Short-term forecasting of electricity demand of smart homes and distribution grids

Alexis Gerossier

PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à MINES ParisTech

Prévision à court terme de la demande électrique des maisons intelligentes et des réseaux de distribution

Short-Term Forecasting of Electricity Demand of Smart Homes and Distribution Grids

Soutenue par

Alexis GÉROSSIER

Le 23 mai 2019

Ecole doctorale n° 621

ISMME – Ingénierie des Systèmes, Matériaux, Mécanique, Énergétique

Spécialité

Énergétique et procédés

Composition du jury :

Jean-Michel POGGI Professeur, Université Paris-Sud	*Président*
Jovica MILANOVIC Professeur, University of Manchester	*Rapporteur*
Zita VALE Professeure, Instituto Politécnico do Porto	*Rapportrice*
Tomasz ZABKOWSKI Docteur, Warsaw University of Life Sciences	*Examinateur*
Yannig GOUDE Docteur, EDF R&D	*Examinateur*
George KARINIOTAKIS Directeur de recherche, MINES ParisTech	*Directeur de thèse*
Robin GIRARD Chargé de recherche, MINES ParisTech	*Maître de thèse*

MINES ParisTech | PSL

Acknowledgment

Since research is hardly possible without collaboration and discussion stimulation ideas, I would like to thank the following people without whom this thesis would have not be possible.

First and foremost, I would like to thank Armines and MINES Paristech who financed this work, and the research lab of PERSEE where top research is carried out in the field of smart grids. I praise my two supervisors George Kariniotakis and Robin Girard, who welcome me in the lab, for their guidance throughout my PhD journey, whether is on the global orientation of the research, the stirring discussion of ideas, or the technical advice. Under their direction, I had the chance to discover a gargantuan amount of research papers, and to meet many partners tackling the challenges of the energy world during several international conferences, workshops, and through the collaboration of the European H2020 project SENSIBLE. These partners range from academics, such as the people from DTU, the University of Nottingham, INESC, and the University of Manchester, to industry, such as Enedis, Siemens, EDF, EDP and Engie. They all contributed to this work, by providing precious use cases and data, along with insightful approaches.

This research would have not be possible without the wholesome environment I found in the MINES centre in Sophia Antipolis, and the people there who make their best in order to improve the world and make the science progress. There would be too many people to cite them individually. Therefore I will just point out the people I work closely, in no particular order: Evgeniya Ishkina and Alexis Bocquet, managing to transform my messy algorithms into practical software; Andrea Michiorri, capable of the excruciating but essential planning of scientific projects; Robin Girard, for debating mathematical and philosophical issues, George Kariniotakis; always giving relevant direction while multitasking; Thibaut Barbier, leading me into the mysterious paths

of electricity. On a more personal note, I would like to thank the fellow PhD students who shared an office with me at some point, all making it a pleasant place: Maxime, Guillaume, Papa, Fabien, Romain, Di, Charlotte, Pedro, and Kévin.

Part of the work in this thesis was carried out in the frame of the research and innovation project SENSIBLE (Storage ENabled SustaInable energy for BuiLdings and communitiEs — www.projectsensible.eu, grant agreement No. 645963) supported by the European Union under the Horizon 2020 Framework Programme.

This short acknowledgment merely reflects the debt I owe for the realization of this thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context: Evolution of the Electrical Network

Historically, the current electrical network was developed in France and Europe at the end of the Second World War. Large infrastructure works were built by operators, e.g. transformers and lines, in order to connect power plants to consumers, first through a transmission network (high voltage) then through a distribution network (medium and low voltage). Thanks to these networks, operators deliver electricity from the production sites to the consumption sites. At first, only a small number of operators existed, sometimes as nationally-subsidized monopolies, such as Électricité de France (EDF) in France. Because of this monopolistic (or oligopolistic) situation, the production is generally highly centralized with large power plants[1] fueled by various types of energy: water, gas, coal, nuclear etc. This centralization paradigm has been recently disrupted by the European Union which encourages the *energy market liberalization.* The operators in charge of the transmission and the distribution are now clearly separated into transmission system operators (TSO) and distribution system operators (DSO), e.g. in France, EDF has been divided into RTE and Enedis. Consequently, many smaller operators are emerging, as well as new roles: producers, retailers, aggregators etc. All of this require a local management of the network, for instance to localize where the network losses occur (Barbier, 2017). Figure 1.1 illustrates this electrical grid evolution (International Energy Agency, 2011).

[1] the very term for power plants in the romance language, e.g. *centrale électrique* in French, embeds the idea of centralization.

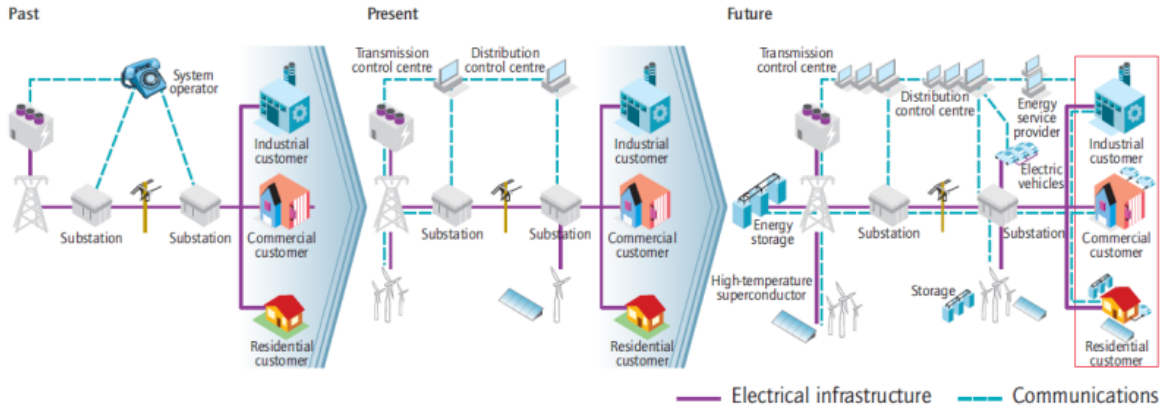


Figure 1.1 – Past, present, and future of the grid. Source: (International Energy Agency, 2011).

The liberalization means that the number of operators increases, and since they need a market to exchange, international marketplaces are created in the 2000s, such as EPEX Spot and Nord Pool in Europe. Such places simplify the trading of electricity production, in coming from smaller power plants, notably the ones fueled by *renewable energies* (biomass, wind, solar, and hydraulic). The development of renewables is stimulated by governments in order to decrease they carbon intensity. For instance, the European Union has set up a goal of 27% share of renewable energy to be reached by 2030. The carbon intensity reduction depends on the current impact of electricity production in greenhouse gas emissions. For instance, the production in Poland is currently highly polluting, and this pollution can be significantly reduced thanks to greener renewable energies. The reduction extent is less in nuclear countries whose carbon impact is relatively low, e.g. France. In any case, for other reasons, France reaches a 22.8% green generation in 2018 (Réseau de Transport d'Électricité (RTE), 2018b) and is considering a 100% green energy generation by 2050 (Krakowski et al., 2016). Similarly, the Danish island of Samsø anticipates 100% generation from renewables by 2030 (Mathiesen et al., 2015). All this cause *a decentralization of the production* which is challenging to integrate to the network for several reasons: e.g. the need of an advanced communication system, the more complex infrastructure, the harmonics produced by multiple production sites, etc. These technical issues add up with the inevitable intermittence caused by some energy sources (wind, solar). This intermittence can be diminished by aggregating several small units in a virtual

power plant (VPP), which is illustrated in Figure 1.2 (Statkraft, 2017). However, the

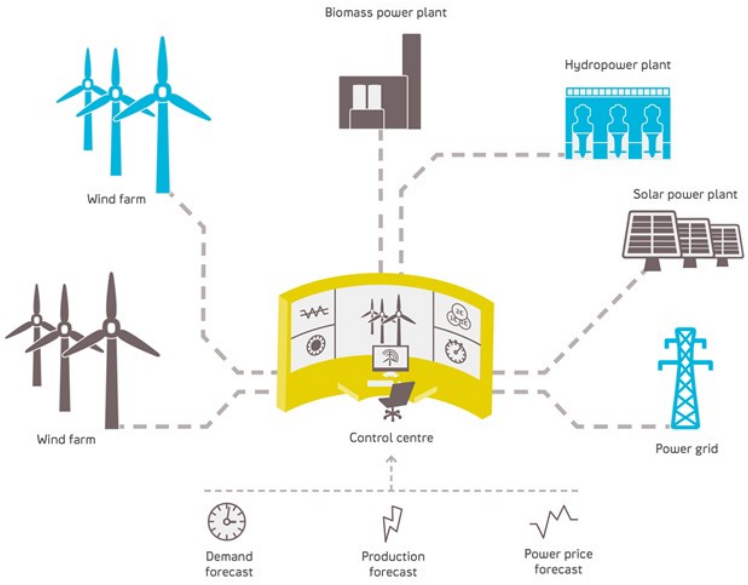


Figure 1.2 – Diagram of virtual power plant. Source: Statkraft

designing and management of an optimal VPP necessitate a precise local management of the grid.

The modernization of the grid comes with its digitalization, requiring that an abundant monitoring should be undertaken, with an exhaustive *smart-meter roll-out*. Smart meters measure and transmit the electricity consumption made by individuals during a given period — e.g. for every one-hour period. A large deployment therefore gives a precise vision of the whole grid. The European Union acted that, in 2020, countries with positive cost-benefit analysis should have a smart-meter roll-out for 80% of households (European Parliament, 2009). In most cases, the costly investments for roll-outs are compensated by precisely quantifying the exact impact of energy programs and appliance standards (Armel et al., 2013) and by improving the customer's billings (Cour des Comptes, 2018). The consumers should also benefit from the smart-meter feedback. By providing more frequent broken down information about their consumption (hourly breakdown every day instead of monthly total demand), householders can adapt their behavior in real time to reduce their consumption or save money (Nordic Energy Regulators, 2014).

1.2 Motivation: Usage of the Forecasts

The recordings of the individual hourly electricity demand values with smart-meters make it possible to analyze and forecast them for short horizons. Accurate forecasts have no value on their own, but gain it when they are used as inputs for further applications. We give some recent examples of the typical applications exploiting such inputs.

Grover-Silva et al. use forecasting scenarios of the electricity load to optimize the day-ahead scheduling of a microgrid (Grover-Silva, Heleno, et al., 2018). A Home Energy Management System (HEMS) combines several information sources (demand and PV forecasts, battery aging, market price and so on) to optimize the electricity usage in a house, and then a neighborhood, as shown in Figure 1.3 (Correa-Florez et al., 2018). Advanced price schemes are proposed, dwelling on individual demand

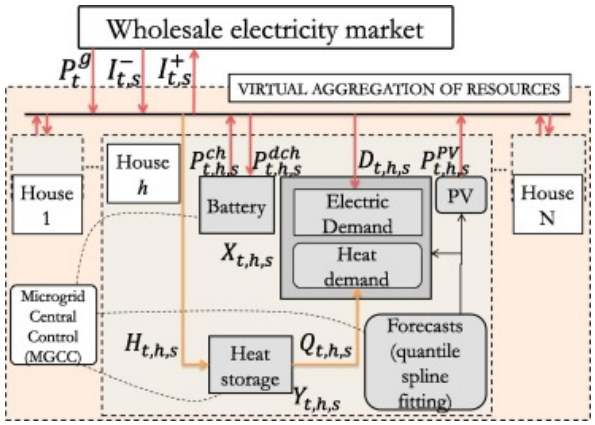


Figure 1.3 – Diagram of the Home Energy Management System. Source: (Correa-Florez et al., 2018)

forecasts, to change the electricity consumption patterns of individuals, i.e. demand response purposes (Le Ray et al., 2018). This goes along with the rising practice of self-consumption ,i.e. strategies to use the most of the local production to less depend on the global grid, and it requires precise load forecasts (Luthander et al., 2015).

These applications are already implemented in the real world, for instance the European project SENSIBLE makes use of the individual demand forecasts to participate in a flexibility market, via a complex ICT infrastructure as shown in Figure 1.4 (SENSIBLE, 2018).

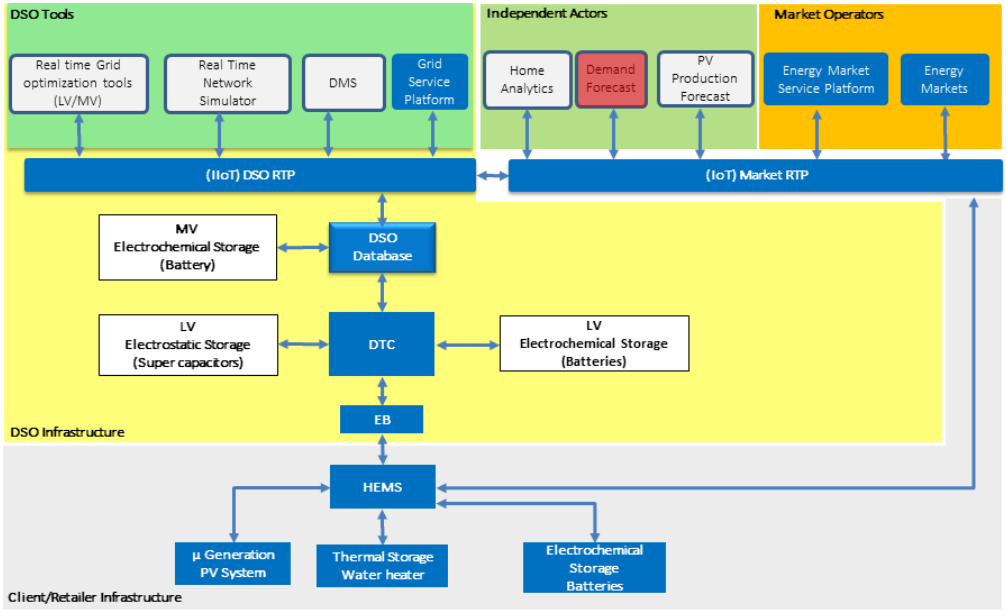


Figure 1.4 – ICT diagram of the SENSIBLE demonstration project in Évora. The demand forecast part is highlighted in red. Source: (SENSIBLE, 2018)

1.3 Scope of the Thesis

This thesis deals with the task of forecasting the electricity demand, i.e. the electricity that users consume in order to run their appliances. Most of the time, this demand is improperly expressed as a power, in watts (W). However, demand rather refers to an energy, e.g. in watt hours (Wh), corresponding to the electric power averaged over a given period. The typical period of interest is one hour. Therefore, our goal is to anticipate the future energy needs during a one-hour period.

With the electricity-related transformations aforementioned, there is a raising interest in the local scale. The scale refers to the size of the geographical area considered, but we conveniently define it according to the mean power of the demand series at hand. In this document, the local scale is defined as ranging from a single appliance (100 W) up to a feeder (1 MW), with a focus on the household demand (1 kW).

The very term of local scale conveys the idea of a short timescale: most local applications (e.g. demand response, battery scheduling, etc.) are designed for short horizons, typically for the next day. Accordingly, our work is devoted to short-term

forecasting. Short-term horizons are not precisely defined in the community, and authors use slightly different terminology: short term is surrounded by very short and medium terms, with forecasting horizons roughly ranging from one hour to one week.

1.4 Challenges

To exhibit the challenges of the forecasting task, Figure 1.5 represents the hourly electricity demand of a US household during one day. The 24 successive points are connected: the resulting demand curve is non-smooth, and highly erratic. This household demand sometimes increases threefold between two successive hours. An inspection on a longer time frame shows no clear pattern, and the shape of the successive daily curves are completely different from one day to another. Moreover, contrary to the demand at large scale, such as the national demand, external factors have a weak impact on the local demand. For instance, while the correlation between outside temperature and national demand is significant (e.g. the temperature determines the value of the peak demand), the influence is subtler for one household. These preliminary observations need to be corroborated by an in-depth comparison between the demand characteristics at different geographical scales.

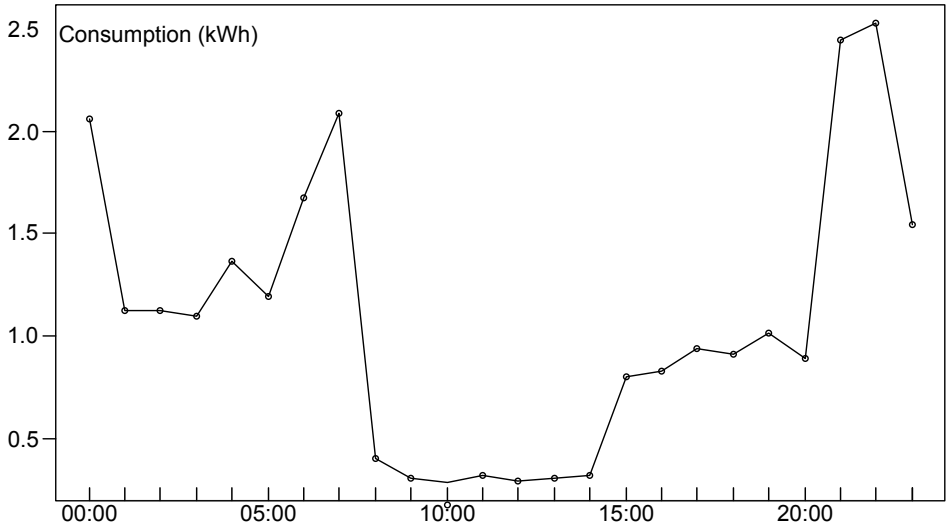


Figure 1.5 – Hourly electricity demand made by a US household during the 24 hours of one day.

The existing forecasting models for the national demand make use of the obvious

demand patterns and driving factors. Since both significantly fade at the local scale, new forecasting models are to be designed for our task. Furthermore, the large-scale models usually produce deterministic forecasts, such as predicting that the demand tomorrow between 9 and 10 will be 4 GW. State-of-the-art models achieve high accuracy with relative errors around 2%, e.g. the actual observation is 4.1 rather than 4 GW. Due to the erratic aspect of the household demand, coming with the almost-random activities of the residents, the same forecasting accuracy is unreachable at this scale. The forecaster should therefore quantify the uncertainty of the prediction, e.g. by stating that the household demand between 9 and 10 will most probably be between 100 and 600 W.

The pending applications require the implementation of the local forecasting model in a real environment, e.g. with roll-outs for hundred to thousands of households at once. This means that the model must be replicable, an adjective that encompasses several features. First, the forecasting model must work under low to no maintenance: no direct intervention is desirable for obvious cost reasons. The model must be adaptable to multiple situations: the electricity demand dynamics completely changes between two households. It means that the models' parameters must be carefully tuned case by case for optimal performance, during scheduled training periods. Finally, a real-life implementation means that the forecasts must be produced online and at all times, i.e. with model that is robust to missing or absurd input data. All of these features necessarily degrade the forecasting performance (compared to preliminary tests in laboratory) and a challenge is to limit this degradation.

The forecasts have no practical value on their own, but gain it by being used in later profitable applications. Those often work on a daily time frame, and thus require forecasts at multiple instants, e.g. through daily demand scenarios. This slightly differs from the common forecasting methods, that lead to complex multidimensional probabilistic forecasts. The scenarios should indeed encompass the multitemporal coherence of the demand values while dealing with the inevitable forecasting uncertainty inherent to the local demand.

1.5 Objectives

From the challenges identified, we highlight four scientific objectives addressed in this work:

1. *Characterization of the electricity demand at the local scale.* What are its specific features? To what extent does its dynamics differ from the demand at larger scale? How well is it driven by exogenous factors?

2. *Development of probabilistic forecasting models.* How to design efficient models? What does it mean to be efficient in a probabilistic framework? What level of performance can the forecaster expect?

3. *Ensuring the replicability of the models.* What features are required for a real implementation? How to deal with missing data? How to reconcile robustness and accuracy?

4. *Generation of daily forecasting scenarios.* How to turn probabilistic forecasts into efficient scenarios at the local scale? How to ensure daily coherence of the forecasting scenarios? Can users' habits produce realistic and accurate demand scenarios?

1.6 Document Outline

The present document is made of this introductory chapter (labeled 1), four chapters (2 to 5), and a conclusion chapter (6). Short abstracts, in English and in French, are provided at the beginning of each chapter (2 to 5). Five appendices (A to E) are attached after the bibliography to provide various mathematical details and supplementary results. An extended chapter-by-chapter French summary, labeled *Résumé*, is included at the end of the document.

Chapter 2 is a general presentation of the forecasting models in the context of electricity demand: we present the general framework of statistical models and introduce the common ones; the evaluation process to assess the forecasting performance is then discussed; and, finally, the literature on electricity demand forecasting is reviewed.

Chapter 3 deals with the electricity demand at the feeder level. An original decomposition algorithm that recovers elementary demand profiles of customer categories is

introduced. The obtained profiles identify the demand patterns throughout the day of each category, allowing to do medium-term forecasting and prospective analysis, such as the future evolution of the peak demand timing and value. However, very specific features are necessary to improve forecasting performance, requiring the study of the demand at the household scale.

This household demand is explored in Chapter 4. The data obtained with smart meters in three worldwide areas is thoroughly analyzed to point out the forecasting challenges. The specificity of this demand requires the designing of new models of probabilistic nature that forecast household demand for the next day. We introduce a reference model, namely a gradient boosting model, designed with state-of-the-art techniques. Its forecasting performance is assessed and compared at different scales: from one household to one feeder; and at different time resolution: from one minute to one week. In spite of its top performance, this reference model cannot be used in a real context. A hierarchical forecasting framework combining several robust models is developed and analyzed. This framework has been implemented on a demonstration site and operated in real time. The project feedback highlights some key points for practical applications.

In Chapter 5, we dwell on using scenarios to produce probabilistic forecasts of the future demand at multiple horizons. The issue of generating these household demand scenarios (scenarios generation), and obtaining a small set of representatives scenarios (scenarios reductions) is discussed in length. The forecasting scale is then pushed one step further with the analysis of the electricity demand of a single domestic appliance, namely the electric vehicle (EV). A day-ahead forecasting model is designed specifically for this appliance demand, building on a deep analysis of the habits of the users regarding the charging of their EV's battery.

1.7 Communication related to the thesis

During the work, four papers have been published in peer-reviewed journal (three as main author):

- Gerossier, A., Barbier, T., and Girard, R. (2017). A novel method for decomposing electricity feeder load into elementary profiles from customer informa-

tion. *Applied Energy, 203, 752-760*; https://doi.org/10.1016/j.apenergy.2017.06.096

- Gerossier, A., Girard, R., Bocquet, A. and Kariniotakis, G. (2018). Robust day-ahead forecasting of household electricity demand and operational challenges, *Energies, 11(12), 3503*; https://doi.org/10.3390/en11123503

- Gerossier, A., Girard, R., and Kariniotakis, G. Modeling and forecasting electric vehicle consumption profiles. *Energies, 2019, vol. 12, no 7, p. 1341*; https://doi.org/10.3390/en12071341

- Correa-Florez, C. A., Gerossier, A., Michiorri, A., and Kariniotakis, G. (2018). Stochastic operation of home energy management systems including battery cycling. *Applied Energy, 225, 1205-1218*; https://doi.org/10.1016/j.apenergy.2018.04.130

Furthermore, I presented early versions of these works during international conferences:

- Gerossier, A., Girard, R., Kariniotakis, G., and Michiorri, A. (2017). Probabilistic day-ahead forecasting of household electricity demand. *CIRED-Open Access Proceedings Journal, 2017(1), 2500-2504*; http://doi.org/10.1049/oap-cired.2017.0625

- André, R., Mendes, G., Neto, A., Castro, P., Madureira, A., Sumaili, J., ... and Michiorri, A. (2017). Energy services bridging the gap between residential flexibility and energy markets. *CIRED-Open Access Proceedings Journal, 2017(1), 2726-2730*; https://doi.org/10.1049/oap-cired.2017.0365

- Correa-Florez, C. A., Gerossier, A., Michiorri, A., Girard, R., and Kariniotakis, G. (2017). Residential electrical and thermal storage optimisation in a market environment. *CIRED-Open Access Proceedings Journal, 2017(1), 1967-1970;* https://doi.org/10.1049/oap-cired.2017.1086

- Correa, C. A., Gerossier, A., Michiorri, A., and Kariniotakis, G. (2017, June). Optimal scheduling of storage devices in smart buildings including battery cycling. *In PowerTech, 2017 IEEE Manchester (pp. 1-6). IEEE;* https://doi.org/10.1109/PTC.2017.7981199

- Gerossier, A., Girard, R., and Kariniotakis, G. (2018, November). Modeling electric vehicle consumption profiles for short-term forecasting and long-term simulation. *In MedPower 2018, Dubrovnik.*

- Correa-Florez, C. A., Michiorri, A., Gerossier, A., and Kariniotakis, G. (2018, November). Day-ahead management of smart homes considering uncertainty and grid flexibilities. *In MedPower 2018, Dubrovnik.*

Chapter 2

Statistical Forecasting Models

Summary Statistical models have been designed to forecast future phenomena based on what happened in previous situations. These models ground from the theory of statistics developed within the last two hundred years, and the very recent increase of computing power. Common forecasting models are introduced in Section 2.1. The evaluation of the quality, i.e. performance, of a forecasting model is made with indices comparing numerical outputs of the model with actual measurements of the phenomenon. In Section 2.2, the main performance indices, for deterministic and probabilistic forecasts, are described and analyzed. The models have been quickly adapted to forecast future electricity demand. We review the body of research focused on this task, specifically for short-term horizons and intermediate temporal granularity — typically, the forecasts of hourly demand loads for the next day. In the scope of this thesis, we focus on the research devoted to electricity demand at the local scale, recently enabled by the availability of smart-meters recordings. We compare the forecasting performance reported in the literature at different scales. On average, the relative errors increase from 3% at the national scale to 30% at the household scale.

Résumé À l'aide de la théorie statistique développée au cours des deux derniers siècles et la récente augmentation de la puissance de calcul, des modèles statistiques ont été créés pour prédire un phénomène futur en fonction des phénomènes précédemment observés. Les modèles de prédiction usuels sont présentés dans la Section 2.1. L'évaluation de la qualité, c.-à-d. la performance, des modèles de prédiction est faite à l'aide d'indices comparant les sorties numériques des modèles aux vraies mesures du phénomène. Les principaux indices de performance, pour des prédictions déterministes et probabilistes, sont décrits et analysés dans la Section 2.2. Ces modèles ont rapidement été adaptés pour la prédiction de la demande électrique future. Dans la Section 2.3, nous passons en revue les travaux consacrés à cette prédiction pour des horizons courts sur de moyenne temporalité (généralement, une prédiction faite un jour à l'avance de la demande électrique moyenne mesurée toutes les heures). Nous nous intéressons particulièrement aux travaux portant sur la demande électrique locale, permise par les fines mesures réalisées avec des compteurs intelligents. Nous fournissons un aperçu de la performance de prédiction rapportée dans la littérature ainsi que son évolution en fonction de l'échelle considérée. En moyenne, l'erreur relative passe de 3% à l'échelle nationale à 30% à l'échelle d'un ménage.

Forecasting a phenomenon is closely related to modeling a phenomenon. In the traditional deterministic philosophy as thought by Laplace, when one understands all the underlying processes of a phenomenon, one is able to perfectly predict its future (Laplace, 1829). Of course, finding all the underlying processes is a harsh — if not impossible — task, and forecasters relied on simplifying hypotheses to perform sufficiently accurate forecasts. This conception has been proven false by quantum theory and the famous "no hidden variables" by Von Neumann (Bub, 2010): quantum mechanics is not deterministic. In additional to the non-deterministic aspect of nature, its chaotic aspect presents another difficulty. For instance, the weather forecast is paramount from an economical viewpoint (Regnier, 2008). Therefore, scientific community devoted a lot of work into the weather forecast problem, but came to understand that the meteorological phenomena are chaotic: a small difference in the initial state cause a large difference in the outputs. Accurate forecasts are then tedious to obtain.

In the following, we start by presenting the basic framework to predict a phenomenon: the equations and the structure of common models (Section 2.1). In Section 2.2, usual methods to assess forecasts quality are presented. Finally, a review of the different forecast models for electricity load forecasts in the literature is drawn in Section 2.3.

2.1 Overview of the Different Forecasting Models

2.1.1 Basic Framework

At an instant t, we denote by i_t the state of the world, or everything that is known up to instant t (past and present). To forecast the future is to predict the state of the world i_{t+h} at instant $t+h$, where $h > 0$ denotes the time horizon.

In most cases, we focus on a single phenomenon, e.g. electricity demand, and not the whole state of the world. With mathematical approaches, it is convenient to express a single phenomenon with a real value (economical cost in €, electricity demand in kWh, event outcome by 0 or 1 and so on). Therefore, a phenomenon can be described by $y_t \in \mathbb{R}$ at instant $t \in \mathbb{R}$. Since measuring y_t is made at a finite number of instants, often regularly, the successive values form a real-valued discrete time series $(y_t)_{t\in\mathbb{N}}$.

At a given instant t, a future value of a phenomenon y_{t+h} depends on the known

state of the world i_t up to instant t. When y_{t+h} is considered to be the realization of a random variable $Y_{t+h|t}$, one uses the probability that the random value is below y knowing i_t

$$F_{t+h|t}(y) = \mathbb{P}\left[Y_{t+h|t} \leq y | i_t\right] \tag{2.1}$$

to define function $F_{t+h|t}$. From this definition, we see that (1) $F_{t+h|t}(\cdot)$ is a right-continuous and non-decreasing function, (2) $F_{t+h|t}(y) \to_{-\infty} 0$ and (3) $F_{t+h|t}(y) \to_{+\infty} 1$. It defines the cumulative distribution function (CDF) of the random variable $Y_{t+h|t}$. One obtains the probability density function (PDF) by differentiating the CDF, i.e.

$$f_{t+h|t}(y) = F'_{t+h|t}(y) \qquad \forall y \in \mathbb{R}. \tag{2.2}$$

As a density function, $f_{t+h|t}$ is non-negative and its integral sums up to 1.

It should be noted that the random variable $Y_{t+h|t}$ depends on the horizon h. In general, the larger is the horizon, the more uncertain is the future value and the more spread is the corresponding PDF. It means that forecasting errors amplify with time. We show this amplification with a simple random walk process in the Example 1.

Example 1 We define a random walk process $\{y_t\}$: for every t, realization y_{t+1} is drawn from the random variable $Y_{t+1|t}$ which follows a Gaussian distribution centered on y_t with unit variance, i.e. $Y_{t+1|t} \sim \mathcal{N}(y_t, 1)$. For larger horizon, we have $Y_{t+h|t} \sim \mathcal{N}(y_t, \sqrt{h})$. The spread of the density function hence increases with the square root of h. Figure 2.1 exhibits three predictive densities of the random walk for horizon of 5, 3, and 1. The actual realization y_{t+5}, that is to be forecast, is shown with a vertical gray line. This realization falls in high-density zone for all three predictive density functions, but the spread of the density is narrower for the most recent density f_1.

In real applications, cumulative and density functions cannot be observed. Exactly obtaining these functions requires an infinite number of realizations of $Y_{t+h|t}$ with the exact same state i_t. This is never the case — worse, there is usually only one observation —, and the forecaster tries to approximate the state i_t by an input subset $s_{t+h|t}$ for generalization purposes[1]. This subset should be suited to the phenomenon studied and the horizon h. Its selection rely on the forecaster's expertise who has to balance a

[1]From here on, we assume that the forecast is carried out at instant t and, as such, we drop the $|t$ to simplify notations

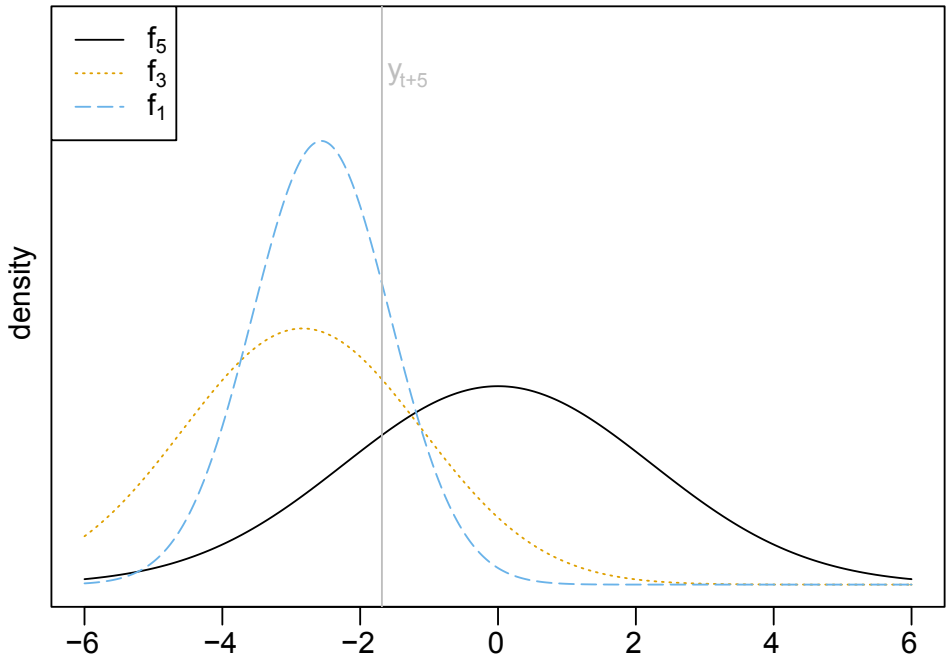


Figure 2.1 – For the random walk process, density function predicted for different horizons: black solid line for horizon of 5, orange dotted line for horizon of 3, blue dashed line for horizon 1. The actual realization y_{t+5} is represented by a gray vertical line.

precise description of the current situation and the generalization capacity. The usage of this subset leads to the estimation of the distributions, noted with a hat, i.e. $\hat{F}_{t+h}(\cdot)$ and $\hat{f}_{t+h}(\cdot)$.

There are two main categories of forecasts: *point* and *probabilistic* forecasts. Point forecasts historically appeared before probabilistic ones. The historian Stigler situates the transition from point to probabilistic forecasts in the nineteenth century when assessing uncertainty become of utmost importance (Stigler, 1986). However, the two types still coexist in the forecasting literature.

Point Forecasts Forecasting a single value for a future event is the more natural approach. It is easy to understand and is sufficient for many applications. It is sometimes conveniently referred to as *deterministic* forecast as opposed to probabilistic[2].

[2]Forecasting a single quantile value is, in fact, a probabilistic point forecast. We choose to classify it in the point forecasts category.

The following point forecasts are the most used ones:

- The *expected mean value* $\hat{y}_{t+h}$ is the estimated mean of the random variable Y_{t+h}. It can be computed from the estimated density function so that $\hat{y}_{t+h} = \int_{\mathbb{R}} y\hat{f}_{t+h}(y)\mathrm{d}y$.
- The *median value* $\hat{y}^{0.5}_{t+h}$ is the estimated median of the random variable Y_{t+h}. It can be computed from the estimated cumulative function so that $\hat{F}_{t+h}(\hat{y}^{0.5}_{t+h}) = 0.5$.
- The *quantile value* $\hat{y}^{\tau}_{t+h}$ for $\tau \in [0, 1]$ gives the estimated threshold value such that the probability of obtaining an actual realization is below quantile level τ, i.e. $\mathbb{P}[y_{t+h} \leq \hat{y}^{\tau}_{t+h}] = \tau$. It can be computed from the estimated cumulative function so that $\hat{F}_{t+h}(\hat{y}^{\tau}_{t+h}) = \tau$.

Probabilistic Forecasts Forecasting a probabilistic distribution is useful to assess the uncertainty of the future event. It gives the confidence we have in the future: a spread distribution shows large uncertainty, and conversely a narrow one indicates a strong confidence. The estimated density (or cumulative) function uniquely and completely characterizes the forecasts and any point forecasts can be deduced from them. Since this estimation is often hard to do, probabilistic forecasters may use the less informative following forecasts:

- A *Monte Carlo sample* $\{\hat{y}^{(1)}_{t+h}, ..., \hat{y}^{(n)}_{t+h}\}$ of size n. Realizations are drawn from the random variable Y_{t+h}. This drawing can be more convenient to do than giving the density function.
- A *list of quantiles* $\{\hat{y}^{\tau_1}_{t+h}, ..., \hat{y}^{\tau_k}_{t+h}\}$ of size k for $0 \leq \tau_1 < ... < \tau_k \leq 1$ which can be deduced from a Monte Carlo sample or from the CDF.
- A *prediction interval* $[\hat{y}^{a}_{t+h}, \hat{y}^{b}_{t+h}]$ for $0 \leq a < b \leq 1$ which assesses the chance that the actual value falls inside this interval, i.e. $\mathbb{P}[\hat{y}^{a}_{t+h} \leq y_{t+h} \leq \hat{y}^{b}_{t+h}] = b - a$. The interval is usually, but not necessarily, centered ($a + b = 1$).

Some forecasting models focus on point forecasts while others are designed to obtain probabilistic forecasts. Generally, obtaining probabilistic forecasts more challenging and more compute-intensive. The current growth of literature on probabilistic forecasts is favored by the increase in computing performance and development of new estimation methods (Gneiting & Katzfuss, 2014).

2.1.2 Common Forecasting Models

We present some forecasting models that are prevalent in the energy community. The order of presentation chosen is based on the ease-of-interpretation of the models: from understandable ("white-box models") to more abstruse models ("black-box models"). We first present some basic facts about the models training. Then the theoretical foundation of each model is quickly drawn, along with the advantages and drawbacks of the methods. The interested reader should refer to reference handbook, such as *The Elements of Statistical Learning* (Friedman et al., 2001).

2.1.2.1 Model Training

A model forecasting for an event at instant $t+h$ uses information set s_{t+h} about the past (up to instant t) as inputs to forecast an event y_{t+h}. A model combines this information set to carry out an output. This combination is made with a parameter set β. The values of the elements in β depend on the event to forecast and the horizon h. The estimation of the parameters is made by comparing the values obtained with the model $g^h_\beta(s_{t+h}) = \hat{y}_{t+h}$. the parameters β depend on the horizon h and are optimized by comparing the outputs with actual values y_{t+h}. This comparison is made with a loss function $\mathcal{L}(\cdot,\cdot)$ on a *training set* $\{1,\dots,T\}$ when the actual values y_{t+h} are known, and leads to the minimization problem

$$\beta^* = \underset{\beta}{\operatorname{argmin}} \sum_{t=1}^{T} \mathcal{L}\left(g^h_\beta(s_{t+h}), y_{t+h}\right). \tag{2.3}$$

There exist multiple loss functions resulting in different kind of outputs: for instance, in 1 dimension, using the quadratic loss, i.e. $\mathcal{L}(x,y) = (x-y)^2$, retrieves the *expected mean value* of the event, while the absolute loss, i.e. $\mathcal{L}(x,y) = |x-y|$, retrieves the *median value* of the event. In general, the loss function $\mathcal{L}(x,y)$ is minimal when $x = y$. Consequently, when the set β is large, there is a solution such that $g^h_\beta(s_{t+h}) = y_{t+h}$, for all $t = 1,\dots,T$. This is not desirable since such models often lead to poor forecasting performance: they are *overfitted* to the training set and poorly generalize to other data. The size of the parameter set, noted $|\beta|$, and hence the complexity of the model, should be kept small, $|\beta| \ll T$. Figure 2.2 shows the typical evolution of errors when one increases the complexity of the model. For low complexity, the errors on the training set are high and comparable to the errors on a

test set (observations not used when minimizing Equation (2.3), see (Tashman, 2000)). As the model complexity increases, the errors on the training set continuously decrease to 0, but the errors on the test set attain a minimum before increasing. The position of this minimum indicates the optimal complexity of the model. Finding this optimum is delicate and relies on the forecaster's decisions.

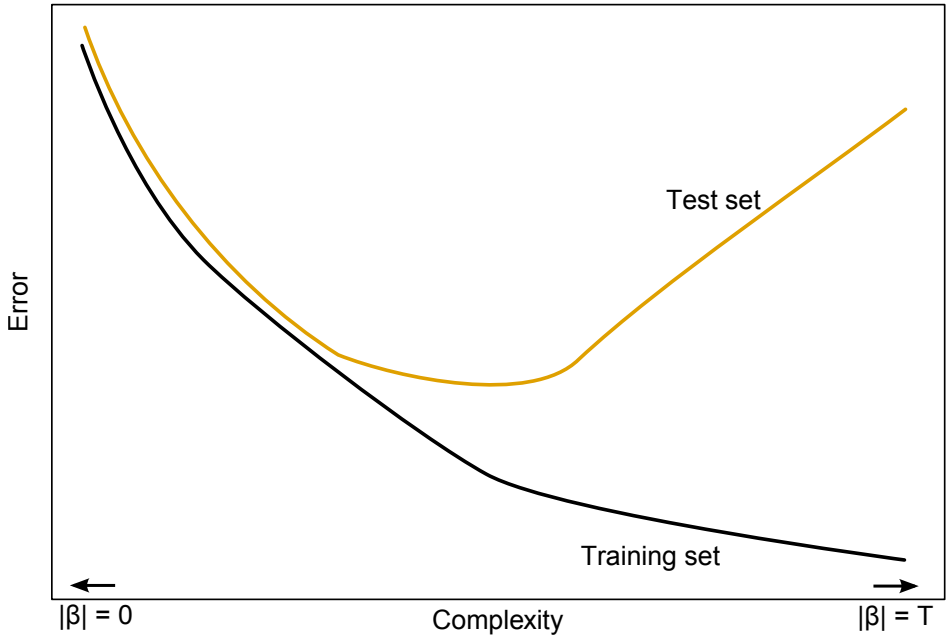


Figure 2.2 – Typical error of a forecasting model for the training set (in black) and the test set (orange) when the complexity of the model increases

2.1.2.2 Linear Model (LM)

Linear regression models were the first methods developed, well before the computer era: it is unclear who was the first discoverer, Legendre or Gauss, but the modern formulation occurs at the turn of the nineteenth century (Plackett, 1972). The calculation is straightforward and the model is easily interpretable.

With such models, the point forecast is supposed to linearly depend on the p inputs $s_t = \{x_t^{(1)}, ..., x_t^{(p)}\}$, according to the parameters $\beta = \{\alpha_0, \ldots, \alpha_p\}$. Hence

$$g_\beta^h(s_{t+h}) = \alpha_0 + \sum_{j=1}^{p} \alpha_j x_{t+h}^{(j)}. \tag{2.4}$$

When the inputs are historical values of the phenomenon, e.g. $x_{t+h}^{(1)} = y_{t+h-1}, x_{t+h-2}^{(2)} = y_{t-2}$, the linear model is called autoregressive of order p, or AR(p). The parameter set β estimated provides valuable information about what influence the phenomenon (a coefficient α_j close to 0 indicates that variable $x^{(j)}$ has small influence).

There are plenty of variants of linear models:

- The β estimation depends on the loss function selected, and leads to different kind of forecasts. For instance, with a pinball loss, Koenker and Bassett obtain a quantile regression (Koenker & Bassett Jr, 1978).
- The addition of a regularization term in the β estimation. The minimization problem (2.3) is often ill-posed and parameters found might explode, leading to poor forecasting performance. A common regularization term is the one proposed by Tikhonov (Tikhonov, 1943).
- The recursive estimation of β, e.g. with the recursive least squares method, one sees the parameters evolution through time.
- The analysis of residual errors turns the deterministic forecasts into probabilistic ones.

Advantages – mature method, fast parameter computation, easy interpretation.
Drawbacks – linear hypothesis, non-flexible, mediocre performance additional work for probabilistic forecasts.

2.1.2.3 Additive Model (AM)

The event to forecast y_{t+h} is generally non-linear in the input set s_{t+h}, so one can non-linearly transform the inputs. In the most basic framework, the effects of each input are supposed to be independent of the others, hence the additive framework

$$g_\beta^h(s_{t+h}) = \alpha_0 + \sum_{j=1}^{p} \alpha_j l_j \left(x_{t+h}^{(j)} \right). \tag{2.5}$$

Functions l_j transform the shape of the input $x^{(j)}$. One may infer the functions based on their data, e.g. if $x^{(j)} > 0$ one uses $l_j(x^{(j)}) = \log(x^{(j)})$.

In most of the cases, however, the forecaster does not know the best transform functions to use. A common approach is thus to use spline functions to find the best

functions. The idea is to fit polynomials to the data on a series of intervals, and to match the polynomials and their derivatives on the boundary points. A penalty term is added to obtain smooth functions. The natural cubic splines are a widely used type of splines, but other kinds of splines exist, see Rodriguez's review (Rodriguez, 2001) or Wahba's reference book (Wahba, 1990).

By analyzing the parameter set made of the function l_j and parameters α_j, one observes the influence of input variables $x^{(j)}$, which provides a useful insight into how the model works. As for the linear models, there are plenty of variants of the additive model framework: recursive estimation, probabilistic extension; but also more complex framework: observation transformation (hence the Generalized AM), multi-dimensional functions which analyze combine effects of multiple variables. One should however be careful, because the number of parameters quickly increases with the number of inputs, leading to possible overfitting.

Advantages – easy interpretation, flexible, good performance, adaptable to non-linear effects.
Drawbacks – small number of predictors, additional work for probabilistic forecasts.

2.1.2.4 Kernel Density Estimator (KDE)

The forecaster uses the historical observations $y_1, \ldots, y_T$ of the training set to anticipate future value y_{T+h}. The historical observations having input set s_t similar to the one to forecast s_{T+h} are favored through a kernel function $K_\Lambda(s_t, s_{T+h})$ measuring situations proximity. The forecast density is then

$$\hat{f}_{T+h}(y) = \frac{\sum_{t=1}^{T} K_\Lambda(s_t, s_T) y_t}{\sum_{t=1}^{T} K_\Lambda(s_t, s_T)}. \tag{2.6}$$

The parameter set β contains description of this density function, and the expected mean value forecast can be computed $g_\beta^h(s_{T+h}) = \int_{\mathbb{R}} y \hat{f}_{T+h}(y) \mathrm{d}y$.

The most common kernel types are the rectangular uniform kernel, the Gaussian kernel, and the Epanechnikov kernel. All kernels have a bandwidth matrix Λ, of the same size as the input set s_t, that determine the proximity metric. Selecting the matrix structure, e.g. symmetric, diagonal etc., and the coefficients is usually more crucial than selecting the kernel type. When the coefficients are low, the neighboring window is narrow and fewer points are found close. It may be an issue when the input

space is large and the training set small. Conversely, when the coefficients are large, the neighboring window is wide and a lot of points are found close. It may be an issue since some irrelevant points then influence the forecasts when they should not. A vast literature is devoted to this bandwidth problem. A popular method is to use plugin bandwidth matrices such as the one proposed by Chacón and Duong (Chacón & Duong, 2010).

Advantages – fast parameter estimation, direct probabilistic forecasts.

Drawbacks – small number of predictors, limited performance, biased at the edge of input space.

2.1.2.5 Gradient Boosting Model (GBM)

Boosting model is a recent machine learning technique that produces forecasts with an ensemble of weak forecasting models, such as regression trees. The weak models are successively trained on the residual errors in a stage-wise fashion. Therefore the weak models need to be all used at once to carry out the forecasts — unlike other ensemble approaches, such as random forest, that work in a parallel fashion. Gradient boosting model generalizes this line of thought by making use of an arbitrary loss function $\mathcal{L}(\cdot, \cdot)$.

The original gradient boosting algorithm proposed by Friedman et al. is as follows (Friedman et al., 2000). One has a training set of observations $y_1, \dots, y_T$ and corresponding input sets $s_1, \dots, s_T$ and wishes to find a function $g^h_\beta(\cdot)$ — noted $g(\cdot)$ for clarity but specific to horizon h and parameter set β —, that carry out accurate forecasts of y_{T+h} with the input set s_{T+h}. The training is made recursively, for step $j = 1, \dots, J$, starting with fixed ${}^{(0)}g(s_1) = \cdots = {}^{(0)}g(s_T) = \text{constant}$

1. Compute the negative gradient, for $t = 1, \dots, T$,

$$ {}^{(j)}z_t = -\frac{\partial}{\partial g(s_t)} \mathcal{L}(y_t, g(s_t)) \Big|_{{}^{(j-1)}g(s_t)} . \tag{2.7} $$

2. Fit a weak regression model ${}^{(j)}w(\cdot)$ forecasting ${}^{(j)}z_t$ from s_t.

3. Choose a gradient step

$$ \rho^* = \underset{\rho}{\operatorname{argmin}} \sum_{t=1}^{T} \mathcal{L}(y_t, {}^{(j-1)}g(s_t) + \rho^{(j)}w(s_t)). \tag{2.8} $$

4. Update estimation, for $t = 1, \ldots, T$,

$$^{(j)}g(s_t) = {}^{(j-1)}g(s_t) + \rho^{*(j)}w(s_t). \tag{2.9}$$

All the successive weak learners are necessary to product a forecast $^{(J)}g(s_{T+h}) = \hat{y}_{T+h}$ according to the input set s_{T+h}. The typical weak learners $^{(j)}w(\cdot)$ should be quickly fitted, such as with a regression tree. Numerous refinement tricks exist to this basic algorithm. The selection of the optimal number of step $J^* < J$ is necessary and requires evaluation on an out-of-sample test set — in order to avoid overfitting. Besides, Friedman shows that the weak learners should be fitted with a subsample of the learning to improve performance (Friedman, 2002). A popular implementation of gradient boosting model has been made with package `gbm` on **R** (Ridgeway, 2017).
Advantages – highly flexible, high performance.
Drawbacks – computation-intensive, point forecasts, obscure interpretation.

2.2 Performance of a Forecasting Model

The quality of a forecasting model should be evaluated before being used in practice. This quality strongly depends on the needs of the users. In an electrical network, the needs of a grid planner differ from the ones of a electricity retailer: the former is more interested in peak consumption, whereas the latter focuses on consumption evolution throughout the day. Hence a good forecasting model for the retailer might be a poor one for the planner. This multi-aspect of forecasting quality calls for different evaluation method. Murphy (Murphy, 1993) explains that a forecasting model should be good on three aspects:

1. *consistency* means that the forecaster makes the best use of the knowledge base;
2. *quality* means that the forecast values are close to the observations;
3. *value* means that the forecasts benefit the users in his or her later decisions.

The first two aspects are part of the model and can be evaluated with statistical tools, but not the value which depends on the decision to make.

In the following are described the recurrent indices that are used in the rest of the thesis. First is introduced the evaluation of point forecasts, then the probabilistic forecasts. A simulation example is then detailed to show how one can use the evaluations.

2.2.1 Evaluation of Point Forecasts

A forecasting model produces a time series $\{\hat{y}_t\}_{t=1,\dots,T}$ that predicts the actual time series $\{y_t\}_{t=1,\dots,T}$. At each instant t, there exists a given error

$$e_t = y_t - \hat{y}_t. \tag{2.10}$$

An error close to 0 indicates that the forecast value is accurate. The error term is in the same unit as the time series $\{y_t\}$, e.g. in kW for electricity demand power. The error term is often normalized in order to have a dimensionless error. Different types of normalization exist. The normalization by the average value of the time series, noted $\operatorname{mean} y$, is the one we favor when studying household electricity demand, i.e.

$$e'_t = \frac{e_t}{\operatorname{mean} y}. \tag{2.11}$$

Of course a single error on one instant is not relevant to assess overall quality, so one uses the successive error of the time series time series $\{e_t\}_{t=1,\dots,T}$. Three main indices are obtained from the series and the normalized series:

- The systematic error of the model, the *Bias*, is the average error made by the model, i.e.

  $$\text{Bias} = \frac{1}{T}\sum_{t=1}^{T} e_t. \tag{2.12}$$

  The normalized version, using the normalized error, is called NBias. Closer to 0 is the bias, better is the model quality. If a model is known to be biased, one corrects forecasts by shifting so as to have a null bias.

- The *Mean Absolute Error* is the average absolute error, i.e.

  $$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T} |e_t|. \tag{2.13}$$

  The normalized version is called NMAE. The MAPE (Mean Absolute Percentage Error) is often reported, when each error is divided by the corresponding phenomenon value, i.e. $\text{MAPE} = 1/T \sum_{t=1}^{T} |e_t|/y_t$ (Poggi, 1994). However, this division is troubling when the phenomenon y_t is close to 0 at a given instant. Unlike the Bias, the MAE prevents the model from balancing positive errors with negative errors, so it cannot be shifted to 0. This score is negatively oriented, i.e. the closer to 0 the better is the model.

- The *Root Mean Square Error* is the square root of the average of the quadratic errors, i.e.

$$\text{RMSE} = \left(\frac{1}{T} \sum_{t=1}^{T} (e_t)^2 \right)^{\frac{1}{2}}. \tag{2.14}$$

  The normalized version is called NRMSE. Compared to the MAE, the RMSE strongly penalizes large errors. This score is negatively oriented, i.e. the closer to 0 the better is the model.

2.2.2 Evaluation of Probabilistic Forecasts

Point indices are not suited for probabilistic forecasts since the evaluation is based on a single value, rather than on the whole distribution. As stated by Gneiting et al., a good probabilistic model should maximize the *sharpness* of the forecast distribution subject to *calibration* (Gneiting et al., 2007). It relates to the first two aspects — consistency and quality — stated by Murphy (Murphy, 1993).

The calibration quality is often referred as reliability, e.g. by Pinson et al. (Pinson, Nielsen, et al., 2007). This property consists in checking that the shape of the predictive distribution is close, and ultimately convergent, to the observed distribution. To study the calibration one may use a reliability graph. It gives the frequency of obtaining an observation between two predicted quantiles. Formally, for a list of quantiles $\tau_0 = 0 < \tau_1 < ... < \tau_k < \tau_{k+1} = 1$

$$\text{Rel}(\tau_l) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}(\hat{y}_t^{\tau_{l-1}} < y_t \leq \hat{y}_t^{\tau_l}), \tag{2.15}$$

and a perfectly calibrated model is such that $\text{Rel}(\tau_l) = \tau_l - \tau_{l-1}$. The limited evaluation set results in slight fluctuations that can analyzed regarding the sample statistical errors, and the serial correlation (Pinson et al., 2010).

Candille and Talagrand propose a single ratio value to assess whether a probabilistic distribution is calibrated relatively to the T observations (Candille & Talagrand, 2005). Supposing that the quantile levels $\tau_0, \tau_1, \ldots, \tau_{k+1}$ are regularly spaced, then the expected values of $\text{Rel}(\tau_1), \ldots, \text{Rel}(\tau_{k+1})$ are all $T/(k+1)$. The value

$$\Delta = \sum_{l=1}^{k+1} \left(\text{Rel}(\tau_l) - \frac{T}{k+1} \right)^2 \tag{2.16}$$

then quantities the deviation of the reliability from flatness. Due to the statistical variation, the expected value of Δ is

$$\Delta_0 = \frac{kT}{k+1}. \tag{2.17}$$

The ratio Δ/Δ_0 is then used as a measure of the reliability of a probabilistic distribution. A ratio close to 1 indicates a correct reliability, while a value significantly larger than 1 us a proof of unreliability.

Calibration is thus a joint property of the forecasts and the observations. Similarly to the systematic bias, that can be removed by shifting values, one increases calibration by dilating or compressing forecast distributions.

Sharpness is the ability to concentrate the forecast distribution around the future observations. For prediction intervals, sharpness means that sizes of the intervals are not too large so as to give meaningful information to the users. In fact, interval sizes reflect the uncertainty put into the forecast. Point forecasts can be seen as probabilistic forecasts with null width (infinite sharpness), but are poorly calibrated since all quantiles coincide and are not consistent with the observations. Between two models similarly calibrated, the one with the lowest interval sizes (greatest sharpness) should be preferred. It means that uncertainty is reduced because the model makes better usage of its inputs.

In practice, a trade-off between calibration and sharpness should be made, like the trade-off between bias and variance during statistical estimation of parameters. The trade-off can be made by evaluating both calibration and sharpness simultaneously. Several authors proposed single scores to reflect efficiency, see, for example, Gneiting and Raftery (Gneiting et al., 2007) and Bickel (Bickel, 2007). Among a multitude, the *Continuous Ranked Probability Score* (CRPS) is a prominent one gaining popularity in the recent years. It corresponds to the integral of the Brier scores for all values and is derived from Cramér-von Mises divergence. The CRPS is estimated on a period $\{1, \ldots, T\}$,

$$\mathrm{CRPS} = \frac{1}{T} \sum_{t=1}^{T} \int_{\mathbb{R}} \left(\mathbb{1}(y \geq y_t) - \hat{F}_t(y) \right)^2 \mathrm{d}y. \tag{2.18}$$

For each instant, it compares the proximity between the forecast distribution $\hat{F}_t$ to the Dirac cumulative function $\mathbb{1}(\cdot \geq y_t)$ of the event — which is equal to 0 on $(-\infty, y_t)$

and to 1 on $(y_t, +\infty)$. Similarly to point forecasts, the CRPS can be normalized by the average of the observations, resulting in the NCRPS.

Laio and Tamea (Laio & Tamea, 2007) transformed expression (2.18) to introduce the quantile score at level $\tau \in (0,1)$, noted $\mathrm{QS}_\tau(\cdot,\cdot)$,

$$\mathrm{CRPS} = \int_0^1 \underbrace{\frac{1}{T}\sum_{t=1}^{T} \mathrm{QS}_\tau\left(\hat{y}_t^\tau, y_t\right)}_{\mathrm{QS}_\tau} \mathrm{d}\tau, \tag{2.19}$$

where

$$\mathrm{QS}_\tau\left(\hat{y}_t^\tau, y_t\right) = 2\left(\mathbb{1}\left(y_t \leq \hat{y}_t^\tau\right) - \tau\right)\left(\hat{y}_t^\tau - y_t\right). \tag{2.20}$$

Details on the equality of the two expressions are given in Appendix B.

While closed-form expressions of the indices (CRPS and QS_τ) exist for specific distribution (see Jordan (Jordan et al., 2017)), most of the time, one relies on numerical calculation. Expression (2.20) is useful to numerically compute the integral by using regularly spaced quantile levels $\tau_0 = 0 < \tau_1 < ... < \tau_k < \tau_{k+1} = 1$, i.e.

$$\mathrm{CRPS} = \frac{1}{k+1}\sum_{l=0}^{k+1} \mathrm{QS}_{\tau_l}. \tag{2.21}$$

A graph plotting the quantile score against the quantile level τ informs on the quality of a quantile point forecast (Gneiting, 2011), and is thus a useful diagnostic tool when analyzing performance. In general, the curve is bell shaped and the middle quantile scores are larger than the extreme ones. When the y_t are sampled from a standard distribution $\mathcal{N}(0,1)$, and forecast with the theoretical quantiles of the standard distribution, then the quantile scores are exactly $\mathrm{QS}_\tau = 2\phi\left(\Phi^{-1}(\tau)\right)$, where ϕ and Φ are the density and cumulative function of the standard distribution. For some applications, such as wind power trading (Pinson, Chevallier, & Kariniotakis, 2007), one wants to emphasize certain parts of the distribution, e.g. higher quantiles. Gneiting and Ranjan (Gneiting & Ranjan, 2011) proposed a weighted version of the quantile score. Accordingly, the weighted CRPS writes

$$\mathrm{CRPS}_w = \frac{1}{kT}\sum_{t=1}^{T}\sum_{l=1}^{k} w(\tau_l)\mathrm{QS}_{\tau_l}\left(\hat{y}_t^{\tau_l}, y_t\right). \tag{2.22}$$

Table 2.1 gives examples of weight functions. The 'Uniform' version is used to compute the regular CRPS. The other ones lead to weighted verions of the CRPS: the

'Upper Tail' (resp. 'Lower Tail') weights takes only high quantiles aboe 95% (resp. low quantiles below 5%) into account; the 'Standard' weights put the same influence of the quantile score at every quantile level for a $\mathcal{N}(0,1)$ distribution.

Table 2.1 – Examples of weight functions, as suggested by Gneiting and Ranjan (Gneiting & Ranjan, 2011) that define weighted versions of the CRPS, that emphasize different parts of the distribution (see Equation (2.22)).

Name	**Weight function**
Uniform	$w(\tau) = 1$
Lower Tail (LT)	$w(\tau) = 20 \cdot \mathbb{1}(\tau \leq 0.05))$
Upper Tail (UT)	$w(\tau) = 20 \cdot \mathbb{1}(\tau \geq 0.95))$
Standard (S)	$w(\tau) = 1/\phi(\Phi^{-1}(\tau))$

The CRPS is non negative and negatively oriented: the lower the CRPS, the better is the model. Diebold and Mariano (Diebold & Mariano, 1995) proposed a statistical test to favor one model over another by comparing score like CRPS. However, most of the time, the quantile scores of 2 models cross at certain quantile levels, and therefore one model is better for one part of the distribution, and another is better for the other part, suggesting the usage of one or the other model depending on the part of the distribution to forecast. Ehm et al. (Ehm et al., 2016) explain that a model is to be rejected only if all of its quantile scores are beaten by another model.

2.2.3 Simulation Study

We make use of the random walk of Example 1 (page 35) to show how to assess quality of competitive models. A total of five models produce forecast distributions of y_t. The first four distributions use information up to instant $t-1$: three of them are centered on y_{t-1} with correct, too low, or too high variances; the fourth one is centered on a biased value. The fifth model use information up to instant $t-2$ with the correct variance. Models are detailed on Table 2.2. On the table are reported the scores estimated with a simulation running for a period $T = 10^5$. Some conclusions can be drawn from the scores:

- As expected, the optimal model has the lower scores and perform the best.

Table 2.2 – Comparison of 6 forecasting models of a standard random walk. Five indices (Bias, MAE, CRPS, CRPS_{UT}, and CRPS_{LT}) are estimated with a simulation for a duration of $T = 10^5$.

Name	**Probabilistic forecast**	**Bias**	**MAE**	**CRPS**	**CRPS_{UT}**	**CRPS_{LT}**
Optimal	$\mathcal{N}(y_{t-1}, 1)$	**0**	**0.80**	**0.58**	**0.13**	**0.13**
Narrow	$\mathcal{N}(y_{t-1}, 1/2)$	**0**	**0.80**	0.62	0.24	0.24
Wide	$\mathcal{N}(y_{t-1}, 2)$	**0**	**0.80**	0.66	0.22	0.22
Biased	$\mathcal{N}(y_{t-1} + 1, 1)$	-1	1.17	0.85	0.17	0.26
Past	$\mathcal{N}(y_{t-2}, \sqrt{2})$	**0**	1.13	0.81	0.19	0.19

- Since the first 3 models are centered on the same values, one cannot discriminate their performance by evaluating point forecasts (Bias and MAE).
- The CRPS is less penalizing for the narrow model than the wide model, the upper and lower tail versions of the CRPS emphasize the tails and show that wide model is more efficient to estimate extreme values.
- A correct variance is paramount for the upper and lower tail CRPS. Indeed, even though the narrow and wide models perform better on MAE and CRPS compared to the past model, this past model is more efficient on distribution tails.
- A biased model leads to good performance on extreme parts of the distribution. One should be careful when examining only the performance at high quantiles since a model may be off for the rest of the distribution (Lerch et al., 2017).

In addition to these indices, a visual inspection of the quantile scores gives a good overview of the performance of a forecasting model. Figure 2.3 plots reliability $\text{Rel}(\tau_l)$ for the example models. Models with correct variances (optimal and past) are perfectly reliable. A narrow model has a U-shape while a wide model has an inverse U-shape. A biased model is a downward slope when bias is negative (and upward when the bias is positive).

Figure 2.4a shows the QS_τ for different quantile levels, and Figure 2.4b presents a zoom on the highest quantiles. The lower the quantile scores are, the better the model. One observes that the narrow model is better than the wide for the middle

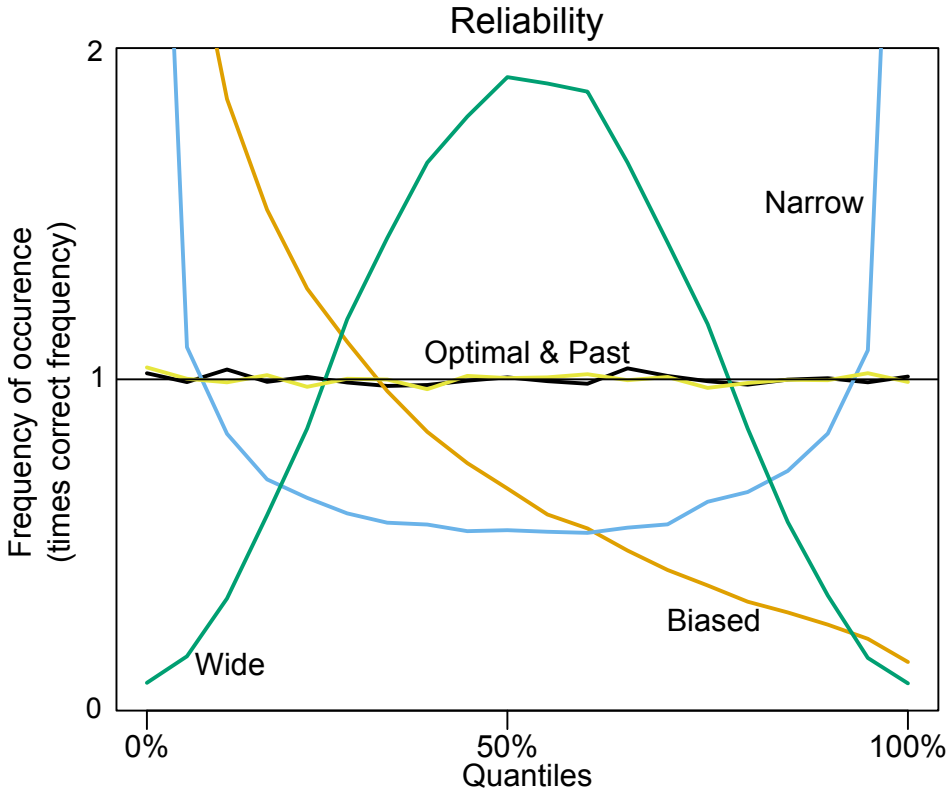
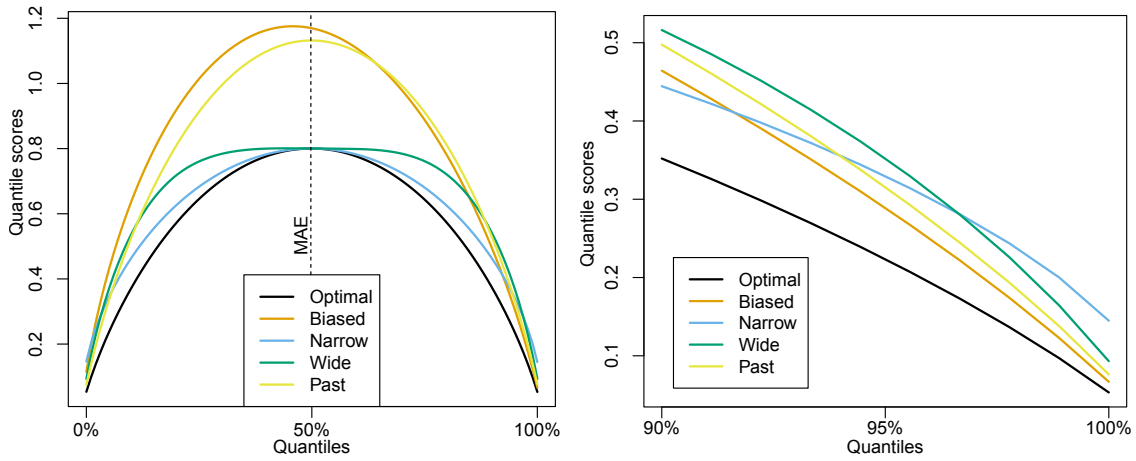


Figure 2.3 – Reliability of the 5 models of Table 2.2 with the random walk example. For perfectly reliable models (optimal and past models in black and yellow), the reliability line is constant. A narrow model (blue) has a U-shape. A wide model (green) has an inverse U-shape. A biased model (orange) has a downward slope when the bias is negative (and upward when the bias is positive).

part (5—95%) but not for the extreme parts (0—4% and 96—100%). Let us finally note that the MAE can be read directly from the graph by looking at the quantile scores at the 50% level.

2.3 Review of Electricity Demand Forecasting Models

Load forecasting is a crucial for the planning and operation of electric utilities. The forecasting horizon depends on the usage one makes of the forecasts: energy policy anticipates demand in the following years, while typical unit commitment problems



(a) Whole distribution

(b) Upper tail zoom

Figure 2.4 – Value of the quantile scores for different quantile levels for the 5 models of Table 2.2. Subfigure (a) shows the whole distribution, while subfigure (b) focuses on the upper tail of the distribution. The lower the quantile score, the more efficient is the model. The MAE corresponds to the value at quantile level 50%, and the CRPS to the integral of these curves.

are studied a few days in advance (Hong & Fan, 2016). Here we focus on a horizon of 1 hour to 1 week, referred to as a short-term load forecasting. Some authors prefer to talk about middle-term load forecasting when the horizon is larger than 1 day (Srinivasan & Lee, 1995). The main quantity of interest is the hourly demand, which is the average power called on a 1 hour time period, expressed in kWh or, abusively, in kW. Some related quantities are present in the literature, such as the daily or monthly load, the value and instant of the peak load and so on.

A short review of forecasting models that have been proposed is sketched in the following sections. First, we present demand forecasting model at large scale (high voltage and power in the MW—GW range), then forecasting model focusing on the local scale (low voltage and power in the kW—MW range), such as a neighborhood or a building.

2.3.1 Large Scale Forecasts

Large-scale demand is fairly regular, and therefore relatively easy to forecast. Typical day ahead forecasting error is around 1%-3% for power in the GW magnitude. However, even a small improvement decreases the costs significantly. Hobbs et al. (Hobbs et al., 1999) find that a 1% error reduction may result up to hundreds of thousands dollars savings for a typical utility, mostly due to better unit commitment.

Large-scale demand forecast has been studied extensively for a long time. Matthewman and Nicholson (Matthewman & Nicholson, 1968) propose a review of the load forecasting in 1968. Effect of the meteorology is mathematically formulated, see Dryar's early attempt (Dryar, 1944), and is used in multivariate regression with factors such as temperature, wind, cloudiness and precipitation. Since monitoring meteorology is costly, both in price and in memory for first computers, models using only historical data with no meteorological inputs have been and still are developed. Taylor and McSharry (Taylor & McSharry, 2007) compare several time series methods without weather inputs and obtain results competitive with weather-based models, with an average MAPE of 1.8% reported for day-ahead demand of different European countries. Dordonnat et al. (Dordonnat et al., 2008) forecast hourly demand of France with detailed trend and dynamics effects and obtain a MAPE around 1.5%. Misiti et al. (Misiti et al., 2010) optimally cluster industrial customers in a two-step process, and forecast each cluster separately to find the industrial demand, with a short-term MAPE around 1.5%. The effect of weather on electricity demand is challenging for forecasters and its precise impacts have been extensively studied. The most important factor is undoubtedly the temperature and thus the most investigated effect. Bessec and Fouqueau (Bessec & Fouquau, 2008) conduct a comparative study of the effect of temperature on national load of several European countries and show that the non-linear influence depends on the country considered. However, it is not clear what are the exact parameters that matter for forecasting performance. For instance, Wang et al. (P. Wang et al., 2016) investigate the effect of lagged temperature values on forecasting errors. The authors observe that using the last two daily temperature and the average of the last 12 hourly temperature values reduce MAPE from 5% to 3.5%.

Alfares and Nazeeruddin (Alfares & Nazeeruddin, 2002) publish a more recent review in 2002 where they classify the load forecasting techniques in 9 groups and they

identify a paradigm shift from time series methods toward more complex methods such as neural network and knowledge-based models. In addition to this shift, nature of forecasts evolve from deterministic to probabilistic. In this respect, the 1993 paper of Hendricks and Koenker (Hendricks & Koenker, 1992), where they demonstrate the use of hierarchical spline functions and quantile regression for household electric demand, is seminal.

Large-scale forecast remains of vivid interest in the community with important research competition. Academically, the global energy forecasting competition (GEFCom) attracted more than 500 forecasting teams from all over the world and challenged to forecast load of a utility of average power around 100 MW: the 2012 version focused on hierarchical forecasting (Hong et al., 2014), and the 2014 one on probabilistic forecasts (Hong et al., 2016). French DSO RTE organized two competitions in 2017 and 2018 to forecast national and regional demands: one with point forecasts, and one with probabilistic forecasts. Such competitions lead to the development of efficient and practical models. Charlton and Singleton (Charlton & Singleton, 2014) precisely model the effect of temperature with polynomial regression and then improved quality with practical adjustments. Xie and Hong (Xie & Hong, 2016) combine simple linear model with temperature scenarios to simulate the residual errors. Gaillard et al. (Gaillard et al., 2016) design a quantile version of the generalized additive model to forecast a temperature distribution useful to improve forecasting performance. Liu et al. (Liu et al., 2017) propose a quantile regression average based on sister forecasts: first they train multiple point forecasting models based on slightly different training sets (the sister models), then they do regression on the sister forecasts to obtain a probabilistic output.

Other searchers develop complex hybrid models that implement the most recent techniques in forecasting. He et al. predict Singapore demand (around 5 GW) with a density function through a kernel-based support vector quantile regression method (Y. He et al., 2017). Clements et al. develop a refined time series framework, with cooling and heating degree variables, in order to forecast a 5 GW load in Australia with a MAPE around 1.4% (Clements et al., 2016).

2.3.2 Local Scale Forecasts

Here, we present models proposed in the literature to forecast electricity demand at the local scale. We mostly focus on the household level but some authors extend their models to incorporate larger scales (neighbourhood, aggregation of households etc.). Day-ahead forecasting errors generally increase when narrowing the scale studied such as the household level. Forecasting errors greatly vary between studies and are reported going from 2% to 85%. The errors strongly depend on the average power level of the demand time series to forecast. Sevlian and Rajagopal highlight a relation between forecasting error and average power — error decreases when average power increases —, and identify a critical power and an irreducible error (Sevlian & Rajagopal, 2014). This relation is concurred by our own case study, in Section 4.3.

The number of articles devoted to forecast household or residential demand is becoming quite large. Since most studies use different — and private — datasets, one abstains from definitive conclusion about competitive forecasting models. Hong and Fan explain in their tutorial for probabilistic forecasts (Hong & Fan, 2016) that there is no universal best technique: "it is the data and jurisdictions that determine what technique we should use, rather than the other way around". Therefore, in the following, we briefly review recent papers on the subject, so as to provide a broad spectrum of forecasting techniques.

The impact of temperature on the local scale demand is ambiguous. Average household demand undoubtedly increases when the outside temperature is low and high, due to heating and cooling devices, but using it as an input for short-term forecasting does not always improve performance. In fact, it is not clear whether the temperature information is encapsulated in other inputs, such as time of the year. Some studies specifically focus on this temperature usage and report an almost null improvement (Haben et al., 2018). We elaborate on the temperature impact on the household demand forecasting performance in Section 4.2.4.2.

Consequently, some authors do not use temperature input in their short-term forecasting models. For instance, Ben Taieb et al. develop a hierarchical probabilistic forecasting model with no temperature input: individual demand is forecast with KDE and then combined together with copula to forecast different levels of aggregation (S. B. Ben Taieb et al., 2017). The residential electricity time series is sometimes

modeled as a mix of normal and log-normal processes, which enables a convenient probabilistic approach to forecasting (Shepero et al., 2018). Advanced neural networks techniques are adapted by Wang et al. to capture the volatile features of the mechanisms governing the individual demand dynamics (Y. Wang et al., 2019). Also with no temperature input, Mocanu et al. develop deep learning methods and evaluate forecasting performance for different resolutions — 1 minute to 1 week — at different horizons — 15 minutes to 1 year (Mocanu et al., 2016). This temperature independence is sometimes highlighted as an advantage by Rodrigues et al., who develop an artificial neural network, relying solely on historical values and hour of the day, in order to accurately forecast both daily and hourly demands of individual households in Lisbon, Portugal (Rodrigues et al., 2014).

However, most researchers rely on temperature values for their models. In a forecasting application, the forecaster generally retrieved temperature values forecast by other organisms such as Weather Underground[3] or European Centre for Medium-Range Weather Forecasts[4] (ECMWF). Arora and Taylor use KDE to forecast Irish smart-meter demand (Arora & Taylor, 2016). Kavousian et al. identify the most important factors of residential electricity consumption (Kavousian et al., 2013). Lusis et al. study the forecasting performance of neural networks of a set of 27 Australian households at different temporal granularity (Lusis et al., 2017). Bina and Ahmadi model aggregated appliance usage for demand response applications (Bina & Ahmadi, 2015). Gajowniczek and Ząbkowski consider states of appliances to forecast individual household demand (Gajowniczek & Ząbkowski, 2016), and attempt to model activity patterns to improve performance (Gajowniczek & Ząbkowski, 2017). Hsiao investigated households in Taiwan and proposed advanced methodology forecast daily demand profile with great accuracy (Hsiao, 2015). Ghofrani et al. adapt a Kalman filtering method to predict residential demand at very short term horizon (Ghofrani et al., 2011). Bennett et al. forecast demand at the low voltage level using a hybrid three step algorithm with clustering, neural network and post-treatment to obtain a low MAPE around 12% (Bennett et al., 2014). Some research focus on the performance when aggregating multiple household demands together. Humeau et al. use neural networks and support vector machine techniques to forecast demand at different level of aggregation in

[3] https://www.wunderground.com

[4] https://www.ecmwf.int

Ireland (between one and a hundred of households) (Humeau et al., 2013). Wijaya et al. forecast different household demand independently and find that forecasting them simultaneously does not improve aggregated forecasting performance (Wijaya et al., 2014). Tidemann et al. compare forecasting at different aggregation levels with little-used advanced techniques (echo state network, wavelet and case-based reasoning) (Tidemann et al., 2013). Detailed articles make extensive reviews of the current forecasting techniques at the household level: such as Yildiz et al. who provide insightful advice for demand modeling and forecasting (Yildiz et al., 2017), Ahmad et al. who review neural networks and support vector machine to forecasting building electricity demand (Ahmad et al., 2014), or Lefieux's exhaustive work on demand forecasting semi-parametric models (Lefieux, 2007).

Another popular approach is to rely on clustering techniques to carry out the forecasts. Some opt to cluster the daily profiles of individual household. Yu et al. use a dictionary of past demand profile along with a polynomial temperature fit for individual households in the United States (Yu et al., 2017). Similarly, Abreu et al. identify daily profile characteristics of a single household with a principal component analysis, then cluster typical daily profile(Abreu et al., 2012). Others try to cluster customers together based on their demand time series, which is an interesting application for targeted pricing system. Giasemidis et al. monitor only a fraction of total feeder demand, then extrapolate from the clustering of available customers to forecast the total load (Giasemidis et al., 2017). Wijaya et al. develop a similar methodology to forecast aggregate consumption by summing demand of individual households with similar features (Wijaya et al., 2015). Haben et al. cluster residential demand based on the shape of the curves, and notice that time-of-use tariff has practically no influence on demand levels (Haben et al., 2016). Dent et al. normalize daily profile before classifying customers with various methods (Dent et al., 2013). Viegas et al. propose a similar method but make use of survey information to reduce the training period (Viegas et al., 2016). Quilumba et al. construct customer groups and train neural network to obtain MAPE around from 2 to 5% depending on the forecasting horizon (Quilumba et al., 2015). Other researchers investigate to cluster together using exclusively sociodemographic information. Beckel et al. identify the most relevant sociodemographic (number of residents, employment status, cooking appliances, floor area etc.) information for clustering customers (Beckel, Sadamori, et al., 2014). Javed

et al. also use anthropological and structural factors (such as number of people in the house, or the living space of the house) to demonstrate that a forecasting model trained on multiple similar households is more efficient that trained on a single one (Javed et al., 2012).

Another promising path for household demand forecasting is bottom-up approaches, i.e. construct the demand profile from scratch by summing up the various appliances demand of the household. It involves precise modeling of every appliance, which is a challenging task. Sancho-Tomás et al. model only partial demand due to small household appliances (audio-visual, computing, small kitchen and others) (Sancho-Tomás et al., 2017). Stokes propose to model household demand with multiple appliances at different timescales (Stokes, 2005). A limiting issue of such bottom-up methods is the scarcity of precise measurements at the appliance level. Such datasets emerge, such as the UK-DALE dataset (Kelly & Knottenbelt, 2015) or the ECO dataset in Switzerland (Beckel, Kleiminger, et al., 2014). Since the data collection is made on very short granularity (every second), only a small sample is available: only 6 households in both datasets. Some broader measurement campaigns are conducted but are often specific to given appliances, such as the cold appliances usage in 100 French households (Ademe, 2008). Since the campaigns are made separately during different periods, combining the information is rather cumbersome. Large samples are also available at coarser granularity (every minute) in the US with the Pecan Street project (*Pecan Street Inc. Dataport*, 2018). We use this data to model and forecast an individual electric vehicle charging time series in Section 5.2. Nevertheless, most researchers use the scarce data in combination with more general information. For instance, Paatero and Lund reflect the variety of users by taking social variables into account (Paatero & Lund, 2006), and Dickert and Schegner retrieve appliances usage statistics regarding the power level and frequency of use (Dickert & Schegner, 2010). The precise modeling is also enabled by time-use survey (TUS) providing valuable information, e.g. when people turn their dryer on in Sweden. Richardson et al. use TUS information to model various household activities with probability of transition from one activity to another calibrated by average frequency of use (Richardson et al., 2010). Widén and Wäckelgård proposed a fine-grained model with Markov chains for Swedish household demand based on TUS (Widén & Wäckelgård, 2010). However, the validation of such bottom-up models is often difficult, for instance, Tanimoto et al. modeled several appliances along with TUS

to generate demand profile but validate it with only data from two days (Tanimoto et al., 2008). The relative lack of appliance data is hoped to be filled by disaggregating the total household demand, measured for instance by smart-meters, rather than by multiplying the number of measurement campaigns. In particular, the non-intrusive load monitoring (NILM) shows promising results if the time resolution of the measurements is high. The interested reader can explore this topic is referred to Parson's work (Parson, 2014).

2.3.3 Errors and Average Power

On Figure 2.5, we represent short-term forecasting demand errors (y-axis) versus the average power of load (x-axis) using articles reviewed in the two previous sections. We keep only models doing short-term — mostly day-ahead — forecasts of hourly electricity demand values. We keep only the best performing models proposed in the article. When the auithors study different level of demand aggregation, multiple points appears on the figure. The performance evaluation differs from one author to the other, in particular regarding the indices reported. Consequently, we apply the following equivalence coefficients between indices, based our own work, see detailed results in Appendix D.3,

- the MAPE is multiplied by 1.3 to obtain the NMAE;
- the RMSE is multiplied by 0.6 to obtain MAE, and then normalized by average power;
- the CRPS is multiplied by 1.4 to obtain MAE, and then normalized by average power.

Since authors rarely report the average power of the demand time series they study, we do our best to deduce this information.

The purpose of the figure is to show the level of performance that can be expected with state-of-the-art forecasting models given an average power of the demand time series. A power law is fitted to the data. For an average power W — expressed in kW —, then

$$\mathrm{NMAE}(W) = \sqrt{\frac{\beta_0}{W^p} + \beta_1}. \tag{2.23}$$

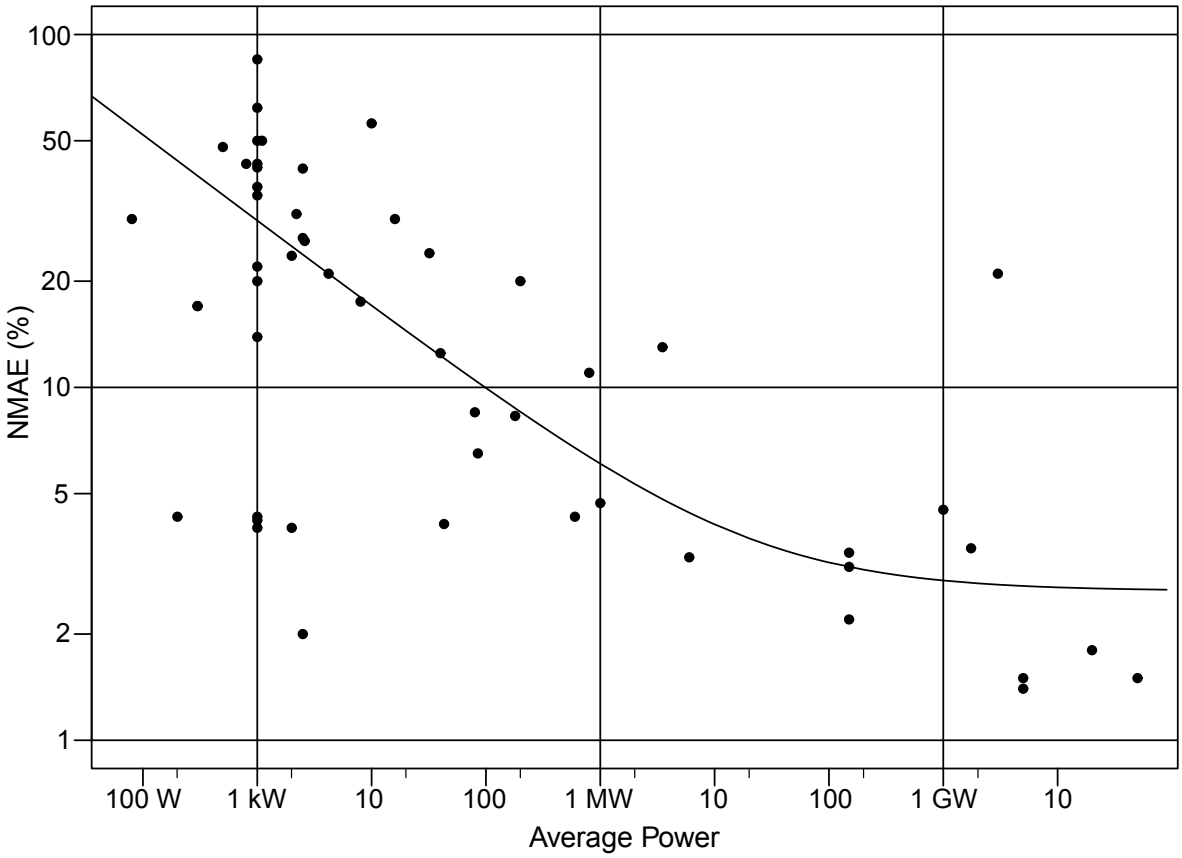


Figure 2.5 – Scatterplot of the short-term forecasting errors for models proposed in the literature reviewed in Section 2.3.1 and Section 2.3.2. The NMAE is on the x-axis in % and logarithmic scale. The average power of the demand time series is on the y-axis in logarithmic scale. The solid line is obtained with a robust power fit of Equation (2.23).

The best parameters found with a robust estimation are: $p = 0.48$, $\beta_0 = 875$, and $\beta_1 = 7$. According to Sevlian and Rajagopal, two regimes are then defined: a scaling law for $W < W^*$ where NMAE strongly decreases with growing average power, and a saturation law for $W > W^*$ where performance no longer improves — attaining an irreducible error (Sevlian & Rajagopal, 2014). The threshold obtained here is when $\beta_0/W^{*p} = \beta_1$, i.e. $W^* = 19$ MW. From this average power on, the performance demand forecasting models plateau around an irreducible error of $\sqrt{\beta_1} \approx 2.65\%$. However, due to the large heterogeneity of the data in the literature, one wants to conduct similar analysis with one's own data. This is what we do in Section 4.3 with US demand data forecast at different temporal granularity and aggregation levels.

Chapter 3

Electricity Demand at the Feeder Scale

Summary The feeder electricity demand sums up all of the individual demand of the clients connected to an electric feeder. The number of clients connected ranges from 1000 to 10,000 depending on the population density. Compared to the features of the individual electricity demand measured by smart meters, the measures of this feeder demand have several advantages: they are exhaustive — all clients are included, non-invasive — individual demand is hidden among the others, and have been collected over a long period — decades or so. Substantial research has been devoted to the demand at this scale, and what drives it. These driving effects have been clearly identified at this aggregated scale, such as the temperature influence. In fact, the aggregation smooths out the individual behavior and reveals the effects, even marginal, that are indistinguishable at the individual scale. This means that forecasting the feeder demand for short to medium horizons, up to several weeks in advance, is quite efficient: state-of-the-art relative errors are around 10%. On the other hand, forecasting for longer horizon necessitates a better understanding of the underlying mechanisms of the demand. Such task is yet necessary for planning the network infrastructure. In Section 3.1, we propose novel algorithm disaggregating the feeder demand in elementary profiles. The algorithm makes use of demand of multiple feeders along with their corresponding customer information systems. An elementary profile depicts the demand of a cluster of customers on regular intervals, e.g. every ten minutes. The clustering process is based on the customer information so that customers of the same

clusters have approximately the same electricity demand dynamics. In Section 3.2, we introduce two typical applications enabled by the elementary profiles. Firstly, the demand of a new unmeasured feeder is estimated with relative errors between 12 and 15%. Secondly, the daily demand peak is examined when additional customers are connected to the feeder. Depending on the kind of customers, the peak shifts to noon or the late evening. However, as we explain in Section 3.3, the elementary profiles reflect average dynamics and cannot be used to improve the short-term forecasting performance. This is due to the variation among the customers of the same cluster. Identifying such variation requires the usage of individual smart-meter measurements.

Résumé Nous nous intéressons à la demande électrique à l'échelle d'un départ HTA (Haute Tension de type A), c.-à-d. la demande totale d'un ensemble de 1000 à 10 000 clients. À l'inverse des mesures individuelles des compteurs intelligents, les mesures d'un départ sont exhaustives (tous les clients sont inclus), non-intrusives (la demande d'un client est noyée parmi celle de tous les autres), et couvrent une longue période (enregistrement depuis plusieurs décennies). Ainsi, cette demande agrégée a beaucoup été étudiée et ses caractéristiques sont bien comprises, notamment l'influence de la température. Ces caractéristiques sont bien visibles sur cette demande agrégée puisque l'effet de foisonnement atténue les comportements individuels, lisse la courbe, et fait ainsi ressortir les mécanismes communs de chaque client, aussi minimes soient-ils. Cela permet une prédiction à court et moyen terme (jusqu'à quelques semaines à l'avance) efficace avec des erreurs relatives de l'ordre de 10%. En revanche, la prédiction à plus long terme, nécessaire pour la planification du réseau, est problématique car elle requiert une compréhension plus fine des mécanismes régissant cette demande agrégée. Nous proposons dans la Section 3.1 un algorithme de décomposition de cette demande agrégée en profils élémentaires. Ces profils sont obtenus grâce à l'analyse combinée des mesures de multiple départs ainsi que d'un descriptif des clients raccordés. Un profil

élémentaire décrit la demande moyenne d'un groupe de clients à intervalle régulier (p. ex. toutes les 10 minutes). Les groupes sont constitués à partir du descriptif afin que les clients d'un même groupe aient des caractéristiques similaires, ainsi tous les restaurants sont dans le même groupe. Ces profils permettent plusieurs analyses prospectives pour l'évolution future du réseau. Nous présentons deux applications typiques dans la Section 3.2, à savoir la détermination du profil de demande d'un nouveau départ (erreurs relatives de l'ordre de 15%), et l'évolution du pic de demande journalier suite au raccordement de nouveaux clients. Comme nous l'expliquons dans le Section 3.3, les profiles obtenus sont seulement moyennés et ne réflètent pas la variation interne au sein d'un groupe. Il faut pourtant détecter ces variations pour affiner la prédiction à court terme, et cela passe par l'utilisation de mesures individuelles provenant de compteurs intelligents.

3.1 Disaggregating Feeder Electricity Demand in Elementary Profiles

3.1.1 Introduction

Electricity consumption represents 18% of total final energy consumption in 2013 (International Energy Agency, 2016b). This share is expected to increase to around 25% by 2040 (International Energy Agency, 2015). This consumption is responsible for an important part of global CO_2 emissions: the International Energy Agency estimates that 42% of all emissions in 2012 are due to electricity and heat production (International Energy Agency, 2016a). Reducing the electricity production is then seen as an important objective in most energy policy goals. For instance, the European Union energy policy aims to reduce by 20% the greenhouse gas emissions by 2020, including the electricity production, with even more stringent landmarks for the future. This energy transition involves significant changes to electricity distribution network, e.g. decentralized production, improved efficiency of buildings and appliances, new uses and demand response enabling energy consumption management (Jin et al., 2017).

These changes impact the planning process of distribution system operators (DSOs). The current network planning processes, such as infrastructure construction, consider only the two most extreme situations occurring at the feeder level, i.e. maximum demand with minimum supply, and maximum supply with minimum demand (Ding, 2012). The medium-voltage feeder delivers electricity for a few thousands customers, and is a prevalent scale for distribution purposes. Planning with such methods does not require a deep modeling of the electricity demand dynamics, since only isolated events are considered. The exact underlying processes governing the demand are not considered at all, so the exact electricity demand phenomenon remains unclear. In particular, the aggregation effect, i.e. how the electricity demand evolves when considering different number of consumers, is still an obscure phenomenon (Dickert & Schegner, 2010).

In the following, we investigate about the demand dynamics specifically at the feeder level, namely we aim to disaggregate this demand in elementary profiles to understand the underlying processes.

The feeder demand dynamics needs to become clearer in order to meet energy

reduction targets.

3.1.2 Related Works on Modeling Feeder Demand

Several research studies examining the feeder electricity demand have been published. While the ultimate goals slightly differ between studies, e.g. forecasting, simulating or characterizing, they all try to model the electricity demand. There exist two kinds of approaches: the global approach, relying on demand measurements made at the feeder level; and the bottom-up approach, building up the feeder demand from individual demand profiles.

3.1.2.1 Global Approach

With a global approaches, models are designed from historical demand measurements and related explanatory variables, such as the temperature or economic growth (Shao et al., 2015).

Most DSOs have been recording the electricity power delivered by their medium-voltage feeders for several years. These measurements are exhaustive, i.e. they take into account the losses made in the distribution grid in addition to the sum of all individual demands; but, since they are also aggregated, they hide the exact demand made by the individual consumer. By collating the successive measurements, one obtains an feeder electricity demand time series. This time series is rather complex, and described as a "nonlinear, non-stationary series, and is often made up by a superposition of several distinct frequencies" (Shao et al., 2017). Some authors consequently define daily to monthly seasonality in their models (Boroojeni et al., 2017). Others determine more precise precise temporal indicators, e.g. working time or holidays (Boroojeni et al., 2015; Goude et al., 2014). In addition to these temporal information, recurrent explanatory variables are integrated in the models, such as the temperature (Shao et al., 2017). However, to our knowledge, no feeder-specific features are directly used as explanatory variables, e.g. the number of restaurants connected to the feeder, for modeling the feeder demand time series. In fact, identifying the impacting feeder-specific features is complex.

In any case, models obtained from the global approach lead to high performance since the losses are integrated in the historical measurements. However, this does not

provide any way to design a prospective model for an unmeasured new feeder, i.e. with no historical measurements. The same issue arises when the set of consumers connected to the feeder evolves. The whole dynamics of the electricity demand changes due to this consumer evolution, and the historical measurements become obsolete. New consumers impact on (1) the mean electricity level and (2) the shape of the demand profile, and these impacts depend on the households' size, e.g small or large buildings, and type, e.g. residential or commercial. Yet, such models are essential for planning network infrastructure, such as its sizing.

3.1.2.2 Bottom-Up Approach

At the other end of the spectrum, some studies try to build the feeder demand from the individual consumers with the so called bottom-up approaches. Measuring the electricity demand of individual consumers is a simple way to establish their load profiles and dynamics, and therefore a necessary step in bottom-up modeling. The current smart-meter roll-out in Europe will provide precise measurements of individual demand profiles. Around 80% of customers are scheduled to receive a smart-meter by 2020 (European Parliamentary Research Service, 2015). In 2014, only 23% of smart-meters in the European Union were installed in localized areas for private customers (European Commission, 2014). In some countries, this share is still insufficient to be representative, and the corresponding deployment is too recent to adequately cover long periods. Pending the smart-meter roll-out, operators finance large measurements campaigns to obtain individual measurements. Since these campaigns are restrained to a small set of consumers, a clustering, or classification, of the individual profiles is then performed. With this clustering, every individual is assigned a demand profile, even in the absence of any individual measurement.

The clustering of electricity demand profiles is a flourishing research topic, see reviews (Zhou et al., 2013), (Rhodes et al., 2014). Researchers apply various clustering methods to the smart-meter time series (Viegas et al., 2016). Other include specific characteristics of the consumers, such as the information in the Customer Information System (CIS) — which is in possession of the DSO (Mutanen et al., 2011). This clustering reduces the size of the data to model, i.e. 1 model for 1 cluster of several individuals (McLoughlin et al., 2015). With the resulting clusters, a precise identification of load profiles can be performed (Räsänen et al., 2010). This identification is then used in

various applications.

Firstly, decision-makers can design personalized policies, such as time-use tariffs, for targeted consumers (Bassamzadeh & Ghanem, 2017). Secondly, the classification allows a DSO to plan its network and anticipate its investments (Mutanen et al., 2011; Seppälä, 1996). For example, the French DSO uses a model named "Bagheera" combining about 50 customer categories to plan its low-voltage network (Ding, 2012). Classification is combined with the evolution of category distributions to forecast aggregated demand in prospective scenarios (Andersen, Larsen, & Gaardestrup, 2013). Lastly, classification allows us to understand the contribution made by each category to aggregated demand (Seppälä, 1996).

The performances of the bottom-up models is generally lower than those of global models. This comes, in part, from the measurements campaigns: limited in size and time so demand profile clusters are difficult to update and to not describe the most recent consumption habits (Andersen, Larsen, & Boomsma, 2013; Räsänen et al., 2010). The poorer performance also comes from the losses of the distribution grid which are not measured at the individual scale.

3.1.3 Proposed Model: Feeder Demand Decomposition into Elementary Profiles

We take an intermediate approach that makes use of the historical measurements made at the feeder level, along with the individual information in the CIS. The CIS provides a wide range of precise information about the individual consumers: feeder connection, annual energy consumption, type of contract, and contracted power. To protect consumer's privacy (McKenna et al., 2012), we do not use the individual information, but information aggregated at the feeder level. It enables to identify consumer categories, e.g. residential category. This provides a description of the feeder, namely the exact mix of the categories forming each feeder, e.g. this feeder has 60% of residential consumers and 40% of offices.

We then propose an algorithm that decomposes the feeder demand time series into elementary profiles corresponding to each consumer category. The primary assumption is that the elementary profiles are the same across feeders, e.g. electricity demand of offices in feeder 1 is similar to the one in feeder 2. The decomposition is possible

when using multiple feeders with varying category mixes. Out algorithm optimizes the elementary profiles by minimizing quadratic differences. The exact algorithm is an advanced optimization method known as alternating direction method of multipliers (ADMM) (Boyd et al., 2011).

Unlike bottom-up methods requiring measurements at the individual levels, our method requires multiple feeder demand measurements and the associated category mixes of the feeders. The advantages of aggregated measurements compared to a set of individual load curves are: (1) the availability of long-term historical data, (2) exhaustiveness, and (3) continuous elementary demand profile updates. We compare the accuracy of our decomposition by trying to anticipate the demand of a new unmeasured feeder, and show that the performance is similar than one of a state-of-the-art bottom-up method. We also demonstrate the use of the elementary profile with a case study of around 300 feeders operate by the French DSO Enedis, specifically evaluating the evolution of the peak load when adding different kind of new customers to a feeder.

3.1.4 Electricity Demand Transforms

Each feeder f delivers electricity for a different number of consumers — from 100 to 10,000 —, so the raw feeder demand time series, noted $d_0^f(t)$ at instant t, are not comparable across different feeders. It means that the dynamics of small feeders are hidden among the dynamics of large feeders. A solution is that all feeder demands are transformed so as to have similar average level. Consequently, two transforms are applied to the raw feeder demand series: removal of the thermal effect, and normalization by average demand.

3.1.4.1 Thermal effect

The impact of the local outside temperature on electricity demand is generally recognized (Bessec & Fouquau, 2008) (Lefieux, 2007, pp 11—12). Since this impact depends on local features, we wish to remove this thermal effect for each feeder. In our case study, relation between the temperature and the electricity demand is easily observable and can be approximated with a linear effect: when the temperature is below a certain threshold, then demand increases linearly for each degree cooler. The exact linear threshold and trend depend on the hour of the day, so an hourly correction is

applied.

For each feeder f and each hour of the day $h = 0, \ldots, 23$, we fix the threshold temperature and fit a linear regression between demand and temperature when the temperature is below the threshold. The best fit results in a trend $a_h^f > 0$ and threshold temperature b_h^f which are used to transform the demand time series. At instant t, corresponding to a hour of the day h, a new demand is defined

$$d_1^f(t) = \begin{cases} d_0^f(t) & \text{if } T^f(t) > b_h^f, \\ d_0^f(t) - a_h^f \left(b_h^f - T^f(t) \right) & \text{otherwise.} \end{cases}, \tag{3.1}$$

where $T^f(t)$ is the temperature at instant t. The new demand time series d_1^f is then supposed to behave similarly independently of the temperature.

3.1.4.2 Normalization

The demand of each feeder is then normalized by its average value during the period $\{1, \ldots, T\}$ considered, e.g. one week,

$$d_2^f(t) = \frac{d_1^f(t)}{\int_{t=1}^{T} d_1^f(t) \mathrm{dt}}. \tag{3.2}$$

The average demand value, i.e. the total energy of the period, can be precisely predicted using different models as long as the period is long enough, e.g. a week (Andersen et al., 2014), and is thereafter supposed to be known. At the end of the day, each feeder's demand $d_2^f(t)$ fluctuates around a dimensionless value 1. We later drop the index, and simply note $d^f(t)$.

3.1.5 Disaggregation Algorithm

3.1.5.1 Recovering Demand Profiles

We collect the electricity demand values, noted $d^f(t)$, of a feeder $f \in \{1, \ldots, F\}$ measured at regular intervals, e.g. every 10 minutes, labeled by a time index t. For a feeder f, the mix of customers categories is defined by the share of each category, $p_1^f, \ldots, p_K^f$, summing up to 1. We assume that the value $d^f(t)$ aggregates K elementary profiles, namely $d_1(t), \ldots, d_K(t)$, corresponding to the categories of customers, i.e.

$$d^f(t) = \sum_{k=1}^{K} p_k^f d_k(t) + \varepsilon^f(t). \tag{3.3}$$

We take the elementary profiles $d_k(t)$ to be common to all feeders, while the weights vary from one feeder to another. The corresponding residual term $\varepsilon^f(t)$ is meant to be small. The decomposition consists in recovering unknown elementary demand profiles $d_k(t)$ for a given period, say for $t = 1, \ldots, T$. Note that the profiles are virtual and only represent the average demand profiles of all of the consumer in a given category. For each feeder f, $d^f(t)$ is observed and, thanks to the CIS, for each category $k \in \{1, \ldots, K\}$, we also have access to the share p_k^f. The process of obtaining shares from the CIS and defining categories is the categorization step, and is described in Section 3.1.5.2. Once the K profiles have been obtained on a set of feeders, it is possible to turn Equation (3.3) into a simulation algorithm. The process is described in Figure 3.1. In the signal processing community, the corresponding problem is called blind signal separation and is well-known, see e.g. (Cardoso, 1998).

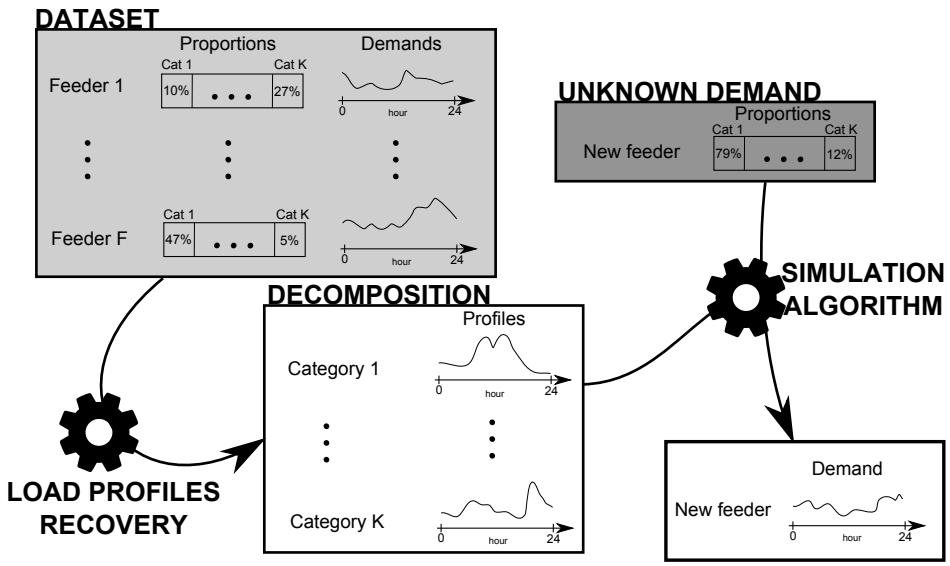


Figure 3.1 – Diagram detailing the disaggregation. A dataset of F feeder measurements is used to find the K category profiles. Once the load profiles recovery is operated, a new feeder whose category distribution is known can be run through the simulation algorithm to find an expected demand profile.

3.1.5.2 Categorization of Consumers

The feeder demand $d^f(t)$ of a feeder f aggregates a large group of consumers, i.e. a few thousands. The CIS provides general features on these consumers, i.e. annual

consumption, type of contract, and contracted power, which can be used to group them into K different categories, e.g. one residential category of all the residential consumers. The sum of all the individual yearly consumption form the annual category consumption c_k^f. We then obtain the category share

$$p_k^f = \frac{c_k^f}{\sum_{k=1}^{K} c_k^f} \in (0,1). \tag{3.4}$$

It is important that the number of feeders F in the dataset is larger than the number of categories K. Empirically, it was observed that the condition $F > 5K$ is preferable in order to obtain various category mix, and thus more precise results. One should keep a reasonably low K for three reasons: (i) to obtain a robust profile, (ii) to avoid an excessively long computing time, and (iii) to ensure that user privacy is not violated.

3.1.5.3 Optimization Problem

The aim is to find the elementary profiles $d_k(t)$ from aggregated demand $d^f(t)$ according to Equation (3.3). We write and solve the following optimization problem.

To mathematically formulate this optimization problem, we define a matrix A of size (F, K) whose elements are category proportions p_k^f. Aggregated demands $d^f(t)$ for all feeders and instants $t \in \{1, \ldots, T\}$ are gathered in a matrix X of size (F, T). We are trying to compute demand profile $d_k(t)$ for all categories and instants: these unknown values can be put in a matrix B of size (K, T). It is useful to define β (resp. x), the column vector obtained by stacking rows of B (resp. X) on top of each other. Two constraints limit the values of matrix B:

1. Each component of β is an electricity demand. Since we only examine feeders with electricity consumers exclusively, components must be positive.

2. For each category k, components should have an average unit, i.e. $\sum_t d_k(t) = T$, to have comparable profiles between categories. To write this constraint in mathematical terms, we define the column of length K, $u = (1, \ldots, 1)^\intercal$, and the column of length T, $v = (T^{-1}, \ldots, T^{-1})^\intercal$ in order to write the average unit constraint, with a Kronecker product $\otimes$, as $(I_K \otimes v^\intercal)\beta = u$.

The optimization problem then writes

$$\begin{aligned} \min_{\beta} \quad & \|x - (A \otimes I_T)\beta\|^2 \\ \text{s.t.} \quad & \beta \geq 0 \\ & (I_K \otimes v^\intercal)\beta = u \end{aligned} \tag{3.5}$$

An ADMM algorithm (Boyd et al., 2011) is implemented to recursively solve problem (3.5):

1. minimize the function with the equality constraint by employing the augmented Lagrangian method,
2. retain only positive components to satisfy the positivity constraint,
3. adjust a penalty variable balancing positivity and the minimization.

The algorithm is implemented with the **R** language. Special care is taken on the first step, since the minimization requires inverting a large matrix of size $K(T+1)$. With advantageous formulation and use of Kronecker product rules, only a matrix of size K is to be inverted, dividing the number of flops by approximately T^3. Details about the algorithm can be found in Appendix C.

3.1.6 Case Study

3.1.6.1 Data description

In this case study, we use electricity feeder demand measured every 10 minutes in 3 geographical regions in France. Data come from the main French DSO, Enedis. The three regions encompass a large French city and its surrounding countryside. The three cities are Blois, Lyon and Rennes. Each region is divided into around 500 feeders, and each of these feeders provides electricity for about 1,000 customers. For each feeder, the demand during 4 years are collected, from 2010 to 2013. We discard some feeders because the measurements are too scarce and their overall quality is not sufficient. This can result from database errors or from network reconfiguration or physical injuries on the grid (Goude et al., 2014). Ultimately, between 200 and 400 feeders are selected for each region.

3.1.6.2 Category Mix

For an efficient disaggregation, the category mix among the feeders of a single region must widely vary. Such variation comes from the local disparities: for instance, there are more restaurants in a city center than in a rural area, hence different restaurant shares.

Figure 3.2 sets out four different categorizations in Lyon, based on information from the CIS. The first categorization divides the total energy into two groups: residential and tertiary. The second splits the tertiary into 7 categories to make a total of 8 categories, i.e. residential, agriculture, commercial, public equipment, office and hospital, industry, restaurant and hotel, and medium-voltage (MV) customers (e.g. large buildings that have a specific contract with the operator). A 9-group division results from splitting the residential share into two groups: base tariff and special tariff[1]. Finally, an even more precise categorization, i.e. 12 groups, is proposed. Commercial buildings are split into 2 categories reflecting low and high annual consumption. Similarly, MV consumers are divided into 3 groups: low, medium and high. The category heights for a category k represent the shares across the feeders for a given category, $m_k = \frac{1}{F}\sum_{f=1}^{F} p_k^f$.

3.1.6.3 Profiles

As previously described (see Figure 3.1), we disaggregated the electricity demand in order to recover a load profile $d_k(t)$ for each category $k \in \{1, \ldots, K\}$. The number of overall categories depends on the customer categorization: 2, 8, 9 and 12 categories were tried out (see Figure 3.2). A total of 12 datasets is formed — for each region: Blois, Lyon and Rennes; and for each year: from 2010 to 2013 — and separately used as input into matrix X in problem (3.5).

Figure 3.3 presents the category elementary profiles obtained for $K = 9$ with only 4 categories shown: commercial, public equipment, restaurant and hotel, industry. Profiles are computed with the demand dataset of Lyon in 2011. Profiles are presented for a typical weekday ($24 * 6 = 144$ values, once every 10 minute). Since we have normalized the data, the variations around the average weekly consumption are displayed.

[1]Special tariff charges less during fixed off-peak periods, i.e. during the night, but more during peak hours.

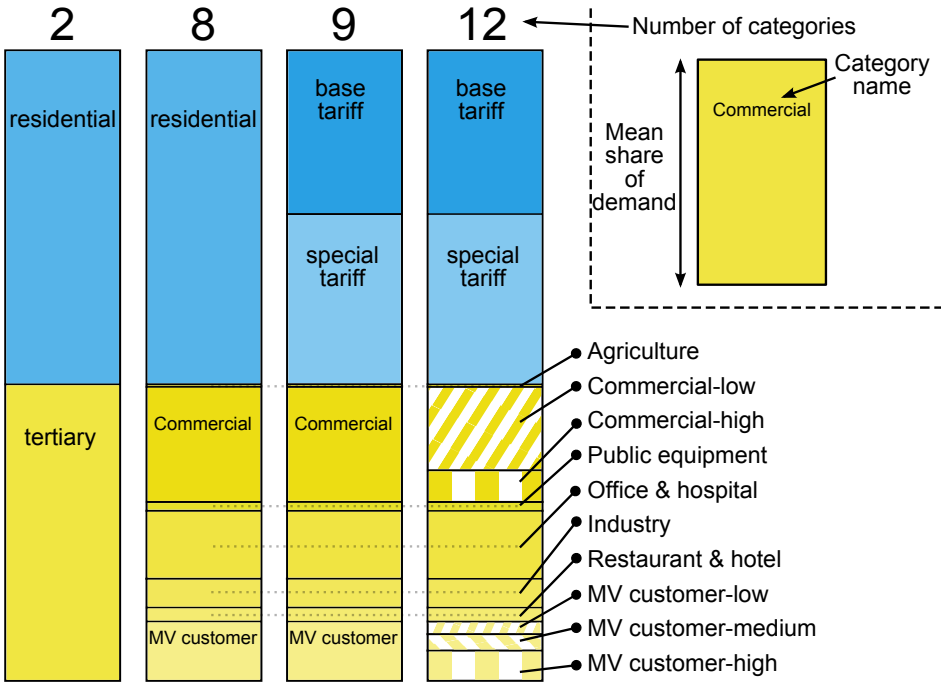


Figure 3.2 – Example of different categorizations (in 2, 8, 9 or 12 groups) for Lyon. There are $F = 320$ feeders. The height of a division shows the mean share of the category in all feeders in the dataset.

Different effects are noteworthy, e.g. the electricity consumption of commercial buildings increases by around 75% during working hours, and decreases by 50% during the night. Conversely, the consumption of public equipment (mainly public lighting and lifts) greatly increases at night.

3.2 Usage of Elementary Profiles

3.2.1 Simulation of a New Feeder

3.2.1.1 Illustration

With the computed elementary profiles $d_1(t), \ldots, d_K(t)$ during a given period $t \in \{1, \ldots, T\}$, associated to categories $1, \ldots, K$, one simulates the demand that a feeder f^* with a given category mix, $p_1^{f^*}, \ldots, p_K^{f^*}$. The simulated demand at instant t is

$$\hat{d}^{f^*}(t) = \sum_{k=1}^{K} p_k^{f^*} d_k(t). \tag{3.6}$$

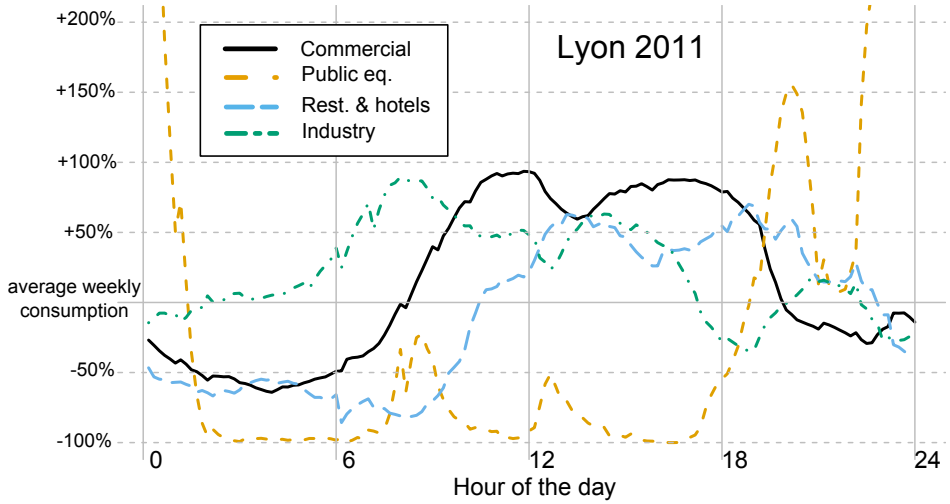


Figure 3.3 – Weekday profiles of 4 different categories computed with the algorithm (9 overall categories) using aggregated consumption data relating to Lyon in 2011. Plots represent the variations around the average weekly consumption and not absolute consumptions.

Note that the category mix is the online information required for this simulation: no need for historical demand data specific to the feeder f^*.

We show a simulation example in Figure 3.4. Demand is simulated during a period of 3 days with only two categories: residential (green area) and tertiary (orange area). In this case, the category mix is: 75% residential and 25% tertiary. The respective contribution of the two categories at each time step is clearly observable on the aggregated demand. An actual demand curve of a feeder with such a mix is superimposed in black.

3.2.1.2 Accuracy

We evaluate the accuracy of the simulated demand with a leave-50-out approach. For a fixed region and year, we randomly select a set of $F - 50$ feeders to perform the disaggregation in K elementary profiles, with $K \in \{2, 8, 9, 12\}$. With the resulting profiles, we use the category mixes of the remaining feeders to simulate 50 demand time series for the $T_0 = 52,560$ 10-minute intervals of the year. The simulated demand is then compared to the real demand with the Mean Absolute Error, i.e. for $f = 1, \ldots, 50$,

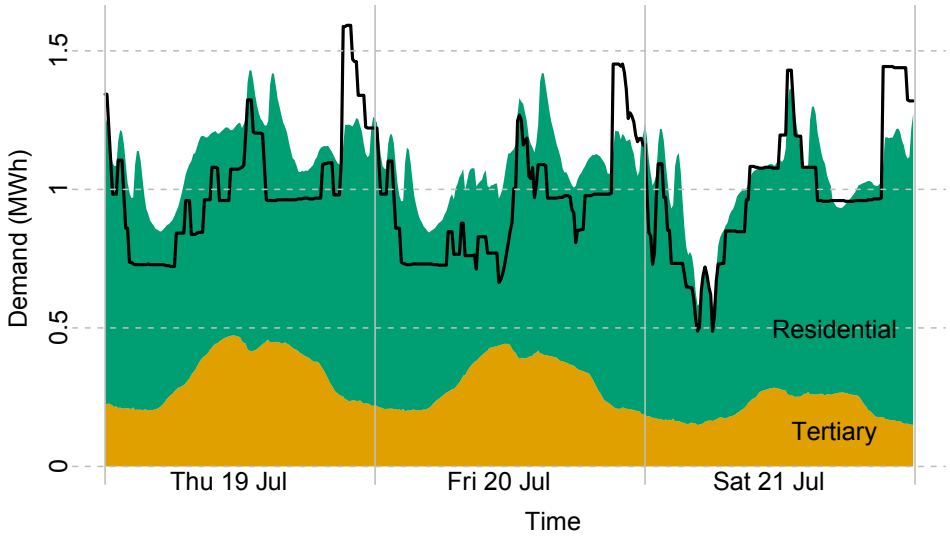


Figure 3.4 – Simulation for one feeder. The profiles were obtained using demand data from Blois for 2012. The black line represents the actual consumption of the unknown feeder (not used in the training dataset). Our algorithm obtained two profiles: the orange part represents the tertiary demand and the green part the residential demand.

and

$$\mathrm{MAE}_f = \frac{1}{T_0} \sum_{t=1}^{T_0} |d^f(t) - \hat{d}^f(t)|. \tag{3.7}$$

The mean $\mathrm{MAE} = 1/F \sum_{f=1}^{F} \mathrm{MAE}_f$ among the 50 left-out feeders is computed. The whole process is repeated with 10 times so as to smooth out extreme feeders by changing the left-out subsets. One run takes roughly 16 hours for each one of the 12 datasets, and the 4 categorizations proposed.

Table 3.1 reports the average MAE and its empirical deviation across runs for the Blois, Lyon and Rennes during the 4 years for different numbers of categories. As a reminder, with consumption normalization, average consumption is dimensionless and equal to 1 (see Section 3.1.4). Hence, the MAE reported is also dimensionless, and is expressed as a percentage.

3.2.1.3 Category Choice

Errors strongly depend on the regions: errors around 15% in Blois, 12% in Lyon, and 15% in Rennes. In fact, the errors depend on the specificity of each feeder. For a feeder

Region	Year	2 categories	8 categories	9 categories	12 categories
Blois	2010	16.5 (0.3)	16.0 (0.4)	16.1 (0.1)	**15.7 (1.3)**
	2011	16.2 (0.3)	15.6 (0.2)	15.5 (0.1)	**15.4 (1.3)**
	2012	15.6 (0.1)	**14.9 (0.1)**	**14.9 (0.2)**	**14.9 (1.0)**
	2013	14.8 (0.2)	14.5 (0.1)	14.6 (0.3)	**13.8 (0.2)**
	Average	15.8 (0.2)	15.3 (0.2)	15.3 (0.2)	**15.0 (0.9)**
Lyon	2010	13.1 (1.2)	13.1 (1.2)	**12.3 (1.0)**	12.9 (1.1)
	2011	11.8 (0.4)	11.6 (0.6)	11.5 (0.3)	**11.1 (0.9)**
	2012	11.9 (0.4)	12.5 (2.1)	**10.9 (0.6)**	12.0 (1.4)
	2013	**10.3 (0.7)**	10.9 (0.4)	11.1 (1.5)	10.7 (0.6)
	Average	11.8 (0.7)	12.0 (1.1)	**11.5 (0.9)**	11.7 (1.0)
Rennes	2010	15.2 (1.4)	**15.0 (0.8)**	**15.0 (0.6)**	15.2 (0.7)
	2011	16.5 (1.2)	16.4 (0.5)	**15.8 (1.3)**	16.8 (0.9)
	2012	15.5 (1.1)	15.2 (0.6)	**15.0 (0.5)**	16.2 (0.9)
	2013	16.4 (1.0)	**15.1 (0.8)**	15.5 (0.8)	16.6 (1.2)
	Average	15.9 (1.2)	15.4 (0.7)	**15.3 (0.8)**	16.2 (0.9)

Table 3.1 – Accuracy of the simulated demand for the 3 different regions over the 4 years with a different number of categories. The simulation is run 5 times. We report the average MAE and its standard deviation among runs between parentheses. The best results over the 4 numbers of categories are written in bold.

f, we define a variation by computing the norm 1 of the feeder demand

$$V^f = \frac{1}{T}\sum_{t=1}^{T}|d^f(t) - 1|. \tag{3.8}$$

This variation reflects how much the feeder demand fluctuates with time. A very smooth (resp. erratic) curve has a low (resp. great) variation. Figure 3.5 represents the feeder MAE (y-axis) in regard of its variation (x-axis) for the three regions. We logically see that feeders with high variation are more difficult to model. We compute the average total variation of all feeders in each dataset (year and region) and reports the value (expressed in %) in Table 3.2. This table in this line with the results in Table 3.1, i.e. when the variation of a dataset is lower, so is the error, e.g. the Lyon

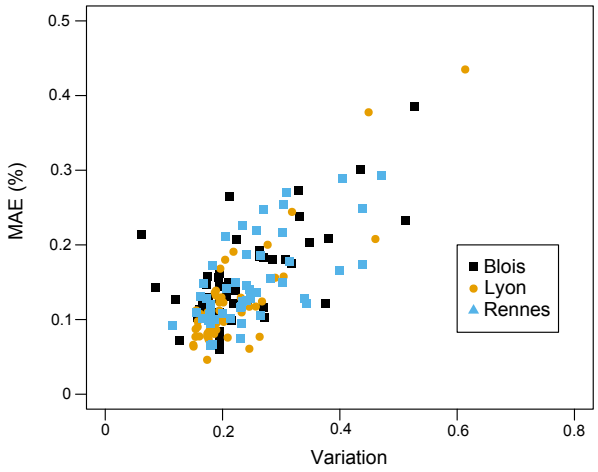


Figure 3.5 – Accuracy of the model (MAE on y-axis) for each feeder depending on the V of the corresponding feeder.

dataset compared to the others.

V (%)	Blois	Lyon	Rennes
2010	23.5	22.0	25.1
2011	23.5	21.3	24.0
2012	22.7	21.9	23.4
2013	22.1	20.7	25.2

Table 3.2 – Average feeder demand variation by dataset.

The question of the optimal number of categories is complex. On one hand, adding categories helps modeling complex demand dynamics. Consequently, using 12 categories outperform other categorization when the total variation of the feeder is large. On the other hand, using too many categories when the variation is low leads to overfitting: a 12-category scheme performs poorly for simple demand time series. On average, the 8- and 9-category schemes are the most efficient for all variations observed.

In addition to the impact of the variation of the feeder demand, the category mix also has an impact on performance. When the category mix of a feeder f is very different from the average category mix, with which the disaggregation in elementary profiles is made, then the simulation demand of this particular feeder is rather inaccurate. We note m_k the average share of the category k among the feeder with which

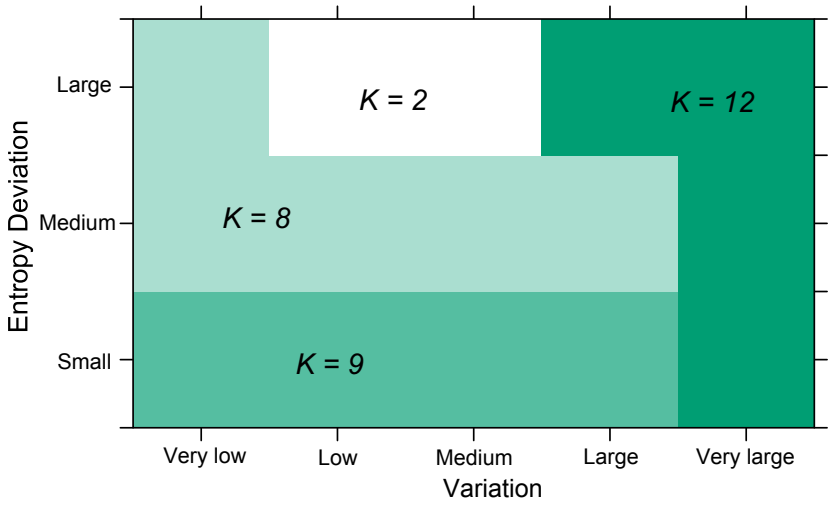


Figure 3.6 – Diagram indicating what is the most efficient categorization to use depending on the feeder demand variation V (x-axis) and the entropy deviation from the training set used for the disaggregation in elementatory profiles (y-axis).

the disaggregation is performed. The set $m_1, \ldots, m_K$ then forms the average category mix. The information entropy of this average category mix writes

$$H_0 = -\sum_{k=1}^{K} m_k \log(m_k). \tag{3.9}$$

The entropy of a specific feeder f writes

$$H^f = -\sum_{k=1}^{K} p_k^f \log(p_k^f). \tag{3.10}$$

When the entropy deviation $|H_0 - H^f|$ is great, it means that the feeder f has a category mix very different from the average, meaning that its demand is more difficult to model.

In Figure 3.6, we represent a diagram showing the most efficient categorizations to use according to the characteristics of the feeder to simulate, as observed with the Lyon 2011 dataset. The variation of the feeder demand is divided in 5 groups defined from the observed values in the dataset (x-axis); the entropy deviation from the average entropy is divided in 3 groups defined from the observed values in the dataset (y-axis). When the feeder demand is complex, i.e. large V, complex categorization is better to capture the complex the dynamics. When the variation is lower, the optimal number of categories depend on the entropy deviation. The small is this deviation,

the large should be the number of categories K. This is logical: when the feeder demand to simulate has a very different category mix compared to the dataset used for the disaggregation, then one should rely on a simple, and thus robust, categorization. Unfortunately, the variation of a new feeder is not known, and so one should simulate with $K = 2$, 8, or 9 categories depending on the category mix, namely on its entropy deviation, to perform the simulation.

Some studies question the categorization of consumers based on the information of the CIS since these information are often incomplete and do not reflect the electricity demand of the consumers (Chicco et al., 2006).

3.2.1.4 Comparison to Similar Works

Our method is less accurate than middle-term forecasting methods at this aggregated scale, relying on historical measurements. Such framework lead to NRMSE between 7 to 10%, or NMAE around 5% see e.g. (Boroojeni et al., 2017), (Goude et al., 2014). The discrepancy with our results — we roughly find twice this error, comes from the exact framework. Most models labelled as forecasting model use historical data in their framework, which is not our case. Consequently, the feeders on which authors evaluate accuracy of their middle-term forecasts are not "new".

Framework of Andersen et al. is more similar to ours (Andersen, Larsen, & Gaarde-strup, 2013). This presents "a model calculating local consumption by categories of customer with specific consumption profiles and different weights in local areas". Unlike us, their profiles are obtained by clustering representative smart-meter measurements, i.e. a bottom-up method. Their results from simulating local areas without using past measurements are expressed with R^2 value and are between 0.95 and 0.56 (their mean R^2 is 0.84). In their case study, the mean consumption of areas is 55.3 MW while in our case, for a given feeder it is between 0.5 and 7 MW. In order to compare our method with their method, we aggregated our areas to obtain similar average power levels and computed the R^2 between prediction and measurements. The results are shown in Table 3.3. The performance of our method is slightly better than Andersen et al.'s method in the Lyon and Rennes datasets, and similar in the Blois one.

Area	Avg. demand 2010 (MW)	R^2
Blois	31.5	0.82
Lyon	46.2	0.88
Rennes	37.4	0.87

Table 3.3 – Coefficient of determination R^2 for different areas showing the predictive performance of our method with a 9-category breakdown. The prediction of a group of 20 feeders is compared to the measured demand of the 20 feeders.

3.2.2 Evolution of the Peak Hour

Before connecting new consumers to a feeder, the DSO has to estimate the future peak demand, i.e. when it occurs and to reach what value. The profiles obtained enable to quantify and forecast the contribution of the new set of consumers to the peak demand.

Let us assume that a feeder f has a category mix $p^f_{1,y_0}, \ldots, p^f_{K,y_0}$ at year y_0. The disaggregation is performed with a large set of F feeders, so as to obtain $d_1(t), \ldots, d_K(t)$. The actual demand observed for this feeder during year y_0 writes

$$d^f_{y_0}(t) = \sum_{k=1}^{K} p^f_{k,y_0} d_k(t) + \varepsilon^f_{y_0}(t). \tag{3.11}$$

The residuals $\varepsilon^f_{y_0}(t)$ depict the specificity of feeder f compared to the others of the set, i.e. explaining the remaining 15% errors previously assessed. If one anticipates an evolution of the category for year y_1, i.e. new category mix $p^f_{1,y_1}, \ldots, p^f_{K,y_1}$, then the elementary profiles and residuals provide a precise estimation

$$\hat{d}^f_{y_1}(t) = \sum_{k=1}^{K} p^f_{k,y_1} d_k(t) + \varepsilon^f_{y_0}(t). \tag{3.12}$$

Figure 3.7 depicts the peak change obtained with this formula in the case of different evolutions for both offices and special-tariff residential consumers. In this case study, the considered feeder is from the Lyon region and has the following distribution of customers: 30% commercial, 15% offices, 30% basic residential and 20% special special-tariff residential. The initial peak occurs at 12:10 and is 650 kW. The profiles used are taken from the 9-category breakdown. We quantify the influence on the peak value (black lines with value added to the initial peak value, per 50 kW) by adding

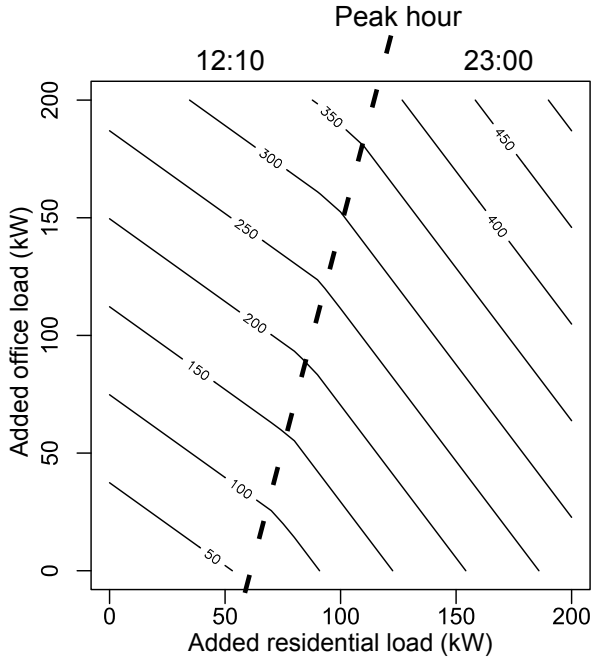


Figure 3.7 – Contour plot representing the load added to the peak value when on adds office consumers (y-axis) or residential consumers with special-tariff (x-axis).

an office category load (y-axis) and a special-tariff residential load (x-axis). We also depict the evolution of the peak hour (black dashed line). Adding offices contributes to increasing the 12:10 peak, whereas the residential load increases the 23:00 peak, which corresponds to the start of the special-tariff period.

This is an illustration of an application of the method that can for example help decision-makers to choose between two projects (offices or a new residential area) and quantify the impact on the existing feeder demand.

3.3 Conclusion

3.3.1 Disaggregation

The model proposed perform the disaggregation of the electricity demand measured at the feeder level to obtain elementary profiles. It was assumed that all feeders aggregates the same elementary profiles, although in various shares, that determine its demand dynamics. The profiles are optimally found by minimizing prediction errors in a novel

ADMM adapted to our demand disaggregation case.

Unlike bottom-up methods that require individual demand curves, our method only requires several feeder demand curves and a description of the consumers. One of the advantages of using aggregated measurements on a set of individual load curves is that they are fully representative.

The method has been applied in a case study comprising three regions in France, with around 300 available feeder measurements over 4 years per region. The elementary profiles describe the dynamics of each categories of consumers, i.e. when the category demand is high or low. We have shown that each load profile gathers intrinsic features of the given category.

We first assess the accuracy of the decomposition by (1) performing the disaggregation with a large set of feeders, (2) using the elementary profiles to simulate left-out feeders and their respective category mix, and (3) comparing the simulated demand with the actual ones. The accuracy of our method performs similarly or better than a bottom-up method in the literature to predict a new local area. Different categorizations are proposed and the respective advantages are drawn up: simulating the demand of a feeder with atypical category mix should be made conservatively with a small number of categories.

Secondly, we see how the usage of the elementary profiles can be used to anticipate the evolution of the peak demand of the day through the years. This evolution reveals both in term of peak value and peak timing. In our example, the addition of offices consumers or special-tariff residential consumers impact differently the peak: the former shifts the peak timing around noon, while the later shifts it around 23:00.

An improved framework can be developed to select the optimal categorizations for the disaggregation depending on the feeders' characteristics to simulate. However, as pointed out in other studies, the information brought by the CIS are not optimal to create meaningful consumer categories. It is believed than the addition of socio-demographic statistics, such as the income of the consumers, should provide more efficient categories. Unfortunately, such statistics do not currently exist at the feeder level.

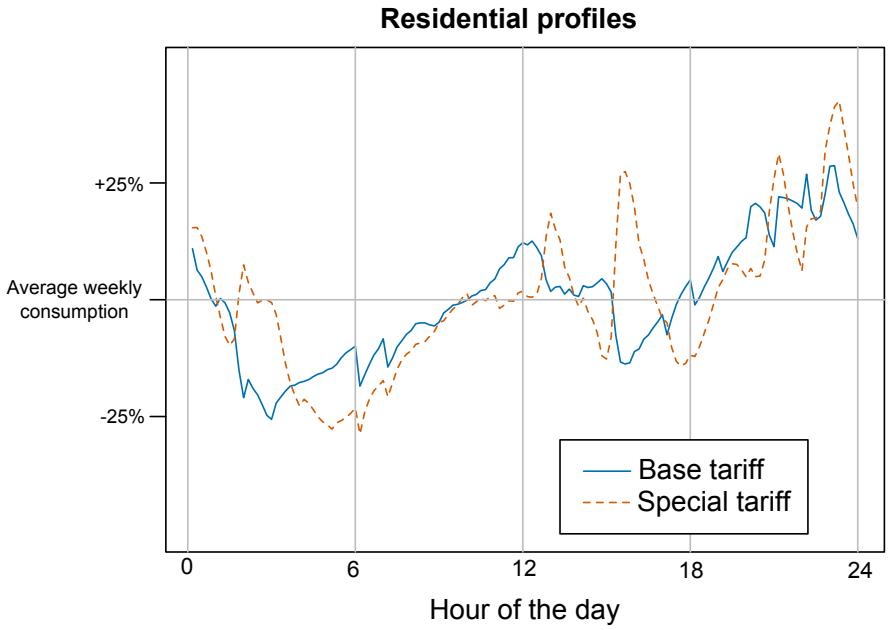


Figure 3.8 – Two residential profiles obtained from the Lyon 2012 dataset: with base tariff (blue solid line), and with special tariff (red dashed line) on a typical weekday.

3.3.2 Residential Demand Profile

The elementary profiles obtained from the feeder demand obtained reflect the average demand of the elementary category. However, since the average does not exist (Rose, 2016), these profiles not provide a description of the individuals. This is notable for the residential category (visible in Figure 3.8) which accounts for about half of the feeder demand, but the average profiles are unrealistic,. In particular, the variation among the residential consumers are smoothed in these profiles. Yet:

- the influence on the tariff signal is noticeable on the average profile, e.g. the peak demands are visible at 16:00 or 23:00 when electricity is cheaper for special-tariff consumers. This influence is more subtle at the individual level.

- The intraday variability is greatly underestimated in the profiles: on average, the demand slightly fluctuates between -25% (during the night) and +25% (in the evening) around an average demand value. This comes from the aggregation since, in fact, the typical load factor of an individual is below 0.5, meaning that his or her electricity demand goes from simple to double throughout the day. This extreme variation is non-visible at the feeder level due to the weak correlation

among user.

The average profiles are useful from an aggregated point-of-view and general decision-making. However, for emerging applications, such as demand response, the individual demand needs to understood in details. This requires individual measurements and the designing of forecasting model specifically devoted to the household level.

Chapter 4

Household Electricity Demand Forecasting

Summary Regular measures of household electricity demand are now obtained with smart meters, with resolution of thirty minutes or one hour. The characteristics of the corresponding time series differ from those of electricity demand at larger scale, e.g. the feeder demand characteristics. We analyze in details these characteristics in Section 4.1. We note that the random behavior of the individuals has a prominent impact on the demand. The specific characteristics of the household demand challenge the traditional short-term forecasting models. In Section 4.2, we introduce a gradient boosting model for the next-day household demand. We detail how this model operates and its performance on three datasets. On average, the relative errors are of 28%. Since such errors are quite large, we emphasize the need of a probabilistic framework that quantifies forecasting uncertainty by adapting the standard gradient boosting model. However, the significant uncertainty impedes the emergence of business models at this household scale. In Section 4.3, the forecasting performance is compared at different levels of aggregation and time resolution. Our experiments show that the optimal aggregation level is around 15 households and that the forecasting errors increase by about 25% when forecasting demand averaged every 15 minutes rather than every hour. The gradient boosting model cannot always operate in practice. Consequently, in Section 4.4, we introduce an additive model along with a hierarchical forecasting framework. The framework has been implemented on a real case in an online demonstration project. We report the online performance and analyze it in regard with tests

made offline.

Résumé Les compteurs intelligents mesurent la demande électrique d'un ménage à intervalles réguliers, généralement toutes les 30 minutes ou toutes les heures. Les séries temporelles obtenues ont des caractéristiques très différentes de celles faites à plus grande échelle, p. ex. au niveau des départs. Nous comparons dans le détail ces caractéristiques dans la Section 4.1. En particulier, nous remarquons l'importance cruciale des comportements individuels sur le niveau de la demande. Ces caractéristiques spécifiques à l'échelle d'un ménage font de la prédiction à court terme un défi complexe. Dans la Section 4.2, nous proposons un modèle de type *gradient boosting* pour effectuer des prédictions. En moyenne, l'erreur relative est de 28% quand on prédit la demande électrique d'un ménage pour le jour suivant. Comme cette erreur est importante, nous adaptons le modèle pour fournir des prévisions probabilistes qui quantifient l'incertitude que l'on a pour les futures valeurs. Néanmoins, l'incertitude qui demeure reste un frein pour le développement d'applications spécifique à l'échelle d'un ménage. Dans le Section 4.3, nous étudions l'évolution de la performance de prédiction quand on change de niveaux d'aggrégation et d'échelles temporelles. Nos tests montrent qu'il est optimal de prévoir la demande agrégée d'un groupe de 15 maisons à la fois, et que l'erreur de prédiction augmente d'environ 25% quand on prédit la demande moyennée toutes les 15 minutes plutôt que toutes les heures. Ce modèle de *gradient boosting* ne peut pas toujours être utilisé en pratique. Par conséquent, un modèle additif ainsi qu'une structure de prédiction sont présentés dans la Section 4.4. Cette structure a été implémentée sur un cas pratique et fonctionne en temps réel. Nous analysons la performance obtenue en pratique et celle obtenue lors de tests en laboratoire.

4.1 Characteristics of Household Electricity Demand

In the following section, we analyze the characteristics exhibited by household electricity time series. Specifically, we want to understand the profile, i.e. the temporal shape, of the time series.

4.1.1 Data

4.1.1.1 Electricity Demand

Electricity demand is strongly dependent on where the electricity is used (Nejat et al., 2015). It has been observed that the electricity consumption is strongly correlated with the economic growth (Wolde-Rufael, 2006): industry requires more electricity for their operation, infrastructure are modernized with electric devices and so on. The overall wealth increase is also reflected in the increasing residential consumption: inhabitants have larger houses with more electric devices either for comfort (heating and cooling devices, household appliances, etc.), or entertainment leisure (television, computers, etc.). Since electricity demand measurements require mature electricity infrastructure and efficient meters, the data used is from 3 rich countries: France, Portugal, and the United States of America:

- In the French dataset (not publicly available), hourly demand of 176 residential houses have been recorded between January and March 2015. The buildings are in a rural neighborhood located in Tours, Centre-Val de Loire region. A single feeder is specifically devoted the electricity delivery of the neighborhood.
- In the Portuguese dataset (not publicly available), hourly demand of 226 buildings have been recorded during year 2015. All the buildings are located in one neighborhood in the vicinity of Évora, in southern Portugal. Most of the buildings are individual residential houses, but a few of them are SMEs (a mini-market, a few restaurants, a small factory etc.). The same feeder delivers electricity to the whole neighborhood.
- In the USA dataset, hourly demand of 175 residential households have been recorded during year 2017. Data is freely available for research purposes in the frame of the Pecan Street Inc. project (*Pecan Street Inc. Dataport*, 2018). Most

buildings are individual houses, with a small number of apartments, and are located in the city of Austin, Texas. Inhabitants voluntarily signed to be part of the research project: they are climate conscious, and 80% have photovoltaics panels. The measurements are made such that electricity load is drawn from the network or from the panels, and so the time series exactly reflects the household electricity demand.

For each location, a time series summing all the household demand is made up to define a "feeder" time series[1].

Additional data is retrieved to compare the characteristics of the household demand with demand at a larger scale. Part of the data introduced (948 time series) in Chapter 3 is used to analyze demand at the feeder level. The French TSO, RTE (Réseau de Transport d'Électricité), freely publishes the electricity demand made at the regional scale and national scale (Réseau de Transport d'Électricité (RTE), 2018a). The Texas DSO, ERCOT (Electric Reliability Council of Texas), freely publishes the electricity demand made in 8 weather regions of Texas (Electric Reliability Council of Texas (ERCOT), 2018a).

All the time series are pre-processed to remove negative values and absurdly high values, defined as when a value is 10 times higher than the average value of the time series. When data are missing, the values are either linearly interpolated (when the period of missing data is less than 5 hours), or labeled as `NA` and not used in the subsequent analyses.

4.1.1.2 Outside Temperature

Since outside temperature has a strong impact on electricity demand, we retrieve the temperature time series corresponding to the location where each electricity demand time series are measured.

The European Centre for Medium-Range Weather Forecasts (ECMWF) provides this data. Two kinds of temperature values are retrieved: the values forecast at 12:00 for the next day, and the exact temperature measured. Since the forecasts and measured are only once for a time period of 3 hours, series are linearly interpolated to obtain

[1] Although it is a real physical feeder in France, it only a part of the feeder in Portugal, and is a virtual feeder in the USA since households are spread across the city of Austin.

an hourly time series. The exact locations of the Portuguese, French, and American neighborhoods are used to obtain a single temperature time series for all the households of each neighborhood, denoted $(\hat{\theta}_t)$ for the day ahead forecasts and (θ_t) for the measures. Additional temperature time series are retrieved to obtain temperature time series related to various French and American regions.

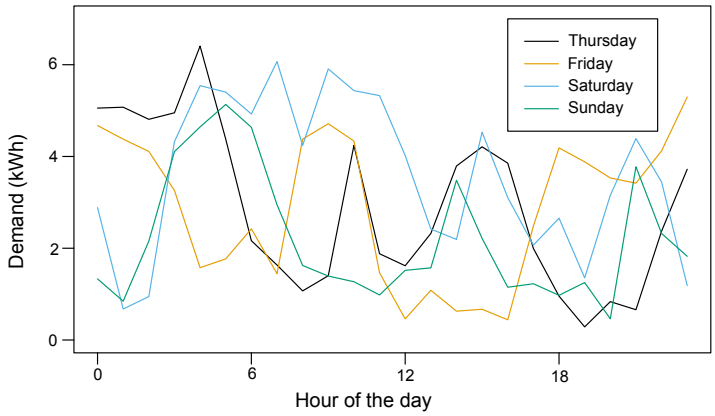
We also construct a climatology time series $(\hat{\theta}_t^0)$ for each location. It is made by computing an average temperature profile and repeating it for every day of the year. The level of this profile is adapted according to the month of the year, so that temperatures are higher in July than in January.

4.1.2 Characteristics Analysis

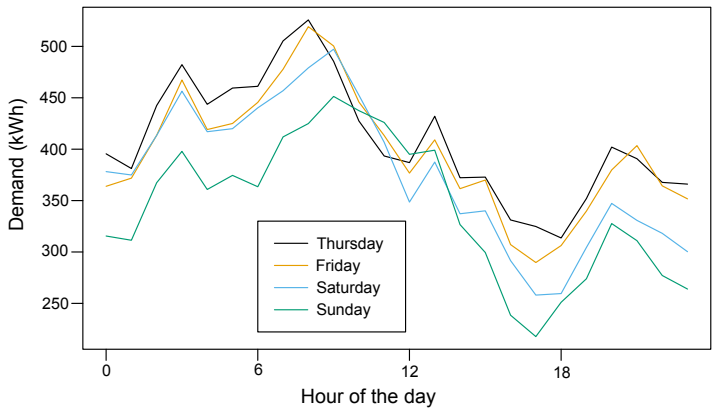
The electricity demand recorded at the household level exhibits specific characteristics that are challenging for modeling the time series. Figure 4.1 depicts the hourly electricity demand of the French dataset during 4 successive days in March 2015 (Thursday 5 to Sunday 8). The top graph represents demand made by an individual household, while the bottom one represents the aggregate demand of the whole neighborhood (176 residential buildings). For the household, the daily time series is highly volatile. The profile shapes have a lot of peaks and valleys that succeed erratically. The daily profiles widely change on successive days, so that trends and patterns are hardly visible on a quick inspection. Conversely, the shape of the neighborhood demand is smoother. The four profiles plotted have roughly the same patterns: high level during the night and morning — due to scheduled cycles of large electric appliances —, a valley in the afternoon and a higher consumption when people are back home in the evening. In particular, some peaks are clearly apparent, e.g. at 13:00, describing a recurrent behavior in the neighborhood, easily explainable: most people use their cooking devices precisely at that time.

Since the neighborhood demand is the sum of household demand, such as the one represented in Figure 4.1a, the resulting profile is statistically smoothing out the erratic household profile. Therefore, the underlying patterns hidden at the household level emerge at the neighborhood level. This *aggregation effect* eases the demand modeling and leads to better forecasting performance.

Researchers have tried to quantify this aggregation effect and the consequences for



(a) Household



(b) Neighborhood

Figure 4.1 – Demand time series of 4 successive days in March 2015. Figure 4.1a depicts the electricity demand of a household near Tours, France, and Figure 4.1b depicts the aggregation of electricity demand in the close neighborhood (total of 176 residential buildings).

forecasting performance, such as Humeau et al. (Humeau et al., 2013) who represent the forecasting error as a function of the number of households considered in the aggregation. With the same idea, Sevlian and Rajagopal (Sevlian & Rajagopal, 2014) propose a mathematical formulation of the forecasting error as a function of the average power demand. They notice that the forecasting errors strongly vary between two households, and advocate for analyzing aggregation effect relatively to the average power demand rather than the number of households.

In the following, we provide illustration that the characteristics of demand time

series evolve with the average power to identify forecasting challenges at the local scale. Five aspects of the demand time series are studied: the smoothness, the hourly spread, the regularity, the temperature influence, and the vacation influence.

4.1.2.1 Daily Smoothness

The smoothness of the demand curve is a visual concept. For continuous function, it is linked to the concept of differentiability: a smooth function is infinitely differentiable. It does not apply for discrete function such as time series, and the concept is more vague. In the electricity power field, the related concept of *load factor* is prominent. During an address before the Finance Forum of the Young Men's Christian Association in 1914, Insull argues for the centralization of energy supply (Insull, 1914). He explains that electricity suppliers should diversify their customers so electricity to produce is smooth over time, reducing the production costs (infrastructure and power plants are fully used at all times). His demonstration stems from the study of the demand curves for different type of users: department stores, office buildings, steel factories, cement works and so on. In particular, he makes use of the load factor to illustrate his point. It refers to the ratio between the mean demand of a time period, e.g. one day, and the peak demand observed during the day. Mathematically, let (y_t) be a demand time series, then the load factor of period $\mathcal{T}$ is

$$\mathrm{LF}_{\mathcal{T}}(y) = \frac{\mathrm{mean}_{\mathcal{T}}\, y_t}{\max_{\mathcal{T}} y_t}. \tag{4.1}$$

With this definition, the load factor is a dimensionless value in the interval $(0, 1)$. The higher it is, the smoother is the demand curve, and when the load factor is 1, the demand is constant throughout the period. The lower it is, the more peaky is the demand curve. In general, the operators prefer a high load factor.

Mean daily load factors are calculated on the data described in Section 4.1.1, and plotted in Figure 4.2. Each dot represents the mean daily load factor for a specific time series plotted versus the average power of the series. The solid black lines represent linear regression for quantile levels 10, 50 and 90%, exhibiting that the daily load factor increases with the average power. As expected, the demand curve is smoother at aggregated level (load factor up to 0.8 for average power over 1 MW) than at the household level (average load factor around 0.5). Let us also note that load factor values are widely spread for low average power. The spread is partially due to the use

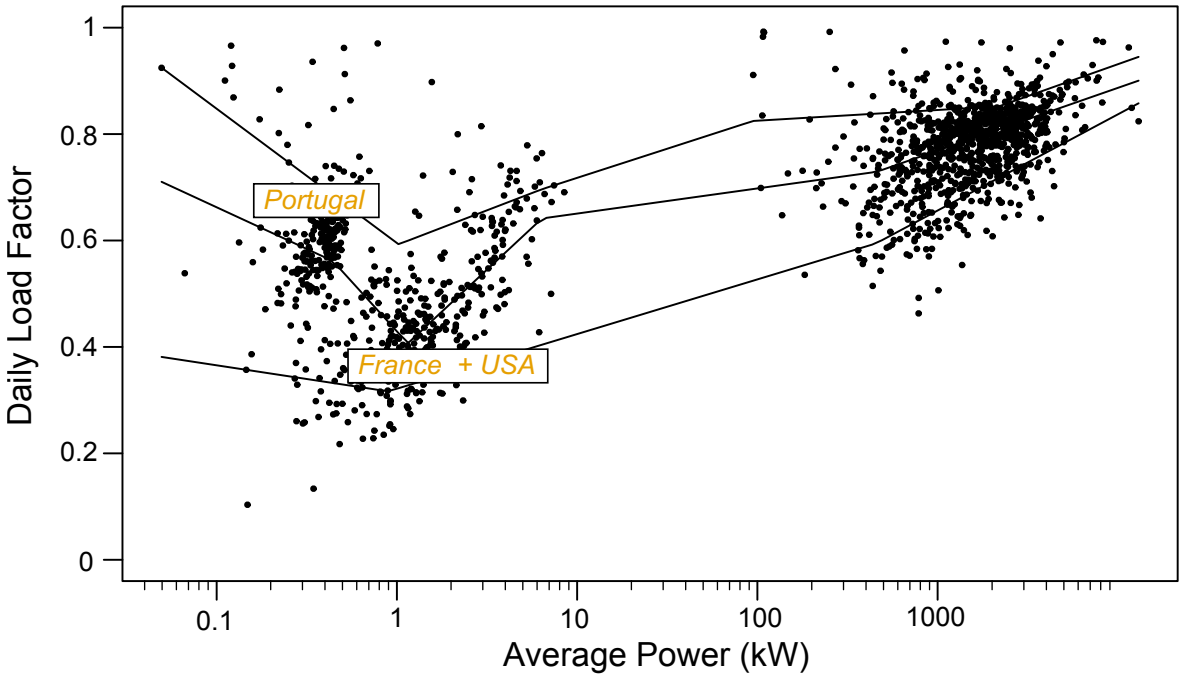


Figure 4.2 – Scatterplot of the mean daily load factor (y-axis) against the average power of a time series (x-axis with a logarithmic scale). Household demand time series from Portugal and from France + USA (with average power below 10 kW) are clearly separate. Black lines represent linear regression for quantile levels — 10, 50 and 90% —, they show that the load factor gets closer to 1 for higher average power.

of several datasets — load factors of Portuguese households are clearly separated from these of France and USA —, but load factors still go from 0.2 to 0.9.

The coefficient of variation (CV) is a statistical measure conceptually close to the load factor. It is equal to the ratio between the standard deviation, noted with function $\mathrm{sd}(\cdot)$, and the mean of the demand during a period $\mathcal{T}$, i.e.

$$\mathrm{CV}_{\mathcal{T}}(y) = \frac{\mathrm{sd}_{\mathcal{T}}(y_t)}{\mathrm{mean}_{\mathcal{T}}(y_t)}. \tag{4.2}$$

Similarly to the load factor, this quantity is a dimensionless positive value but is not upper bounded. When the coefficient is higher than 1 (resp. lower than 1), the series is said to be overdispersive (resp. underdispersive). There is a strong negative link with the load factor: a high load factor corresponds to a low CV. But the latter has the advantage to be more robust to absurdly high values that remain in the time series. Figure 4.3 represents the daily CV observed for the time series of the data studied.

The same conclusions are drawn from the graph: the coefficient decreases with average power, and the values are widely spread for low average power.

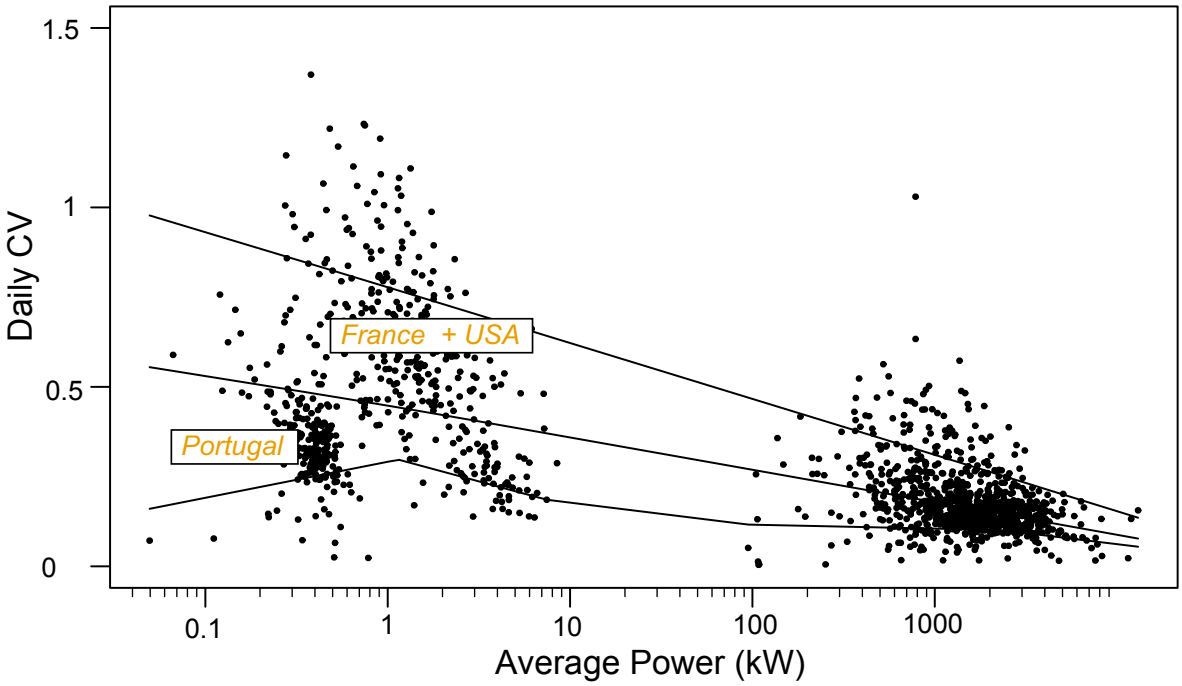


Figure 4.3 – Scatterplot of the mean daily coefficient of variation (y-axis) against the average power of a time series (x-axis with a logarithmic scale). Household demand time series from Portugal and from France + USA (with average power below 10 kW) are clearly separate. Black lines represent linear regression for quantile levels — 10, 50 and 90% —, showing that the coefficient of variation decreases to 0 for higher average power.

Table 4.1 reports the values of the average power, daily load factor (in %) and daily coefficient of variation (in %) with the standard deviation between parentheses. Taken separately, the load factors increase, and the CV decrease, when analyzing larger scale. Moreover, the variation of the indices is rather large for household data.

From this daily smoothness analysis, we conclude that the individual household demand curve is less smooth and more peaky than at larger scale. Furthermore, the smoothness greatly depends on the household considered. Consequently, forecasting models already existing for the large scale cannot be directly used for the households. Secondly, the models' parameters should be tuned specifically to account for the variety of demand, ad opposed to define the same global parameters for all households.

Table 4.1 – Daily smoothness indices.

Type	Location	Size	Av. Power	LF (%)	CV (%)
Household	Portugal	226	0.4 (0.2) kW	61 (10)	36 (13)
	France	176	2.4 (1.8) kW	50 (16)	53 (34)
	USA	175	1.3 (0.7) kW	64 (16)	42 (8)
Feeder	Portugal	1	95 (—) kW	91 (—)	5 (—)
	France	949	1.8 (1.2) MW	78 (8)	18 (9)
	USA	1	231 (—) kW	66 (—)	30 (—)
Region	France	12	4.5 (2.0) MW	88 (2)	10 (2)
	USA	8	5.1 (4.9) MW	85 (3)	13 (3)

4.1.2.2 Periodicity

The non-smoothness of the household demand curve is not in itself an issue for forecasting purposes. The demand curves made by factories are also non-smooth (low to null demand during the night but high constant demand during workdays) but future demand is fairly easy to anticipate thanks to its periodicity. In fact, most electricity demand curve are periodic. There exist clear periodic patterns in the demand time series: daily periodicity, weekly periodicity, and even yearly. These come in part from the day and night periodicity, and in part from the human habits. For instance, inhabitants wake up every day around the same time and then use electrical appliances. The weekly patterns is visible between week days and weekend days: people generally wake up earlier to go to work during the week. The yearly periodicity is caused by the fairly regular weather conditions, daylight duration, and cultural events, such as holidays. Other types of less visible periodicity also exist such as intraday and have been used at the national level (Taylor & Snyder, 2012).

The persistence model is a widely used benchmark for electricity demand forecasting models. It specifically makes use of the periodic patterns that exist in the time series. For a periodicity s, the persistence forecast of the demand at instant t

$$\hat{y}_t = y_{t-s} \tag{4.3}$$

is taken equal to the value of the demand measured at instant $t - s$. The value of s is chosen by the forecaster and depends on the horizon of the forecast: when forecasting

demand for the next day, s should be greater than 24 hours. Intuitively, the most promising value of s is the lowest value possible while accounting for the strongest periodicity of the demand. Hence, for the day-ahead forecast, best value is $s = 24$ hours. However, when doing a very short-term forecasts, for horizon of 3 hours, the choice of s is less clear: should one use the most recent value available $s = 3$ hours or should one prefer the daily periodicity $s = 24$ hours?

Table 4.2 presents the average performance (and standard deviation between time series) of the persistence model for the various periodicity and datasets. The score used for evaluation is the Normalized Mean Absolute Score. For a periodicity s and a demand time series $(y_t)_{t=1,\dots,T}$, it writes

$$\mathrm{NMAE}(s) = \sum_{t=s+1}^{T} \frac{|y_{t-s} - y_t|}{\operatorname{mean} y_t}. \tag{4.4}$$

This score is expressed in % and negatively oriented, i.e. the lower is the NMAE the more efficient is the persistence forecasting model.

Table 4.2 – Forecasting performance (NMAE) of persistence models with various periodicity s.

Type	Location	Size	1 hour	3 hours	1 day	1 week
Household	Portugal	226	31 (16)	36 (12)	34 (13)	35 (12)
	France	176	35 (23)	48 (29)	37 (18)	43 (19)
	USA	175	34 (11)	54 (14)	46 (13)	53 (13)
Feeder	Portugal	1	3 (—)	5 (—)	3 (—)	3 (—)
	France	949	7 (3)	14 (6)	12 (7)	14 (6)
	USA	1	9 (—)	23 (—)	13 (—)	20 (—)
Region	France	12	4 (1)	9 (2)	7 (1)	7 (1)
	USA	8	3 (1)	10 (3)	6 (1)	10 (2)

As expected, for a given periodicity, the performance of persistence model is more efficient at larger scale. Considering the 1 hour persistence model in France, errors go from 35 % at the household level, to 7% at the feeder level, and to 4% at the regional level. Considering a specific location, the 1 hour persistence is the most efficient, closely followed by the 1 day and 1 week. Comparing 1 hour and 1 day efficiency,

one notices that the relative errors reduction is about 10% at the household level, but 40% at the regional level. Therefore, the daily cycle is relatively more important at the household level. For day-ahead forecasting, the 1 hour persistence model is not available, and therefore the 1 day persistence model is the one to use[2]. For intraday forecasting, the pertinence of the very recent demand values however quickly falls: the 3 hour persistence model is far less efficient that the 1 day model. This observation points out that intraday forecasting for average horizons (between 3 and 24 hours) is little influenced by the most recent available demand values. As expected, the performance of the 1 week persistence model is poorer than the 1 day, but causing only a marginal errors increase (around 15%). As expected, the weekly periodicity is quite significant and is a strong indicator of the future demand. In particular, one expects that information given by the 1 week old and 1 day old values are fairly uncorrelated, and thus that using both as inputs provide original information to a forecasting model.

On the other hand, the absolute performance of persistence models at the household level is very poor: average errors are higher than 35%, and above 100% for some households. This benchmark shows that forecasting the demand of a specific household is a difficult task, and reveals that deterministic forecast is of little relevance with such errors. Consequently, one should favor probabilistic forecasts to reflect the inherent uncertainty.

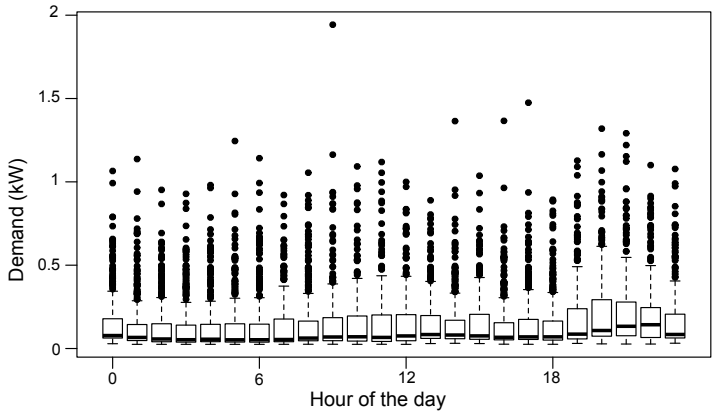
From this periodicity analysis, we conclude that the daily periodicity is very strong for electricity demand, especially at the household level. However, the weekly periodicity is also relevant and should be accounted for when designing forecasting models.

4.1.2.3 Hourly Distribution

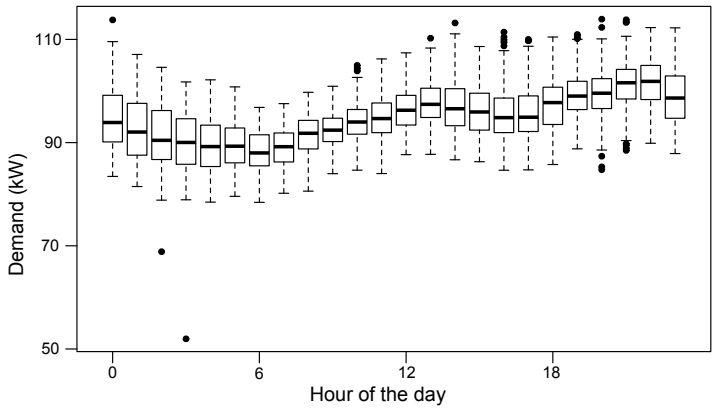
In addition to the daily indices, the statistical distribution of the demand values measured for each hour of the day is noteworthy. Figure 4.4 shows the distribution variation in hourly demand for one household and the neighborhood in Portugal. The graph is a standard boxplot representation: for each hour of the day, the wide horizontal line represents the median demand measured, the rectangle indicates the interquartile range, the dashed line and points indicate a range for values and the points show the detected outliers. We see that the standard definition of outliers (higher than quantile value at

[2]Other cycles were tested, such as 3 days, 2 weeks or 1 month, but errors were much higher than those reported.

level 75% + 1.5 times the interquartile) is not suited for household demand: a lot of values are detected as outliers. Two important features are visible. First, the hourly distribution is more spread for the household than for the neighborhood. This is due to he electric habits of a household which are generally spread in time. Inhabitants usually do no care if they start their washing machine at 17:00 rather than at 18:00. Second, hourly distribution of the household demand is positively skewed, meaning that the upper tail is longer than the lower. This asymmetry is due to the very demand phenomenon: some electric appliances are rarely used but require important electricity power, e.g. dryer, and so are responsible for the peakiness of the demand curve and the asymmetry of the distribution.



(a) Household



(b) Neighborhood

Figure 4.4 – Boxplot of the hourly demand values for a household (4.4a) and a neighborhood (4.4b) in Portugal.

The visual example is generalized for all the time series described in Section 4.1.1

and systematized with statistical indices. Two indices are used: the coefficient of variation (CV) and the nonparametric skew (S). The indices are computed using the demand measured at the same hour $h = 0, \ldots, 23$. First coefficient

$$\mathrm{CV}_h(y) = \frac{\mathrm{sd}_h(y_t)}{\mathrm{mean}_h(y_t)} \tag{4.5}$$

is a dimensionless positive value. Distribution is underdispersive (resp. overdispersive) when the coefficient is below (resp. above) 1. Secondly, the nonparametric skew

$$S_h(y) = \frac{\mathrm{mean}_h(y_t) - \mathrm{median}_h(y_t)}{\mathrm{sd}_h(y_t)} \tag{4.6}$$

is a dimensionless value in interval $(-1, 1)$, positive (resp. negative) when the upper (resp. lower) tail is longer than the other. A null skewness indicates that the distribution is symmetrical. Average CV and S over the 24 hours of the day are computed for each time series data, and the average results are reported in Table 4.3 along with the standard deviation measured on all of the time series. The coefficients of variation

Table 4.3 – Hourly distribution indices.

Type	Location	Size	CV (%)	S (%)
Household	Portugal	226	39 (14)	+10 (14)
	France	176	50 (28)	+14 (14)
	USA	175	74 (18)	+29 (9)
Feeder	Portugal	1	5 (—)	+5 (—)
	France	949	36 (14)	+17 (21)
	USA	1	37 (—)	+30 (—)
Region	France	12	21 (2)	+25 (6)
	USA	8	17 (4)	+20 (8)

of the hourly distributions significantly decrease with the average power. However, perhaps surprisingly, this is not the case of the smoothness which remains more or less constant at every power. There exist factors, common to all the households of a location, that are responsible for a correlation of abnormally high demand on several households at a certain instant, and consequently, the aggregate demand is also extremely high. These common factors are not always identifiable. The weather conditions are a usual cause: for instance, Sunday 30 July 2017 is a really hot day in Austin

— more than 37°C at 18:00 — and so the electricity demand of several households are simultaneously abnormally high — skewing the 18:00 demand distribution — and this skewness is also visible on the neighborhood demand. Cultural influence is also a possible cause: electricity demand on the July 4 2017 is abnormally higher than usual for most household, and thus abnormally higher at the neighborhood level.

From this hourly distribution analysis, we conclude that the hourly demand distributions are widely spread and that these distributions exhibit large positive skew. Consequently, the probabilistic aspect of household demand forecasting is prominent to capture the possible variation and the forecaster should be cautious with probabilistic models. In particular, the direct use of symmetrical parameters is to be avoided.

4.1.2.4 Temperature Influence

The outside temperature has a strong impact on the electricity demand. Most people in developed adjust the temperature inside their home in order to be comfortable, and therefore switch on heating or cooling devices — referred to Heating, Ventilation and Air-Conditioning (HVAC) in the following — when the outside temperature is low or high. Since HVAC devices are not always electrical, the temperature influence on electricity demand depends on the situation. Bessec and Fouqueau (Bessec & Fouquau, 2008) examine this influence on national demand for the countries in the European Union. According to this study, there is a demand increase when the weather is either too cold or too hot, and a plateau when outside temperature is around 16°C. The increase is superlinear, meaning that the electricity demand increases even more for extreme temperature. In Europe, the temperature effect is more pronounced for cold than hot temperatures — air conditioning is rarer than heating. In America, where electrical air conditioning is more widespread than electrical heating, the demand increase is larger for hot than cold temperatures.

The scatterplot in Figure 4.5 represents the hourly demand made in the South Central region in Texas, USA (black points) and in the Grand Est region in France (orange points) versus the outside temperature measured in the region. The two regions are randomly selected among our data but have similar power demand, so as to be represented on the same figure. The temperature effect is clearly visible with high demand values when the outside temperature is low and high. The lowest demand values are observed for intermediate temperature (between 10 and 20°C), while the

South Central demand almost triple when it is more than 30°C. For the French region, the heating effect appears almost linear and the cooling effect is very slight. On the other hand, both the heating and cooling effects are significant the Texan region and these effects look superlinear.

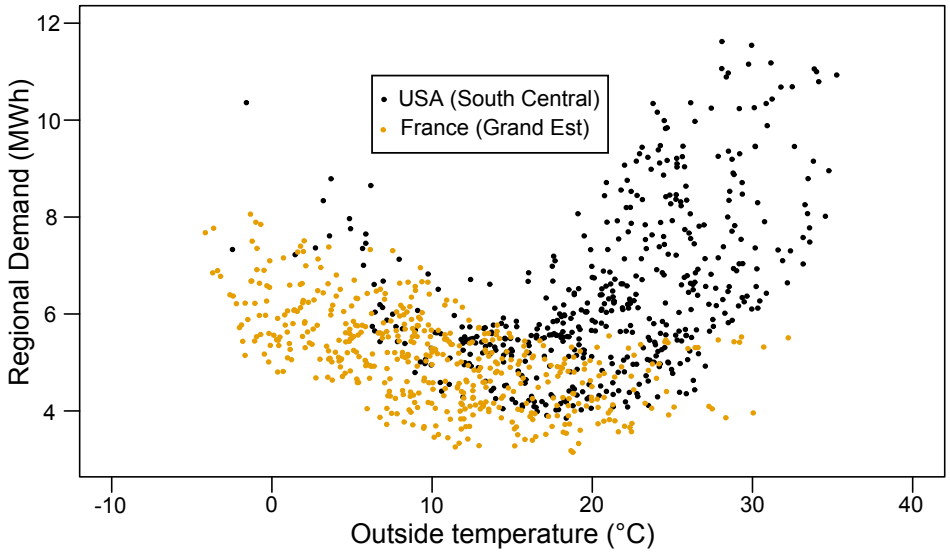


Figure 4.5 – Scatterplot of the hourly electricity demand made by South Central region in Texas, USA (black), and by Grand Est region, France (orange) (y-axis) against the outside temperature measured in the region (x-axis).

While the temperature greatly impacts demand for high demand levels, its effects are less clear at the household scale. In particular, these effects are highly dependent on the hour of the day: a high temperature at 4:00 has less impact than at 18:00. Figure 4.6 represents scatterplot of the hourly electricity demand made by a US household versus the outside temperature measured at 4:00 and at 18:00. Like at the region level, demand increases when it is hot or cold and is at the lowest around 20°C. Considering two hours of the day distinctively separate the temperature effects: we see that the cooling effect is more pronounced at 18:00 when people are home and directly affected by the outside hot temperature. The solid lines represent spline regressions of the data points for the points at the 4:00 and 18:00.

For each demand time series, 24 spline regressions of the hourly demand versus temperatures are made — one for each hour of the day. The resulting spline functions are used as demand forecasts $\hat{y}_t^\theta$, and then compared to the actual hourly demand data.

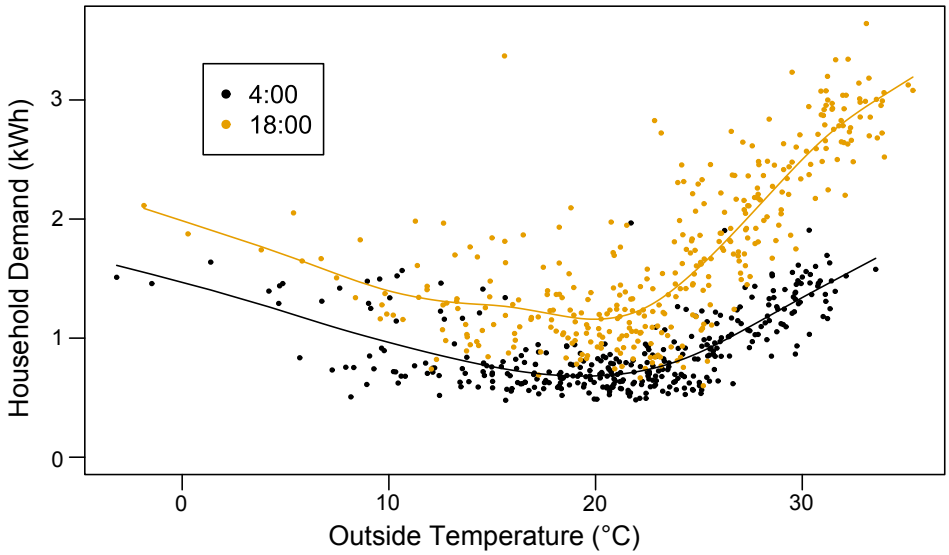


Figure 4.6 – Scatterplot of the hourly electricity demand made at 4:00 (black points) and at 18:00 (orange points) by one household in Austin, Texas (y-axis) against the outside temperature measures (x-axis). The solid lines are non-linear spline regressions.

The error made by each time series is measured with the Normalized Mean Average Error (NMAE), i.e.

$$\text{NMAE}_\theta = \sum_{t=1}^{T} \frac{|\hat{y}_t^\theta - y_t|}{\text{mean}\, y_t}. \tag{4.7}$$

The NMAE is a dimensionless value expressed in %. This score is negatively oriented, meaning that the lower is the NMAE, the better is the demand forecast based only on outside temperature. The process is made for 3 temperature time series: the measured one (θ_t), the day ahead forecasts ($\hat{\theta}_t$), and a climatology series ($\hat{\theta}_t^0$).

Although it is expected that the temperature measures should lead to better forecasting than forecast and than climatology, results reported in Table 4.4 lead to mixed conclusions. Average NMAE is written with the standard deviation between parentheses. For all datasets, forecasting performance is similar using either of the three temperature time series, except for the feeder consumption of USA dataset. Even more surprisingly, performance is better with climatology time series for the Portugal datasets. It shows that precise temperature-metering devices are not necessary and even a basic time series is sufficient to take the temperature impact into account in forecasting models. When comparing forecasting performance of the 1 day persistence

Table 4.4 – Forecasting performance (NMAE) of models using either measured, forecast, or climatology temperatures.

Type	Location	Size	(θ_t)	$(\hat{\theta}_t)$	$(\hat{\theta}^0_t)$
Household	Portugal	226	27 (10)	27 (9)	26 (8)
	France	176	30 (16)	31 (16)	31 (14)
	USA	175	39 (12)	40 (12)	41 (10)
Feeder	Portugal	1	3 (—)	3 (—)	2 (—)
	France	949	16 (10)	15 (10)	15 (8)
	USA	1	9 (—)	10 (—)	15 (—)
Region	France	12	8 (1)	8 (1)	8 (1)
	USA	8	5 (1)	5 (1)	8 (2)

model (Table 4.2) and temperature-only models (Table 4.4), we note that the first model is more efficient at the region scale, both models have same performance at the feeder level, but the temperature-only models are more efficient at the household scale, even for persistence of 1 hour in Portugal and France datasets. The worse performance at higher scales is explained by the use of a single temperature when the weather fluctuates in large geographical zone. Even though, one may hope to obtain smaller forecasting errors at the household scale with temperature-only models, the absolute errors remain large, around 30%.

From this temperature influence analysis, we conclude that using exact temperature values are not essential, and basic temperature profiles are often sufficient to capture the influence.

4.1.2.5 Vacation Influence

The presence and absence of residents in an household have a logical influence on the electricity consumed by the household. Researchers investigate this occupancy influence on the household electricity demand. In particular, detecting whether the residents are home or not from the household electricity demand are primordial for energy savings. Kleiminger et al. set up exhaustive metering in 5 households in Switzerland: their hidden Markov model classifier detects occupancy 80% of the time in average (Kleiminger et al., 2013). However, their datasets is made of demand measure every

second. Approaches directly using occupancy sensors are also used (Duarte et al., 2013) but primarily in large office buildings due to the installation and maintenance costs of sensors.

In our datasets, such occupancy classifying approaches are too ambitious: demand is measured every one hour, and training data for occupancy are unavailable. In consequence, occupancy cannot be used as an hourly inputs in forecasting model. However, the absence of residents during a prolonged period of time are visible on the time series, e.g. during week-long vacation. Figure 4.7 shows the hourly demand time series of an individual household on September 2017. The curve is almost flat from 11 to 28 indicating that the house is empty during this period. However, the demand is

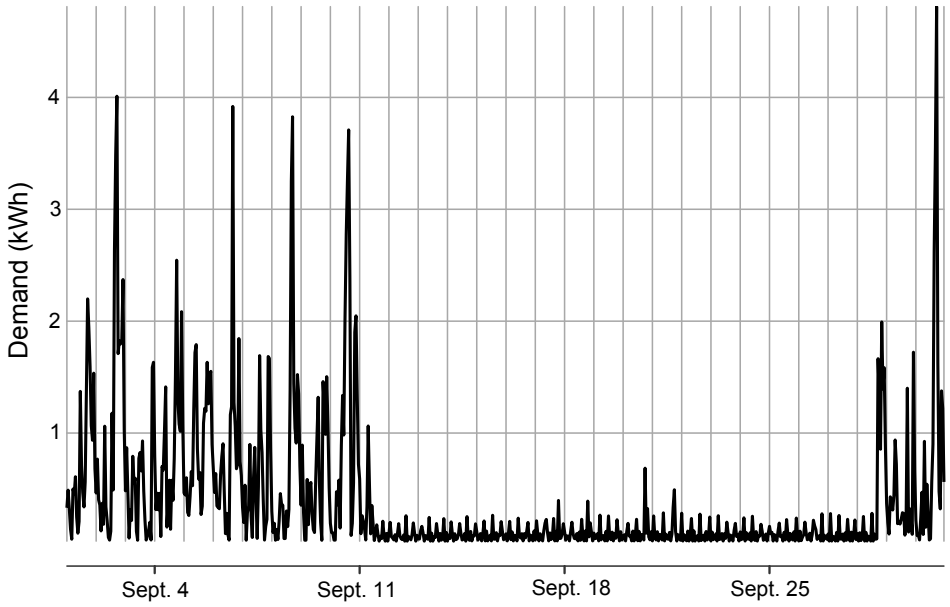


Figure 4.7 – Hourly electricity demand of a US household in September 2017.

not null even during prolonged period of non-occupation. There is significant residual electricity demand caused by standby mode of ICT equipment. Due to the growing number of such equipment, the standby consumption is accountable for about 10% of overall electricity demand (Gram-Hanssen, 2010). Furthermore, the daily demand curve is not smooth and presents non-negligible peak value. In consequence, identifying such periods is difficult and strongly depends on the processing. We apply the following processing:

1. a threshold is fixed at 100 Wh,

2. day is flagged when the difference between the hourly demand and the daily average demand is less than the threshold for all of the 24 hours of the day,

3. when 3 consecutive days are flagged, the according period is defined as a non-occupation period.

With this method, the period in Figure 4.7 between the 11 and the 28 September is correctly detected. According to this process, 16% of the Portuguese households, 5% of the French households, and 20% of the US one have a vacation period. Among those, the average vacation period is around 10 days.

Visual inspection agree with this: vacation periods identifiable on the household electricity demand are quite rare. With our processing, only days when demand fluctuates a little are detected, but remote appliances may be switched-on during these periods. Therefore, the separation between a vacation day and a day with little electricity activity is not clearly visible on the time series.

From this vacation influence analysis, we conclude that using prolonged vacation periods are rare events: around one week per year for 15% of the households. Furthermore, the daily curve of electricity demand during these days is not clearly distinguishable from any other day with little demand. These periods are therefore kept in the datasets and are used to evaluate the forecasting models introduced in this chapter.

4.2 Gradient Boosting Model

The Gradient Boosting Model is a recent machine learning model that proved to be efficient. A short description is provided in Section 2.1.2.5. Hereafter, we give details about the implementation used in practice, i.e. the function `gbm` (Ridgeway, 2017).

4.2.1 Framework

4.2.1.1 Theoretical

The information set of size J, i.e. $s_t = \{x_t^1, \dots, x_t^J\}$, is used, along with the observation y_t to train a day-ahead forecasting model during the training period $t = 1, \dots, T$. Since we wish to have probabilistic forecasts, multiple models are trained independently with different loss functions. To retrieve a set of quantiles, we therefore define loss functions

equal to the quantiles score at level $\tau = 0.01, \ldots, 0.99$. At a given instant t, we look for the value $g^\tau(s_t) = \hat{y}_t^\tau$ than minimizes the following quantile score[3] regarding the observation y_t,

$$\mathrm{QS}_\tau\left(\hat{y}_t^\tau, y_t\right) = \left(\mathbb{1}\left(y_t \leq \hat{y}_t^\tau\right) - \tau\right)\left(\hat{y}_t^\tau - y_t\right). \tag{4.8}$$

The implementation in package `gbm` uses a variant of the original gradient boosting algorithm adapted to the quantile score loss function (Kriegler & Berk, 2010). One initializes the forecasting model at a given quantile level τ, ${}^{(0)}g^\tau(s_1) = \cdots = {}^{(0)}g^\tau(s_T) =$ constant, and then repeat the follow steps recursively for $n = 1, \ldots, N$:

1. Compute the negative gradient, for $t = 1, \ldots, T$,

$$\begin{aligned} {}^{(n)}z_t &= -\frac{\partial}{\partial g^\tau(s_t)}\mathrm{QS}_\tau(g^\tau(s_t)), y_t)\Big|_{{}^{(n-1)}g^\tau(s_t)} && (4.9)\\ &= \tau\mathbb{1}(y_t > {}^{(n-1)}g^\tau(s_t)) - (1-\tau)\mathbb{1}(y_t \leq {}^{(n-1)}g^\tau(s_t)), && (4.10) \end{aligned}$$

2. Randomly select a subsample of the dataset of rate $p \in (0, 1)$ to be used in steps 3 and 4,

3. Fit a regression tree with K terminal nodes, $S_1, \ldots, S_K$ forecasting ${}^{(n)}z_t$ from s_t.

4. Compute the optimal terminal node forecasts, ρ_k for $k = 1, \ldots, K$ as

$$\rho_k = \underset{\rho}{\mathrm{argmin}} \sum_{s_t \in S_k} \mathrm{QS}_\tau({}^{(n-1)}g^\tau(s_t) + \rho, y_t), \tag{4.11}$$

   where S_k is the set of observations in node k.

5. Update estimation, for $t = 1, \ldots, T$,

$${}^{(n)}g^\tau(s_t) = {}^{(n-1)}g^\tau(s_t) + \lambda\rho_{k(s_t)}, \tag{4.12}$$

   where $k(s_t)$ indicates the index of the terminal node where observation at time t falls, and λ the shrinkage.

[3] We give the package definition which drops the leading multiplier of 2 compared to our definition in Equation (2.20). This package quantile score at 50% is equal to half the MAE; our quantile score at 50% is exactly equal to the MAE. Nevertheless, this multiplier does not impact the optimization, and the later forecasting performance discussion is made with our definition.

4.2.1.2 Tuning Meta-Parameters

Several meta-parameters are to be selected when training a gradient boosting model in order to have optimal performance. Three parameters are related to the complete structure of the model: the number of trees $\tau_{\max}$, the shrinkage parameter λ, and the subsampling rate $p \in (0, 1)$. In the package `gbm`, the weak learners are standard regression trees which mean than two more parameters devoted to the individual tree are to be selected: the interaction depth Δ, and the minimal number of observations required in a terminal node ν. The tuning of the meta-parameters is analyzed in Appendix E.

4.2.2 Day-Ahead Forecasting Model

4.2.2.1 Model Inputs

A total of 3 datasets with electricity demand time series at the household level are used to test the forecasting model (see Section 4.1.1):

- the Portuguese dataset contains hourly electricity demand of 226 buildings (mostly residential) in 2015, located in a neighborhood close to Évora;
- the French dataset contains hourly electricity demand of 176 buildings (exclusively residential) in January—March 2015, located near Tours;
- the USA dataset contains hourly electricity demand of 175 buildings (exclusively residential) in 2017, located in Austin, Texas.

A household time series are denoted by $(y_t)_{t=1,\dots,T}$.

The 3 following temperature time series are retrieved, for each location,

- the temperature measurements (θ_t),
- the temperature forecast at 12:00 the previous day $(\hat{\theta}_t)$,
- a climatology temperature $(\hat{\theta}^0_t)$.

To do the day-ahead forecasting of future household demand value y_t, only information known up to instant $t-24$ are known, and usable in real conditions[4] Following

[4] In consequence, the temperature time series used is the forecast one $(\hat{\theta}_t)$.

the observations made in Section 4.1, a total of 6 model inputs are selected for their relevance to forecast value at instant t:

1. demand measured on the day, $x_t^1 = y_{t-24}$;
2. median demand on the 7 previous days at the same hour

$$x_t^2 = \text{median}\left(y_{t-24}, y_{t-48}, \ldots, y_{t-168}\right),$$

3. hour of the day, $x_t^3 \in \{0, \ldots, 23\}$,
4. day of the week, $x_t^4 \in \{\text{Sunday}, \ldots, \text{Saturday}\}$,
5. temperature forecast, $\hat{\theta}_t$,
6. smoothed temperature forecast, $\hat{\Theta}_t$.

The smoothed temperature forecast is defined as in Gaillard et al. (Gaillard et al., 2016)

$$\hat{\Theta}_t = \alpha\hat{\theta}_t + (1-\alpha)\hat{\Theta}_{t-1}, \tag{4.13}$$

with the smoothing parameter $\alpha \in (0, 1)$, so that the resulting time series reflects the current temperature of the season. The smoothing parameter is optimized for each household time series so that the correlation between $\hat{\Theta}_t$ and the demand y_t is maximized, after the effect of temperature $\hat{\theta}_t$ on demand y_t is removed. Therefore $(\hat{\Theta}_t)$ brings the most original information in addition to $(\hat{\theta}_t)$. The optimal smoothing parameter ranges from 0.01 to 0.15 depending on the household with a median value of 0.08, meaning that temperature values have a lasting influence on the electricity demand, concurring with Wang et al. who exhibit a recency effect of the temperature, which peaks for temperature measured two days before the instant to forecast (P. Wang et al., 2016).

4.2.2.2 Quantile Outputs

The inputs are fed to the gradient boosting model at different quantile levels $\tau \in \{0.01, \ldots, 0.99\}$, to obtain quantile forecasts

$$\hat{y}_t^\tau = g^\tau\left(x_t^1, x_t^2, x_t^3, x_t^4, x_t^5, x_t^6\right). \tag{4.14}$$

Since the values are carried out independently for each quantile level, it occurs that the logical order of the values is not followed, i.e. we do not have $\hat{y}_t^{0.01} \leq \hat{y}_t^{0.02} \leq \cdots \leq \hat{y}_t^{0.99}$. This quantile crossing phenomenon is well documented (Chernozhukov et al., 2010) and multiple solutions have been proposed (Schnabel & Eilers, 2013). In our case, we simply sort the set of values obtained so that the order of the 99 quantile forecasts is logical.

4.2.3 Evaluation

4.2.3.1 Indices

The quality of a forecasting model is evaluated with different indice: the MAE, the Quantile Score, the CRPS (and its weighted version), and reliability. A complete description of these indices are given in Section 2.2.

4.2.3.2 Benchmark Models

Since some demand time series are easier to forecast than others, e.g. the ones with regular patterns, benchmark models are important to assess a basic level of performance that one expects when forecasting a demand time series. We propose two benchmark models: the 1 day persistence model, and a climatology model:

Persistence Model A 1 day persistence model is a fast and fairly efficient model that forecasts the demand value at instant t by the demand measured the previous day at the same hour, i.e. ${}^{\mathrm{p}}\hat{y}_t = y_{t-24}$. This model does not provide probabilistic forecasts, so only its deterministic performance can be evaluated. Furthermore, this model is greatly impacted by missing values: if measures have not been made on day d, then no forecast are carried out for day $d+1$. Therefore, the evaluation is slightly biased compared to other models that produce forecast even in these cases.

Climatology Model The climatology model computes a climatology profile for a single day. The model computes, for each hour of the day, the values of the demand measured at this hour of the day at quantile level $\tau = 0.01, \ldots, 0.99$ noted ${}^{\mathrm{c}}\hat{y}_t^{\tau}$. These values are computed on a training period, made of half the data available randomly selected. The climatology is therefore a probabilistic model, and it produces the similar forecasts for every day.

4.2.4 Results

4.2.4.1 Optimal Number of Trees

The performance of the gradient boosting model strongly depends on the correct tuning of its meta-parameters. With the tuning introduced in Section 4.2.1.2, we obtain the optimal number of trees to build to do the forecast at quantile levels $\tau = 0.01, \ldots, 0.99$. Figure 4.8 depicts the optimal number of trees found in average to train forecasting models for the French dataset for each quantile value, determining the distribution function of the model. The optimization is made with cross validation in 5 folds, the shrinkage parameter is fixed at $\lambda = 0.05$, the minimum size of leafs is fixed at 10, and only 1 variable is used for splitting. After preliminary trials, the upper bound for the number of trees is fixed at 2000. This graph shows that to obtain best accuracy, the

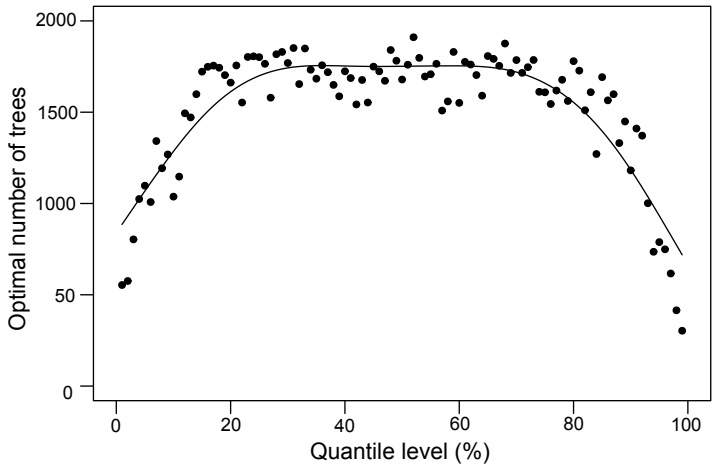


Figure 4.8 – Optimal number of trees found by the gradient boosting model (y-axis) regarding the quantile level (x-axis). Points represents the average optimal number of trees found for the 176 households in the French dataset. The solid line is a smoothing regression.

model needs less trees for extreme quantiles to avoid overfitting. It comes from the fact that extreme values are rarely observed, and forecasting models should be conservative regarding these rare events.

4.2.4.2 Temperature Influence

To assess the influence of temperature on household electricity demand demand, we compare 4 versions of the forecasting model: (1) one with the temperature measures, θ_t as input, (2) one with the temperature forecast at 12:00 the day before, $\hat{\theta}_t$ as input, (3) one with a climatology temperature, $\hat{\theta}_t^0$ as input, and (4) one without any temperature input at all. Intuitively, the forecasting performance of the models should be ordered, i.e. version (1) more efficient than version (2) and so on.

The 4 versions of the model are computed for the 176 households of the French dataset and evaluated with the quantile scores normalized by the average power of each household demand. The normalized scores are then multiplied by the standard weights, i.e. the theoretical quantile scores for the Gaussian distribution, as as to have comparable scores at every quantile levels. The average normalized weighted quantile scores obtained are plotted in Figure 4.9. The logical order of the performance

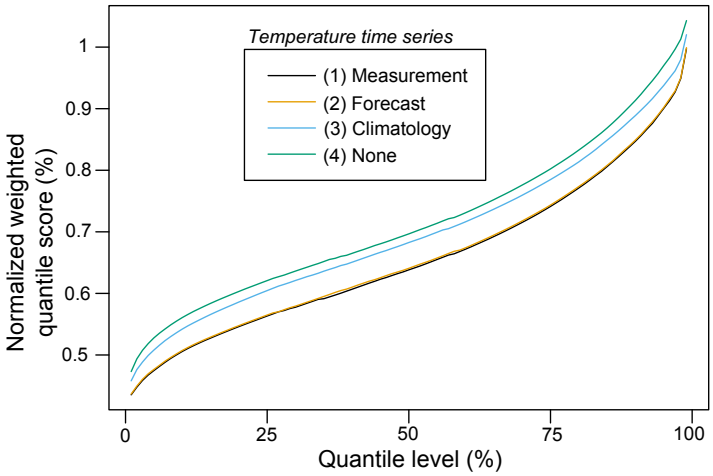


Figure 4.9 – Normalized quantile scores (y-axis) of the 4 versions of forecasting model regarding the quantile levels (x-axis) for the 176 households of the French dataset.

is respected: using temperature measurements is slightly more efficient than using temperature forecasts (improvement around 0.3%), which is more efficient than using a climatology time series (improvement around 6.4%), which is more efficient than using no temperature at all (improvement around 9.1%). The usage of an accurate forecast temperature is therefore an informative input of the models and should be consider when designing a model. This goes against the preliminary analysis (see Section 4.1.2.4) in which a basic temperature-dependent model is deemed as efficient with any type

of temperature: only the combination of historical values and temperature reveals the need of precise forecast values. To mimic the real condition and have good performance, we later use the forecast temperature $\hat{\theta}_t$.

Another observation is that the quantile score curves is asymmetrical, meaning that higher demand values are relatively more difficult to forecast than the lower ones. Furthermore, the quantile scores of these extreme parts (0—5 and 95—100%) are relatively better for versions (3) and (4) than the middle part. Such extreme demand are in fact more related to extraordinary appliances usage than to temperature.

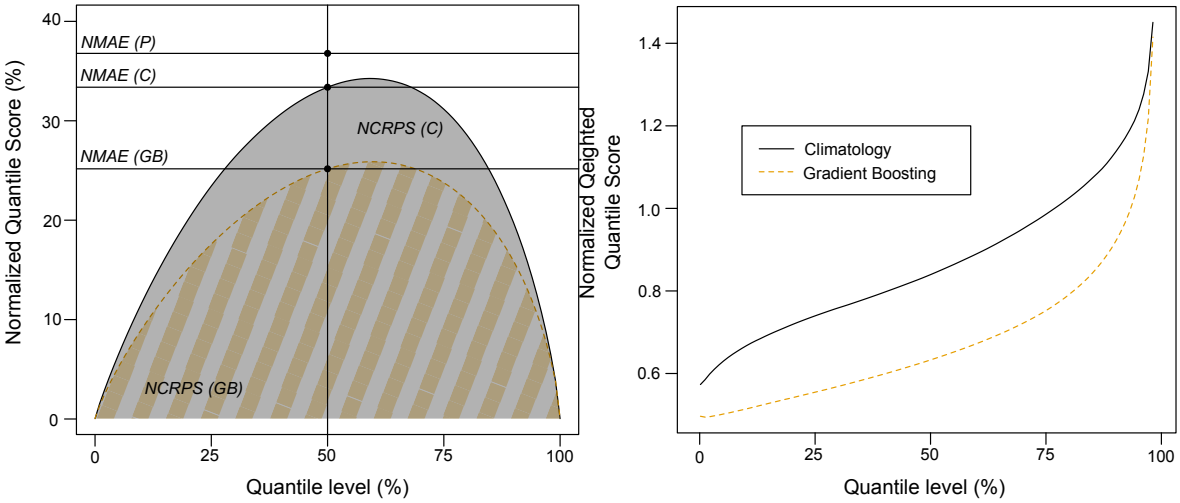
4.2.4.3 Detailed Results

Table 4.5 summarizes the average performance — and standard deviation between parentheses — of the persistence, climatology, and gradient boosting models for the 3 datasets. More thorough results are reported in Appendix D.

Table 4.5 – Performance of persistence, climatology, and gradient boosting models.

Dataset	Model	NMAE (%)	NCRPS (%)
Portugal	Persistence	33 (11)	—
	Climatology	30 (11)	21 (7)
	Gradient Boosting	**25 (9)**	**17 (6)**
France	Persistence	37 (18)	—
	Climatology	33 (13)	24 (10)
	Gradient Boosting	**25 (13)**	**18 (9)**
USA	Persistence	46 (13)	—
	Climatology	52 (9)	36 (6)
	Gradient Boosting	**33 (9)**	**24 (7)**

In average, the gradient boosting outperforms the benchmarking models, as it can be seen in Figure 4.10 in the French case. The left panel represents the average quantile scores curves obtained with the two probabilistic models: climatology (C) in black and the gradient boosting (GB) in orange. The GB curve is below the C curve at all quantile levels, which is reflected in the lower NCRPS — i.e. better performance —, which is of 24% for the climatology and 18% for the gradient boosting. The NMAE is read at the 50% level, and represented with large points. The NMAE is around



(a) Normalized quantile scores

(b) Normalized weighted quantile scores

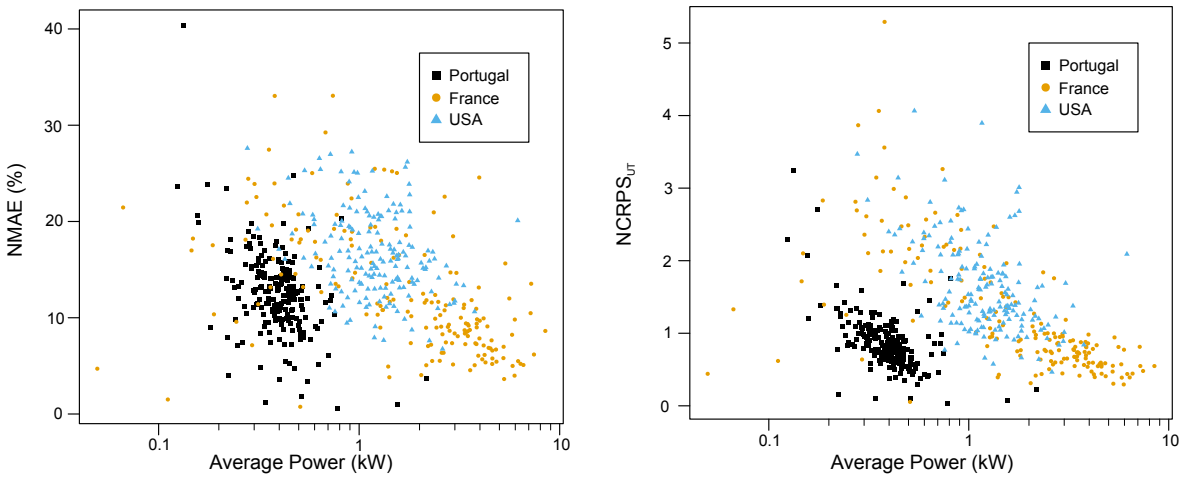
Figure 4.10 – Average quantile score curves obtained with the 176 households of the French dataset. In Figure 4.10a, the quantile scores are normalized and averaged over all of 176 households: the NMAE is read at the 50% quantile level, and the NCRPS is read with the integral of the curves. In Figure 4.10b, the quantile scores are multiplied by the standard weights. In both figures, the climatology model is plotted in black and the gradient boosting model in orange.

25% for the gradient boosting, which decreases the error by around 1/4 compared to the climatology NMAE (33%) and by around 1/3 compared to the persistence NMAE (37%). The improvement is comparable for the Portugal and USA datasets, but the relative performances of the persistence and climatology model are reversed — for instance, the persistence model in the USA is more efficient than the climatology, because of the stronger daily seasonality.

The right panel represent the same average curve but weighted by the standard quantile score. We see that the curves are increasing, meaning that the upper part of the distribution is more difficult to forecast than the lower part. This effect is especially strong for the extreme upper part, so the rare peak demand are notably difficult to anticipate even with the highly flexible gradient boosting model. The relative improvement of the gradient boosting over the climatology is more visible on the middle part of the forecast distribution, by ingeniously making use of the numerous observations.

Figure 4.11 shows the day-ahead forecasting performance, measured by the NMAE

and NCRPS_{UT}, for each one of the buildings in the Portugal (black squares), France (orange circles), and USA (blue triangles) datasets. The performance is compared to the average power of the building (logarithmic x-axis). We note the slight decrease in errors — i.e. better performance — when the average power increases. This effect is more clear when assessing the quality of the upper tail of the distribution, i.e. with NCRPS_{UT}. At comparable average power, the performance greatly changes between buildings which reflects the large variety of electricity habits. We nonetheless notice that the buildings in the same location form visible clusters in the graph: overall the Portugal households have lower average power and lead to better performance; the France and USA buildings are more energy intensive but more difficult to forecast. While the negative correlation between average power and forecasting performance is abundantly documented, see Section 2.3.3, it seems this correlation is more ambiguous for a small power range. Such ambiguity reminds of the Simpson's paradox widespread in statistics (Blyth, 1972).

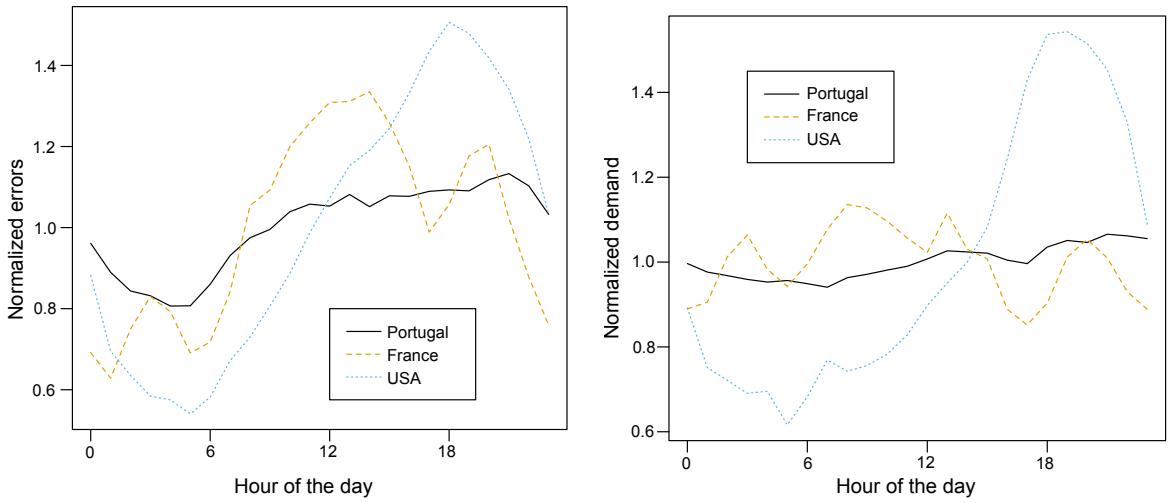


(a) NMAE (b) NCRPS_{UT}

Figure 4.11 – Performance scores of each individual building (y-axis) regarding the average power of the corresponding building (logarithmic x-axis) for the three datasets: Portugal (black squares), France (orange circles), and USA (blue triangles).

Figure 4.12a shows the forecasting performance as a function of the hour of the day. The computation is done by averaging all the normalized absolute errors of a given hour over all the buildings in each dataset: Portugal (black), France (orange),

and USE (blue). Each resulting curve of 24 values is then divided by its mean to obtained a normalized errors fluctuating around 1 to be comparable between the 3 datasets — lower values indicate better forecasting performance. Similar errors profiles are obtained when examining the extreme part of the distributions rather than the middle part. In all 3 datasets, we see that the nighttime is easier to predict than the rest of the day, due to the more regular habits — householders are sleeping. On the other hand, the errors fluctuations throughout the day are much more pronounced for the France and USA datasets indicating that households have more varied and unpredictable habits. We can draw a parallel between these hourly errors and the average daily demand profile, see Figure 4.12b. These average profiles are extremely close to the the error profiles in the Portugal and USA datasets: it is indeed more difficult to forecast values when the values to be forecast are higher. Such correlation is not visible for the France dataset. This comes in part from the relatively smooth demand profile, obtained with large number of devices scheduled during the night, e.g. water heater. Such scheduling is precisely grasped by the forecasting model, as opposed to the activities during the rest of the day.



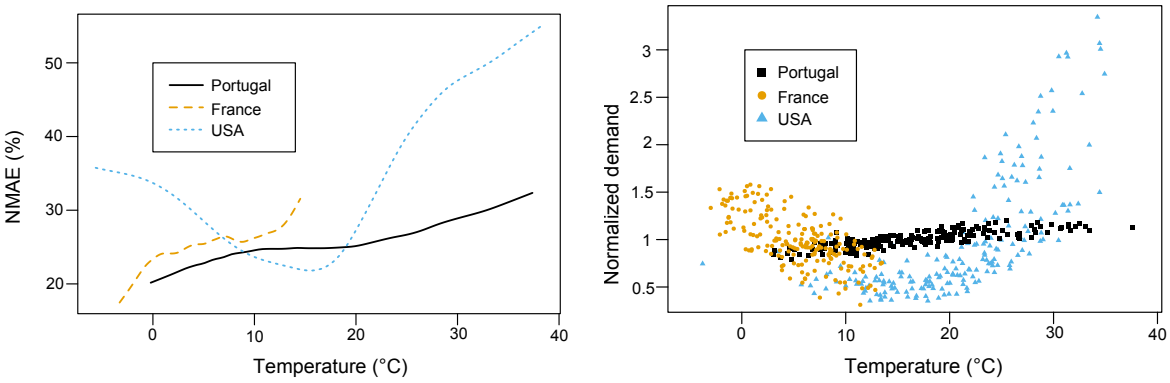
(a) Normalized errors profile

(b) Average demand profile

Figure 4.12 – For the three datasets: Portugal (black solid lines), France (orange dashed lines), and USA (blue dotted lines): (a) The normalized profile errors throughout the day; (b) the average demand profile throughout the day.

Figure 4.13a depicts the relation between the NMAE and the local temperature.

This graph is made by superimposing all the normalized absolute errors observed and the corresponding temperature and by doing a spline regression on all the points of each 3 datasets. The temperature range varies between the three datasets because of the different measurement periods, e.g. the French data covers only the winter period. When drawing a parallel between the NMAE as a function of the temperature, see Figure 4.13b, and the average demand as a function of the temperature, no definite conclusion can be drawn about the forecasting performance. In general, when electricity demand increases, so does the errors: this is seen during warm periods in Portugal and USA and cold periods in USA. In the France case, the relation is reversed: and the cold periods are easier to forecast, probably because of automatic scheduled heating devices well capture by the forecasting model. In any case, these observations show that the relation between temperature and forecasting performance is not clear.



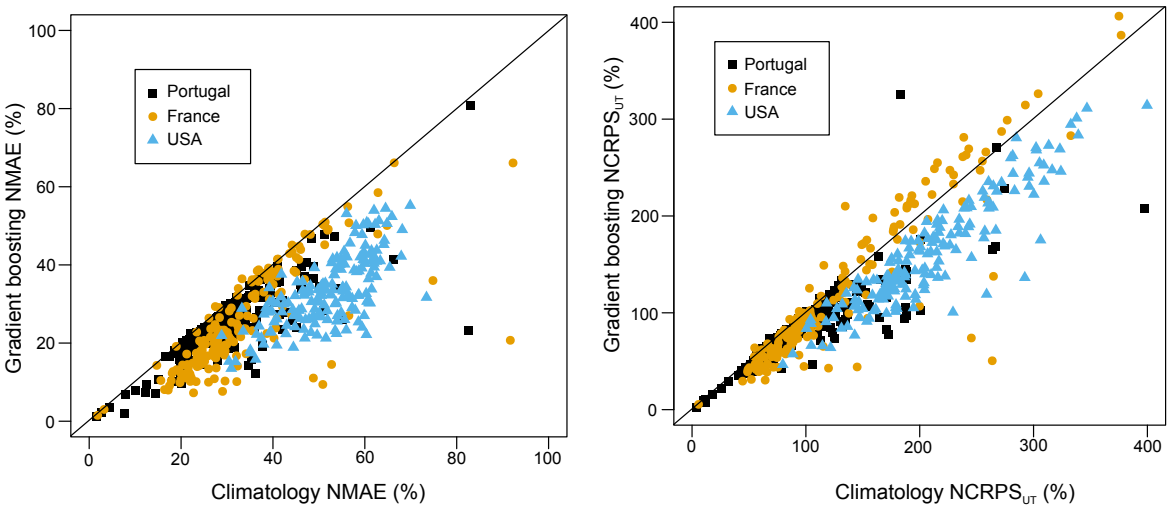
(a) NMAE vs. temperature

(b) Average demand vs. temperature

Figure 4.13 – For the three datasets: Portugal (black), France (orange), and USA (blue): (a) the average NMAE as a function of the local temperature; (b) the normalized demand as function of the local temperature.

Figure 4.14 compares the performance of the gradient boosting (y-axis) and that of the climatology models (x-axis). Figure 4.14a examines the deterministic performance with the NMAE index, while Figure 4.14b compares the probabilistic forecasts, and specifically the quality of the upper tail with the $\mathrm{NCRPS}_{\mathrm{UT}}$ score. Most points, i.e. buildings, are below the diagonal, meaning that the gradient boosting is more efficient. For the deterministic performance, the absolute decrease of the NMAE is higher when the climatology NMAE is high, but the relative improvement remains

stable — around 25%. Regarding the upper tail of the forecast distribution, for a large amount of Portuguese and French buildings, the climatology model is more efficient than the gradient boosting model. It indicates that in some not-so-rare cases, producing conservative forecasts of the upper tail with the climatology lead to better results than advanced machine learning technique. Such effect is exacerbated when the the individual demand peak is more difficult to forecast, i.e. greater comparatively to the average demand, observed for high climatology $\mathrm{NCRPS_{UT}}$. However, the effect is not visible for the USA dataset. One possible explanation is that complex models, such as the gradient boosting, underperform when forecasting rare event with few observations such as the France dataset — 2,136 hourly values —, but their performance improve with larger training sets, such as with the USA dataset — 8,760 hourly values.



(a) NMAE (b) $\mathrm{NCRPS_{UT}}$

Figure 4.14 – Individual performance of the gradient boosting model (y-axis) and the climatology model (x-axis) for the 3 datasets: Portugal (black squares), France (orange circles), and USA (blue triangles). Panel (a) shows the deterministic NMAE score, and panel (b) shows the probabilistic evaluation of the upper tail with $\mathrm{NCRPS_{UT}}$.

4.3 Forecasting Performance and Aggregation Effect

4.3.1 Introduction

It has been observed that the performance of an electricity demand forecasting model depends on the level of demand aggregation (Wijaya et al., 2014). In general, forecasting the electricity demand of one household is more difficult than of a region. Sevlian and Rajagopal propose a popular scaling law of the expected NMAE (in %) depending on the average power of the demand time series (Sevlian & Rajagopal, 2014). This law writes

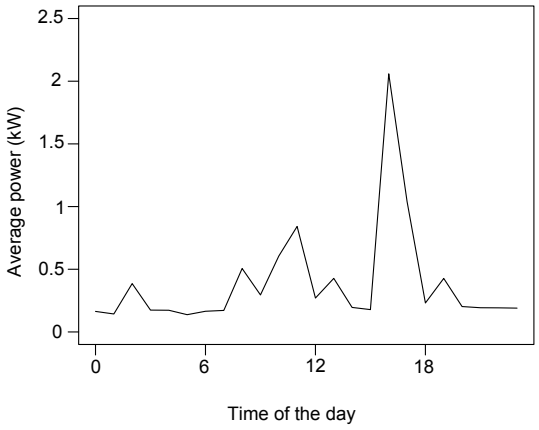
$$\mathrm{NMAE}(W) = \sqrt{\frac{\beta_0}{W^p} + \beta_1}, \tag{4.15}$$

with parameters $p > 0$, β_0, and β_1 to be fitted with the case study. This expression gives the expected error in % when forecasting a demand time series of average power W in kW. While the performance reported in our literature review, see Section 2.3.3, roughly follows this law, recent research suggests that it does not always apply in some specific situations, e.g. feeders with large overnight consumption (Haben et al., 2018).

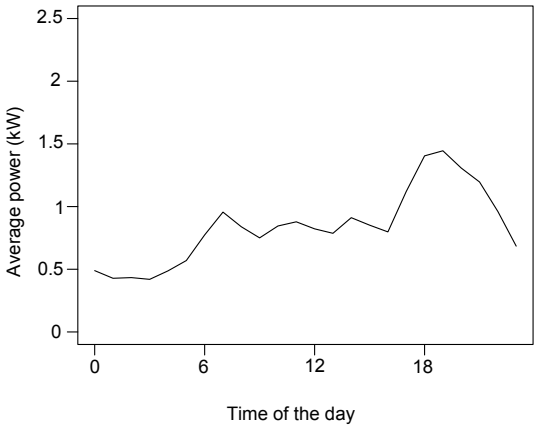
The time resolution of the data is also a key point when studying forecasting performance. It is known that better performance comes with coarser resolution as illustrated by Rodrigues et al. who obtain a MAPE around 20% when forecasting hourly demand values and around 5% when forecasting daily demand values (Rodrigues et al., 2014). Lusis et al. study precisely how the day-ahead forecasting performance at different time resolution evolves: they observe that all forecasts models perform better, but not much better, when the time resolution is changed from 30 minutes to 120 minutes (Lusis et al., 2017). A visual inspection clearly highlights the forecasting challenges induced by the different aggregation levels and time resolutions, see Figure 4.15.

4.3.2 Case Study

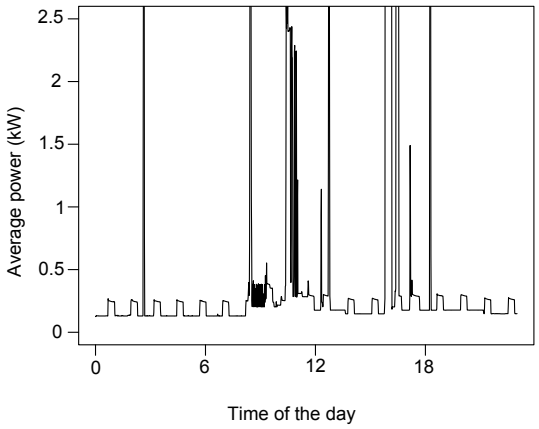
We investigate precisely the effect on forecasting performance when we consider demand time series at different average power levels and different time resolutions. We randomly select a subset of 92 households from the US dataset, introduced in Section 4.1.1, from which we retrieve measurements made every minute. When summing the 92 individual series, we obtain a power demand time series just short of 100 kW, and consequently only the range of performance for aggregation level between 1 kW and



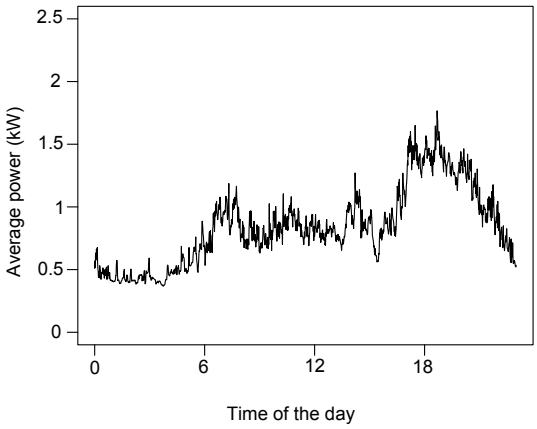
(a) Hourly power of 1 household



(b) Hourly average power of 100 households



(c) Minute-by-minute power of 1 household



(d) Minute-by-minute average power of 100 households

Figure 4.15 – Power demand measurements recorded during one day for 1 household (left panels), and for the average of 100 households (right panels), at a time resolution of 1 hour (top panels) and 1 minute (bottom panels).

100 kW is possible here. Different virtually aggregated power time series are obtained by summing any number of households from the complete dataset, meaning that the performance at a certain aggregation level is strongly correlated with the performance at a lower aggregation level. From the minute-by-minute time series, we create series for different resolutions, specifically for 1 minute, 5 minutes, 15 minutes, 30 minutes, 1 hour, 6 hours, 1 day, 3 days and 1 week by averaging the power recorded during the corresponding periods. This also means that the performance at one time resolution is strongly correlation with other time resolutions. We adapt the gradient boosting model and its inputs introduced in Section 4.2 to produce the day-ahead forecasts: for instance the hour input is removed for the daily forecasts, and the temperature input is averaged over the future week to obtain a unique temperature in the weekly model. The default meta-parameters of the **gbm** package are used, meaning that the absolute performance are not optimal. The loss function used to train the model is the quantile score at 50%. The performance index is thus the NMAE assessing solely the deterministic performance. This index is averaged over multiple out-of-sample test sets, thanks to a cross-validation approach.

4.3.3 Performance as a Function of Average Power and Time Resolution

Once the performance is evaluated on a grid of average power W (in kW) and time resolution τ (in minute), we fit a 2D additive model $f(\cdot,\cdot)$ such that $\text{NMAE}(W,\tau) = f(W,\tau)$, where W is the average power in kW, and τ the time resolution in minute. The resulting 3D graph is represented in Figure 4.16 with the **R** function `levelplot`. Specific curves are represented in Figure 4.17 for 3 time resolutions. Considering a demand time series of average level 10 kW, the average NMAE is 11.7% at a 4 hour resolution, 20.1% at a 1 hour resolution, and 24.7% at a 15 minute resolution. Considering a demand time series at a 1 hour resolution, to forecast one of average power of 2 kW leads to a NMAE of 32.3%, one of 10 kW to 20.1%, and one of 50 kW to 12.0%. The power law, see Equation (4.15) roughly fits the shape of the 1 hour resolution curve. A robust estimation of the coefficient leads to: $\beta_0 = 3.4$, $\beta_1 = 0.1$, and $p = 1.2$. Therefore, the threshold power, from which the forecasting performance plateaus, is $W^* = 16$ kW. It indicates that aggregating around 15 US residential households seems

optimal from a day-ahead forecasting performance point of view. Let us note that since our average power range is 1—100 kW and our dataset limited to 92 residential households, the power threshold greatly differs from the 19 MW found in the literature review, see Section 2.3.3.

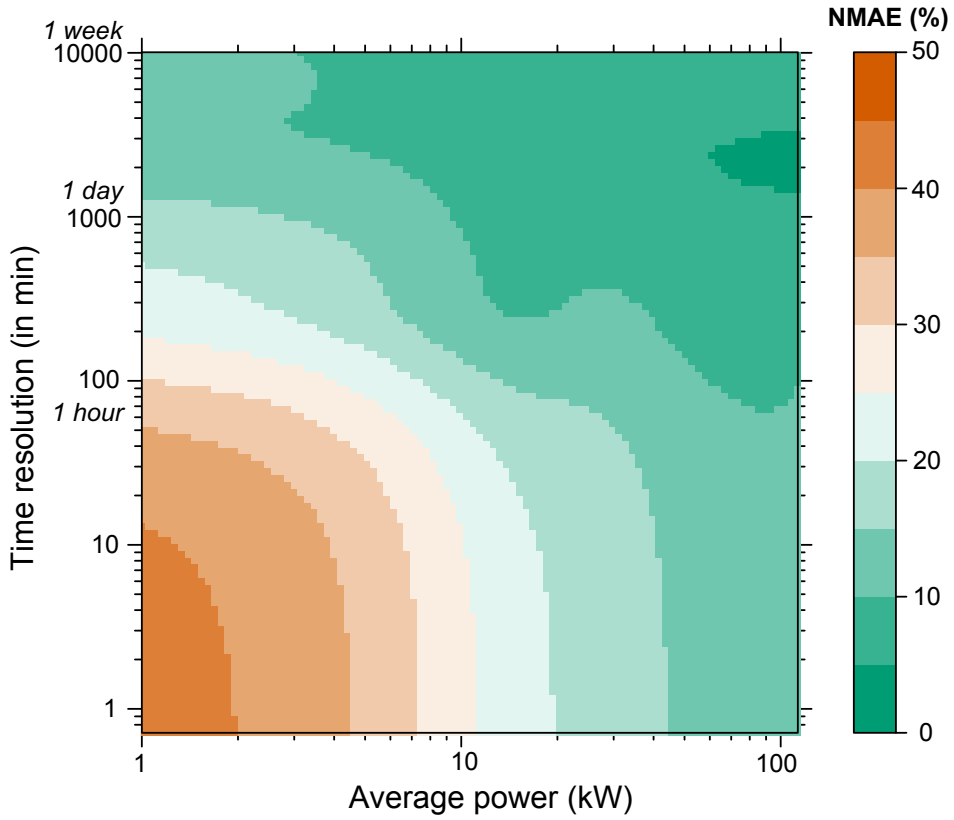


Figure 4.16 – Average performance, measured by the NMAE, of a reference day-ahead forecasting model at different power aggregation levels (logarithmic x-axis) and at different time resolution (logarithmic y-axis).

4.3.4 Forecasting Hourly Values with Various Resolution Data

The previous results indicate that it is easier to forecast a demand time series at coarser resolution than at finer resolution, meaning that one is better off forecasting the future weekly energy than the precise power required at 16:32 tomorrow. However, the required resolution is usually not the responsibility of the forecaster but rather that of a later user. So, when this user requires a forecast value every hour, the forecaster abides and provides one forecast per hour. A relevant question is then: what is the

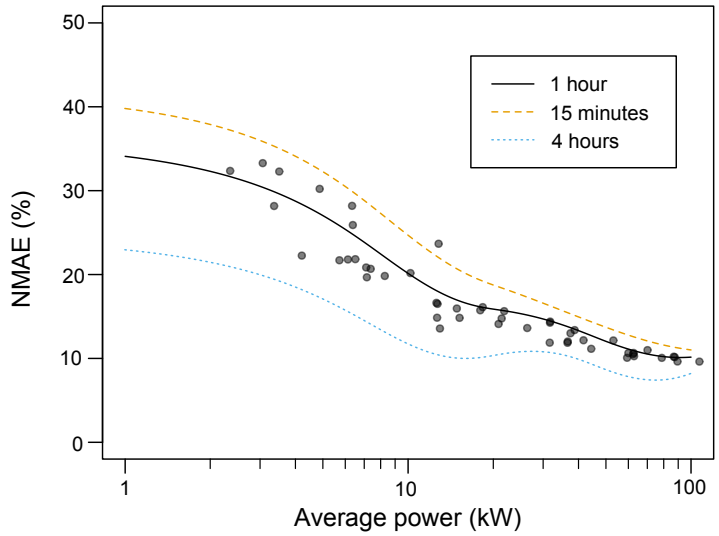


Figure 4.17 – Average NMAE performance (y-axis) regarding the average power of the time series (logarithmic x-axis), for 3 time resolutions: 1 hour (solid black line), 15 minutes (dashed orange line), 4 hours (dotted blue line). The points represent the actual performance assessed for the 1 hour resolution.

optimal resolution of the training data to forecast demand at a specific resolution, e.g. 1 hour?

Let us detail by taking the 48 forecast values produced by a day-ahead forecasting model trained at a 30 minute resolution, labeled $\hat{y}_{0:00}, \hat{y}_{0:30}, ..., \hat{y}_{23:30}$. For a typical household of average power 2 kW, when we compare this forecast series with the observation series, i.e. successively compare $\hat{y}_{0:00}$ with observation $y_{0:00}$, $\hat{y}_{0:30}$ with observation $y_{0:30}$ and so on, we obtain a MAE around 718 W ($35.9\% \times 2$ kWh). If the user wants one forecast value every hour, we rather compare $\frac{\hat{y}_{0:00} + \hat{y}_{0:30}}{2}$ with $\frac{y_{0:00} + y_{0:30}}{2}$. The triangular inequality tells us that, in theory, the resulting MAE is lower or equal than 718 W. This claim concurs with our experimental results (see Figure 4.17). Nevertheless no theory tells us if this resulting MAE is lower than 646 W, i.e. the MAE obtained using the 1 hour time series.

Figure 4.18 shows the performance obtained when forecasting hourly values with demand time series recorded at different time resolutions: from 5 minutes to 1 day. The NMAE is averaged over subsets of US households[5]. The NMAE obtained for

[5]The subsets are randomly sampled among the complete dataset.

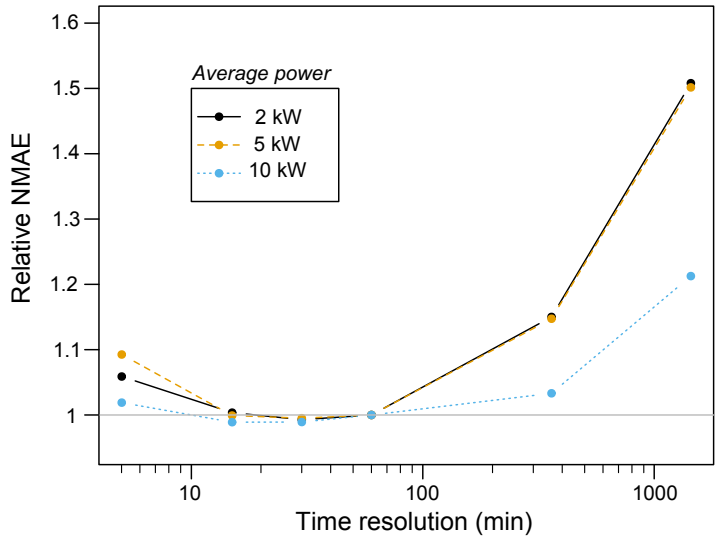


Figure 4.18 – The average NMAE obtained when forecasting the hourly demand time series using demand time series at different resolution, from 5 minutes to 1 day (x-axis). The NMAE obtained is divided by the NMAE of the hourly time series. The evaluation is done at 3 average power: 2 kW (solid black line), 5 kW (orange dashed line), and 10 kW (blue dotted line).

different resolutions is divided by the reference NMAE obtained with the hourly time series. For all average power — 2, 5, or 10 kW — we see that the optimal performance is found by using data recorded every 30 minutes, marginally improving the reference NMAE by less than 1%. Logically, the forecasting performance using coarser resolution is strongly degraded. However, perhaps surprisingly, using demand recorded every 5 minutes to forecast the hourly values increases the forecasting errors. The additional information brought by the 5 minute values is, in fact, detrimental to the quality of hourly forecasts. In general, it is more efficient to forecast a phenomenon by metering directly this very phenomenon, and not indirectly through divided causes.

Similar behavior is observed when we forecast half-hourly values: using the same resolution as required by the forecast's user is almost optimal, often slightly outperformed with just finer resolution, that is using 15 minute values slightly improve performance when forecasting 30 minute values. In any case, one should not use coarser resolution than the one desired by the user of the forecasts, since it substantially decreases forecasting performance.

4.4 Robust Forecasting Model and Operational Challenges

4.4.1 Presentation

In the frame of the SENSIBLE project (SENSIBLE, 2018), part of EU Horizon 2020, we develop a day-ahead forecasting model to be used on a real-time platform. The project demonstrates the use of energy storage for buildings and communities. It requires the deployment, for each household, of a day-ahead electricity demand forecasting model . Since the performance of demand forecasting is known to be quite poor at the household level — state-of-the-art errors range from 5% to 60% (see Section 4.3 and review (Yildiz et al., 2017)) —, a probabilistic output is employed to quantify the uncertainty, following a current trend in the forecasting literature (Hong & Fan, 2016). An operational load forecasting platform was set up to predict the demand of each household at the demonstration site of the city of Évora in Portugal. The platform retrieves information from the smart meters at each household through appropriate application programming interfaces (APIs). The outputs of the forecasting models are then transmitted to other applications to be used as inputs, such as Home Energy Management Systems (Correa-Florez et al., 2018). Our model should provide the probabilistic forecasts at 12:00 on day $D-1$ of the future demand expected on day D at 0:00, 1:00, ..., and 23:00, i.e. for horizons of 12, 13, ..., and 35 hours. In such a use case, several features are required for the forecasting model to be implemented:

- *High robustness:* demand forecasts are required at all times in all situations, e.g. new house, faulty meter, etc., with reasonable performance.
- *Fast computation:* the model should carry out demand forecasts in a reasonable time for a potentially large number of households than can range from hundreds to thousands.
- *Easy replicability:* the model should be easily replicable for a large number of household typologies and demand profiles.
- *Remote control:* no direct intervention is possible in situ.

- *Easy interpretation:* finally, among two competitive models with equivalent performance, some end-users may have preference for a model that is understandable by anyone, instead of a black-box approach.

Consequently, given the operational requirement for high availability in the forecasts, a robust approach is proposed based on the operation of alternative models of varying complexity through a hierarchical framework. In contrast to mot academic studies, here we compare the simulation results under ideal conditions (i.e. in terms of input data availability) with field tests featuring erroneous or missing data. This provides a realistic view of the level of load predictability at local scale.

To address these requirements, in Section 4.4.2, we introduce 5 forecasting models — and a reference model based on machine learning — at the household level. These are combined in a hierarchical framework so that they can always provide a forecast output. In Section 4.4.3, we (1) analyze the respective performance of each model with an offline dataset and (2) identify the possible situations preventing the usage of a specific forecasting model so as to (3) propose a hierarchical framework to design a foolproof forecasting model. After deployment in 2018 at the demonstration site, the field experience is used to evaluate the performance of the forecasting hierarchical framework. A comparison between this online performance and the offline performance is drawn and discussed in Section 4.4.4.

4.4.2 Case Study and Models

Firstly, we describe the offline dataset collected in bulk with the smart meters set up as part of the SENSIBLE project. Secondly, we define the selected input values that are to be fed into the forecasting models we then introduce. Finally, we present the different scores that are used to assess the forecasting performance of the models.

4.4.2.1 Offline Data Set

As part of the SENSIBLE project, smart meters are set up in a localized neighborhood in Évora, Portugal, see Figure 4.19, and record the hourly electricity demand of each of the 226 households of the neighborhood. The recordings collected during the 8,760 hours in 2015 form the offline data set, made up of 226 individual time series. A

mean demand time series is created by averaging the demand of the 226 individual households,

Figure 4.19 – Localization of the neighborhood comprising the 226 households at the demonstration site in Évora, Portugal. Source: Google Maps.

Following common practice, this dataset is divided into a training period to fit the models' parameters, from 1st January to 30th September — 6,552 values, and a test period from 1st October to 31st December — 2,208 values. This separation is made to emulate real-life conditions where a model is trained and then installed for operational use. This is opposed to other approaches, e.g. cross-validation, which do not provide a realistic performance assessment. The opposition is illustrated in Figure 4.20. In this

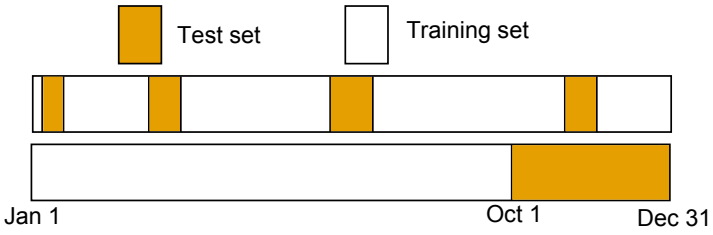


Figure 4.20 – The cross-validation method (top graph) randomly selects a fold of the whole period to use as a test set. In a real application (bottom graph) the test set follows the training set.

case, the forecasting model is trained with historical data, and then deployed at a given instant, on the 1st October 2015, to be tested over 3 months. The recordings collected

during the 8,760 hours in 2015 form the offline dataset, made up of 226 individual time series. Advanced learning techniques exist, such as a recursive training process that regularly refines the model parameters with the most recent data, blurring the lines between the training and test periods (Rydén, 1997). We do not consider such techniques here since they require high maintenance.

4.4.2.2 Input of the Forecasting Models

An efficient forecasting model makes use of informative inputs in order to produce relevant forecasts. Based on the electricity demand forecasting literature and to keep a small input set, we select only two kinds of information: historical data of demand measurements, and local outside temperature.

Historical Demand Measurements Recent demand measurements, i.e. lagged values of the time series, constitute precious information when forecasting future demand y_t (Gerossier, Girard, et al., 2017). Selecting the most informative lagged values is tricky and is ideally made for each household separately. A common practice is to analyze the partial auto-correlation function. This function quantifies how much each lagged value is correlated to the current value independently of the values in between, e.g. how much y_{t-2} is correlated to y_t after removing the correlation effect between y_{t-1} and y_t (Brockwell & Davis, 2013). However, selecting automatically how many lagged values and which ones for each household is often cumbersome, and hinders the replicability of the model. For instance, the number of relevant lags change with household, and as a consequence, they modify the complexity of the models.

Here we consider that the primary interest is to develop a model that is easily replicable for a (very) large number of households that range from hundreds to thousands. We therefore opt to keep only two lagged values that proved efficient on average:

1. The measurement made 24 hours before the instant to forecast y_{t-24}, which is highly informative due to the strong daily seasonality. When the forecasting horizon is superior to 24 hours, the measurement made 48 hours before is used as a direct surrogate.

2. The median demand made on the previous week $\bar{y}_t = \text{median}(y_{t-24}, \dots, y_{t-168})$, which reflects the recent behavior.

While these two historical inputs are related, both are insightful: the value observed the previous day is volatile and depends on the specific inhabitant's activity on this particular day, the median value of the previous week conveys the recent habits in a smoother manner.

Outside Temperature The impact of the local outside temperature on electricity demand is generally recognized (Bessec & Fouquau, 2008). For forecasting purposes, we retrieve the local temperature predictions made on the previous day from a Numerical Weather Prediction (NWP) model. In this case study, we consider NWPs provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Buizza, 2014). For the offline dataset, and to mimic the real application, we retrieve the deterministic forecasts made at 12:00 for the next day, i.e. with forecasting horizons of $12, 13, \ldots, 35$ hours[6]. Therefore, the temperature forecasts produced at 12:00 on 31st December 2014, 1st January 2015, ..., 30th December 2015 are collated in a time series, noted $(\hat{T}_t)$, comprising the 8,760 hourly temperature values in the neighborhood in 2015. For the online usage, the NWPs are directly retrieved at the household level through an API. Although some studies show that lagged values of the temperature slightly improve the electricity demand forecasting performance, we select only one single value to keep the model simple and interpretable (P. Wang et al., 2016).

4.4.2.3 Forecasting Models

In order to provide a day-ahead probabilistic forecast of the electricity consumption of a household at all times, we propose a total of 5 alternative models of increasing complexity: 2 "climatology"models, 2 temperature-dependent models, and 1 additive model. These models are meant to be used in a hierarchical manner to always provide the most accurate forecast depending on the situation. Additionally, a reference model based on machine learning is introduced as a benchmark. The models' parameters are fitted to the data from the training period, so as to keep out-of-sample the data from the test period (Tashman, 2000). Each model is probabilistic and produces a set of forecasts for instant t at quantile levels $\tau = 0.1, 0.2, \ldots, 0.9$. The median probabilistic forecast at level $\tau = 0.5$ is used as the point/deterministic forecast.

[6]In fact, ECMWF provides only one forecast value every 3 hours, and, hence, the gap hours are filled with a linear interpolation.

Climatology Models We create a "climatology" type of model for each one of the 226 households. This kind of model was early introduced in the weather community (Murphy, 1977) and consists in computing quantile forecasts based on all the historical observations unconditionally. In our case, all the demand measurements of the training period made on a given day of the week and hour of the day are used to compute fixed quantile values for this hour and day, independently of the recent demand values or the weather conditions. This method means that the forecasts for every Monday are always the same, be it in August or in December. The 1512 ($7 \times 24 \times 9$) values computed from the training period provide a quantile forecast of the demand for any future instant t, noted ${}^{c}\hat{y}_t^{\tau}$.

This climatology model is then referred to as M_0^i for household $i = 1, \ldots, 226$, or just M_0. Additionally, we create an average climatology model, referred to as A_0, based on the mean demand time series.

Temperature-dependent Models Since the temperature time series is retrieved from a different source than the smart meter measurements, the presence of this input is expected to have a different reliability. Usually, given a good internet connection, the availability of Numerical Weather Predictions is high. They are also provided several times per day and even if once they are not available one can use forecasts from previous runs of the NWP model. For this reason, it is useful to design a forecasting model relying solely on this information. Quantile smoothing spline functions are fitted[7] by optimizing with the quantile score as a loss function. Since the temperature has a different impact on demand depending on the hour of the day, a total of 24×9 functions $a_h^{\tau}(\cdot)$ are fitted, for $h = 0, \ldots, 23$, so that

$$ {}^{\theta}\hat{y}_t^{\tau} = a_h^{\tau}(\hat{T}_t) \tag{4.16} $$

is the quantile forecast of the demand y_t at level $\tau = 0.1, \ldots, 0.9$, where the instant t to be forecast is associated with the hour h of the current day. In practice, the function is not fitted to the actual demand y_t, but rather to the residual errors after shifting the demand value by the median climatology forecasts. Our experiments, non reported here, show that proceeding as such slightly refines the spline fitting process.

[7] The fit is done with function `rqss` implemented in the **R** package *quantreg* (Koenker, 2012).

This temperature-dependent model is then referred to as M_1^i for household $i = 1, \ldots, 226$, or simply M_1. An average temperature-dependent model, using the mean demand time series, is also fitted, and noted A_1.

Additive Model Three independent quantile smoothing spline functions are fitted[8] to the data of the training period to reflect the effects of three inputs: the demand measured 24 hours before, the median demand during the 7 previous days, and the temperature forecast. An additive structure is selected to simplify the fitting process, and a fit is done for each hour of the day h, so that

$$\hat{y}_t^\tau = b_h^\tau(\hat{T}_t) + c_h^\tau(y_{t-24}) + d_h^\tau(\bar{y}_t) \tag{4.17}$$

is the quantile forecast of the residual error y_t at level $\tau = 0.1, \ldots, 0.9$, where the instant t to be forecast is associated with the hour h of the current day. Similarly as for the temperature-dependent model, the fitting is made on the residual errors rather than the actual demand. The fitting process for the 6,552 points of the training period is fast, i.e. less than 5 seconds on an average 2013 laptop. In the literature, this kind of additive framework proves efficient when forecasting electricity demand (Gerossier, Girard, et al., 2017).

This additive model is then referred to as M_2^i for household $i = 1, \ldots, 226$, or simply M_2. No average model is created because it would involve gathering individual smart meter data in real time in order to compute the mean demand time series. Such gathering is strongly invasive of privacy and thus to be avoided (McKenna et al., 2012). Advanced methods to protect user privacy exist, such as employing a consensus framework (Boyd et al., 2011), but are not considered in this study.

Reference Model Based on Machine Learning Additionally, we train a gradient boosting model that makes use of the same inputs as the additive model, i.e. the demand measured 24 hours before, the median demand value during the 7 previous days, the temperature forecast, and the hour of the day. A total of 9 versions are computed for quantile levels $\tau = 0.1, 0.2, \ldots, 0.9$. The meta-parameters of the gradient boosting model are adjusted in such a way that the computation time for the training phase is approximately the same as for the additive model, i.e. about 5 seconds. This

[8] The fit is done with function `rqss` implemented in the **R** package *quantreg* (Koenker, 2012).

gradient boosting model is then referred to as G_2^i for household $i = 1, \ldots, 226$, or simply G_2. This machine learning model is used as a benchmark due to its established performance (S. Ben Taieb & Hyndman, 2014). Note that this black-box model cannot be used in the demonstration project due to its somewhat obscure behavior.

4.4.2.4 Forecasting Performance Scores

To assess the performance of a forecasting model, we compare the forecast values with the observations during a test period, i.e. for $t \in \{1, \ldots, T\}$. We compute three common indices to assess the deterministic performance, with $\hat{y}_t$ the point forecast for instant t, and y_t its corresponding observations. First, the Normalized Mean Bias Error (NMBE)

$$\text{NMBE} = \frac{1}{T} \sum_{t=1}^{T} \frac{y_t - \hat{y}_t}{\text{mean } y_t}, \tag{4.18}$$

should be close to 0, then, the Normalized Mean Absolute Error (NMAE)

$$\text{NMAE} = \frac{1}{T} \sum_{t=1}^{T} \frac{|y_t - \hat{y}_t|}{\text{mean } y_t}, \tag{4.19}$$

and the Normalized Root Mean Square Error (NRMSE)

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{T} \sum_{t=1}^{T} (y_t - \hat{y}_t)^2}}{\text{mean } y_t}, \tag{4.20}$$

should be as low as possible.

For the probabilistic performance, we first calculate the reliability (Rel) between two successive quantile levels $\tau_0 = 0 < \tau_1 < \cdots < \tau_{K+1} = 1$

$$\text{Rel}_k = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}(\hat{y}_t^{\tau_{k-1}} < y_t \leq \hat{y}_t^{\tau_k}), \tag{4.21}$$

for $k = 1, \ldots, K+1$, where $\mathbf{1}$ is the Heaviside function, and $\hat{y}_t^{\tau}$ is the forecast quantile at level τ. To ensure that the forecast distribution is reliable, or calibrated, the reliability for interval k must be close to the theoretical frequency $\tau_k - \tau_{k-1}$. This frequency is never exactly observed due to natural statistical fluctuation, so Candille and Talagrand propose a reliability ratio Δ/Δ_0 that quantifies how well-calibrated the forecast distribution is, see (Candille & Talagrand, 2005, Section 3). In addition to the

reliability, we compute the Normalized Quantile Score (NQS) to check the accuracy of the probabilistic forecasts. Specifically,

$$\mathrm{NQS}_\tau = \frac{1}{T}\sum_{t=1}^{T} \frac{2(\mathbf{1}(y_t \leq \hat{y}_t^\tau) - \tau)(\hat{y}_t^\tau - y_t)}{\text{mean } y_t}. \tag{4.22}$$

The NQS_τ is negatively oriented: a lower value indicates a better performance at quantile level τ. Note that $\mathrm{NQS}_{0.5} = \mathrm{NMAE}$.

4.4.3 Hierarchical Forecasting Framework

We first select a subset of 20 households with high-quality smart-meter data to assess the performance of each forecasting model. Then, we identify the problematic situations occurring in practice, before finally designing a hierarchical forecasting framework combining the models based on their respective performance and robustness to problematic situations.

4.4.3.1 Offline Forecasting Performance of a Subset of Households

For each household, we have 6 alternative day-ahead forecasting models, A_0, A_1, M_0^i, M_1^i, M_2^i, and G_2^i. Based on their respective level of complexity and the forecasting literature, we expect similar performance from G_2^i and M_2^i, and that both will outperform M_1^i, then M_0^i, then A_1, and then A_0. We wish to assess their respective performance during the test period going from 1st October to 31st December 2015. To perform this evaluation, we select a subset of households based on two criteria:

1. The availability of the smart-meter data of the household should be almost perfect. We only retain households whose demand data are available at least 95% of the time in both the training and the test periods.

2. There should be no abrupt change in demand patterns between the training and the test periods. This is assessed by examining the climatology probabilistic forecasts computed during the training period. With such a model, the reliability of the forecast should be fairly correct during the test period when no abrupt changes occur. Therefore, we check that the reliability ratio defined by Candille and Talagrand, see Section 4.4.2.4, is close to the ideal ratio of 1. Somewhat

arbitrarily, we choose that a household passes this reliability test when the ratio is below 20.

A subset of only 20 out of the 226 households fulfill the two criteria, later denoted subset Ξ. In fact, most of the 226 households exhibit abrupt changes in their demand patterns that are quite difficult to anticipate, and that do not reflect the intrinsic performance of the forecasting model. For the 20 households in the subset Ξ, we compute the forecasting performance scores defined in Section 4.4.2.4, for the 6 models introduced. The average results are shown in Figure 4.21 and in Table 4.4.3.1.

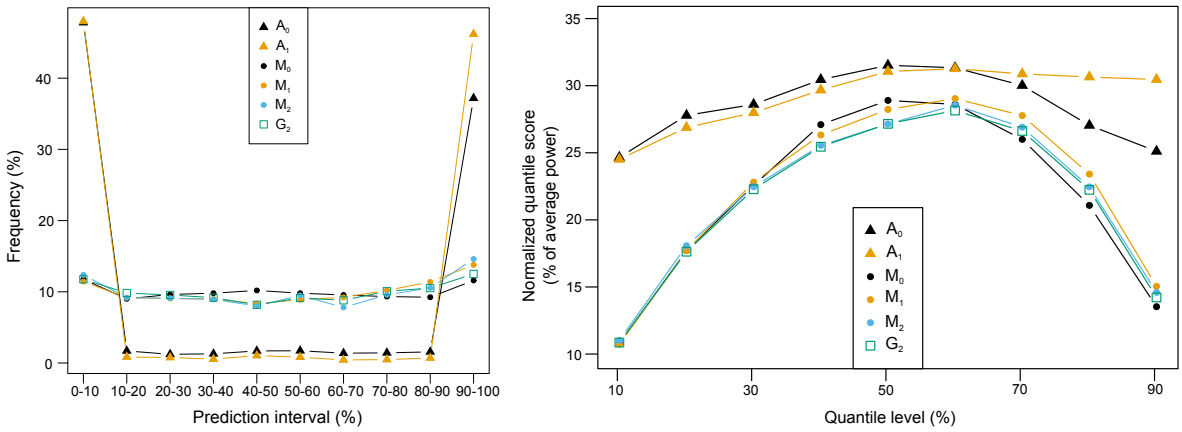


Figure 4.21 – Reliability graph (left panel) and quantile score curves (right panel) for the 6 models in the selected subset Ξ.

We first examine the deterministic performance. Regarding the NMAE, we see that the models' performances are ordered as expected, with a top performance of 27.2% for M_2. The hypothesis that all 6 models have similar performance is rejected according to the Friedman statistical test (p-value of 10^{-4}) (Fan et al., 2018). Additionally, we note that the most efficient model M_2 has similar performance to the reference model G_2 when comparing the individual household errors: the nonparametric Wilcoxon test does not reject the null hypothesis claiming similar performance (p-value of 0.54) (Derrac et al., 2011). The NRMSE are slightly larger than NMAE, with a minimum of 35.0% obtained with the G_2 model. The performance order is marginally altered: the temperature-dependent models are poorer than their climatology counterpart, and the machine learning model is found more efficient than the additive model. The difference between NRMSE and NMAE is due to the fact that

Table 4.6 – Median performance indices (in %) and reliability ratio for various day-ahead forecasting models among the subset Ξ.

Type	Model	NMBE	NRMSE	$NQS_{0.1}$	NMAE	$NQS_{0.9}$	Δ/Δ_0
Average climatology	A_0	-2.9	38.7	24.7	31.6	25.2	741
Ave. temperature-dependent	A_1	**1.2**	38.8	24.6	31.1	30.6	871
Specific climatology	M_0	2.7	36.7	10.9	29.0	**13.6**	**6**
Spec. temperature-dependent	M_1	6.2	37.1	**10.8**	28.3	15.1	11
Spec. additive	M_2	4.6	35.9	11.1	**27.2**	14.7	11
Machine Learning	G_2	5.3	**35.0**	10.9	**27.2**	14.3	8

quadratic errors strongly penalize large deviation between forecasts and observations. Such deviations are fairly common for the electricity demand since demand distribution exhibits heavy tails. Furthermore, this demand distribution is usually positively skewed, meaning that the upper tail is longer than the lower tail. This positive skew means that the point forecasts — that are optimized on the median values — underestimates the real values (Groeneveld & Meeden, 1977). This positive skew is visible on the NMBE obtained with our models. We note that, on average, the models specifically trained for households decrease the errors by around 10% in comparison with the average models.

We then examine the probabilistic performance. When looking at the reliability ratio, we observe that the specific models are reasonably calibrated — with reliability ratio between 6 and 11 — but that the average models are not — ratio above 700. The whole forecast distribution of the average models either overestimates or underestimates the demand. Consequently, providing point forecasts of the demand of an unknown household is reasonably efficient — NMAE around 31.1% —, but providing average probabilistic forecasts makes no sense and requires specific measurements of the corresponding household. The quantile score curves, visible on the right panel in Figure 4.21, depict the performance at different quantile levels, i.e. for different parts of the forecast distribution. We remind that the NMAE scores are readable at quantile level 50%. The curves crossings between the models suggest that forecasters should use the additive model for lower quantile levels (10—60%) and then switch to the specific climatology model for higher levels (70—90%). This observation highlights that it is, perhaps surprisingly, more efficient to carry out conservative forecasts for the upper part of the forecast distribution. However, this conclusion should be adapted depending on the household considered. For instance, for about one third of the households, the models with a temperature input, i.e. M_1 and M_2, clearly outperform the climatology M_0 at all levels. Identifying these households that benefit from the temperature input is quite straightforward: they are equipped with heating or cooling electrical devices, i.e. they have clear thermal sensitivity (Gerossier, Girard, et al., 2017). This sensitivity is measured by retrieving the correlation between the electricity demand and the outside temperature. Thermal sensitivity is defined as the squared correlation and so a high (resp. low) sensitivity depicts a strong (resp. weak) demand–temperature correlation. The households with high sensitivity show a clear increase in electricity demand when

it is cold outside. In these cases, the spline functions fitted to the temperature visibly show the effect. In particular, one sees in Figure 4.23 that the sensitivity depends on the hours considered, i.e. the occupancy of the house. Consequently, the forecasts are more accurate as illustrated in Figure 4.22, where the evening demand is well anticipated by the temperature-dependent model M_1 (orange) since it is a cold day, but not by the climatology model M_0 (black).

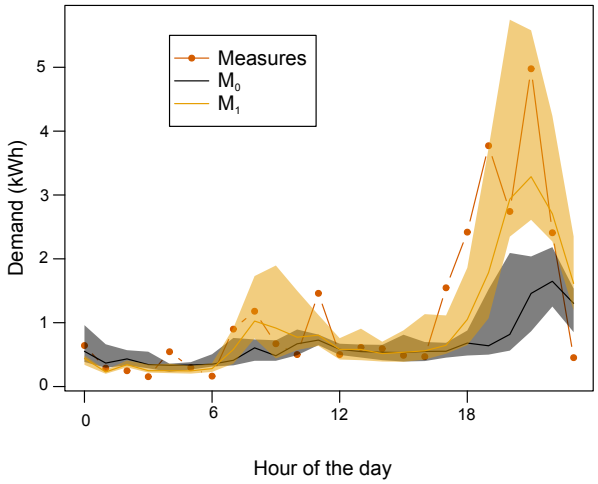
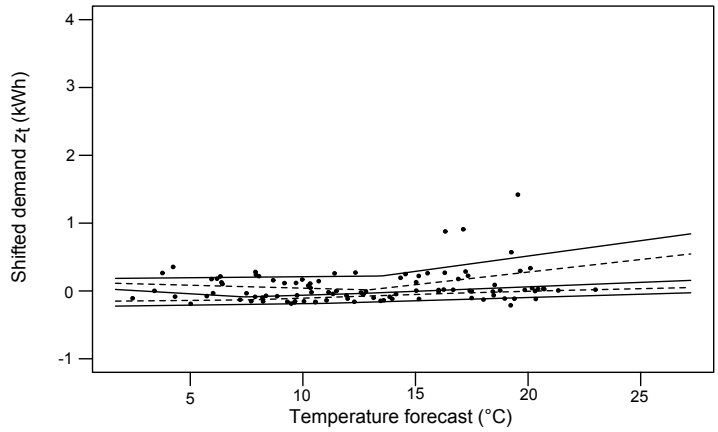


Figure 4.22 – Day-ahead forecasts of hourly demand of an individual household on Sunday 22nd November 2015 with the specific climatology model M_0 and the specific temperature-dependent model M_1: solid lines depict the median forecast, and the filled-in areas show the interval prediction 30—70%. The actual demand measurements are represented by the red line connecting the circles.

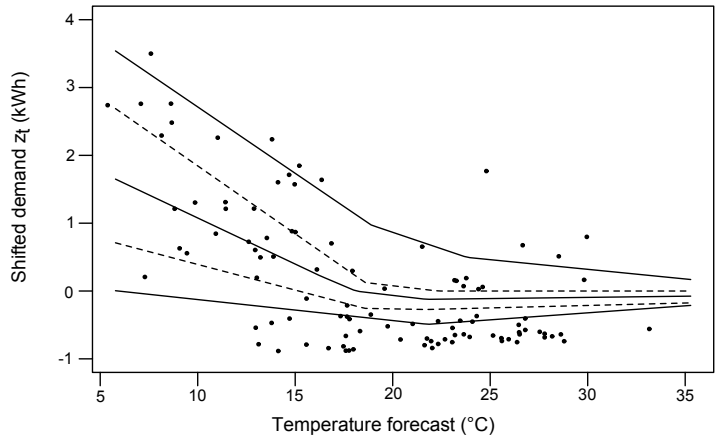
4.4.3.2 Problematic Situations

Although the additive model provides the best performance, it is also the least robust model and a number of problematic situations occasionally prevent its usage. This is often the case for similar type of models based on time-series approach. The following situations are identified to be problematic when forecasting the demand of household $i \in \{1, \ldots, 226\}$:

- *No data in the training period.* There is no way to create the specific models M_0^i, M_1^i, and M_2^i.



(a) 03:00



(b) 20:00

Figure 4.23 – The functions fitted to forecast the demand 03:00 and 20:00, given the temperature forecast, see Equation (4.16). The lines represent the functions fitted at quantile levels $\tau = 0.1, 0.3, 0.5, 0.7, 0.9$. The points represent the actual observations of the training set.

- *No temperature forecast.* Models making use of the temperature A_1, M_1^i, and M_2^i are missing an input and cannot properly carry out a forecast.
- *No recent measurements.* Input values y_{t-24} or $\bar{y}_t$ are then unavailable, meaning that M_2^i cannot operate.
- *Unknown situation.* A drawback of the smoothing splines is that extrapolation is known to perform poorly, affecting the activation of A_1, M_1^i, and M_2^i. For instance, if recently observed demand values have never been this low in the

training set, it is better to refrain from using the additive model M_2^i.

4.4.3.3 Hierarchical Framework

Flowchart The respective performance of each model coupled with the identification of problematic situations enable us to design a forecast hierarchical framework represented in Figure 4.24. In the implementation, when producing a forecast for instant t for a household i, we successively check:

1. Are there historical measures specific to this household?
2. Is there a temperature forecast $\hat{T}_t$ available?
3. Are the recent measures y_{t-24} and $\bar{y}_t$ available?
4. Is the future situation known, i.e. do the inputs values extrapolate from the ones that occurred during the training period?

Performance We implement the hierarchical framework for each of the 226 households in the neighborhood. The flowchart detailing the model usage according to the situation allows us to always provide day-ahead probabilistic forecasts for each hour of the day in the test period — from 1st October to 31st December 2015. We assess the performance by comparing these forecasts to the available data. Since some households have missing demand measurements, the length of the test period is not exactly the same for all the households. For instance, one household has no measurement at all in December and so the performance is estimated with a test subperiod going from 1st October to 30th November.

Figure 4.25 depicts the NMAE observed for each hour of the day among all of the 226 households. The points show the median NMAE, and the segments show the variation 20—80% among households. The errors follow the same trend as the actual demand values: lower in the nighttime, and higher in the evening. However, the fluctuation throughout the day is minor. Since all the forecasts are carried out at 12:00 on the previous day, forecasts for a specific hour of the day represents a specific horizon. It means that errors at 0:00 correspond to a forecasting horizon of 12 hours, errors at 1:00 correspond to a forecasting horizon of 13 hours, and so on.

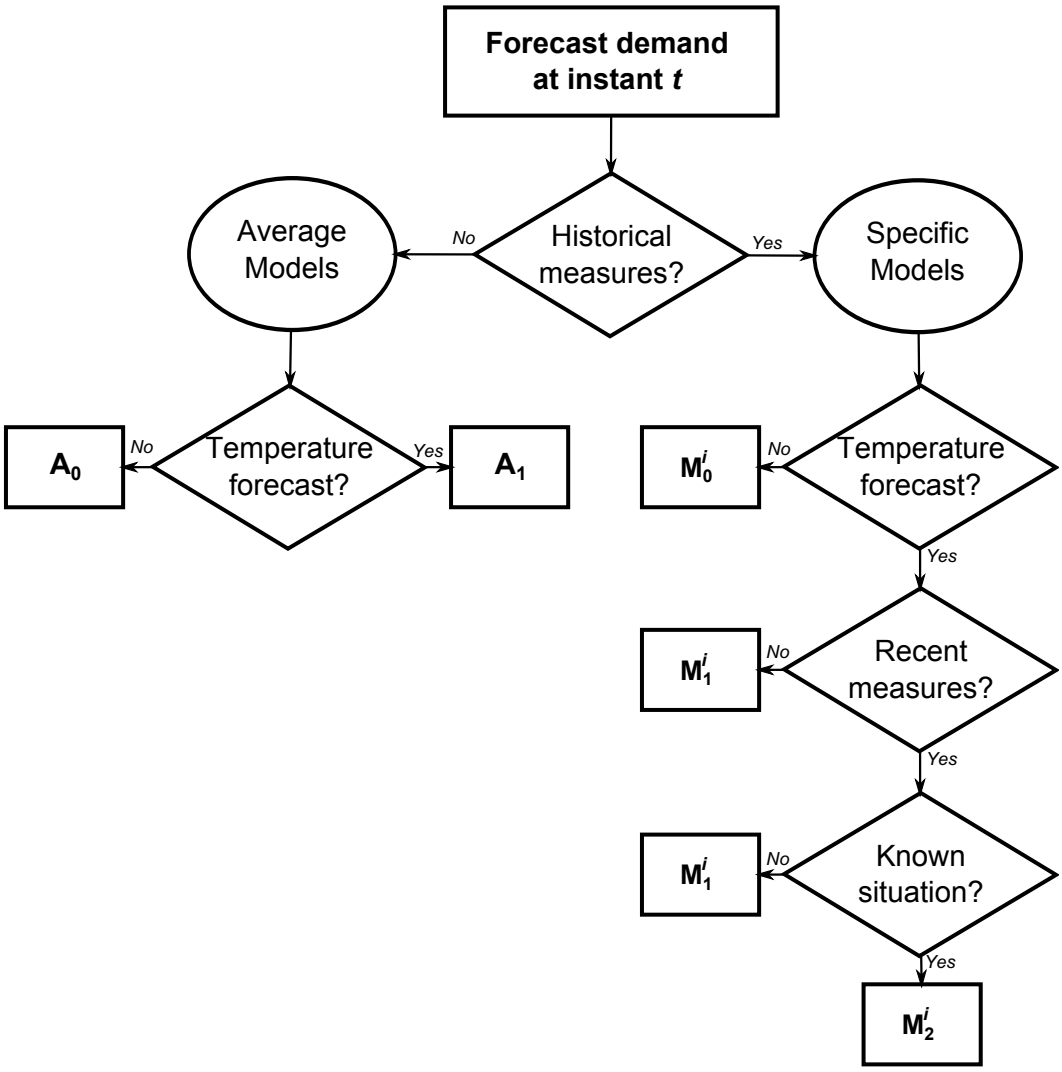


Figure 4.24 – Flowchart of the hierarchical framework indicating which forecasting model is used.

We then represent the NMAE, averaged over the 24 hours, as a function of the thermal sensitivity in Figure 4.26. The households in the subset Ξ are represented by the orange dots, and the rest by black dots. We can see that the model performs slightly better on the subset Ξ: the median NMAE decreases from 29.9% to 27.7%. The graph also logically shows that households with greater thermal sensitivity are easier to forecast. Additionally, we can see that performances greatly vary between households with similar sensitivity: errors range from 2% to 51% for low sensitivity (below 0.1). This is due to the unknown behaviors of the householders and other cultural factors,

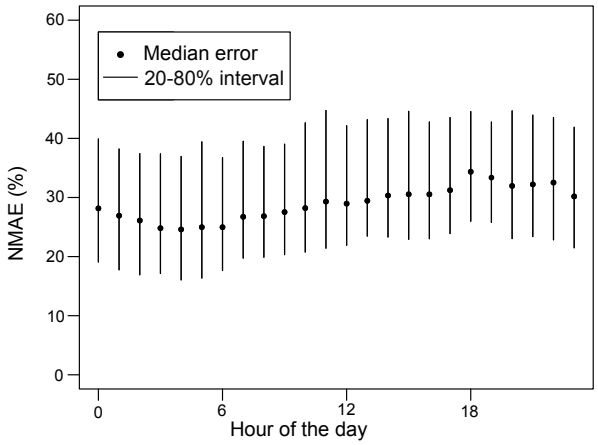


Figure 4.25 – Hourly errors distribution (NMAE in % on the y-axis) depending on the hour of the day (x-axis).

e.g. the number of appliances in the house. It highlights that anticipating a forecasting performance for a different use case should be done with caution.

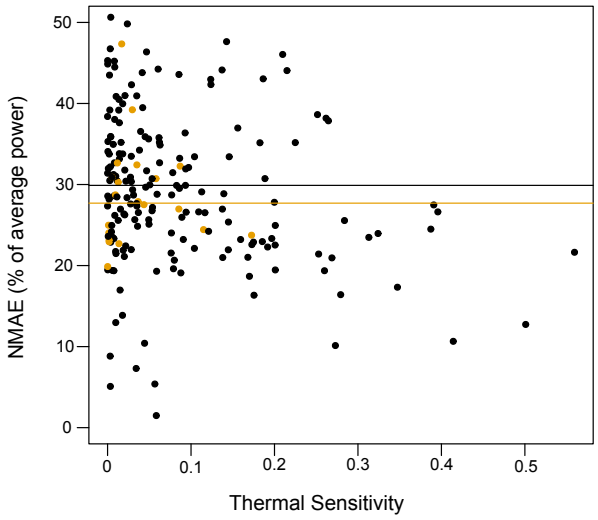


Figure 4.26 – Forecasting performance (y-axis) for each of the 226 households, regarding their respective thermal sensitivity (x-axis). The 20 households of the selected Ξ subset are depicted in orange, and the rest in black. The lines represent the median value of the households.

4.4.4 Offline and Online Performances

We first draw a household-by-household comparison of the offline and online forecasting performances. Then, we discuss and quantify in detail the factors that cause a noticeable performance degradation with precise test cases.

4.4.4.1 Performance Comparison

The hierarchical forecasting framework is implemented at the Évora demonstration site. The forecasts produced and smart-meter measurements are retrieved, providing a recent online dataset. This dataset is made up of two parts: a training period going from July to December 2017, and a test period from April to August 2018.

We first analyze the frequency with which each one of the 5 models that compose the framework, depicted in the flowchart in Figure 4.24, are activated as a function of the available data. The results are given in Table 4.7. It is noted that, at each instant, a single model produces the final forecast, according to the situation. The most efficient model M_2 is activated in about three quarters of the cases. We observe similar model activation frequencies in the online and offline cases.

Model	Offline	Online
A_0	0	0
A_1	3	0
M_0	3	3
M_1	18	19
M_2	76	78

Table 4.7 – Average usage frequency (rounded in %) of the various models on the offline dataset (226 households) and on the online dataset (20 households).

The online data is collected from the 20 households of the Ξ subset introduced in Section 4.4.3.1. Figure 4.27 compares the performance of these 20 households obtained during online test period — 1st April to 31st August 2018 —, and during the offline test period — 1st October to 31st December 2015. We compare the NMAE obtained during the two periods[9] with our forecasting framework and divide this error by the NMAE

[9]Note that the normalization in the NMAE score comes from the mean value observed from the

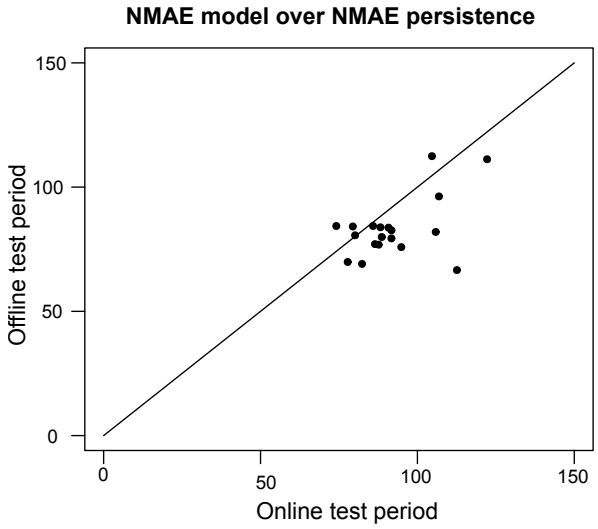


Figure 4.27 – Each point represents the household performance on the online test period (x-axis) — 1st April to 31st August 2018 — compared to the performance on the offline test period (y-axis) — 1st October to 31st December 2015. The performance is the ratio between the NMAE obtained with the forecasting framework and the NMAE obtained with a 1-day persistence model.

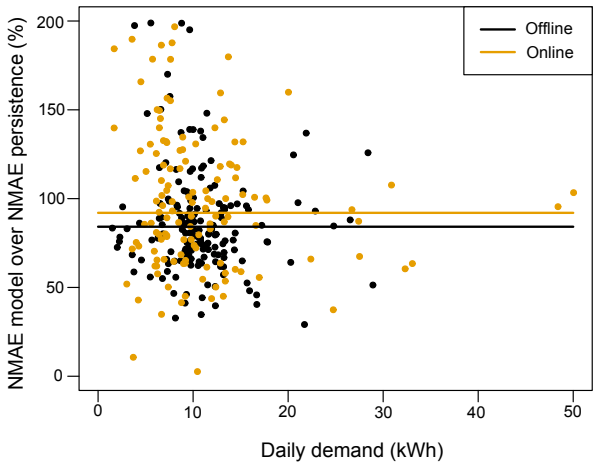


Figure 4.28 – Each point represents the forecasting performance computed over a single household and single day. The NMAE ratio between our model and the persistence model (in %) is on the y-axis, and the total daily demand (in kWh) is on the x-axis. The horizontal lines represents the average performance over all households and all days.

sets studied, and so the normalization value evolves between the offline and online test sets.

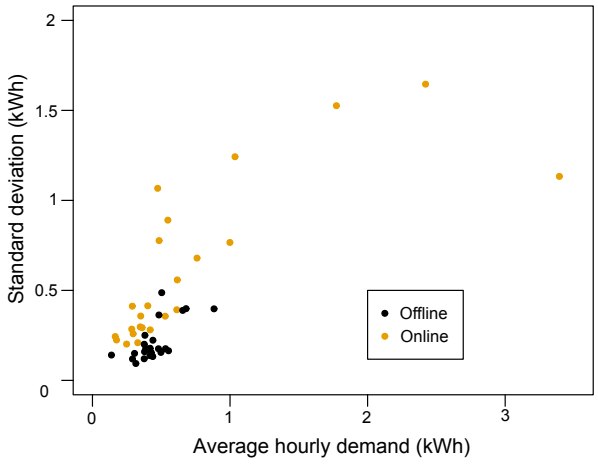


Figure 4.29 – Characteristics of the individual time series of the 20 households in the offline (black points) and online (orange points) cases. The standard deviation of the series (y-axis) is represented in regard with its mean hourly demand (x-axis).

obtained with a 1-day persistence model. For most households, the errors made by our model is lower than the persistence errors (average of 0.90 offline and 0.97 online). Furthermore, for 17 out of 20 households, the individual NMAE obtained offline is lower than online, meaning that the model performance has decreased between the two test cases. We also provide in Figure 4.28 the NMAE computed over a single day. Each point, in black for the offline case and in orange for the online test, represents the ratio between the NMAE of our forecasting framework and the NMAE of the persistence model (y-axis). The daily demand of the day (in kWh) is represented on the x-axis. We see that the daily performance is more volatile when the demand of the day is low than when this demand is important. In fact, this performance volatility is due to the persistence forecasts performance that also widely range for low-demand day: performance is either very good (when the previous day is also a low-demand day) or very poor (when the previous day is not a low-demand day). The improvement over persistence is more clear for high-demand days in online and offline cases.

On average, the online performance is worse than the offline performance. In absolute values, the average NMAE goes from 34.8% on the offline test to 58.5% on the online test. This comes from the demand characteristics that are quite different between two cases. Figure 4.29 provides an indicative illustration. For the same set of households in the two cases, one point represents the average hourly electricity de-

mand of the household (x-axis) and its standard deviation (y-axis). Both the mean and deviation largely increase between the two cases. This evolution directly influences the forecasting performance since it denotes the usage of more appliances, hence more demand volatility and forecasting complexity.

4.4.4.2 Discussion

We investigate the possible reasons for the performance degradation between the offline and online tests: the evolution of the demand time series, the availability rate in the test period, the duration and recency of the training period, the position during the year of the test period. The subsequent tests are made using our offline 2015 dataset with the Ξ subset of 20 households to quantify the possible performance degradation.

Evolution of the Demand Since there is a considerable time gap between the offline test, in 2015, and the online test, in 2018, the behaviors of the householders living in the 20 households have evolved: new people, new appliances, new habits, etc. This evolution is reflected in the electricity demand patterns which modify the intrinsic complexity of the forecasting task. Defining this complexity is not straightforward: we examine the performance of a 1-day persistence model — by which we use the demand measured on the current day to provide point forecasts for the next day. We observe that this persistence model has an average NMAE of 45% from April to August 2015, and this error increases to 69% from April to August 2018. This means that forecasting the 2018 time series is roughly 50% more difficult than forecasting the 2015 time series.

Availability Rate in the Test Period For each one of the 20 households in the Ξ subset, we randomly discard a certain amount of available measurements in the test set, obtaining an availability rate between 0 and 1. This mimics the case when a specific hourly observation is missing, and so the forecast cannot be compared to the actual observation. We compute the forecasting performance of M_2 with the NMAE and $NQS_{0.9}$ indices on the available subperiod. In Figure 4.30, we represent the performance fluctuation (in %) regarding the availability rate. Logically, we see that the average performance is constant, i.e. at a reference level of 100%, whatever the availability rate. However, note that the missing values introduce variability in the performance evaluation. This variability logically increases when the availability rate decreases. It

goes up to 2% when examining the NMAE. This effect is emphasized for the distribution tails, as seen on the $\mathrm{NQS}_{0.9}$ going up to 4 % for low rates, that are more difficult to estimate accurately.

We conclude that missing values in a test set induces limited performance fluctuation. However, the missing values here are assumed to be uniformly spread throughout the period, which is the case in the actual online dataset retrieved. Another use case may result in different missing value distribution, e.g. when a smart meter is disconnected during a contiguous period of time.

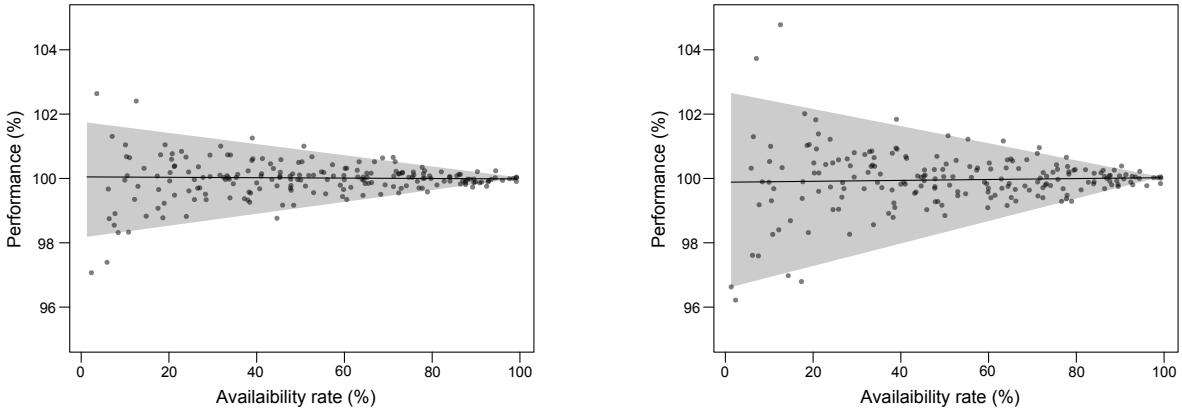


Figure 4.30 – Variety of the performance (y-axis) according to the data availability in the test period (x-axis). One point represents one trial run for a given availability rate. The solid line represents the median spline, while the grey filled zone represents the confidence interval 5—95% induced by the availability randomness.

Training Period Position For each of the 20 households in the Ξ subset, we train the forecasting models M_2 and G_2 at quantile level 50% with different training periods. Figure 4.31 represents the average NMAE achieved on the test period, fixed from 1st October to 31st December 2015, relatively to the minimal NMAE obtained with the longest training period going from January to September. The beginning of the training period is selected on the x-axis, and the end is selected on the y-axis. The left panel represents the performance with the M_2 model while the right panel represents the performance with the G_2. Since the additive model M_2 is not designed for extrapolation, the training period necessarily should include the first months of the year, so as to observe similar temperature as during the test period, to produce forecasts. It

means that only a limited range of training periods could be evaluated. On the other hand, the machine learning model G_2 is designed for such extrapolation, so we can extend the performance on more diverse training periods. While both models produce the same performance when using the 9 months (January to September) as training sets, we see that G_2 does a better job with reduced periods. We logically see that reducing the duration of the period damage the performance of both models. We see that the degradation can be up to 10% for M_2 when the period lasts only 3 months (February to April) with a time gap between training and test, instead of 9 months (January to September).

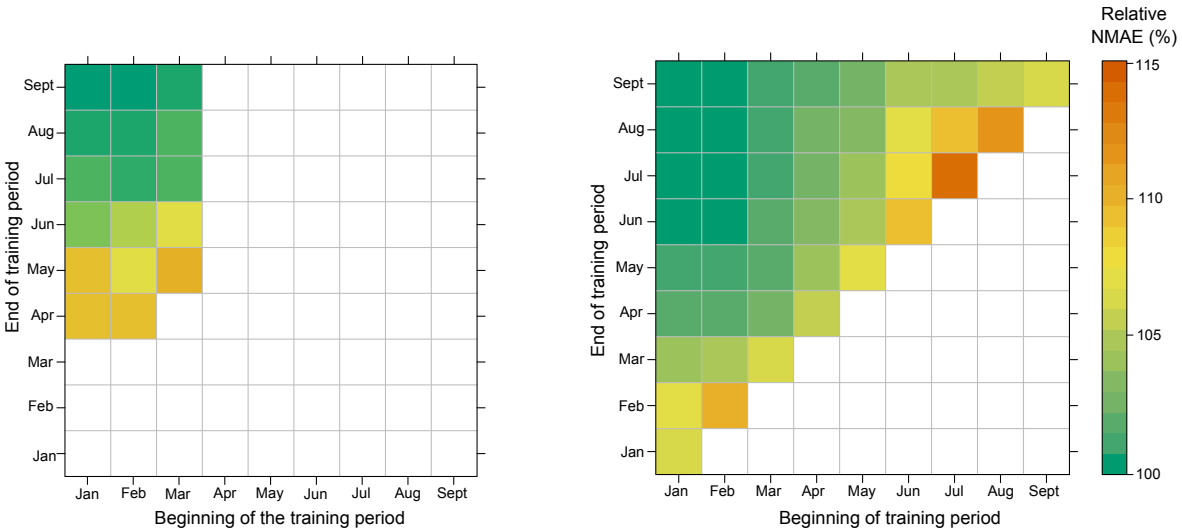


Figure 4.31 – Forecasting performance depending on the exact period of the training set, i.e. the beginning of the training period (x-axis), and its end (y-end). For each training period, the relative NMAE is equal to the average NMAE over the Ξ subset divided by the minimal NMAE obtained with the maximal training period. The left panel represents the results obtained with M_2, the right panel represents the results with G_2.

We conclude that training with the all of the data, and using as recent data as possible, is the best way to grasp the various recent demand patterns. Furthermore, we stress the importance of using data collected during similar situations to those to be forecast, especially regarding the temperature. For instance, to efficiently forecast summer 2018 ideally means training the model with data collected in summer 2017.

Test Period Position The test period's position in the year impacts the performance. Figure 4.32 represents the forecasting performance with model M_2 obtained using, in turn, each month of the year 2015 as the test period, using the remainder as the training period[10]. For each household in the Ξ subset, the NMAE obtained for each month of the year is divided by the average over the whole year, so as to obtain a relative NMAE. The boxplot representation indicates the variation in the subset. We can see that, on average, the summer period, i.e. June to August, produces a slightly better performance than the other months, with a NMAE decrease of around 5%.

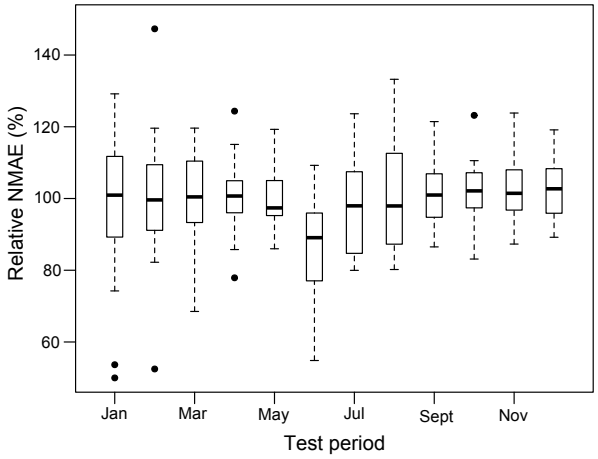


Figure 4.32 – Boxplot of the forecasting performance depending on the exact test period. Each month of the year is, in turns, selected as the test period while the rest of the year is used as the training period. For each household in the Ξ subset, the NMAE obtained for each month is divided my the mean value obtained across the 12 months.

4.4.4.3 Summary

As a reminder: (1) the offline training period goes from 1st January to 30th September 2015, the offline test period from 1st October to 31st December 2015, and the offline NMAE is 34.8%; (2) the online training period goes from 1st July to 31st December 2017, the online test period from 1st April to 31st August 2018, and the offline NMAE is 58.5%.

[10]This framework implies that, while the test period is always out-of-sample, it is surrounded by the training period, which prevents any major deviation, possible in a real case.

We identify that the main cause of this 68% relative performance degradation is due to the intrinsic evolution of the time series. Thanks to a simple persistence forecasting model, we assess that the demand time series in the online case are roughly 50% more difficult to forecast than those of the offline case. To a great extent, we remove this intrinsic time series evolution by analyzing the performance improvement of the forecasting framework over the persistence model. On average, we have seen that the NMAE is reduced to 90% of the persistence NMAE in the offline dataset, but only 97% in the online dataset. This remaining relative performance discrepancy of 8% is due to the mismatch of the training and test period positions in the online case. In fact, the models are trained with fall data, but tested with spring data, which causes a relative degradation of around 15%. This effect is counterbalanced by around 5% due to the position of the test period, since the spring period (online case) is easier to predict than the fall period (offline case).

4.4.5 Conclusion

We present 5 probabilistic forecasting models that employ small input sets — day of the week, hour of the day, recent smart-meter data, temperature prediction — to produce day-ahead forecasts of electricity demand at the household level. We compare the performance of the models on an offline dataset collected at a demonstration site in a Portuguese neighborhood. We observe that the more flexible, and thus more complex, model logically results in better overall performance, similar to that of a machine learning benchmark.

However, many problematic situations arise and prevent the usage of this flexible model in real time. We therefore propose a hierarchical forecasting framework, combining the 5 models introduced, that addresses the following requirements: high robustness, fast computation, easy replicability, remote control, and easy interpretation. These requirements are essential for deployment of a forecasting model for a large number of households in real-world applications. After deployment in 2018, in the demonstrator in the frame of SENSIBLE project, the feedback data collected at the demonstration site are analyzed in order to provide an online forecasting performance. A household-by-household comparison with the performance assessed using an offline dataset shows a considerable relative degradation. We quantify the possible reasons

for this degradation. Although it is due, in part, to the mismatch between the online training and test periods, the main cause is the evolution of the demand. From the distance in time between the initial offline testing of the model and its implementation for real operation, we observed an evolution of the characteristics of the physical process itself. The complexity of the demand pattern has greatly increased, meaning that the forecasting task is found to be about 50% intrinsically more complex during the online test. This observation highlights the fact that assessing forecasting performance at the household level is challenging. While forecasting performance has been observed to vary greatly between two households, even when located in the same neighborhood, our experimental feedback shows that this performance also significantly evolves with time. This evolution is caused by unknown abrupt characteristics changes in the household, such as new people, additional appliances, changing habits of the householders, etc.

This raises the question of the adaptability of forecasting models at the household scale. We recommend incorporating the most recent data into a training period, to which the forecasting models are regularly fitted. The regularity of this training process can be quite coarse, e.g. every month, since most recent demand patterns are only slight deviations of older ones. Such a framework still implies a degree of model maintenance, like reviewing the validity of the most recent smart-meter data recorded and starting the training process. A more intricate issue is caused by occasional abrupt changes in demand patterns. These changes are difficult to observe solely from the electricity demand time series. We advise using external input information about such changes, e.g. moving-in of new householders, in order to discard obsolete data and train using only smart-meter data recorded after the changes.

Chapter 5

Forecasting Electricity Demand with Scenarios

Summary Probabilistic forecasts are usually generated independently from one instant to the next. However, the household electricity demand between two consecutive instants is strongly correlated. For instance, when the demand is more important than expected at 15:00, the demand tends to also be more important than expected at 16:00. This correlation is due to the house occupancy and the electricity-related activities spreading over long periods. It means that, even if the two probabilistic forecasts for the 15:00 and 16:00 demand are optimal when evaluated separately, their combination is usually sub-optimal. Demand scenarios addresses this issue by providing a coherent set of e.g. 24 values for the 24 hours of the day. In Section 5.1, a scenario generation is proposed. We conclude that atleast 400 daily scenarios are necessary to reach optimal performance for each independent forecasts, while ensuring the overall consistency of the successive values. The demand scenarios are then used as inputs in later applications, such as optimal battery scheduling. Since those applications can require important computation time, only a few number of representatives scenarios can be processed. We introduce a method to reduce the number of scenarios, e.g. from 400 to 5, relying on an original metrics based on the hourly electricity price on the market. A comparison is drawn between the performance resulting from the usage of the reduced and complete sets of scenarios. On the other hand, in Section 5.2, we focus on the demand of a single appliance, namely the battery charging of an electric vehicle. A precise study of the appliance usage is made thanks to minute-by-minute

power data collected at the charging point. A stochastic model then describes the collection of all the start-up time and charging duration observed for one user. This model enables to generate forecasting scenarios of the next-day demand due to the electric vehicle. These forecasts result in fairly good performance, with relative errors of 43%. Such a study demonstrates how appliance power collection can be turned into accurate short-term forecasting models through scenarios. Additionally, the demand scenarios anticipate how the household demand is modified when a new appliance is integrated.

Résumé Les prédictions probabilistes sont généralement produites indépendamment d'un intervalle au suivant. Pourtant, la demande électrique d'un ménage est fortement corrélée d'un instant à l'autre, étant données la présence des habitants et leurs activités. Ainsi, quand la demande à 15h00 est plus importante que prévue la veille, elle est également plus importante que prévue à 16h00. Par conséquent, même si la prédiction probabiliste à chaque instant est optimale, l'ensemble de plusieurs de ces prédictions ne l'est pas. Ce problème peut être résolu par l'utilisation de scénarios regroupant, p. ex. 24 valeurs de la demande horaire pour les 24 heures d'une journée. Dans la Section 5.1, nous proposons une méthode pour générer ce type de scénarios de demande. À partir de 400 scénarios, la qualité de chaque prédiction probabiliste évaluée indépendemment des autres est optimale, tout en respectant la corrélation interne entre les valeurs successives. Ces scénarios de demande sont ensuite utilisés come entrées par d'autres applications faisant intervenir des algorithmes d'optimisation. Comme ceux-ci nécessitent parfois un temps de calcul important, le nombre de scénarios doit rester faible tout en étant représentatifs des variations envisageables. Nous présentons une méthode de réduction, pour passer p. ex. de 400 à 5 scénarios, s'appuyant sur plusieurs métriques, en particulier une métrique liée au tarif horaire de l'électricité. Nous évaluons la performance obtenue avec cet ensemble réduit de scénarios et la comparons avec celle de l'ensemble complet. D'autre part, dans la Section 5.2, nous prédisons des

scénarios de demande pour un usage particulier, celui du rechargement de la batterie d'un véhicule électrique. Cette prédiction passe par une modélisation précise de l'usage à partir de données recueillies à l'échelle de l'usage, concrètement par l'étude stochastique de l'instant et la durée pendant laquelle un usager recharge son véhicule. De cette manière, nous sommes capables de prédire la demande du jour suivant due au véhicule électrique avec une bonne précision, c.-à-d. avec une erreur relative autour de 43%. Cette méthode ouvre la voie à des analyses prospectives pour anticiper l'évolution de la demande totale d'un ménage quand celui-ci se dote de nouveaux appareils entraînant de nouveaux usages.

5.1 Day-ahead Household Demand Scenarios

In the two previous chapters, models investigated carried out forecasts for a single time instant (e.g. future load at 15:00) independently of forecasts made at adjacent instants. Alongside with this independence approach, forecasts are increasing of probabilistic nature to account for uncertainty, as discussed by Hong and Fan (Hong & Fan, 2016). However, when forecasting electricity demand at multiple instants (e.g. for the whole day), a probabilistic approach becomes very complex. Consequently, applications using multiple demand values as inputs, such as unit commitment (Dvorkin et al., 2014), home energy management (Correa-Florez et al., 2018), or battery sizing and placement (Grover-Silva, Girard, & Kariniotakis, 2018), are not able to deal with these multiple probabilistic forecasts. Researchers therefore rely on Monte Carlo methods and sample from the independent marginal forecast distributions to generate several deterministic trajectories or scenarios, e.g. a set of 24 values for the 24 hours of the day. These scenarios are then used as inputs of researchers' application to assess its robustness and sensitivity to demand variation. However, since the developed applications may be computationally intensive, the number of scenarios should be small, raising the question of which scenarios to pick to accurately describe the future demand profile. This picking process may be done empirically, such as by creating a few unrealistic scenarios that are extreme for all hours of the day. Such heuristics usually do not observe the optimal marginal forecast distributions and, consequently, degrade the forecasting performance. More valid methods have been proposed. For instance, one generates a large number of basic scenarios and then clustering them in homogeneous groups. This is a scenario reduction method (Dupačová et al., 2003).

This scenario issue has been addressed early by the meteorological community with ensemble forecast methods: the idea is to slightly alter the inputs and the parameters of the numerical weather prediction models to carry out multiple forecast values (Leutbecher & Palmer, 2008). Later, scenarios that forecast renewable energy production were proposed. Pinson and Girard identify that scenarios for multiple lead times improve the detection of wind power production gradients compared to climatology, and that statistically generated scenarios exhibit comparable performance as ensemble method (Pinson & Girard, 2012). Bruninx and Delarue compare different methods to reduce the number of wind production scenarios with a subsequent stochastic unit

commitment problem (Bruninx & Delarue, 2016). Luis proposed a complete overview of the photovoltatic production scenarios generation and reduction applied to unit commitment (Luis, 2018). No work has been proposed for the household electricity demand specifically.

Household demand scenario is particularly challenging because the uncertainty in the future values is large (errors up to 100%). Contrary to meteorological forecasting models, which are based on precise physical modeling, the household demand models are mostly statistical and hence, ensemble methods are ill-suited to the problem. The uncertainty is more related to non-measurable human decisions than measurable external causes, such as the wind speed or the air humidity for meteorology. Pure statistical approach is therefore adequate to forecast accurate scenarios.

Thereafter, we address this question by issuing multiple scenarios describing precisely the possible variation of the next day household load. This is done in three steps. First, a machine learning forecasting model is designed to do probabilistic forecasts of the demand at every hour on the next day. Then, the marginal forecast distributions found at all hours are transformed into a scenario, i.e. a set of 24 points, through a scenario generation method. Once we have a large number of scenarios, large enough to have optimal statistical performance, a scenario reduction method is proposed to obtain a small set of representative scenarios. Reducing the number of scenario relies on a proximity distance, or proximity metric. Two original distances are crafted: a profile characteristics distance, emphasizing four key characteristics of household demand; and a price-weighted household demand distance, that is suited for household demand, and that takes into account the electricity price on the market. Performance of the scenarios are evaluated by analyzing how the scenarios anticipate the various characteristics and the future costs due to next-day load.

5.1.1 Day-Ahead Forecasting of Household Demand

5.1.1.1 Data

We retrieve hourly recordings made in the year 2017 of 175 households located in Austin, Texas, thanks to the Dataport project run by Pecan Street Inc. (*Pecan Street Inc. Dataport*, 2018). Insight is given on this set of households: most of the inhabitants are active families living in individual buildings that are rather large (average of 200

m^2) and recently built (two thirds were built after year 2000), so that average hourly demand is 1.3 kWh, i.e. annual electricity consumption of 11.5 MWh. Data is treated so that negative or absurdly high values are removed, and missing values (less than 1% of all values) are filled in with linear interpolation. The hourly demand of one household is then referred to as time series (y_t) and is expressed in kW.

5.1.1.2 Forecasting Method

For each household, a gradient tree boosting model is set up to forecast the 24 hourly demand values of the next day (Ridgeway, 2017). Six inputs are used for the model in order to do the prediction: (1) last available demand value recorded at the same hour, (2) median demand recorded at the same hour during the last week, (3) hour of the day, (4) weekday, (5) temperature forecast, and (6) exponentially smoothed temperature forecast with smoothing factor fixed at 0.35. Temperature forecasts made at 12:00 the previous day are obtained with the European Centre for Medium-Range Weather Forecasts (ECMWF). Meta-parameters of the model (number of trees, tree depth, shrinkage parameter, and tree width) are carefully tuned by balancing performance and computational time, so that model fitting takes less than 2 minutes for one household. A total of 99 trees are then fitted with loss functions equal to quantile score (see Equation (5.3) for quantile levels $\tau \in \{0.01, 0.02, \ldots, 0.99\}$. Each tree is trained and produces a single forecasting value for each hour of the day. A cross-validation approach is taken to select different training set and obtain only out-of-sample forecasts. Since the training is done independently for each quantile level, the 99 values are reordered to avoid any absurd quantile crossing situation (Chernozhukov et al., 2010), so as to obtain, for instant t, a set of forecast quantile values $\hat{y}_t^{0.01} \leq \hat{y}_t^{0.02} \leq \cdots \leq \hat{y}_t^{0.99}$.

5.1.1.3 Performance

The quality of the set of quantiles forecast by our model is evaluated with 3 scores: the Mean Absolute Score (MAE), the Prediction Interval Coverage Probability (PICP) for interval 10—90%, and the Continuous Ranked Probability Score (CRPS). The scores are computed separately for each household and averaged across the time period of one

year, minus a burn-in periods of 50 days for later usage, indexed by $t = 1, \dots, T$,

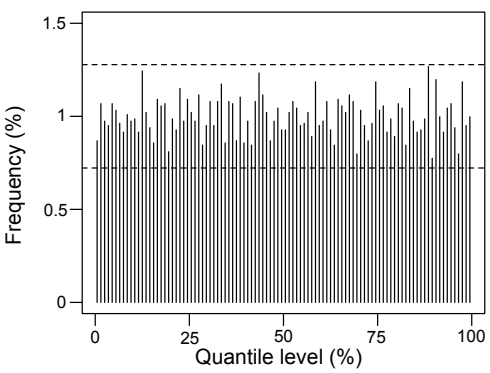
$$\text{MAE} = \frac{1}{T} \sum_{t=1}^{T} |\hat{y}_t^{50\%} - y_t|, \tag{5.1}$$

$$\text{PICP} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left(\hat{y}_t^{10\%} < y_t \leq \hat{y}_t^{90\%}\right), \tag{5.2}$$

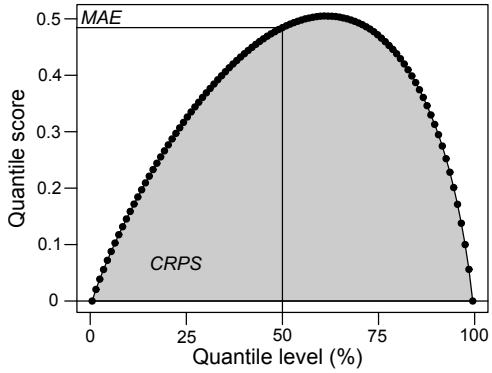
$$\text{CRPS} = \frac{1}{101} \sum_{\tau=0,0.01,\dots 1} \underbrace{\frac{1}{T} \sum_{t=1}^{T} 2\left(\mathbb{1}\left(y_t < \hat{y}_t^{\tau}\right) - \tau\right)\left(\hat{y}_t^{\tau} - y_t\right)}_{\text{QS}_\tau}. \tag{5.3}$$

PICP should be as close as possible to the theoretical coverage probability (i.e. 80%): when PICP is lower (resp. higher), predicted distribution is overdispersive (resp. underdispersive) indicating a wrong calibration (Chu & Coimbra, 2017). Interval 10—90% is selected since it is a standard interval for robust optimization (Correa-Florez et al., 2018). MAE and CRPS are positive and negatively-oriented, meaning that the closer to 0 are the scores, the better is the model. MAE is a deterministic score taken only the median forecast value into account, while CRPS is a probabilistic score evaluating the quality of the complete forecast distribution. CRPS is numerically computed by averaging the quantile scores (noted QS_τ) over the 101 uniformly distributed values of quantiles, i.e. $\tau = 0, 0.01, 0.02, \dots, 0.99, 1$ (see Appendix B). In order to assess overall performance on all of the household consumption, MAE and CRPS are normalized (called NMAE and NCRPS) by the average hourly demand of the household so as to obtain a dimensionless value (expressed in %).

Figure 5.1 represents two common graphs for forecasting performance. For a randomly selected household, Figure 5.1a represents the Probability Integral Transform (PIT) histogram, also called Talagrand histogram (Candille & Talagrand, 2005). Although a perfect PIT histogram should be flat, there exists inevitable statistical errors due to the limited sample. Dashed lines represent the confidence interval of 99% of the histogram (Pinson et al., 2010). Figure 5.1b represented the quantile scores on the y-axis against the quantile levels $\tau = 0, 1, \dots, 100\%$. The quantile score curve is bell shaped, with higher values on the middle part of the distribution than on the extremes. Scores are directly read from these graphics: PICP can be read by summing all bars of the histogram between 10% and 90%, MAE corresponds to the quantile score at level 50%, and CRPS is equal to the integral between 0% and 100% of the quantile score



(a) PIT histogram



(b) Quantile scores

Figure 5.1 – Day-ahead forecasting performance for one specific household. 5.1a shows the PIT histogram on 100 regular intervals. Confidence bars show the theoretical statistical sample error. 5.1b shows the quantile score for different quantile levels. MAE is read on quantile level of 50%, and CRPS is equal to the area under the curve.

curve.

Correct calibration of the forecasting models is assessed with the PIT histogram. In the perfect case, 99% of all bars should fall in the 99% confidence intervals. With our forecasting models, the frequency observed in the interval is 98.1%, suggesting that calibration is correct. Figure 5.2 shows performance in terms of NMAE and NCRPS

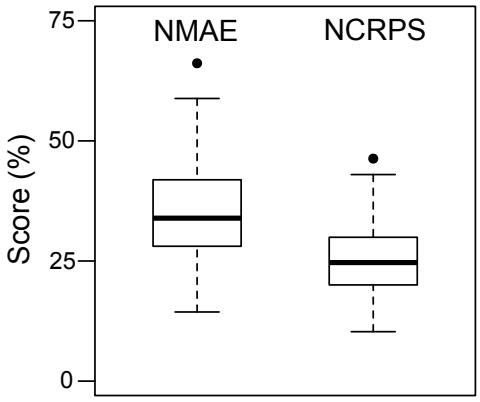


Figure 5.2 – Boxplot of the performance of day ahead forecasting model of each household hourly demand.

for each one of the 175 households. Average score is equal to 35% (resp. 25%) for the NMAE (resp. NCRPS), ranging from 14 to 66% (resp. 10 to 46%) depending

on the household considered. Such performance is comparable to other studies on the literature for household day ahead forecasting, such as study by Gajowniczek and Ząbkowski (Gajowniczek & Ząbkowski, 2017) and review of Yildiz et al. (Yildiz et al., 2017). In conclusion, our day-ahead forecasting models are probabilistically correct and show good performance. The quantile values produced are then used to generate scenarios.

5.1.2 Scenarios Generation

5.1.2.1 Introduction

The forecasting models previously introduced compute probabilistic distributions of future demand independently for successive hours. This independence is a simplifying assumption since, ideally, forecasting must be done for multiple horizons simultaneously to reflect the actual dependence between successive hourly demands. However, forecasting models at multiple horizons are highly complex: a very large dataset is required due to the high dimension of the problem, and computational cost is important. A demand scenario turns the hourly probabilistic forecast distributions into a set of deterministic values to describe a possible trajectory of electricity demand. We focus on daily trajectory or scenario, from 00:00 to 23:00 of one day. Therefore, instead of 24 distribution functions forecasting the hourly demand, one obtains a scenario, i.e. a 24-dimensional point, which is easier to handle in further applications. Multiple scenarios are used in practice to describe the probabilistic nature of the forecasting. All the 24 elements of a trajectory are drawn from the marginal distributions forecast. However, there are different ways to draw the set of 24 elements. Quality of the scenarios is analyzed by comparing their characteristics to the actual demand profiles.

We analyze the demand profiles by defining a residual time series (z_t) that indicates the part of the forecast distribution where the actual demand falls in. With the forecast cumulative distribution function (CDF) $\hat{F}_t$ — in our case retrieved from the set of forecast quantile values —, then

$$z_t = \hat{F}_t^{-1}(y_t) \in (0, 1). \tag{5.4}$$

When the forecasting model is calibrated, time series (z_t) is overall uniformly distributed in the interval $(0, 1)$. In parallel, time series (z'_t) is defined as the transforma-

tion of (z_t) with the inverse standard cumulative function Φ^{-1}, i.e.

$$z'_t = \Phi^{-1}(z_t) \sim \mathcal{N}(0,1). \tag{5.5}$$

The trajectories may be indexed by time index t, or by the day d and the hour of the day h, noted z_h^d or just z_h.

A scenario s is therefore made of a 24-dimensional point $(\hat{z}_0^{(s)}, \dots, \hat{z}_{23}^{(s)})$, and is good if it has characteristics similar to to an actual daily trajectory $(z_0, \dots, z_{23})$. Four generating methods are presented in the following. Their respective performance is evaluated by comparing the generated scenario to actual trajectories. This comparison is done in Section 5.1.4.

5.1.2.2 Benchmarking Methods

Two benchmarking methods are first proposed to generate scenarios. The *Connect-the-Quantiles* method supposes that successive values are completely correlated, and conversely the *Uniform Random Sampling* assumes that successive values are completely independent.

Connect-the-Quantiles In this method, one assumes that successive demand values are exactly on the same part of the forecast distributions, i.e. at a quantile level $\tau \in (0,1)$,

$$\hat{z}_0^{(s)} = \hat{z}_1^{(s)} = \dots = \hat{z}_{23}^{(s)} = \tau. \tag{5.6}$$

As seen in Figure 5.3 (orange lines), scenarios obtained with this method are very smooth. Moreover, such scenarios are completely ordered, meaning that demand for one scenario is higher than another scenario for all 24 hours of the day. It implies that some of these scenarios — e.g. the scenario for τ close to 0.5 — is more likely to occur. Indeed, it is very unlikely that the 24 hourly demands will actually fall on the extreme parts of the forecast distribution, since it would mean that extreme activities happen on all of the 24 hours of the day.

Uniform Random Sampling With the Uniform Random Sampling method, one assumes that successive demand values are completely independent between each others. The elements of the scenarios are therefore independently drawn for the uniform

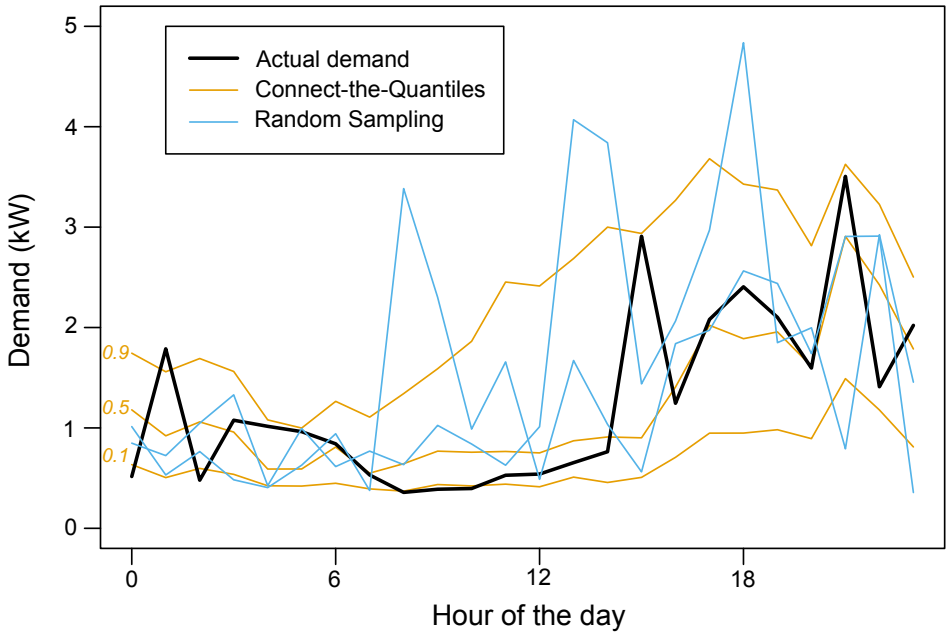


Figure 5.3 – Example of the scenarios obtained with the two benchmarking methods compared with the actual demand profile (black line) for a specific day and household. Three scenarios are depicted with the Connect-the-Quantiles method (orange lines) at level $\tau = 0.1, 0.5, 0.9$: such scenarios are smooth and ordered. Two scenarios are depicted with the Uniform Random Sampling method (blue lines): such scenarios have more variation than the actual demand profile.

distribution $\mathcal{U}(0,1)$, i.e.

$$\hat{z}_h^{(s)} \sim \mathcal{U}(0,1) \qquad \forall h \in \{0,\dots,23\}. \tag{5.7}$$

As seen in Figure 5.3 (blue lines), scenarios obtained with this method are erratic with large variation.

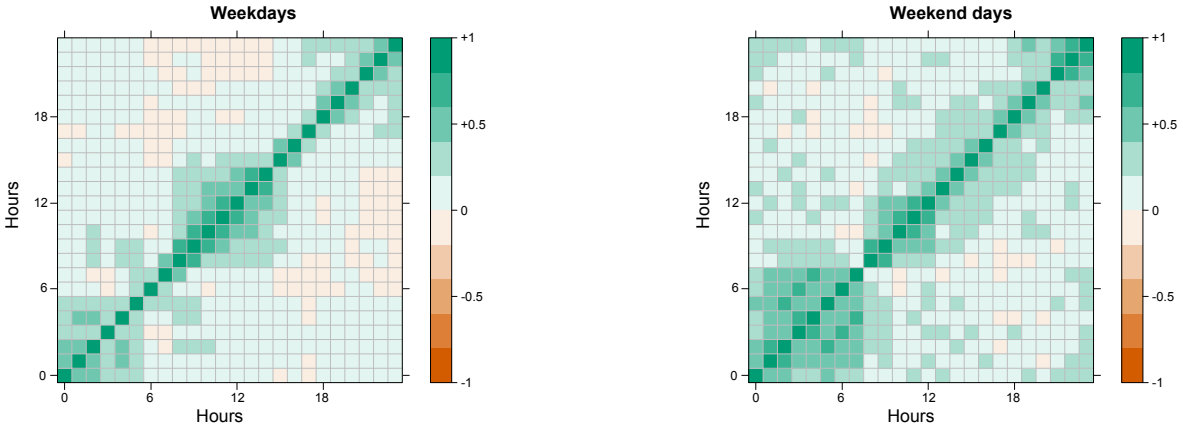
5.1.2.3 Covariance of Residual Time Series

Introduction The two benchmarking scenario generation methods describe the two extreme behaviors of the residual time series, i.e. complete dependence or complete independence between successive hours. In fact, the behavior of historical daily trajectory $(z_0,\dots,z_{23})$ is in between the two extreme methods. There is a strong correlation observed in values between successive hours, but this correlation decreases when the temporal difference increases. Figure 5.4 depicts observed residual correlation between

each of the 24 hours of the day for one household computed for all of weekdays and all of weekend days of the year. The greener is an area, the stronger is the correlation. Observed correlation fades away with temporal difference and is negligible when this difference is over 5 hours. An efficient way to generate scenarios is to use this covariance matrix, or equivalently correlation. Note that this is equivalent to use a Gaussian copula.

The residual correlation matrix reflects daily activities of the household. Correlation clusters are visible during certain periods when residual values are strongly correlated between each others, but not with other periods, e.g. between 08:00 and 15:00 on weekdays. Household activity during such periods is quite regular, but the exact level of electricity demand is not predictable on the day before. This is caused for instance by the exact temperature of the day which is slightly different than the one forecast. This slight difference impacts electricity demand via heating appliances. Such correlation clusters usually relate to the household's occupancy: when people are not at home — or asleep — electricity demand is regular. Conversely, when successive values are weakly correlated, e.g. between 16:00 and 20:00 on weekdays, it means that people are at home, and have various activities (e.g. cooking, using dryer, etc.) that require rather large amount of electricity during short periods of time. Consequently residual time series (z_t) considerably varies, hence the weak correlation observed.

Since household activities depend on the day of the week, correlation depends on the daily trajectory subset used. In Figure 5.4, correlation matrices are computed on the one hand for weekdays only, and on the other hand for weekend days only. The obtained matrices are visually different. In fact, 2-dimensional Bartlett tests show that the weekdays/weekend difference is significant for most households (D'Agostino SR & Russell, 2005). However, since datasets are limited — there are only 356 daily profiles in one year —, using exclusively weekend days produces noisy matrix. It is visible in Figure 5.4b where correlation is wrongly found large between 12:00 and 23:00. Consequently, computing the correlation matrix is not straightforward and impacts quality of scenarios. In particular, a burn-in period of 50 days is necessary to have a representative set of trajectories for a meaningful correlation matrix. We present two methods called *Basic Covariance* and *Refined Covariance*.



(a) Weekdays (b) Weekend days

Figure 5.4 – Correlation matrices between the 24 hourly values of a day of the residual time series (z'_t) for a specific household. In Figure 5.4a, correlation is computed using trajectories observed on weekdays; in Figure 5.4b, correlation is computed using trajectories observed on weekend days.

Basic Covariance The Basic Covariance method is a straightforward way to compute covariance matrix. Each day, covariance matrix is computed using trajectories observed for all of the previous days of the year. Covariance between hour h_1 and h_2 of the day is then computed using all of the days up to the current day $D \geq 50$

$$\Sigma^D_{h_1 h_2} = \frac{1}{D-1} \sum_{d=1}^{D-1} z'^d_{h_1} z'^d_{h_2}. \tag{5.8}$$

Refined Covariance Covariance can be refined in two aspects. Firstly, habits evolve with time. Therefore, an exponentially decreasing weighting is added to favor more recent points in the covariance calculation. Secondly, as observed in Figure 5.4, covariance matrix is visually and statistically different according to the subset used. Therefore, second weights favoring observations on the same day of the week (i.e. $7, 14, \ldots$ days ago) is added. These weights depend on the day of the week since this weekly correlation is usually stronger for non-working days: e.g. Sunday is relatively more similar to previous Sundays than Tuesday is to previous Tuesdays. The refined covariance value between two hours computed on day $D \geq 50$ therefore writes

$$\Sigma'^D_{h_1 h_2} = \frac{1}{(D-1)W_D} \sum_{d=1}^{D-1} w^1_d w^2_d z'^d_{h_1} z'^d_{h_2}, \tag{5.9}$$

where

$$w_d^1 = \exp\left(-\frac{D-d}{\lambda}\right),$$

$$w_d^2 = \begin{cases} \alpha_{\mathrm{wd}(D)} & \text{if } (D-d) = 0 \bmod 7, \\ 1 & \text{otherwise,} \end{cases}$$

$$W_D = \sum_{d=1}^{D-1} w_d^1 w_d^2,$$

with $\mathrm{wd}(D)$ indicating the day of the week (between 0 and 6) of day D. The 8 unknown parameters $\lambda, \alpha_0, \ldots, \alpha_6$ are estimated by maximum likelihood estimation on historic data of each specific household, and are updated once per month. Optimal values of λ fluctuate between 300 and 1000 days depending on the household. Such high values mean that habits slowly evolve through time: trajectory observed one day ago matters only twice as trajectory one year before. Second weights specific to the day of the week are found approximately equal to 1.7 for Saturday and Sunday and to 1.2 on other days, corroborating the statement that Sunday is more similar to previous Sundays than Tuesday is of previous Tuesdays.

Scenarios from Covariance of Residuals Once the covariance matrix Σ, or similarly Σ', is computed, one draws a trajectory of residuals, i.e. a scenario s, according to a multivariate Gaussian distribution

$$\left(\hat{z'}_0^{(s)}, \ldots, \hat{z'}_{23}^{(s)}\right) \sim \mathcal{N}(0, \Sigma). \tag{5.10}$$

Each element is then transformed with the standard CDF Φ and forecast marginal distribution $\hat{F}_h$

$$\hat{z}_h^{(s)} = \Phi\left(\hat{z'}_h^{(s)}\right) \tag{5.11}$$

$$\hat{y}_h^{(s)} = \hat{F}_h\left(\hat{z}_h^{(s)}\right) \tag{5.12}$$

for $h = 0, \ldots, 23$. The obtained scenario $\left(\hat{y}_0^{(s)}, \ldots, \hat{y}_{23}^{(s)}\right)$ follows the 24 marginal distributions, and thus each hourly demand is correctly forecast.

5.1.2.4 Number of Scenarios

With the methods presented, generating a lot of daily scenarios is computationally cheap. Therefore, we wish to generate a sufficient number of scenarios in order not

to degrade performance compared to the original forecast distributions. In our case, this original forecast distribution of hourly demand is approximated by 99 quantiles regularly spaced. However, when we draw from such distribution to obtain scenarios, the resulting sample is not regularly spaced and degrade probabilistic quality. When analyzing solely quality of hourly demand forecasts, performance of scenarios is necessarily poorer than the performance of the original forecasting model: scenarios degrade independent forecasting performance since they ensure the multi-temporal, e.g. daily, consistency of forecasts. When the number of scenarios increases, the performance degradation reduces to 0. The objective is to find a value for this number to ensure a limited degradation.

To estimate this number, we compare the quantile scores of hourly demand at 3 levels, for $\tau = 0.01, 0.10, 0.50$, according to Equation (5.3). On the one hand, we have the scores obtained with the original and regularly spaced forecast distribution. On the other hand, we have the quantile scores at the same levels for various numbers of scenarios. The ratio of the two scores indicates the degradation due to scenario sampling. For instance, a ratio of 1.1 means that degradation is of 0.1 (or 10%). We want to limit the degradation to 0.5% (i.e. ratio error below 0.005). Figure 5.5 depicts this ratio error, averaged over all 175 households, against the number of scenarios for the three quantile levels. The scenarios are here generated with the Uniform Random Sampling method; results are similar with the covariance methods. The errors logically decrease with the number of scenarios generated and cross the 0.5% threshold for around 300 scenarios. Furthermore, the performance degradation is larger for quantile level of 1% than 10%, and than 50%. This is due to the fact that the extreme parts of distribution are harder to approximate than the middle part. In fact, the natural quantile estimator at level τ converges to a Gaussian distribution with a variance depending on the density f, and CDF F, of the phenomenon distribution, precisely

$$\mathcal{N}\left(F^{-1}(\tau), \frac{\tau(1-\tau)}{n \cdot f^2(F^{-1}(\tau))}\right) \tag{5.13}$$

with n the sample size (see Appendix A and (Xu & Miao, 2011)). This limiting distribution exhibits that the convergence rate, i.e. variance of limiting distribution, depends on the quantile level τ. In the Gaussian case, at equal sample size, the standard deviation at level $\tau = 0.5$ is one third of the standard deviation at level $\tau = 0.01$. This fact explains that the quantile score at level $\tau = 0.01$ is roughly three times higher

than the score at level $\tau = 0.5$. Performance for the higher part of the distribution is symmetrically similar, i.e. performance for 99% is worse than 90%, and worse than 50%. However, although the two levels are as extreme, performance degradation is less important for level 1% than 99%. This is expected since electricity demand distribution is usually right skewed, and so the upper tail is longer than the lower tail.

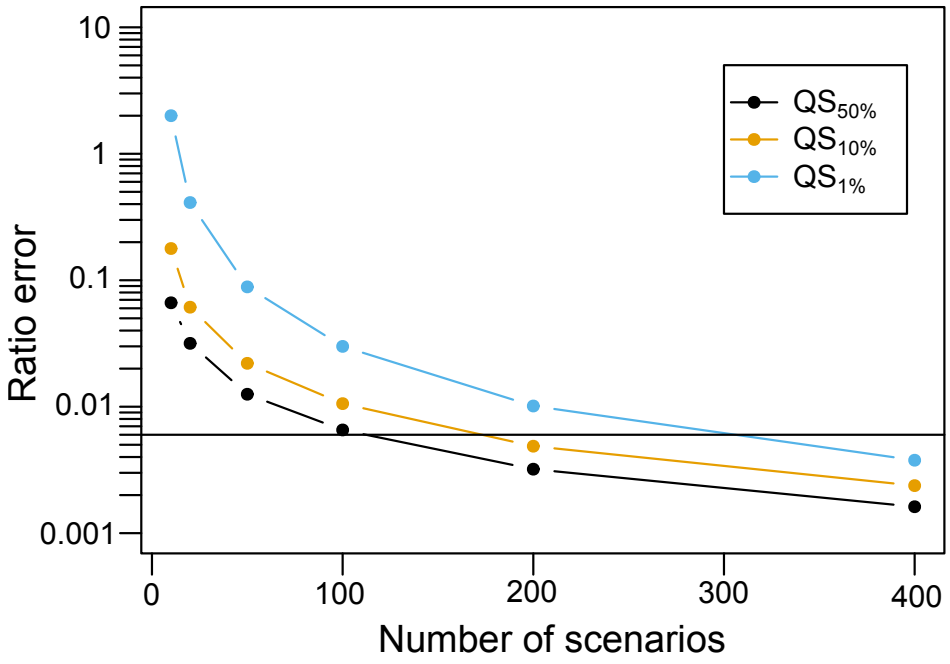


Figure 5.5 – Average ratio errors of the quantile scores of hourly demand forecasting between the original forecast distributions and the scenarios. A ratio error close to 0 means that the hourly forecast distributions are well approximated by the scenarios. The horizontal line indicates the chosen threshold at 0.5%.

5.1.3 Scenario Reduction

5.1.3.1 Framework

Section 5.1.2 shows how to generate daily scenarios of household demand from probabilistic forecasts. Methods presented are fast and a lot of scenarios can be generated for a low computational cost. Each scenario is then used by an application leading to an optimal decision, denoted $\hat{o}^{(s)}$ depending on the scenario s, such as "should the battery be charged right now?". With a sufficient number of scenarios, e.g. $S = 400$,

the different decisions $\hat{o}^{(1)}, \ldots, \hat{o}^{(S)}$ provide a probabilistic way to make a decision.

However, the computation of quantity $\hat{o}^{(s)}$ for one scenario may be long and expensive to do, so using all the generated scenarios may be unpractical. One therefore wants to find a smaller set of representative scenarios. This is usually done with a scenario reduction process, i.e. reducing the number of scenarios from S to $K \ll S$. A widespread method of scenario reduction is to cluster scenarios together when they are close, and to only consider one representative scenario per cluster. This clustering process implies that a proximity metric is to be defined. Since proximity depends on the application, there is no universal metric. One person, who is studying profitability of battery with cycling due to charge and discharge, is interested in the possible ramps in the profiles, whereas a trader, who is optimizing her purchases on the market, is more interested in the hourly demands when electricity is expensive on the market. The reduced set of scenarios should dwell on the driving features of the later application and cluster scenarios accordingly.

While daily profiles generated in Section 5.1.2 are all equally probable, this is usually not the case with a reduced set of scenarios. Each representative scenario $\hat{y}^k$, for $k = 1, \ldots, K$, is associated to a probability π_k. This probability is set equal to the ratio between the number of scenarios in cluster k and the total number of scenarios S.

Let us note that, since the reduced set of scenarios is only an approximation of the complete set of scenarios, quality of the reduced set is inevitably worse than quality of the complete set. In the following, general framework of scenario reduction is introduced. A proximity metric between scenarios is to be chosen. Three metrics are presented in the following: a point-wise distance, i.e. a metric measuring distance between forecast and observation at the same instants; a characteristics distance, based on main characteristics of household daily profile; a price-weighted household demand distance, specifically crafted for household demand scenarios and taking the electricity market price into account.

5.1.3.2 Methodology

In order to select representative scenarios for the reduced set of scenarios, a proximity metric should be defined to assess if two scenarios are close with each other. Indeed, when two scenarios are close, only one should be included in the reduced set, the second

one being represented by the first one. To describe the variety of possible trajectories, the reduced set should be comprised of scenarios that are fairly far from each other. However, scenarios, i.e. multidimensional points, are not fully ordered and several metrics exist.

Point-Wise Distance The point-wise distance measures the proximity between two demand trajectories; i.e. between $\hat{y}^{(s_1)} = (\hat{y}_0^{(s_1)}, \dots, \hat{y}_{23}^{(s_1)})$ and $\hat{y}^{(s_2)}$ by point-wise comparison, i.e. the distance between two values at the same hour. Different underlying distances may be used, such as the absolute difference, so the distance between two scenarios is

$$d_0(s_1, s_2) = \sum_{h=0}^{23} |\hat{y}_h^{(s_1)} - \hat{y}_h^{(s_2)}|. \tag{5.14}$$

This metric is a straightforward way to compute distance between two time trajectories, i.e. multidimensional point. However, such metric is known to be ill-suited for irregular time series such as household electricity demand (Keil & Craig, 2009) and is expected to provide a reduced set that poorly reflects demand dynamics.

Profile Characteristics Distance For each demand scenario $\hat{y}^{(s)}$, we identify four key parameters describing profile characteristics:

- Total daily demand $\chi_1^{(s)} = \sum_h \hat{y}_h^{(s)}$,
- Peak demand $\chi_2^{(s)} = \max_h \hat{y}_h^{(s)}$,
- Maximal ramp between successive hours $\chi_3^{(s)} = \max_h |\hat{y}_{h+1}^{(s)} - \hat{y}_h^{(s)}|$,
- Peak demand hour $\chi_4^{(s)} = \operatorname{argmax}_h \hat{y}_h^{(s)}$.

Total daily demand is often a prominent parameter since it is the total energy quantity to produce for the particular profile. Coupled with the peak demand, it defines the load factor, which has long been identified as a key parameter for the profitability of an electricity supplier (Insull, 1914). A high load factor indicates that demand is stable throughout the day hence infrastructure (e.g. distribution lines) are fully used at all times. Visually, a demand trajectory with high load factor is smooth with low "peakiness" (Barker et al., 2012). Additionally, maximal ramp is relevant to describe variations occurring throughout the day. Large ramps are more stringent for

the infrastructure. In the case of household energy management, large ramps impact the depth of discharge of batteries, deteriorating their performance (Correa-Florez et al., 2018).

Due to the different natures and dimensions of the characteristics, each one is centered and rescaled according to observed mean and deviation on all the scenarios. Therefore, all 4 characteristics matter equivalently for future clustering applications and the distance between $\hat{y}^{(s_1)}$ and $\hat{y}^{(s_2)}$ is

$$d_1(s_1, s_2) = \sum_{i=1}^{4} (\chi_i^{(s_1)} - \chi_i^{(s_2)})^2. \tag{5.15}$$

Price-Weighted Household Demand Distance Haben et al. craft a metric specially designed for household electrical energy demand (Haben et al., 2014). It is presented as a solution to the double penalty effect often observed when forecasting household electricity demand. When measuring error between forecast and actual demand, the double penalty effect penalizes twice a demand peak correctly forecast in amplitude but not in time. For instance, a peak forecast at 07:00 but actually occurring at 08:00 is wrong at 07:00 and at 08:00. Therefore, forecasting models prefer to produce flat forecasts with no peak at all, so as to be penalized only once at 08:00. However, flat forecasts are usually less informative than peaky forecasts for further applications (Molderink et al., 2010). Let $\mathfrak{S}_{24}^*$ denotes the group of permutations of size 24 excluding permutations between elements that are more than w hours apart. The household demand distance between two daily scenarios s_1 and s_2 is

$$\min_{\sigma \in \mathfrak{S}_{24}^*} \sum_{h=0}^{23} \left(|\hat{y}_{\sigma(h)}^{(s_1)} - \hat{y}_h^{(s_2)}|^p \right)^{1/p}. \tag{5.16}$$

This distance is symmetric but usually does not obey the triangle inequality, hence is a semi-metric rather than a metric. We fix meta-parameters of the distance $p = 1$ for the distance to be robust to outliers and $w = 2$ hours to have a reasonable amount of time for intraday adjustment. As suggested by Haben et al., minimization problem of Equation (5.16) is solved in polynomial time with the Hungarian method (Papadimitriou & Steiglitz, 1998).

A natural refinement of this household demand distance is to take into account the price of electricity. The idea is that a demand difference is more crucial when electricity

price is high, so difference should be weighted by the electricity price on the market. When working at the regional scale, this weighting is less necessary because electricity prices are mainly driven by regional demand, and thus, the weighting is implicitly described by the level of demand. However, household demand does not follow the regional demand trends, and this price weighting is important to reflect trajectory differences.

ERCOT provides observed hourly electricity prices (in \$/MWh) on the market in the year 2017 (Electric Reliability Council of Texas (ERCOT), 2018b). These prices greatly vary from day to day, due to the high fluctuations on the market. Average price is around 25 \$/MWh, ranging from less than 2 \$/MWh in the night of the 11th January to 250 \$/MWh on 28th July afternoon. In Figure 5.6, median price profiles are represented for weekdays (black) and weekend days (orange) of the year. According to Anderson-Darling non parametric tests on price distribution on different days of the week (Scholz & Stephens, 1987), the distinction between weekdays and weekend days is statistically significant (p-value < 0.01)[1]. In particular, morning prices are higher on weekdays than on weekend days. These price profiles follow the general trends of the regional demand of Texas. Electricity prices are higher in the afternoon (16:00 and 17:00) due to high industrial and residential demand, and low production generated by photovoltaics. These median profiles provide prices $\hat{p}_h$ used to weight hourly demand difference. Additionally, in the problem of Equation (5.16), hourly demands are exchangeable without penalty. When taking electricity prices into account, a penalty is required when permuting demand during expensive and cheap hours.

In conclusion, the *price-weighted household demand distance* follows the 3 principles:

1. distance is function of sum of absolute difference of hourly demands weighted by the price on the electricity market,

2. permutation between successive hours is possible only if the time gap is less than or equal to 2 hours,

3. when a permutation between hour h and $\sigma(h)$ is done, distance increases with the price absolute difference $|\hat{p}_{\sigma(h)} - \hat{p}_h|$.

[1] Wednesday price profiles fall in between weekdays and weekend days: we opt to group it with the weekdays profiles.

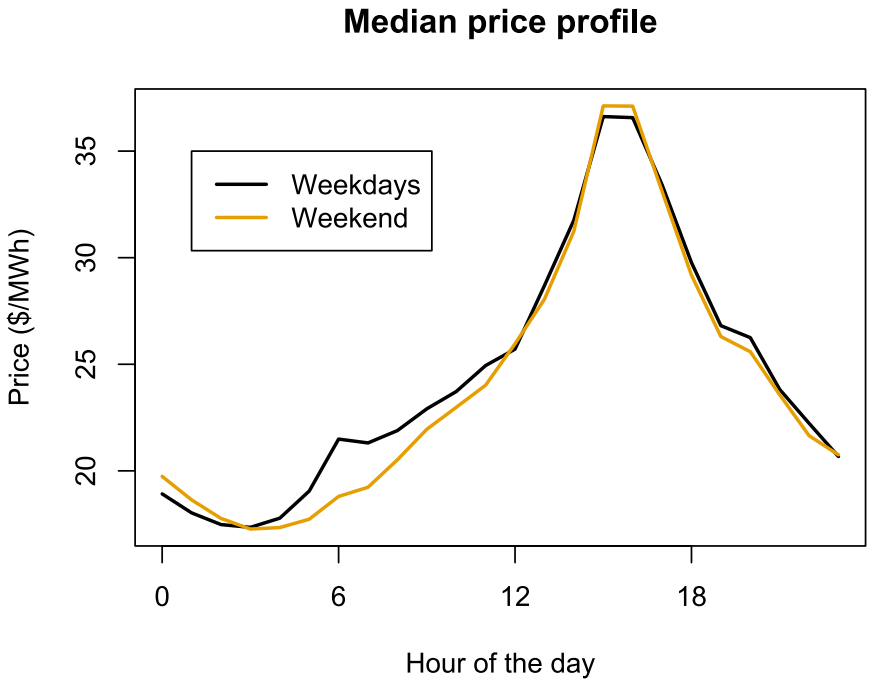


Figure 5.6 – Median electricity price profile for weekdays (black line) and weekend days (orange line) observed in Texas in 2017.

The metric proposed between scenarios s_1 and s_2 then writes

$$d_2(s_1, s_2) = \min_{\sigma \in \mathfrak{S}^*_{24}} \sum_{h=1}^{24} \frac{\hat{p}_{\sigma(h)} + \hat{p}_h}{2} |\hat{y}^{(s_1)}_{\sigma(h)} - \hat{y}^{(s_2)}_h| + \frac{\hat{y}^{(s_1)}_{\sigma(h)} + \hat{y}^{(s_2)}_h}{2} |\hat{p}_{\sigma(h)} - \hat{p}_h|. \tag{5.17}$$

Distances between every pair of scenarios is computed. This is a computationally intensive part of the reduction, taking around 20 seconds to compute all of the distances in a set 400 daily scenarios, with a CPU of 3 GHz. When all the distances are computed, the selection of the representative scenarios can be made with various methods, such as the fast forward scenario reduction (Dupačová et al., 2003).

Illustration In order to identify the advantages of each distance introduced, we illustrate how they discriminate scenarios.

Figure 5.7 depicts two schematic scenarios in orange and blue. They have a peak demand of 5 kW at 12:00 or 13:00 and are flat at different demand levels during the rest of the day. Consequently, the two scenarios demand the exact same daily energy,

have the same peak demand at almost the same hour, and the maximal ramp is equal in both cases (an increase of 4 kW between at 11:00 or 12:00). Profile Characteristics distance d_1 between the two scenarios is therefore very small. This is not the case for the Point-Wise distance d_0, which emphasizes the different demand levels during flat periods, nor for the Price-Weighted Household distance d_2, which emphasizes the demand differences during peak price hours (around 16:00). The black dotted line represents a scenario that is equidistant to the other two scenarios according to d_0. This equidistant scenario is of low interest since it does not anticipate the peak at 5 kW.

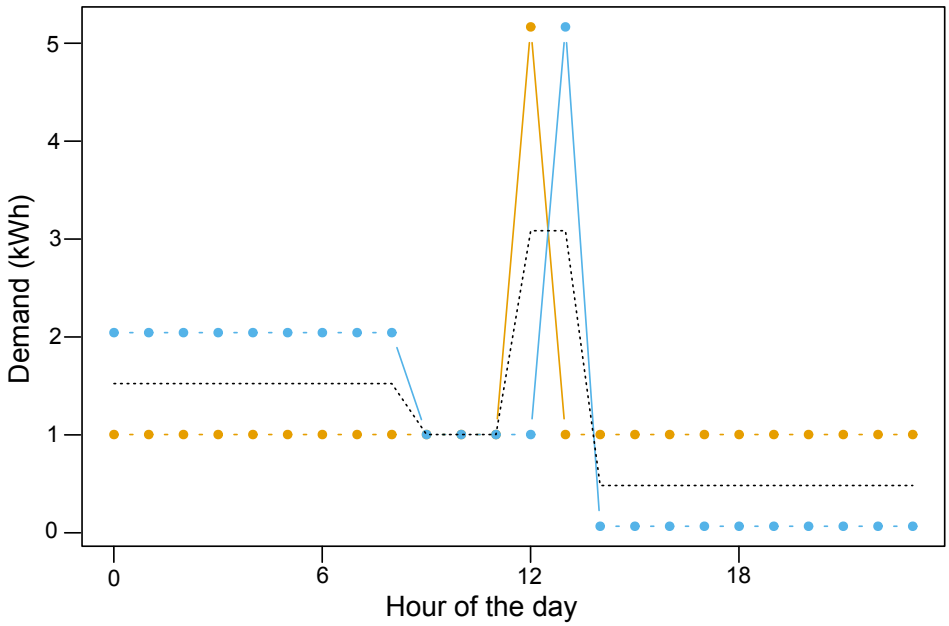


Figure 5.7 – The Profile Characteristics distance between the scenarios in blue and orange is small. Black dotted line represents a scenario that is equidistant to the two other colored scenarios according to Point-Wise distance.

Figure 5.8 also depicts two schematic scenarios in orange and blue. They are flat during most periods of the day, except during the night when fluctuations occur. Since, these fluctuations happen when prices are low and in a short period of time (less than 3 hours), the Price-Weighted Household distance d_2 between the two scenarios is very low. This is not the case for the Point-Wise distance d_0, nor for the Profile Characteristics distance d_1, which penalizes the different peak demand hours. The black dotted line represents a scenario that is equidistant to the other two scenarios according to d_0. As

before, this scenario is of low interest because it flattens out the fluctuations.

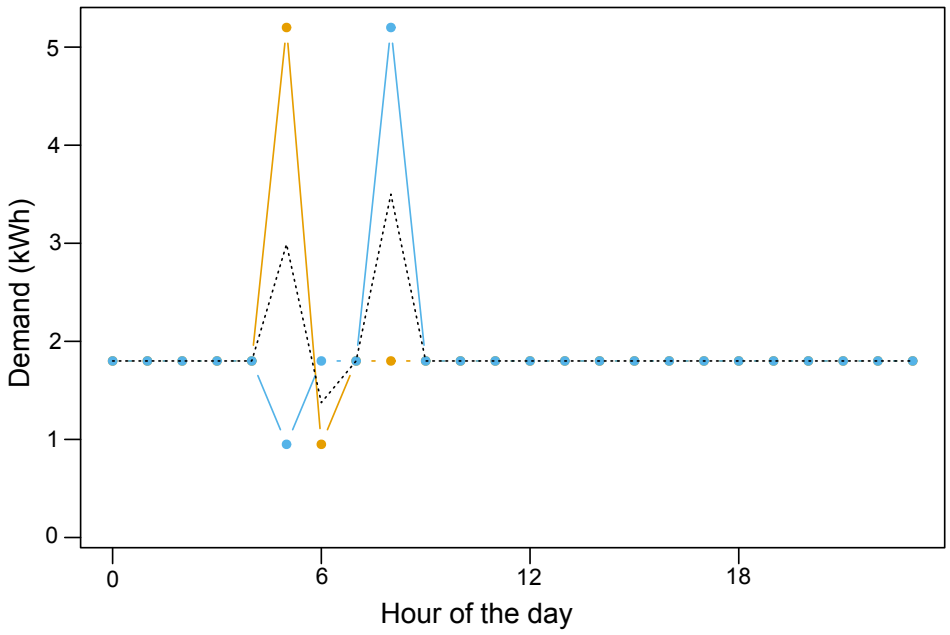


Figure 5.8 – The Price-Weighted Household Demand distance between the scenarios in blue and orange is small. Black dotted line represents a scenario that is equidistant to the two other colored scenarios according to Point-Wise distance.

Fast Forward Scenario Reduction Bruninx and Delarue (Bruninx & Delarue, 2016) use a fast forward scenario reduction algorithm to reduce the number of wind power scenarios for a unit commitment problem. The algorithm is based on the Monge-Kantorovich mass transport problem. In our case, distances d_0, d_1 or d_2 are used as cost function of the mass transport problem.

The algorithm selects representative scenario among all the scenarios available in an iterative way:

- The first representative is the scenario that is the most equidistant from all other scenarios;
- representative k is the scenario that, if taken as a representative, minimizes total distance between all of the scenarios and their closest representatives;
- the probability assigned to each representative, π_k, is then taken equal to the

ratio of scenarios that are closest to representative k and the total number of scenarios.

The number of representatives K is defined a priori and depends on the required accuracy for later applications.

5.1.4 Quality of Scenarios

5.1.4.1 Criteria

The quality of scenarios refers to the ability of the scenario to correctly forecast the future unknown load. Different methods to generate and reduce scenarios induce different quality. To correctly assess the quality, the evaluation should be made over the complete day, and not independently for each hour[2]. We first look at the characteristics of the daily profile and, secondly, at the daily cost due to the demand profile.

Profile Characteristics The 4 characteristics described in Section 5.1.3.2 — total daily demand, peak demand, maximal ramp between successive hours, and peak demand hour —, are used to assess the quality of the scenarios generated and the reduced set obtained. We evaluate the 4 characteristics independently. However, since characteristics are often correlated between each other, e.g. maximal ramp and peak demand, scenario sets efficient regarding one characteristic are also efficient regarding the others.

Daily Cost The daily cost is equal to the sum of the 24 hourly demands of the day multiplied by the corresponding hourly electricity market price. For a specific day and scenario $s = 1, \ldots, S$

$$\hat{o}^{(s)} = \sum_{h=0}^{23} p_h \hat{y}_h^{(s)}. \tag{5.18}$$

These daily costs $\hat{o}^{(s)}$ are not forecast since exact market prices $p_0, \ldots, p_{23}$ are unknown in advance. However, such quantities provide a post hoc forecast of the next day cost without taking into account the price forecasting errors, and thus conveniently focus on errors caused by inaccuracy of demand scenarios.

[2]The marginal distribution forecast for each hour are only due to the quality of the forecasting methods and is not impacted by the scenarios method.

Evening Demand The evening demand is defined as the sum of the hourly demand made between 18:00 and 23:00, i.e.

$$\hat{\nu}^{(s)} = \sum_{h=18}^{23} \hat{y}_h^{(s)}. \tag{5.19}$$

In most cases, the inhabitants are home during this period, and therefore the evening demand is an important fraction of the daily total demand. This period exhibits strong variety depending on the behavior of the inhabitants. Consequently, the period offers important flexibility opportunities, e.g. by shifting water heater, or by injecting remaining energy in electric vehicle's battery.

5.1.4.2 Performance Scores

The criteria presented in Section 5.1.4.1 are evaluated with performance scores. Except for the peak demand hour (χ_4), the same three scores used in Section 5.1.1.3, assess the quality of the scenarios: NMAE, PICP and NCRPS. NMAE and and NCRPS are normalized so the household scores can be compared between each other: daily total demand χ_1 is normalized by the average daily total demand of the household, peak demand χ_2 is normalized by the average hourly demand of the household, maximal ramp χ_3 is normalized by the average hourly demand of the household, daily cost o is normalized by the average daily cost of the household, and the evening demand ν is normalized by the average evening demand of the household.

The NMAE score evaluates the deterministic aspect of the scenarios, while PICP and NCRPS analyze their probabilistic aspect. The score definitions are adapted for scenarios characteristics and daily cost: the characteristics and costs obtained for each scenario are ordered, then the quantile values of the quantities, for quantile levels $\tau = 0.01, \ldots, 0.99$, are computed and used in the scores definitions.

Any method for computing the value at all quantile levels is accurate enough for a large number of scenarios. However, an issue arises when the number of scenarios is small, which is the case after a scenario reduction. Figure 5.9 shows, for a random household and a random day, the cumulative distribution functions of daily cost obtained with the $S = 400$ (blue) scenarios generated with the refined covariance methods, and with $K = 5$ representatives (oranges) obtained with the d_2 distance. The actual daily cost of the day is depicted by the vertical black line. The empirical cumulative

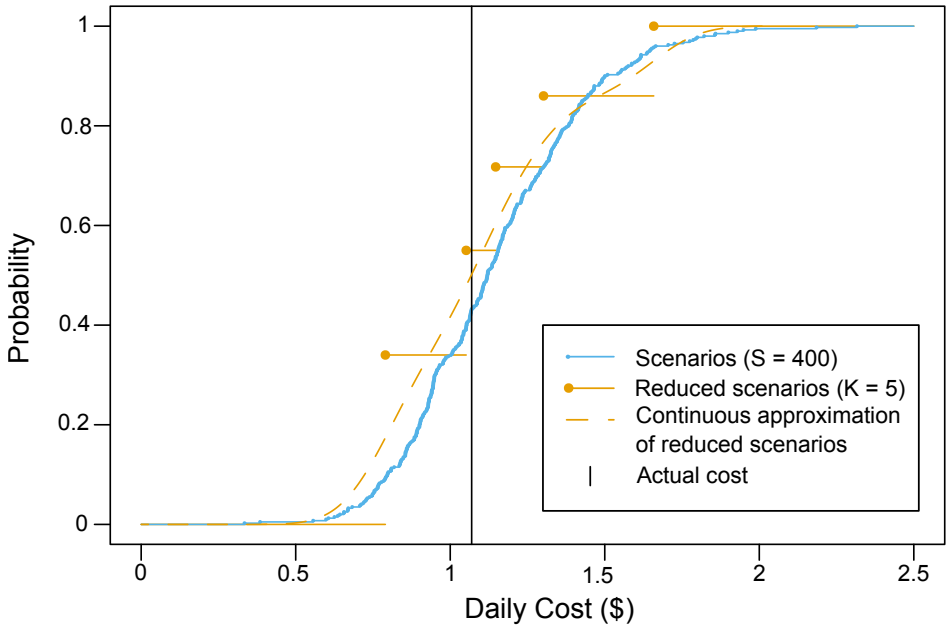


Figure 5.9 – Example of the forecast CDFs obtained with scenarios for daily cost of next day of one specific household. The empirical CDF of the complete set of 400 scenarios (blue) provide an almost continuous function, while the empirical CDF of the reduced set of 5 representative scenarios (orange) is notably discontinuous. A continuous approximation of this latter CDF is depicted in orange dash lines. The actual cost is represented by the vertical black line.

distribution function is an increasing simple function with gap at every realization. When the number of scenarios is large, simple functions are very close to continuous functions and can numerically be used as such. However, when the number of representatives is small, e.g. $K = 5$, continuous CDF should be estimated with caution. Therefore the density function is estimated with a uni-dimensional kernel density estimator lower bounded by 0 with bandwidth selected with the Silverman rule-of-thumb rule (Silverman, 1986), and corresponding probability assigned to each representative. This result function is shown in orange dash line in Figure 5.9.

To evaluate the quality of scenario regarding the peak demand hour χ_4, the previous scores are ill-suited. For some households, the peak demand hours is either in the morning or during the evening, but never in the afternoon. Consequently, half of the scenarios have a peak in the morning, and half have a peak in the evening. Therefore, computing the average to compute the NMAE, leading to an average demand hour in

the afternoon, is nonsensical. We choose to report solely the frequency of days when the peak hour is in the night and morning (0:00 to 8:00), in the middle of the day (9:00 to 15:00), during the peak price hours (16:00 and 17:00), or in the evening (18:00 to 23:00).

5.1.4.3 Scenario generation

A total of $S = 400$ are generated for each day of the year (excluding the first 50 days to compute covariance matrices) from the different scenario generation methods introduced: connect-the-quantiles, uniform random sampling, basic covariance of residuals, refined covariance of residuals. The performance of each method is evaluated by computing the scores seperately for each household. Table 5.1 reports the scores averaged over the 175 households along with their standard deviation between parentheses. Table 5.2 details the quality of the peak demand hour of the scenarios. Best performance are in bold.

Key information can be drawn from these report tables:

- The Basic Covariance and Refined Covariance methods are the most efficient methods: NMAE and NCRPS values are the lowest, and the PICP values are closest to theoretical 80%. The peak demand hours generated are a bit off the measured peak demand hours, with an under-representation of profiles with peaks during the night and at midday. There is a minor advantage for the Refined Covariance method but both covariance methods clearly outperform the two benchmarking methods.

- The probabilistic aspect of the scenarios generated by the Connect-the-Quantiles method is extremely poor, which is especially visible on the PICP score. This come from the unrealistically smooth scenarios generated. The scenarios are overdispersive for daily cost o, daily total demand χ_1, and evening demand ν since the extreme scenarios simulate extreme event for all the 24 hours. Conversely, the scenarios are underdispersive for maximal ramp χ_3 because of the smoothness of the scenarios. Peak demand occur too frequently during the evening and peak price hours because of the very similar shapes of scenarios.

- The second benchmark, with the Uniform Random Sampling method, has good deterministic performance, as seen on the NMAE, but not in probability. The

Table 5.1 – Scores obtained with the 4 scenario generation methods

$S = 400$ scenarios	NMAE	PICP (10—90%)	NCRPS
Daily total demand (χ_1)	%	%	%
Connect-the-Quantiles	19.9 (5.3)	97.8 (2.0)	15.4 (5.3)
Uniform Random Sampling	18.1 (5.0)	48.8 (7.4)	13.8 (5.0)
Basic Covariance	**18.0 (4.5)**	77.8 (3.7)	**12.8 (4.5)**
Refined Covariance	**18.0 (4.5)**	**77.9 (3.7)**	**12.8 (4.5)**
Peak demand (χ_2)	%	%	%
Connect-the-Quantiles	100.1 (56.8)	66.5 (8.2)	71.3 (38.9)
Uniform Random Sampling	73.6 (44.2)	62.1 (7.4)	52.1 (29.2)
Basic Covariance	70.6 (40.9)	74.9 (4.6)	49.3 (27.1)
Refined Covariance	**70.6 (40.8)**	**74.9 (4.4)**	**49.2 (27.1)**
Maximal ramp (χ_3)	%	%	%
Connect-the-Quantiles	92.2 (47.3)	19.3 (12.9)	81.8 (40.2)
Uniform Random Sampling	71.5 (39.2)	56.8 (12.7)	50.7 (26.3)
Basic Covariance	57.9 (33.1)	67.9 (8.1)	**40.7 (22.2)**
Refined Covariance	**57.9 (33.0)**	**67.9 (8.0)**	**40.7 (22.2)**
Daily cost (o)	%	%	%
Connect-the-Quantiles	23.1 (8.4)	97.4 (2.1)	17.7 (6.0)
Uniform Random Sampling	**21.4 (7.5)**	49.5 (7.2)	16.2 (5.8)
Basic Covariance	**21.4 (7.5)**	**78.1 (3.7)**	**15.1 (5.2)**
Refined Covariance	**21.4 (7.5)**	**78.1 (3.7)**	**15.1 (5.2)**
Evening demand (ν)	%	%	%
Connect-the-Quantiles	24.1 (8.4)	91.7 (3.5)	17.6 (6.0)
Uniform Random Sampling	23.7 (8.4)	59.2 (6.4)	17.5 (6.2)
Basic Covariance	**23.6 (8.1)**	**78.7 (3.1)**	16.9 (5.8)
Refined Covariance	**23.6 (8.1)**	**78.7 (3.1)**	**16.8 (5.8)**

Table 5.2 – Peak demand hour frequency (in %) observed in actual measurements and scenarios

Period	0 to 8	9 to 15	16 to 17	18 to 23
Actual Measurements	13.5	22.0	15.6	48.9
Connect-the-Quantiles	6.3	15.1	27.4	**51.1**
Uniform Random Sampling	9.2	27.9	**20.3**	42.5
Basic Covariance	**10.9**	**26.4**	20.6	42.1
Refined Covariance	**10.9**	26.5	20.6	42.1

scenarios are generally overdispersive (PICP greatly below ideal 80%) since scenarios fluctuate too much over and under the real demand curve. The peak demand hours are too spread across the day, with too many peaks occurring in the midday compared to actual measurements.

- For the two covariance methods, scenarios are slightly overdispersive (PICP always slightly below ideal 80%), indicating that extreme events are difficult to anticipate, especially the maximal ramp χ_3 with a PICP of 67.9% with the Refined Covariance method. This is due to the non-perfect hourly forecasts, and their small hourly biases that add up throughout the day.

- Important standard deviations are visible on NMAE and NCRPS among the households for peak demand χ_2 and maximal ramp χ_3, with a coefficient of variation going up to 58%. This high deviation is mainly due to the variety of the daily household load factors: some households triple their mean demand on peak hour, i.e. load factor of 33%, when others have load factor up to 70%. This load factor strongly impacts the evaluation of χ_2 and χ_3.

From this analysis, we later opt for the 400 scenarios generated with the Refined Covariance method and proceed to the scenario reduction process.

5.1.4.4 Scenario Reduction

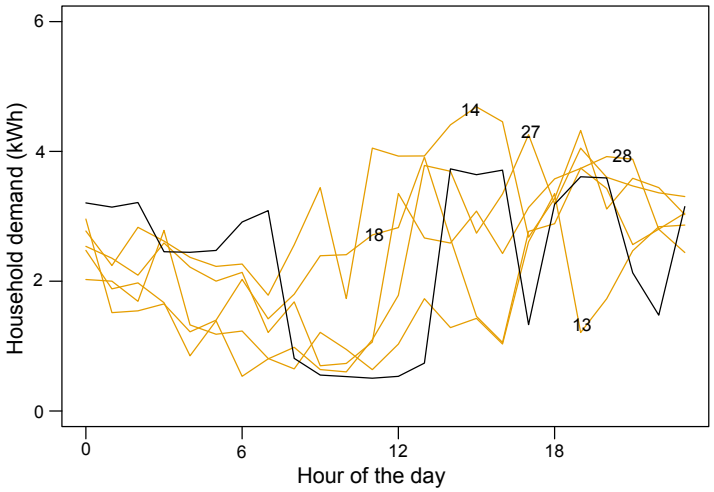
The scenario reduction is made with the fast forward method relying on a distance function between two scenarios, see Section 5.1.3.2. The 3 distances introduced are

tested: point-wise distance d_0, profile characteristics distance d_1, and price-weighted household demand distance d_2. In addition, a benchmark is tested by randomly selected K scenarios among the $S = 400$ total scenarios, labeled Random Representatives. The quality of the reduced set of scenarios is bounded by that of the total set of scenarios. We expect that the quality of the reduced set improves when the number of representatives K increases.

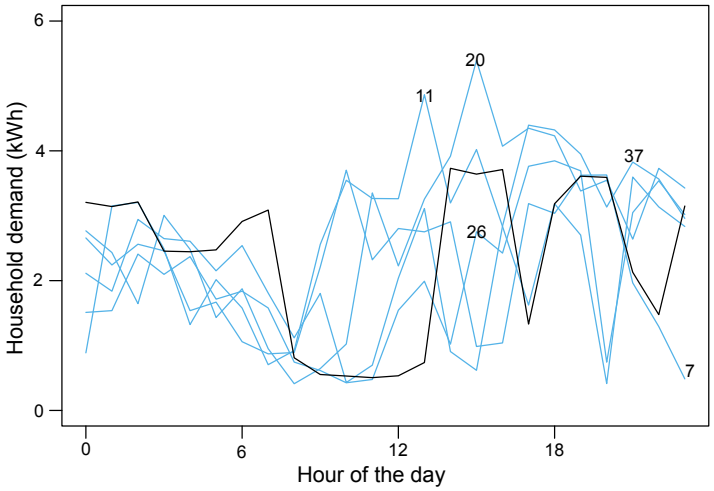
Impact of the Distance used for Reduction The selection of the distance used to perform the fast forward reduction algorithm is key to obtain an efficient reduced set. This distance must discriminate the scenarios between each other. Figure 5.10 depicts the reduced set of 5 representatives scenarios for the 3 distances, for a given household on a given day. The labeled percentage indicates the probability of occurrence (in %) of the corresponding representative scenario. The actual demand, unknown when the scenarios are generated then reduced, is plotted in black. Some observations can be made from this example:

- The representative scenarios obtained with the Point-Wise distance (see Figure 5.10a) have a similar shape throughout the day with rare crossings between them. The range of the representatives is rather constant, even during the night. Besides, the representative scenarios can be almost ordered according to the total daily energy forecast: the scenario requiring the least daily energy is the one requiring the least hourly energy for every one of the 24 hours. Furthermore, the probabilities assigned to the scenarios are fairly homogeneous, from 13 to 28%. It all comes from the known fact that point-wise distance is not well suited to compare household demand time series. It barely discriminates scenario shape, is conservative regarding the peak demand, and the reduction results in almost ordered representative scenarios.

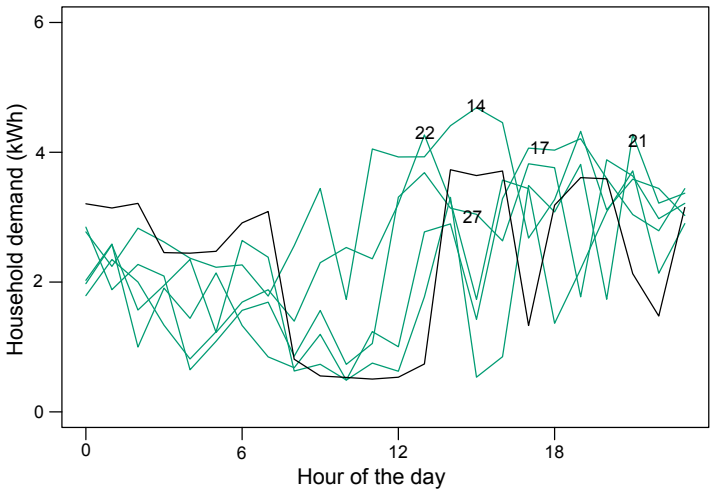
- The representative scenarios obtained with the Profile Characteristic distance (see Figure 5.10b) are more diversified. Due to the distance definition, the night values do not matter and so the curves are indistinguishable during this period. The 5 curves then often cross and have various peak hours with important fluctuations. Probabilities obtained are varied, from 7 to 37%. Since daily energy is one characteristic of the distance, the scenarios are fairly ordered. Reduced sets



(a) Point-Wise Distance (d_0)



(b) Profile Characteristics Distance (d_1)



(c) Price-Weighted Household Demand Distance (d_2)

Figure 5.10 – Reduced sets 5 representatives scenarios obtained using various distances. Probabilities (in %) of scenarios are labeled. The black lines represent the actual demand profile.

obtained with this distance have large spread during the evening, capturing the situations occurring on real demand profiles. However, due to the distance definition, the curves after the peak hours seem quite unrealistic. The demand values seem extremely spread but some scenarios seem very smooth in this period.

- The representative scenarios obtained with the Price-Weighted Household Demand distance (see Figure 5.10c) are also diversified. Since the electricity price during the night is low, scenarios are indistinguishable during the night. There are a lot of crossings between scenarios, and the ramps described are realistically depicting the actual demand profiles observed. However, scenarios are almost equally probable, from 14 to 27%, and the peak demand values obtained are quite similar. The permutation allowed in the distance definition causes a rather conservative anticipation of the peak demand. On the other hand, and in opposition to the representative scenarios obtained with d_1, the fluctuations in the late evening are important as observed on the real curve.

Detailed Results We validate these observations by computing performance scores on a larger scale. For the $S = 400$ scenarios generated with the Refined Covariance method, for every day and household, we apply the reduction process with distances d_0, d_1, and d_2 to obtain $K = 5$ representative scenarios. The performance scores separately evaluate the quality of the anticipated daily total demand (χ_1), the peak demand (χ_2), the maximal ramp (χ_3), the daily cost (o), and the evening demand (ν). The average scores, and their standard deviation in parentheses, are reported in Table 5.3. They are put in parallel with benchmark scores obtained with Random Representatives, and the optimal scores obtained with the set of 400 scenarios obtained with Refined Covariance.

Some conclusions can be drawn by examining the results:

- The reduction method based on the characteristics (distance d_1) is logically the more efficient regarding the characteristics scores. Both the deterministic and probabilistic evaluation show that the performance expected are close to the optimal performance, i.e. the one obtained with the $S = 400$ scenarios. The relative degradation ranges between 1% for the daily total demand and 3% for the maximal ramp, while the benchmark degradation is around 10%.

Table 5.3 – Scores obtained with the 4 reductions methods with $K = 5$ representatives from the scenarios generated by the Refined Covariance method

$K = 5$	NMAE	PICP (10—90%)	NCRPS
Daily total demand (χ_1)	%	%	%
Random Representatives	19.9 (7.1)	71.3 (3.3)	14.6 (5.1)
Point-Wise d_0	19.6 (7.4)	69.4 (6.3)	14.3 (5.4)
Profile Characteristics d_1	**18.2 (6.6)**	**75.1 (3.8)**	**13.0 (4.7)**
Price-Weighted d_2	18.8 (7.0)	74.6 (4.2)	13.5 (5.0)
Refined Covariance	*18.0 (6.5)*	*79.8 (3.5)*	*12.8 (4.5)*
Peak demand (χ_2)	%	%	%
Random Representatives	77.2 (44.9)	69.9 (3.6)	55.7 (31.1)
Point-Wise d_0	82.3 (45.8)	62.9 (9.8)	60.5 (31.5)
Profile Characteristics d_1	**72.1 (43.3)**	**76.6 (4.5)**	**50.2 (28.3)**
Price-Weighted d_2	76.0 (42.7)	69.9 (7.4)	54.5 (28.5)
Refined Covariance	*70.8 (41.4)*	*77.4 (4.0)*	*49.3 (27.3)*
Maximal ramp (χ_3)	%	%	%
Random Representatives	62.6 (36.5)	**63.7 (6.2)**	45.6 (25.5)
Point-Wise d_0	68.3 (36.0)	52.0 (9.9)	52.5 (25.7)
Profile Characteristics d_1	**59.4 (35.4)**	63.0 (7.8)	**42.1 (23.4)**
Price-Weighted d_2	61.8 (33.5)	60.1 (8.2)	45.8 (23.0)
Refined Covariance	*58.1 (33.5)*	*70.7 (7.6)*	*40.7 (22.4)*
Daily cost (o)	%	%	%
Random Representatives	23.5 (8.0)	71.3 (3.2)	17.3 (5.9)
Point-Wise d_0	22.9 (8.6)	70.5 (6.2)	16.6 (6.2)
Profile Characteristics d_1	**21.5 (7.7)**	75.7 (3.7)	**15.4 (5.5)**
Price-Weighted d_2	22.0 (7.5)	**76.6 (4.0)**	15.8 (5.7)
Refined Covariance	*21.3 (7.5)*	*79.9 (3.5)*	*15.1 (5.2)*
Evening demand (ν)	%	%	%
Random Representatives	26.4 (9.5)	**72.2 (3.1)**	19.4 (6.9)
Point-Wise d_0	24.2 (8.4)	66.9 (8.5)	17.9 (6.0)
Profile Characteristics d_1	24.9 (8.6)	72.1 (4.9)	18.2 (6.2)
Price-Weighted d_2	**24.0 (8.5)**	68.6 (7.4)	**17.6 (6.0)**
Refined Covariance	*23.6 (8.1)*	*80.9 (2.9)*	*16.8 (5.8)*

- Although the characteristics distance does not take into account the electricity price, its reduction leads to better daily cost anticipation than the price-weighted distance d_2. While this seems surprising, this is due to the discrepancy between real prices p_h (unknown when the scenarios are created) and median prices $\hat{p}_h$ (used in the definition of metric d_2).

- Since d_1 does not consider demand levels after the peak hour (usually around 18:00), the evening demand is poorly predicted. For this score, the price-weighted distance is the most efficient reduction method.

- As expected, the point-wise distance d_0 poorly forecast the different characteristics of the daily demand profile. According to our evaluation, it performs only slightly better than picking random representative scenarios.

Impact of the Number of Representatives We perform the same tests, but with a larger number of representatives, i.e. $K = 20$, and report the result in Table 5.4. As expected, all the scores are lower than those for $K = 5$ representatives, meaning that the performance is increased. In most cases, the optimal method lead to results very close to those of the 400 scenarios generated with the Refined Covariance Method. We show the typical impact of the number of representatives K for a typical household in Figure 5.11. We compare the performance on two criteria: the maximal ramp χ_3 and the evening demand ν; with two reduction methods: one based on distance d_1 and the other with random representatives (RR). The relative CRPS errors are on the y-axis, i.e. the ratio between the CRPS of the reduced set of scenarios of K representatives and the optimal CRPS (obtained with 400 scenarios). We see that the optimal performance is reached between $K = 5$ and 20, depending on the criterion examined, with the reduction based on d_1

5.1.5 Conclusion

With the hourly electricity demand values during one year of a dataset of 175 US households, we perform a day-ahead probabilistic forecasts of each hourly values. The probabilistic forecasts assess the possible demand range for a specific hour through a probabilistic function. However, when one studies the daily profile, i.e. the collection of the 24 hourly values, one analyzes the collection of 24 deterministic values rather than

Table 5.4 – Scores obtained with the 4 reductions methods with $K = 20$ representatives from the scenarios generated by the Refined Covariance method

$K = 20$	NMAE	PICP (10—90%)	NCRPS
Daily total demand (χ_1)	%	%	%
Random Representatives	18.5 (6.6)	**79.8 (3.2)**	13.2 (4.7)
Point-Wise d_0	19.0 (7.1)	77.9 (4.2)	13.5 (5.0)
Profile Characteristics d_1	**18.1 (6.5)**	81.1 (3.3)	**12.8 (4.5)**
Price-Weighted d_2	18.5 (6.8)	80.4 (3.4)	13.1 (4.7)
Refined Covariance	*18.0 (6.5)*	*79.8 (3.5)*	*12.8 (4.5)*
Peak demand (χ_2)	%	%	%
Random Representatives	72.9 (43.3)	77.6 (3.8)	50.9 (28.6)
Point-Wise d_0	76.1 (42.5)	74.6 (6.4)	53.6 (28.3)
Profile Characteristics d_1	**71.6 (42.6)**	**79.7 (3.9)**	**49.6 (27.9)**
Price-Weighted d_2	73.1 (42.0)	76.9 (5.2)	51.1 (27.6)
Refined Covariance	*70.8 (41.4)*	*77.4 (4.0)*	*49.3 (27.3)*
Maximal ramp (χ_3)	%	%	%
Random Representatives	59.7 (35.1)	71.1 (6.9)	42.0 (23.6)
Point-Wise d_0	63.5 (34.4)	66.1 (8.3)	45.7 (23.2)
Profile Characteristics d_1	**59.0 (35.1)**	**71.5 (7.1)**	**41.1 (23.2)**
Price-Weighted d_2	60.4 (34.2)	68.3 (7.3)	42.9 (22.7)
Refined Covariance	*58.1 (33.5)*	*70.7 (7.6)*	*40.7 (22.4)*
Daily cost (o)	%	%	%
Random Representatives	22.0 (7.7)	**79.8 (3.3)**	15.7 (5.4)
Point-Wise d_0	22.2 (8.2)	78.8 (4.0)	15.8 (5.7)
Profile Characteristics d_1	21.4 (7.5)	81.6 (3.3)	**15.2 (5.3)**
Price-Weighted d_2	**21.3 (7.5)**	81.5 (3.2)	15.4 (5.4)
Refined Covariance	*21.3 (7.5)*	*79.9 (3.5)*	*15.1 (5.2)*
Evening demand (ν)	%	%	%
Random Representatives	24.3 (8.5)	**80.6 (2.9)**	17.5 (6.1)
Point-Wise d_0	**23.9 (8.4)**	76.9 (5.2)	17.2 (6.0)
Profile Characteristics d_1	**23.9 (8.4)**	81.2 (3.3)	17.2 (6.0)
Price-Weighted d_2	**23.9 (8.5)**	78.3 (4.5)	**17.1 (6.0)**
Refined Covariance	*23.6 (8.1)*	*80.9 (2.9)*	*16.8 (5.8)*

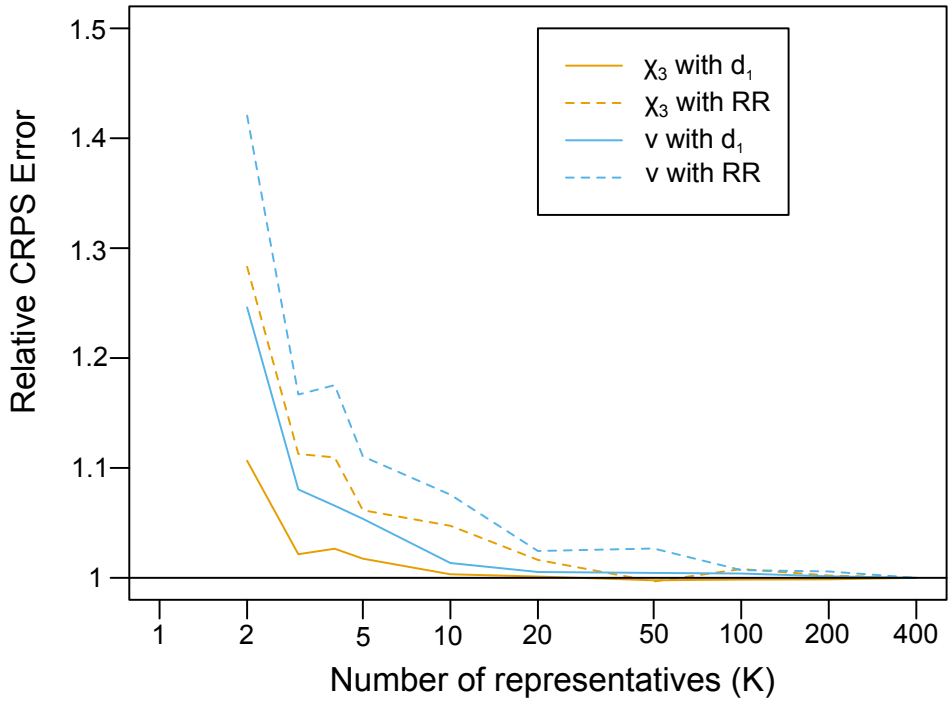


Figure 5.11 – The ratio of CRPS error (y-axis) according to two criteria: χ_3 and ν for two reduction methods: one based on distance d_1, and one with random representatives (RR). The horizontal black line represents the optimal performance, i.e. the one achieved when using all the 400 scenarios. The performance is depicted for various number of representatives K (logarithmic x-axis).

a complex 24-dimensional function. Generating a scenario consists in picking multiple collections of these values while maintaining the probabilistic quality of the hourly forecasts. A basic method, that we call Connect-the-Quantiles, consists in supposing that the successive values are completely independent, so if the actual demand at 15:00 is high — i.e. falls in the upper part of the forecast distribution — then the demand at 16:00 is also high. Conversely, the Uniform Random sampling method supposes that the successive values are completely independent. In reality, the demand profiles fall in between. Certain hours are strongly correlated between each other but not with others, thus creating clusters of correlated hours depicting the habits of a household. These habits can be mimicked by the use of a 24-dimensional Gaussian distribution with an adequate covariance matrix. While the most basic covariance matrix generates accurate scenarios, we introduce a refined covariance matrix that lead to minor improvement.

We show that, on average, generating a total of 400 scenarios is to necessary to have optimal probabilistic forecasts for each hour of the day.

Since the forecasting demand scenarios as generally used in later computational-intensive applications, one wants to reduce the number of scenarios by finding a small set of representative scenarios, i.e. finding $K \ll S = 400$ representative scenarios, that accurately anticipate the different phenomena that may occur. The reduction is based on a distance metric that compares how different are two demand profiles. This metric is to be defined according to the later application. We describe 3 metrics and each one of them results in a different reduced set. We examine the performance of the reduced sets according to different criteria and show that a metric based on key characteristics of the demand profiles result in good performance. Regarding the number of representatives, we find that between 5 and 20 representatives are sufficient to describe the variety of the original set of 400 scenarios.

5.2 Electric Vehicle Charging Scenarios

5.2.1 Introduction

In this section, we focus on the electric demand made by an Electric Vehicle (EV). This device operates on a switch-on mode: it can be either in charge or not. It is therefore comparable to major domestic appliances ('white goods') such as stove or laundry dryer. EVs require a very large amount of energy in a small amount of time: power drawn from the grid is large during short definite periods. Dickert and Schegner represented typical appliances on a 2D graph with one axis for the power of the appliance, and one for the annual frequency of use (Dickert & Schegner, 2010). In such a graph EVs would be on the right side of the graph with power drawn higher than stove. We represent a similar graph on Figure 5.12 where we show energy versus peak power for main domestic appliances of a typical US household. Such characteristics (important energy and high power) are demanding for the power network, and operators therefore are interested in modeling how EV charges occur.

Electric vehicles are used in a multitude of context depending on the owner's culture (Glerum et al., 2013). For instance, professional vehicles and privately-owned vehicle have different usage depending on the amount of people driving and their schedules.

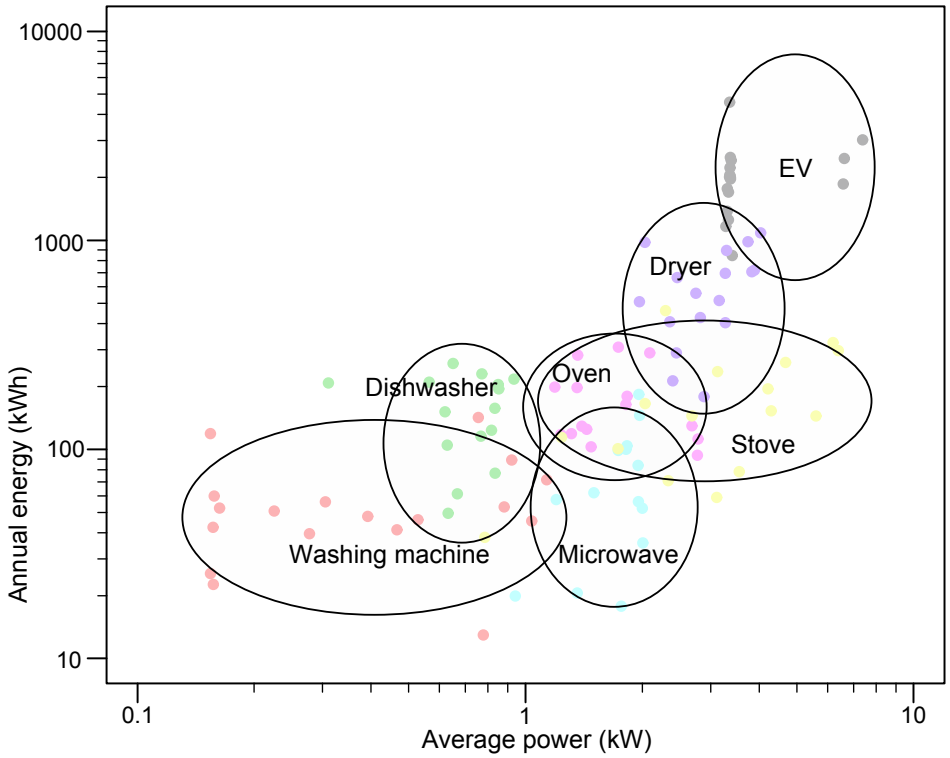


Figure 5.12 – Scatterplot of annual demand (in kWh) versus the average power (in kW) when device is switched on for typical appliances of a US household. Source: processing of raw data from Pecan Street.

This variety of usage highly reflects in the charging cycles of EVs and challenges the modeler. Due to its nature, an EV can be charged in different places (at home and at the workplace) which impede a traditional switch-on appliance model. Bae and Kwasinski propose a spatial model to account for different charging stations (Bae & Kwasinski, 2012). Modeling a single EV is difficult to validate with real data since power called is measured at the charging station level and not at the vehicle level, making it difficult to know exactly the power drawn by a specific vehicle. Conversely, if one models the charging station, there can be multiple vehicles charging sporadically. However, this latter approach is the one we take with our data-driven study: we analyze power drawn by private EVs at a private charging station, and model only the power from this plug. In the following, we will refer to EV demand to denote the power delivered by this single plug.

EV charging is a controllable load such as the washing machine or the water heater.

As such, EV offers offer advantageous flexibility for demand response purposes. For instance, shifting charging cycles during the night when electric demand is low. EVs can also be used as a battery to be injected on the grid (Gough et al., 2017), or to stabilize the system (Tomić & Kempton, 2007). While most works study flexibility of all the appliances of the house, for an individual household (Florez et al., 2017) or at the aggregated level (Ponoćko & Milanovic, 2018), we will thereafter focus only on the power drawn solely by EV, independently of the rest of the household's demand.

In the following, we model charging with the help a dataset of 46 EVs located in Austin, Texas, from Pecan Street (*Pecan Street Inc. Dataport*, 2018). Power drawn by each vehicle is measured for every minute of the year 2015.

5.2.2 Detection of Charging Blocks

Time series of power drawn by an EV is modeled as a simple function with two states. Power drawn is either null, when the EV is not charging, or equal to a certain nominal power, when the EV is charging. Figure 5.13 shows an extract of a time series where the two states are visible. Since real measures are noisy, power drawn slightly fluctuates around nominal power when charging. Time series is therefore not comprised of perfect rectangles. An ideal charging block has three characteristics schematized in Figure 5.14:

- *Nominal power*: the power drawn from the grid is constant during the whole charging period. This power is defined by the type of battery and charging station.
- *Duration* of the charging period.
- *Start-up time*: the instant of the day when EV charging starts.

From our observations, nominal power is always the same as long as there is no technological change (i.e. battery or charging station replacement). Most current private charging stations do not offer different charging power levels. In measurements, the ramp up to the nominal power is not infinite, and it takes some time to reach this value. For most EVs (35/46), it takes less than 15 minutes to reach nominal power, for the others, it takes between 15 and 60 minutes. We thus model it with a perfect rectangle.

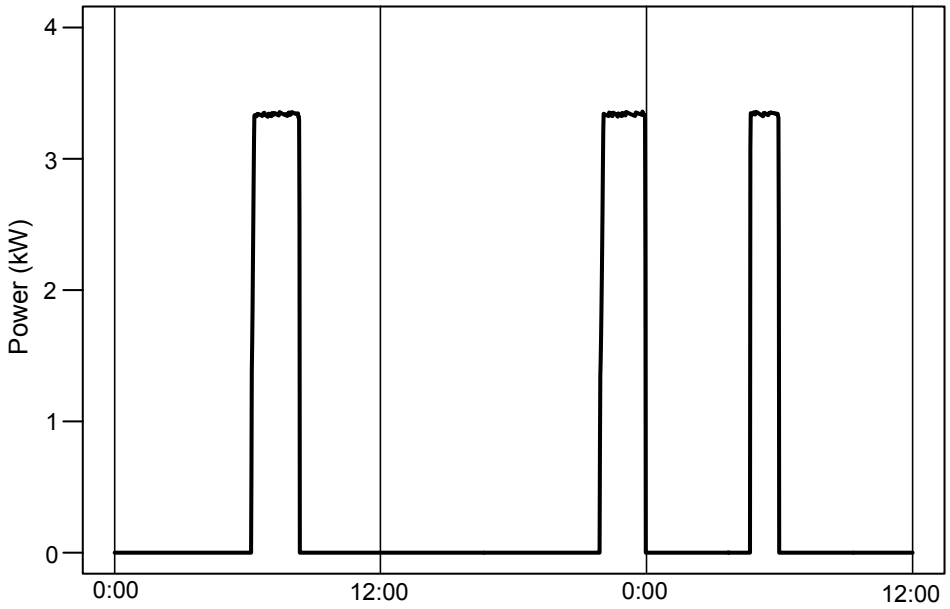


Figure 5.13 – Power drawn every minute by an EV during 36 successive hours. Power is null when the EV is not charging, and is very close to a nominal power when charging.

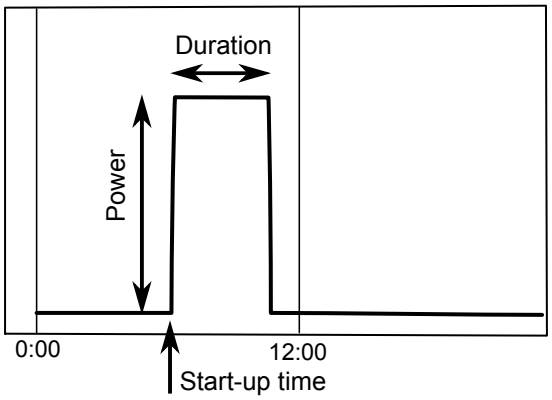


Figure 5.14 – Charging block model with 3 characteristics: power, duration and start-up time.

On the other hand, duration and start-up time are not fixed. Since we analyze daily profile, we assume that the minute when charging starts is between 1 and 60×24. In any case, charging blocks do not start at the exact same time each day, and do not last the same duration: it depends on the unknown user's habits. The different values of these parameters are to be detected on the time series of power measures in order to characterize habits of a particular EV/user.

The following procedure is implemented to model charging habits of an EV:

(1) *Detecting nominal power.* Density of all the strictly positive values is estimated, and the maximum of this function (i.e. the statistical mode) is retrieved as the nominal power[3].

(2) *Transforming time series in simple time series.* A threshold defined as 80% of the nominal power is defined. The raw time series is transformed in a series with two values, equal to 0 when power measured is below the threshold, and 1 when it is above.

(3) *Pre-processing the simple time series.* Refinement is made on the transformed time series to account for error measures. Too short remaining blocks (less than 20 minutes) are removed from the time series.

(4) *Detecting duration and start-up time.* A straightforward data treatment is operated to list all the start-up time and associated duration from the time series.

The whole procedure runs fast on an average computer: less than 10 seconds to go through the 525,600 points of an EV yearly time series.

5.2.3 Analysis of Charging Characteristics

The procedure is run for each of the 46 EV time series. We thereafter review the results obtained.

Most EVs have a nominal power between 3.2 and 3.7 kW (see Table 5.5). Table 5.6

Table 5.5 – Nominal power of the vehicles.

Nominal power (kW)	1.5	3.2 to 3.7	6.2 to 7.3
Number of EVs	1	37	8

reports the number of days with certain number of charging blocks (0, 1 or more than

[3] As explained, nominal power is unique but can suddenly change with technological replacement. It happens on 2/46 of our EVs. We do not model such a rare event, since such an abrupt event is unpredictable with the power time series. We manually define two levels for the 2 troublesome time series.

2) for two randomly selected EVs, and the average for the 46 EVs. Most days, people charge their EV 0 or 1 time. Furthermore, when considering solely days with more than 2 charging blocks, the longest one accounts for two thirds of daily energy. It shows that the other charging blocks are residual and modeling the largest block largely prevail.

Table 5.6 – Number of days with 0, 1, or more than 2 charging blocks, for two random EVs and in average.

Number of blocks	Number of days		
	EV α	EV β	Average EV
0	98	209	150
1	233	108	158
≥ 2	34	48	57
Total	365	365	365

To understand EV charging habits, a scatterplot of duration against start-up time is useful. Figure 5.15 represents every charging blocks of an individual EV during one year, detected with our procedure. The x-axis tells the minute of the day when charging starts, and the y-axis tells us the corresponding duration of the block. Colors and shapes of the points indicate if the charging block represented is occurring during a weekday or weekend day, and if this is the longest block of the day.

As it can be seen from the graph, there is a clear relation between duration and start-up time. For this vehicle, charging blocks occurring during the evening last longer (up to 10 hours) than charging blocks during the morning (usually around 50 minutes). A tentative explanation is that user charges completely her or his vehicle in the evening after work, and on other occasions, she or he charges it rapidly in the morning before leaving home. Colors and shape give us more information about the habits pattern. There is a notable difference between longest and residual blocks in duration. However, start-up time is approximately the same for longest or residual blocks. Difference in habits is not clear between weekdays and weekend. The graph shows no real distinction between circle and triangle.

To support this visual analysis, a statistical test is computed for different distribution. The null hypothesis $\mathcal{H}$ = "Both block samples come from the same distribution"

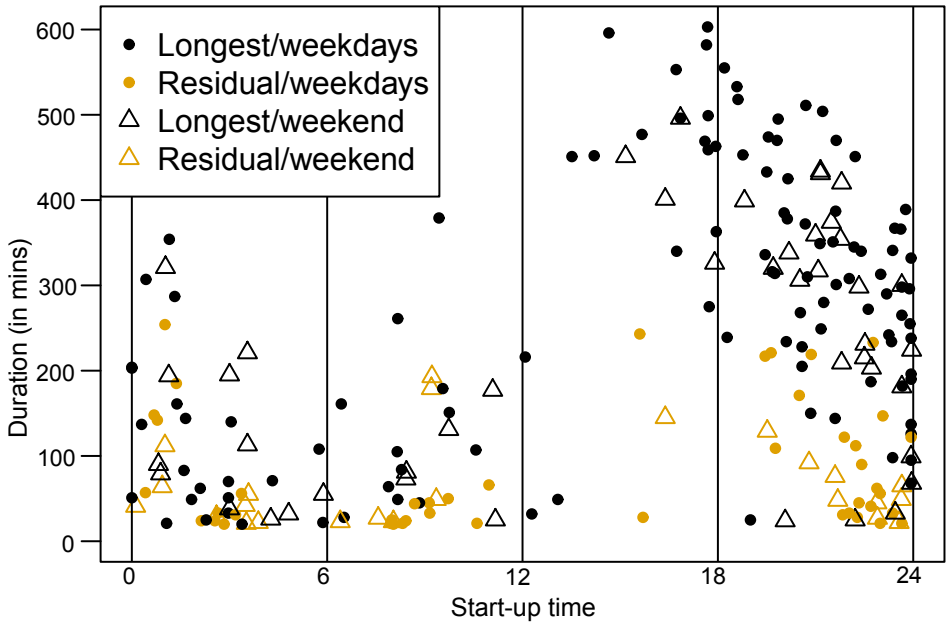


Figure 5.15 – Each point represents a charging block of a specific EV during one year. Minute of the start-up time is on the x-axis, and duration of the block on the y-axis. Filled circles, resp. empty triangles, indicate that the charging occurred during a weekday, resp. a weekend day. Colors indicate if this is the longest block of the day or a residual block.

is tested for 3 cases :

$\mathcal{H}_1$: Longest and residual blocks come from the same 2D distribution (duration $\times$ start-up time);

$\mathcal{H}_2$: Longest and residual blocks come from the same 1D distribution (start-up time);

$\mathcal{H}_3$: Week days and weekend blocks come from the same 2D distribution (duration $\times$ start-up time);

In each case, function `kde.test`, implemented in R package *ks* (Duong et al., 2012), estimates density functions for both samples and computes the integrated squared error to obtain the statistics. In the case of the EV charging blocks on Figure 5.15, respective p-values of the three tests are $3 \cdot 10^{-16}$, 0.39 and 0.60. As visually observed, there is no statistical difference in patterns between weekdays and weekend, and in start-up time between longest and residual blocks. On the other hand, hypothesis $\mathcal{H}_1$ is rejected with strong confidence.

Tests are computed for the 46 EVs[4], and results are given in Table 5.7. For the distinction between longest and residual blocks, 2D samples are significantly different in most cases ($\mathcal{H}_1$). However, in half the cases, the difference comes only from duration ($\mathcal{H}_2$), and start-up time are almost the same for the two kind of charging blocks. It means that in half the cases, there is a trend such as "longest block in the night & residual blocks in the morning". In the other cases, one does not know if a block starting at a certain instant is a long or a residual block. This result is interesting for intraday forecasting, where one wants to know how long a charging block that just started will last. Concerning the distinction between weekdays and weekend, hypothesis $\mathcal{H}_3$ is never statistically rejected (p-values < 0.01). Additional test hypotheses on different days (e.g. "is a Friday different from the rest of the week?") are almost never rejected (e.g. Friday is similar to the rest of the week). It means that EV users do not change their charging patterns (duration×start-up time) for any day of the week.

Table 5.7 – Number of EVs for which the three hypotheses are rejected or not (p-value < 0.01)

Hypothesis	Rejected	Not rejected	Total
$\mathcal{H}_1$	11	33	44
$\mathcal{H}_2$	24	20	44
$\mathcal{H}_3$	0	45	45

The fact that charging patterns are similar for all days of the week is convenient for training since all blocks provide a description of the habits. On the other hand, describing longest and residual blocks is more troublesome.

5.2.4 Bottom-up Forecasting

We propose to forecast the next-day profile of a fleet of EVs thanks to the precise analysis of charging characteristics of each EV. Contrary to forecasting methods that do not take into account the nature of the consumption presented in Chapter 4 and consider demand as a whole, we try here to use individual EV models (i.e. the patterns

[4]For 2 EVs, there is never more than 1 charging block per day, so $\mathcal{H}_1$ and $\mathcal{H}_2$ are not tested. For 1 EV, there is no charging blocks at all during any weekend, so $\mathcal{H}_3$ is not tested.

learnt during analysis) to generate individual scenarios and construct the aggregated consumption of the fleet by summing up individual profiles.

For each individual EV, we forecast a scenario for next day consumption profile in 3 steps:

1) forecast number of charging blocks;

2) forecast possible characteristics (duration×start-up time) for each block;

3) add charging blocks multiplied by EV's nominal power to the consumption profile.

For step 1, the forecasting model we propose is a random forest using the following inputs : weekday, number of blocks 1 day ago, number of blocks 7 days ago, median number of blocks during the 7 previous days, mean temperature of the previous day. These inputs have been selected based on demand forecasting model and experience. A probability random forest is implemented in the R package *ranger* (Wright & Ziegler, 2017), details of the algorithm is given in (Malley et al., 2012). Parameters of the forest are kept at their default values after observing that they were close to optimal. Random forest provides a convenient way to draw a random number of charging blocks according to forecast probabilities. For step 2, we select characteristics (duration×start-up) according to the 2D distribution observed. New block characteristics are drawn from previous charging blocks. These blocks are weighted by a decreasing exponential law of parameter λ, i.e. ancient blocks are forgotten when time goes by. A Gaussian noise, with covariance matrix estimated from all the previous charging blocks, is added to the 2D point drawn. Several checks are operated to rule out impossible situations: overlapping blocks, negative duration and so on.

The forgetting factor λ is delicate to tune: if it is too small, only the most recent blocks are considered and variety is low; conversely, if it is too large, too many blocks are used and recent effects are not considered. Value of λ describes the speed with which behavior changes. The issue is detecting this speed. If it is specific to each user, the optimal value is difficult to select since there is no straightforward way to assess quality of forecasts at the individual level. When forecast quality is assessed on aggregated consumption, only a single value of λ common to all users is optimized. In any case, our experience shows that a forgetting parameter λ between 1 and 50 days gives approximately the same results. Therefore, λ is set up equal to 50 days in the following.

5.2.5 Forecasting Performance

We compare the forecasting performance of our bottom-up method with two benchmark methods which do not model individual EVs but considers only the aggregated consumption: a persistence model, and a gradient bossting tree model.

The first benchmark model is a persistence model. Value of aggregated at the same minute of the previous day is used as point forecast. An artificial probabilistic forecast is obtained by using the point forecast as a Dirac probability distribution function.

An advanced benchmark is also proposed. Specifically, a gradient boosting tree (labeled as GBM, from package *gbm* (Ridgeway, 2017)) to directly forecast aggregated consumption of the 46 EVs. The 5 following inputs are selected based on previous experience on aggregated consumption: the minute of the day, the weekday, temperature forecast, consumption 1 day ago, median consumption during the 7 previous days. Parameters of the model are carefully tuned (i.e. number of trees, shrinkage parameter, and tree width) and probabilistic forecasts are made using pinball loss: a total of 19 boosting trees, for quantiles $\tau = \{0.05, 0.10, ..., 0.95\}$, are computed. A cross-validation approach is made, therefore training and test sets are randomly selected, across the whole year. According to function `relative.influence` implemented in the package *gbm*, median consumption during the 7 previous days is the most important input, with a relative influence between 80 and 90%, then the minute of the day around 10%. As previously observed, temperature and weekdays have almost no influence whatsoever.

For our bottom-up approach, the three-step procedure described is done for each EV. The number of blocks forecast (step 1) is done with a probabilistic random forest. The `variable.importance` implemented in the package *ranger* shows that forecasting performance of next day number of charging blocks are the most influenced by the temperature of the previous day (51%), then the 3 inputs regarding the number of blocks during the previous week (30%), and then the day of the week (19%). Contrary to the gradient boosting tree at the aggregated level, influence of temperature and day of the week is much more important to forecast number of charging blocks at the individual level. The second step consists on selecting duration and start-up time of the blocks forecast. As explained, these parameters are drawn from previous historical blocks characteristics. Blocks are weighted with an decreasing exponential of parameter $\lambda = 50$ days to favor more recent blocks. These 3 steps are used to create S scenarios

for each day of the year. To assess quality, scenarios are turned into probabilistic forecasts by computing quantiles $\tau = \{0.05, 0.10, ..., 0.95\}$ from the S values of each minute.

Figure 5.16 illustrates the interest of the bottom-up approach by furnishing a precise decomposition of the total load of the fleet by EVs. The graph represents each individual EV consumption scenario in filled areas. The sum of all the 46 individual scenarios is used as an aggregated scenario that forecasts the actual aggregated consumption (orange dashed line). High consumption during the night is correctly forecast, as well as the very low consumption in the early afternoon. Highly volatile consumption and high peaks (between 4 and 6 in the morning) are difficult to grasp with our bottom-up model since too short charging blocks (under 20 minutes) are not simulated at all to avoid noisy measures.

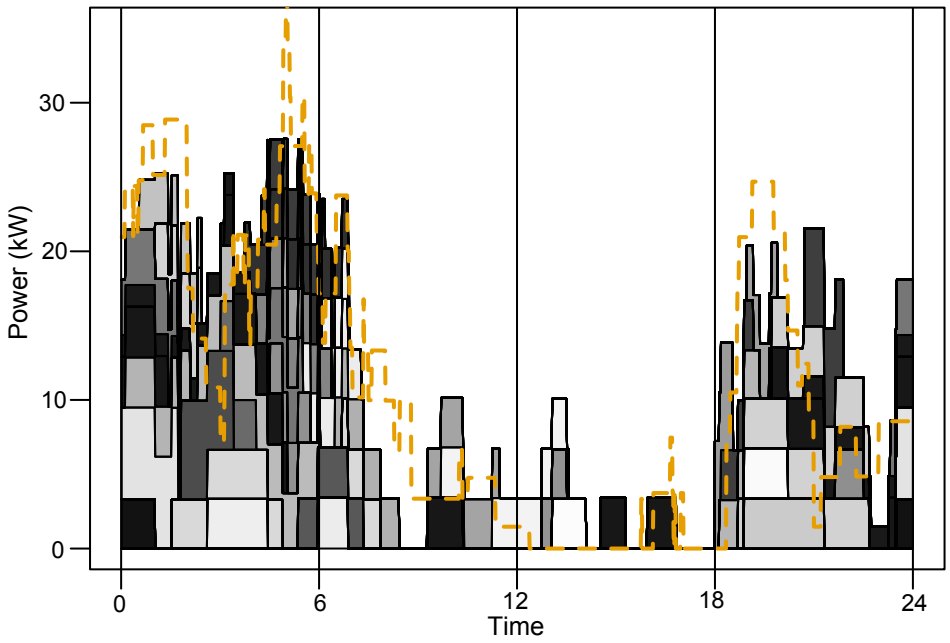


Figure 5.16 – Day-ahead minute scenario forecast of a fleet of 46 EVs on Saturday 12th December 2015. Orange dashed line shows the actual consumption to be forecast. Each individual scenario is represented by a filled area. The sum of all these scenarios is used to forecast the aggregated consumption.

Multiple scenarios for each day of the year (minus the three first months of the year used as a burn-in period) are generated and the forecast performance is evaluated. Table 5.8 reports results. We compute MAE and CRPS by comparing forecast series

and measured series every minute. Both scores are in the same unit (kWh) as the aggregated consumption. This consumption usually fluctuates between 3 and 20 kWh, with a mean of 11 kWh and peaks up to 50 kWh. As it can be seen, performance of the gradient boosting tree method is around 45%, with a MAE around 4.9 kWh. We report performance of our bottom-up approach for two different numbers of scenarios generated, 20 and 400. Our bottom-up approach always beats the persistence model, but need sufficient number of scenarios to reach GBM's performance. Figure 5.17 depicts the decrease of CRPS depending on the number of scenarios S[5]. There is an irreducible error equal to 3.6 kWh for the CRPS (resp. 4.9 for the MAE) attained for around 100 scenarios. Probabilistic forecasts are more favorable to our bottom-up approach which better captures the load distribution. Figure 5.18 depicts the quantile scores for the benchmark persistence model, the GBM as an advanced benchmark, and our bottom-up approach for 20 and 400 scenarios. As it can be seen, for probabilistic method, there is asymmetry in the results, i.e. lower tail is better approximated that the upper tail. This inevitable behavior is due to the positive skewness of the actual consumption. This asymmetry is stronger for our approach than the GBM. Therefore, although performance is very similar for the upper tail, it is undeniably advantageous to prefer bottom-up approach over GBM for the lower part of the distribution.

Table 5.8 – Forecasting performance of aggregated consumption of 46 EVS of 4 models: a persistence model (previous day), a gradient boosting tree (GBM) model, and our bottom-up forecast for 20 and 400 scenarios generated.

Score	Persistence	GBM	Bottom-up	
			$S = 20$	$S = 400$
MAE	6.24	4.86	5.03	4.87
CRPS	6.24	3.63	3.75	3.59

5.2.6 Conclusion

We analyze 46 minute-by-minute time series of the power drawn by individual Electric Vehicle (EV) at a residential charging station. We implement a procedure to detect the

[5]The theoretical decreasing rate for a standard distribution is in $S^{-1/2}$, see Appendix A.

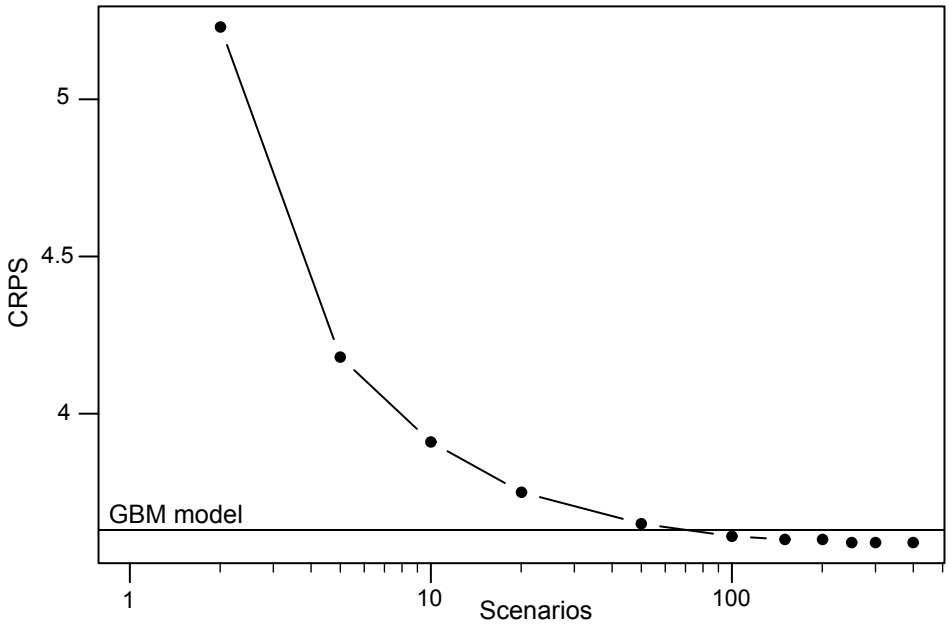


Figure 5.17 – CRPS obtained with different number of scenarios for the bottom-up method. Horizontal line indicates the performance of the GBM model.

charging periods visible on the time series, i.e. the start-up time and duration of all the "blocks" of the series. It enables us to model the charging habits of the user on the 2D graph, and so to grasp the individual EV demand behavior. With this model, we propose a day-ahead forecasting model of this individual demand. We first forecast the number of charging blocks for the next day in a probabilistic manner, specifically we use a probabilistic random forest using carefully selected inputs such as previous charging blocks, weekday, and temperature. We then simulate the corresponding number of blocks according to the habits of the user. Since the whole process is of probabilistic nature, forecasting scenarios of the demand are generated for the whole day. In order to validate these scenarios, we examine the aggregated consumption of the EV fleet, i.e. we sum up all the individual scenarios to forecast an aggregated scenario. This bottom-up forecasting method is compare to a machine-learning forecasting method that deals only with the aggregated demand, i.e. without decomposing the fleet demand in individual EV demand. We obtain similar deterministic and probabilistic performance, with an absolute error around 5 kW, which roughly represent a relative error of 40%.

The forecasting of the individual EV demand requires a preliminary analysis of the EV usage based on measurements at the plug level. This measurement process

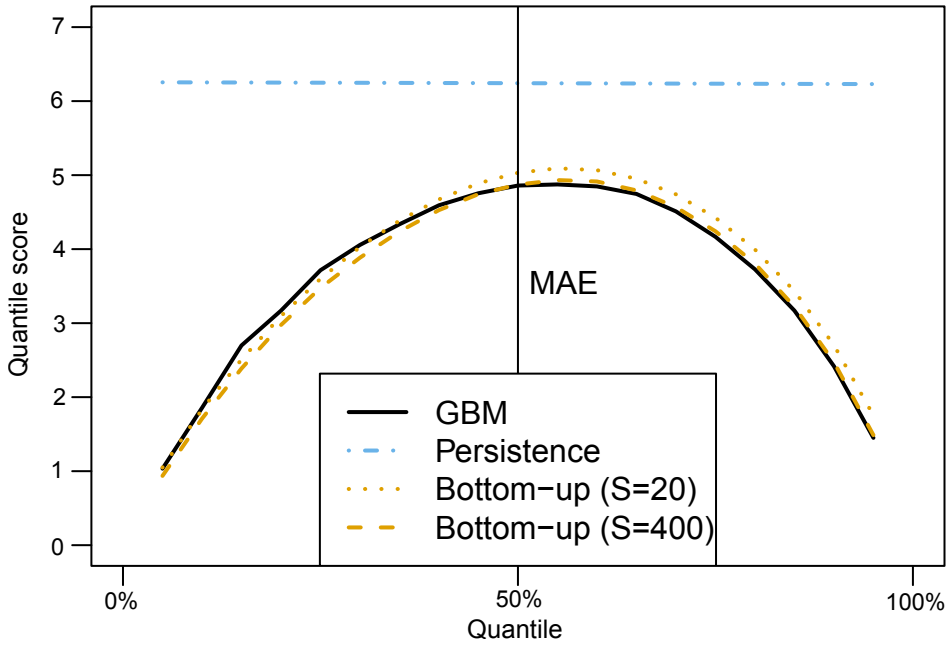


Figure 5.18 – Quantile scores for the persistence model (blue dash-dotted line), the GBM model (black solid line), and the bottom-up model with 20 (orange dotted line) and 400 scenarios (orange dashed line). Intersections between curves and the vertical line at quantile 50% indicate the MAE of each model.

is rather expensive and impractical. However, by analyzing demographic features of the householder (such as work schedule), one can cluster habits so as to anticipate the habits of a similar person, even when this person does not necessarily have an EV yet. The method introduced may therefore anticipae in detail the electricity demand of a new appliance, e.g. an EV, that an individual does not yet possess. We note that such methods can be adapted for other appliances than EV, even though the impact on total household demand is lower, and their usage cycles are less clear than the perfect rectangles of the EV charging blocks. In future work, we expect to model of other major appliances (i.e. dryer, stove, etc.) by detecting their habits in a similar way. This would provide a bottom-up approach to forecast scenarios of the total household demand. Promising applications can emerge, especially regarding short-term flexibility, e.g. quantify how flexible the stove demand of the next day is and find the necessary incentives in order to shift this appliance demand.

Chapter 6

Conclusions and Perspectives

6.1 Conclusions

In Chapter 1, the reasons of the current interest in the local electricity grid are explained, necessitating this work on short-term forecasting of the local demand : decentralization of the electricity production, liberalization of the market, integration of renewable energies, smart-meter roll-outs, emergence of self-consumption. We define the exact scope of the thesis: short-term refers to forecasting horizons going from 1 hour to 1 week, local scale refers to the average power of the case studied, going from 1 kW to 1 MW. We dwell on the challenges of the forecasting task to identify four objectives:

1. Characterization of the electricity demand at the local scale.
2. Development of probabilistic forecasting models.
3. Ensuring the replicability of the models.
4. Generation of daily forecasting scenarios.

In Chapter 2, an introduction to statistical forecasting models is provided, including the most common types of models and how to assess their forecasting performance, be it deterministic or probabilistic. The presentation is made in the electricity demand forecasting context. An overview of the literature on the subject is then drawn, focusing on the short-term horizons. We analyze and compare the forecasting performance

reported to exhibit a known scaling law connecting the forecasting performance with the average power of the case studied: the relative forecasting errors decrease from 30% at the household scale (power around 1 kW) to 3% at the national scale (power around 1 GW).

In Chapter 3, we focus on the electricity demand of a feeder, i.e. the aggregated demand of 1000 to 10,000 people. Demand data at this scale have been measured for a long time, and so the driving effects on the demand are clearly identified, notably the temperature influence. The short-term forecasting of this scale is mature with relative errors around 10 %. We contribute to the understanding of this demand by proposing an algorithm disaggregating the feeder demand in elementary profiles corresponding to the demand of a cluster of similar customers. The algorithm makes use of demand measurements of multiple feeders along with their corresponding customer information system. The resulting elementary demand profiles can be used in various applications, that we illustrate on multiple datasets: forecasting the demand of a new unmeasured feeder, with a relative error ranging from 12 to 15%; and analyzing the evolution of the daily demand peak when new customers are connected to a feeder. Parts of this chapter have been published in the *Applied Energy* journal (Gerossier, Barbier, & Girard, 2017).

In Chapter 4, we deal with household electricity demand, and specifically how to design a short-term forecasting model. The exact characteristics of the demand at this low scale are analyzed in detailed, and compared to these of larger scales for three datasets worldwide. A gradient boosting model is developed and its quality assessed with the datasets: an average deterministic error of 28% for the next day hourly demand values. This model constitutes a reference model for the household scale. With thorough testing, we assess its performance at different levels of aggregation (demand of a single household to aggregated demand of 200 households) and time resolution (demand averaged over 1 minute to 1 week). We conclude that the forecasting errors logically decrease when considering a coarser time resolution, and a larger level of aggregation. Since the error decreasing is not linear, an optimal aggregation level is found when forecasting the aggregated demand of a group of 15 households at once. Moreover, to address the issue of robustness, a hierarchical forecasting framework is then introduced. It combines multiple models to produce probabilistic forecasts in all situations. Once deployed on a real project, we analyze the online performance of the

framework. This work has been presented at the CIRED 2017 conference (Gerossier, Girard, et al., 2017), and in the *Energies* journal (Gerossier et al., 2018). Correa-Florez made use of these household demand forecasts in order to optimize a smart home energy management system (Correa-Florez et al., 2018).

In Chaper 5, forecasting demand with scenarios is analyzed. Common methods produce the forecasts for a single instant, which are suboptimal if one wants to use them for multiple instants. Scenarios address this issue. We present a generation method, by computing the correlation between demand values throughout the day, and a reduction method, by clustering the scenarios in groups according to a designed metric. We find accurate scenarios that are coherent over a daily period and practical for later applications. We then forecast demand scenarios at the appliance scale, specifically for the charging of a residential electric vehicle. Habits related to the charging are analyzed and used in order to forecast the demand of the vehicle for the next day. This study has been presented at the MedPower 2018 conference (Gerossier et al., 2018).

6.2 Perspectives

We identify two promising perspectives on forecasting local electricity demand that can be built upon the research made in this document.

Generation of comprehensive demand forecasting scenarios at multiple scales. We illustrated how to produce accurate probabilistic forecasts at the household and at the neighborhood scale (Chapter 4), and turn them into large, or reduced, sets of scenarios (Section 5.1). However, we proposed and evaluated forecasts independently for the neighborhood and the household scales, and the question of forecasting demand scenario at multiple scale remains open. In fact, the issue of the grid losses inherently prevents the reconciliation of the neighborhood and household scales. Indeed, the two underlying objectives are partially conflicted and one should favor: either an aggregated point of view, for which the grid losses are partially transferred to each household and, thus, the household demand does not correspond to the smart-meter measurement; either an individual point of view, for which there is no loss between the two scales, i.e. the aggregated demand is exactly the sum of all the individual demands. When adopting the aggregated point of view (such as in Chapter 3), the forecaster relies on the

measurements made at this aggregated scale (with losses), and any method to obtain the household demand scenarios will be inaccurate regarding the smart-meter measurements (without losses). When adopting the individual point of view, the forecast relies on household smart-meter measurements and artificially creates a lossless aggregated demand. Generating demand scenarios valid at the household and neighborhood scales even in this lossless case is then a challenging task. Due to the aggregation effect, the demand scenarios at the neighborhood scale are less diverse than those at the household scale. Consequently, the forecaster cannot sum individual scenarios to obtain an accurate aggregated scenario. The interdependence between households should be taken into account, e.g. with consensus constraints on the individual scenarios.

Bottom-up forecasts of the household electricity demand by data-driven habits analysis. The household electricity demand is strongly influenced by the habits of the resident. However, it has been noted that very detailed surveys about the householder are necessary to identify the electricity-related habits, if not impossible since the habits sometimes root in unconscious practices (Gram-Hanssen, 2014). Fortunately, they are reflected in the electricity demand patterns. On the one hand, disaggregation algorithms are able to detect these patterns and their corresponding appliance with high frequency smart-meter measurements. On the other hand, invasive monitoring infrastructure can be set up to retrieve the electricity demand of specific appliances. In any case, it is not clear how these data can be used to produce short-term forecasts. We proposed (in Section 5.2) a precise modeling of an appliance habits, namely of an electric vehicle, and showed how the habits can be turned into accurate day-ahead scenarios, atleast as accurate as a machine-learning forecasting model ignoring the habits. Forecasting appliance scenarios for these switch-on appliances (combined with scenarios for the remaining electricity demand) is expected to provide a bottom-up approach to generating scenarios for the total household demand with optimal efficiency. We do not expect that such approach will improve the forecasting performance of the household demand. However, the disaggregated scenarios enable a precise assessment of the flexibility of the householders, e.g. by anticipating that two electricity-consuming appliances will potentially be turned on simultaneously. Since the framework is developed for short-term horizons and according to the exact current situations, this assessment is believed to be adaptable and highly reliable. Moreover,

the habits analysis for a large numbers of people and appliances exhibit that such habits can be clearly clustered by usage time and duration. This provides a way to build on an existing forecasting model, already tuned for a specific household, and add the additional demand required by a yet-unmeasured appliance. This would solve a common issue with statistical forecasting models and the need of a training period to tune their parameters.

Bibliography

Abreu, J. M., Pereira, F. C., & Ferrão, P. (2012). Using pattern recognition to identify habitual behavior in residential electricity consumption. *Energy and Buildings*, *49*, 479–487.

Ademe, E. (2008). ENERTECH : campagne de mesures des appareils de production de froid et des appareils de lavage dans 100 logements. *Projet AEE*. (In French)

Ahmad, A., Hassan, M., Abdullah, M., Rahman, H., Hussin, F., Abdullah, H., & Saidur, R. (2014). A review on applications of ANN and SVM for building electrical energy consumption forecasting. *Renewable and Sustainable Energy Reviews*, *33*, 102–109.

Alfares, H. K., & Nazeeruddin, M. (2002). Electric load forecasting: literature survey and classification of methods. *International Journal of Systems Science*, *33*(1), 23–34.

Andersen, F., Larsen, H., & Boomsma, T. (2013). Long-term forecasting of hourly electricity load: Identification of consumption profiles and segmentation of customers. *Energy Conversion and Management*, *68*, 244–252.

Andersen, F., Larsen, H., & Gaardestrup, R. (2013). Long term forecasting of hourly electricity consumption in local areas in Denmark. *Applied Energy*, *110*, 147–162.

Andersen, F., Larsen, H., Juul, N., & Gaardestrup, R. (2014). Differentiated long term projections of the hourly electricity consumption in local areas. the case of Denmark West. *Applied Energy*, *135*, 523–538.

Armel, K. C., Gupta, A., Shrimali, G., & Albert, A. (2013). Is disaggregation the holy grail of energy efficiency? The case of electricity. *Energy Policy*, *52*, 213–234.

Arora, S., & Taylor, J. W. (2016). Forecasting electricity smart meter data using conditional kernel density estimation. *Omega*, *59*, 47–59.

Bae, S., & Kwasinski, A. (2012). Spatial and temporal model of electric vehicle charging demand. *IEEE Transactions on Smart Grid*, *3*(1), 394–403.

Bahadur, R. R. (1966). A note on quantiles in large samples. *The Annals of Mathematical Statistics*, *37*(3), 577–580.

Barbier, T. (2017). *Modélisation de la consommation électrique à partir de grandes masses de données pour la simulation des alternatives énergétiques du futur* (Doctoral dissertation). PSL Research University. (In French)

Barker, S. K., Mishra, A. K., Irwin, D. E., Shenoy, P. J., & Albrecht, J. R. (2012). Smartcap: Flattening peak electricity demand in smart homes. In *Percom* (pp. 67–75).

Bassamzadeh, N., & Ghanem, R. (2017). Multiscale stochastic prediction of electricity demand in smart grids using Bayesian networks. *Applied Energy*, *193*, 369–380.

Beckel, C., Kleiminger, W., Cicchetti, R., Staake, T., & Santini, S. (2014). 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* (pp. 80–89).

Beckel, C., Sadamori, L., Staake, T., & Santini, S. (2014). Revealing household characteristics from smart meter data. *Energy*, *78*, 397–410.

Ben Taieb, S., & Hyndman, R. J. (2014). A gradient boosting approach to the kaggle load forecasting competition. *International journal of forecasting*, *30*(2), 382–394.

Ben Taieb, S. B., Taylor, J. W., & Hyndman, R. J. (2017). *Hierarchical probabilistic forecasting of electricity demand with smart meter data.*

Bennett, C. J., Stewart, R. A., & Lu, J. W. (2014). Forecasting low voltage distribution network demand profiles using a pattern recognition based expert system. *Energy*, *67*, 200–212.

Bessec, M., & Fouquau, J. (2008). The non-linear link between electricity consumption and temperature in Europe: a threshold panel approach. *Energy Economics*, *30*(5), 2705–2721.

Bickel, J. E. (2007). Some comparisons among quadratic, spherical, and logarithmic scoring rules. *Decision Analysis*, *4*(2), 49–65.

Bina, M. T., & Ahmadi, D. (2015). Stochastic modeling for the next day domestic demand response applications. *IEEE Transactions On Power Systems*, *30*(6),

2880–2893.

Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, *67*(338), 364–366.

Boroojeni, K. G., Amini, M. H., Bahrami, S., Iyengar, S., Sarwat, A. I., & Karabasoglu, O. (2017). A novel multi-time-scale modeling for electric power demand forecasting: From short-term to medium-term horizon. *Electric Power Systems Research*, *142*, 58–73.

Boroojeni, K. G., Mokhtari, S., Amini, M. H., & Iyengar, S. S. (2015). Optimal two-tier forecasting power generation model in smart grids. *CoRR*, *abs/1502.00530*.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, *3*(1), 1–122.

Brockwell, P. J., & Davis, R. A. (2013). *Time series: theory and methods*. Springer Science & Business Media.

Bruninx, K., & Delarue, E. (2016). Scenario reduction techniques and solution stability for stochastic unit commitment problems. In *Energy Conference (ENERGYCON), 2016 IEEE International* (pp. 1–7).

Bub, J. (2010). Von Neumann's 'no hidden variables' proof: a re-appraisal. *Foundations of Physics*, *40*(9-10), 1333–1340.

Buizza, R. (2014, November). *The TIGGE global, medium-range ensembles* (Technical Memorandum No. 739). ECMWF.

Candille, G., & Talagrand, O. (2005). Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society*, *131*(609), 2131–2150.

Cardoso, J.-F. (1998). Blind signal separation: statistical principles. *Proceedings of the IEEE*, *86*(10), 2009–2025.

Chacón, J. E., & Duong, T. (2010). Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *Test*, *19*(2), 375–398.

Charlton, N., & Singleton, C. (2014). A refined parametric model for short term load forecasting. *International Journal of Forecasting*, *30*(2), 364–368.

Chernozhukov, V., Fernández-Val, I., & Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, *78*(3), 1093–1125.

Chicco, G., Napoli, R., & Piglione, F. (2006, May). Comparisons among clustering

techniques for electricity customer classification. *IEEE Transactions on Power Systems*, *21*(2), 933-940.

Chu, Y., & Coimbra, C. F. (2017). Short-term probabilistic forecasts for direct normal irradiance. *Renewable Energy*, *101*, 526–536.

Clements, A. E., Hurn, A., & Li, Z. (2016). Forecasting day-ahead electricity load using a multiple equation time series approach. *European Journal of Operational Research*, *251*(2), 522–530.

Correa-Florez, C. A., Gerossier, A., Michiorri, A., & Kariniotakis, G. (2018). Stochastic operation of home energy management systems including battery cycling. *Applied Energy*, *225*, 1205–1218.

Cour des Comptes. (2018). *Les compteurs communicants Linky : tirer pour les consommateurs tous les bénéfices d'un investissement coûteux.* (In French)

D'Agostino SR, R. B., & Russell, H. K. (2005). Multivariate Bartlett test. *Encyclopedia of Biostatistics*, *5*.

Dent, I., Aickelin, U., & Rodden, T. (2013). Application of a clustering framework to UK domestic electricity data. *arXiv preprint arXiv:1307.1079*.

Derrac, J., García, S., Molina, D., & Herrera, F. (2011). A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation*, *1*(1), 3–18.

Dickert, J., & Schegner, P. (2010). Residential load models for network planning purposes. In *Modern Electric Power Systems (MEPS), 2010 Proceedings of the International Symposium* (pp. 1–6).

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*, 253–263.

Ding, N. (2012). *Load models for operation and planning of electricity distribution networks with metering data* (Doctoral dissertation). Université de Grenoble.

Dordonnat, V., Koopman, S. J., Ooms, M., Dessertaine, A., & Collet, J. (2008). An hourly periodic state space model for modelling French national electricity load. *International Journal of Forecasting*, *24*(4), 566–587.

Dryar, H. A. (1944). The effect of weather on the system load. *Electrical Engineering*, *63*(12), 1006–1013.

Duarte, C., Van Den Wymelenberg, K., & Rieger, C. (2013). Revealing occupancy

patterns in an office building through the use of occupancy sensor data. *Energy and Buildings*, *67*, 587–595.

Duong, T., Goud, B., & Schauer, K. (2012). Closed-form density-based framework for automatic detection of cellular morphology changes. *Proceedings of the National Academy of Sciences*, *109*(22), 8382–8387.

Dupačová, J., Gröwe-Kuska, N., & Römisch, W. (2003). Scenario reduction in stochastic programming. *Mathematical Programming*, *95*(3), 493–511.

Dvorkin, Y., Wang, Y., Pandzic, H., & Kirschen, D. (2014). Comparison of scenario reduction techniques for the stochastic unit commitment. In *PES General Meeting – Conference & Exposition, 2014 IEEE* (pp. 1–5).

Ehm, W., Gneiting, T., Jordan, A., & Krüger, F. (2016). Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *78*(3), 505–562.

Electric Reliability Council of Texas (ERCOT). (2018a). *ERCOT: hourly load data archives.* http://www.ercot.com/mktinfo/prices. (Online; accessed 24-August-2018; website blocked in France)

Electric Reliability Council of Texas (ERCOT). (2018b). *Market Prices: day-ahead market prices.* http://www.ercot.com/mktinfo/prices. (Online; accessed 20-November-2018; website blocked in France)

European Commission. (2014). *Benchmarking Smart Metering Deployment in the EU-27 with a Focus on Electricity.* Brussels.

European Parliament. (2009). Directive 2009/72/EC concerning common rules for the internal market in electricity. *Official Journal of the European Union*.

European Parliamentary Research Service. (2015). *Smart Electricity Grids and Meters in the EU Member States.* Brussels: European Parliament.

Fan, G.-F., Peng, L.-L., & Hong, W.-C. (2018). Short term load forecasting based on phase space reconstruction algorithm and bi-square kernel regression model. *Applied Energy*, *224*, 13–33.

Florez, C. A. C., Gerossier, A., Michiorri, A., Girard, R., & Kariniotakis, G. (2017). Residential electrical and thermal storage optimisation in a market environment. In *CIRED 2017-24th International Conference on Electricity Distribution.*

Friedman, J. (2002). Stochastic gradient boosting. *Computational Statistics & Data*

Analysis, *38*(4), 367–378.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The Elements of Statistical Learning* (Vol. 1). Springer, New York.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2000). Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, *28*(2), 337–407.

Gaillard, P., Goude, Y., & Nedellec, R. (2016). Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. *International Journal of forecasting*, *32*(3), 1038–1050.

Gajowniczek, K., & Ząbkowski, T. (2016). Short term electricity forecasting based on user behavior from individual smart meter data. *Journal of Intelligent & Fuzzy Systems*, *30*(1), 223–234.

Gajowniczek, K., & Ząbkowski, T. (2017). Electricity forecasting on the individual household level enhanced based on activity patterns. *PloS one*, *12*(4), e0174098.

Gerossier, A., Barbier, T., & Girard, R. (2017). A novel method for decomposing electricity feeder load into elementary profiles from customer information. *Applied Energy*, *203*, 752–760.

Gerossier, A., Girard, R., Bocquet, A., & Kariniotakis, G. (2018). Robust day-ahead forecasting of household electricity demand and operational challenges. *Energies*, *11*(12). Retrieved from http://www.mdpi.com/1996-1073/11/12/3503 doi: 10.3390/en11123503

Gerossier, A., Girard, R., & Kariniotakis, G. (2018). Modeling electric vehicle consumption profiles for short-term and long-term simulation. *MedPower*.

Gerossier, A., Girard, R., Kariniotakis, G., & Michiorri, A. (2017). Probabilistic day-ahead forecasting of household electricity demand. *CIRED-Open Access Proceedings Journal*, *2017*(1), 2500–2504.

Ghofrani, M., Hassanzadeh, M., Etezadi-Amoli, M., & Fadali, M. S. (2011). Smart meter based short-term load forecasting for residential customers. In *North American Power Symposium (NAPS), 2011* (pp. 1–5).

Giasemidis, G., Haben, S., Lee, T., Singleton, C., & Grindrod, P. (2017). A genetic algorithm approach for modelling low voltage network demands. *Applied Energy*, *203*, 463–473.

Glerum, A., Stankovikj, L., Thémans, M., & Bierlaire, M. (2013). Forecasting the demand for electric vehicles: accounting for attitudes and perceptions. *Trans-*

portation Science, *48*(4), 483–499.

Gneiting, T. (2011). Quantiles as optimal point forecasts. *International Journal of Forecasting*, *27*(2), 197–207.

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(2), 243–268.

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, *1*, 125–151.

Gneiting, T., & Ranjan, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, *29*(3), 411–422.

Goude, Y., Nedellec, R., & Kong, N. (2014). Local short and middle term electricity load forecasting with semi-parametric additive models. *IEEE Transactions on Smart Grid*, *5*(1), 440–446.

Gough, R., Dickerson, C., Rowley, P., & Walsh, C. (2017). Vehicle-to-grid feasibility: A techno-economic analysis of EV-based energy storage. *Applied Energy*, *192*, 12–23.

Gram-Hanssen, K. (2010). Standby consumption in households analyzed with a practice theory approach. *Journal of Industrial Ecology*, *14*(1), 150–165.

Gram-Hanssen, K. (2014). New needs for better understanding of household's energy consumption–behaviour, lifestyle or practices? *Architectural Engineering and Design Management*, *10*(1-2), 91–107.

Groeneveld, R. A., & Meeden, G. (1977). The mode, median, and mean inequality. *The American Statistician*, *31*(3), 120–121.

Grover-Silva, E., Girard, R., & Kariniotakis, G. (2018). Optimal sizing and placement of distribution grid connected battery systems through an socp optimal power flow algorithm. *Applied Energy*, *219*, 385–393.

Grover-Silva, E., Heleno, M., Mashayekh, S., Cardoso, G., Girard, R., & Kariniotakis, G. (2018). A stochastic optimal power flow for scheduling flexible resources in microgrids operation. *Applied energy*, *229*, 201–208.

Haben, S., Giasemidis, G., Ziel, F., & Arora, S. (2018). Short term load forecasts of low voltage demand and the effects of weather. *arXiv:1804.02955*.

Haben, S., Singleton, C., & Grindrod, P. (2016). Analysis and clustering of residential

customers energy behavioral demand using smart meter data. *IEEE Transactions on Smart Grid*, *7*(1), 136–144.

Haben, S., Ward, J., Greetham, D. V., Singleton, C., & Grindrod, P. (2014). A new error measure for forecasts of household-level, high resolution electrical energy consumption. *International Journal of Forecasting*, *30*(2), 246–256.

He, B. S., Yang, H., & Wang, S. L. (2000). Alternating direction method with self-adaptive penalty parameters for monotone variational inequalities. *Journal of Optimization Theory and applications*, *106*(2), 337–356.

He, Y., Liu, R., Li, H., Wang, S., & Lu, X. (2017). Short-term power load probability density forecasting method using kernel-based support vector quantile regression and copula theory. *Applied Energy*, *185*, 254–266.

Hendricks, W., & Koenker, R. (1992). Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American Statistical Association*, *87*(417), 58–68.

Hobbs, B. F., Jitprapaikulsarn, S., Konda, S., Chankong, V., Loparo, K. A., & Maratukulam, D. J. (1999). Analysis of the value for unit commitment of improved load forecasts. *IEEE Transactions on Power Systems*, *14*(4), 1342–1348.

Hong, T., & Fan, S. (2016). Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, *32*(3), 914–938.

Hong, T., Pinson, P., & Fan, S. (2014). *Global Energy Forecasting Competition 2012.* Elsevier.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., & Hyndman, R. J. (2016). *Probabilistic Energy Forecasting: Global Energy Forecasting Competition 2014 and Beyond.* Elsevier.

Hsiao, Y.-H. (2015). Household electricity demand forecast based on context information and user daily schedule analysis from meter data. *IEEE Transactions on Industrial Informatics*, *11*(1), 33–43.

Humeau, S., Wijaya, T. K., Vasirani, M., & Aberer, K. (2013). Electricity load forecasting for residential customers: Exploiting aggregation and correlation between households. In *Sustainable Internet and ICT for Sustainability (SustainIT), 2013* (pp. 1–6).

Insull, S. (1914). Centralization of energy supply. *Central Station Electricity Service: Its Commercial Development and Economic Significance as Set Forth in*

the Public Addresses (1897-1914) of Samuel Insull, 475.

International Energy Agency. (2011). *Technology Roadmap – Smart Grids*. Paris: IEA Publishing.

International Energy Agency. (2015). *World Energy Outlook*. Paris: IEA Publishing.

International Energy Agency. (2016a). *CO_2 Emissions from Fuel Combustion Highlights 2016*. Paris: IEA Publishing.

International Energy Agency. (2016b). *Key World Energy Statistics*. Paris: IEA Publishing.

Javed, F., Arshad, N., Wallin, F., Vassileva, I., & Dahlquist, E. (2012). Forecasting for demand response in smart grids: An analysis on use of anthropologic and structural data and short term multiple loads forecasting. *Applied Energy*, *96*, 150–160.

Jin, M., Feng, W., Liu, P., Marnay, C., & Spanos, C. (2017). MOD-DR: Microgrid optimal dispatch with demand response. *Applied Energy*, *187*, 758–776.

Jordan, A., Krüger, F., & Lerch, S. (2017). Evaluating probabilistic forecasts with the R package scoringRules. *arXiv preprint arXiv:1709.04743*.

Kavousian, A., Rajagopal, R., & Fischer, M. (2013). Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior. *Energy*, *55*, 184–194.

Keil, C., & Craig, G. C. (2009). A displacement and amplitude score employing an optical flow technique. *Weather and Forecasting*, *24*(5), 1297–1308.

Kelly, J., & Knottenbelt, W. (2015). The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. *Scientific Data*, *2*, 150007.

Kleiminger, W., Beckel, C., Staake, T., & Santini, S. (2013). Occupancy detection from electricity consumption data. In *Proceedings of the 5th ACM Workshop on Embedded Systems for Energy-Efficient Buildings* (pp. 1–8).

Koenker, R. (2012). *quantreg: Quantile regression. R package version 4.79*. Department of Economics, University of Illinois.

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, 33–50.

Krakowski, V., Assoumou, E., Mazauric, V., & Maïzi, N. (2016). Reprint of feasible

path toward 40–100% renewable energy shares for power supply in France by 2050: A prospective analysis. *Applied Energy*, *184*, 1529–1550.

Kriegler, B., & Berk, R. (2010). Small area estimation of the homeless in Los Angeles: An application of cost-sensitive stochastic gradient boosting. *The Annals of Applied Statistics*, 1234–1255.

Kuhn, H. W., & Tucker, A. W. (2014). Nonlinear programming. In *Traces and Emergence of Nonlinear Programming* (pp. 247–258). Springer.

Laio, F., & Tamea, S. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences Discussions*, *11*(4), 1267–1277.

Laplace, P.-S. (1829). *Essai philosophique sur les probabilités.* H. Remy. (In French)

Le Ray, G., Larsen, E. M., & Pinson, P. (2018). Evaluating price-based demand response in practice—with application to the EcoGrid EU Experiment. *IEEE Transactions on Smart Grid*, *9*(3), 2304–2313.

Lefieux, V. (2007). *Modèles semi-paramétriques appliqués à la prévision des séries temporelles. Cas de la consommation d'électricité.* (Doctoral dissertation). Université Rennes 2. (In French)

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., & Gneiting, T. (2017). Forecaster's dilemma: extreme events and forecast evaluation. *Statistical Science*, *32*(1), 106–127.

Leutbecher, M., & Palmer, T. N. (2008). Ensemble forecasting. *Journal of Computational Physics*, *227*(7), 3515–3539.

Liu, B., Nowotarski, J., Hong, T., & Weron, R. (2017). Probabilistic load forecasting via quantile regression averaging on sister forecasts. *IEEE Transactions on Smart Grid*, *8*(2), 730–737.

Luis, L. M. (2018). *Framework for scenario generation and reduction in photovoltaic-integrated generation commitment* (Doctoral dissertation). The University of North Carolina at Charlotte.

Lusis, P., Khalilpour, K. R., Andrew, L., & Liebman, A. (2017). Short-term residential load forecasting: Impact of calendar effects and forecast granularity. *Applied Energy*, *205*, 654–669.

Luthander, R., Widén, J., Nilsson, D., & Palm, J. (2015). Photovoltaic self-consumption in buildings: A review. *Applied Energy*, *142*, 80–94.

Malley, J. D., Kruppa, J., Dasgupta, A., Malley, K. G., & Ziegler, A. (2012). Probability machines. *Methods of Information in Medicine*, *51*(01), 74–81.

Mathiesen, B. V., Hansen, K., Ridjan, I., Lund, H., & Nielsen, S. (2015). Samsø energy vision 2030: Converting Samsø to 100% renewable energy.

Matthewman, P., & Nicholson, H. (1968). Techniques for load prediction in the electricity-supply industry. In *Proceedings of the Institution of Electrical Engineers* (Vol. 115, pp. 1451–1457).

McKenna, E., Richardson, I., & Thomson, M. (2012). Smart meter data: Balancing consumer privacy concerns with legitimate applications. *Energy Policy*, *41*, 807–814.

McLoughlin, F., Duffy, A., & Conlon, M. (2015). A clustering approach to domestic electricity load profile characterisation using smart metering data. *Applied Energy*, *141*, 190–199.

Misiti, M., Misiti, Y., Oppenheim, G., Poggi, J.-M., et al. (2010). Optimized clusters for disaggregated electricity load forecasting. *Revstat*, *8*(2), 105–124.

Mocanu, E., Nguyen, P. H., Gibescu, M., & Kling, W. L. (2016). Deep learning for estimating building energy consumption. *Sustainable Energy, Grids and Networks*, *6*, 91–99.

Molderink, A., Bakker, V., Bosman, M. G., Hurink, J. L., & Smit, G. J. (2010). A three-step methodology to improve domestic energy efficiency. In *Innovative Smart Grid Technologies (ISGT), 2010* (pp. 1–8).

Murphy, A. H. (1977). The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Monthly Weather Review*, *105*(7), 803–816.

Murphy, A. H. (1993). What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, *8*(2), 281–293.

Mutanen, A., Ruska, M., Repo, S., & Jarventausta, P. (2011, July). Customer classification and load profiling method for distribution systems. *IEEE Transactions on Power Delivery*, *26*(3), 1755–1763.

Nejat, P., Jomehzadeh, F., Taheri, M. M., Gohari, M., & Majid, M. Z. A. (2015). A global review of energy consumption, CO_2 emissions and policy in the residential sector (with an overview of the top ten CO_2 emitting countries). *Renewable and Sustainable Energy Reviews*, *43*, 843–862.

Nordic Energy Regulators. (2014). *Recommendations on Common Nordic Metering*

Methods.

Paatero, J. V., & Lund, P. D. (2006). A model for generating household electricity load profiles. *International Journal of Energy Research*, *30*(5), 273–290.

Papadimitriou, C. H., & Steiglitz, K. (1998). *Combinatorial Optimization: Algorithms and Complexity*. Courier Corporation.

Parson, O. (2014). *Unsupervised training methods for non-intrusive appliance load monitoring from smart meter data* (Doctoral dissertation). University of Southampton.

Pecan Street Inc. Dataport. (2018). https://dataport.cloud/. (Online; accessed 27-April-2018)

Pinson, P., Chevallier, C., & Kariniotakis, G. N. (2007). Trading wind generation from short-term probabilistic forecasts of wind power. *IEEE Transactions on Power Systems*, *22*(3), 1148–1156.

Pinson, P., & Girard, R. (2012). Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy*, *96*, 12–20.

Pinson, P., McSharry, P., & Madsen, H. (2010). Reliability diagrams for non-parametric density forecasts of continuous variables: Accounting for serial correlation. *Quarterly Journal of the Royal Meteorological Society: A Journal of the Atmospheric Sciences, Applied Meteorology and Physical Oceanography*, *136*(646), 77–90.

Pinson, P., Nielsen, H. A., Møller, J. K., Madsen, H., & Kariniotakis, G. N. (2007). Non-parametric probabilistic forecasts of wind power: required properties and evaluation. *Wind Energy*, *10*(6), 497–516.

Plackett, R. L. (1972). The discovery of the method of least squares. *Biometrika*, *59*(2), 239–251.

Poggi, J.-M. (1994). Prévision non paramétrique de la consommation électrique. *Revue de Statistique Appliquée*, *42*(4), 83–98. (In French)

Ponoćko, J., & Milanovic, J. V. (2018). Forecasting demand flexibility of aggregated residential load using smart meter data. *IEEE Transactions on Power Systems*.

Quilumba, F. L., Lee, W.-J., Huang, H., Wang, D. Y., & Szabados, R. L. (2015). Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities. *IEEE Transactions on Smart Grid*, *6*(2), 911–918.

Regnier, E. (2008). Doing something about the weather. *Omega*, *36*(1), 22–32.

Rhodes, J. D., Cole, W. J., Upshaw, C. R., Edgar, T. F., & Webber, M. E. (2014). Clustering analysis of residential electricity demand profiles. *Applied Energy*, *135*, 461–471.

Richardson, I., Thomson, M., Infield, D., & Clifford, C. (2010). Domestic electricity use: A high-resolution energy demand model. *Energy and buildings*, *42*(10), 1878–1887.

Ridgeway, G. (2017). gbm: Generalized boosted regression models [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=gbm (R package version 2.1.3)

Rodrigues, F., Cardeira, C., & Calado, J. M. F. (2014). The daily and hourly energy consumption and load forecasting using artificial neural network method: a case study using a set of 93 households in Portugal. *Energy Procedia*, *62*, 220–229.

Rodriguez, G. (2001). *Smoothing and Non-Parametric Regression.*

Rose, T. (2016). *The End of Average: How to succeed in a world that values sameness.* Penguin UK.

Réseau de Transport d'Électricité (RTE). (2018a). *Eco2mix.* https://www.rte-france.com/fr/eco2mix/eco2mix-telechargement. (Online; accessed 05-August-2018; in French)

Réseau de Transport d'Électricité (RTE). (2018b). *Panorama de l'électricité renouvelable au 31 mars 2018.* (In French)

Räsänen, T., Voukantsis, D., Niska, H., Karatzas, K., & Kolehmainen, M. (2010). Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. *Applied Energy*, *87*(11), 3538–3545.

Rydén, T. (1997). On recursive estimation for hidden Markov models. *Stochastic Processes and their Applications*, *66*(1), 79–96.

Sancho-Tomás, A., Sumner, M., & Robinson, D. (2017). A generalised model of electrical energy demand from small household appliances. *Energy and Buildings*, *135*, 350–366.

Schnabel, S. K., & Eilers, P. H. (2013). Simultaneous estimation of quantile curves using quantile sheets. *AStA Advances in Statistical Analysis*, *97*(1), 77–87.

Scholz, F. W., & Stephens, M. A. (1987). K-sample anderson–darling tests. *Journal of the American Statistical Association*, *82*(399), 918–924.

SENSIBLE. (2018). *Évora demonstrator site.* https://www.projectsensible.eu/demonstrators/evora/. (Online; accessed 14-December-2018)

Seppälä, A. (1996). *Load Research and Load Estimation in Electricity Distribution* (Theses). Technical research center of Finland, VTT Publications.

Sevlian, R., & Rajagopal, R. (2014). Short term electricity load forecasting on varying levels of aggregation. *arXiv:1404.0058*.

Shao, Z., Chao, F., Yang, S.-L., & Zhou, K.-L. (2017). A review of the decomposition methodology for extracting and identifying the fluctuation characteristics in electricity demand forecasting. *Renewable and Sustainable Energy Reviews*, *75*, 123–136.

Shao, Z., Gao, F., Zhang, Q., & Yang, S.-L. (2015). Multivariate statistical and similarity measure based semiparametric modeling of the probability distribution: A novel approach to the case study of mid-long term electricity consumption forecasting in China. *Applied Energy*, *156*, 502–518.

Shepero, M., van der Meer, D., Munkhammar, J., & Widén, J. (2018). Residential probabilistic load forecasting: A method using gaussian process designed for electric load data. *Applied Energy*, *218*, 159–172.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis* (Vol. 26). CRC Press.

Srinivasan, D., & Lee, M. (1995). Survey of hybrid fuzzy neural approaches to electric load forecasting. In *Systems, Man and Cybernetics, 1995. Intelligent Systems for the 21st Century., IEEE International Conference* (Vol. 5, pp. 4004–4008).

Statkraft. (2017). *The virtual power plant and how it works.* https://www.statkraft.co.uk/power-purchase-agreements/virtual-power-plant/. (Online; accessed 18-December-2018)

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900.* Harvard University Press.

Stokes, M. (2005). Removing barriers to embedded generation: a fine-grained load model to support low voltage network performance analysis.

Tanimoto, J., Hagishima, A., & Sagara, H. (2008). Validation of probabilistic methodology for generating actual inhabitants' behavior schedules for accurate prediction of maximum energy requirements. *Energy and Buildings*, *40*(3), 316–322.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and

review. *International Journal of Forecasting*, *16*(4), 437–450.

Taylor, J. W., & McSharry, P. E. (2007). Short-term load forecasting methods: An evaluation based on European data. *IEEE Transactions on Power Systems*, *22*(4), 2213–2219.

Taylor, J. W., & Snyder, R. D. (2012). Forecasting intraday time series with multiple seasonal cycles using parsimonious seasonal exponential smoothing. *Omega*, *40*(6), 748–757.

Tidemann, A., Høverstad, B. A., Langseth, H., & Öztürk, P. (2013). Effects of scale on load prediction algorithms.

Tikhonov, A. N. (1943). Об устойчивости обратных задач [On the stability of inverse problems]. In *Doklady Akademii Nauk SSSR* (Vol. 39, pp. 195–198). (in Russian)

Tomić, J., & Kempton, W. (2007). Using fleets of electric-drive vehicles for grid support. *Journal of Power Sources*, *168*(2), 459–468.

Viegas, J. L., Vieira, S. M., Melício, R., Mendes, V., & Sousa, J. M. (2016). Classification of new electricity customers based on surveys and smart metering data. *Energy*, *107*, 804–817.

Wahba, G. (1990). *Spline Models for Observational Data* (Vol. 59). Siam.

Wang, P., Liu, B., & Hong, T. (2016). Electric load forecasting with recency effect: A big data approach. *International Journal of Forecasting*, *32*(3), 585–597.

Wang, Y., Gan, D., Sun, M., Zhang, N., Lu, Z., & Kang, C. (2019). Probabilistic individual load forecasting using pinball loss guided lstm. *Applied Energy*, *235*, 10–20.

Widén, J., & Wäckelgård, E. (2010). A high-resolution stochastic model of domestic activity patterns and electricity demand. *Applied Energy*, *87*(6), 1880–1892.

Wijaya, T. K., Humeau, S., Vasirani, M., & Aberer, K. (2014). *Residential electricity load forecasting: evaluation of individual and aggregate forecasts.* Citeseer.

Wijaya, T. K., Vasirani, M., Humeau, S., & Aberer, K. (2015). Cluster-based aggregate forecasting for residential electricity demand using smart meter data. In *2015 IEEE International Conference on Big Data* (pp. 879–887).

Wolde-Rufael, Y. (2006). Electricity consumption and economic growth: a time series experience for 17 African countries. *Energy Policy*, *34*(10), 1106–1114.

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1),

1–17.

Xie, J., & Hong, T. (2016). GEFCom2014 probabilistic electric load forecasting: An integrated solution with forecast combination and residual simulation. *International Journal of Forecasting*, *32*(3), 1012–1016.

Xu, S., & Miao, Y. (2011). Limit behaviors of the deviation between the sample quantiles and the quantile. *Filomat*, *25*(2), 197–206.

Yildiz, B., Bilbao, J., Dore, J., & Sproul, A. (2017). Recent advances in the analysis of residential electricity consumption and applications of smart meter data. *Applied Energy*, *208*, 402–427.

Yu, C.-N., Mirowski, P., & Ho, T. K. (2017). A sparse coding approach to household electricity demand forecasting in smart grids. *IEEE Transactions on Smart Grid*, *8*(2), 738–748.

Zhou, K., lin Yang, S., & Shen, C. (2013). A review of electric load classification in smart grid environment. *Renewable and Sustainable Energy Reviews*, *24*, 103–110.

Appendices

Appendix A

Quantile estimation

Let X be a univariate real random variable. Under reasonable and common assumptions, there exist its cumulative distribution function F (CDF), and its probability density function f (PDF).

A.1 Quantile definition

A quantile value of order τ, i.e. at quantile level τ, of the random variable X is a real value x^τ such that $\mathbb{P}[X \leq x^\tau] = \tau$. Quantiles can be obtained from the inverse of the CDF $x^\tau = F^{-1}(\tau)$. When F is left-discontinuous, quantiles are not unique and can be any value on the interval where F is discontinuous. Whence this issue occur in practical applications, linear interpolation is made to obtain a unique quantile value. When F is constant on an interval, a single value can be the quantile at multiple levels.

A.2 A natural estimator

When CDF F is unknown, an estimator of the quantile is needed. Let $x_1, ..., x_n$ be independent samples of distribution X. There exists a permutation of these samples such that $x_{(1)}, ..., x_{(n)}$ are in an increasing order. A natural estimator of the quantile of order τ is

$$\hat{x}_n^\tau = x_{(\lfloor n\tau \rfloor)}, \tag{A.1}$$

where $\lfloor t \rfloor$ denotes the floor function of t.

Let us show that this estimator converges to the quantile value of order τ by using the empirical cumulative distribution function $\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i \leq t)$ to write an upper bound of the difference,

$$|F(\hat{x}_n^\tau) - F(x^\tau)| \leq |F(\hat{x}_n^\tau) - \hat{F}_n(\hat{x}_n^\tau)| + |\hat{F}_n(\hat{x}_n^\tau) - F(x^\tau)| \tag{A.2}$$

$$\leq \sup|F - \hat{F}_n| + \left| \frac{\lfloor n\tau \rfloor + 1}{n} - \tau \right|. \tag{A.3}$$

The second term clearly goes to 0 when n grows, and the first term converges to 0 from Glivenko–Cantelli theorem. Therefore, when F is continuous in x^τ,

$$\hat{x}_n^\tau \xrightarrow[n\to\infty]{} x^\tau, \qquad \text{almost surely.} \tag{A.4}$$

A.3 Convergence rate

Supposing that PDF f is differentiable and strictly positive in a neighborhood around x^τ, Bahadur (Bahadur, 1966) proves that

$$\hat{x}_n^\tau = x^\tau + \frac{\tau - \hat{F}_n(x^\tau)}{f(x^\tau)} + \mathcal{O}\left(n^{-3/4}\left(\log n\right)^{3/4}\right). \tag{A.5}$$

From this equation and the Linderberg central limit theorem, applied on the empirical cumulative distribution function, we have the following distribution convergence

$$\hat{x}_n^\tau \xrightarrow{d} \mathcal{N}\left(0, \frac{\tau(1-\tau)}{n f^2(x^\tau)}\right). \tag{A.6}$$

Further results on the rate of convergence have been established, e.g. see (Xu & Miao, 2011).

The distribution convergence provides an asymptotic interval of order $1-\alpha$, with a_α the quantile of order $1-\alpha/2$ of the standard law $\mathcal{N}(0,1)$,

$$x^\tau \in \mathcal{I}_n^\alpha = \left[\hat{x}_n^\tau \pm a_\alpha \frac{\sqrt{\tau(1-\tau)}}{\sqrt{n} f(\hat{x}_n^\tau)}\right]. \tag{A.7}$$

When density f is unknown, one may prefer to use two samples of the order statistics to obtain a confidence interval. For n large enough, the two integers

$$i_n = \lfloor n\tau - a_\alpha \sqrt{n\tau(1-\tau)} \rfloor \tag{A.8}$$

$$j_n = \lfloor n\tau + a_\alpha \sqrt{n\tau(1-\tau)} \rfloor \tag{A.9}$$

are between 1 and n. It provides the following asymptotic interval of order $1-\alpha$

$$\mathcal{J}_n^{\alpha} = \left[x_{(i_n)}, x_{(j_n)}\right]. \tag{A.10}$$

In summary, the rate of convergence is in $n^{-1/2}$ and is proportional to $\dfrac{\sqrt{\tau(1-\tau)}}{f(F^{-1}(\tau))}$.

Appendix B

On the Equality Between CRPS and QS

B.1 Problem

We want to prove that the *Continuous Ranked Probability Score* (CRPS) is equal to the integral of the *Quantile Score* (QS_τ over all quantile levels $\tau \in (0, 1)$), i.e.

$$\mathrm{CRPS} = \int_0^\infty \mathrm{QS}_\tau \mathrm{d}\tau. \tag{B.1}$$

The CRPS assesses the proximity between a cumulative density function F and a real value y

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{+\infty} (\mathbb{1}(z \geq y) - F(z))^2 \mathrm{d}z, \tag{B.2}$$

where $\mathbb{1}(E)$ is the indicator function equal to 1 if statement E is true, or 0 if E is false. The QS assesses the quality of a quantile value y^τ at level τ compared to a real value y

$$\mathrm{QS}_\tau(y^\tau, y) = 2(\mathbb{1}(y \leq y^\tau) - \tau)(y^\tau - y) \tag{B.3}$$

Let us note that factor 2 is sometimes omitted in the literature. However, with this factor, we conveniently have an equality with the MAE, namely $\mathrm{QS}_{0.5} = \mathrm{MAE}$.

B.2 Proof

The proof is straightforward and involves an integration by parts and a variable change under the integral.

To avoid dealing with the discontinuous function $\mathbb{1}(\cdot)$, we separate the CRPS integral in 2 terms:

$$\mathrm{CRPS}(F, y) = \underbrace{\int_{-\infty}^{y} F^2(z)\mathrm{d}z)}_{(A)} + \underbrace{\int_{y}^{+\infty} (1 - F(z))^2\mathrm{d}z)}_{(B)}. \tag{B.4}$$

We first transform term (A). An integration by parts is made using $u(z) = F^2(z)$ and $v'(z) = 1$. So that $u'(z) = 2f(z)F(z)$, and a practical integrand is $v(z) = z - y$. Therefore

$$(A) = \left[F^2(z)(z-y)\right]_{-\infty}^{y} - \int_{-\infty}^{y} 2(z-y)f(z)F(z)\mathrm{d}z. \tag{B.5}$$

The first term is null for most distributions[1]. The second term is an integral for which we introduce $\tau = F(z)$. Thus, $\mathrm{d}\tau = f(z)\mathrm{d}z$, and $z = F^{-1}(\tau) = y^\tau$. The new integral goes from 0 to $F(y)$, so

$$(A) = -\int_{0}^{F(y)} 2\tau(y^\tau - y)\mathrm{d}\tau. \tag{B.6}$$

Similar transformations lead to

$$(B) = \int_{F(y)}^{1} 2(1-\tau)(y^\tau - y)\mathrm{d}\tau. \tag{B.7}$$

Thus, the sum of the two terms

$$\mathrm{CRPS}(F, y) = \int_{0}^{F(y)} 2(-\tau)(y^\tau - y)\mathrm{d}\tau + \int_{F(y)}^{1} 2(1-\tau)(y^\tau - y)\mathrm{d}\tau. \tag{B.8}$$

The writing is simplified by using the indicator function $\mathbb{1}(\tau \geq F(y))$, which is equal to $\mathbb{1}(y^\tau \geq y)$. So

$$\mathrm{CRPS}(F, y) = \int_{0}^{1} 2(\mathbb{1}(y^\tau \geq y) - \tau)(y^\tau - y)\mathrm{d}\tau, \tag{B.9}$$

and we find equation (B.1). □

[1] when the first term is not null, it exactly compensates with a similar term in (B).

Appendix C

Demand Disaggregation Algorithm

C.1 Problem

A total of F feeders ($F \approx 1000$) deliver electricity to multiple individual consumers. For each feeder $f \in \{1, \ldots, F\}$, electricity load is recorded every 10 minutes during a given period, say one week, so to define a time index $t \in \{1, \ldots, T\}$ with $T = 1008$. In practice, two transforms are applied to the load measurements made in kW or kWh: (1) removal of the thermal effect, and (2) normalization by average feeder demand during the period of length T. These transforms are necessary to obtain dimensionless value of unit average so as to compare the demand across feeders, see Section 3.1.4. Each feeder is associated to a mix of K consumer categories. For instance, with $K = 2$, one feeder has a mix: 80% share of residential consumers and 20% share of tertiary consumers. We suppose that the each feeder demand is entirely composed by elementary category demand profile, and so each feeder demand is obtained just with the category mix of the feeder. Since the exact mixes vary between feeders, it is possible to separate electricity of one category from the others using the measurements of the F feeders.

The goal is therefore to find the profile of electricity demand for each one of the K categories at instant t, noted $d_k(t)$. Defining $p_1^f, \ldots, p_K^f$ as the category shares of one feeder f (they sum to 1) then the demand $d^f(t)$ at time t is supposed to be the sum of K values, independent of the feeder:

$$d^f(t) = \sum_{k=1}^{K} p_k^f d_k(t). \tag{C.1}$$

C.2 Disaggregation Algorithm

C.2.1 Unknown Matrix

We want to find the unknown matrix B with demands for every category and instant

$$B = \begin{pmatrix} d_1(1) & \dots & d_1(T) \\ \vdots & \ddots & \vdots \\ d_K(1) & \dots & d_K(T) \end{pmatrix}.$$

It is also convenient to define the associated column vector β stacking columns on top of each other

$$\beta = \begin{pmatrix} d_1(1) \\ \vdots \\ d_1(T) \\ \vdots \\ d_K(T) \end{pmatrix}.$$

C.2.2 Data Matrices

We define matrix X containing the dimensionless demand values of every feeder at every time step of the week

$$X = \begin{pmatrix} d^1(1) & \dots & d^1(T) \\ \vdots & \ddots & \vdots \\ d^F(1) & \dots & d^F(T) \end{pmatrix}.$$

Column vector x is associated

$$x = \begin{pmatrix} d^1(1) \\ \vdots \\ d^1(T) \\ \vdots \\ d^F(T) \end{pmatrix}.$$

We define the proportion matrix A containing the category mixes of every feeder, i.e.

$$A = \begin{pmatrix} p_1^1 & \dots & p_K^1 \\ \vdots & \ddots & \vdots \\ p_1^F & \dots & p_K^F \end{pmatrix}.$$

Each row contains positive elements summing to 1.

Matrix Y is defined as

$$Y = A \otimes I_T \tag{C.2}$$

where $\otimes$ denotes the Kronecker product and I_n the identity matrix of size n. Let us note that Y is a matrix of size (FT, KT), which can be huge, and it is better not to compute it.

We see that the square distance between our unknown variable β and the data x can be written $\|x - Y\beta\|^2$. However the problem cannot be resolved just by minimizing this function of β:

— To have a physical interpretation, every categorical demand value in β should be positive.

— Each row of B, i.e. each weekly categorical profile, should have a unit mean. We use column vectors $v = (T^{-1}, \dots, T^{-1})^\intercal$ of size T, and $u = (1, \dots, 1)^\intercal$ of size K, in order to write the constraint $(I_K \otimes v^\intercal)\beta = u$.

C.2.3 Optimization problem

The optimization problem thus writes

$$\begin{aligned} \min_{\beta} \quad & \|x - Y\beta\|^2, \\ \text{s.t.} \quad & \beta \geq 0, \\ & (I_K \otimes v^\intercal)\beta = u. \end{aligned}$$

By defining $P = 2Y^\intercal Y$ and $q = -2Y^\intercal x$, problem becomes

$$\begin{aligned} \min_{\beta} \quad & \frac{1}{2}\beta^\intercal P \beta + q^\intercal \beta, \\ \text{s.t.} \quad & \beta \geq 0, \\ & (I_K \otimes v^\intercal)\beta = u. \end{aligned} \tag{C.3}$$

C.2.4 Alternating Direction Method of Multipliers

The problem (C.3) is exactly dealt with in (Boyd et al., 2011, Section 5.2). We consequently follow Boyd et al., and an alternating direction method of multipliers (ADMM)

is employed to solve the constrained problem. The positivity constraint is resolved by introducing α, the *dual* variable of β. The variables α and β ultimately converge to the same optimal point. A third variable γ, the *scaled* variable, assesses the closeness of α and β and balances this error compared to the error between β and the data.

Therefore, with a real $\rho > 0$, the algorithm writes at step l:

(i) $\beta^{l+1} = \arg\min_\beta \left(\frac{1}{2}\beta^\intercal P\beta + q^\intercal\beta + \frac{\rho}{2}\|\beta - \alpha^l + \gamma^l\|^2\right)$ with the constraint $(I_K \otimes v^\intercal)\beta = u$.

(ii) $\alpha^{l+1} = \left(\beta^{l+1} + \gamma^l\right)_+$, i.e. keeping only the positive values and assigning 0 to the others.

(iii) $\gamma^{l+1} = \gamma^l + \beta^{l+1} - \alpha^{l+1}$.

C.2.4.1 Details for Step (i)

Matrix System To carry out the optimization at each step, we use the Lagrangian function, with multiplier ν on the constraint $(I_K \otimes v^\intercal)\beta - u = 0$,

$$\mathcal{L}(\beta, \nu) = \frac{1}{2}\beta^\intercal P\beta + q^\intercal\beta + \frac{\rho}{2}\beta^\intercal\beta - \rho\beta^\intercal\left(\alpha^l - \gamma^l\right) + \nu^\intercal((I_K \otimes v^\intercal)\beta - u).$$

By differentiating with respect to β and ν, a known Karush-Kuhn-Tucker system is obtained (Kuhn & Tucker, 2014),

$$\begin{pmatrix} P + \rho I_{KT} & (I_K \otimes v) \\ (I_K \otimes v^\intercal) & 0 \end{pmatrix} \begin{pmatrix} \beta^{l+1} \\ \nu \end{pmatrix} = \begin{pmatrix} -q + \rho\left(\alpha^l - \gamma^l\right) \\ u \end{pmatrix}. \tag{C.4}$$

The leftmost multiplying matrix needs to be inverted. The size of this square matrix is $K(T+1)$ so computing its inverse can be computation-expensive. Using its simple shape, this matrix can however be efficiently inverted.

Matrix to Invert One has to compute the inverse of the square matrix of size $K(T+1)$

$$\begin{pmatrix} P + \rho I_{KT} & (I_K \otimes v) \\ (I_K \otimes v^\intercal) & 0 \end{pmatrix}.$$

Instead of a direct solving, advantageous writing can notably speed up the computation. The upper-left block is the most voluminous part of the matrix: a square matrix of side KT. Using the definition of P, it rewrites as a Kronecker product

$$P + \rho I_{KT} = (2A^\intercal A + \rho I_K) \otimes I_T.$$

Therefore, by defining $N = 2A^\intercal A + \rho I_K$, we write matrix

$$M = \begin{pmatrix} N \otimes I_T & I_K \otimes v \\ I_K \otimes v^\intercal & 0 \end{pmatrix}.$$

Block Inversion The inverse of block matrix M is also a block matrix

$$M^{-1} = \begin{pmatrix} M_1 & M_2 \\ M_3 & M_4 \end{pmatrix}.$$

By applying the Helmert-Wolf block formulae, and noting that $v^\intercal v = T^{-1}$, we obtain, starting by M_2 for convenience,

$$\begin{aligned}
M_2 &= \left(N^{-1} \otimes I_T\right) (I_K \otimes v) \left[(I_k \otimes v^\intercal) \left(N^{-1} \otimes I_T\right) (I_K \otimes v)\right]^{-1} \\
&= \left(N^{-1} \otimes I_T\right) (I_K \otimes v) \left(T(N \otimes I_T)\right) \\
&= T(I_K \otimes v), \\
M_1 &= N^{-1} \otimes I_T - M_2 \left(I_K \otimes v^\intercal\right) \left(N^{-1} \otimes I_T\right) \\
&= N^{-1} \otimes I_T - T \left(N^{-1} \otimes vv^\intercal\right) \\
&= N^{-1} \otimes \left(I_T - Tvv^\intercal\right), \\
M_3 &= M_2^\intercal \\
&= T(I_K \otimes v^\intercal), \\
M_4 &= -TN.
\end{aligned}$$

Therefore

$$M^{-1} = \begin{pmatrix} N^{-1} \otimes (I_T - Tvv^\intercal) & T(I_K \otimes v) \\ T(I_K \otimes v^\intercal) & -TN \end{pmatrix}.$$

Necessary Inversion At the end of the day, we only need to invert N, a matrix for size K. Moreover, since N is a symmetric definite matrix, the inversion is quick with a Cholesky decomposition.

Writing q As seen in equation (C.4), you need to compute $q = -2Y^\intercal x$. You do not want to explicitly do this matrix multiplication for memory reason. Instead, we define Q the matrix of size (K, T) associated to q obtained by taking the first T components of q and putting them in the first column, then the next T components of q and so on. We write

$$\begin{aligned} q &= -2Y^\intercal x \\ &= -2\left(A^\intercal \otimes I_T\right) x. \end{aligned}$$

But it can be seen that $\left(A^\intercal \otimes I_T\right) x$ is the same operation as $A^\intercal X$ although the first one gives the outcomes in an column vector while the second one gives a matrix. Therefore, we can compute $Q = -2A^\intercal X$ and reshape this matrix as a column vector to get q.

Summary To summarize, we compute q and N^{-1} at the initial step. Then, at each iteration l,

$$\begin{aligned} \beta^{l+1} =& \rho\left[N^{-1} \otimes \left(I_K - Tvv^\intercal\right)\right]\left(\alpha^l - \gamma^l\right) \\ & \underbrace{-\left[N^{-1} \otimes \left(I_K - Tvv^\intercal\right)\right] q + H\left[I_K \otimes v\right] u}_{\text{constant term}} \end{aligned} \tag{C.5}$$

is updated. A matrix multiplication $(KT, KT) \times (KT, 1)$ has to be computed at each iteration and added to a constant term to update β.

C.2.4.2 Convergence

Convergence of the algorithm is guaranteed by checking that both the primal residual $\rho\|\alpha^{l+1} - \alpha^l\|$ and secondary residual $\|\alpha^l - \beta^l\|$ are both below an arbitrarily fixed threshold of 10^{-6}. To ensure that both residuals similarly contribute to the overall errors, parameter ρ is adjusted following the iterative scheme proposed by He et al. (B. S. He et al., 2000, Strategy 3).

Appendix D

Detailed Forecast Performance

In this appendix, we provide detailed forecast performance about day-ahead forecasting performance on the three smart-meter datasets with the persistence, climatology, and GBM model introduced in Section 4.2.

D.1 Extended Results

The evaluation is made on the 3 datasets: Portugal (226 households), France (176 households), and USA (175 households). The following point/deterministic indices are reported: NBias, MAPE, NMAE, NRMSE; the following probabilistic indices are reported: NCRPS, NCRPS_{LT}, NCRPS_{UT}, NCRPS_{S}. These indices are defined in Section 2.2. The indices are computed on each hourly value of an household and the mean over the whole period of available measurements is taken as the forecasting performance for this household. Since the hourly value of MAPE is sometimes absurdly large — due to a division by a demand value close to 0 —, the median over the period is computed rather than the mean. All the indices are dimensionless and expressed in %. Once the indices are computed for each household of a dataset, the average value and standard deviation — between parentheses — are reported in Table D.1.

Table D.1 – Average forecast performance measured with various indices over multiple datasets and models.

Dataset	Model	NBias	MAPE	NMAE	NRMSE	NCRPS	NCRPS_{LT}	NCRPS_{UT}	NCRPS_{S}
Portugal	Persistence	0 (1)	25 (10)	33 (11)	47 (17)	—	—	—	—
	Climatology	+6 (9)	24 (10)	30 (11)	41 (17)	21 (7)	63 (13)	100 (48)	75 (25)
	Gradient Boosting	+3 (3)	19 (8)	25 (9)	35 (12)	17 (6)	57 (13)	86 (37)	63 (21)
France	Persistence	-1 (1)	25 (12)	37 (18)	60 (37)	—	—	—	—
	Climatology	+9 (13)	25 (11)	33 (13)	53 (31)	24 (10)	67 (15)	141 (110)	87 (38)
	Gradient Boosting	+7 (8)	18 (10)	25 (13)	43 (29)	18 (9)	55 (19)	124 (85)	68 (35)
USA	Persistence	0 (0)	29 (8)	46 (13)	79 (30)	—	—	—	—
	Climatology	+22 (11)	43 (9)	52 (9)	82 (23)	36 (6)	75 (8)	213 (69)	129 (24)
	Gradient Boosting	+10 (6)	23 (7)	33 (9)	59 (22)	24 (7)	66 (10)	164 (63)	89 (24)

D.2 Positive Bias

In most cases, the climatology and the gradient boosting models have a non-negligible positive bias. This is due to the asymmetrical electricity demand distribution which has longer upper tail than lower tail. This fact is concurred by the value of the lower tail and upper tail version of the CRPS: the former being roughly half of the latter. This fact indicates that the upper tail is about twice more difficult to forecast, because the demand distribution of the upper part is more spread. This asymmetry is physically logical: the highest demand values, i.e. peak demand, are relatively farther from the average demand than the lowest demand values, hence the positive skew of the demand distribution. This skew, along with the generally unimodal demand distribution, implies that the median of the distribution is lower than the mean (Groeneveld & Meeden, 1977). Since the deterministic forecasts of the GBM are based on the Laplace distance, i.e. the quantile score at quantile level 50%, the forecasts are optimized for the median value rather than the mean. Consequently, the mean of the bias is expected to be positive , i.e.

$$\begin{aligned} \mathbb{E}\left[y_t - \hat{y}_t^{0.5}\right] &= \operatorname{mean} y_t - \operatorname{median} y_t \\ &\geq 0 \end{aligned} \tag{D.1}$$

D.3 Equivalence Coefficients

From these detailed results, we observe a proportionality between the different indices — except the bias —, meaning that either one of the indices are often enough — and more convenient — to characterize the performance of a forecasting model. This comes from the reasonable design of the forecasting models generally proposed in the literature, that do not exploit the relative drawbacks of each indices. One may imagine probabilistic models resulting in very poor CRPS but good NMAE.

In any case, we find the following equivalence coefficients between the indices, computed with the gradient boosting model,

$$\mathrm{NMAE} = 1.3 \times \mathrm{MAPE}, \tag{D.2}$$

$$\mathrm{NMAE} = 0.6 \times \mathrm{NRMSE}, \tag{D.3}$$

$$\mathrm{NMAE} = 1.4 \times \mathrm{NCRPS}. \tag{D.4}$$

Note that the relations in Equation (D.3) and (D.4) are also valid for the non-normalized version, i.e. $\mathrm{MAE} = 0.6 \times \mathrm{RMSE}$ and $\mathrm{MAE} = 1.4 \times \mathrm{CRPS}$.

Appendix E

Tuning Gradient Boosting Model

E.1 Parameters Analysis

We thereafter present an iterative method to select the meta-parameters of the gradient boosting model detailed in Section 4.2.1: the number of trees $\tau_{\max}$, the interaction depth Δ, the shrinkage parameter λ, the minimal node size ν, and the subsampling rate p. We detail the key parameters and the sensitivity to the fine tuning of the meta-parameters. The performance tests are made with a randomly selected subset of 10 US households, that we train to forecast next-day hourly electricity demand values with 6 input variables, see the details in Section 4.2.2. The performance is assessed by using the NMAE, i.e. the $\text{NQS}_{0.5}$, index.

E.1.1 Number of Trees

The iterative structure of the gradient boosting model allows flexibility regarding the choice of the number of trees to stack. When $\tau_{\max}$ trees are computed, one can evaluate the model performance for any number of trees, i.e. for $1, 2, \ldots, \tau_{\max}$ trees. However, when the performance is assessed solely on the training set, the NMAE keeps on decreasing when the number of trees increases. Therefore, a cross-validation approach to assess performance is recommended by Ridgeway (Ridgeway, 2017). We select 5 folds and use, in turns, 1 fold as an out-of-sample data to evaluate performance. This cross-validation error is then assumed to correctly assess the performance of the model. With this approach, we logically observe that, the cross-validation error decreases up

to an optimal number of trees, noted $\tau^* \leq \tau_{\max}$, and eventually increases when the model ends up overfitting the data. Although there is no theoretical proof that a minimum is reached, we assume that there exists an optimal number of trees that one can reach by setting a large enough upper bound $\tau_{\max}$. Naturally, the greater $\tau_{\max}$ is, the more computation time it necessitates. The computation time to train the model is proportional to the value $\tau_{\max}$. Therefore, to minimize computation time while maximizing performance, one should aim for a value just larger than τ^*, since remaining trees, i.e. trees $\tau^* + 1, \ldots, \tau_{\max}$, are discarded. In the following, the same optimal performance is assigned for any number of trees greater than τ^*, meaning that the overfitting performance degradation is not visible.

E.1.2 Interaction Depth

The interaction depth Δ is used by the individual regression trees. A depth of 1 creates a tree that uses only 1 input variable, a depth of 2 only 2 variables, and so on. The value of Δ therefore determines the complexity of each weak learner, and hence the computation time. Table E.1 reports the computation time that our average laptop (2 chores at 2 GHz) needs to compute 2,000 trees regarding the interaction depth. The computation time for tree of depth 1 is taken as a reference to compute a time factor, that is independent of the computer used. We roughly observe that

$$\text{Time Factor} = 0.5 \times (\Delta + 1). \tag{E.1}$$

Table E.1 – Computation time for 2,000 trees

Depth Δ	CPU Time (s)	Time Factor
1	65	1
2	108	1.6
3	137	2.1
4	171	2.6
5	197	3.0
6	231	3.5

We then fix other parameters $\tau_{\max} = 2,000$, $\lambda = 0.05$, $\nu = 10$, and $p = 0.5$ and

compute the cross-validation error (NMAE) for our subset of 10 US households for depth going from 1 to 6 — which is the total number of input variables selected[1]. Figure E.1 depicts the average results. The performance is on the y-axis, and the number of trees $\tau_{\max}$ is on the x-axis. We see that, for a given number of trees, increasing the depth leads to better performance (lower NMAE). However, the computation time depends on this depth, and one may prefer to compare performance for the same computation time. In the figure, the points indicate the best performance obtained for a fixed computation time, i.e. for 2000 trees of depth 1, 1215 trees of depth 2, 952 trees of depth 3, 764 trees of depth 4, 662 trees of depth 5, and 565 trees of depth 6. We see that, for the same computation time, using interaction of depth 6 relatively decrease NMAE by 10% compared to depth 1. Depending on the household considered, the top performance for a given computation time is sometimes obtained with depth of 4 or 5. Additionally, if one can afford the computation time, the larger depths eventually reach a lower NMAE.

While fitting such complex trees with $\Delta = 6$ seems to go against the "weak" learners philosophy of the gradient boosting, we explain this fact by noticing that our input selection is carefully done: each variable is highly relevant to the electricity demand, and fairly uncorrelated between each other. It therefore makes sense that use complex learners that well approximate real behavior.

E.1.3 Shrinkage Parameter

The shrinkage parameter λ determines the learning rate of the model at each iteration. A low value leads to better ultimate performance but necessitates more regression trees, and thus more computation time. We study the performance of the model regarding the shrinkage parameter with the fixed parameters, $\Delta = 6$, $\nu = 10$, and $p = 0.5$. The performance for different values of the shrinkage parameter λ averaged over the 10 US households is represented in Figure E.2. One sees that the minimal NMAE is quickly reached for high value of λ, e.g. after about 40 trees for $\lambda = 1$, but the ultimate NMAE reached by lower shrinkage value is lower. If one can afford the computation time, one should opt for the lowest value. In the case of limited computation time, there

[1]Note that the number of input variables does not necessarily correspond to an upper bound for the weak learners: one input can be used multiple times in a single weak learner.

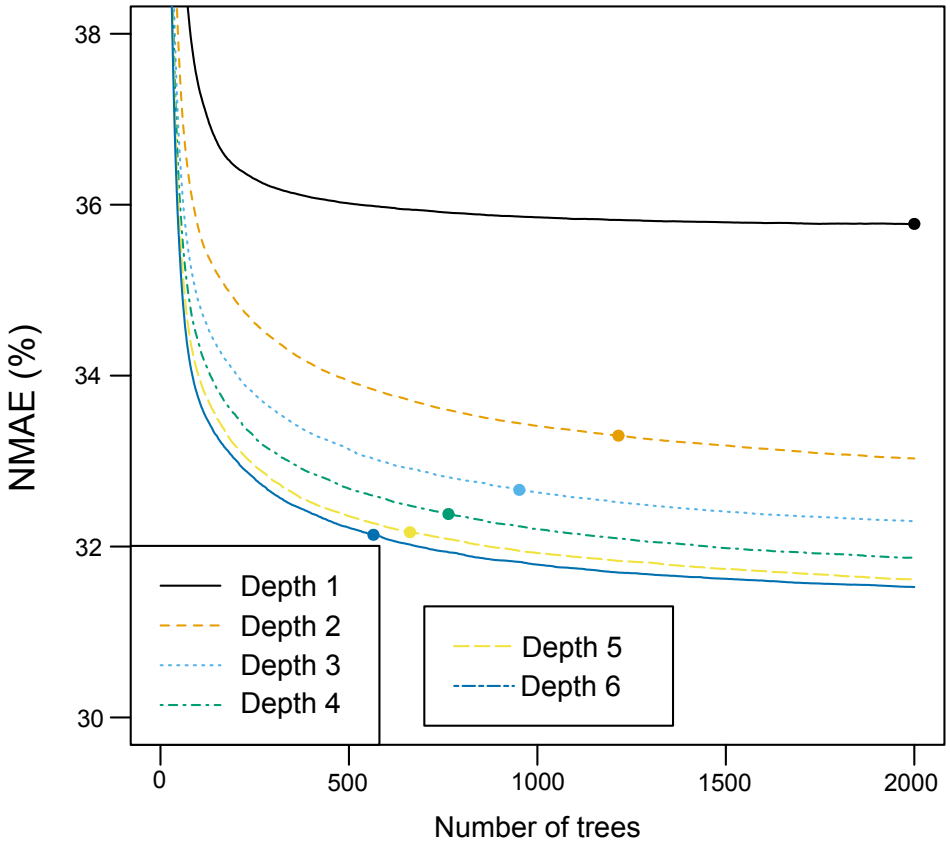


Figure E.1 – Forecasting performance of the gradient boosting model for various interaction depths, and number of trees stacked.

exists a trade-off between performance and number of trees optimizing the NMAE. For instance, for a depth $\Delta = 6$ and $\tau_{\max} = 565$, the value $\lambda = 0.1$ is the most efficient in average.

E.1.4 Minimal Node Size

The minimial node size ν is used to ensure that sufficient number of observations are present in the terminal nodes of the regression trees. Conceptually, larger ν leads to more conservative forecasts, i.e. less impacted by outliers but less flexible. In our case, the value of ν has almost no influence on the training process: we observe that the computation time remains the same for any minimal node size; computed trees are almost never rejected for too few observations. Consequently, the average performance is almost independent to this minimal node size. For other parameters fixed to $\Delta = 6$,

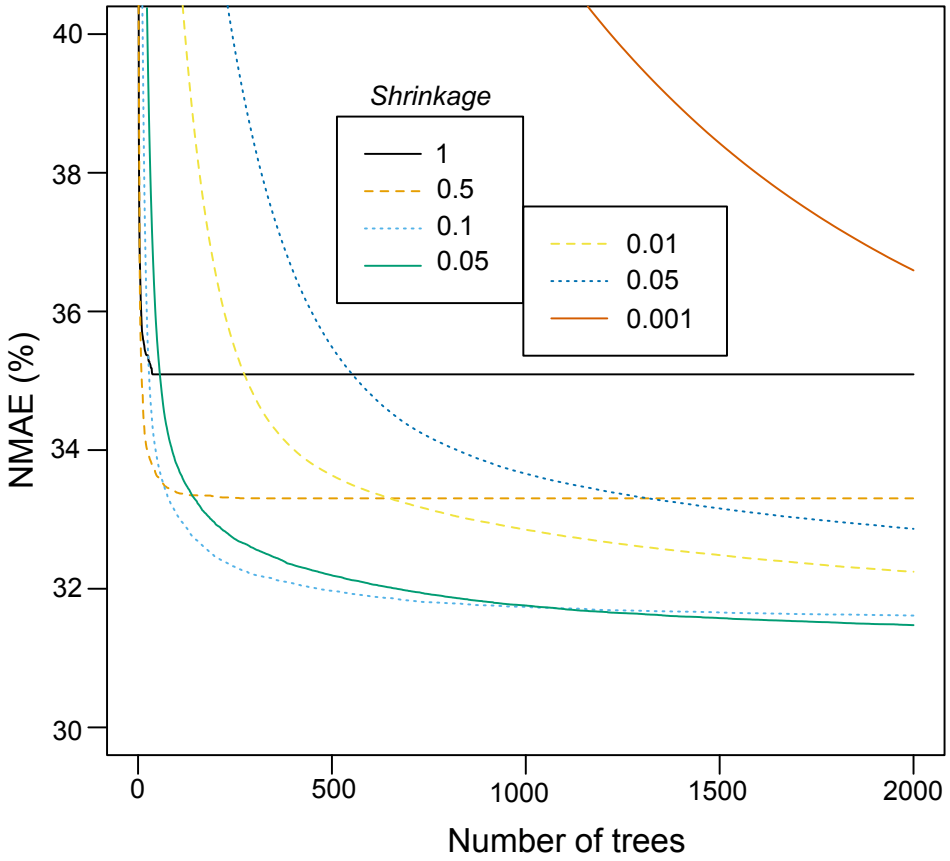
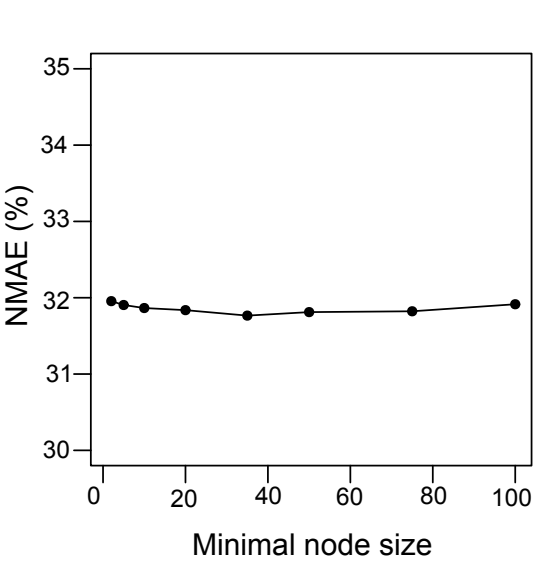


Figure E.2 – Forecasting performance of the gradient boosting model for various shrinkage parameters, and number of trees stacked.

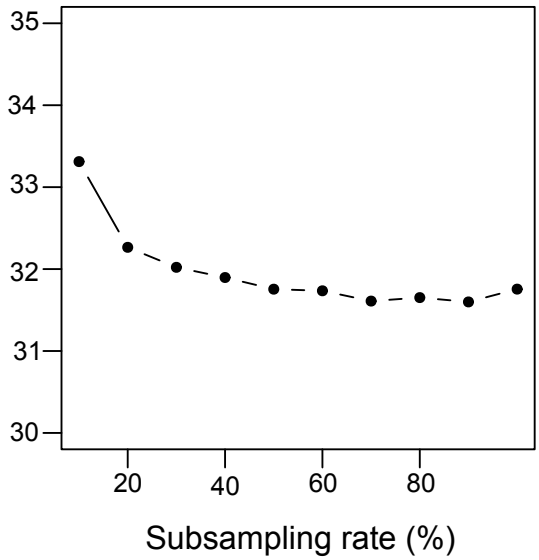
$\tau_{\max} = 565$, $\lambda = 0.1$, and $p = 0.5$, the performance (lack of) evolution is represented in Figure E.3a. This feature is due to our rather large dataset (8,760 values) compared to our small input set (6 variables. In general, we see a minor performance degradation for too small values, so we recommend to keep the default value $\nu = 10$.

E.1.5 Subsampling Rate

The subsampling rate p introduces a stochastic framework to the model training which generally reduces computation time (Friedman, 2002). The idea is; when p is close to 1, the model quickly overfits the data since all the dataset is used at each time. This issue is prevented by subsampling. In practice, with the cross-validation approach that selects the optimal number of trees, the subsampling value is less useful. We fix other parameters to $\Delta = 6$, $\tau_{\max} = 565$, $\lambda = 0.1$, and $\nu = 10$, and examine the



(a) Influence of ν



(b) Influence of p

Figure E.3 – Forecasting performance of the gradient boosting model for various minimal node size (a), and subsampling rate (b).

performance for different values of p, see Figure E.3b. We see that, when p is too low, the performance is degraded. In fact, for low subsampling rate, e.g. $p = 10\%$, one needs more trees for comparable performance, since the tree is fitted with only $p = 10\%$ of all data at each iteration. However, any value of p above 30% leads to similar performance, and so we recommend the default value of $p = 0.5$.

E.2 Performance for Various Configurations

The tuning process described in Section E.1 is made by analyzing performance in terms of NMAE averaged over a subset of 10 households, resulting in a meta-parameters configuration. A similar process is done specifically for each one of the 10 households to describe the various optimal configurations and reported in Table E.2. The whole tuning process is also repeated by using the normalized quantile score at 95% ($\text{NQS}_{0.95}$ to examine the variation between the middle and the extreme parts of the forecast distributions. The optimal configurations are reported in Table E.3.

All of these configurations are then used to train a gradient boosting model at quantile levels 50% and 95% individually for each one of the 10 households. The performance

Table E.2 – Meta-parameters configuration optimized over the NMAE of the gradient boosting model

Optimization w/ NMAE	Δ	$\tau_{\max}$	λ	ν	p
Default	1	2000	0.001	10	0.5
Global	6	565	0.1	35	0.9
Specific	5	662	0.1	5	0.7
	6	565	0.1	20	0.8
	6	565	0.1	35	0.4
	6	565	0.1	50	0.9
	6	565	0.1	20	0.6
	5	662	0.1	35	0.9
	6	565	0.1	35	0.7
	6	565	0.05	1	0.4
	5	662	0.1	75	0.6
	6	565	0.1	50	0.8

Table E.3 – Meta-parameters configuration optimized over the $\text{NQS}_{0.95}$ of the gradient boosting model

Optimization w/ $\text{NQS}_{0.95}$	Δ	$\tau_{\max}$	λ	ν	p
Default	1	2000	0.001	10	0.5
Global	6	565	0.05	75	0.8
Specific	5	662	0.1	100	0.7
	6	565	0.1	20	0.6
	6	565	0.1	75	0.5
	4	764	0.01	75	0.8
	6	565	0.05	35	1
	4	764	0.05	35	0.5
	4	764	0.05	20	0.5
	6	565	0.05	75	0.4
	4	764	0.05	05	0.8
	6	565	0.05	100	1

are then averaged and normalized by the performance obtained with the default configuration, so a performance of 80% means that you decrease the error obtained with the default configuration by 20%. The results are reported in Table E.4. We see that one may reduce by around 25% the performance by using optimized parameter rather than the default configuration. Moreover, one sees, quite surprisingly, that a global optimization, i.e. averaged over multiple households, is slightly more efficient than specific optimization for each household. This is due to the usage of more data that prevent major impact of outlying data points.. Optimizing the parameters at each quantile level also slightly improves the performance. However such a minor improvement is too computation intensive and an overall optimization with an increased number of trees $\tau_{\max}$ is relatively more efficient.

Table E.4 – Average forecasting performance obtained with diverse meta-parameters configurations.

Configuration	Optimization w/	NMAE	$\text{NQS}_{0.95}$
Default	—	100	100
Global	NMAE	**76.6**	74
Specific		76.8	74.2
Global	$\text{NQS}_{0.95}$	77.5	**73.1**
Specific		77.8	73.3

Résumé en français

Hereafter, we provide an extended chapter-by-chapter summary of the thesis in French. These complete the abstracts included at the beginning of each chapter (written in both English and French).

Chapitre 1 : Introduction

Contexte

Un intérêt croissant est porté à la demande électrique locale au vu des récentes évolutions du réseau électrique. Nous mettons quatre changements majeures en avant qui attise l'attention :

- *La décentralisation de la production et la libéralisation du marché.* Alors que les premiers réseaux électriques d'envergure concentraient la production d'électrique sur certains pôles bien identifiés, la libéralisation du marché enclenchée à la fin des années 1990 en Europe a changé la donne. Les grandes entreprises qui s'occupaient de la gestion globale du réseau (comme EDF en France) se scindent en plusieurs entités, et de nouvelles entreprises plus petites apparaissent, p. ex. des producteurs éparpillés sur les territoires. La libéralisation permet également l'émergence de nouveaux rôles (revendeurs, aggrégateurs) qui nécessite alors une plus étroite collaboration entre les acteurs, notamment au niveau locale pour tenir compte des particularités des zones.

- *Les énergies renouvelables.* Dans beaucoup de pays (à l'exception notable de la France), la production d'électricité émet d'importantes quantités de gaz à effet de serre. Dans le cadre des mesures actuelles pour réduire la pollution, ces pays souhaitent s'extraire de la dépendance au charbon et au pétrole en stimulant le développement des énergies renouvelables pour la production électrique : biomasse, éolien, solaire et hydraulique. Comme cette production est souvent locale et intermittente (éolien et solaire), une gestion très précise de l'offre et de la demande est nécessaire au niveau local.

- *L'installation de compteurs intelligents à grande échelle.* Grâce à leur mesure de la demande électrique à fines échelles temporelles (mesures toutes les 15 minutes, ou toutes les heures) et spatiales (demande d'un seul ménage), les compteurs intelligents donnent une vision détaillée de l'état du réseau. Ces mesures devraient permettrent de diminuer la consommation électrique globale, p. ex. en identifiant les pertes du réseau et en fournissant des informations aux utilisateurs pour adapter leurs habitudes.

- *Le développement de l'autoconsommation.* De nouveaux moyens de production (p. ex. des panneaux solaires) permettent à des particuliers, ou des voisinages, de produire une énergie locale. Avec une législation adaptée, cette pratique permet à ces petits producteurs d'utiliser directement la production sans passer par le réseau. Ils économisent alors une partie du tarif d'utilisation du réseau, mais conservent la sécurité apportée par celui-ci en cas de panne. Un développement optimale de cette pratique repose sur la gestion très précise de la demande et de l'offre locale.

Enjeux et objectifs

Nous désignons par demande électrique, la puissance moyenne appelée pendant une certaine période, p. ex. une heure. Nous définissons l'échelle locale comme allant du simple appareil (puissance de 100 W) au départ HTA (puissance de 1 MW), avec une attention spéciale pour la demande d'un ménage (puissance autour de 1 kW). Nous définissons le court terme comme des horizons de prédiction allant de 1 heure à 1 semaine.

Prédire la demande électrique locale, p. ex. d'un ménage, est une tâche plus délicate que prédire celle à une plus grande échelle, p. ex. celle d'un pays. La figure 1.5 montre les 24 demandes horaires d'un ménage états-unien pendant une journée : la courbe obtenue est très volatile avec une demande parfois triplée en l'espace d'une heure. De plus, une visualisation sur une plus grande période indique que ces courbes changent du tout au tout entre deux jours successifs. Certains facteurs, qui ont une influence marquée sur la demande nationale (p. ex. la température), ont un impact imperceptible sur la demande d'un seul ménage. Ces observations préliminaires doivent être étayer par une analyse plus détaillée afin de *caractériser la demande électrique à l'échelle locale.*

Les modèles de prédiction à l'échelle nationale tirent parti des motifs récurrents et des facteurs externes ayant une forte influence sur le niveau de la demande. Les modèles les plus avancés atteignent aujourd'hui une précision de l'ordre de 2%. Traditionnellement, les prédictions à grande échelle sont faites de manière détermiste, p. ex. la demande entre 9 et 10 heures sera de 4 GW demain, et est proche de la vraie valeur mesurée *a posteriori*, la vraie demande est de 4.1 GW. Une telle approche n'est

pas pertinente à l'échelle d'une maison à cause du caractère volatile de cette demande, fortement liée aux comportements imprévisibles des habitants. Par conséquent, une approche probabiliste est nécessaire pour quantifier l'incertitude que l'on a dans la future demande, p. ex. prédire que la demande du ménage sera probablement entre 100 et 600 W. Il faut donc *développer des modèles de prédictions probabilistes*.

Pour une installation des modèles de prédiction sur un cas réel pour des centaines de ménages à la fois, ces derniers doivent faire preuve de réplicabilité. Ce concept englobe plusieurs caractéristiques : fonctionnement avec peu de maintenance, adaptation à différents cas, et grande robustesse pour donner une prédiction en temps réel et en toute circonstance. Toutes ces caractéristiques doivent être réunies pour *garantir la réplicabilité des modèles*, au risque de dégrader légèrement la performance en comparaison à des tests en laboratoire.

Enfin, les prédictions n'ont que peu de valeur en elles-mêmes mais en obtiennent avec des applications ultérieures. Ces dernières reposent souvent sur des optimisations sur la journée complète, et nécessitent par conséquent des prédictions de demande durant plusieurs périodes successives, généralement pendant la journée complète. Cela veut dire qu'il faut *générer des scénarios prédictifs de la demande*. Ces scénarios doivent assurer la cohérence multi-temporelle des valeurs de la demande, tout en représentant l'incertitude liée à la prédiction.

Plan

La chapitre 2 présente une introduction aux modèles de prédictions statistiques pour la demande électrique avec les méthodes pour évaluer leur performance. Une revue de la littérature consacrée à ce sujet est ensuite dressée.

Le chapitre 3 est consacré à la demande électrique au niveau d'un départ HTA. Nous proposons une méthode de désagrégation de la demande totale du départ en profils élémentaires. Nous illustrons l'intérêt de cette désagrégation avec (1) la prédiction de la demande à moyen terme, et (2) une analyse prospective de l'évolution du pic de demande.

Le chapitre 4 s'intéresse à la demande électrique d'un ménage. Nous étudions trois jeux de données pour créer et évaluer un modèle de prédiction pour le lendemain. La performance de ce modèle est ensuite étudiée pour d'autres cas d'études, notamment

sur différents niveaux d'agrégation et de temporalité. Enfin, une structure de prédiction est d'abord développée, puis utilisée en temps réel sur un projet de démonstration, et enfin évaluée.

Le chapitre 5 étudie la génération de scénarios de demande journaliers. Nous comparons diverses méthodes pour la génération, puis la réduction, de scénarios à l'échelle d'une maison. Enfin, nous proposons une méthode pour générer les scénarios de la demande d'un seul appareil après une analyse des habitudes des usagers faite uniquement à partir de mesures de la demande.

Le chapitre 6 conclut en résumant les travaux et en présentant quelques perspectives de recherche.

Chapitre 2 : Modèles statistiques de prédiction

Introduction aux modèles de prédiction

Statistiquement, nous considérons qu'un phénomène future, à l'instant $t+h$, noté y_{t+h}, est une réalisation d'une variable aléatoire $Y_{t+h|t}$ qui dépend des informations i_t connues jusqu'à l'instant t. Ce phénomène est par exemple la demande électrique. La fonction de répartition de cette variable aléatoire s'écrit

$$F_{t+h|t}(y) = \mathbb{P}\left[Y_{t+h|t} \leq y | i_t\right].$$

On utilise aussi la densité de probabilité, c.-à-d. sa dérivée $f_{t+h|t}(y) = F'_{t+h|t}(y)$. Nous notons que la variable aléatoire $Y_{t+h|t}$ dépend de l'horizon h considéré. De manière générale, plus cet horizon est grand, plus la valeur future est incertaine, ce qui se traduit par une fonction de répartition de plus grand support. Cela veut aussi dire que la performance d'une prédiction diminue généralement quand l'horizon augmente.

En pratique, les fonctions de répartitions et densités de probabilité ne sont pas connues , même *a posteriori* : nous n'observons qu'une seule réalisation dans un état donné. Pour estimer ces fonctions, on rassemble des observations faites dans des cas similaires, en utilisant un sous-ensemble d'informations $s_{t+h|t}$ plutôt que toutes les informations i_t[2]. Ce sous-ensemble dépend de l'horizon et doit être choisi avec soin par le prévisionniste. Mathématiquement, nous indiquons cette approximation par des chapeaux, c.-à-d. $\hat{F}_{t+h}(\cdot)$ et $\hat{f}_{t+h}(\cdot)$.

Il existe deux sortes de prédiction : une prédiction *ponctuelle* et une prédiction *probabiliste*. Historiquement, les prédictions ponctuelles ont précédé les probabilistes, avec une transition que Stigler date au XIV^{e} siècle (Stigler, 1986). Les deux sortes de prédiction coexistent aujourd'hui.

Les prédictions ponctuelles sont parfois appelées *déterministes*, par opposition à probabiliste. Ces prédictions peuvent être données sous différentes formes : l'espérance future, la valeur médiane future, ou le quantile future à un niveau donné.

Les prédictions probabilistes permettent d'indiquer la confiance que l'on a dans la future en quantifiant l'incertitude dans la valeur future du phénomène. Les prédictions peuvent être données sous différentes formes : un échantillon de Monte Carlo, une liste de quantiles, ou un intervalle de prédiction .

[2]Nous omettons l'indice $|t$ par la suite.

Un modèle statistique s'écrit sous une forme générale avec des paramètres à sélectionner en fonction du cas d'étude. Cette sélection repose sur la comparaison systématique entre ce que le modèle prédit et les observations connues. Cela veut dire qu'il faut un *ensemble d'apprentissage* pour lequel on connait à la fois les informations s_{t+h} et les vraies observations y_{t+h}. La comparaison est faite avec une fonction de perte que l'on cherche à minimiser en faisant varier les paramètres. Le type de prédiction obtenue dépend de cette fonction de perte, p. ex. la perte absolue conduit à une prédiction médiane. Toutefois, il faut veiller à bien choisir son modèle statistique pour éviter le phénomène d'*overfitting*, quand les paramètres du modèle sont trop spécifiques à l'ensemble d'apprentissage et ne conduisent pas à des prédictions précises dans d'autres cas.

Parmi les modèles de prédiction les plus courants que nous utilisons par la suite, nous citons le modèle linéaire, le modèle additif, l'estimateur par noyau de densité, ou le modèle *gradient boosting*. Une description en anglais est donnée au paragraphe 2.1.2.

Performance d'un modèle de prédiction

Un modèle de prédiction doit être évaluer avant d'être utiliser en temps réelle dans une application pratique. Cependant, la performance d'un modèle dépend fortement des besoins de l'utilisateur des prédictions ; des prédictions idéales pour une application ne le sont pas pour une autre applications. Murphy (Murphy, 1993) explique qu'il y a trois aspects concernant la performance des prédiction : la cohérence (le previsionniste fait le meilleur usage de ses connaissances), la qualité (la proximité entre les prédictions et les observations), et la valeur (le fait d'être utile à l'utilisateur des prédictions). D'un point de vue purement statistique, seuls les deux premiers aspects peuvent être évalués à l'aide de scores, autrement nommés indices. De même que les prédictions, il existe deux sortes de scores, adaptés aux prédictions ponctuelles ou probabilistes.

Nous mettons en avant les scores ponctuels suivants : le biais (Bias en anglais), l'erreur absolue moyenne (MAE est l'abréviation anglaise), et la racine de l'erreur quadratique moyenne (RMSE est l'abréviation anglaise). De même, nous mettons en avant les scores probabilistes suivants : la fiabilité (Rel est la version anglaise raccourcie), le *continuous ranked probability score* (abrégé en CRPS), et le score quantile (QS est l'abréviation anglaise). Ces scores sont généralement normalisés pour obtenir des

scores comparables entre les cas d'étude. Nous choisissons une normalisation par la valeur moyenne du phénomène. Nous définissions précisément ces scores et introduisons des variantes dans le paragraphe 2.2.

État de l'art des modèles de prédiction de la demande électrique

Nous avons appuyé nos recherches sur des travaux majoritairement écrits en langue anglaise. Ainsi, nous signalons simplement les grandes lignes ici, et renvoyons à l'état de l'art plus complet proposé en anglais au paragraphe 2.3.

De multiples travaux sur la prédiction de la demande à court terme ont été réalisés. En général, les auteurs distinguent des modèles de prédiction à échelle régionale (pour la demande électrique d'une région ou d'un pays), et celles à échelle locale (pour la demande électrique d'un ménage ou d'un quartier). Étant donné la disponibilité plus anciennes des données, les modèles à échelle régionale ont été développés depuis plus longtemps et arrivent maintenant à maturité. Les chercheurs ont noté et cherché à quantifier l'influence de la température sur la demande (quand il fait très chaud ou très froid, la consommation augmente). Aussi bien les méthodes de série temporelle traditionnelle (p. ex. modèle ARMA) que des méthodes plus modernes (p. ex. les réseaux de neurones) conduisent à de bonnes performance une fois leurs paramètres convenablement réglés. La question de la prédiction de la demande d'un menage est plus récente et dynamique aussi bien dans le monde académique qu'industrielle, avec l'organisation de compétitions internationales. Les mêmes modèles qu'à l'échelle régionale ont été testés et évalués avec les données nouvellement disponibles. Les chercheurs notent qu'un même modèle conduit à des performances très différentes selon le ménage et la variabilité de la demande (les erreurs relatives passent de 2% à 85%). En particulier, l'intérêt d'incorporer la température pour produire des prédictions à court terme est souvent remis en cause. Des méthodes originales sont proposées pour effectuer la prédiction, comme la classification des profiles individuels de demande.

Cette revue de la littérature confirments différents travaux qui notent que la performance augmente, ou de manière équivalent que les erreurs de prédiction diminuent, quand la puissance moyenne de la demande à prédire augmente : il est plus facile de prédire la demande électrique d'une région que celle d'un ménage. Nous reprenons la performance de tous les travaux étudions et uniformisons les scores de performance

(sous la forme du NMAE) pour représenter la figure 2.5. Une loi d'échelle est ajustée à ces données (Sevlian & Rajagopal, 2014) : l'erreur décroît selon l'inverse de la puissance à la racine quatrième pour les petites puissances jusqu'à atteindre une erreur irréductible autour de 2,5% à l'échelle d'un pays (1 GW ou plus).

Chapitre 3 : La demande électrique à l'échelle d'un départ

Désagrégation de la demande électrique d'un départ en profils élementaires

Nous nous intéressons à la demande électrique d'un départ HTA (Haute Tension de type A), c.-à-d. la demande totale faite par un ensemble de 1000 à 10 000 clients. Les mesures de la demande à cette échelle ont plusieurs avantages : elles sont exhaustives (les demandes individuelles et les pertes sont inclus), non-intrusives (la demande d'un client est dissimulée par celle de tous les autres) et couvrent une longue période.

La modélisation de cette demande, à des fins d'analyse ou de prédiction, se fait généralement avec deux approches distinctes. D'abord, une approche globale qui cherchent à identifier précisément les variables influençant la demande électrique (p. ex. la température), ainsi que l'analyse historique détaillée de la dynamique de la demande (p. ex. avec des cycles hebdomadaires). Parallèlement, une approche inductive, constructiviste ou *bottom-up*, cherche à construire la demande du départ en sommant les demandes de sous-groupes constituant l'ensemble des clients du départ. Cela passe par la collecte des demandes de ses sous-groupes ou par une extrapolation de mesures partielles avec des méthodes de classification. De manière générale, l'approche globale donne des prédictions plus précises que l'approche inductive, mais cette dernière permet souvent une compréhension plus fine des mécanismes en jeu.

Nous proposons une méthode intermédiaire pour modéliser la demande des départs sur des cas d'études en France. Nous nous appuyons à la fois sur les demandes historiques au niveau des départs, et sur les informations (comme la consommation annuelle ou le type de contrat de chaque client) à propos des clients connectés au départ. Nous proposons de créer de larges catégories qui représentent les clients d'un départ. De cette façon, nous caractérisons la démographie connectée à chaque départ, notons que la taille des catégories reposent sur la consommation d'énergie annuelle et non sur un simple décompte. L'utilisation de cette caractérisation pour de multiples départs permet alors de décomposer la demande du départ, mesurée toutes les 10 minutes, en sous-demande pour chaque groupe. Ainsi, la dynamique de chaque groupe apparaît (p. ex. la demande de clients résidentiels diffèrent de celle de bureaux).

Cette désagrégation n'est possible que si les demandes de plusieurs départs sont comparables. Ce n'est pas le cas des valeurs brutes : la demande d'un départ peut passer du simple ou double selon la situation. Nous effectuons donc deux transformations. D'abord, nous prenons en compte la thermosensibilité qui varient selon chaque départ. L'influence de la température sur la demande a été abondemment remarquée. Dans le contexte français, cet effet est visible uniquement pour les températures froides. Nous définissions un palier de température en-dessous duquel la demande augmente linéairement quand la température descend (le palier trouvé est généralement autour des 16°C, la température de confort généralement reconnue). Le palier ainsi que la pente sont trouvés automatiquement pour chaque départ et chaque heure de la journée. De plus, nous normalisons la demande de puissance électrique de chaque départ par l'énergie de ce départ sur la période considérée, la journée dans notre cas. De cette façon, chaque départ possède un profil journalier (composé de $6 \times 24 = 144$ valeurs, une toute les 10 minutes) de valeurs adimensionnelles fluctuant autour de 1.

Après ces transformations, la décomposition s'écrit sous la forme d'un problème d'optimisation sous contraintes, voir l'équation (3.5). Pour le résoudre en un temps raisonnable, nous adaptons un algorithme d'*alternating direction method of multipliers* à notre problème (Boyd et al., 2011).

Nous proposons plusieurs caractérisations (ou catégorisations) des départs : en 2, 8, 9, ou 12 catégories en utilisant plus ou moins de détails sur les clients connectés. Un exemple de profils obtenus est visible sur la figure 3.3 pour un jour de la semaine avec un jeu de données de la région lyonnaise. Les dynamiques variées des différents groupes sont bien visibles : les commerces sont actifs durant la journée, tandis que les équipements publics (éclairage, ascenceurs, etc.) le sont la nuit.

Utilisation des profils élementaires

Nous pouvons utiliser les profils élémentaires pour faire de la prédiction de la demande d'un nouveau départ dont nous connaissons uniquement la caractérisation. Un exemple est donné sur la figure 3.4 pour un départ dont la caractérisation est faite en deux catégories : 75% de l'énergie est pour des clients résidentiels, et 25% de l'énergie est pour des clients tertiaires. Nous réalisons des simulations exhaustives sur nos trois jeux de données avec différentes caractérisations. Nous notons qu'il existe un nombre

optimal de catégories : trop petit et la prédiction est trop grossière et imprécise, trop grand et les profils obtenus sont sur-appris (*overfitted*) sur l'ensemble d'apprentissage et la performance sur un ensemble test est dégradée. Généralement, nous observons qu'utiliser 9 catégories est, en moyenne, le plus efficace avec des erreurs variant entre 12 et 15%. En réalité, connaître le nombre optimal de catégories *a priori* pour un départ est délicat puisque ce nombre dépend (a) de la variabilité de la demande, et (b) de la taille respective des catégories. Dans la plupart des cas, le nombre optimal augmente quand la variabilité augmente, et quand les tailles des catégories sont comparables.

De plus, ces profils élémentaires permettent d'anticiper les caractéristiques futures d'un départ, selon l'évolution des clients connectés au départ. Nous nous intéressons par exemple à l'instant et la valeur du pic de demande quotidien. Quand on raccorde de nouveaux clients à un départ, la valeur du pic de demande augmente naturellement, mais de façon plus ou moins importante selon le type de ces clients supplémentaires. Nous prenons un cas d'étude spécifique où le pic de demande est actuellement à 12:10. Nous constatons que si l'on ajoute beaucoup de clients résidentiels, ce pic va finir se produire à 23:00 et grandir de façon importante. À l'inverse, si l'on ajoute beaucoup de buraux, le pic restera à 12:10 et augmente moins rapidement.

Chapitre 4 : Prédiction de la demande électrique d'un ménage

Caractéristiques de la demande électrique d'un ménage

Les compteurs intelligents (baptisés *smart meters* en anglais) mesurent la demande électrique, sous forme de puissance moyenne (en kilowatts) d'un ménage à intervalles réguliers, toutes les 30 minutes ou une heure. Leur installation récente permettent de récolter de nombreuses séries temporelles. Nous appuyons nos analyses sur trois jeux de données composés de : 176 séries pour des maisons vers Tours en France, 226 séries pour des bâtiments à Évora au Portugal, et 175 séries pour des ménages à Austin aux États-Unis (Texas). Nous comparons ces séries à l'échelle individuelle à celles mesurées à l'échelle des départs et des régions correspondantes. Nous remarquons cinq différences majeures entre ces échelles qui nécessitent la création de modèle de prédiction spécifique à l'échelle individuelle :

1. *Le lissage de la courbe quotidienne.* La variabilité des courbes de demande s'évanouit à mesure que l'on somme les demandes par un effet de foisonnement. On note que le facteur de charge passe de 50% pour la demande d'un ménage à 90% pour la demande d'une région.

2. *La périodicité.* Il existe des cycles périodiques marqués pour la demande électrique : horaire, journalier, hebdomadaire, etc. À l'aide de modèles de persistance, nous constatons que les périodicités sont plus marquées pour une région que pour une maison.

3. *La distribution horaire.* La distribution statistiques des demandes horaires est plus étalée pour la demande des ménages que pour celles des départs. Une analyse superficielle des données des compteurs individuelles peut laisser penser à tort que certains pics sont des valeurs aberrantes.

4. *L'influence de la température.* Tandis que cette influence est marquée à grande échelle, elle est plus subtile à l'échelle d'un ménage. Nous montrons que cet impact peut être capturé par un simple profil de température figurant les principaux cycles.

5. *Les périodes de vacance.* Contrairement à l'échelle d'un départ où les vacances sont indiscernables sur la courbe de demande, nous constatons, pour environ 15% des ménages, une inactivité prolongée sur la courbe de demande (en moyenne 10 jours).

Modèle de type *gradient boosting*

Nous développons un modèle de prédiction pour le lendemain, adapté pour la demande électrique d'un ménage, de type *gradient boosting* implémentée dans la librairie `gbm` disponible dans le langage **R** (Ridgeway, 2017). Nous sélectionnons trois types d'entrées (pour un total de 6 *inputs*) : des mesures historiques (demande la veille, et demande médiane de la semaine écoulée), des informations contextuelles (heure de la journée, et jour la semaine), des prédictions de température (température ponctuelle, et température lissée). Pour obtenir des prédictions probabilistes, nous créeons de multiples modèles en parallèle définis avec différents scores quantiles comme fonctions de perte, si bien que nous produisons un ensemble de 99 valeurs quantiles pour la prédiction de la demande à un seul instant. Les méta-paramètres des modèles (nombre d'arbres, profondeur, etc.) sont précisément ajustés (*cf.* les paragraphes 4.2.1 et 4.2.2, et l'annexe E). Nous comparons la performance de notre modèle pour prédire la demande d'un ménage pour le lendemain et la comparons à deux modèles de référence : un modèle de persistance, et un modèle climatologique. Pour tous les cas d'étude, notre modèle est le plus efficace, améliorant la qualité d'environ 30%, pour atteindre une erreur déterministe (NMAE) de 28% (*cf.* le paragraphe 4.2.4). Concernant l'aspect probabiliste, nous atteignons une erreur (NCRPS) de 20%, mais observons que la queue de distribution sur la partie supérieure est généralement délicate à prédire, si bien que le modèle climatologique, c.-à-d. des prédictions très conservatrices, est parfois plus efficace que notre modèle *gradient boosting*.

Performance de la prédiction et l'effet de foisonnement

Avec ce modèle, nous effectuons les prédictions de la demande pour différents groupes de maison (puissance moyenne entre 1 et 100 kW) et résolutions temporelles (intervalles de 1 minute à 1 semaine). Cela nous permet d'étudier systématiquement l'effet de foisonnement, et de mesurer à quel point il joue sur la performance de prédiction. La

figure 4.16 représente les erreurs NMAE obtenues avec le jeu de données des États-Unis. Comme attendu, il est plus facile de prédire les niveaux d'agrégation supérieurs : p. ex. il est plus facile d'anticiper la demande hebdomadaire d'un groupe de maison (NMAE de 4%) que la demande exacte d'une seule maison demain à 16:32 (NMAE de 43%). Concernant le niveau d'agrégation, nous trouvons un optimum, au-delà duquel la performance stagne, quand nous prédisons la demande agrégée de 15 maisons à la fois.

Concernant la résolution temporelle, une question proche est de savoir s'il faut utiliser la même résolution pour l'apprentissage du modèle et la prédiction en elle-même. Nous trouvons que, quel que soit le niveau d'agrégation, il faut apprendre avec une résolution légèrement plus précise, mais non pas trop précise ni moins précise (*cf.* le paragraphe 4.3.4). Par exemple, utiliser une résolution de 30 minutes plutôt qu'une heure (pour prévoir la demande horaire du lendemain) diminue les erreurs de l'ordre de 5%, tandis qu'utiliser une résolution de 5 minutes ou de 6 heures les augmente de l'ordre de 10%.

Modèle de prédiction robuste et défis opérationnels

Dans un contexte pratique, nous ne pouvons pas utiliser notre modèle de *gradient boosting* car le modèle de prédiction doit satisfaire plusieurs critères : grande robustesse, calculs rapides, large réplicabilité, intervention à distance, résultats interprétables. Nous développons ainsi plusieurs modèles de complexité, et précision, variée que nous combinons dans une structure de prédiction probabiliste originale (voir figure 4.24).

Le modèle le plus efficace que nous propons est un modèle additif, que nous avons déjà étudié en détails (Gerossier, Girard, et al., 2017), tirant partie de seulement 3 variables d'entrée pour concilier efficacité et interprétabilité. Sa performance est comparable à celle du modèle *gradient boosting* avec des erreurs NMAE de 27% dans notre cas d'étude portugais. La structure de prédiction est installée sur un projet de démonstration et est actuellement en fonctionnement. Notre structure fait bien usage de tous les modèles implémentés dans la structure pour produire une prédiction probabiliste à chaque instant et pour chacune des maisons. La performance des prédictions en temps réel est inférieure à celles faites dans une évaluation au préalable : les erreurs augmentent de l'ordre de 68%. Quand on prend en compte l'impact des deux situations

un peu différentes (périodes de test variable, évolution des habitudes des maisons, etc.), la dégradation est mineure, autour des 5%.

Chapitre 5 : Prédiction de la demande électrique avec des scénarios

Scénarios de la demande électrique d'un ménage pour le lendemain

Les prédictions probabilistes sont généralement produites indépendamment d'un intervalle au suivant. Pourtant, la demande électrique d'un ménage est fortement corrélée d'un instant à l'autre, étant données la présence des habitants et leurs activités. Ainsi, quand la demande à 15h00 est plus importante que prévue la veille, elle est également plus importante que prévue à 16h00. Par conséquent, même si la prédiction probabiliste à chaque instant est optimale, l'ensemble de plusieurs de ces prédictions ne l'est pas. Ce problème peut être résolu par l'utilisation de scénarios regroupant, p. ex. 24 valeurs de la demande horaire pour les 24 heures d'une journée.

Nous effectuons des prédictions probabilistes pour le lendemain pour chacune des 175 ménages de notre jeu de données des États-Unis (voir le détail au paragraphe 5.1.1). Suivant l'exemple de ce qui est proposé pour la prédiction de la production éolienne (Pinson & Girard, 2012), nous étudions la série résiduelle de la demande, notée (z_t), plutôt que les valeurs brutes. Cela revient à regarder dans quelle partie de la distribution prédite tombent les vraies valeurs de demande. Notons qu'il n'y a, en principe, aucune façon d'anticiper les vraies valeurs ; quand les prédictions probabilistes sont bien calibrées, la série résiduelle est uniformément distribuée sur $(0, 1)$. Créer des scénarios quotidiens de la demande revient alors à générer un ensemble de 24 points (un par heure de la journée), $(\hat{z}_0, \hat{z}_1, \cdots, \hat{z}_{23})$ qui aie des caractéristiques proches de celles d'une vraie trajectoire $(z_0, z_1, \cdots, z_{23})$. Ces points sont ensuite transformés en vraies valeurs de demande (en kilowatt).

Nous proposons quatre méthodes pour générer les scénarios :

1. *Relier les quantiles.* Toutes les valeurs de $\hat{z}_t$ sont prises égales, et donc les valeurs entre les heures sont complètement corrélées.

2. *Tirage uniforme.* Les valeurs de $\hat{z}_t$ sont indépendemment tirés selon une loi uniforme entre 0 et 1, et donc les valeurs entre les heures sont complètement décorrélées.

3. *Matrice de covariance standard.* La matrice de covariance (de taille 24) est calculée avec les trajectoires observées et utilisé pour générer l'ensemble des 24 valeurs.

4. *Matrice de covariance raffinée.* Deux matrices de covariance sont calculés et utilisés selon le jour de la semaine, et un paramètre d'oubli est défini pour ne conserver que les trajectoires, et donc les dépendences, récentes lors de la génération.

En règle générale, les scénarios 1) sont trop lisses, les scénarios 2) sont trop irréguliers, et seuls les scénarios 3) et 4) fournissent des courbes de demande réalistes au premier abord. Quand on examine une heure précise, l'ensemble des valeurs des scénarios générées constituent un échantillon de Monte Carlo prédictif. Nous trouvons qu'il faut au moins 400 scénarios pour garantir une performance optimale, voir le paragraphe 5.1.2.4. Plusieurs critères sont définis pour évaluer précisément la qualité des scénarios, notamment évaluant la cohérence des 24 valeurs entre elles. Ils viennnt confirmer que les scénarios 4) sont les meilleurs, avec une légère amélioration en comparaison des scénarios 3) (moins de 2%), et une nette amélioration en comparaison des scénarios 1) et 2) (environ 20%).

Pour diminuer ce nombre de scénarios nécessaires, nous proposons une méthode de réduction, nommée *fast forward reduction scenarios* (Bruninx & Delarue, 2016). L'algorithme se base une métrique qui mesure la distance entre les 400 scénarios générées pour conserver uniquement les plus représentatifs. La métrique dépend de ce que veut faire l'utilisateur des scénarios prédictifs. Par exemple, nous en définissions trois : une distance ponctuelle, une distance selon les caractéristiques du scénario (énergie totale, valeur du pic, heure du pic, énergie de la soirée), et une distance adaptée à la demande d'un ménage pondérée par le prix. Nous comparons la qualité des scénarios réduits à celles de l'ensemble complet de 400 scénarios. Nous constatons qu'il suffit d'une vingtaine de scénarios représentatifs pour obtenir la même efficacité (voir p. ex. la figure 5.11).

Scénarios de la demande d'un véhicule électrique

Dans un second temps, nous cherchons à produire des scénarios prédictifs de la demande électrique d'un seul usage, en l'occurrence pour le chargement de la batterie d'un véhicle électrique (VÉ). Le nombre de VÉ est censé augmenté très largement dans un futur,

et quand la batterie est chargée chez l'habitant, la consommation des ménages va fortement évoluer. Nous recueillons les mesures de la demande faites au niveau de la station de recharge de 46 véhicules aux États-Unis. Les mesures sont faites minute par minute, si bien que la série temporelle des demandes est extrêmement détaillée, avec 1440 valeurs pour une journée.

Dans notre cas d'étude, la recharge de la batterie se fait à une puissance nominale constante, et donc la série temporelle est faite de deux niveaux si la recharge est en cours ou non. Nous développons un algorithme qui passe en revue la série temporelle sur l'année complète afin de détecter (1) la puissance nominale, (2) la durée de la recherge, et (3) la minute de la journée quand la recharge commence. Cela permet de modéliser précisément les habitudes de chargement de l'utilisateur du VÉ.

Chapitre 6 : Conclusion

Résumé

Dans le chapitre 1, nous mettons en avant les évolutions du réseau électrique qui nécessite une compréhension au niveau local, ce qui passe notamment par l'étude de la prédiction de la demande locale à court terme. D'abord, la décentralisation de la production électrique ainsi que la libéralisation du marché de l'énergie contribuent à une augmentation des interactions entre les opérateurs travaillant à une échelle plus locale. Par ailleurs, l'intégration des énergies renouvelables est un défi majeur à cause de leur taille moindre et l'incertitude sur leur niveau exact de production. Cette gestion locale passe par la mise en place de compteurs intelligents à grande échelle pour (1) mesurer la demande de façon détaillée, et (2) développer de nouvelles pratiques, comme l'autoconsommation, promettant d'alléger les contraintes sur le réseau de distribution.

Dans le chapitre 2, une introduction aux modèles statistiques de prédiction est donnée. Nous nous attardons sur les modèles les plus courants et les façons d'évaluer la qualité des prédictions d'un modèle, que celui-ci soit déterministe ou probabiliste. Nous passons en revue la littérature consacrée à la prédiction de la demande électrique à court terme. Nous analysons et comparons les performances indiquées dans les travaux étudiés, et identifions une loi d'échelle qui met en rapport la précision des prédictions et la puissance moyenne du cas étudié : les erreurs relatives de prédiction passent de 30% pour une maison (1 kW) à 3% pour un pays (1 GW).

Dans le chapitre 3, la demande électrique d'un départ HTA, comprenant entre 1000 et 10 000 consommateurs, est étudiée. Des mesures de la demande à cette échelle existent depuis longtemps, si bien que les facteurs exogènes qui impactent le niveau de demande sont bien connus, comme la température. Les erreurs relatives des modèles de prédiction à court terme sont autour des 10%. Nous proposons un algorithme pour décomposer la demande en profils élementaires. Chaque profil correspond à la demande moyenne d'une certaine classe de consommateurs. Pour effectuer la décomposition, nous utilisons les mesures de plusieurs départs HTA en même temps, ainsi qu'un descriptif des consommateurs connectés. Sur plusieurs cas d'études, nous illustrons l'intérêt de ces profils de demande : nous prédisons la demande d'un départ HTA jamais mesuré (avec des erreurs entre 12% et 15%), et nous anticipons l'évolution du pic de demande. Nous avons publié une partie de ces travaux en 2017 dans le journal

Applied Energy (Gerossier, Barbier, & Girard, 2017).

Dans le chapitre 4, nous traitons de la prédiction de la demande électrique au niveau d'une maison. Les caractéristiques de cette demande (sa dynamique, ses mécanismes, sa régularité, etc.) sont détaillées et mises en relation avec celles observées à une échelle plus large. Un modèle de type *gradient boosting* est développé et sa performance est analysée à l'aide de plusieurs jeux de données : l'erreur déterministe moyenne (indice NMAE) est de 28% pour la prédiction de la demande horaire du lendemain. Ce modèle constitue une référence à l'échelle d'une maison. Avec des tests exhaustifs, nous l'utilisons pour prédire la demande pour une large gamme de niveaux d'agrégation (allant de la demande d'une simple maison à celle d'un groupe de 200 maisons) et de résolutions temporelles (demande moyennée sur une période allant d'une minute à une semaine). Nous concluons que les erreurs de prédiction diminuent quand l'agrégation est importante et quand la résolution diminue. Un optimum est trouvé pour la prédiction d'un groupe de 15 maisons. Par ailleurs, nous développons une structure de prédiction pour répondre au défi de la réplicabilité. La structure est composée de plusieurs modèles de prédictions afin de produire une prédiction probabiliste en toute circonstance. Ce travail a été présenté à la conférence CIRED 2017 (Gerossier, Girard, et al., 2017), et dans un article publié dans le journal *Energies* (Gerossier et al., 2018). Correa-Florez et al. optimisent l'utilisation d'un *smart home energy management* à l'aide des prédictions obtenues dans le cas d'étude présenté (Correa-Florez et al., 2018).

Dans le chapitre 5, nous prédisons la demande future sur une journée complète à l'aide de scénérios. Nous présentons (1) une méthode de génération, en utilisant la matrice de corrélation entre les demandes horaires de la journée, et (2) une méthode de réduction, en regroupant les scénarios identifiés comme équivalents avec des métriques adaptées. Nous produisons ensuite des scénarios de la demande pour un seul appareil, en l'occurrence la recharge de la batterie d'un véhicule électrique. Cette production passe par l'analyse et la modélisation des habitudes de l'utilisateur du véhicule. Cette dernière étude a été présentée à la conférence MedPower 2018 (Gerossier et al., 2018).

Perspectives de recherche

Génération de scénarios prédictifs de la demande à plusieurs échelles. Produire des scénarios à plusieurs échelles en même temps est en réalité irréconciliable à cause des pertes du réseau. D'une part, ces pertes apparaissent quand on mesure la demande au niveau agrégé, mais, d'autre part, elles ne sont pas visibles dans chacune des demandes individuelles. De ce fait, l'objectif de produire des scénarios aux deux échelles est impossible et nécessite de définir un point de vue. Du point de vue agrégé, les pertes totales sont réparties entre chaque maison, et viennent s'ajouter à la demande propre (mesurée par le compteur intelligent). Du point de vue indiviudel, les pertes du réseau sont laissées de côté et la demande agrégée est exactement égale à la somme des demandes individuelles. Ce second point de vue pose un défi intéressant. En effet, l'effet de foisonnement rend la demande d'un voisinage moins volatile que celle d'une maison. Cet aspect doit se refléter dans les scénarios, et donc, on ne peut pas sommer simplement les scénarios individuels pour obtenir un scénario agrégé. La dépendance entre les maisons doit être prise en compte, p. ex. avec une méthode de consensus entre les scénarios.

Prédiction *bottom-up* de la demande électrique d'une maison par une analyse des habitudes. Plutôt que de prendre une approche déclarative pour rendre compte des habitudes d'une maison (c.-à-d. avec des sondages), l'analyse de la demande électrique est une voie encourageante, car les habitudes y sont inscrites. Pour les analyser, on peut utiliser des mesures à haute fréquence de la demande totale pour détecter la signature électrique de chaque appareil. D'autre part, on peut mettre en place une infrastructure complète pour mesurer directement la demande de chaque appareil. Dans tous les cas, transformer ces données en habitudes utilisables pour la prédiction à court terme n'est pas évident. Nous avons montré (au paragraphe 5.2) comment les utiliser dans le cas des véhicules électriques, pour obtenir des prédictions au moins aussi précises qu'avec une méthode d'intelligence artificielle (qui est aveugle aux usages). Nous pensons que la mise en place de cette méthode pour tous les appareils d'une maison permet d'obtenir des prédictions *bottom-up* précises de la demande totale de la maison. Bien que nous n'imaginons pas améliorer les performances pour la demande totale, nous anticipons clairement la flexibilité de la maison, en s'appuyant sur la demande de chaque appareil. De plus, comme les habitudes sont similaires entre

certains consommateurs, nous pouvons nous servir des scénarios d'un appareil d'un usager pour un autre et, ainsi, anticiper l'impact sur la demande (déjà enregistrée d'une maison) causé par l'ajout d'un nouvel appareil. Cela permet de se passer de la période d'entraînement des modèles statistiques et d'obtenir immédiatement des prédictions efficaces.

RÉSUMÉ

Cette thèse s'intéresse à la prévision à court terme de la demande électrique d'une maison intelligente et des réseaux de distribution. Les données mesurées par les compteurs intelligents permettent de caractériser la demande électrique à l'échelle d'une maison et de la comparer à la demande régionale, pour étudier notamment l'effet de foisonnement. Cette analyse permet de développer des modèles de prévision de cette demande. Ces modèles sont de nature statistique et font usage de méthodes d'apprentissage automatique. Un soin particulier est porté à la sélection de variables d'entrée pertinentes. Afin d'être déployés dans un environnement opérationnel, les modèles doivent faire preuve de réplicabilité : fonctionnement autonome, aptitude à s'adapter à de multiple situations, et robustesse face aux données erronées. Plusieurs produits de prévision sont développés et évalués avec plusieurs jeux de données : des prévisions probabilistes à différentes résolutions, et des scénarios journaliers de la demande. Enfin, les habitudes relatives à un usage électrique particulier, à savoir le chargement d'une batterie de véhicule électrique, sont modélisées pour produire des scénarios prédictifs de la demande de cet usage spécifique.

MOTS CLÉS

Demande électrique d'une maison; Compteur intelligent ; Maison intelligente ; Effet de foisonnement ; Réseau de distribution ; Enjeux opérationnels ; Apprentissage automatique ; Prévision à court terme ; Prévision probabiliste ; Génération de scénarios ; Réduction de scénarios ; Véhicule électrique ; Flexibilité électrique ; Smart grid.

ABSTRACT

This thesis is devoted to the short-term forecasting of electricity demand of smart homes and distribution grids. The household demand data provided by smart meters is analyzed to characterize the electricity demand at the local scale and compared to this at the regional scale, so as to examine the aggregation effect. This thorough analysis enables the designing of models that forecast the future demand. The models make use of advanced statistical tools and machine-learning techniques. The inputs are selected with special care for their relevancy to the household demand. To be deployed in an operational environment, the models must be replicable: low to no maintenance, adaptability to various situations, and robustness to the lack of data. Several demand forecasting products are developed and compared to actual datasets: probabilistic forecasts at different temporal and spatial resolutions, and daily demand scenarios. Finally, the habits related to a domestic appliance, namely the charging of an electric vehicle battery, are modeled in order to generate forecasting scenarios of the appliance demand.

KEYWORDS

Household Electricity Demand; Smart Meters; Smart Homes; Aggregation Effect; Distribution Grid; Operational Challenges; Machine Learning; Short-Term Forecasting; Probabilistic Forecasting; Scenario Generation; Scenario Reduction; Electric Vehicle; Electric Flexibility; Smart grid.