Statistics for Industry, Technology, and Engineering

Anestis Antoniadis
Jairo Cugliari
Matteo Fasiolo
Yannig Goude
Jean-Michel Poggi

Statistical Learning Tools for Electricity Load Forecasting

Birkhäuser

Statistics for Industry, Technology, and Engineering

The *Statistics for Industry, Technology, and Engineering* series will present up-to-date statistical ideas and methods that are relevant to researchers and accessible to an interdisciplinary audience: carefully organized authoritative presentations, numerous illustrative examples based on current practice, reliable methods, realistic data sets, and discussions of select new emerging methods and their application potential. Publications will appeal to a broad interdisciplinary readership including both researchers and practitioners in applied statistics, data science, industrial statistics, engineering statistics, quality control, manufacturing, applied reliability, and general quality improvement methods.

Principal Topic Areas:

* Quality Monitoring * Engineering Statistics * Data Analytics * Data Science * Time Series with Applications * Systems Analytics and Control * Stochastics and Simulation * Reliability * Risk Analysis * Uncertainty Quantification * Decision Theory * Survival Analysis * Prediction and Tolerance Analysis * Multivariate Statistical Methods * Nondestructive Testing * Accelerated Testing * Signal Processing * Experimental Design * Software Reliability * Neural Networks *

The series will include professional expository monographs, advanced textbooks, handbooks, general references, thematic compilations of applications/case studies, and carefully edited survey books.

Anestis Antoniadis • Jairo Cugliari •
Matteo Fasiolo • Yannig Goude •
Jean-Michel Poggi

Statistical Learning Tools for Electricity Load Forecasting

Birkhäuser

Anestis Antoniadis
Naples Unit
Institute of Applied Sciences and Intelligent Systems 'Eduardo Caianiello'
Naples, Italy

Matteo Fasiolo
School of Mathematics
University of Bristol
Bristol, UK

Jean-Michel Poggi
Laboratoire de Mathematiques
Université Paris-Saclay
GIF SUR YVETTE, France

Jairo Cugliari
Lab. ERIC EA 3083
Lumière University Lyon 2
LYON CEDEX 07, France

Yannig Goude
Laboratoire de Mathematiques
Universite Paris-Sud
Orsay, France

ISSN 2662-5555 ISSN 2662-5563 (electronic)
Statistics for Industry, Technology, and Engineering
ISBN 978-3-031-60338-9 ISBN 978-3-031-60339-6 (eBook)
https://doi.org/10.1007/978-3-031-60339-6

This book is published under the imprint Birkhäuser, www.birkhauser-science.com by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Chapter 1
Introduction

This book is the result of our experience in the field of short-term forecasting using "high-resolution" data sets generated in the context of industrial and scientific applications. We illustrate the basic theory and practical utility of several up-to-date statistical methods, with particular emphasis on their functional data analysis aspect. The book should be useful beyond the electrical context because it discusses methods and models that extend to other applications, such as forecasting of seasonal phenomena, possibly influenced by external factors (e.g., call centers activity, public hot water supply, airport passenger traffic, etc.).

The use of statistical and machine learning techniques to solve real-life problems has been undergoing tremendous change in the last few decades. More and more branches of science and engineering require modern approaches merging statistics, machine learning, and software tools, especially in the context of forecasting, expert systems and Big Data. The main aim of the present work is introducing a set of statistical and machine learning tools, and showing how they can be used effectively for applied data analysis in the context of electricity load forecasting. While some of the topics addressed in the manuscript are rather classical, our presentation follows an unconventional approach, which has been inspired by our research and applied work on forecasting electricity demand in an industrial setting. Indeed, we aim at guiding the reader through a number of modern forecasting ideas, from an industrial and applied perspective, centered around a collection of case studies. Several of the examples treated in the book are based on sizeable high-resolution data sets, thus giving a realistic account of the managerial, descriptive, and predictive challenges faced by real-world statisticians and engineers.

This work is the result of the long-lasting friendship and scientific collaboration of the five authors with Électricité de France (EDF); hence we consider mostly problems and data sets related to the electricity industry. However, the workflows and tools presented in this book are applicable more broadly and reflect our views of what the practice of functional data analysis for high-dimensional data should

A. Antoniadis et al., *Statistical Learning Tools for Electricity Load Forecasting*, Statistics for Industry, Technology, and Engineering,
https://doi.org/10.1007/978-3-031-60339-6_1

be, with particular focus on predictive analytics and the presence of external or exogenous factors. Hence, in the following treatment we aim at striking a balance between illustrating how to use certain statistical tools and R packages to solve specific applied problems, and providing the theoretical and methodological background that justifies and motivates their use.

Being based on the authors' collaboration on load forecasting with EDF, this book is biased toward the statistical models and methods that they have found most useful in this context. For example, generalized additive models (GAMs) have been used extensively in their EDF-related work; hence several chapters of this book feature such models. In contrast, classical time series methods, such as autoregressive models, are covered only marginally. The reason is that, in their load forecasting work, the authors have relied more heavily on models that are able to flexibly capture the effect of exogenous variables, such as GAMs, than on time series models designed to more rigorously handle the time series aspect of the data. Further, the use of time series models for load forecasting has already been covered by Weron (2007), a reference that the interested reader could use to fill the gaps in the present book. The selection of methods and models presented here was also dictated by the industry-specific need to use models characterized by a high degree of interpretability. In particular, to be trusted for use in operations, a model must produce a load forecast that can be (at least approximately) decomposed into separate effects (of, e.g., temperature, seasonality, etc.). Hence, hard-to-interpret models, such as deep neural networks, are not covered in this book.

The book is suitable for several audiences. As should be clear from its title, it is intended for practitioners, researchers, and post-graduate students working on electricity load forecasting. More broadly, it can be of interest to applied academics or scientists (e.g., biostatisticians, econometricians, and quantitative scientists in general) wanting to learn about cutting-edge forecasting tools, but not necessarily interested in the electricity industry. Another potential audience consists of statistically oriented scientists, having already a working knowledge of traditional forecasting tools for industrial applications, and wanting to explore more flexible semi-parametric functional data analysis models. The most advanced material contained in this book should be of interest even to experienced data science professionals.

The book assumes that the reader is familiar with standard statistical concepts such as random variables, probability density functions, and expected values. It also assumes that the reader has some minimal statistical modelling experience. Little or no prior knowledge of the electricity industry is assumed. Some topics covered in the book are more advanced than others and might not be easily accessible to all readers. An example is Chap. 6 on aggregation of experts, which should be considered an advanced topic. Conversely, the book contains several chapters on GAMs which, if read in the right order, should be accessible to readers with little background on such models. In particular, Chap. 2 introduces this model class, which is then extended in Chaps. 3 and 7. Having covered the material in these chapters, the reader should be able to get the most out of the case studies covered in Chaps. 10 and 12.

1.1 Industrial Motivation

Forecasts are of fundamental importance for energy markets. To guaranty the equilibrium between consumption and production at any time on the grid, accurate electricity load forecasting at different horizons are needed. Load forecasting is usually performed at different horizons of time and different spatial resolutions. Horizons of forecasts range from intraday (10 minutes to 24 hours ahead) to daily, weekly, monthly, or even a few years in advance, whereas spatial resolution can go from an individual house, substations on the grid, regional, or national level. Industrial needs cover production planning, demand response, grid management, electricity trading, risk management, optimization of production units maintenance, and commercialization. Due to the energy transition and the development of IT technologies, energy systems face many challenges impacting forecasting activities.

At an aggregated level, salient features presented in Sect. 1.2.2 are well known but can face some changes due to the evolution of socio-economic conditions, climate variations, or energy market regulation. The consumption habits constantly evolve due to technological progress and energy transition: Low-energy bulbs, air conditioners, TVs, smartphones, better insulated buildings, electric cars are all examples of new usages which affect or will affect electricity demand. Another recent example of social changes is the increase of people working from home during the COVID 19 lockdown. Economic development also plays an important role regarding the consumption of energy-intensive industrial producers (chemistry, paper industry, steel industry, etc.). On the regulatory side, the further opening of the energy market which occurred in France in 2004 for professional consumers and in 2007 for all consumers entails a need for electricity provider to forecast the consumption of time-varying portfolios of customers.

Electricity production is progressively moving to more intermittency and complexity with the increase of renewable energy and the development of small distributed production units such as photovoltaic panels or wind farms. To maintain the electricity quality, energy stakeholders are developing smart grids, the next-generation power grid including advance communication networks and associated optimization and forecasting tools. A key component of the smart grids is smart meters. They allow two-sided communication with the customers, real-time measurement of consumption, and a large scope of demand side management services. These new technologies structure intelligent networks, with more local and high-resolution information making available in real time individual load curves or geographical aggregate. This results into new opportunities such as local optimization of the grid, demand side management, and smart control of storage devices.

Exploiting the smart grid efficiently requires advanced data analytics and optimization techniques to improve forecasting, unit commitment, and load planning at different geographical scales. Massive data sets are and will be produced: energy consumption measured by smart meters at a high-frequency (a few minutes mainly) data from the grid management (e.g., Phasor Measurement Units); data from energy markets (prices and bidding, TSOs and DSOs data like balancing and capacity); data

from production units and equipment for their maintenance and control (sensors, periodic measures, so on). A lot of efforts are made by utilities to develop data lakes and IT structures to gather and make these data available for their business units in real time. Designing new algorithms to analyze and process these data at scale is a key activity and a real competitive advantage. In particular, due to the lack of stationarity, greater volatility, and the need to consider more flexible, adaptive strategies motivate the need to develop and apply new statistical learning methods. As they are acting in the real world and used to make critical decisions, these methods and models need to be at the same time more and automatic and adaptive to preserve their accuracy but also robustness, trustworthiness, and interpretability.

1.2 Data Sets

Electrical load forecasting strategies are the result of quite different approaches coming from economics, statistics, engineering, and computer science. In general, we can state the objective as the most early and accurate anticipation of the electrical system needs at some aggregation and resolution levels. One may be interested in both pointwise predictions (e.g., peak, median, or mean loads), the whole distribution, or at least some prediction interval. In any case, the main inputs for the prediction strategies are historical records of electricity demand as well as other data describing factors that drive the electricity demand. We highlight one data set that is the running data set, the most frequently used. In addition, we complete with other real data sets.

1.2.1 General Considerations

Electricity data sets we will manipulate in this chapter are composed of two kind of information: time series and static features. In a lot of machine learning applications to time series, observations are considered as independent and identically distributed (i.i.d.) observations, but we can deal with them in many ways as detailed bellow.

Time Series Data In the time series category we include electricity consumption itself but also related time series like meteorological observations/forecasts, calendar information, or economical indicators (other time series data related to other kinds of consumption like gas consumption, water consumption, waste production, phone communications, etc., could be considered as interesting information to forecast or model the electricity load, but we will not consider it in this book). Electricity load could be measured at different time and spatial scales. The time scale is the frequency at which the observations are measured. Most of the electricity data sets consist of half-hourly measurement (mean of electricity load during 30 minutes), but recent works in the field of nonintrusive load monitoring consider

high-frequency data at the under second time frequency (see, e.g., the UK Dale data presented in Kelly and Knottenbelt (2015)). There are so many alternative ways to deal with these time series data.

Functional Data The above time–frequency measurements of electricity load may lead to two different views for handling the daily aspect of such time series data, a longitudinal one (LDA) and a functional data analysis (FDA) approach. Measurements treated in the FDA literature typically are recorded by high-frequency automatic sensing equipment, whereas those treated in the LDA literature are more typically sparsely, and often irregularly, spaced measurements. Despite the differences between these two aspects, there are many common aims in their objectives, many of these objectives entailing smoothing data, either explicitly or implicitly, characterizing average or "typical" time trends, and assessing the relationships of shapes of daily loads to covariates. Commonly, time series data are treated as multivariate data because they are given as a finite discrete time series. In the LDA literature, smoothers and extrapolators typically arise from stochastic models, in particular from classical time series models, and the data analysis is concerned with daily time series in the form of random vectors. This usual multivariate approach completely ignores important information about the smooth functional behavior of the generating process that underpins the data. The basic idea behind FDA is to express discrete observations arising from time series in the form of a function that represents the entire measured function as a single observation, using smoothing and interpolation procedures, and then to draw modelling and/or prediction information from a collection of functional data by applying appropriate statistical concepts from data analysis. Therefore, functional data analysis (FDA) goes one big step further than LDA, focusing on data that are infinite-dimensional, such as curves or shapes. However, proven methods of longitudinal analysis form the backbone for some of the prominent techniques of FDA. Areas that FDA draws upon besides longitudinal analysis include nonparametric regression, functional analysis (linear operators in Hilbert space), and properties of square integrable stochastic processes. Several chapters in this book examine the feasibility of suitable nonparametric functional models allowing to handle various configurations of observed data.

Spatial Data Regarding the spatial scale there are two main directions we will look at. One way is to consider individual data coming from the development of smart meters, data at the household spatial resolution is thus available and can be viewed as the basic unit we will manipulate on many kinds of problems: clustering, bottom-up forecasting, and peak forecasting. Sub-household data, e.g., by usage, can be considered in the field of nonintrusive load monitoring, but we will not deal with such a low level of aggregation in this book. An other way is to consider data from the grid; thus spatial resolution is constrained by the structure of the network. This is the case of substation data we will consider here in the GEFCOM data set. Meteorological data used in forecasting model design consists mainly of observed features like temperature, cloud cover, wind, humidity, and solar radiation at some meteorological stations in an area. Association with load data raises the issue of time and spatial correspondence. Time resolution of meteorological data is

often not the same as electricity load one leading to interpolation questions. Space correspondence could be a complex problem depending on whether we have access to the localization of electricity data (location of household for the smart meter data, end users for grid data).

Static Data Other information is often provided with time series electricity data. For smart meters data this could be household characteristics such as the type of heating, tariff, social class, or location. For substation data this could be socio-economic indicators from census, or static technical data such as the characteristics of the lines, and the number and types of customers connected to a node.

1.2.2 Salient Features of Electricity Demand

Among other, common salient features of the electricity demand are: a *long-term trend*, describing population increase and intensification of electricity usages; *annual cycles*, due to socio-economic and meteorological seasonal patterns; and *exogenous factors*, as for instance the dependence against temperature—explained by cooling and heating systems. A more detailed account of these factors can be found in Cugliari and Poggi (2020). Notice that these factors and not static ones and so may evolve. Load varies across different hours of the day and different days of the week, thus making it a highly time-varying quantity. Additional seasonal patterns, which accommodate the evolution of winter daily profiles to those during summer, add more complexity to the time series. Effects of weather conditions, in particular temperature, complicate matters further. Thus, finding reliable models to forecast electricity load is a challenging task. Several chapters in this book aimed at developing appropriate models for electricity load forecasting.

1.2.3 Irish Individual Electrical Demand Data

1.2.3.1 Data Presentation

This data set was published by the Commission for Energy Regulation (CER) Commission for Energy Regulation (2011) as part of the Smart Metering Project, which started in 2007. The main goal of the project was assessing the performance of smart meters, analyzing their impact on customers' consumption and quantifying the related costs. The full data set is composed of 4623 of individual demand time series, each containing 48 half-hourly observations per day and covering the period from July 2009 to December 2010. The data set has been studied in a number of papers, for example, Quilumba et al. (2015) and Wang et al. (2018a) use the individual loads to improve the forecast of the aggregate demand, while Capezza et al. (2020) focus on forecasting the individual time series directly.

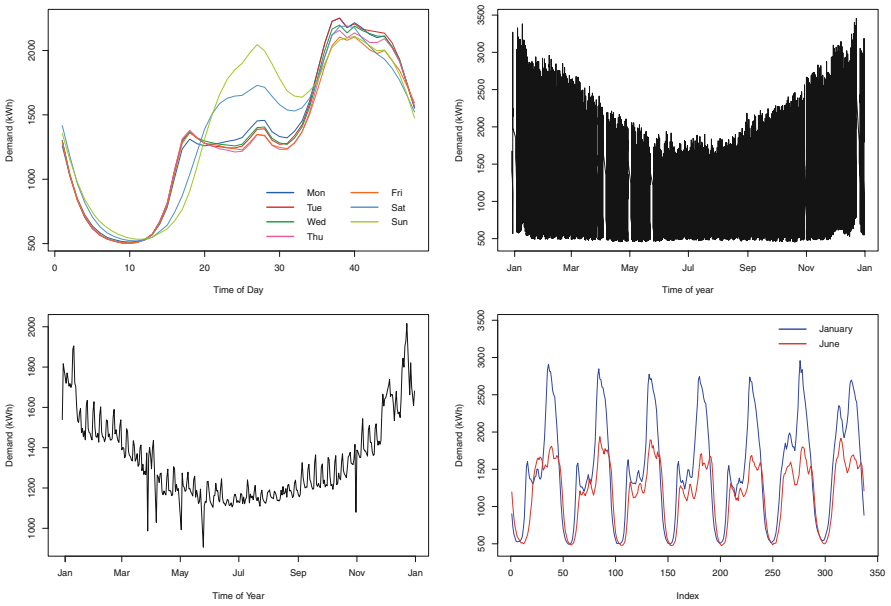

Fig. 1.1 Characteristics of the aggregated load of the 2672 Irish residential customers. See the main text for details

The full data set contains data from both residential and commercial customers. In this book we work only on the data from 2672 households. For each household the data set contains information about the customer's tariff as well as survey data specifying, among other variables: the type of heating used, whether the windows are double-glazed, the number of appliances, and the year of construction. The location of each client is unknown to preserve privacy.

Figure 1.1 shows the characteristics of the aggregated demand of all the customers. The top-left plot shows the daily profile of the mean consumption for each day of the week, the top-right and bottom-left the consumption over the year at an half-hourly and daily resolution, and the bottom-right two weeks of consumption in the winter and in the summer. We clearly see the three calendar patterns of the load: daily, weekly, and yearly.

Figure 1.2 shows that the signal-to-noise ratio of the demand data predictive decreases with the level of granularity. In particular, Plots 1.2a to d show that, while the daily profile is smooth when demand is averaged across the customers, disaggregating the demand leads to rough, less predictable profiles. The low signal-to-noise ratio characterizing individual household demand might suggest an individual demand modelling strategy based on predicting the data from several customers using a single model, to reduce the noise. However, Plots 1.2e and f show that the behavior of customers is highly heterogeneous; hence naïve aggregation would induce much bias. See Capezza et al. (2020) for a discussion of the challenges of modelling individual demand trajectories.

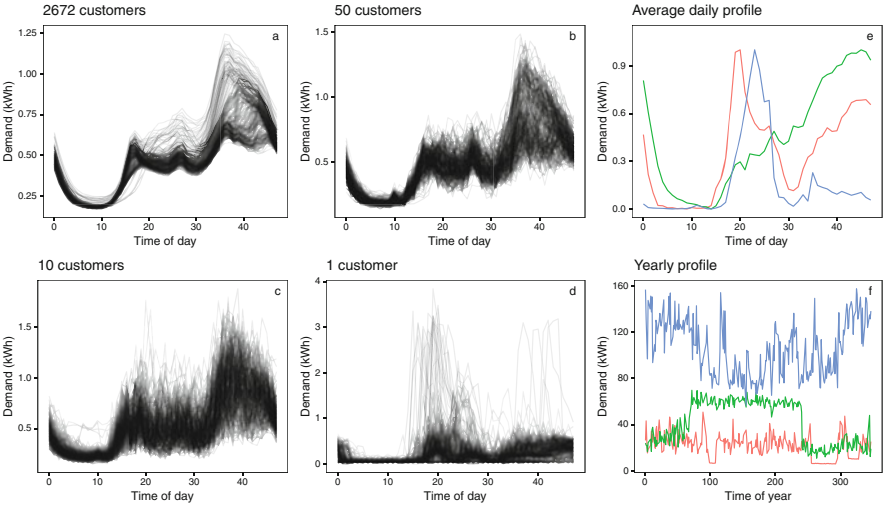

Fig. 1.2 Plots (**a**–**d**) show the daily profiles of the demand averaged over increasingly small groups of customers from the CER trial (Commission for Energy Regulation 2011). Plots (**e**) and (**f**) show the average daily and yearly demand profiles of three customers. The blue profile in Plot (**f**) has been vertically shifted for visibility

1.2.3.2 Data Processing

The data contains a number of bad observations, mainly missing or duplicated values, which need to be dealt with. We simply remove them from the data set. As the first 6 months of data correspond to a control period, after which the customers are subjected to incentives aimed at modifying their consumption pattern, we keep only data from year 2010. We integrate the demand data with hourly temperatures, interpolated at the half-hourly resolution, from the National Centers for Environmental Information (NCEI). The temperature data was measured at ten different locations in Ireland. We built a single temperature variable by averaging these temperatures with uniform weights.

Modelling the individual or aggregate demand requires dealing with bank holidays and other special days, for which a forecast could be produced only if we have had several years of data (i.e., to estimate the Christmas effect we would need several observed Christmases). In particular, modellers must take into account days of the year equal to 1, 2 (first two days of the year), 87 (Sunday before Easter), 94, 95, 96 (Easter and two following days), 120, 121, 122 (May Day and two days before), 143, 144, 145 (Pentecost Monday), 304 (Halloween), 358, 359, 360 (24, 25, and 26 December), and 365 (New Year's Eve).

Depending on the model being used, modelling individual demand data might require excluding customers for which the demand data is anomalous. For example, Capezza et al. (2020) exclude customers for which the 99-th quantile of the electricity demand over the entire year is less than 0.4kWh. The aim is removing

customers whose demand is near zero for most of the year. They also exclude all customers for which the vector of differences of consecutive demand values contained more than 2500 zeros, over the entire year. These are customers whose demand is constant for long period. The two filtering criteria just mentioned lead to a data set of 2565 customers.

1.2.3.3 Getting the Data

The full data set, containing demand data from both individual and commercial customers, can be obtained from the Commission for Energy Regulation website (https://www.ucd.ie/issda/data/commissionforenergyregulationcer/). A preprocessed data set, covering residential customers only, is contained in the `electBook` `R` package and can be loaded by typing:

```
library(electBook)
data(Irish)
```

`Irish` is a list with three elements. In particular, `Irish$indCons` is a matrix where each row is the demand for an individual household over one year, while `Irish$survey` is a `data.frame` containing the survey information:

```
head(Irish$survey)
```

```
##      ID   meanDem SOCIALCLASS OWNERSHIP BUILT.YEAR HEAT.HOME
## 1 I1002 0.2081436          DE         O       1975     Other
## 2 I1003 0.6215765          C1         O       2004     Other
## 3 I1004 0.9617103          C1         O       1987     Other
## 4 I1005 0.6402214          C1         O       1930     Other
## 5 I1013 0.2414805          C2         O       2003     Other
## 6 I1015 0.4631413          DE         R       1989      Elec
##   HEAT.WATER WINDOWS.doubleglazed HOME.APPLIANCE..White.goods. Code
## 1       Elec                 All                             1    1
## 2      Other                 All                             5    1
## 3       Elec                 All                             5    1
## 4      Other                 All                             4    1
## 5       Elec                 All                             3    1
## 6      Other                 All                             2    1
##   ResTariffallocation ResStimulusallocation
## 1                   E                     E
## 2                   A                     4
## 3                   A                     2
## 4                   D                     4
## 5                   D                     4
## 6                   C                     3
```

Here `ID` is the customer identifier, `meanDem` is the demand of each customer, while the definition of other variables can be found at the link provided above. The entry `Irish$extra` is a `data.frame` containing meteorological and calendar information:

```
head(Irish$extra)
```

```
##   time       toy dow  holy tod temp            dateTime
## 1    1 0.9863014 Wed FALSE   0    4 2009-12-30 00:00:00
## 2    2 0.9863014 Wed FALSE   1    4 2009-12-30 00:30:00
## 3    3 0.9863014 Wed FALSE   2    4 2009-12-30 01:00:00
## 4    4 0.9863014 Wed FALSE   3    4 2009-12-30 01:30:00
## 5    5 0.9863014 Wed FALSE   4    4 2009-12-30 02:00:00
## 6    6 0.9863014 Wed FALSE   5    4 2009-12-30 02:30:00
```

In particular:

- `time` is aprogressive time counter.
- `toy` is the time of year from 0 (Jan 1) to 1 (Dec 31).
- `dow` is a factor variable indicating the day of the week.
- `holy` is a binary variable indicating holidays.
- `tod` is the time of day, ranging from 0 to 47, where 0 indicates the period from 00:00 to 00:30, 1 the period from 00:30 to 01:00, and so on.
- `temp` is the external temperature in degrees Celsius.

The `electBook` data frame also contains the `IrishAgg` data frame, where the individual consumption trajectories have to be aggregated to produce a single trajectory:

```
data(IrishAgg)
head(IrishAgg)
```

```
##   time       toy dow  holy tod temp            dateTime      dem
## 1    1 0.9863014 Wed FALSE   0    4 2009-12-30 00:00:00 1674.398
## 2    2 0.9863014 Wed FALSE   1    4 2009-12-30 00:30:00 1404.605
## 3    3 0.9863014 Wed FALSE   2    4 2009-12-30 01:00:00 1180.766
## 4    4 0.9863014 Wed FALSE   3    4 2009-12-30 01:30:00 1022.626
## 5    5 0.9863014 Wed FALSE   4    4 2009-12-30 02:00:00  877.018
## 6    6 0.9863014 Wed FALSE   5    4 2009-12-30 02:30:00  775.936
##   dem48 temp95
## 1    NA      4
## 2    NA      4
## 3    NA      4
## 4    NA      4
## 5    NA      4
## 6    NA      4
```

Here the new variables are:

- `dem48` is the demand in the same half-hourly period of the previous day.
- `temp95` is an exponential smooth of `temp`, that is, `temp95[i] = a*temp[i] + (1-a)*temp95[i-1]` with `a = 0.05`.

1.2.4 French National Demand Data

1.2.4.1 Data Presentation

This data set contains aggregate French electricity demand data, at an half-hourly resolution and covering the period from the 1st of January 2012 to the 18th January 2021. The raw demand data was obtained from https://www.rte-france.com/fr/eco2mix/eco2mix. At the time of writing, the demand data corresponding to 2020 is provisional.

Figure 1.3 shows some of the characteristics of the data. Plot 1.3a shows that the long-term demand trend is negative. This is because the demand contained in the data set is net of embedded production from, e.g., solar panels and wind turbines. That is, net French demand is decreasing because embedded generation is increasing faster than gross demand. The curves in plot 1.3b show the daily demand profiles for

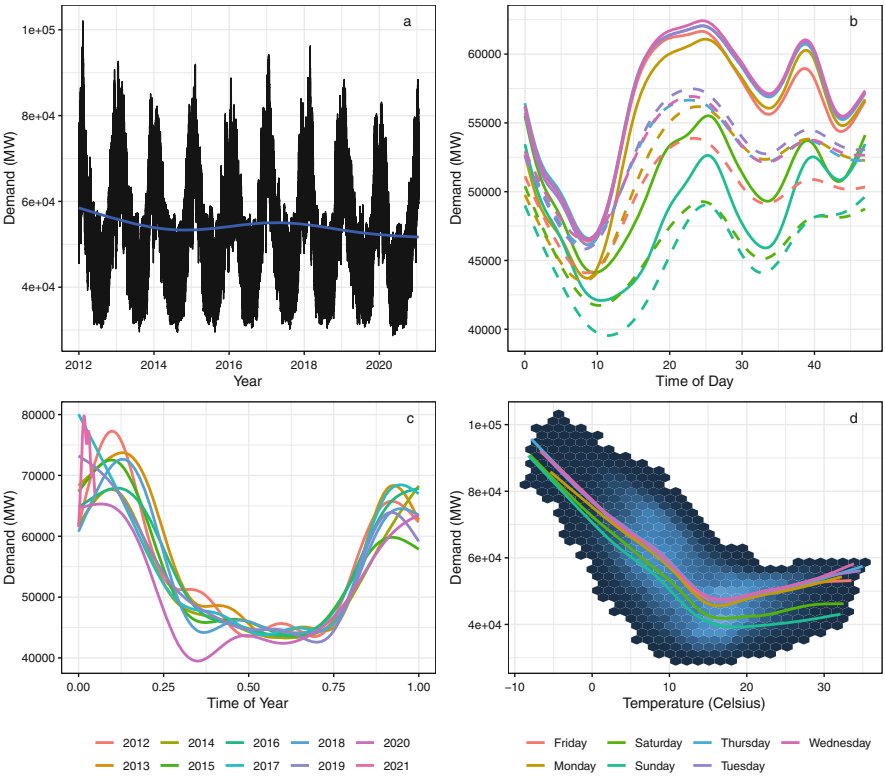

Fig. 1.3 The plots show: (**a**) demand vs. time and the long-term trend; (**b**) the daily demand profiles for each day of the week, the dashed lines are the profiles corresponding to the two national lockdowns; (**c**) seasonal demand dynamics for each year; and (**d**) temperature effect for each day of the week. See the main text for more details

each day of the week. The consumption is lower during the night than during the day, starting increasing around instant 11 (5h30 a.m.) during week days and instant 15 (7h30 a.m.) on Sunday. Two peaks occur around 12h and 20h30 at lunch and dinner times. Another peak at the end of the day is attributable to the water heaters working during the off-peak tariff period. Unsurprisingly, demand is lower on weekends with a less pronounced morning ramp-up, as people wake up later. The dashed lines are the daily profiles during the two national lockdowns that took place in 2020 (17th March to 11th of May and 30th of October to 11th of December). As expected the demand is lower during these periods, especially during peak hours. Plot 1.3d shows how demand varies with temperatures, for each day of the week. The shape of the temperature effect is similar across the week days, the cooling effect ($t > 17\text{C}^\circ$) appearing to be stronger on weekend than during the week. Figure 1.3c shows that demand is higher during the winter than during the summer (time of year is 0 on January 1 and 1 on December 31). Between-years discrepancies in yearly demand profiles are substantial, especially in the winter. The drop in demand corresponding to the first national lockdown is clearly visible.

1.2.4.2 Data Processing

We integrate the demand data with calendar information such as weekdays, time of day, bank holidays, and time of year. We also add temperature data from Metéo France (https://donneespubliques.meteofrance.fr/). The temperature data was measured in the proximity of several large French cities. We built a single temperature variable by averaging these temperatures with weights that are proportional to the population of each city.

1.2.4.3 Getting the Data

This data set is available on the French National Grid website (https://www.rte-france.com/fr/eco2mix/eco2mix). While we did not include it in the `electBook` `R` package due to the lack of a redistribution license, the `R` code we used to process the data is available at https://github.com/cugliari/ML4ELF/tree/master/website.

1.2.5 US Regional Demand Data from the GEFCOM 2014 Competition

1.2.5.1 Data Presentation

The Global Energy Forecasting Competition 2014 (GEFCom2014) was a probabilistic energy forecasting competition comprising of four tracks: load, price, wind, and solar production forecasting. Here we describe data from the load forecasting

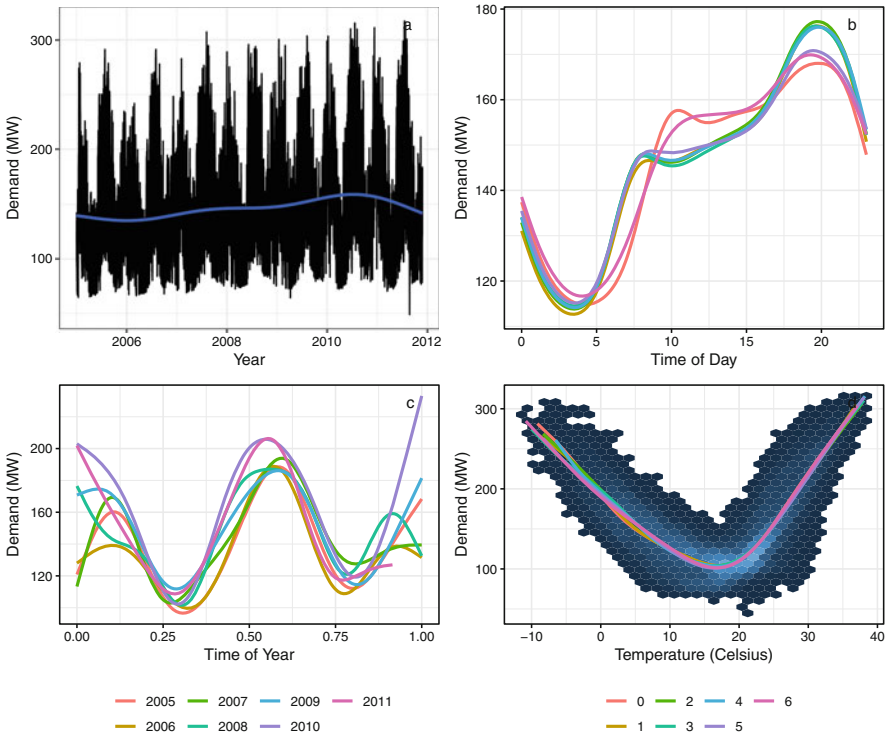

Fig. 1.4 The plots show: (**a**) demand vs. time and the long-term trend; (**b**) the daily demand profiles for each day of the week; (**c**) seasonal demand dynamic for each year; and (**d**) temperature effect for each day of the week. See the main text for more details

track, where the aim was to forecast several conditional quantiles of the hourly loads for a (undisclosed) US utility. The organizers provided hourly historical load and weather data, and the participants were allowed to use calendar information on public US federal holidays. The data discussed here covers the period from the 1st of January 2005 to the 1st of December 2011.

Figure 1.4a shows that the long-term demand trend is positive up to around the beginning of 2011 and then turns negative. Plot 1.4b shows the daily demand profiles for each day of the week. As for the Irish data, the morning peak is delayed on weekends, but it reaches higher values, while the evening peak is flatter than that on working days. Plot 1.4d shows that the relation between temperature and demand has a similar shape across different days of the week and that the heating effect is as strong as the cooling effect. Figure 1.4c shows the presence of a winter and a summer demand peak. These plots suggest that the data comes from a residential area, in a region characterized by cold winters and hot summers.