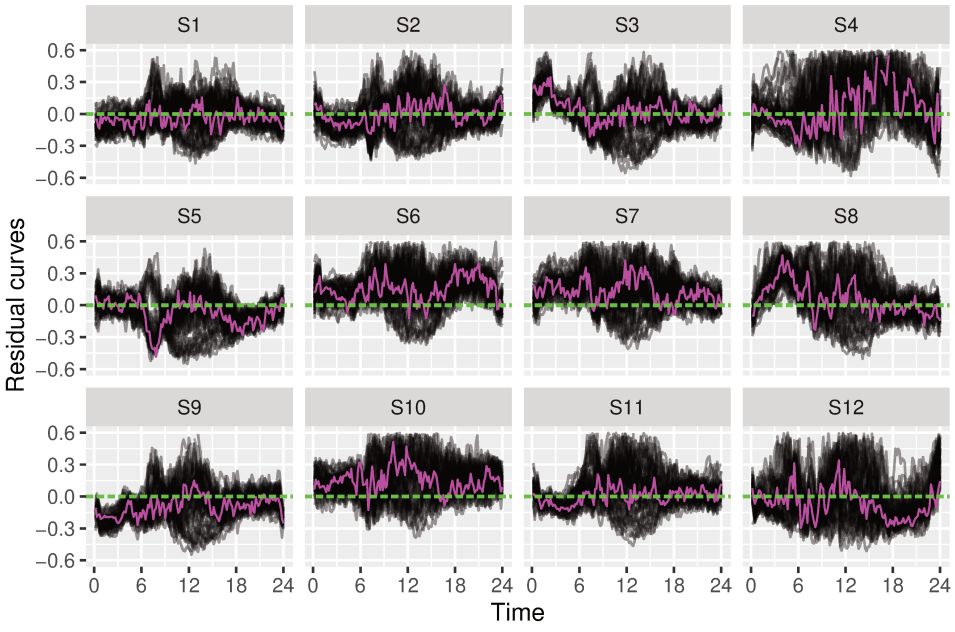
Table 21 – Functional mean squared relative error of fitted and observed data under the two- and three-cluster models at each substation.

Substation	Clusters	fMSRE
S1	2	0.1330
	3	0.1403
S2	2	0.1552
	3	0.1569
S3	2	0.1676
	3	0.1663
S4	2	0.3009
	3	0.2354
S5	2	0.1603
	3	0.1603
S6	2	0.1818
	3	0.1678
S7	2	0.1971
	3	0.1851
S8	2	0.2284
	3	0.2636
S9	2	0.1750
	3	0.1584
S10	2	0.2689
	3	0.1739
S11	2	0.1590
	3	0.1633
S12	2	0.1977
	3	0.1977

cluster model seems to be a reasonable choice to group the electrical energy substations in the UK electrical load data.



(a) Observed aggregated load data over estimated aggregated curves.



(b) Relative residuals curves.

Figure 26 – Full aggregated data model typical curves results: (a) Observed aggregated load data (in tones of magenta) in kWh over estimated aggregated curves (in gray) and (b) relative error curves (in gray), median residual curves (in green) and zero reference line (in magenta) for the 12 substations using the full aggregated data model.

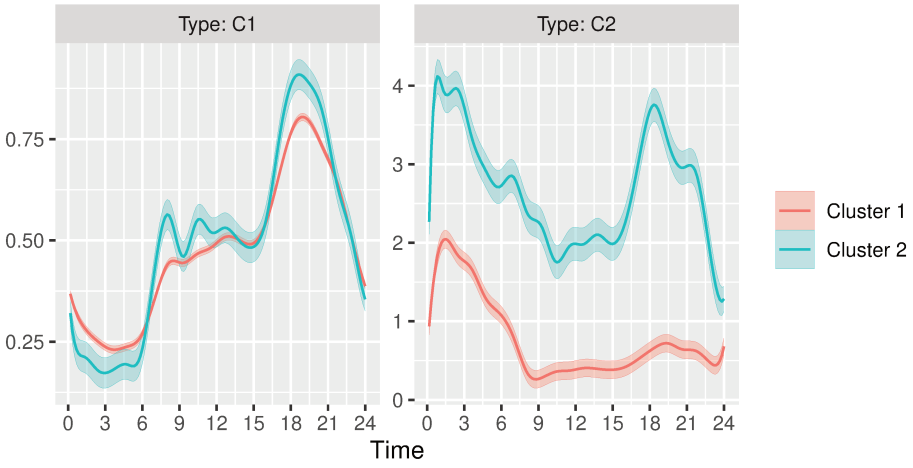
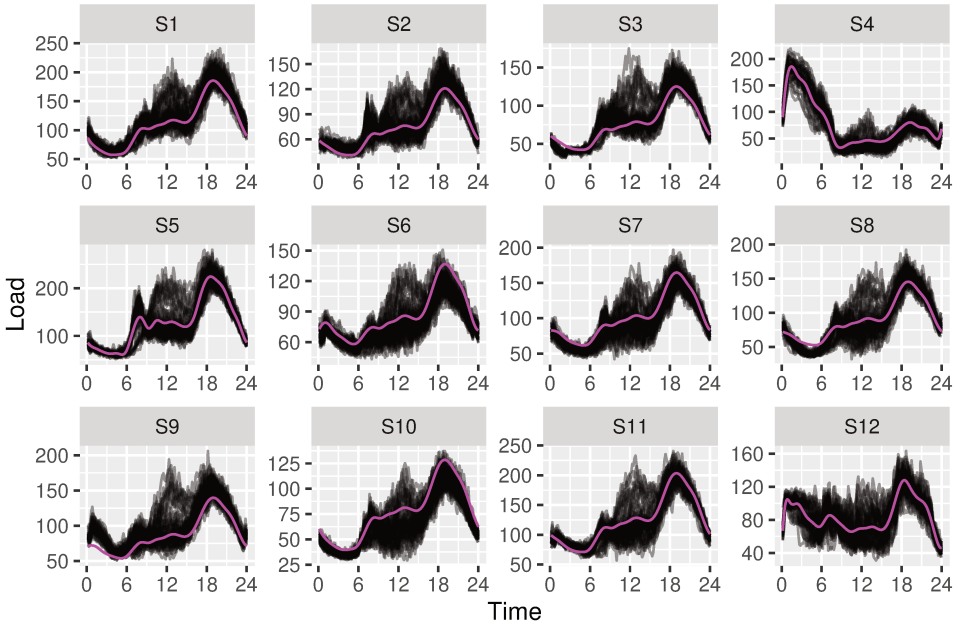
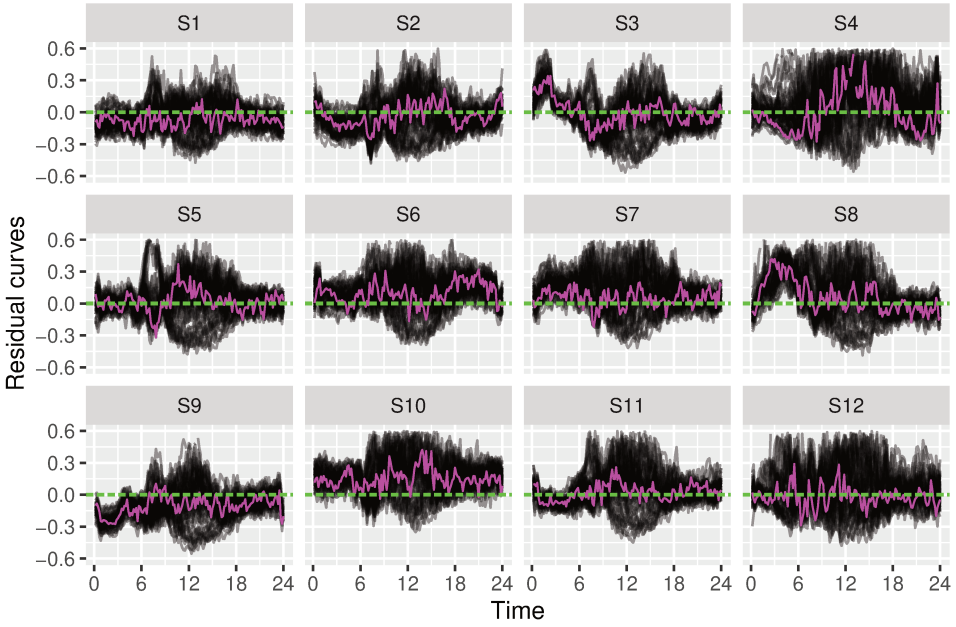


Figure 27 – Estimated typical curves in kWh and their confidence band for unrestricted (C1) and "Economy 7" (C2) domestic customers under two clusters aggregated data model fit.



(a) Observed aggregated load data over estimated aggregated curves.



(b) Relative residual curves.

Figure 28 – Results of the clustering aggregated data model with two clusters: (a) Observed aggregated load data (in tones of magenta) in kWh over estimated aggregated curves (in gray) and (b) relative error curves (in gray), median residual curves (in green) and zero reference line (in magenta) for the 12 substations.

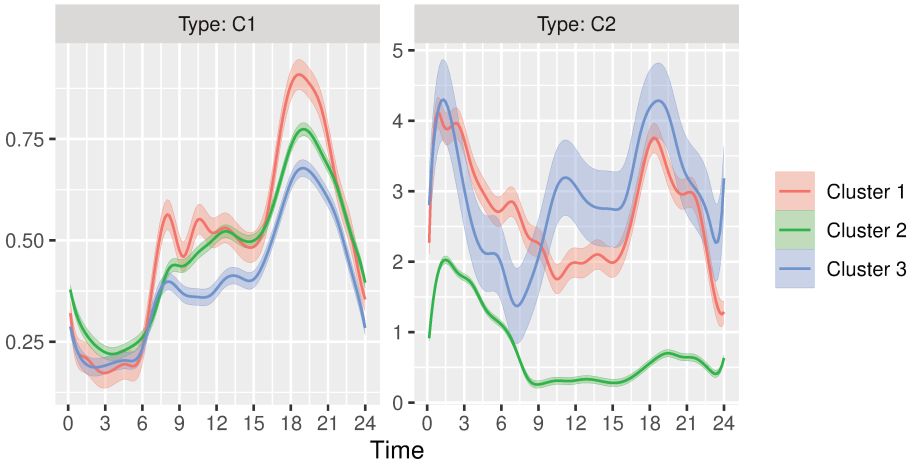
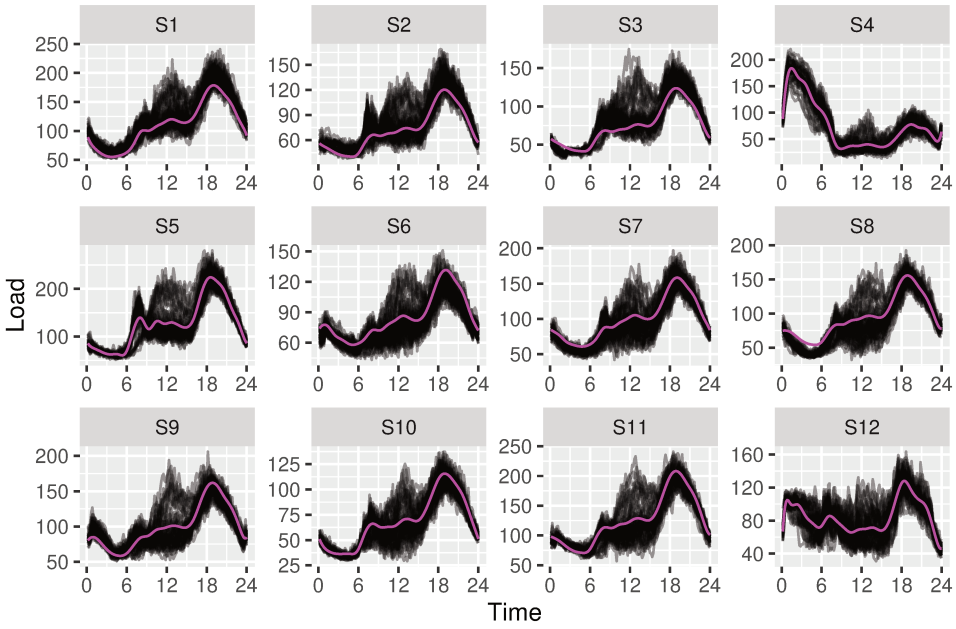
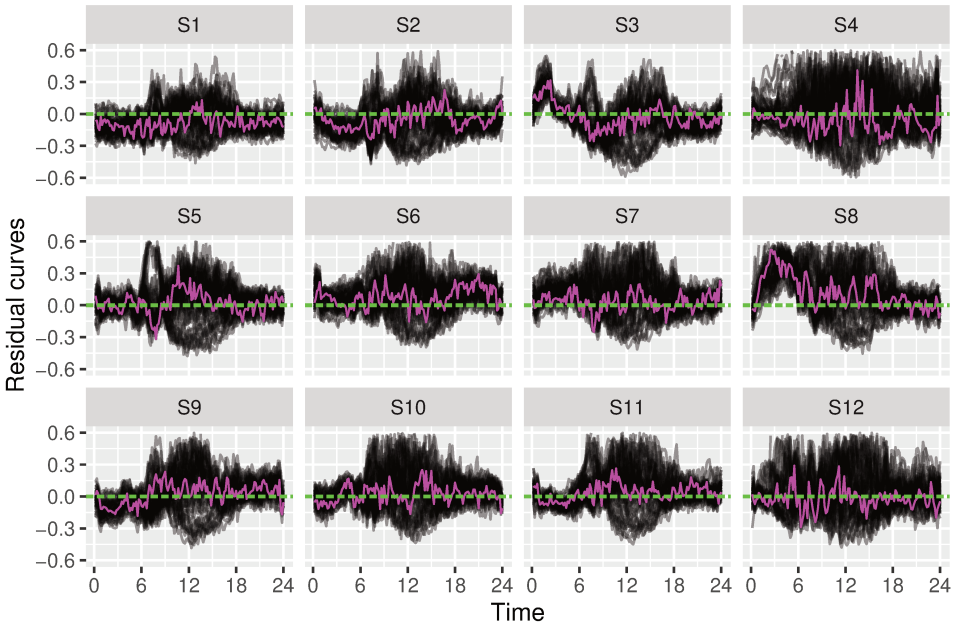


Figure 29 – Estimated typical curves in kWh and their confidence band for unrestricted (C1) and "Economy 7" (C2) domestic customers under three clusters aggregated data model fit



(a) Observed aggregated load data over estimated aggregated curves.



(b) Relative residual curves.

Figure 30 – Results of the clustering aggregated data model with three clusters: (a) Observed aggregated load data (in tones of magenta) in kWh over estimated aggregated curves (in gray) and (b) relative error curves (in gray), median residual curves (in green) and zero reference line (in magenta) for the 12 substations.

5 Final considerations

The proposed aggregated data model has proved to be a useful tool to separate substation aggregated electrical load data into typical curves for each type of supplied customer and to comprehend their covariance structure. The model also proposes novel approaches such as typical surface estimation as a function of time and temperature and explanatory variables and substation clustering based on the similarity of their estimated typical curves. The methodology based on basis function expansion and a Gaussian process assumption brings the model into a family of functional models with favourable mathematical properties and well-established inference methodology.

With both simulated and real data, the estimated typical curves demonstrated robustness to wrong covariance structure assumptions. Under the assumption of an incorrect number of clusters, the model groups substations with greater similarity. Furthermore, this thesis has discussed the results of mis-specified scenarios and how they relate to the true parameters, for example, when scalar dispersion parameters are assumed instead of functional variances.

Scenarios based on more observed days tend to have better estimates, especially if markets are balanced. In the clustering approach, clusters containing more substations have improved precision in parameter estimation compared to clusters with the minimum number of elements. Scenarios with few observed days still deliver good estimated typical curves on average, but covariance parameter estimation may be a challenge. Nonetheless, the proposed model provides standard errors to assess the variability of its estimated parameters.

Some estimation methods were crucial to the success of the proposed model, such as the least-squares estimator for the typical curves and the proposed initial value evaluation in the clustering approach. This latter method drastically reduced computing time and improved estimation performance by providing clustering setups close to reality (see Section 2.4.2.4). With full control of model setup and robust methodology, users can fit the best aggregated data model to their reseach.

The full aggregated model with explanatory variables and additional functional component demonstrated sophistication and flexibility with both real and simulated data. Suggestions on how to use the additional component properly were provided to avoid poor decisions in temperature ranges with little information. In any case, when working with real data, the confidence intervals of the estimated typical curves and surfaces will indicate ranges of uncertainty.

5.1 Future work

Two papers will result from this thesis and will be submitted to top-ranking statistical journals. One paper will present the methodology of the proposed aggregated model, and the other will present the R package aggrmodel to the scientific community.

The model offers several topics for future work. One would be to propose a latent variable Z_{jc} indexed both by group and by subject type. This approach assigns to subject c in station j a probability of belonging to a cluster. With a well-written likelihood function, the model might keep its base estimation routine and gain even more flexibility to perform further analysis of aggregated data models.

Load monitoring of electrical appliances has attracted growing interest in academic journals. The proposed aggregated data model could be used in referential data sets such as ECO Data used in recent studies to evaluate energy disaggregation models (BECKEL et al., 2014). Some adjustments might be necessary to fit the proposed model, but this is an opportunity to compare it with several well-established statistical approaches (SCHIRMER; MPORAS; PARASKEVAS, 2019; SCHIRMER; MPORAS, 2019; SCHIRMER; MPORAS; SHEIKH-AKBARI, 2020).

The relationship between the aggregated data model and blind source separation can be explored as well. Both have the same goal of separating an aggregated signal and could have an interesting mathematical relationship to be investigated. Inserting the aggregated model into an expanding area such as blind source separation may place the method in a visible research area and attract researchers worldwide.

The opportunities mentioned above are perfectly suited for multiple levels of research. Direct application in ECO Data might be an interesting project for an undergraduate research assistant in partnership with a Master's student to adapt the model to appliance data. The new latent variable Z_{jc} and the formal relationship with blind source separation might also become a new PhD project.

Bibliography

ALLWOOD, J. M.; CULLEN, J. M.; MILFORD, R. L. *Options for achieving a 50% cut in industrial carbon emissions by 2050*. [S.l.]: ACS Publications, 2010. Citado na página 13.

ARGHIRA, N.; HAWARAH, L.; PLOIX, S.; JACOMINO, M. Prediction of appliances energy use in smart homes. *Energy*, Elsevier, v. 48, n. 1, p. 128–134, 2012. Citado na página 14.

BASU, K.; DEBUSSCHERE, V.; BACHA, S.; MAULIK, U.; BONDYOPADHYAY, S. Nonintrusive load monitoring: A temporal multilabel classification approach. *IEEE Transactions on industrial informatics*, IEEE, v. 11, n. 1, p. 262–270, 2014. Citado 2 vezes nas páginas 15 and 16.

BECKEL, C.; KLEIMINGER, W.; CICCHETTI, R.; STAAKE, T.; SANTINI, S. The eco data set and the performance of non-intrusive load monitoring algorithms. In: *Proceedings of the 1st ACM conference on embedded systems for energy-efficient buildings*. [S.l.: s.n.], 2014. p. 80–89. Citado na página 88.

BILSKI, P.; WINIECKI, W. Generalized algorithm for the non-intrusive identification of electrical appliances in the household. In: IEEE. *2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*. [S.l.], 2017. v. 2, p. 730–735. Citado na página 15.

BOOR, C. D.; BOOR, C. D.; MATHÉMATICIEN, E.-U.; BOOR, C. D.; BOOR, C. D. *A practical guide to splines*. [S.l.]: springer-verlag New York, 1978. v. 27. Citado na página 18.

BRISTOW, A. L.; TIGHT, M.; PRIDMORE, A.; MAY, A. D. Developing pathways to low carbon land-based passenger transport in great britain by 2050. *Energy Policy*, Elsevier, v. 36, n. 9, p. 3427–3435, 2008. Citado na página 13.

CARDOSO, J.-F. Blind signal separation: statistical principles. *Proceedings of the IEEE*, IEEE, v. 86, n. 10, p. 2009–2025, 1998. Citado na página 16.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 39, n. 1, p. 1–22, 1977. Citado 2 vezes nas páginas 16 and 24.

DEVORE, R.; PETROVA, G.; TEMLYAKOV, V. Best basis selection for approximation in l p. *Foundations of Computational Mathematics*, Springer, v. 3, n. 2, p. 161–185, 2003. Citado na página 19.

DIAS, R. Density estimation via hybrid splines. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 60, n. 4, p. 277–293, 1998. Citado na página 19.

DIAS, R.; GARCIA, N. L. Consistent estimator for basis selection based on a proxy of the kullback–leibler distance. *Journal of econometrics*, Elsevier, v. 141, n. 1, p. 167–178, 2007. Citado na página 19.

DIAS, R.; GARCIA, N. L.; LUDWIG, G.; SARAIVA, M. A. Aggregated functional data model for near-infrared spectroscopy calibration and prediction. *Journal of Applied Statistics*, Taylor & Francis, v. 42, n. 1, p. 127–143, 2015. Citado 2 vezes nas páginas 16 and 19.

DIAS, R.; GARCIA, N. L.; MARTARELLI, A. Non-parametric estimation for aggregated functional data for electric load monitoring. *Environmetrics: The official journal of the International Environmetrics Society*, Wiley Online Library, v. 20, n. 2, p. 111–130, 2009. Citado 3 vezes nas páginas 16, 18, and 19.

DIAS, R.; GARCIA, N. L.; SCHMIDT, A. M. A hierarchical model for aggregated functional data. *Technometrics*, Taylor & Francis Group, v. 55, n. 3, p. 321–334, 2013. Citado 4 vezes nas páginas 16, 19, 20, and 21.

DONOHO, D. L. Unconditional bases are optimal bases for data compression and for statistical estimation. *Applied and computational harmonic analysis*, Elsevier, v. 1, n. 1, p. 100–115, 1993. Citado na página 19.

D'OCA, S.; CORGNATI, S. P.; BUSO, T. Smart meters and energy savings in italy: Determining the effectiveness of persuasive communication in dwellings. *Energy Research & Social Science*, Elsevier, v. 3, p. 131–142, 2014. Citado na página 14.

EMPRESA DE PESQUISA ENERGÉTICA – EPE. *Balanço Energético Nacional – Relatório Síntese*. Rio de Janeiro, 2018. Disponível em: <"http://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-303/topico-397/Relat%C3%B3rio%20S%C3%ADntese%202018-ab%202017vff.pdf">. Citado na página 13.

FLETCHER, R. *Practical methods of optimization*. [S.l.]: John Wiley & Sons, 2013. Citado 2 vezes nas páginas 26 and 30.

GOUVEIA, J. P.; SEIXAS, J. Unraveling electricity consumption profiles in households through clusters: Combining smart meters and door-to-door surveys. *Energy and Buildings*, Elsevier, v. 116, p. 666–676, 2016. Citado na página 14.

GREEN, P. Reversible jump mcmc computation and bayesian model determination. biomeirika, 82: 711-732 hastings, wk 1970. monte carlo sampling methods using markov chains and their applications. *Biometrika*, v. 57, p. 97–109, 1995. Citado na página 30.

HART, G. W. Nonintrusive appliance load monitoring. *Proceedings of the IEEE*, IEEE, v. 80, n. 12, p. 1870–1891, 1992. Citado na página 14.

HOSSEINI, S. S.; AGBOSSOU, K.; KELOUWANI, S.; CARDENAS, A. Non-intrusive load monitoring through home energy management systems: A comprehensive review. *Renewable and Sustainable Energy Reviews*, Elsevier, v. 79, p. 1266–1274, 2017. Citado na página 15.

HULLAIT, H.; LESLIE, D. S.; PAVLIDIS, N. G.; KING, S. Robust function-on-function regression. *Technometrics*, Taylor & Francis, p. 1–14, 2020. Citado na página 31.

KIM, Y.; KONG, S.; KO, R.; JOO, S.-K. Electrical event identification technique for monitoring home appliance load using load signatures. In: IEEE. *2014 IEEE International Conference on Consumer Electronics (ICCE)*. [S.l.], 2014. p. 296–297. Citado na página 16.

KIRKPATRICK, S.; GELATT, C. D.; VECCHI, M. P. Optimization by simulated annealing. *science*, American association for the advancement of science, v. 220, n. 4598, p. 671–680, 1983. Citado na página 26.

KOHN, R.; MARRON, J. S.; YAU, P. Wavelet estimation using bayesian basis selection and basis averaging. *Statistica Sinica*, JSTOR, p. 109–128, 2000. Citado na página 19.

LENZI, A.; SOUZA, C. P. de; DIAS, R.; GARCIA, N. L.; HECKMAN, N. E. Analysis of aggregated functional data from mixed populations with application to energy consumption. *Environmetrics*, Wiley Online Library, v. 28, n. 2, p. e2414, 2017. Citado 3 vezes nas páginas 16, 19, and 55.

LI, R.; GU, C.; LI, F.; SHADDICK, G.; DALE, M. Development of low voltage network templates—part i: Substation clustering and classification. *IEEE Transactions on Power Systems*, IEEE, v. 30, n. 6, p. 3036–3044, 2015. Citado 2 vezes nas páginas 13 and 63.

_____. Development of low voltage network templates—part ii: Peak load estimation by clusterwise regression. *IEEE Transactions on Power Systems*, IEEE, v. 30, n. 6, p. 3045–3052, 2015. Citado na página 63.

LI, Y.; WANG, N.; CARROLL, R. J. Selecting the number of principal components in functional data. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 108, n. 504, p. 1284–1294, 2013. Citado na página 31.

LIN, Y.-H.; TSAI, M.-S. An advanced home energy management system facilitated by nonintrusive load monitoring with automated multiobjective power scheduling. *IEEE Transactions on Smart Grid*, IEEE, v. 6, n. 4, p. 1839–1851, 2015. Citado na página 15.

LIPP, J. Lessons for effective renewable electricity policy from denmark, germany and the united kingdom. *Energy policy*, Elsevier, v. 35, n. 11, p. 5481–5495, 2007. Citado na página 13.

LUO, Z.; WAHBA, G. Hybrid adaptive splines. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 92, n. 437, p. 107–116, 1997. Citado na página 19.

MCLACHLAN, G.; KRISHNAN, T. *The EM algorithm and extensions*. [S.l.]: John Wiley & Sons, 2007. v. 382. Citado 3 vezes nas páginas 16, 24, and 28.

MENG, X.-L.; RUBIN, D. B. Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, Oxford University Press, v. 80, n. 2, p. 267–278, 1993. Citado na página 28.

MINISTÉRIO DE MINAS E ENERGIA. *Plano Nacional de Energia 2030*. Brasília, 2007. Disponível em: <"http://www.mme.gov.br/documents/10584/1139260/Plano+Nacional+de+Energia+2030+%28PDF%29/ba957ba9-2439-4b28-ade5-60cf94612092?version=1.2">. Citado na página 13.

NOBRE, P.; PEREIRA, E. B.; LACERDA, F. F.; BURSZTYN, M.; HADDAD, E. A.; LEY, D. Solar smart grid as a path to economic inclusion and adaptation to climate change in the brazilian semiarid northeast. *International Journal of Climate Change Strategies and Management*, Emerald Publishing Limited, 2019. Citado na página 13.

PRAHASTONO, I.; KING, D.; OZVEREN, C. S. A review of electricity load profile classification methods. In: IEEE. *2007 42nd international universities power engineering conference*. [S.l.], 2007. p. 1187–1191. Citado na página 13.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2019. Disponível em: <https://www.R-project.org/>. Citado na página 30.

RAMSAY, J.; SILVERMAN, B. *Functional Data Analysis*. [S.l.]: Springer Science & Business Media, 2005. Citado 3 vezes nas páginas 18, 21, and 26.

REIF, U. Orthogonality of cardinal b-splines in weighted sobolev spaces. *SIAM Journal on Mathematical Analysis*, SIAM, v. 28, n. 5, p. 1258–1263, 1997. Citado na página 19.

RUGGIERO, M. A. G.; LOPES, V. L. d. R. *Cálculo numérico: aspectos teóricos e computacionais*. [S.l.]: Makron Books do Brasil, 1997. Citado na página 26.

SCHIRMER, P. A.; MPORAS, I. Integration of temporal contextual information for robust energy disaggregation. In: IEEE. *2019 IEEE 38th International Performance Computing and Communications Conference (IPCCC)*. [S.l.], 2019. p. 1–6. Citado 2 vezes nas páginas 15 and 88.

SCHIRMER, P. A.; MPORAS, I.; PARASKEVAS, M. Evaluation of regression algorithms and features on the energy disaggregation task. In: IEEE. *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*. [S.l.], 2019. p. 1–4. Citado 3 vezes nas páginas 14, 16, and 88.

SCHIRMER, P. A.; MPORAS, I.; SHEIKH-AKBARI, A. Energy disaggregation using two-stage fusion of binary device detectors. *Energies*, Multidisciplinary Digital Publishing Institute, v. 13, n. 9, p. 2148, 2020. Citado 3 vezes nas páginas 15, 16, and 88.

SCHWARZ, G. et al. Estimating the dimension of a model. *The annals of statistics*, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978. Citado na página 30.

SHI, H.; XU, M.; LI, R. Deep learning for household load forecasting—a novel pooling deep rnn. *IEEE Transactions on Smart Grid*, IEEE, v. 9, n. 5, p. 5271–5280, 2017. Citado na página 13.

SHI, J.; WANG, B. Curve prediction and clustering with mixtures of gaussian process functional regression models. *Statistics and Computing*, Springer, v. 18, n. 3, p. 267–283, 2008. Citado 2 vezes nas páginas 31 and 80.

SHI, J. Q.; CHOI, T. *Gaussian process regression analysis for functional data*. [S.l.]: Chapman and Hall/CRC, 2011. Citado 6 vezes nas páginas 16, 18, 24, 26, 30, and 80.

SHI, J. Q.; MURRAY-SMITH, R.; TITTERINGTON, D. Hierarchical gaussian process mixtures for regression. *Statistics and computing*, Springer, v. 15, n. 1, p. 31–41, 2005. Citado 2 vezes nas páginas 16 and 23.

SOUSA, J. C.; JORGE, H. M.; NEVES, L. P. Short-term load forecasting based on support vector regression and load profiling. *International journal of energy research*, Wiley Online Library, v. 38, n. 3, p. 350–362, 2014. Citado na página 13.

SOUZA, C. P. D.; HECKMAN, N. E.; XU, F. Switching nonparametric regression models for multi-curve data. *Canadian Journal of Statistics*, Wiley Online Library, v. 45, n. 4, p. 442–460, 2017. Citado na página 13.

STATISTICS, O. for N. *2011 Census aggregate data*. DOI: http://dx.doi.org/10.5257/census/aggregate-2011-1, 2016. Citado na página 67.

TRESP, V. Mixtures of gaussian processes. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2001. p. 654–660. Citado 2 vezes nas páginas 16 and 30.

WANG, Y.; CHEN, Q.; KANG, C.; ZHANG, M.; WANG, K.; ZHAO, Y. Load profiling and its application to demand response: A review. *Tsinghua Science and Technology*, TUP, v. 20, n. 2, p. 117–129, 2015. Citado na página 13.

WEAVER, A. J.; ZICKFELD, K.; MONTENEGRO, A.; EBY, M. Long term climate implications of 2050 emission reduction targets. *Geophysical Research Letters*, Wiley Online Library, v. 34, n. 19, 2007. Citado na página 13.

WEI, J.; ZHOU, L. Model selection using modified aic and bic in joint modeling of paired functional data. *Statistics & probability letters*, Elsevier, v. 80, n. 23-24, p. 1918–1924, 2010. Citado na página 31.

WILKS, M. Demand side response: Conflict between supply and network driven optimisation. *A Report to DECC Nov*, 2010. Citado na página 63.

WILLIAMS, J. H.; DEBENEDICTIS, A.; GHANADAN, R.; MAHONE, A.; MOORE, J.; MORROW, W. R.; PRICE, S.; TORN, M. S. The technology path to deep greenhouse gas emissions cuts by 2050: the pivotal role of electricity. *science*, American Association for the Advancement of Science, v. 335, n. 6064, p. 53–59, 2012. Citado na página 13.

ZHAO, H.-x.; MAGOULÈS, F. A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, Elsevier, v. 16, n. 6, p. 3586–3592, 2012. Citado na página 13.

ZHU, Y.; LU, S. Load profile disaggregation by blind source separation: A wavelets-assisted independent component analysis approach. In: IEEE. *2014 IEEE PES General Meeting| Conference & Exposition*. [S.l.], 2014. p. 1–5. Citado na página 16.

APPENDIX A – Covariance structure mispecification study structure

The covariance structure of the aggregated data model describes the correlation between two points of the observed functional response variable. Although it does not influence the expected value of the disaggregated curve least square estimator, the covariance matrix is important to build its confidence intervals.

The experiment below was designed to assess the impact of covariance structure mispecification on typical curves estimation.

A.1 Setup

Four scenarios with $R = 30$ simulated datasets composed by $C = 3$ types of customers and $J = 10$ substations were generated:

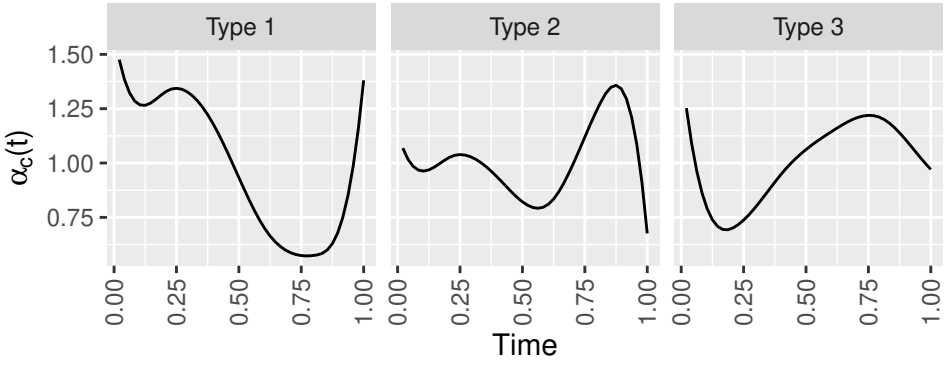
- Scenario 9: Homogeneous uniform covariance structure with 5 days,
- Scenario 10 Homogeneous uniform covariance structure with 30 days,
- Scenario 11 Complete covariance structure with 5 days,
- Scenario 12 Complete covariance structure with 30 days.

For each scenario, a homogeneous uniform, a homogeneous and a complete aggregated data model is fitted to observe the estimated typical curves. For this case, markets of each dataset were generated as random numbers from 5 to 20.

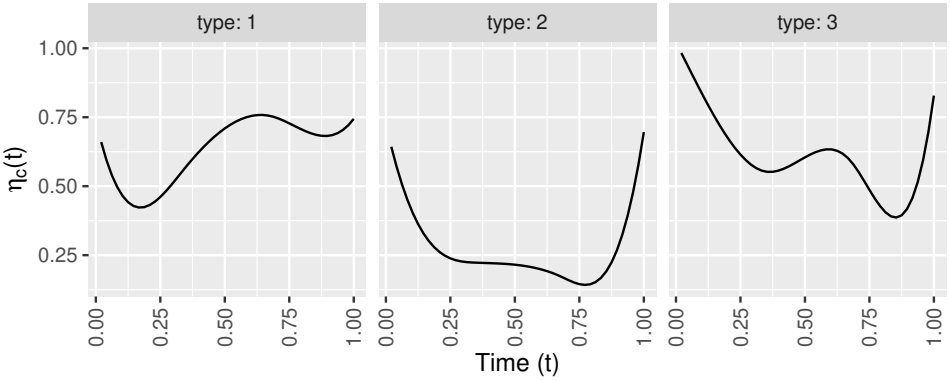
Figure 31a shows the true typical curves used to simulate the datasets in the four scenarios. Homogeneous uniform scenarios have covariance parameters set as $\sigma_c = 0.60$ and $\omega_c = 0.50$ for all customers of type $c = 1, 2, 3$. On the other hand, complete scenarios are composed by decay parameters $\omega_1 = 0.50$, $\omega_2 = 0.25$ and $\omega_3 = 0.125$ and variance functionals displayed in Figure 31b.

A.2 Homogeneous uniform scenarios

In this case, the homogeneous and the complete overparametrize the covariance structure data because they are composed by covariance structures with more parameters than the true homogeneous uniform scenario. Figures 32, 33 and 34 show the homogeneous



(a) True typical curves.



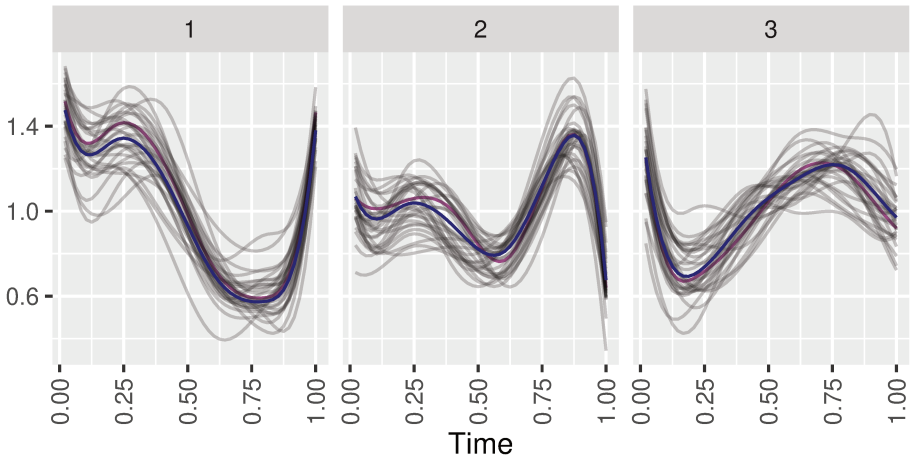
(b) True variance functionals.

Figure 31 – (a) True typical curves for each custumer of type $c = 1, 2, 3$ used to estimate the simulated dataset in the four scenarios and (b) true variance functionals for each customer of type $c = 1, 2, 3$ of scenarios 11 and 12.

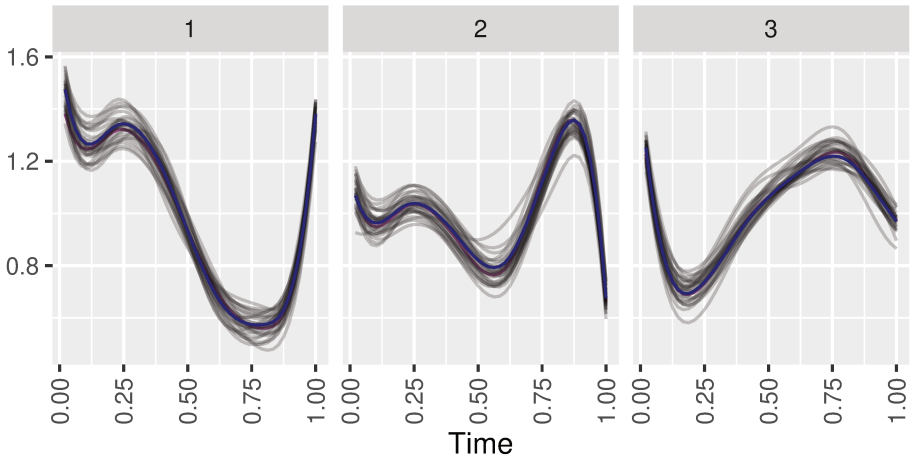
uniform, homogeneous and complete fit, respectively, for Scenarios 9 and 10. The three figures show good typical curve estimation with estimated curves oscillating around the true curve. As expected, the scenario with 30 days presents estimated curves closer to their true curve in the three fitted models.

A.3 Complete scenarios

In this section, the homogeneous uniform and homogeneous fit are simpler cases nested in the complete covariance structure. Figures 35, 36 and 37 show the homogeneous uniform, homogeneous and complete fit, respectively, for Scenarios 11 and 12. Again, The three figures show good typical curve estimation with estimated curves oscillating around the true curve.



(a) 5 days



(b) 30 days

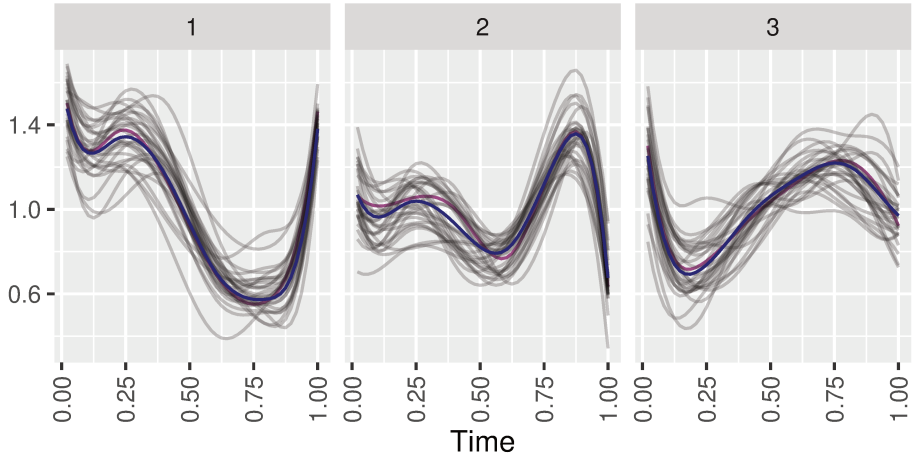
Figure 32 – Estimated typical curves (in gray) under the homogeneous uniform aggregated data model fit for homogeneous uniform scenarios 9 and 10, median curve of the estimated curves in magenta and true curve in blue.

A.4 Conclusion

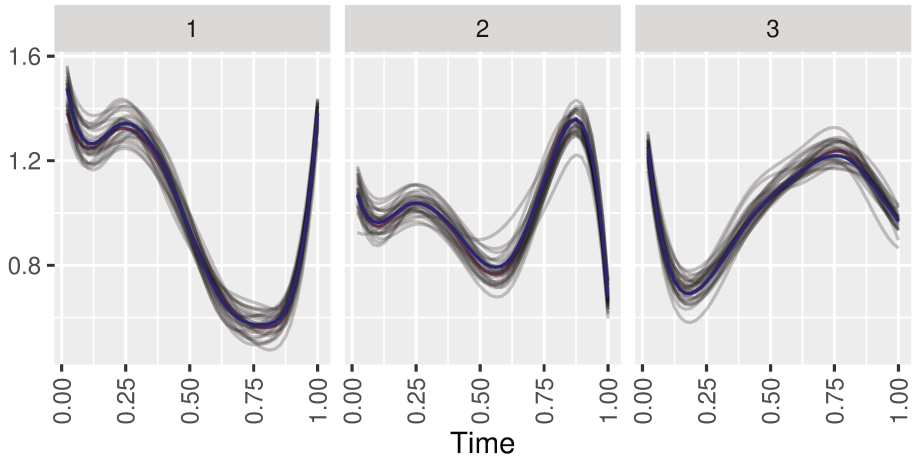
Compared to scenarios 9 and 10, there is no clear distinction in terms of estimated curves. All cases in all scenarios presented good estimates with median curves close to the true curves. This indicates that the least square estimator for typical curves can be robust to covariance mispecification, as mentioned in Section 2.4.1.

Because this section was focused in the estimated typical curves, there is no analyses for covariance parameter estimation under mispecification of covariance structure.

ANEXOS

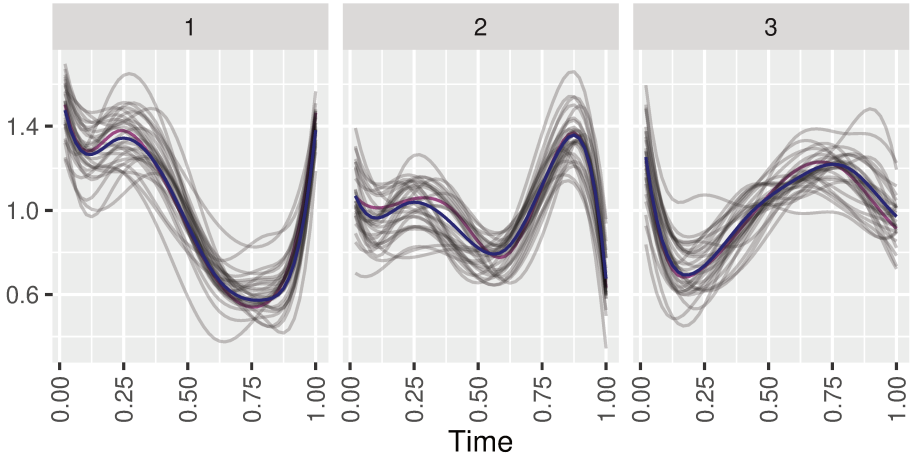


(a) 5 days

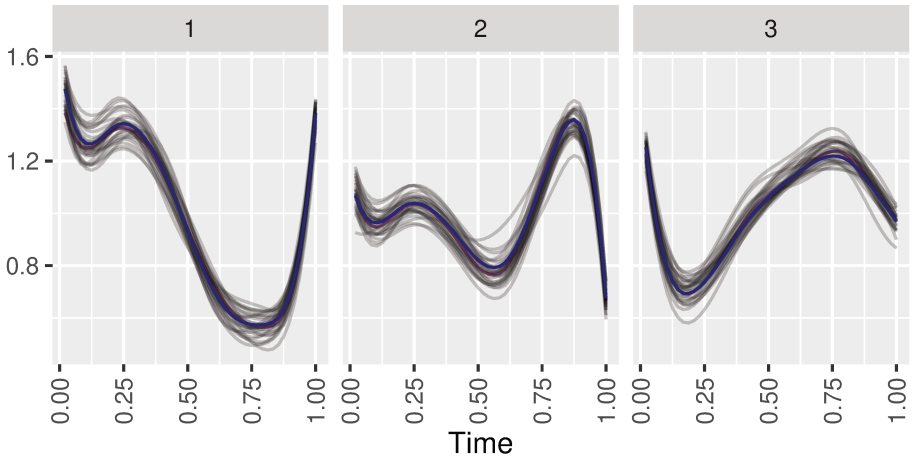


(b) 30 days

Figure 33 – Estimated typical curves (in gray) under the homogeneous aggregated data model fitfor homogeneous uniform scenarios 9 and 10, median curve of the estimated curves in magenta and true curve in blue.

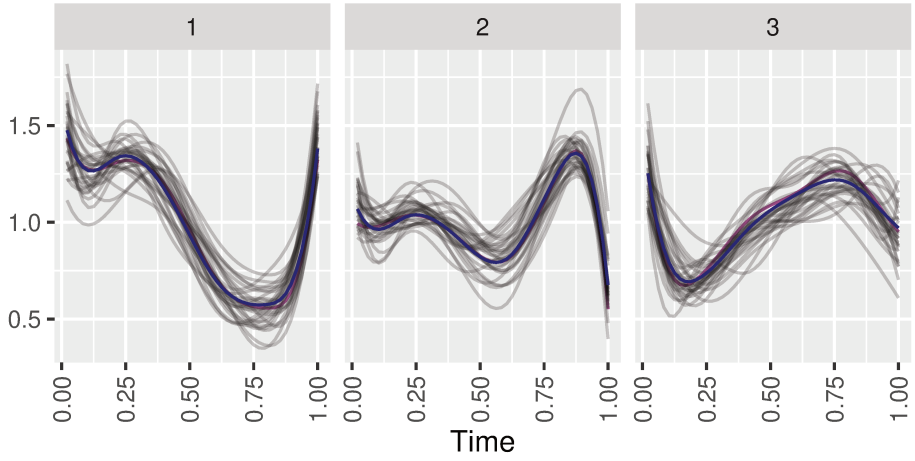


(a) 5 days

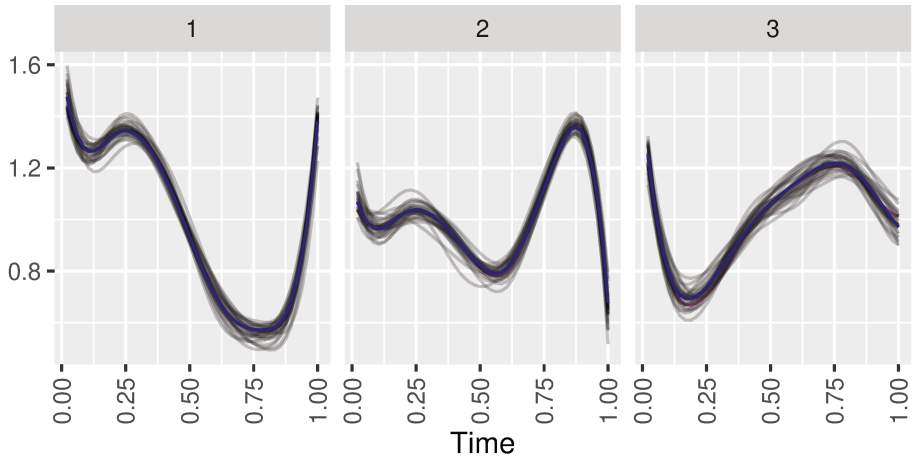


(b) 30 days

Figure 34 – Estimated typical curves (in gray) under the complete aggregated data model fit for homogeneous uniform scenarios 9 and 10, median curve of the estimated curves in magenta and true curve in blue.

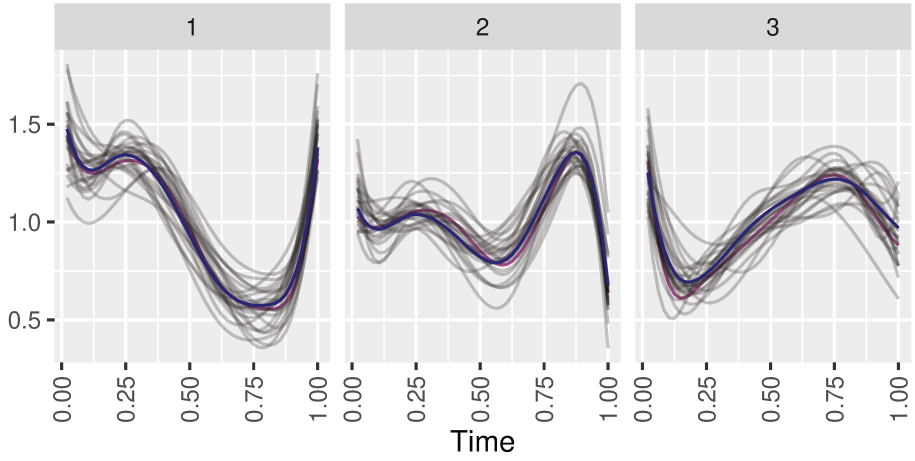


(a) 5 days

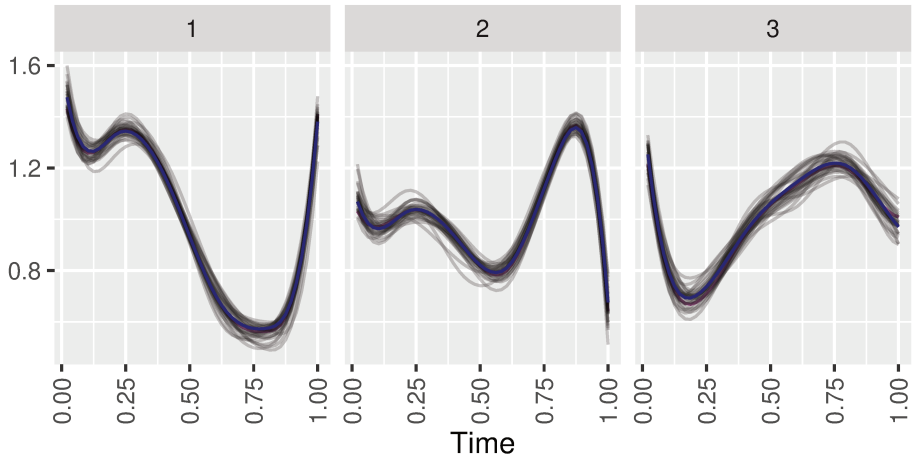


(b) 30 days

Figure 35 – Estimated typical curves (in gray) under the uniform homogeneous aggregated data model fit for complete scenarios 11 and 12, median curve of the estimated curves in magenta and true curve in blue.

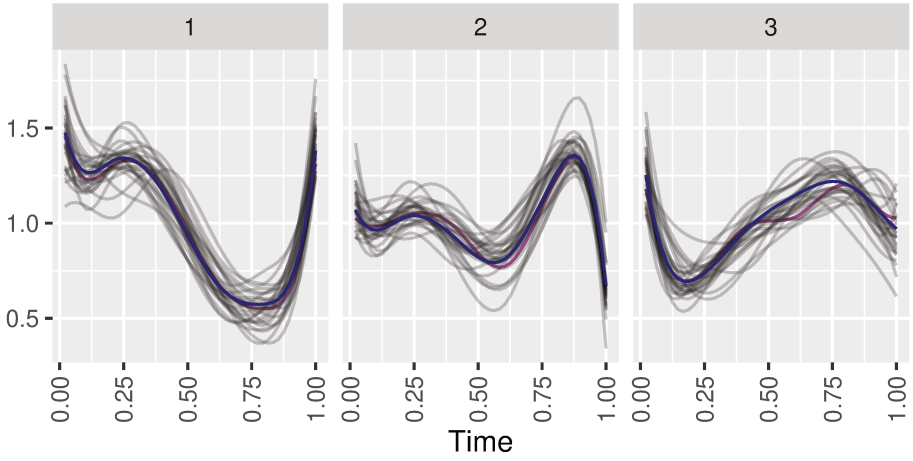


(a) 5 days

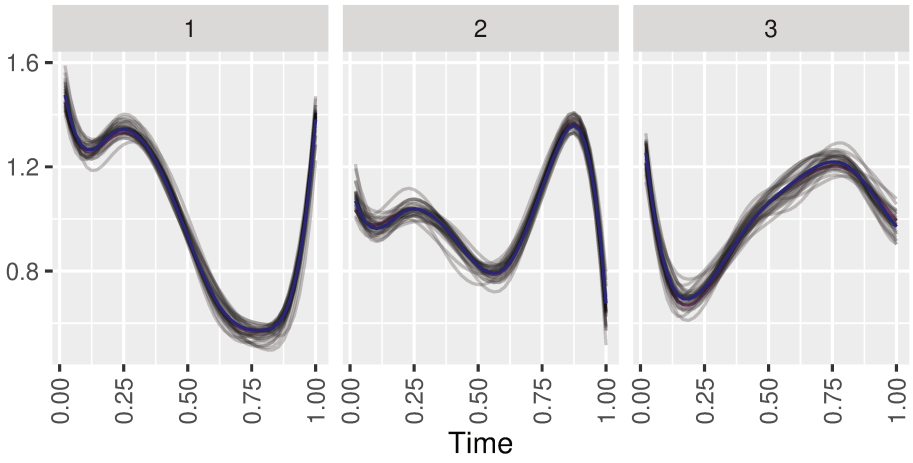


(b) 30 days

Figure 36 – Estimated typical curves (in gray) under the homogeneous aggregated data model fit for complete scenarios 11 and 12, median curve of the estimated curves in magenta and true curve in blue.



(a) 5 days



(b) 30 days

Figure 37 – Estimated typical curves (in gray) under the complete aggregated data model fit for complete scenarios 11 and 12, median curve of the estimated curves in magenta and true curve in blue.

ANNEX A – Supplementary tables

Table 22 – Likelihood ratio test comparison table of homogeneous and complete aggregated data models in simulated datasets for each experimental run. The degrees of freedom used to compute the p-value is 10 for all comparisons.

		Log-likelihood			
Days	Run	Homogeneous	Complete	Test statistic	p-value
5 days	1	11356.44	11206.11	300.6677	<0.0001
5 days	2	11468.83	11279.91	377.8489	<0.0001
5 days	3	11329.26	11129.60	399.3154	<0.0001
5 days	4	11393.79	11237.21	313.1679	<0.0001
5 days	5	11380.05	11161.95	436.2019	<0.0001
5 days	6	11412.08	11240.27	343.6055	<0.0001
5 days	7	11409.96	11230.16	359.5910	<0.0001
5 days	8	11380.91	11221.02	319.7964	<0.0001
5 days	9	11296.43	11111.48	369.8967	<0.0001
5 days	10	11293.50	11131.60	323.7867	<0.0001
5 days	11	11297.66	11132.56	330.1824	<0.0001
5 days	12	11380.86	11204.74	352.2391	<0.0001
5 days	13	11305.42	11135.14	340.5573	<0.0001
5 days	14	11357.41	11182.27	350.2775	<0.0001
5 days	15	11337.32	11153.20	368.2447	<0.0001
5 days	16	12252.88	12167.35	171.0486	<0.0001
5 days	17	12199.07	12079.56	239.0298	<0.0001
5 days	18	12315.08	12222.54	185.0706	<0.0001
5 days	19	12394.19	12316.81	154.7764	<0.0001
5 days	20	12254.24	12160.62	187.2429	<0.0001
5 days	21	12286.01	12189.36	193.3048	<0.0001
5 days	22	12314.49	12225.54	177.8949	<0.0001
5 days	23	12173.57	12073.10	200.9313	<0.0001
5 days	24	12248.57	12150.72	195.6961	<0.0001
5 days	25	12197.71	12090.13	215.1580	<0.0001
5 days	26	12304.61	12196.86	215.4922	<0.0001
5 days	27	12135.35	12034.37	201.9598	<0.0001

5 days	28	12385.44	12302.31	166.2589	<0.0001
5 days	29	12359.07	12253.33	211.4677	<0.0001
5 days	30	12249.15	12142.30	213.7031	<0.0001
30 days	1	68412.15	67418.28	1987.7512	<0.0001
30 days	2	68555.41	67692.07	1726.6639	<0.0001
30 days	3	68409.85	67507.56	1804.5693	<0.0001
30 days	4	68848.74	67854.20	1989.0811	<0.0001
30 days	5	68853.23	67912.53	1881.3950	<0.0001
30 days	6	68254.05	67181.14	2145.8196	<0.0001
30 days	7	68428.22	67442.46	1971.5111	<0.0001
30 days	8	68720.86	67794.68	1852.3555	<0.0001
30 days	9	68143.11	67241.85	1802.5107	<0.0001
30 days	10	68639.51	67645.38	1988.2693	<0.0001
30 days	11	68532.07	67568.32	1927.5025	<0.0001
30 days	12	68452.58	67430.66	2043.8409	<0.0001
30 days	13	68697.84	67687.06	2021.5609	<0.0001
30 days	14	68183.63	67221.51	1924.2479	<0.0001
30 days	15	68828.41	67972.03	1712.7790	<0.0001
30 days	16	73834.91	73325.45	1018.9162	<0.0001
30 days	17	74309.19	73894.94	828.4990	<0.0001
30 days	18	74396.67	73964.25	864.8214	<0.0001
30 days	19	74607.51	74136.77	941.4789	<0.0001
30 days	20	74741.17	74284.07	914.1883	<0.0001
30 days	21	74686.06	74273.96	824.1941	<0.0001
30 days	22	74678.44	74242.28	872.3236	<0.0001
30 days	23	74295.85	73798.20	995.2966	<0.0001
30 days	24	73439.84	72817.88	1243.9196	<0.0001
30 days	25	74510.71	74113.48	794.4602	<0.0001
30 days	26	74232.01	73790.03	883.9603	<0.0001
30 days	27	74783.54	74378.31	810.4463	<0.0001
30 days	28	74510.23	74068.60	883.2588	<0.0001
30 days	29	74963.07	74588.12	749.8957	<0.0001
30 days	30	74913.23	74506.08	814.3098	<0.0001

Table 23 – BIC values for the clustering aggregated data models in simulated datasets at each experimental runs. NA values represent the runs that did not converge.

		BIC		
Days	Run	2 Clusters	3 Clusters	BIC diff
5 days	1	22892.98	22653.75	239.24
5 days	2	23037.71	22728.32	309.38
5 days	3	22765.23	22522.55	242.68
5 days	4	23002.98	22632.54	370.44
5 days	5	NA	NA	NA
5 days	6	22867.82	22589.01	278.82
5 days	7	22928.48	22674.78	253.70
5 days	8	23057.19	NA	NA
5 days	9	23110.73	22797.53	313.20
5 days	10	NA	NA	NA
5 days	11	23023.11	22788.12	234.99
5 days	12	22893.37	22633.57	259.80
5 days	13	23033.60	22798.68	234.93
5 days	14	22996.42	22713.55	282.87
5 days	15	23127.33	22832.22	295.11
5 days	16	24320.71	23984.69	336.02
5 days	17	24930.08	24283.66	646.42
5 days	18	24508.43	NA	NA
5 days	19	24306.36	24053.74	252.62
5 days	20	24582.05	24364.63	217.42
5 days	21	24138.96	23911.80	227.16
5 days	22	24580.06	24236.24	343.83
5 days	23	24477.64	NA	NA
5 days	24	24365.16	24106.78	258.37
5 days	25	NA	NA	NA
5 days	26	24842.38	24266.45	575.93
5 days	27	23893.08	23763.69	129.38
5 days	28	24568.87	24068.67	500.20
5 days	29	24718.34	24275.80	442.54
5 days	30	24304.65	24030.62	274.03
30 days	1	140732.75	139731.99	1000.75
30 days	2	137862.55	137674.83	187.73

30 days	3	141129.86	140153.32	976.53
30 days	4	139429.37	NA	NA
30 days	5	138120.81	137360.21	760.60
30 days	6	138909.99	NA	NA
30 days	7	138244.85	NA	NA
30 days	8	139216.20	138601.06	615.15
30 days	9	138597.82	137882.19	715.63
30 days	10	137915.18	NA	NA
30 days	11	139329.72	138463.56	866.16
30 days	12	138265.27	NA	NA
30 days	13	139177.05	138376.03	801.02
30 days	14	137716.41	136986.04	730.38
30 days	15	138839.29	138544.00	295.30
30 days	16	148793.81	NA	NA
30 days	17	151598.28	149765.27	1833.01
30 days	18	148195.67	146904.43	1291.24
30 days	19	146444.55	145965.37	479.17
30 days	20	150741.62	148847.58	1894.04
30 days	21	145479.22	144462.94	1016.28
30 days	22	147632.69	146775.49	857.20
30 days	23	149746.58	148789.10	957.48
30 days	24	NA	148840.17	NA
30 days	25	151627.47	150116.68	1510.80
30 days	26	NA	150376.76	NA
30 days	27	150394.95	NA	NA
30 days	28	149657.30	147460.43	2196.86
30 days	29	147228.13	146344.21	883.91
30 days	30	152994.93	150468.24	2526.69