Machine Learning Approaches for Energy Distribution and Planning

Author **Emanuele Fabbiani**		Supervisor **Prof. Giuseppe De Nicolao**

PhD Thesis - Cycle XXXIII
November 2020

University of Pavia

Department of Electrical, Computer
and Biomedical Engineering

Emanuele Fabbiani
`emanuele.fabbiani@gmail.com`
Machine Learning Approaches for Energy Distribution and Planning
November 2020
Supervisor: Prof. Giuseppe De Nicolao
Cover design: Juliane Trier

University of Pavia
Department of Electrical, Computer and Biomedical Engineering
Via Ferrata 1
27100, Pavia, Italy

English Summary

The shift towards more sustainable energy generation, transportation, and storage will be a major challenge in the next decades. Following the global trend, both academic and industrial communities are exploiting all the available tools to facilitate the transition. Machine learning is undoubtedly one such tool: substantial advancements in the last years enabled its application to several aspects of energy production and management.

We selected two problems that can be addressed with machine learning. In collaboration with A2A, the third largest Italian utility, we studied the prediction of natural gas demand; with the Ecole Polytechnique Fédérale de Lausanne, we tackled the identification of the topology and the electrical parameters of distribution power networks.

Both topics have deep practical implications.

As nations are decommissioning coal and oil plants, natural gas becomes the ideal candidate to complement renewable yet intermittent power sources. Moreover, natural gas covers a relevant portion of the energy consumption of residential and industrial buildings. The accurate prediction of the demand can make both transportation and storage more efficient, reducing environmental and financial costs.

As the electrification of transportation and domestic heating gains traction, power networks are put under heavy stress. Moreover, the bidirectional power flows created by distributed generation must be carefully managed. New paradigms, such as microgrids and smart grids, are set to replace the current infrastructure. Yet, the complex control algorithms required by such designs require complete knowledge of the network structure.

We deal with the prediction of residential, industrial and thermoelectric gas demand at country level. We present a comprehensive explorative study, which lays the foundation for feature selection and engineering. We then cast a regression problem and compare several base models, highlighting the strengths and weaknesses of each one. For the first time, we propose to apply ensembling, showing how it yields more accurate predictors. Finally, we design a novel model for the influence of weather forecasting errors on the accuracy of residential gas demand predictors, and we demonstrate its effectiveness with experimental evidence.

We propose to solve the identification of distribution networks by means of a novel procedure, complementing an online estimation algorithm with a sequential design of experiment. The approach has two main advantages with respect to traditional methods: it exploits controllable generators to maximize the information content of the samples, and it can seamlessly adapt to changes in topology, which are especially frequent in microgrids. The effectiveness of the proposed approach is substantiated by simulations on standard testbeds.

With respect to both topics, throughout the thesis we highlight the concrete industrial applications of our work and provide directions for future developments.

Italian Summary

La transizione verso un paradigma più sostenibile di generazione, trasporto e stoccaggio dell'energia sarà una delle sfide più critiche dei prossimi decenni. Seguendo le tendenze globali, sia l'accademia che l'industia stanno sfruttando tutti gli strumenti a loro disposizione per facilitare e accelerare tale processo. Il machine learning è uno di tali strumenti: negli ultimi anni, numerose e rilevanti innovazioni hanno portato ad un numero sempre crescente di applicazioni, che ormai comprendono ogni aspetto della produzione e del trasporto dell'energia.

Abbiamo scelto di investigare due problemi che ben si prestano ad essere risolti con tecniche di machine learning: da un lato, in collaborazione con A2A, la terza utility italiana, abbiamo studiato la previsione della domanda nazionale di gas natuale; dall'altro, in collaborazione con l'Ecole Polytechnique Fédérale de Lausanne, abbiamo affrontato l'identificazione della topologia e dei parametri delle reti elettriche di distribuzione.

Entrambi gli ambiti offrono immediate applicazioni. Diverse nazioni – inclusa l'Italia – pianificano di dismettere i generatori a carbone o olio combustibile: gli impianti a gas naturale diventano quindi gli ideali candidati a complementare fonti rinnovabili intermittenti. Inoltre, il gas naturale copre attualemente una larga porzione del fabbisogno primario dei complessi industriali e residenziali. Previsioni accurate della domanda costituiscono un elemento fondamentale nei processi delle utility e dei gestori di rete e promettono di rendere più efficiente il trasporto e lo stoccaggio, diminuendo così i costi finanziari e ambientali. L'identificazione delle reti elettriche, invece, è necessaria agli algoritmi di controllo della generazione distributa, a loro volta moduli fondamentali in strutture ad alta efficienza e basso impatto ambientale, come microgrid e smart grid.

In questo lavoro, affrontiamo la previsione della domanda italiana di gas naturale ad uso residenziale, industriale e termoelettrico. La nostra discussione si apre con un'analisi esplorativa, tesa a guidare la scelta e la creazione delle variabili. Prosegue quindi con la trasformazione della previsione in un problema di regressione e la comparazione di diversi modelli base. Per la prima volta, applichiamo poi in questo ambito la tecnica dell'ensembling e dimostriamo come questa produca predittori più accurati e robusti. Infine, proponiamo un originale modello probabilistico per l'impatto dell'inaccuratezza delle previsioni meteo sull'errore nella previsione della domanda residenziale.

Per quanto concerne l'identificazione delle reti elettriche di distribuzione, proponiamo una nuova procedura, che complementa un algoritmo di apprendimento online con la tecnica del design of experiment. Tale approccio ha due vantaggi rispetto ai metodi esistenti: sfrutta i generatori controllabili per massimizzare l'informazione contenuta nelle misure, senza tuttavia compromettere l'operatività o la sicurezza della rete, ed è capace di adattare la stima a cambiamenti di configurazione, molto comuni nelle microgrid. L'efficacia del metodo viene comprovata da numerose simulazioni numeriche.

Infine, nel corso di tutta la tesi, sottolineamo le applicazioni concrete dei nostri contributi e forniamo indicazioni per possibili sviluppi futuri.

Acknowledgments

The first acknowledgment must go to you, my reader.

I hope something in these pages will catch your attention, teach you something, inspire you, or just make you wonder. I tried to be as rigorous as I must, as clear as I can, as interesting as everyone should be. If I failed in any of these goals, please accept my apologies and enjoy the reading as much as you can.

This work would not have been possible without my students. Quite literally. Life is about passions, thank you all for sharing mine.

But my PhD as a whole would not have started without stubborn relatives, a couple of inspiring bosses and a dynasty of wise professors: thank you for believing in me, even when you had no reason to.

The last three years would have been boring without fellow students and researches: in the homeland and abroad, you transformed a job into an exciting experience – and I know it is all but easy, from experimental evidence.

The same transformation happened in another context, where friends became colleagues, and then colleagues friends. Someone said, someone says, it was a mistake – surely, likely, maybe – still, one worth making.

Friends can be found everywhere, not just in tidy labs or chaotic offices. They can appear on the door of a new flat, on a dark beach, on a sandy field, on a crowded staircase, in a noisy classroom, on the floor of an airport, next to a fire in a wood, or around a wooden table. Your words, your actions, you, are worth to me much more than I have ever admitted.

Finally, my gratitude, appreciation, and much more goes to one of the few certainties sealed in the last years. I will not write what I have never said, for I owe you three lines and not just three words. From the grass to the marble, from the water to the snow, thank you for everything you did with me. Thank you for everything you did for me. I will not forget.

Curriculum Vitae

Education

PhD Diploma 2017 - 2021

Identification and Control of Dynamic Systems Laboratory, University of Pavia, Italy
Thesis: Machine Learning Approaches for Energy Distribution and Planning
Supervisor: prof. Giuseppe De Nicolao

Master Degree 2015 - 2017

University of Pavia, Italy
Diploma of Computer Engineering, 110 cum laude / 110
Thesis: Applications of the Lyapunov Equation to Derivative Pricing
Supervisors: prof. Giuseppe De Nicolao, Andrea Marziali, PhD

Bachelor Degree 2012 - 2015

University of Pavia, Italy
Diploma of Electronic and Computer Engineering, 110 cum laude / 110
Thesis: Automatic Detection and Classification of Defects on Silicon Wafers
Supervisors: prof. Giuseppe De Nicolao, Simone Pampuri, PhD

Work Experience

Co-Founder and Chief Data Scientist Jan 2018 - present

xtream, Milan, Italy
Main topics: forecasting of natural gas demand, power load, wind power generation, and financial indices; creation of scenarios for power markets; application of deep reinforcement learning to conversational agents.

Data Scientist Jan 2017 - Dec 2017

Techedge, Milan, Italy – part-time position
Main topics: forecasting in power markets; detection of market abuses; predictive maintenance.

Data Scientist Mar 2017 - May 2017

Techedge, Madrid, Spain

Main topics: data modelling, processeing and visualization for petrochemical plants; predictive maintenance.

Software Developer Oct 2015 - Jul 2016
Techedge, Milan, Italy – part-time position
Main topics: data visualization; predictive maintenance.

Data Scientist Oct 2015 - Dec 2016
Statwolf, Pavia, Italy – part-time position
Main topics: analysis of market surveys; sentiment analysis on reviews of household appliances.

Data Scientist Jul 2015 - Sep 2015
Statwolf, Dublin, Ireland – internship
Main topics: fault detection for silicon wafer, modelling of the effect of environmental variables on semiconductor manufacturing.

Teaching

Lecturer Dec 2020
Department of Neurological and Behavioural Sciences, University of Pavia, Italy
Introduction to Python for Data Analysis.

Lecturer Nov 2020 - Jan 2021
Almo Collegio Borromeo, Pavia, Italy
Introduction to R programming.

Lecturer Jul 2019
Department of Neurological and Behavioural Sciences, University of Pavia, Italy
Python for Data Science.

Academic Tutor Oct 2013 - Jul 2016
University of Pavia, Italy
Calculus I, Calculus II, Physics I, Physics II, Computer Science, and Electronics.

Student Supervision

A Survey on the Application of Deep Leaning to Natural Gas Demand Forecasting by Erik Turricelli, University of Pavia, bachelor project.

AI Gamer: Deep Reinforcement Learning for Arcade Games by Barbier De La Serre Nicolas Marc Eugène, Tchatat Njieyep Victoire Sephora, Nicoletti Francesca Paola, and Raita Omar, École polytechnique fédérale de Lausanne, bachelor project.

Publications

Emanuele Fabbiani, Andrea Marziali, and Giuseppe De Nicolao (2021). "Fast calibration of two-factor models for energy option pricing". Applied Stochastic Models in Business and Industry [1].

Emanuele Fabbiani, Andrea Marziali, and Giuseppe De Nicolao (2020). "vanilla-option-pricing: Pricing and market calibration for options on energy commodities". Software Impacts, Vol. 6, pp. 100043 [2].

Emanuele Fabbiani, Andrea Marziali, and Giuseppe De Nicolao (2021). "Forecasting residential gas demand: machine learning approaches and seasonal role of temperature forecasts". International Journal of Oil, Gas and Coal Technology, Vol. 6, No. 2, pp. 202-224 [3].

Andrea Marziali, Emanuele Fabbiani, and Giuseppe De Nicolao (2021). "Ensembling methods for countrywide short term forecasting of gas demand". International Journal of Oil, Gas and Coal Technology, Vol. 6, No. 2, pp. 184-201 [4].

Maurizio Polano, Emanuele Fabbiani, Eva Adreuzzi, Federica D. Cintio, Luca Bedon, Davide Gentilini, Maurizio Mongiat, Tamara Ius, Mauro Arcicasa, Miran Skrap, Michele Dal Bo, Giuseppe Toffoli (2021). "A New Epigenetic Model to Stratify Glioma Patients According to Their Immunosuppressive State". Cells, Vol. 10 No. 3, pp. 576 [5].

Preprints

Emanuele Fabbiani, Pulkit Nahata, Giuseppe De Nicolao, and Giancarlo Ferrari-Trecate (2020). "Identification of AC Networks via Online Learning". arXiv preprint arXiv:2003.06210 . Submitted to IEEE Transactions on Control System Technologies and currently under review [6].

Jean-Sébastien Brouillon, Emanuele Fabbiani, Pulkit Nahata, Florian Dörfler, and Giancarlo Ferrari-Trecate (2021). "Bayesian Methods for the Identification of Distribution Networks". Submitted to IEEE Conference on Decision and Control and currently under review.

Jean-Sébastien Brouillon, Emanuele Fabbiani, Pulkit Nahata, Florian Dörfler, and Giancarlo Ferrari-Trecate (2021). "Bayesian Error-in-Variables Models for the Identification of Power Networks". Submitted to IEEE Transactions on Control System Technologies and currently under review.

Conference Talks

Emanuele Fabbiani, *Fast Calibration of Two-Factor Models for Energy Option Pricing*, Energy Finance Italia III, Pescara, Italy, 15-16 February 2018 [7].

Emanuele Fabbiani, *Short-Term Forecasting of Italian Gas Demand*, Energy Finance Italia IV, Milan, Italy, 4-5 February 2019 [8].

Emanuele Fabbiani, *Academic and Buisness Research in AI for Energy*, Applied Machine Learning Days, Lausanne, Switzerland, 25-29 January 2020 [9].

Conference Organization

AI and Sustainable Energy, Applied Machine Learning Days, Lausanne, Switzerland, 26 April 2021, track organizer and host.

Review activity

IEEE Conference on Decision and Control, 2020.

Applied Machine Learning Days, 2021.

Contents

Introduction 1

1.1 The Dark Side of Green Energy

On 10th September 2018, governor Edmund J. Brown signed the Clean Energy Bill, committing California to achieve carbon-free power generation by 2045. It was a bold move by the old politician, one in clear contrast with the withdrawal from the Paris climate agreement ordered by the US president Donald Trump the year before. The law was criticized by many: Pacific Gas and Electric, a major utility, deemed it "not affordable nor sustainable" [10]. But further, even more impactful changes were yet to come.

On 1st January 2020, California became the first state in the USA to mandate photovoltaic devices on new residential buildings. The law was once again disputed: for someone, it was a necessary step to ditch fossil fuels, for others, a smart way to inflate the price of new houses.

As controversial as it was, the energy policy adopted by the Golden State was undoubtedly effective. According to official accounts, the share of renewable generation grew from 12% in 2009 to an estimated 36% in 2019, overperforming the target of 33% set for 2020 (Fig. 1.1). The figures become even more impressive when considering that large hydroelectric facilities are not included in the renewable share. All the carbon-free sources - solar, wind, geothermal, hydroelectric and nuclear - covered an estimated 63% of the retail demand in 2019 [11].

The effort to achieve a complete transition towards sustainable energy gained California the admiration of many but also the interest of researchers and engineers from other nations. As several countries intend to follow similar paths, the problems and solutions found in the Golden State may prove valuable for the whole world.

Indeed, California faced several challenges caused by its reforms.

The daily pattern of power demand is quite repetitive and predictable: it peaks in the morning and in the evening, while hitting a low in the middle of the day. This pattern, known as the "camel" curve, results from the daily productive cycle and is very stable in its shape. As electric energy is difficult and expensive to store at scale,

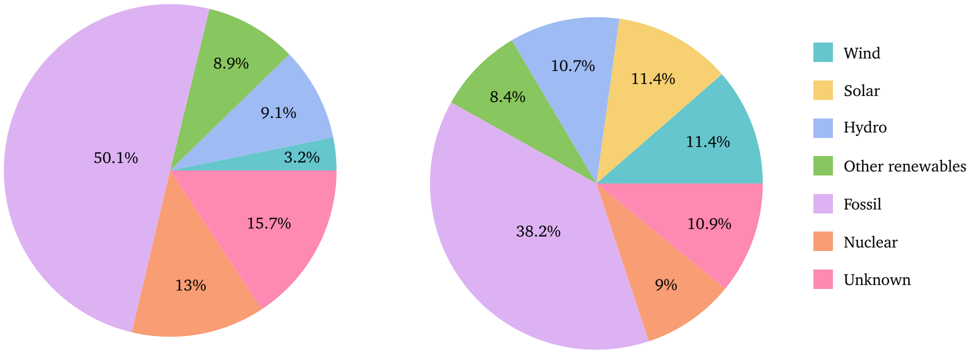


Figure 1.1. Energy mix in California in 2009 (left) and in 2018 (right). Solar increased by more than 11% while wind grew by 8.2%.

during the last decades of the XX century and the first of the XXI century, power operators optimized the generation to meet the pattern of the demand.

However, in the 2010s, technological advancements and an increased sensibility for environmental matters led to strong incentives for renewable sources, mostly wind and photovoltaic. The latter was particularly popular, as solar panel became affordable to individual households. The case of California is exemplary: due to a favourable climate and accommodating fiscal policies, photovoltaic generation grew from a negligible share to more than a tenth of the overall demand in just nine years (Fig. 1.1).

Such a change was unprecedented and carried several issues. With a bit of simplification, it is safe to assume that, before the advent of cheap photovoltaic devices, nodes in the power grid were sharply divided between generators, usually large and controllable plants, and consumers. Solar panels enabled consumers to turn into intermittent generators, thus complicating grid operations and control.

However, a different problem became more and more apparent in the early 2010s. Solar and wind production is unreliable and uncontrollable: even worse, the typical generation pattern for photovoltaic panels is very different from the camel curve of demand. Solar production peaks in the early afternoon, when the demand is relatively low, and drops in the morning and in the evening, when the demand is high. The difference between the demand and the renewable generation thus assumes a very characteristic shape: flat during the night, slightly increasing in the morning, a convex hole in the afternoon and a sudden rise followed by a peak in the evening: the "duck" curve. The term was coined in 2012 by the California Independent System Operator (CAISO) to denote the pattern depicted in Fig. 1.2 and became since then topic for debate and research [12, 13].

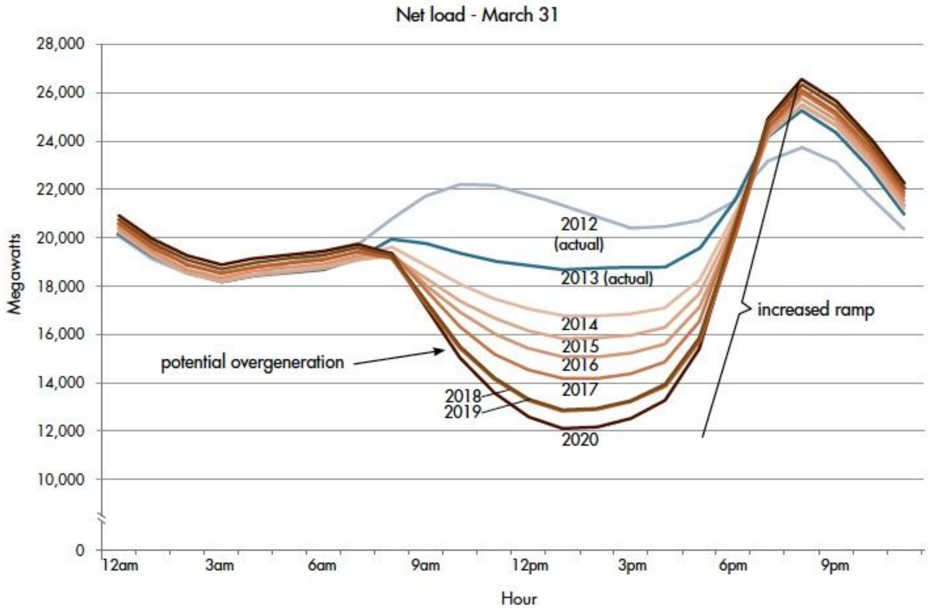


Figure 1.2. A representation of the duck curve [12], that gives the difference between the demand and the generation of renewable energy. The two main issues are highlighted: the potential overproduction during the afternoon and the steep ramp in the evening.

The duck curve highlights two major problems: first, in the middle of the afternoon, the output of photovoltaic devices may result in overproduction, with excess energy that cannot be consumed; second, in the evening, the ramp may became so steep that traditional dispatchable sources, such as gas-fueled thermoelectric plants, struggle to keep up.

These problems are far from being mere theoretical speculation: in California, power curtailment of photovoltaic and wind generation steadily increased from 2014 to 2019. In 2019, 4% of the total solar and wind production was curtailed, the equivalent of the annual need of 400,000 households (Fig. 1.3) [14]. To address the issue, in recent years utilities are turning to large-scale battery packs, paving the way for what is expected to be one of the most relevant technological challenges of the next decade: cheap and persistent energy storage [15, 16].

California is an interesting case study for a process that will likely involve many other countries, and provides valuable lessons. The energy system is complex and any change may result in important yet unforeseen consequences. It is thus critical to apply all the tools of modern research to gain insights and to support the modifications that will happen in the near future.

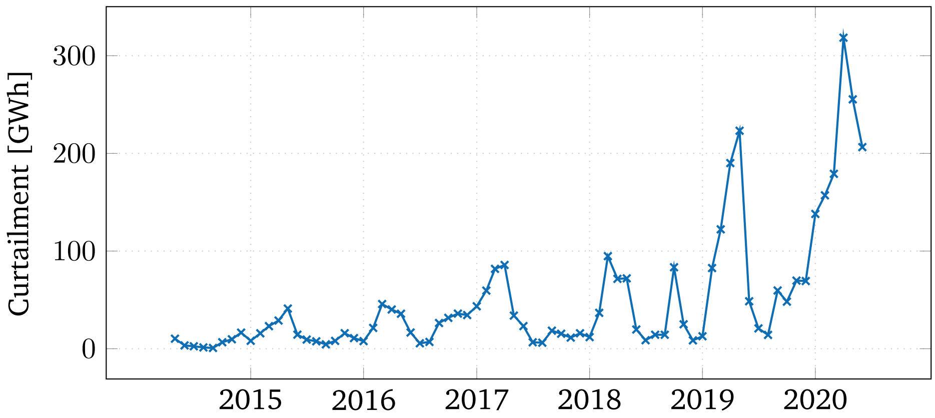


Figure 1.3. Power curtailment of photovoltaic and wind generation [15]. The spike in Spring 2020 is partly due to the drop in industrial demand caused by the COVID-19 shelter-in-place orders.

Machine Learning is one such tool. Its adoption in the energy sector is no novelty and has already transformed the field. However, advancements in research and the availability of ever-increasing computation capacity are enabling new applications. Attenuating the impact of intermittent renewable power production, reducing inefficiencies in gas storage and transportation, managing the charge and discharge of large battery packs, estimating the topology of distribution power network in order to allow for distributed solar generation, nudging users towards more responsible energy saving are but a few examples.

These speculations provide the main motivation to this work. Our intention is to test and experiment applications of machine learning to the energy sector, and possibly verify their effectiveness in real or realistic setups.

The first step towards such goal is to analyze where machine learning has been applied before and where it may deliver the most value.

1.2 Machine Learning Applications to Energy

The term "energy" can assume several different meanings. Apart from the well-known physical connotation, we will use it to indicate the complex system of devices, infrastructure, technology and organizations required to harvest power from the environment and deliver it to end users. In this view, "energy" does not only mean "electric energy", as it encompasses other vectors, such as natural gas, oil, and hydrogen, as well as their support infrastructure.

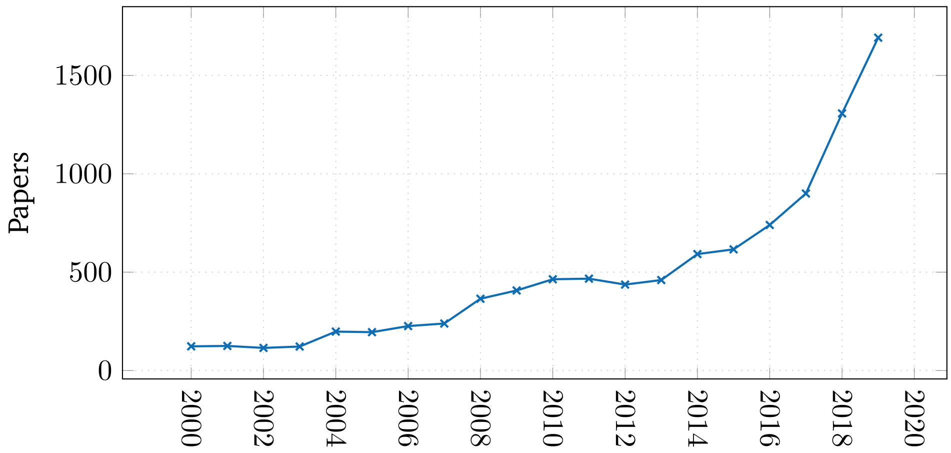


Figure 1.4. Papers dealing with machine learning in energy and power systems. Data were extracted from Scopus using the query: ("machine learning" OR "deep learning" OR "forecasting" OR "statistical learning") AND ("energy systems" OR "power systems")

Statistical methods were successfully applied in energy long before Machine Learning (ML) became mainstream. An early review about power load forecasting dates back to 1972 [17]. Since 2010, the interest of both the academic and industrial community skyrocketed, leading to a fast increase in the papers published every year. According to the survey by Mosavi et al., academic works regarding applications of machine learning to energy problems were a few tens per year until 2008, surpassed one hundred in 2010 and numbered about six hundreds in 2018 [18]. A more general query on the Scopus database reveals that articles about machine learning and deep learning for energy and power systems numbered more than 1,500 in 2019 (Fig. 1.4).

To help scholars and practitioners keep up with the volume of publications, comprehensive surveys were produced. Our analysis is partly based on the work by Mosavi et al. [18], but takes a different approach. Instead of reviewing the main machine learning methods and the proposed usages for each of them, we will group the most common fields of application of machine learning and briefly discuss the proposed approaches.

1.2.1 Time Series Forecasting

Offer, demand, and price of different forms of energy are often the target of extensive forecasting studies.

As far as the electric energy is considered, generation was not worth predicting until the 2000s. In the XX century, thermoelectric, hydroelectric and nuclear were the only widespread power sources and their output was known and, within appropriate constraints, controllable. As electric energy is notoriously difficult and expensive to store at scale, the issue was forecasting the demand, in order to plan for economically optimal generation and guarantee the power balance of the grid.

Classical time series models like multivariate regression, ARMAX and exponential smoothing were applied first [19]. Since the 2000s, different machine learning techniques gained traction: support vector machines (SVM), random forest (RF) and artificial neural networks (ANN) were extensively adopted and compared to classical methods [20, 21]. More recently, of particular importance were studies on different architectures of ANNs as well as on hybrid and ensemble models [22]. The field of application also broadened: the advent of microgirds called for small-scale, localized forecasting [23]. The introduction of smart meters paved the way for the analysis, forecasting, and optimization of the demand of individual residential units [24].

In the late 2000s and in the 2010s, the increase in penetration of wind and solar generation extended the application of forecasting models to power production [25, 26]. For both solar and wind, accurate long-term forecasting (months or years ahead) is impossible at fine granularity (hour or day), while short-term forecasting (hours or days ahead) is of paramount importance for grid balancing and control. Due to their effectiveness in capturing complex, non-linear patterns, neural networks became a popular prediction tool. In particular, recurrent architectures like the long-short term memory (LSTM) are currently considered the state of the art [27, 28]. Both power sources are highly sensitive to weather conditions: models were then adopted also to forecast wind speed and solar radiation. Different from load models, which are mostly developed at country or regional level, production models are usually fit for a specific farm.

While production and load forecasting are often motivated by grid safety and stability, price forecasting is mostly driven by economics. Short-term forecasting, hour- or day-ahead, is important for utilities and trading firms, while long-term predictions move strategic decisions of power operators and energy-intensive manufacturing companies, like steel mills and chemical plants.

Forecasting prices is generally considered more complex than predicting demand and generation, as prices are influenced by both generation and demand as well as by many other factors, such as the cost of fuels (oil and natural gas) and social and political events. Moreover, the price of electricity is known to feature spikes

[29] and to be subject to other phenomena common to energy commodities, like mean reversion [1]. Several different approaches to power price forecasting were proposed: in addition to data-driven methods, like time series models and neural networks, fundamental approaches were proven effective [29].

The interest towards demand, load and price forecasting increased considerably after the liberalization of the energy markets carried out by most European countries in the last decades of the XX century. With the introduction of free markets, accurate price forecasting translated into a competitive advantage and, after the creation of complex auction systems for power generation and pipe capacity, predictions of demand and generation also acquired more and more economic value.

Forecasting of production, demand and prices also interested other energy vectors, like natural gas. However, less attention was paid to them, as hydrocarbons can be easily and effectively stored, thus removing the requirement for a perfect balance between production and consumption. Still, accurate prediction of gas demand can be useful in planning efficient and profitable storage and transportation of resources. A more throughout analysis is deferred to Chapter 2.

1.2.2 Anomaly Detection

Closely related to time series forecasting is anomaly detection. Highlighting abnormal samples or patterns in a series can be useful in many contexts, notably in fault detection and efficiency optimization. Fault detection is particularly important for both power lines and natural gas infrastructure, whose failure may lead to tragic consequences, ranging from economic damage to the loss of human lives. Efficiency optimization has gained popularity in recent years: abnormal patterns may result from accidental or involuntary usage of equipment in a building and proper monitoring and reporting may nudge the users towards a more aware and energy-efficient behaviour.

Several machine learning methods were proposed, mostly relying on a forecasting model trained on the "normal" behaviour and a procedure to detect significant deviations from it. To detect anomalies in the flow of natural gas in small networks, nearest neighbours and a local regression method based on weather variables, complemented with a threshold-based outlier detection method were proposed [30]. A similar structure, involving an hybrid ARIMA-ANN model and a threshold-based detection method was applied to gas demand in buildings [31]. One step further is a probabilistic approach based on a Bayesian maximum likelihood classifier presented for anomaly detection in gas demand, which did not provide just the binary classification "outlier" or "normal", but also the estimated probability of each class [32]. Finally, industrial applications and use cases for anomaly detection were also presented: one of them involved the remote management of schools, where abnormal electrical load may indicate negligence in the application of energy-saving policies [33].

1.2.3 Signal disaggregation

A classification problem based on time series data is signal disaggregation, which recently assumed particular relevance for power load analysis. The increasing penetration of smart meters made it possible to record the demand of residential units and to identify its components. The knowledge of the set of active appliances in a house enables proactive advisory to the end user, promoting energy and cost saving.

Several solutions to the problem were suggested, ranging from optimization techniques, such as quadratic and linear integer programming, to statistical approaches based on hidden Markov models. In recent years, however, more and more attention was paid to machine learning.

Support vector machines, adaboost, fuzzy systems, and other learning algorithms were adopted, but many of them require complex feature engineering steps [34]. Artificial neural networks were then proposed as an end-to-end approach. Signal disaggregation requires both a short-term memory, to understand which patterns are appearing, and a long-term one, to compare current patterns with the ones observed in the past. Such requirements well suit both recurrent architectures, like LSTM, and convolutional architectures [35].

1.2.4 System Identification

Also reliant on time series data, but in a different way, is system identification. In the last decades, system identification was applied to several energy devices, such as batteries [36], or even entire buildings. In recent years, however, a new need for identification emerged in the field of power systems.

The increased penetration of distributed renewable generation called for more and more complex control systems to be applied to distribution networks, the portions of the grid connecting substations to end users. Such algorithms mostly require as an input the exact topology and line parameters, which are seldom known. Hence the requirement for data-driven methods to infer such information from measurements. An in-depth analysis of the topic is presented in Chapter 3.

1.2.5 Control

Besides system identification, the control algorithms themselves were transformed by the rise of machine learning. Data-driven methods were applied in several filed, including control of microgrids, buildings, and batteries, due to their effectiveness in coping with uncertainty in the model and in the parameters, and their ability to adjust to drifts.

Among data-driven control methods in energy, an important role was tributed to deep reinforcement learning. Different flavours of deep reinforcement learning algorithms were applied to problems including energy management, operational control of networks, demand response, and optimization of trading in energy markets [37].

At the present moment, however, the application of deep reinforcement learning to energy is mostly limited to academic research. The lack of theoretical guarantees on performance and explainability of the algorithms are preventing a wide adoption in the industry, although several studies are aiming at improving our understanding of the method.

1.3 Dissertation Topics and Outline

The survey proposed in Section 1.2 shows that applications of machine learning to energy are numerous and diverse. We choose to focus on two of them: time series forecasting and system identification.

The interest in time series forecasting derives from the collaboration with A2A, the third largest Italian utility. A2A provided not only a strong motivation to approach the problem, but, more importantly, a real dataset and a real industrial case to test our methods.

We focus on the analysis and forecasting of natural gas demand in Italy. While natural gas may not seem the most sustainable energy source, it is arguably the cleanest among fossil fuels and it has the potential to complement intermittent renewable generation for years to come. Moreover, the electrification of industrial plants and domestic heating will take long time. Hence the interest in studying tools to make natural gas storage and delivery more efficient.

Despite gas being as important as electricity as an energy vector, the study of natural gas demand received little attention from the academic community. We try to fill the gap, providing an insightful analysis and comparing several forecasting methods. To this end, the Market and Price Forecasting Department in A2A provided extensive support and collaborated in the research in the context of the PhD program of Andrea Marziali, currently head of the department.

The choice of system identification as second topic is motivated by the willingness to contribute and support the effort of the academic community for the transition to clean energy. In particular, the difficulty in identifying the topology and line parameters of a power network is a major issue for the diffusion of smart grids and distributed generation.

The Dependable Control and Decision Group of the Automatic Control Laboratory at the École Polytechnique Fédérale de Lausanne (EPFL) were valuable partners in the research, thanks to their expertise in power systems and smart grids. In particular, Prof. Giancarlo Ferrari-Trecate, Dr. Pulkit Nahata, and Jean-Sébastien Brouillon participated in the research described in Chapter 3.

Reflecting the two topics, this work is divided into two main parts.

Chapter 2 introduces the problem of natural gas forecasting, reviews the related literature and presents our main contributions on the subject. It includes a detailed analysis of the daily series of natural gas demand and the description of the proposed forecasting methods. Feature selection and engineering are discussed and several models are compared. Following recent trends, ensemble approaches are also introduced and tested: the experimental evidence shows that they systematically outperform other methods. A discussion on the influence of the weather forecasting errors on the accuracy of gas demand prediction is carried out, and a novel error propagation model is proposed. Finally, industrial applications of the research are described and discussed.

Chapter 3 starts from a description of the structure of the power grid to motivate the need for accurate identification algorithms. Existing approaches are reviewed and the need for new online methods is shown. Then, our main contributions are proposed: a recursive least squares algorithm is derived and complemented with an online design-of-experiment procedure. The method is validated via simulations on standard testbeds.

1.4 Contributions

The contributions of this work are also split between the two main topics. For what concerns gas demand forecasting, we:

- perform a throughout and insightful analysis of the time series of the Italian daily gas demand, divided into the residential, industrial and thermoelectric sectors, discussing the seasonality and the role of exogenous factors;
- provide guidance about feature selection and engineering, discussing in particular the role of temperature, seasonality and holidays;
- perform a comparison of several different forecasting models and highlight strengths ad weaknesses of each one;
- propose the application of ensembling methods, which, though already popular for power load forecasting, were never tested on the task of gas demand prediction;
- develop a novel model to propagate the effect of the error in temperature prediction to residential gas demand forecasting.

Moreover, our models were integrated into the A2A IT platform and are now adopted to support the daily operations of the utility.

As far as the identification of power distribution network is concerned, we focus on the estimation of the admittance matrix of the grid, which encodes information for both the topology and line parameters. In particular, we:

- propose a parameterization of the admittance matrix which does away with redundant parameters and is suitable both in presence and in absence of slack capacitors;
- propose a novel algorithm based on optimal design of experiment to exploit the controllable generators in the grid;
- describe a recursive algorithm capable of following the changes in time of the grid topology.

1.5 Publications and Presentations

The following works were conceived during the doctoral program, although only items 3, 4, and 6 deal with topics discussed in this dissertation:

1. Emanuele Fabbiani, Andrea Marziali, and Giuseppe De Nicolao (2021). "Fast calibration of two-factor models for energy option pricing". Applied Stochastic Models in Business and Industry [1].

2. Emanuele Fabbiani, Andrea Marziali, and Giuseppe De Nicolao (2020). "vanilla-option-pricing: Pricing and market calibration for options on energy commodities". Software Impacts, Vol. 6, pp. 100043 [2].

3. Emanuele Fabbiani, Andrea Marziali, and Giuseppe De Nicolao (2021). "Forecasting residential gas demand: machine learning approaches and seasonal role of temperature forecasts". International Journal of Oil, Gas and Coal Technology, Vol. 6, No. 2, pp. 202-224 [3].

4. Andrea Marziali, Emanuele Fabbiani, and Giuseppe De Nicolao (2021). "Ensembling methods for countrywide short term forecasting of gas demand". International Journal of Oil, Gas and Coal Technology, Vol. 6, No. 2, pp. 184-201 [4].

5. Maurizio Polano, Emanuele Fabbiani, Eva Adreuzzi, Federica D. Cintio, Luca Bedon, Davide Gentilini, Maurizio Mongiat, Tamara Ius, Mauro Arcicasa, Miran Skrap, Michele Dal Bo, Giuseppe Toffoli (2021). "A New Epigenetic Model to Stratify Glioma Patients According to Their Immunosuppressive State". Cells, Vol. 10 No. 3, pp. 576 [5].

6. Emanuele Fabbiani, Pulkit Nahata, Giuseppe De Nicolao, and Giancarlo Ferrari-Trecate (2020). "Identification of AC Networks via Online Learning". arXiv preprint arXiv:2003.06210 . Submitted to IEEE Transactions on Control System Technologies and currently under review [6].

7. Jean-Sébastien Brouillon, Emanuele Fabbiani, Pulkit Nahata, Florian Dörfler, and Giancarlo Ferrari-Trecate (2021). "Bayesian Methods for the Identification of Distribution Networks". Submitted to IEEE Conference on Decision and Control and currently under review.

8. Jean-Sébastien Brouillon, Emanuele Fabbiani, Pulkit Nahata, Florian Dörfler, and Giancarlo Ferrari-Trecate (2021). "Bayesian Error-in-Variables Models for

the Identification of Power Networks". Submitted to IEEE Transactions on Control System Technologies and currently under review.

The following presentations were held at international conferences during the doctoral program:

1. Emanuele Fabbiani, "Fast Calibration of Two-Factor Models for Energy Option Pricing", Energy Finance Italy III, Pescara, Italy 15-16 February 2018 [7].
2. Emanuele Fabbiani, "Short-Term Forecasting of Italian Gas Demand", Energy Finance Italy IV, Milan, Italy, 4-5 February 2019 [8].
3. Emanuele Fabbiani, "Academic and Buisness Research in AI for Energy", Applied Machine Learning Days, Lausanne, Switzerland, 25-29 January 2020 [9].

1.6 Notation

We define here the notation that will be used throughout the work, in both Chapter 2 and Chapter 3.

1.6.1 Sets, Vectors, and Functions

$\mathbb{N}$, $\mathbb{R}$ and $\mathbb{C}$ are respectively the sets of natural, real and complex numbers; $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x \geq 0\}$ is the set of non-negative real numbers; $\mathbf{j} = \sqrt{-1}$ is the imaginary unit. Scalars are denoted by lowercase letters, while vectors by lowercase bold letters. If $\boldsymbol{x} \in \mathbb{R}^n$, $x_i \in \mathbb{R}$ denotes the element of $\boldsymbol{x}$ in the i-th position, $i = 1 \dots n$. The ℓ_1 and ℓ_2 norms of a vector $\boldsymbol{x}$ are represented by $\|\boldsymbol{x}\|_1$ and either $\|\boldsymbol{x}\|_2$ or $\|\boldsymbol{x}\|$, respectively. Given $\boldsymbol{x} \in \mathbb{C}^n$, $\overline{\boldsymbol{x}}$ is its complex conjugate taken element-wise and $[\boldsymbol{x}]$ the associated diagonal matrix of order n. Throughout, $\mathbf{1}_n$ and $\mathbf{0}_n$ are n-dimensional vectors of all ones and zeros.

A time series is represented by a vector $\boldsymbol{x} \in \mathbb{R}^n$ of samples x_t, where t is the timestamp of each sample.

Matrices are represented by capital letters: let $X \in \mathbb{R}^{n \times m}$, then $x_{ij} \in \mathbb{R}$ is the element in row i and column j. The trace of the matrix X is denoted by $\operatorname{tr}(X)$, its determinant by $\det(X)$, the Kronecker product between matrices by $\otimes$. For an matrix $A \in \mathbb{C}^{n \times m}$, $A^\top$ denotes its transpose, A^H its Hermitian (complex conjugate) transpose, $\boldsymbol{a}_{i\cdot}$ its i^{th} column vector, and $\operatorname{vec}(A) = [\boldsymbol{a}\cdot_1^\top, \dots, \boldsymbol{a}_{n\cdot}^\top]^\top$ the mn-dimensional stacked column vector. If $A \in \mathbb{C}^{n\times n}$ is a square matrix, then $\mathrm{vech}(A)$ is the $n(n+1)/2$-dimensional vector obtained by eliminating all supradiagonal elements of A from $\mathrm{vec}(A)$, and $\mathrm{ve}(A)$ is the $n(n-1)/2$-dimensional vector obtained by removing diagonal elements from $-\mathrm{vech}(A)$. A positive definite matrix A and a positive semidefinite matrix B verify $A \succ 0$ and $B \succeq 0$. The Frobenius and the maximum norm of a matrix A are denoted by $\|A\|_{\mathrm{F}}$ and $\|A\|_{\max}$. The symbols $\mathbb{I}_n$ and $\mathbb{O}_{n\times m}$ represent identity and zero matrices of dimension $n \times n$ and $n \times m$. The unit vector $\boldsymbol{e}_i$, $i = 1 \dots n$ is the i^{th} column of $\mathbb{I}_n$.

Sets are represented by calligraphic capital letters, their elements, when not vectors or matrices, by lowercase letters. For a finite set $\mathcal{V}$, $|\mathcal{V}|$ denotes its cardinality. Let $f : \mathcal{X} \to \mathcal{Y}$ be a function from $\mathcal{X}$ to $\mathcal{Y}$, then $y = f(x)$, $y \in \mathcal{Y}$, represents the image of $x \in \mathcal{X}$ through f. Random variables are denoted by calligraphic letters: $\mathcal{X} \sim \mathcal{N}(\mu, \sigma^2)$ is the Gaussian random variable with expected value $\mathrm{E}[\mathcal{X}] = \mu$ and variance $\mathrm{Var}[\mathcal{X}] = \sigma^2$.

1.6.2 Algebraic Graph Theory

Graphs are also denoted by calligraphic capital letters. In case of ambiguity, the nature of the symbol will be specified in the text.

We denote by $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{W})$ an undirected connected and weighted graph, where $\mathcal{V}$ is the node set and $\mathcal{E} \subseteq (\mathcal{V} \times \mathcal{V})$ the edge set. The adjacency matrix W of order $|\mathcal{V}|$ embeds the edge weights $w_{ij} \in \mathcal{W}$ in positions such that $(i, j) \in \mathcal{E}$ and has zeros in other places. The n-order matrix $L = [W\,\mathbf{1}_{|\mathcal{V}|}] - W$ is the Laplacian matrix associated with $\mathcal{G}$. By construction, a Laplacian matrix $L \in \mathbb{C}^{n\times n}$ is such that $L\mathbf{1}_n = \mathbf{0}_n$.

2 Forecasting of the Italian Gas Demand

2.1 Role of Natural Gas in Italy

Natural gas plays a key role in the Italian energy mix and, according to the National Plan for Energy Development, its importance will not diminish in the near future [38]. Natural gas is commonly used as fuel for domestic heating and cars, powers industrial facilities and is burnt in thermoelectric power plants. Considered the cleanest among all fossil fuels, natural gas is seen as an intermediate step in the transition towards clean energy generation and an ideal complement to intermittent renewable sources.

2.1.1 Demand

Data from SNAM Rete Gas, the Italian transmission system operator (TSO), reveals that the total demand for natural gas oscillated between 85 and 62 billions of standard cubic meters (BSCM) since 2000. An increasing trend insists from 2002 to 2006, a period of economic expansion, while a sudden drop appears between 2010 and 2014, due to the combined effect of the financial crisis and the push for renewable power generation. Since 2014, the demand increased again to the levels of the early 2000s, reaching 74.23 BSCM in 2019 (Fig. 2.1).

The total gas demand (GD) is made of three main components, with negligible residuals: residential gas demand (RGD), industrial gas demand (IGD), and thermoelectric gas demand (TGD). In 2018, RGD accounted for 41.5% of the total consumption, IGD for 25.4% and TGD for the remaining 33.1% (Fig. 2.2) [39].

2.1.2 Infrastructure

In order to deliver natural gas across the country, Italy built a vast infrastructure. The nation is poor in reserves and, in 2019, domestic production covered only 6.5% of

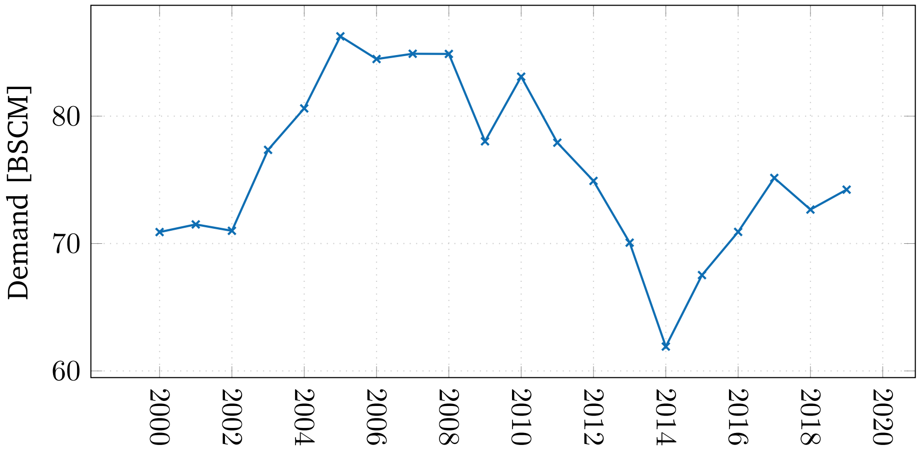


Figure 2.1. Gas demand from 2000 to 2019. The sudden drop starting in 2010 is due to the contemporary effect of the financial crisis and the incentives for renewable power generation.

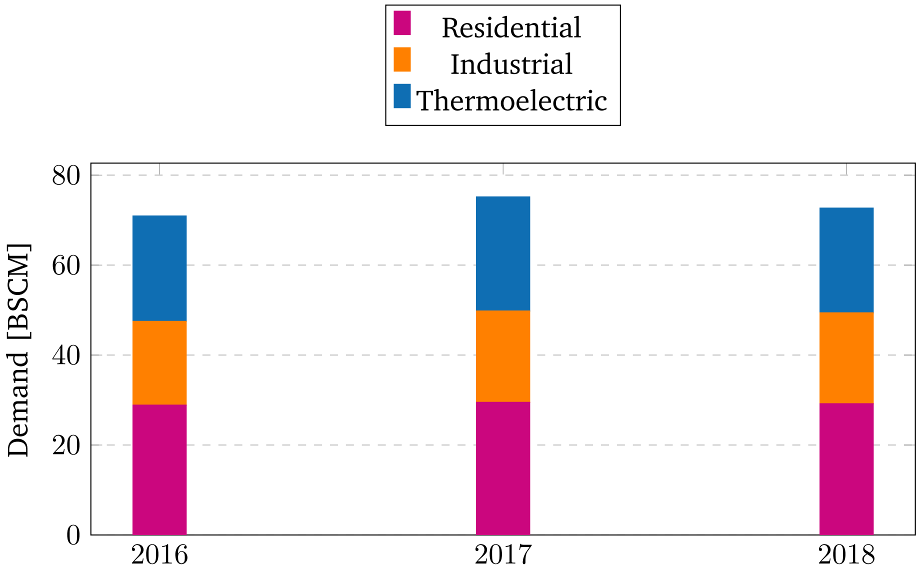


Figure 2.2. The threee main components of gas demand: residential (RGD), industrial (IGD) and thermoelectric (TGD).

the demand. The remaining 93.5% was satisfied by import: in 2018, Russia supplied 45%, Algeria 32% and Libya 8% of the requirement. Most of the gas enters Italy through five international pipelines, while a small fraction is shipped as liquefied natural gas.

Once inside the country, the gas is transported in a system of pipelines. This is conventionally divided into primary and secondary network: the former is made of high-capacity, high-pressure (70 bars) pipes which connect hubs and deposits over long distances, while the latter is a low-pressure (1 bar) network aimed at delivering the product to the final users. SNAM Rete Gas owns 93% of the network, for a total extension of 41,000 km [40].

Although hydrocarbons can be effectively stored, the network must always be kept in balance: injection must equal extraction, up to a small margin, so that the volume of gas inside the network, named line-pack, stays constant. SNAM Rete Gas, the TSO, must then continuously monitor the situation and act to prevent imbalance surging to critical levels.

To alleviate issues deriving from the lack of domestic production, Italy built the largest storage capacity in Europe, with 10 sites collectively capable of holding about 12.5 BSCM, equivalent to 17% of the yearly demand. Deposits are always partially filled: in case of failures in international streams, natural disasters, or political tensions, the TSO taps into reserves to avoid disruption in the delivery.

In addition to its role in emergencies, storage also helps in the optimization of transportation. Deposits are usually filled when piping capacity is available and emptied when the demand is high in order to deliver gas to nearby towns. Moreover, storage provides readily available flexibility in balancing the network.

2.1.3 Market

Natural gas provisioning, distribution and commercialization to final customers used to be a monopoly until the end of the XX century. In 2000, law 164/2000 implemented European directive 90/30/CE and imposed the liberalization of the market.

The liberalization created four main actors: the end users, the TSO, in charge of managing the network, the providers, private companies in the business of provisioning and selling, and the Bureau of Energy Markets ("Gestore Mercati Energetici" or GME in Italian), in charge of supervising trading. An exchange for

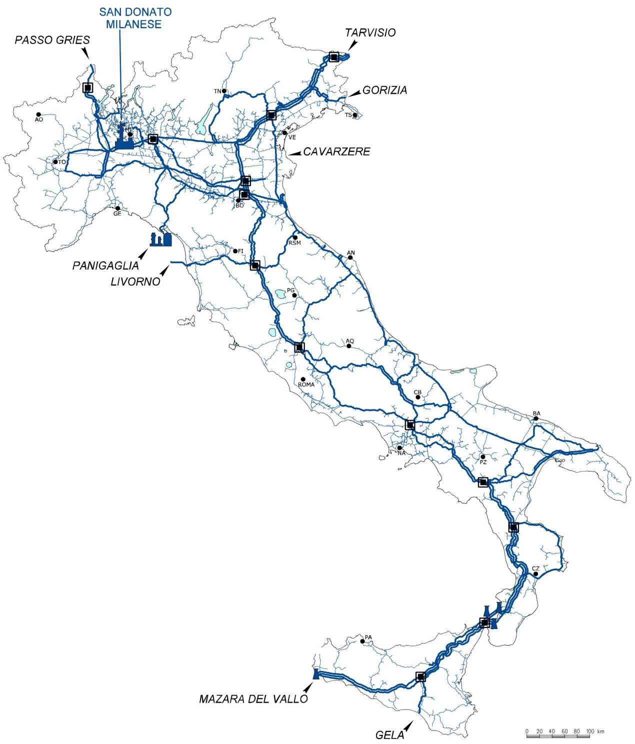


Figure 2.3. Main natural gas facilities in Italy [40].

natural gas was created and, since October 2003, operators can trade gas at the "Virtual Point of Exchange" ("Punto di Scambio Virtuale" or PSV in Italian), an imaginary point logically included within the Italian network.

As of 2020, transactions can occur in two ways: in a regulated market or over the counter, through private agreements between organizations. The regulated market is made of two platforms: the MP-GAS spot market and the MT-GAS market, where future contracts are traded. MP-GAS has two main sessions: in the "day-before" session (MGP-GAS), gas is traded for the following day, while in the "intraday" session (MI-GAS), the commodity is exchanged with immediate effect.

The market plays an important role in keeping the network balanced. An energy provider is said to be balanced if

$$E - I = T, \tag{2.1}$$

where I is the gas it injects in the network, E the extraction by its customers and T the amount traded at the PSV, assumed positive if it is bought and negative if it is sold. If (2.1) does not hold, the imbalance is $U = E - T - I$.

At the end of each day, if its U is positive, a provider pays

$$P_b = \max\left(P_{b,\max}, \bar{P} + S\right) \tag{2.2}$$

for each unit of gas demanded by its customers and not injected in the network and receives

$$P_s = \min\left(P_{s,\min}, \bar{P} - S\right) \tag{2.3}$$

for each unit of excess gas introduced in the network, where $P_{b,\max}$ is the maximum price of gas bought by the TSO on the market in the same day, $P_{s,\min}$ is the minimum price of gas sold by the TSO, $\bar{P}$ is the average market price of the day and S is a small constant, currently set at 0.108 €/MWh.

The imbalance price is designed to be always inconvenient. If a provider has to buy gas, it pays either more than the market price or the maximum TSO price, if it has to sell, it receives either less than the market price or the minimum TSO price. This is an incentive for providers not to unbalance the network and, in case they realize their position is becoming critical, to proactively act on the market to compensate as soon as possible.

2.1.4 Future

According to the National Plan for Energy Development, after the decommissioning of coal-fueled power plants, planned in 2025, natural gas will remain the only fossil fuel used for power generation [38]. Such a strategic choice is motivated by the inferior carbon footprint of combined-cycle gas-fueled plants. According to SNAM, the emission of carbon dioxide per unit energy of gas plants is 25% to 30% lower than oil plants and 40% to 50% lower than coal plants. Thermoelectric plants will thus remain the backbone of the Italian power generation infrastructure for the near future, though steadily lowering their share in favour of photovoltaic and wind. A key role will also be played by the so-called "green" hydrogen: produced by electrolysis powered by renewable sources, it will be then stored and finally burnt in power plants, possibly mixed with natural gas. In the future, hydrogen is expected to replace completely natural gas as a vector for long-term energy storage.

For what concerns the industrial and residential sectors, demand is expected to remain stable. Electrification of domestic heating is not thought to happen quickly: at current prices, gas is still economically competitive and the overwhelming majority of residential buildings in mainland Italy is already equipped with the required infrastructure.

In the next few years, the natural gas network is expected to reach Sardinia. As of 2020, Sardinia is the only region lacking a proper infrastructure for gas treatment and distribution, but both private companies and public institutions committed to substantial investments to align the island with the rest of the country.

Based on such a scenario, it is reasonable to expect that natural gas will remain a relevant energy source for decades. An analysis published by SNAM in 2017 puts the demand at 68.9 BSCM in 2035, with an average annual decrease of 0.5% between 2017 and 2035. While the short- and mid-term effect of the COVID-19 pandemic is difficult to assess, the long-term guidance was not modified.

Moreover, analyses published before the outbreak projected the global natural gas demand rising by 2% every year until 2040, when gas will represent 19% of the global primary energy production.

2.2 Importance of Demand Forecasting

While long-term forecasting is essential for establishing policies and committing to large investments, short-term forecasting is especially useful to the operations of utilities and transmission system operators and for the overall stability of the network.

As pointed out in Section 2.1.3, the Italian gas market is engineered to provide a strong economic incentive for providers not to unbalance the network. Although it is possible to compensate excess or deficiency in gas injection, it is difficult and expensive to perform such operation under short notice. Import capacity must indeed be booked in advance and trades with other operators must be performed on the market, where last-minute deals may turn unfavourable.

Hence, accurate forecasting helps both energy providers and the TSO to optimize import, transportation, storage and delivery of natural gas and has consistent economic implications for utilities.

In order to allow providers to take appropriate action and balance their position, SNAM publishes daily data of demand and injections in the network. Moreover, it releases its own forecast of demand, both at national and local level.

2.3 State of the Art

Due to the relevance of the topic, several works addressed the prediction of natural gas demand: we will use the comprehensive summaries by Soldo [41] and Šebalj et al. [42] as starting points for our review.

Šebalj et al. proposed a classification along four dimensions. The *prediction horizon* can range from hourly to yearly, the *reference area* from single nodes of the network to a whole country; adopted *models* include fundamental approaches, time series, mathematical and statistical models, neural networks, hybrid models, and others; input *features* can be historical demand, temperature, calendar, social and economic indicators or others.

We will split works sharply on the basis of the prediction horizon. Long-term forecasting, months or years ahead, is fundamentally different from short-term forecasting, days or hours ahead, in purpose, methods, and inputs. Consistent with the purpose of this work, we will focus our review mainly on country-wide forecasting, while citing only the most relevant papers dealing with regional or local targets.

2.3.1 Long-term forecasting

Several statistical models have been proposed for long-term gas demand prediction: among them, of particular interest are the so-called grey models, which were applied in different flavours to the Chinese scenario [43, 44, 45]. While such models were proven effective in delivering predictions to guide policy makers, they fail in considering exogenous factors, which may influence gas demand. Hence the need for different and more flexible approaches.

A model based on temperature was proposed to forecast residential Turkish demand [46]. The importance of temperature was also recognized with a different probabilistic approach in a work targeting the demand of Argentinian cities [47]. A study on Bangladesh showed that social and macroeconomic factors, such as population growth and gross domestic product (GDP) are essential drivers for gas demand [48]. Similar conclusions were drawn in a case study on Turkey: in this context, a breeder model was proven superior to other approaches [49].

Apart from grey and statistical models, more complex methods were also proposed. Classical time series models, like ARIMA and Holt-Winter ware applied to predict annual gas demand in Pakistan [50]. Adaptive network-based fuzzy inference systems (ANFIS) were also proven effective: a variant, featuring stochastic frontier analysis, was fed with GDP and population growth and applied to predict the evolution of natural gas demand in Bahrain, Saudi Arabia, Syria, and the UAE [51]. More recently, ANFIS were proposed as part of an hybrid ANFIS-ARIMA model to forecast demand growth in Greece [52].

2.3.2 Short-term forecasting

While long-term forecasting is heavily influenced by socio-economic variables, like economic activity and population growth, short-term forecasting can do away with such information, being it incorporated in recent demand. On the other hand, weather and holidays heavily affect the short-term behaviour of gas demand, while being hardly useful for long-term scenarios, where they are averaged out over seasons. Such differences motivate the lack of works dealing with both short and long-term forecasting - although rare examples exist [47].

Short-term forecasting of UK natural gas demand was addressed using support vector regression (SVR) with false neighbours filtered [53]. The model was fed with historical data, a composite weather variable, and calendar features. The authors showed that it outperformed auto-regressive moving average (ARMA) and neural networks (ANN) approaches. An optimized SVR, combined with factor selection algorithm and life genetic algorithm was also proposed to forecast the demand in Greek cities [54].

ANFIS was applied to predict Iranian gas demand and improved in accuracy over classical time series methods and ANN [55]. Again, historical data were complemented by calendar features, although limited to the day of the week. A more advanced model, combining wavelet transform, genetic algorithm, ANFIS and ANN was applied to forecast the daily demand across different nodes of the Greek gas distribution network [56]. The proposed model was fed with historical data, temperature and calendar features.

Neural networks were particularly popular in the last decade. Different architectures were applied to hourly and daily forecasting problems at a regional and local level [57, 58, 59, 60]. Besides being included in hybrid models (e.g. in ANFIS-ANN), ANNs were combined with principal components correlation analysis (PCCA) to provide robust and precise forecasting of regional demand [61]. Hybrid models based on recurrent neural networks were also proposed and tested on the demand of cities with different climatic conditions [62].

2.3.3 Impact of Inaccurate Temperature Forecasts

Apart from univariate models, most of the short-term approaches use temperature data. The main causal link between temperature and gas demand is the heating of houses, offices and industrial facilities, but other, less intuitive phenomena influence the consumption, as it will be clear in the sequel.

Unfortunately, the origin and the nature of weather data was seldom tributed the required attention. Some papers do not clarify if the data fed into the forecasting models are actual measurements or predictions (see, e.g., [56, 59]), some others use actual measurements both to train and validate models (e.g. [58]), even though such procedure would be impractical in an industrial application, as some authors duly observed [60].

There are few papers in the literature that consider the impact of weather forecasting errors. Potočnik et al. included weather forecasting errors in their risk analysis on gas demand forecasting errors, finding that a noise with a standard deviation of 1 °C resulted in an increase in normalized relative error of 7.7% [63]. Baldacci et al. provided empirical results on the deterioration of gas demand forecasting due to the inaccuracy of temperature predictions, concluding that 8.8% of the demand forecasting error could be attributed to weather forecasting error [30].

2.3.4 Italy

In addition to the already mentioned work by Baldacci et al. [30], a few other papers focus on the Italian scenario. Bianco et al. investigated the long-term evolution of the Italian gas demand [64, 65]: macroeconomic indicators, such as gross domestic product and gas prices, and climatic factors were used to build scenarios of residential, industrial and thermoelectric gas demand up to 2030. A large margin for energy saving was spotted in the residential sectors, where the energy efficiency of buildings plays an important role. As for the other sectors, the GDP was shown to be the most important factor in the evolution of the demand. Also in the field of long-term forecasting is the study by Scarpa and Bianco, highlighting the sensitivity of years-ahead predictions on the accuracy of input data, such as estimates of economic growth [66].

2.3.5 Open points

Our review of the literature reveals a fundamental gap in studies about short-term forecasting of Italian gas demand. Predictive models were designed and are currently in use in every utility and in SNAM Rete Gas, the TSO, but they are not publicly available.

Moreover, despite many different models fed by different set of features were proposed, a comprehensive comparison is lacking. While hybrid models were presented and tested, ensembling approaches seem not to have gained popularity, despite being proven effective in similar tasks, e.g. forecasting power load [22].

Finally, the impact of the intrinsic inaccuracy of weather forecasting on demand prediction needs to be further investigated, both under a theoretical and an experimental perspective.

2.4 Problem Statement

We addressed the prediction of the daily Italian gas demand (GD) starting from its base components. Apart from minor contributions that can be safely neglected, at any date d GD is composed by the sum of industrial gas demand (IGD), thermoelectric gas demand (TGD) and residential gas demand (RGD):

$$\mathrm{GD}_d = \mathrm{IGD}_d + \mathrm{TGD}_d + \mathrm{RGD}_d. \tag{2.4}$$

RGD includes mostly domestic heating, IGD encompasses demand by industrial facilities, while TGD only accounts for the fuel required by thermoelectric power plants.

Four one-day-ahead forecasting problems were then considered, concerning RGD, IGD, TGD, and overall GD. GD can be obtained by summing its base components, as per equation (2.4).

Data for RGD, TGD and IGD, ranging from 2007 to 2018, can be downloaded from the SNAM website. The series of gas demand were augmented with weather data, provided by an Italian specialised company. The data for gas demand is publicly available, but weather forecasts and actual measurements cannot be disclosed due to commercial agreements.

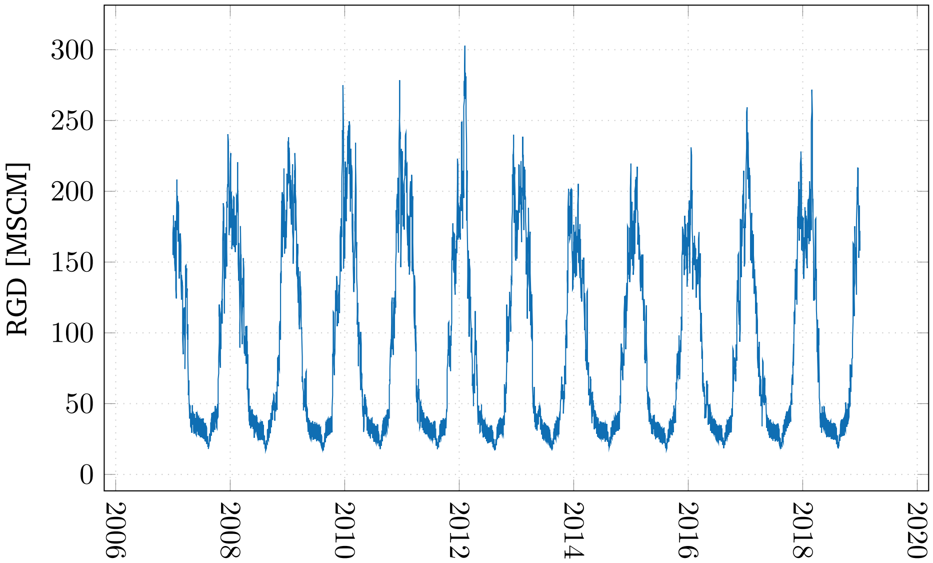


Figure 2.4. Italian residential gas demand (RGD)

The resulting dataset consisted of 5 fields: date (d), predicted average temperature in Northern Italy (t), and gas demand (RGD, IGD and TGD). All the series have daily granularity and are 12 years long. In addition, the actual average temperature in Northern Italy was procured for the years 2015 to 2018.

We considered a population-weighted average of the temperature, precomputed by the weather data provider: the underlying raw data is not available. The Northern Italy was chosen as the region where to compute the average because it experiences the most rigid climate, and is thus more sensible to heating requirements.
In a preliminary analysis, both a population-weighted and a GDP-weighted average on the whole country were also considered, but they were dropped because they showed very similar, yet weaker correlation with RGD and IGD, respectively, after performing the transformations described in the following sections.

Figs. 2.4 to 2.7 display the series of RGD, IGD, TGD, and overall GD.

2.5 Exploratory Analysis

Exploratory data analysis is an essential preliminary for the design of effective predictive models. The insights gathered in this phase drive feature selection, feature engineering and model design.

Due to their peculiar features, RGD, IGD and TGD will be analyzed separately.

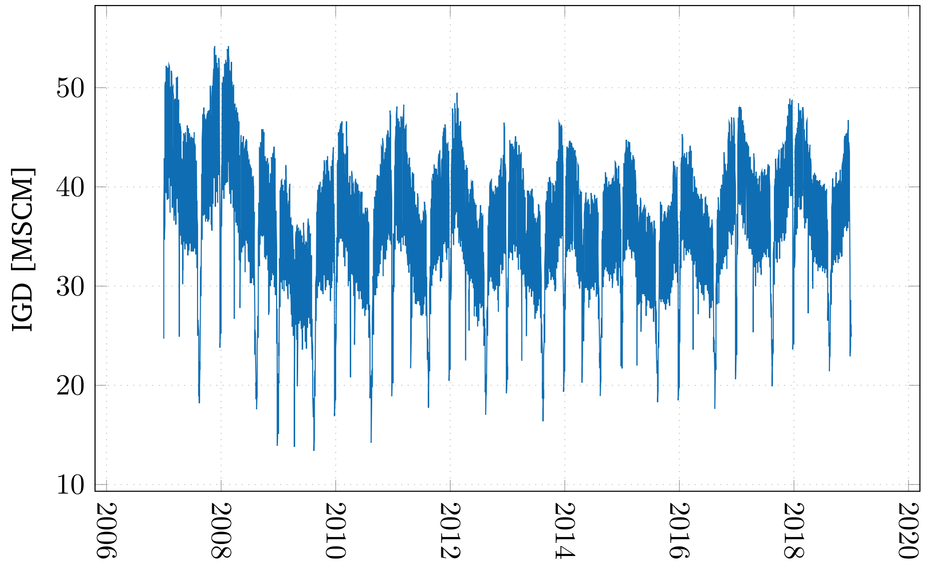


Figure 2.5. Italian industrial gas demand (IGD)

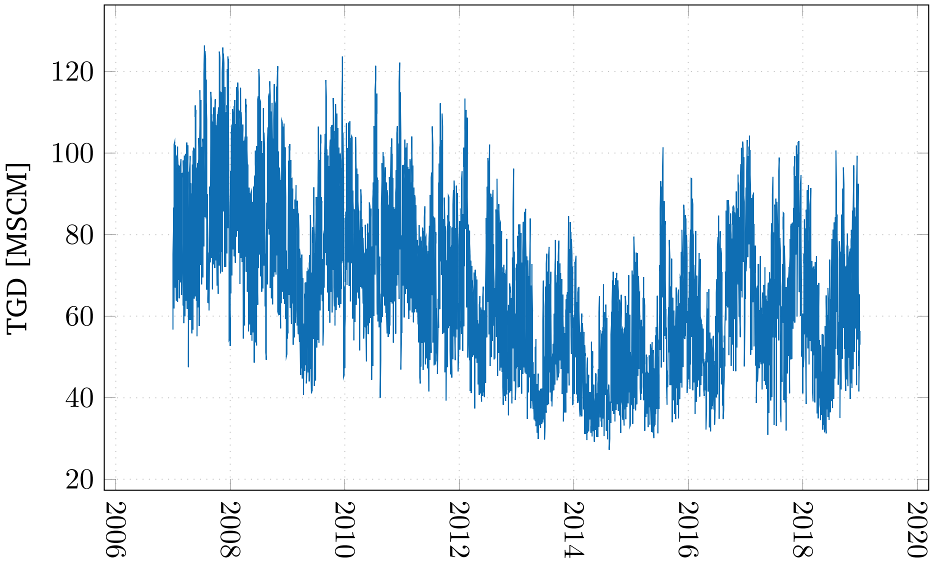


Figure 2.6. Italian thermoelectric gas demand (TGD)

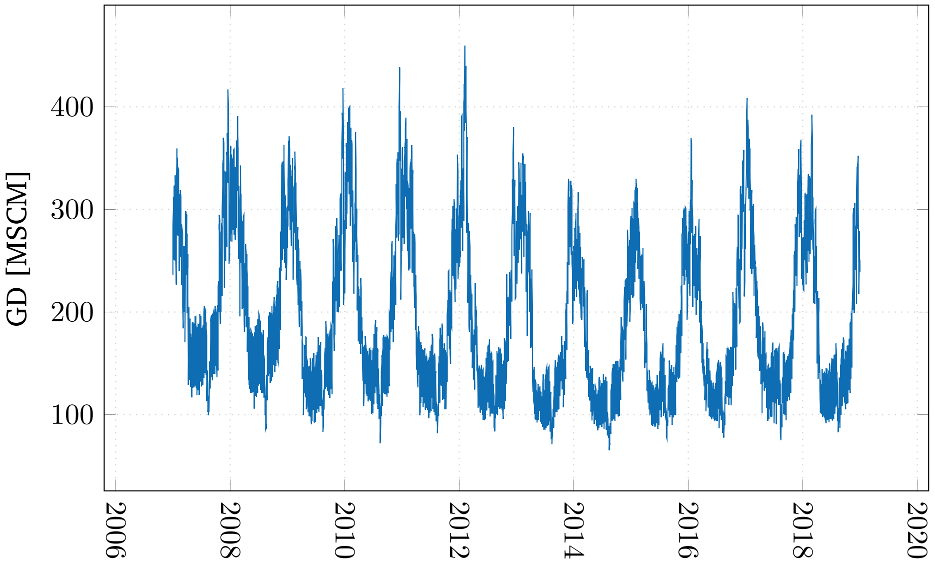


Figure 2.7. overall Italian gas demand (GD)

To keep the dissertation self-contained, a short reminder of the main statistical tools used in the next sections is provided.

2.5.1 Autocorrelation and Spectral Density

The autocorrelation function and the periodogram describe the periodic structure of a time series. For the sake of clarity, we will only present their application to time series analysis, without delving into the theory of stochastic processes from which they both originate. The interested reader can refer to the classical textbook by Box et al. for a rigorous and comprehensive presentation [67].

The empirical autocorrelation function (ACF) is a measure of the linear dependence of a time series from its lagged values. Assuming that $\boldsymbol{x} = \{x_t\} \in \mathbb{R}^n$, $t = 1 \dots n$, is a stationary series with zero mean, its ACF is a function r of the lag k, computed by the formula

$$r(k) = \frac{\sum_{t=k+1}^{n} x_t x_{t-k}}{\sum_{t=1}^{n} x_t^2}. \tag{2.5}$$

Similar to the correlation, the ACF is bounded between -1 and +1: an absolute value of $r(k)$ close to 1 suggests a perfect linear relationship between the series and its k-th lag, while $r(k)$ close to 0 indicates that a linear model is not adequate to capture the relationship. In case the series is not stationary, standard techniques can be adopted to remove the trend before computing the ACF [67].

The periodogram is the discrete Fourier transform of the empirical autocorrelation function and estimates the power spectral density (PSD), which in turn describes which frequencies contribute the most to the periodic structure of the time series. It can be shown that the periodogram can also be computed directly from the samples of the time series x_t as

$$p(f) = \frac{\Delta t}{n} \left| \sum_{t=1}^{n} x_t e^{-2\pi \mathbf{j} f t \Delta t} \right|^2, \quad \frac{-1}{2\Delta t} \leq f \leq \frac{1}{2\Delta t}, \tag{2.6}$$

where Δt is the sampling period of the series, i.e. the interval between two consecutive samples. In the following sections all the periodograms will be computed using the Welch method, a technique consisting in averaging overlapping estimates of the PSD obtained on different subsets on the series. The Welch method was shown to decrease the variance of the periodogram, thus leading to more stable estimations [67].

2.5.2 Residential Gas Demand

The magnitude of RGD oscillates greatly with the season: during the cold months, from October to March, it represents 56% of the overall Italian demand, while it drops to 28% from April to September - see Fig. 2.4.

The pattern is explained by domestic heating, the primary usage of gas in households. During the cold period, the need for maintaining a large difference between the indoor and the outdoor temperature produces a larger RGD while, when the temperature climbs above 17-18 °C, heating is typically switched off, thus reducing the consumption.

Moreover, due to the impact of domestic heating, winter RGD is influenced by weather conditions, and thus changes drastically from year to year. Conversely, in summer, RGD is independent from the weather and its profile is remarkably repeatable over different years. All these features are visible in Fig. 2.8, which displays eleven years of Italian RGD, overlapped with a proper shift to align weekdays.

As expected, the empirical autocorrelation function, estimated on the whole dataset, exhibits a dominant yearly seasonality and a much smaller weekly periodicity (Fig. 2.9). The periodogram also shows that most of the spectral density is concentrated at the period of 365.25 days (Fig. 2.10). A smaller yet relevant peak can be found at a period of 7 days, accounting for the weekly periodicity. In both cases, smaller peaks at lower periods are ascribable to harmonics.

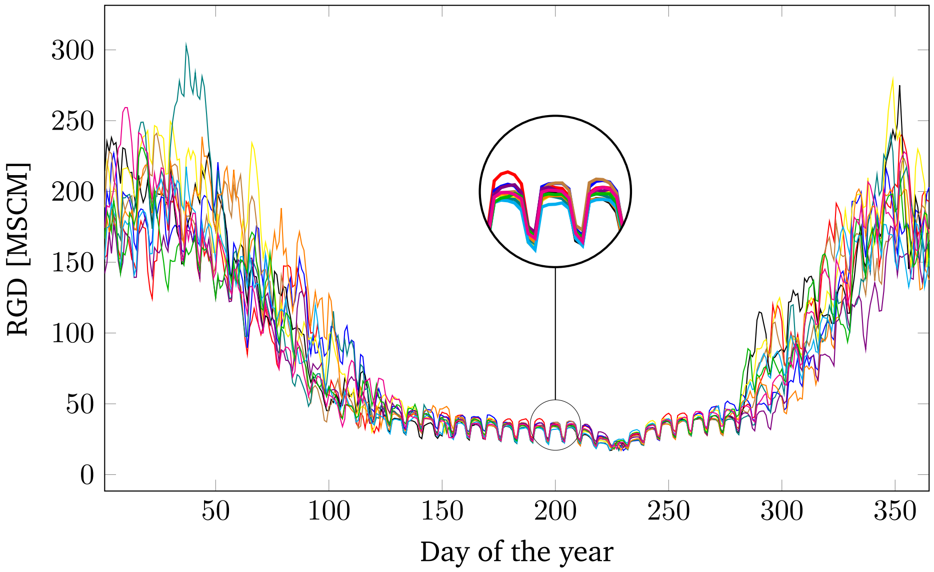


Figure 2.8. Italian residential gas demand (RGD), years 2007 to 2017. The time series are shifted to align weekdays. Weekly periodicity is particularly visible in summer. The yearly seasonal variation is mostly explained by heating requirements. In the inset, three weeks of July are zoomed.

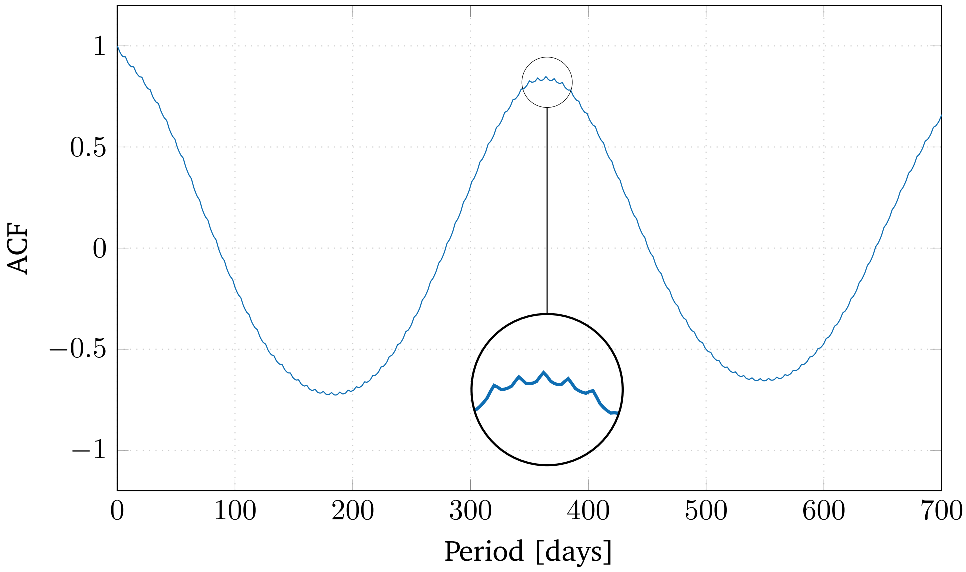


Figure 2.9. Autocorrelation function of RGD. The 365-day yearly periodicity is evident. In the inset, weekly waves witness the presence of a 7-day periodicity of smaller amplitude.

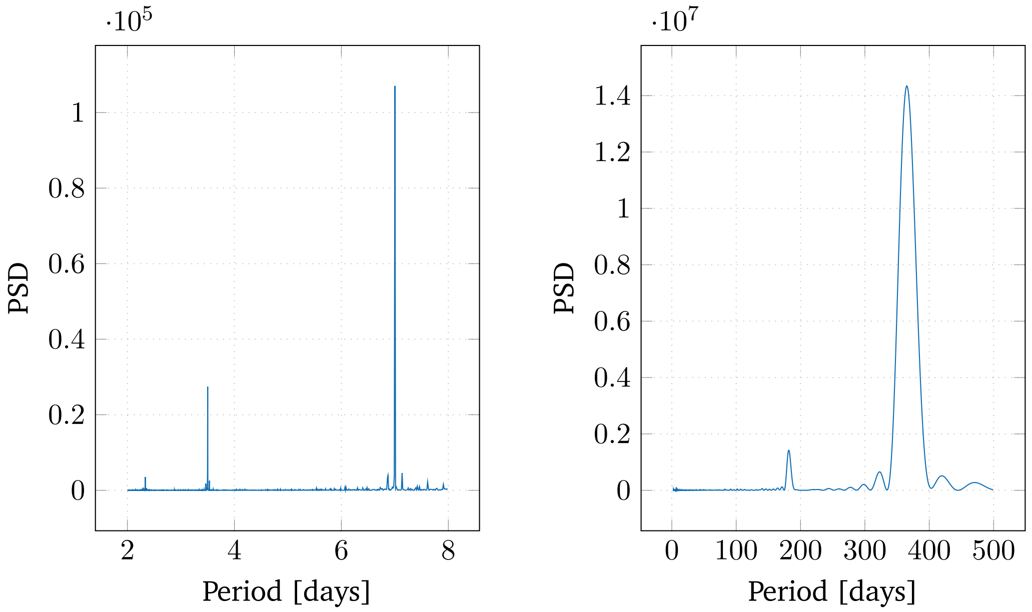


Figure 2.10. Periodogram of RGD. Left panel: periods up to to 8 days; right panel: periods up to 500 days. The yearly periodicity is highlighted by peaks at 365.25 days, while the weekly one by the smaller spike at a period of 7 days. Other spikes correspond to harmonics located at multiples of the main harmonic.

The strength of the lag-1 autocorrelation can be assessed through the scatter plot in Fig. 2.11a, where RGD at time d is plotted against RGD at time $d-1$. The correlation coefficient computed on the entire dataset is 0.988, and it increases to 0.995 if the pairs Saturday-Friday and Monday-Sunday are discarded. This reflects a different behavior between working days and weekends, visually confirmed in the plot, where Monday's RGD stays in the upper part of the cloud whereas Saturday's RGD lies in the lower part.

As for the lag-7 autocorrelation, in Fig. 2.11b the scatter plot of RGD at times d and $d-7$ is displayed. The cloud of points is narrower when the demand is low, that is during the warm season, while it gets more dispersed in winter, when the demand is high. This is possibly due to the variability of weather from one week to the next one.

In order to characterize the yearly seasonality, it is convenient to introduce the following definitions:

$\mathrm{year}(d)$ is the year to which date d belongs;

$\mathrm{weekday}(d)$ is the weekday of date d, e.g. Monday, Tuesday, etc;

$\mathrm{yearday}(d)$ is the ordinal number of date d within $\mathrm{year}(d)$ starting from 1st January, whose yearday equals 1;

$\mathcal{H}$ is the set of holidays. According to the Italian calendar, holidays are 1st January, 6th January, 25th April, 1st May, 2nd June, 15th August, 1st November, 8th, 25th and 26th December, Easter and Easter Monday.

Comparing RGD with the same yearday would not be accurate, mostly in summer, due to the shift in weekday from year to year. Adjusting only for the weekday would also introduce errors, due to the presence of holidays. Traditionally, in the literature, categorical or dummy variables are introduced to account for the issues [53, 55]. Though clearly useful, categorical variables do not carry information about previous samples with the same characteristics. We thus propose a novel characterization, based on the idea of similar day.

Definition 1 (Similar Day). *If $t \notin \mathcal{H}$, its similar day* $\mathrm{sim}(d)$ *is*

$$\mathrm{sim}(d) = \arg\min_{\tau} |\mathrm{yearday}(\tau) - \mathrm{yearday}(d)| \tag{2.7a}$$

$$\text{subject to :} \quad \mathrm{year}(\tau) = \mathrm{year}(d) - 1 \tag{2.7b}$$

$$\mathrm{weekday}(\tau) = \mathrm{weekday}(d) \tag{2.7c}$$

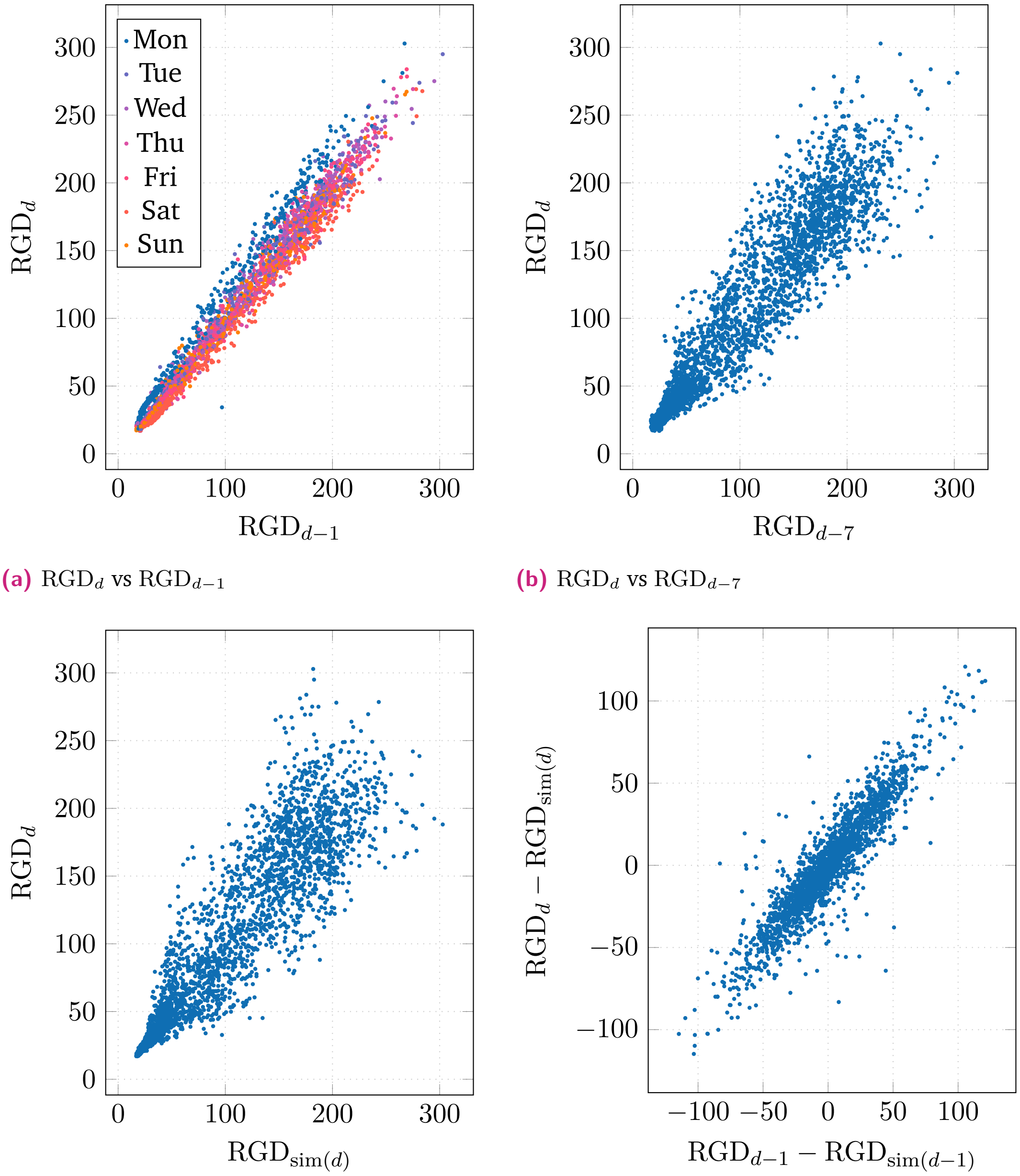
If $t \in \mathcal{H}$, its similar day $\mathrm{sim}(d)$ *is the same holiday in the previous year.*

The relationship between RGD_d and $\mathrm{RGD}_{\mathrm{sim}(d)}$ is shown in Fig. 2.11c: again, the correlation is higher when the demand is lower, due to the smaller influence of temperature, suggesting that the RGD recorded on the similar day may be an important feature during the summer period.

It can also be of some interest to take into account the similar day of $d-1$. Indeed, the scatter plot in Fig. 2.11d shows that the difference $\mathrm{RGD}_{d-1} - \mathrm{RGD}_{\mathrm{sim}(d-1)}$ is a good proxy for the difference $\mathrm{RGD}_d - \mathrm{RGD}_{\mathrm{sim}(d)}$. Therefore, a predictive model may take advantage from the available information in RGD_{d-1}, $\mathrm{RGD}_{\mathrm{sim}(d-1)}$ and $\mathrm{RGD}_{\mathrm{sim}(d)}$ to infer the target RGD_d.

Temperature

The analysis performed so far confirms the intuition that temperature is a major driver for RGD. We thus dig deeper into the relationship between weather and RGD.



(a) RGD_d vs RGD_{d-1}

(b) RGD_d vs RGD_{d-7}

(c) RGD_d vs $\mathrm{RGD}_{\mathrm{sim}(d)}$

(d) Difference between lags.

Figure 2.11. Scatter plots of RGD and its lags or derived variables. All the values are in MSCM.

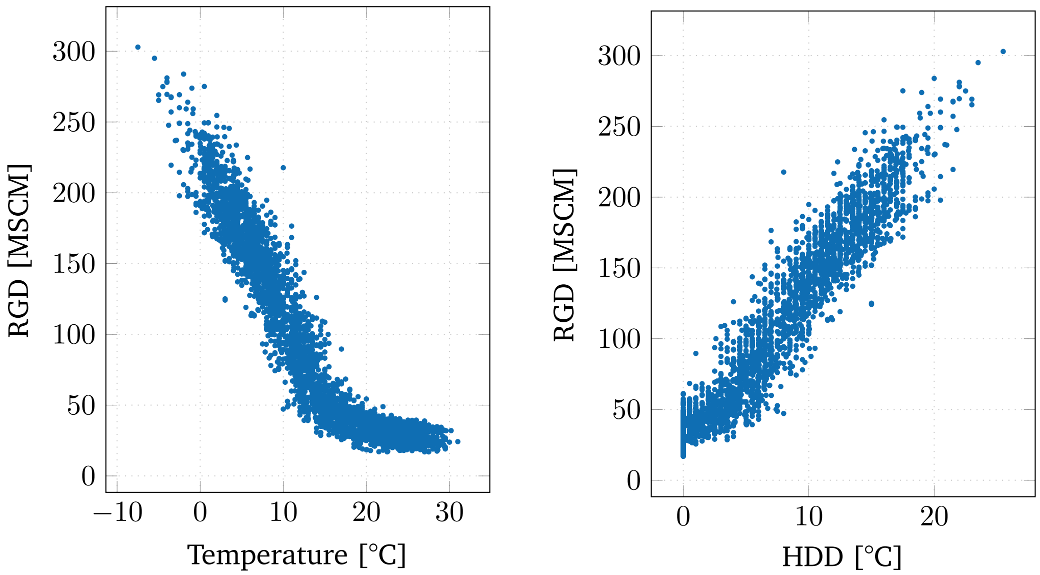


Figure 2.12. Left panel: scatter plot of daily RGD vs average daily temperature. Right panel: scatter plot of daily RGD vs HDD.

The correlation between the two series is strong during the winter season, when the temperature falls and domestic heating becomes relevant. As shown in the left panel of Fig. 2.12, with good approximation the relationship is piecewise linear: a line with a negative slope below $17 - 18$ °C and a constant above $17 - 18$ °C. In order to transform the piecewise linear dependence into a linear one, it is useful to resort to the so-called Heating Degree Days (HDD).

Definition 2 (Heating Degree Days (HDD))**.**

$$\mathrm{HDD}(t) := \max(t_h - t, 0) \tag{2.8}$$

where t_h is the point above which the temperature has no influence on gas demand.

Experiments showed that the maximum correlation between HDD and RGD, excluding points where HDD equals zero, is achieved for $t_h = 18$ °C.

In the right panel of Fig. 2.12, the scatter plot of RGD vs. HDD highlights an approximately linear relationship, with a positive correlation of 0.97.

Despite being effective in capturing the main functional form of the relationship between RGD and temperature, the HDD transformation is not smooth, while the shape in Fig. 2.12 is. Hence, we introduce a novel formulation, called Smooth Heating Degree Days (SHDD).

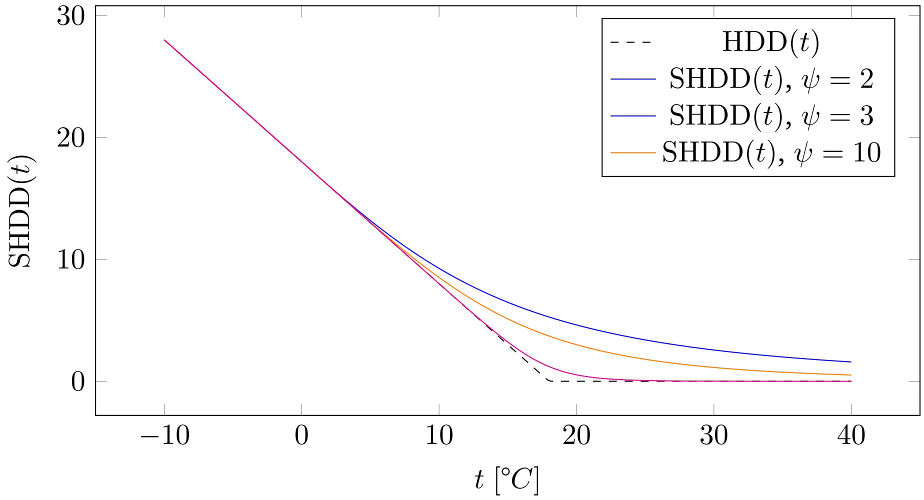


Figure 2.13. HDD(t) vs SHDD(t) for $t_c = 18$ °C. The two functions tend to coincide for large values of ψ.

Definition 3 (Smooth Heating Degree Days (SHDD))**.**

$$\mathrm{SHDD}(t) := \begin{cases} t_h - \dfrac{t}{\left(1+\left(\frac{t}{t_h}\right)^{\psi}\right)^{\frac{1}{\psi}}} & t \geq 0 \\ t_h - t & t < 0 \end{cases} \tag{2.9}$$

where t_h has the same meaning as in Definition 2 while ψ is a suitable constant that controls the smoothness of the curve.

The function SHDD(t) is represented in Fig. 2.13 for different values of ψ. Unfortunately, experiments showed that the introduction of SHDD does not yield any significant advantage, neither in terms of correlation with RGD nor in terms of performance of the final models. Thus, SHDD was dropped and will not be considered in the following.

The link between between HDD and RGD is also evident from the time series of RGD and HDD during 2017 (Fig. 2.14): it is easy to see how spikes in HDD correspond to peaks in RGD on the same date. This is in disagreement with some of the literature, which suggests that the reaction to temperature movements is not instantaneous [30]. One should note, however, that our case study involves daily data and embraces a whole country: it is thus likely that some phenomenon which appears for individual customers on a smaller time scale gets attenuated by the lower granularity and the aggregation.

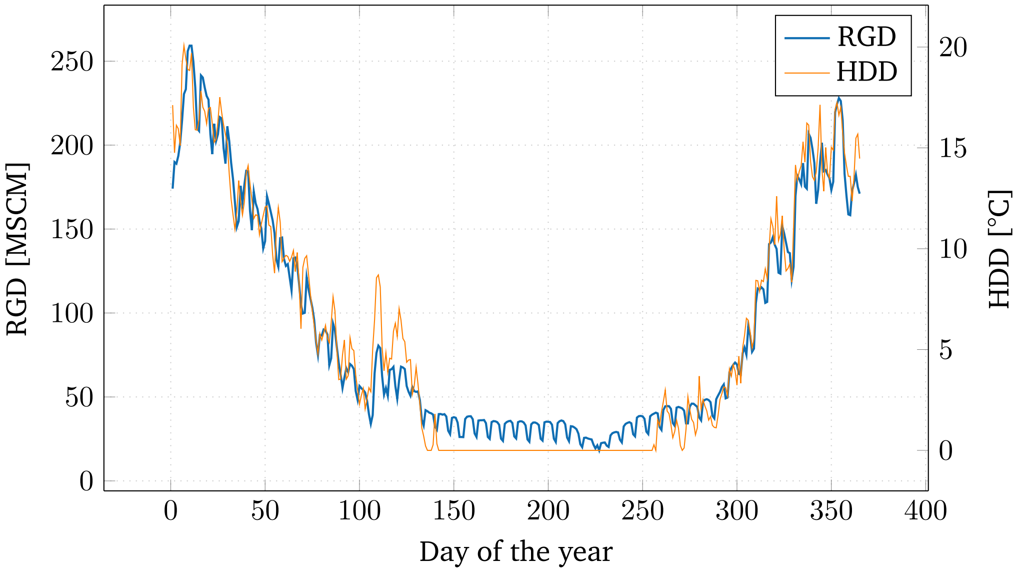


Figure 2.14. Time series of RGD and HDD in 2017: the chart shows the instantaneous correlation between the two series.

It is also interesting to notice the effect of seasonality: the cold period in the second half of April (days 105-120) results in a lower demand with respect to days with similar temperatures in the beginning of March (days 60-70). This is likely due to a combination of the effects of town regulations, which forbid the use of heating in the warm season and psychological bias of people, who are less likely to turn on domestic heating in the warm season.

All the conclusions drawn for the relationships of Italian RGD with temperature intuitively apply also to other countries with similar climatic conditions. For instance, the sharp difference in demand between the cold and warm season is apparent in Greek data [55]. On the other hand, countries with a more rigid climate, like the UK, or where natural gas is not extensively used for domestic heating, show a more constant demand throughout the year [53].

2.5.3 Industrial Gas Demand

Industrial gas demand (IGD) does not exhibit strong trends: a significant decrease is only recorded in 2009, following the global financial crisis. The series presents weekly and yearly seasonal patterns: in particular, as most of the industrial facilities stop or slow down production during the weekends, IGD is lower on Saturdays and Sundays. In August and at the end of December, typical periods for vacation in Italy, IGD drops to about half of its average value. Other holidays, such as Easter and the

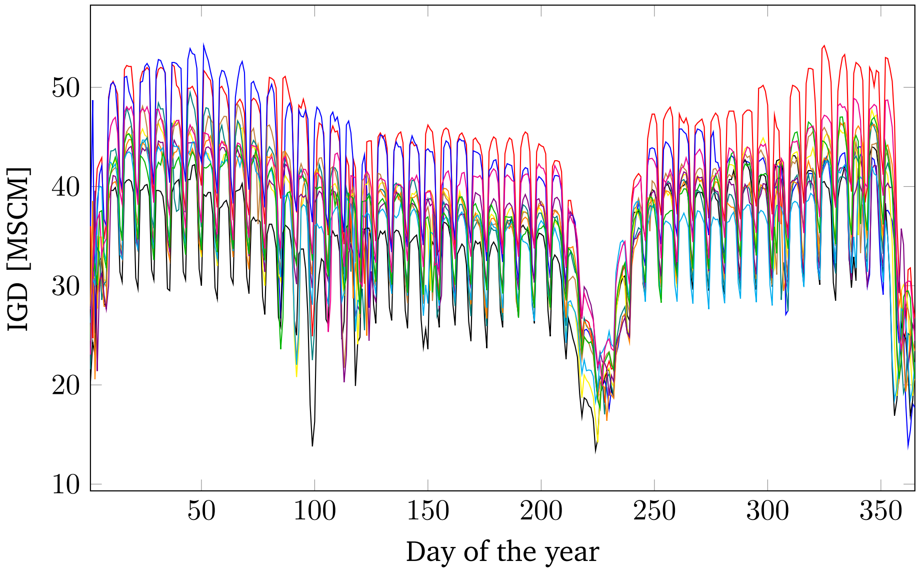


Figure 2.15. IGD, years 2007-2017. The time series has been shifted to align weekdays over different years.

Labour Day, result in similar effects. During the year, IGD shows a decrease from January to August and an increase from September to December, due to the use of gas for environmental heating in factories and manufacturing plants.

All these features can be appreciated in Fig. 2.15, where 11 years of IGD are superimposed, aligning weekdays to better highlight periodic behaviours.

The periodogram, plotted in Fig. 2.16, exhibits peaks at periods of 365.25 and 7 days, while other relevant values are ascribable to multiple harmonics of the fundamental ones. Different from RGD, the weekly seasonality prevails in terms of magnitude and the yearly one is of secondary importance. The analysis of the autocorrelation function is here neglected, as it also yields the same results.

Temperature

It is reasonable to expect the link between IGD and temperature to be less tight than the one of RGD. The natural gas consumed to heat industrial facilities is but small portion of the total IGD, the remainder being used directly to power machinery. Such intuition is confirmed by the scatter plots in Fig. 2.17, which show a much greater dispersion than Fig. 2.12.

Interestingly, the chart highlights other characteristics of IGD: with reference to the left panel of Fig. 2.17, the points in the bottom left (low IGD, low temperature)

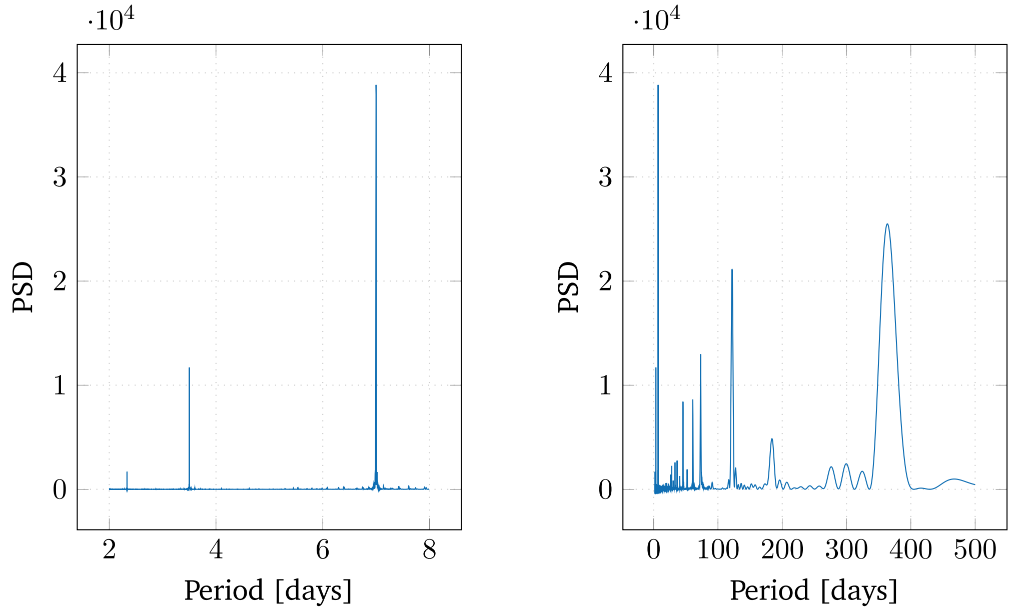


Figure 2.16. Periodogram of IGD. Left panel: periods from 0 to 8 days; right panel: periods from 0 to 500 days.

belong to the Christmas period, while the ones in the bottom right (low IGD, high temperature) to the summer break. The isolation of these two groups of samples from the rest of the cloud suggests the importance of correctly identifing and characterizing the two main vacation periods and other holidays.

2.5.4 Thermoelectric Gas Demand

Different from IGD, thermoelectric gas demand (TGD) shows a clear trend (Fig. 2.6). From 2008 to 2014 TGD decreases, mostly due to the growth in capacity of renewable power sources, substantially subsidized by the government. Since 2014, however, the trend stabilizes, likely due to the decrease in subsidies for the installation of photovoltaic systems.

Moreover, TGD shows a greater variability compared to IGD and RGD, as the year-over-year plot in Fig. 2.18 shows. TGD is indeed influenced by several factors other than temperature and productive cycles, including prices of power, gas, and European emission allowance (EUA) certificates, which exhibit a large volatility [68]. This explains why yearly periodicity is relatively less important in TGD than in IGD and RGD. The periodogram in Fig. 2.19 shows that, also for TGD, the main seasonal component is weekly, which is consistent with the patterns reported in the Italian power demand [69].

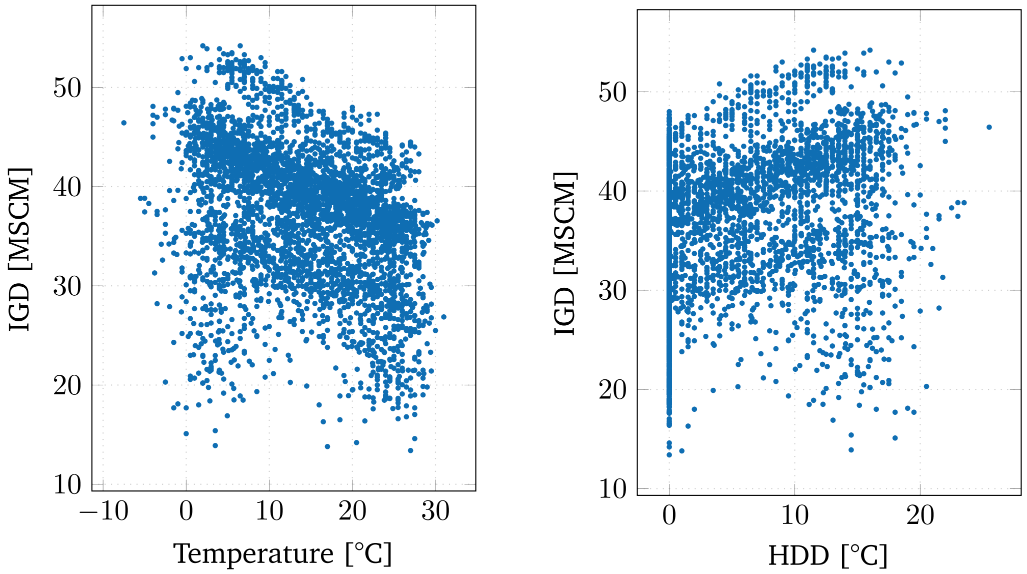


Figure 2.17. Effect of temperature on IGD. Left panel: IGD vs temperature; right panel: IGD vs Heating Degree Days (HDD).

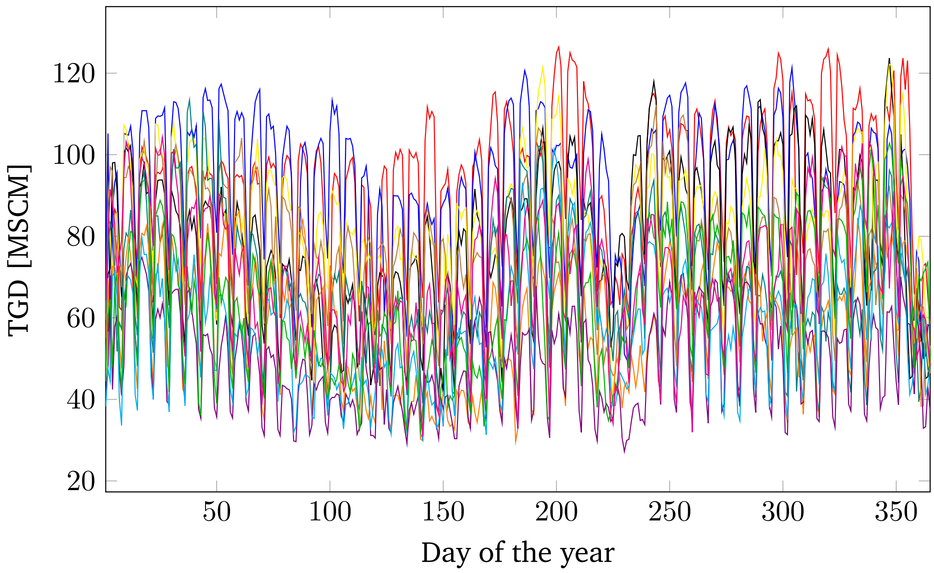


Figure 2.18. TGD, years 2007-2017. The time series has been shifted to align weekdays over different years.

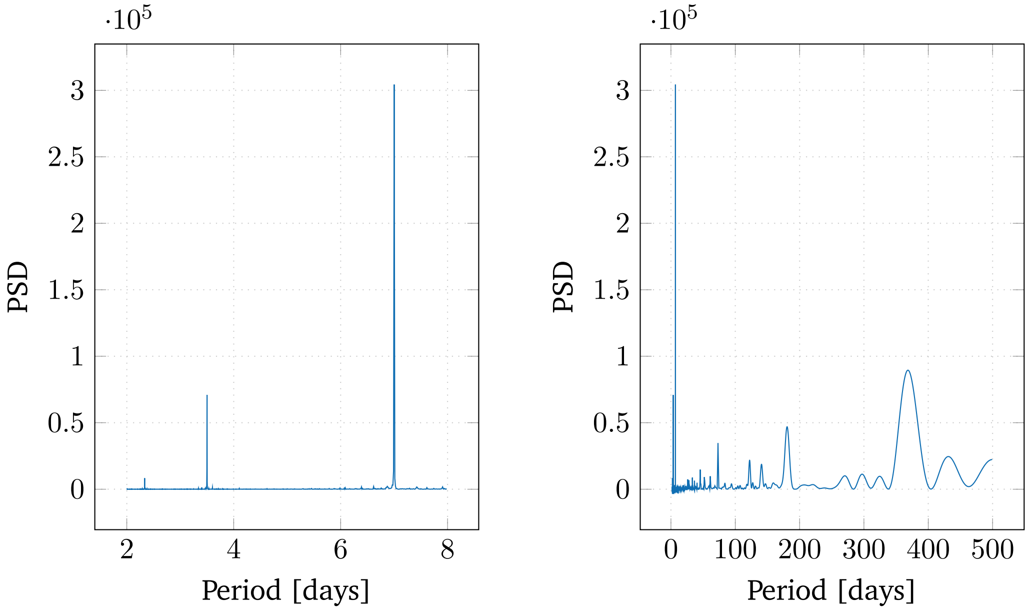


Figure 2.19. Periodogram of TGD. Left panel: periods from 0 to 8 days; right panel: periods from 0 to 500 days.

Temperature

The scatter plot of TGD against temperature, displayed in the left panel of Fig. 2.20, shows a peculiar U-shaped pattern: TGD increases as weather gets colder, but also when it gets warmer. Indeed, during the summer more thermoelectric production is required to match the demand caused by the extensive use of air conditioning. The increase in solar generation typically recorded in the same part of the year in not sufficient to counterbalance the increase in power demand with the current solar installed capacity.

This U-shaped pattern calls for the introduction of a derived feature variable, named Heating and Cooling Degree Days (HCDD).

Definition 4 (Heating and Cooling Degree Days (HCDD))**.**

$$\mathrm{HCDD}(t) := |t_c - t| \tag{2.10}$$

where t_c is a suitable constant.

Experiments showed that $t_c = 16$ °C maximizes the correlation between TGD and HCDD.

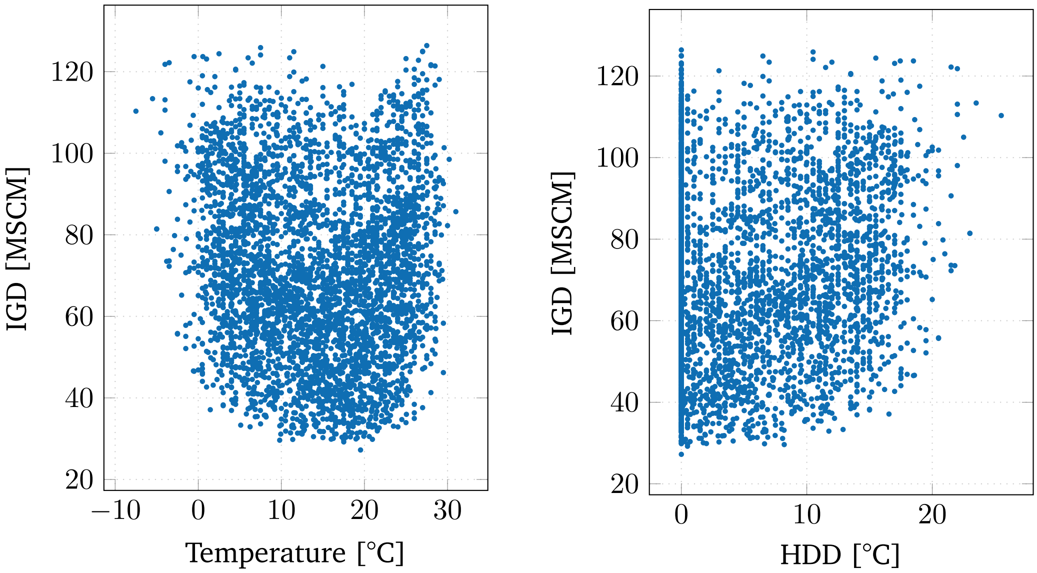


Figure 2.20. Effect of temperature on TGD. Left panel: TGD vs temperature; right panel: IGD vs Heating and Cooling Degree Days (HCDD).

2.6 Feature selection

Based on the insights offered by the exploratory analysis, it is reasonable to conclude that the gas demand is mainly influenced by three factors: past consumption, which incorporates information about the population, economic activity, season, and climate, the weather, and the calendar, in particular holidays and vacations.

We thus decided to feed our models with autoregressive terms, calendar features, temperature and its derived variables HDD and HCDD. Table 2.1 reports the complete list of features.

2.6.1 Autoregressive features

Periodograms and scatter plots of RGD, IGD and TGD clearly show the strong correlation between gas demand at specific dates. To predict $y_d, y \in \{\mathrm{RGD}, \mathrm{IGD}, \mathrm{TGD}\}$, we included y_{d-1}, y_{d-7}, $y_{\mathrm{sim}(d)}$ and $y_{\mathrm{sim}(d-1)}$. The feature y_{d-1} carries information about short-term phenomena, such as particular weather conditions, while y_{d-7} and $y_{\mathrm{sim}(d)}$ aim at capturing the two main seasonalities of RGD, IGD and TGD. The latter feature, $y_{\mathrm{sim}(d-1)}$, is included because, as observed in Section 2.5.2, the difference $y_{d-1} - y_{\mathrm{sim}(d-1)}$ is a good proxy for the difference $y_d - y_{\mathrm{sim}(d)}$.

Feature	Reference time	Type	Series
Gas demand series	$d-1$	continuous	RGD, IGD, TGD
Gas demand series	$d-7$	continuous	RGD, IGD, TGD
Gas demand series	$\mathrm{sim}(d)$	continuous	RGD, IGD, TGD
Gas demand series	$\mathrm{sim}(d-1)$	continuous	RGD, IGD, TGD
Forecasted temperature	d	continuous	RGD, IGD, TGD
Forecasted temperature	$d-1$	continuous	RGD, IGD, TGD
Forecasted temperature	$d-7$	continuous	RGD, IGD, TGD
Forecasted temperature	$\mathrm{sim}(d)$	continuous	RGD, IGD, TGD
Forecasted HDD	d	continuous	RGD, IGD
Forecasted HDD	$d-1$	continuous	RGD, IGD
Forecasted HDD	$d-7$	continuous	RGD, IGD
Forecasted HDD	$\mathrm{sim}(d)$	continuous	RGD, IGD
Forecasted HCDD	d	continuous	TGD
Forecasted HCDD	$d-1$	continuous	TGD
Forecasted HCDD	$d-7$	continuous	TGD
Forecasted HCDD	$\mathrm{sim}(d)$	continuous	TGD
Weekday	d	categorical	RGD, IGD, TGD
Holiday	d	dummy	RGD, IGD, TGD
Day after holiday	d	dummy	RGD, IGD, TGD
Bridge holiday	d	dummy	RGD, IGD, TGD

Table 2.1. Features selected for the short-term forecasting of natural gas demand.

2.6.2 Calendar features

It has been observed that weekdays and holidays have great influence on RGD, IGD and TGD for different reasons. To capture this phenomenon, the autoregressive terms based on the similar day may not be sufficient. Thus, the following categorical features were generated and introduced in the dataset.

Weekday Six binary features which encode the days of the week via dummification.

Holiday A binary feature which takes value 1 in correspondence of holidays.

Day after holiday A binary feature which takes value 1 the first working day after a holiday. A working day is defined as a day different from Saturday and Sunday that is not a holiday.

Bridge holiday A binary feature which takes value 1 on isolated working days, that is working days where both the day before and the day after are either Saturdays, Sundays or holidays.

2.6.3 Temperature features

Based on the analysis of the link between gas demand and temperature, we selected forecasted temperatures t_d, t_{d-1}, t_{d-7} and $t_{\mathrm{sim}(d)}$. The lags correspond to the ones of the autoregressive features and are intended to complement their information. For instance, in the case of RGD, the demand of the day before is itself an accurate predictor if the temperature of the target day is similar, while it has to be corrected if the temperature changes greatly. In view of what shown in Section 2.5 for RGD and IGD, also HDD values at the same times were introduced, while, for TGD, HCDD replaced HDD.

2.7 Modelling

A popular approach to predict gas demand - and time series in general - is to cast a regression problem, where lags of both the target and exogenous series are used as inputs - see, e.g., [53, 61, 56].

Two fundamental question then arise: how to choose the features and what model to select. In order to better answer such questions, it is important to define the purpose of the models.

Our goal is twofold. We aim at designing an accurate predictor for the Italian gas demand, but we also wish to provide an insightful comparison between different methods, understanding when and why some perform better than others.

Two approaches can be adopted to answer the first question: feature can be chosen either with automated procedures or based on exploratory analysis. Both methods were proposed in the literature. For instance, Zhu et al. adopted a nearest-neighbours approach with false neighbours filtered [53] to select the features. Unfortunately, as results presented in the following sections show, the closest neighbours are not good predictors for our series. Moreover, the results of the proposed procedure are hard to explain. The best predictive performance is achieved when lags up to 55 are considered as the starting set for the nearest neighbours selection. In view of the discussion in Section 2.5, it is difficult to justify how such distant lags can deliver some predictive power on gas demand.

Similar considerations make us lean towards an expert selection, based on exploratory analysis: we select the features listed in Table 2.1 based on the analyses of Section 2.5.

In order to answer the second question, we selected nine among the most widespread regression models and we compared their performance. The selected models were already adopted in the literature, either to predict gas demand or to solve similar problems. We decided to disregard in this study hybrid models, i.e. models created by combining one or more base predictors, despite their success in recent years. A single hybridization approach or a similar enhancement can be applied to several models: taking into account also such methods would make the comparison hard to implement and interpret.

Nonetheless, we noted an interesting gap in the literature. To the best of our knowledge, ensembling, a popular technique consisting in the aggregation of the output of different models, was never applied to the problem of gas demand prediction, despite being proven successful on similar tasks - e.g., the prediction of electric load. Thus, we present and discuss four different ensembling approaches and compare their performance with the base models.

Before discussing the experimental setup and the results, we briefly present the statistical learning framework and the adopted models.

2.7.1 A Short Introduction to Statistical Learning

Statistical learning is a mathematical framework aimed at estimating an unknown function from a set of samples drawn from a probability distribution.

Let $\mathcal{X}$ be the vector space of the inputs and $\mathcal{Y}$ the vector space of outputs. The meaning of "input" and "output" will be clarified soon. The first assumption of statistical learning is the existence of a fixed, yet unknown, probability distribution ρ defined over the product space $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, so that $\rho(z) = \rho(x, y)$ represents the joint probability distribution of $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

The second key assumption is that the relationship between x and y can be described by a function $f : \mathcal{X} \rightarrow \mathcal{Y}$, such that $f(x) \sim y$. In this respect, f defines the roles of inputs and outputs. The problem now consists in finding f.

As it would be impractical to search f among all possible function, the choice is restricted to the hypothesis set $\mathcal{H}$. In order to define an ordering among the candidate functions and choose the best one, a real-valued loss function $l : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$, is introduced. The loss l is intended to yield a measure of the error committed by replacing y with $f(x)$: hence, lower values correspond to higher similarity between $f(x)$ and y and thus to a "better" f.

It is now possible to define the expected risk e as the expectation of l over the space of all possible input-output pairs $\mathcal{Z}$:

$$e(f) = \int_{\mathcal{Z}} l(f(x), y)\, d\rho(x, y). \tag{2.11}$$

Having fixed l, the expected risk only depends on the function f. Due to the characteristics of l, the best f is the one corresponding to the lowest expected risk: it is thus possible to cast the optimization problem

$$f^* = \arg\min_{f \in \mathcal{H}} e(f) \tag{2.12}$$

and solve for f.

Unfortunately, it is not feasible to compute $e(f)$ in practice, as it involves the unknown probability density ρ. Thus, a proxy for the expected risk is required. Suppose now that a set $\mathcal{S}$ of n samples drawn from ρ is available: $\mathcal{S} = \{x_i, y_i\}$, $i = 1...n$.

The empirical risk, based on the available samples $\mathcal{S}$, is

$$\hat{e}(f) = \frac{1}{n} \sum_{i=1}^{n} l(f(x_i), y_i). \tag{2.13}$$

The empirical risk does away with ρ and it is thus computable. The best f can finally be found by solving

$$\hat{f} = \arg \min_{f \in \mathcal{H}} \hat{e}(f). \tag{2.14}$$

Once $\hat{f}$ has been chosen, it can be used to predict the output $\hat{y} = \hat{f}(x)$ from inputs x of new samples not included in $\mathcal{S}$.

In the context of statistical learning, a model is thus defined by three elements: a set of candidate functions $\mathcal{H}$, which are supposed to describe the relationship between inputs and outputs, a loss function l, capable of ranking the functions in $\mathcal{H}$, and a training algorithm to solve the empirical risk optimization problem (2.14).

The best function $\hat{f}$ is identified based on a training set of samples $\mathcal{S}$, but a different set of observations, the test set $\mathcal{T}$, is required to verify the goodness of the model. Indeed, the theoretically optimal f^* minimizes the expected risk over the whole space $\mathcal{Z}$: assessing $\hat{f}$ on the same samples used to choose it would not provide any indication about its performance on different inputs and outputs. A situation where a model adapts too tightly to the training set and loses generalization capability is called overfitting.

In this study, $\mathcal{X} = \mathbb{R}^m$, $m = 21$ is the space of the features listed in Table 2.1, while $\mathcal{Y} = \mathbb{R}$ is the space of the target gas demand, either RGD, IGD or TGD. The samples included in $\mathcal{S}$ and $\mathcal{T}$ are extracted from the dataset presented in Section 2.4 with the procedure described in Section 2.8.

Before describing the considered models, we provide two basic classifications which will be useful in the following.

2.7.2 Parametric and non-parametric models

A common practice to describe the hypothesis set $\mathcal{H}$ is by using parametric functions. The parametric function f is fixed but for a vector $\boldsymbol{\theta} \in \mathbb{R}^p$ of real-valued parameters: $f : \mathcal{X} \times \mathbb{R}^p \to \mathcal{Y}$.

With such an approach, problem (2.14) becomes

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} l(f(x_i, \boldsymbol{\theta}), y_i), \tag{2.15}$$

where f is now fixed.

Compared to non-parametric approaches, parametric models trade flexibility for simplicity and stability, as it is in general easier to solve an optimization problem on the set of real numbers, like (2.15), than one on a set of functions, like (2.14). On the other hand, if the initial choice of f is not appropriate, parametric methods may fail to achieve a satisfactory performance.

To allow for further degrees of freedom, some models, both parametric and non-parametric, feature additional parameters which are not chosen by the training algorithm. These are called hyperparameters and must be selected by *ad-hoc* methods, such as cross-validation, or general optimization techniques, like genetic algorithms.

2.7.3 Linear and non-linear models

Among parametric models, a distinction is made between linear and non-linear methods. The former assume that f is a linear function, whereas the latter do not enforce such constraint. Whether a linear or a non-linear model is most appropriate depends on the problem on hand and the previous knowledge about f. In general, linear models are more stable, while being less flexible than non-linear approaches.

In the task of forecasting short-term gas demand, non-linear models were proven more successful - see Section 2.3, but we will consider also linear models as baselines.

2.7.4 Base Models

We selected nine base models, which can be grouped into three categories:

- *linear models*: ridge regression, lasso, elastic net, and torus model [69];
- *non-linear models*: support vector regression, neural networks;
- *non-parametric models*: random forest, Gaussian process, nearest neighbours.

Ridge, lasso, and elastic net

Ridge regression [70], lasso [71] and elastic net [72] are linear models based on the parametric function $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\theta}$, where $\boldsymbol{x} \in \mathbb{R}^p$ is the vector of inputs and $\boldsymbol{\theta} \in \mathbb{R}^p$ are the parameters.

The methods differ in their loss functions, which can be decomposed into two terms, the first being common and the second different across the models. Let us define the input matrix $X \in \mathbb{R}^{n \times p}$ as $X = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \ldots, \boldsymbol{x}_n^\top]^\top$ and the output vector $\boldsymbol{y} \in \mathbb{R}^n$ as $\boldsymbol{y} = [y_1, y_2, \ldots, y_n]^\top$, where the couples $\{\boldsymbol{x}_i, y_i\}$ belong to the training set $\mathcal{S}$. Then, the shared term of the loss l is the quadratic function

$$l_q(\hat{\boldsymbol{y}}, \boldsymbol{y}) = l_q(X\boldsymbol{\theta}, \boldsymbol{y}) = \|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2 \,. \tag{2.16}$$

To prevent overfitting and improve generalization capabilities, a penalty on the magnitude of $\boldsymbol{\theta}$ is added to the loss function. Such approach is known as Tikhonov regularization: the interested reader can find an extensive discussion in the classical textbook by Hastie et al. [73]. The parametric empirical risk minimization problems cast by the three models are

$$\hat{\boldsymbol{\theta}}^{\text{ridge}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_2^2 \,, \tag{2.17a}$$

$$\hat{\boldsymbol{\theta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1 \,, \tag{2.17b}$$

$$\hat{\boldsymbol{\theta}}^{\text{el. net}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2 + \lambda \left(\alpha \|\boldsymbol{\theta}\|_2^2 + (1-\alpha) \|\boldsymbol{\theta}\|_1\right) . \tag{2.17c}$$

The different penalties result in specific shrinking patterns of the parameters $\boldsymbol{\theta}$. The ℓ_2 penalty in (2.17a) shrinks $\boldsymbol{\theta}$ toward the origin; the ℓ_1 penalty in (2.17b) has the effect of zeroing the least relevant parameters, thus enforcing some degree of sparsity, while the mixed ℓ_1 and ℓ_2 term in (2.17c) achieves an intermediate effect. In (2.17), both λ and α are hyperparameters.

Torus model

The torus is a linear model originally proposed to predict power load [69]. The idea is to deconstruct the target series into trend, seasonality and effect of exogenous factors, separately identify and predict the different components, and finally recompose the forecast. Due to its peculiar structure, the torus model does not require all the features of Table 2.1, but only a suitable subset. On the other hand, it introduces other inputs, not included in Table 2.1.

Before estimating the parameters of the model, a logarithmic transformation is applied to the target series, in order to level off the variance. The predictive function of the torus model is

$$\log(f_l(\boldsymbol{x}_d)) = q(d) + \boldsymbol{\theta}_s^\top \boldsymbol{s}_d + \boldsymbol{\theta}_h^\top \boldsymbol{h}_d, \tag{2.18}$$

where the input $\boldsymbol{x}_d = [d, \boldsymbol{s}_d^\top, \boldsymbol{h}_d^\top]^\top$ is composed by the date d of the output, the seasonal terms $\boldsymbol{s}_d$, and the exogenous variables $\boldsymbol{h}_d$. The prediction law defined by f_l does not take into account autoregressive terms, and is thus called long-term model.

In (2.18), the function $q(d)$ accounts for the long-term trend of the output series. The multiperiodic term $\boldsymbol{s}$ is a vector of sinusoidal functions evaluated at time d:

$$\boldsymbol{s}_d = \mathcal{D} \times \mathcal{W}, \tag{2.19}$$

where

$$\mathcal{D} = \{\cos(j\psi d), j = 0 \dots n_d\} \cup \{\sin(j\psi d), j = 1 \dots n_d\}, \tag{2.20a}$$

$$\mathcal{W} = \{\cos(k\omega d), k = 0 \dots n_w\} \cup \{\sin(k\omega d), k = 1 \dots n_w\}, \tag{2.20b}$$

and the frequencies of the sinusoidal functions are tuned to coincide with the peaks of the periodograms presented in Section 2.5: $\psi = \frac{2\pi}{365.25}$ and $\omega = \frac{2\pi}{7}$. The number of harmonics n_w and n_d are hyperparameters of the model.

The exogenous term $\boldsymbol{h}$ includes the temperature on the target date d, its appropriate transformation, either HDD or HCDD, and the calendar features except weekday, which is already captured by $\boldsymbol{s}$.

The optimal parameters $\hat{\boldsymbol{\theta}}_s$ and $\hat{\boldsymbol{\theta}}_h$ are obtained via least squares, that is solving the parametric empirical risk minimization problem (2.15) under the loss function (2.16).

Finally, to take into account the most recent available data, the long-term model is corrected with the gas demand of the previous day:

$$\hat{y_d} = \hat{f}_l(x_d)\frac{y_{d-1}}{\hat{f}_l(x_{d-1})}. \tag{2.21}$$

Support vector regression

The support vector regression (SVR) assumes that the parametric candidate function f is in the form

$$f(\boldsymbol{x}) = \boldsymbol{b}^\top \phi(\boldsymbol{x}) + a, \tag{2.22}$$

where the vector $\boldsymbol{b}$ and the scalar a are parameters, resulting in $\boldsymbol{\theta} = [\boldsymbol{b}^\top, a]^\top$, and $\phi : \mathbb{R}^m \to \mathbb{R}^p$ is a function mapping the input vector into an higher-dimensional space. The idea is that non-linear relationships in the feature space $\mathbb{R}^m$ can be described by a linear function after the projection into $\mathbb{R}^p$. In case ϕ is the identity, SVR is a linear model, while different transformations result in non-linear models.

Although other choices are possible, the most common loss function for SVR is Vapnik's ϵ-insensitive function, defined by

$$l(f(\boldsymbol{x}), y) = \begin{cases} 0 & |y - f(\boldsymbol{x})| < \epsilon \\ |y - f(\boldsymbol{x})| & |y - f(\boldsymbol{x})| \geq \epsilon \end{cases}, \tag{2.23}$$

where ϵ is an hyperparameter. In order to find the optimal parameters $\hat{\boldsymbol{\theta}}$, SVR does not resort to empirical risk minimization, but to the structural risk minimization principle [73, Chap. 12], resulting in

$$\hat{\boldsymbol{\theta}}, \boldsymbol{v}^*, \boldsymbol{u}^* = \arg\min_{\boldsymbol{\theta},\boldsymbol{v},\boldsymbol{u}} \|\boldsymbol{\theta}\|_2^2 + \gamma \sum_{i=1}^{n} (v_i + u_i) \tag{2.24a}$$

$$\text{subject to:} \quad \boldsymbol{b}^\top \phi(\boldsymbol{x}_i) + a - y_i \leq u_i + \epsilon \quad \forall i \tag{2.24b}$$

$$y_i - \boldsymbol{b}^\top \phi(\boldsymbol{x}_i) - a \leq v_i + \epsilon \quad \forall i \tag{2.24c}$$

$$v_i \geq 0 \quad \forall i \tag{2.24d}$$

$$u_i \geq 0 \quad \forall i \tag{2.24e}$$

In (2.24), the slack vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ in (2.24b) and (2.24c) penalize only the errors which are larger than ϵ, the cost function (2.24a) imposes an additional penalty on the ℓ_2 norm of the parameters and the scalar constant γ controls the trade-off between the two penalties. The optimization problem is usually solved in its dual form: using duality, it can be shown that an equivalent formulation for the optimal function $\hat{f}$ is

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} \hat{\alpha}_i k(\boldsymbol{x}, \boldsymbol{x}_i), \tag{2.25}$$

where $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{x}_j)$ is called kernel function and $\hat{\boldsymbol{\alpha}}$ is the vector of optimal Lagrange multiplier and represents an equivalent parametrization of $\hat{\boldsymbol{\theta}}$.

The hyperparameters of SVR are ϵ and γ, while the kernel also affects the characteristics of the model. A popular and flexible choice, adopted also in this study, is the Gaussian kernel

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{\frac{\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2}{\delta^2}}, \tag{2.26}$$

where the scalar δ becomes a third hyperparameter.

Artificial neural network

Artificial neural networks (ANN) are non-linear models capable of capturing complex patterns and relations. A comprehensive explanation of their structure and the most common training algorithms can be found in the textbook by Goodfellow et at. [74]. In this dissertation, we focused on the multi-layer perceptron (MLP) or fully connected ANN. Different architectures, like convolutional neural networks (CNN) and recurrent neural networks (RNN), were also considered and are the main topic of Andrea Marziali's PhD thesis [75].

The MLP uses different nested layers of linear and non-linear transformations to build a parameteric candidate function. For example, a three-layer MLP models the prediction function

$$f(x) = g(W_1 g(W_2 g(W_3 \boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2) + \boldsymbol{b}_3), \tag{2.27}$$

where the weight matrices W_1, W_2, and W_3 and the bias vectors $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, and $\boldsymbol{b}_3$ are parameters. The size of the weight matrices and bias vectors are considered hyperparameters and, despite recent progresses in the field of automatic ANN tuning, are most often chosen by trial and errors.

A common metaphore associates the ANN to the structure of the human brain: using such metaphore, each layer has a number of neurons equal to the size of its bias vector. The selection of the non-linear function g also affects the model: the Rectified Linear Unit (ReLu) is a natural choice for regression problems and, due to its piecewise-linear nature, yields a considerable speedup in the training procedure.

Several loss functions can be adopted, the most popular one for regression problems - and the one we chose in this work - is the Mean Squared Error (MSE). Solving the empirical risk minimization for an ANN is a non-convex optimization problem. First-order methods are usually exploited: in this study, we used the Adaptive Moment Estimation (ADAM) algorithm [76].

ADAM includes some additional hyperparameters, like the number of training epochs and the learning rate, which, like the structure of the network, are often chosen by trial and errors.

Random forest

The random forest (RF) is a non-parametric model based on the classification and regression trees (CART) [77]. CARTs perform a recursive feature-wise partitioning of the input space: each split is based on a subset of inputs chosen by minimizing a suitable criterion, which is usually the variance of the target variable for regression problems. When the tree grows to the maximum allowed depth - an hyperparameter - or no more splitting is possible, the outputs of train samples are averaged in each region of the final partition.

CARTs are known to be unstable and prone to overfitting. In order to overcome these limitations, random forest models grow multiple CARTs, resorting to data and feature bagging. Bagging or bootstrap aggregating is the practice of randomly selecting a subset of the dataset: in the case of random forests, bagging is repeated with replacement for each CART. By applying bagging to both data and features, each tree gets trained on different samples and feature sets. Forecasts performed by all the individual CARTs are then averaged to get the final prediction, leading to a more stable predictor.

Nearest neighbours

The k-nearest neighbours (KNN) is a simple non-parametric model which relies on the distance between samples in the feature space. Given a test sample $\boldsymbol{x}$, $\hat{y} = \hat{f}(\boldsymbol{x})$ is computed by averaging the output of the k training samples y_i, $i = 1 \dots k$, whose feature vectors $\boldsymbol{x}_i$ that are closest to $\boldsymbol{x}$, according to some distance measure. The average can be either arithmetic or weighted.

In order to specify a KNN estimator, one has to choose the distance metric, e.g. Euclidean, Minkowsky, Manhattan, the type of weighted average, e.g. uniform or inverse distance, and the number k of neighbours. Too small values of k lead to overfitting to the training data, while including too many neighbors jeopardizes the flexibility of the model.

Gaussian process

The Gaussian process (GP) is a flexible non-parametric model which aims at directly obtaining the predictive function rather than inferring its parameters. The basic hypothesis of GP is that any two samples y_i, y_j of the output vector $\boldsymbol{y} \in \mathbb{R}^n$ follow a joint Gaussian distribution $\mathcal{N}(0, k(\boldsymbol{x}_i, \boldsymbol{x}_j))$, where k is the kernel function. Moreover, it is assumed that $\boldsymbol{y}$ is corrupted by Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbb{I}_n)$, so that $y_i = f(\boldsymbol{x}_i) + \epsilon_i$ for each sample in the training set.

Let $X = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \dots, \boldsymbol{x}_n^\top]^\top$ denote the input matrix and $K(X, X)$ the covariance matrix of $\boldsymbol{y}$, so that

$$[K(X,X)]_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{2.28}$$

Let also denote with $\boldsymbol{x}$ the inputs of a new sample, and $\hat{\boldsymbol{y}}_s$ the stochastic variable associated with the predicted output. The joint distribution of $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}_s$ is

$$\begin{bmatrix} \boldsymbol{y} \\ \hat{\boldsymbol{y}}_s \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X) + \sigma^2 \mathbb{I}_n & K(\boldsymbol{x}, X) \\ K(X, \boldsymbol{x}) & k(\boldsymbol{x}, \boldsymbol{x}) \end{bmatrix}\right). \tag{2.29}$$

Then, GP uses the expectation of $\hat{\boldsymbol{y}}_s$ as the point estimate $\hat{\boldsymbol{y}}$:

$$\hat{\boldsymbol{y}} = \mathbb{E}[\hat{\boldsymbol{y}}_s] = K(\boldsymbol{x}, X)(K(X,X) + \sigma^2 \mathbb{I}_n)^{-1} \boldsymbol{y}. \tag{2.30}$$

As with SVR, the kernel determines the characteristics of a GP model: a popular and flexible choice is the Matérn kernel, a generalization of the family of radial basis functions defined by

$$k(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \Big(\frac{\sqrt{2\nu}r}{l}\Big)^{\nu} K_\nu\Big(\frac{\sqrt{2\nu}r}{l}\Big), \quad r = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2, \tag{2.31}$$

where K_ν is a modified Bessel function, Γ is the gamma function and the scalars ν and l are hyperparameters.

Different approaches can be followed to estimate ν, l, and σ^2. We adopted the empirical Bayes method and set the hyperparameters as the optimizer of the marginal log likelihood

$$\log p(\boldsymbol{y}|X, \nu, l, \sigma) = -\frac{1}{2}\boldsymbol{y}^\top (K(X,X) + \sigma^2 \mathbb{I}_n)^{-1} \boldsymbol{y} - \frac{1}{2}\log(K(X,X) + \sigma^2 \mathbb{I}_n), \tag{2.32}$$

up to a constant, where K depends on ν and l as per (2.28) and (2.31).

2.7.5 Ensemble Models

Four aggregation techniques were considered: three out of four need the calibration of additional parameters. The training set $\mathcal{S}$ is exploited to identify the parameters of the base models and cannot be used again, while the test set $\mathcal{T}$ is intended for the sake of the final evaluation only. Thus, a third dataset, disjoint by both $\mathcal{S}$ and $\mathcal{T}$ is needed: we call it the validation set $\mathcal{V}$.

Simple Average

The most trivial aggregation is the arithmetic average of the forecasts achieved by base models. Given a test input $\boldsymbol{x}$, and b predicted outputs by the base models $\hat{f}_i(\boldsymbol{x})$, $i = 1, \ldots, b$, the ensemble forecast is

$$\hat{f}^{\mathrm{A}}(\boldsymbol{x}) = \frac{1}{b} \sum_{i=1}^{b} \hat{f}_i(\boldsymbol{x}). \tag{2.33}$$

Despite its simplicity, this ensembling methods is very stable and often outperforms more complex approaches [78].

Weighted Average

A second option is the weighted average of base forecasts:

$$\hat{f}^{\mathrm{W}}(\boldsymbol{x}) = \sum_{i=1}^{b} w_i \hat{f}_i(\boldsymbol{x}), \quad w_i \geq 0, \quad \sum_{i=1}^{b} w_i = 1. \tag{2.34}$$

The weights w_i are obtained by solving a constrained least square problem, where the cost function is the sum of squared residuals between the ensemble forecast and the target vector on $\mathcal{V}$.

Best Subset Average

The third ensemble method computes the average on the best subset of predictors. The best subset is obtained by a exhaustive search: the set of predictors which minimizes a suitable error metric on $\mathcal{V}$ is chosen. Exhaustive search is often deemed too expensive, but, due to the limited amount of base models, the number of combinations to test is relatively small. Excluding the complete subset made of all the nine predictors (already considered as simple average), and the subsets made of a single model, the number of candidate subsets is

$$\sum_{k=2}^{8} \binom{9}{k} = 501. \tag{2.35}$$

SVR Aggregation

The fourth ensemble method trains a SVR model on $\mathcal{V}$, using base forecasts as features. Different models could be chosen in place of the SVR, but empirical evidence suggested that SVR achieves a good balance between accuracy and stability.

2.8 Experiments

In order to compare the thirteen models described in Section 2.7 and assess strengths and weaknesses of each of them, experiments were carried out using the dataset described in Section 2.4.

Due to the presence of ensembling models, which require the estimation of additional hyperparameters, a proper experimental setup had to be put in place in order to guarantee a fair comparison with the base models.

2.8.1 Experimental Setup

The available data ranges from 2007 to 2018. Four one-year-long test sets $\mathcal{T}_a$, $a \in \{2015, 2016, 2017, 2018\}$, were used in order to perform a comprehensive evaluation over multiple years. Each test set was associated to a set of training data, that was organized differently depending on the nature of the considered model, either base or ensemble.

Training of Base Models

Before the training, all the continuous series in the dataset were normalized in the range $[0, 1]$. This procedure is required by many models, like ANN, whose training algorithms assume that all the inputs have the same order of magnitude. At the end of the prediction, the inverse transformation was applied to the outputs.

The base training set $\mathcal{S}_a^{\text{base}}$, is made of the samples previous to the beginning of the test set $\mathcal{T}_a$. For instance, base models tested on $\mathcal{T}_{2017}$ were trained on the $\mathcal{S}_{2017}^{\text{base}}$, containing data ranging from 2007 to 2016 inclusive.

Hyperparameters of the torus model were tuned by maximizing the Akaike Information Criterion (AIC), those of the Gaussian process by maximizing the marginal likelihood, while for the other base models five-fold cross validation was adopted. For each series and training set, a large and coarse grid of possible values was initially considered for each hyperparameter, then a finer grid was adopted in order to choose the final value. The tuning of hyperparameters only exploited data in the training set.

Training of Ensemble Models

In order to train ensemble models, two distinct sets were assembled. A one-year-long validation set $\mathcal{V}_a$, encompassing the year before a was used to tune the hyperparameters of the ensembling procedures, while the remaining data, forming the training set $\mathcal{S}_a^{\text{ens}}$, were used to train the base models that enter the aggregation.

For instance, if $\mathcal{T}_{2017}$ is the test set, the base models were trained on $\mathcal{S}_{2017}^{\text{ens}}$, ranging from 2007 to 2015 inclusive, while the ensemble models were trained on $\mathcal{V}_{2017}$, containing the samples of 2016.

The proposed organization ensures that both base and ensemble model use the same amount of data for training, and have thus access to an equal share of information. A visual representation of the sets is displayed in Fig. 2.21.

Performance Evaluation

Three metrics were used to evaluate the performance of the models: the mean absolute error (MAE), mean absolute percentage error (MAPE) and root mean

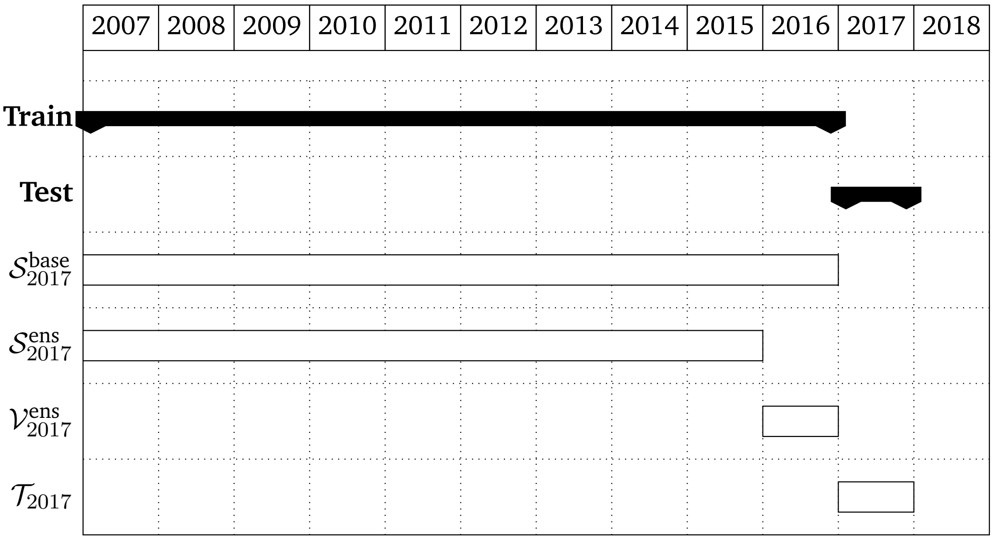

2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
Train
Test
$\mathcal{S}_{2017}^{\text{base}}$
$\mathcal{S}_{2017}^{\text{ens}}$
$\mathcal{V}_{2017}^{\text{ens}}$
$\mathcal{T}_{2017}$

Figure 2.21. Structure of training, validation and test sets for base and ensembling model relative to the test set $\mathcal{T}_{2017}$.

square error (RMSE). For a given $\mathcal{T}_a$, let $q = |\mathcal{T}_a|$ denote its cardinality. Then, the error metrics are given by

$$\text{MAE} = \frac{1}{q}\sum_{d=1}^{q} |y_d - \hat{y}_d|, \tag{2.36a}$$

$$\text{MAPE} = \frac{1}{q}\sum_{d=1}^{q} \frac{|y_d - \hat{y}_d|}{|y_d|}, \tag{2.36b}$$

$$\text{RMSE} = \sqrt{\sum_{d=1}^{q} (y_d - \hat{y}_d)^2}. \tag{2.36c}$$

Among the three metrics, MAE is the most relevant for this study, as it is a good proxy for the financial penalties caused by errors in forecasting. Moreover, relative metrics, such as MAPE, overweight errors during low-demand seasons, while it is during the high-demand cold periods that the economic consequences of inaccurate predictions are the most severe.

2.8.2 Experimental Results

Hyperparameters

The results of the hyperparameter tuning offer interesting insights about the models and the features in the dataset.

For ridge regression, lasso and elastic net, the tuning of the regularization parameter λ yielded values between $\lambda_{\max} = 0.236$ and $\lambda_{min} = 10^{-4}$. To interpret such results, it is convenient to introduce the notion of effective degrees of freedom $\mathrm{df}(\lambda)$. Intuitively, the effective degrees of freedom represent the number of features of the original dataset used by the regularized model, taking into account the shrinkage of the parameters discussed in Section 2.7. In the case of the ridge regression, $\mathrm{df}(\lambda) = \mathrm{tr}(H_\lambda)$, where H_λ is a matrix dependent on λ such that $\hat{\boldsymbol{y}} = H_\lambda \boldsymbol{y}$, while the definitions for lasso and elastic net are more involved - the interested reader may find a detailed explanation in [73, Chap. 3.4]. The effective degrees of freedom corresponding to the chosen values of λ were comprised between 20.94 and 20.99. This means that all the 21 available features were deemed relevant for prediction and regularization played a marginal role.

For what concerns the torus model, the minimization of the AIC led to the choice of $n_w = 3$ for RGD and TGD, $n_w = 4$ for IGD, $n_d = 1$ for RGD and TGD, and $n_w = 3$ for IGD for all the training sets. The requirement for more harmonics for IGD may be due to the higher importance of the periodic structure imposed by weekends and vacations with respect to the influence of the weather.

For SVR, $\gamma = 50$ and $\sigma = 0.1$ were selected for each series and training set, while the value of ϵ oscillated between $4.5 \cdot 10^{-4}$ and $1.5 \cdot 10^{-3}$, with no clear pattern among the series and the training sets.

For the ANN models, a trial and error procedure led to a three-layer architecture with 24, 12, and 4 neurons. More complex structures led to overfitting and loss of predictive performance. By cross-validation, we obtained a learning rate of 0.001 and a batch size of 32. We selected 1000 epochs for training by observing the evolution of the loss function on train and validation sets.

For KNN, we optimized the number of neighbours as well as the weighting strategy, choosing between uniform and inverse of the distance. We obtained optimal numbers of neighbours between 5 and 9, while the "inverse of distance" weights were consistently selected for all the series and training sets.

For the Gaussian process, the maximization of the marginal likelihood yielded $\nu = 1.5$, $l = 10$, and $\sigma^2 = 10$, with minimal variations among all the series and training sets.

Finally, for random forest, the cross-validation selected 500 trees for all series and training sets, a maximum number of features to consider for each split between 12 and 21 and a maximum depth of the tree between 14 and 19. The high values in the maximum number of features suggest that taking into account all the features in the dataset yields to a more accurate characterization of each point, coherently with what already observed for linear models.

Performance

Mean absolute errors (MAEs) for RGD, IGD, TGD, and total GD are shown in Tables 2.4 to 2.7.

Among base models, GP, ANN and SVR achieved the lowest average MAE across all the types of gas demand, with differences between each other smaller than 0.10 millions of standard cubic meters (MSCM). Notably, the performance was also stable across all the test sets. On the other hand, KNN was consistently the worst performer, due to its poor capability of capturing the influence of temperature and holidays. The three regularized linear models, lasso, ridge regression and elastic net, achieved an almost identical performance, a further confirmation of the marginal role of regularization. Moreover, they never matched the accuracy of more complex, non-linear models, thus certifying the consensus of the literature.

As noted in Section 2.5, RGD presents peculiar characteristics: it is mainly driven by temperature during the winter, while being periodic during the summer. It is thus of interest to disaggregate the performance at a monthly level and investigate which models achieve the lowest error in different periods. Table 2.2 reports the monthly averages of MAE throughout the test sets. It appears that GP is the best performer during the warm period, especially from June to October, whereas in the cold months, from December to February, the non-linear models SVR and ANN are more accurate. A possible explanation is that GP is better at capturing the effects of the weekly seasonality, that explains most of the variability during the summer, while ANN and SVR better account for the non-linear effect of temperature, only relevant during the cold months. In this respect, the result may suggest that the HDD transformation is not completely adequate in linearizing the influence of temperature. The analysis of the MAPEs, reported in Table 2.3, yields similar conclusions. Moreover, the smaller values of relative error recorded during the summer by all models suggest that the periodic pattern of RGD in the warm season is relatively easier to capture than the more irregular one shown in the rest of the year.

Ensemble models consistently outperformed base ones. In particular, subset average achieved the best average MAE on all the considered series: RGD, IGD, TGD and GD. A possible explanation builds on the aforementioned differences between models: while different methods are better at capturing specific behaviours, aggregation

Model	1	2	3	4	5	6	7	8	9	10	11	12
Ridge	5.44	4.98	*4.90*	3.58	2.16	1.34	1.07	2.17	1.71	2.49	3.59	5.22
Lasso	5.45	4.99	*4.90*	3.58	2.16	1.34	1.07	2.18	1.71	2.49	3.59	5.22
Elastic net	5.45	4.99	4.91	3.58	2.16	1.34	1.07	2.18	1.70	2.49	3.59	5.22
SVR	*4.93*	*4.34*	5.00	3.15	*1.03*	0.75	0.41	1.52	0.64	2.06	3.67	4.91
GP	5.29	4.89	5.17	*3.04*	1.09	***0.42***	*0.37*	***0.71***	*0.42*	*1.85*	3.77	4.90
KNN	11.38	8.65	9.14	6.46	1.85	0.90	0.55	0.95	0.77	3.99	9.22	8.90
Random forest	6.57	5.43	5.55	4.41	1.76	0.73	0.49	0.76	0.65	2.55	4.44	6.94
Torus	6.03	5.86	5.18	3.43	1.56	1.12	0.48	0.94	0.52	1.94	*3.47*	4.87
ANN	5.33	4.75	5.41	3.33	1.09	0.64	0.44	1.04	0.65	2.25	3.80	*4.85*
Simple average	5.11	4.69	4.85	3.36	1.06	0.70	0.41	1.10	0.70	1.84	3.52	4.67
Weighted average	4.54	4.52	4.60	3.14	1.03	0.61	0.36	0.77	0.40	1.77	3.16	4.23
SVR aggregation	4.76	4.52	**4.55**	3.02	0.98	0.58	**0.31**	0.75	**0.39**	1.86	3.27	4.30
Subset average	**4.16**	**4.16**	4.59	**2.98**	**0.94**	0.57	0.36	0.90	0.44	**1.63**	**3.13**	**3.94**

Table 2.2. Monthly MAEs [MSCM] on Residential Gas Demand test sets 2015-2018: best performers are in **boldface blue**, while the best among base models are in *italics*.

Model	1	2	3	4	5	6	7	8	9	10	11	12
Ridge	3.00	2.85	*3.86*	6.08	5.25	4.21	3.61	8.98	4.79	4.75	*2.96*	3.02
Lasso	3.00	2.85	3.87	6.08	5.25	4.21	3.61	9.00	4.79	4.75	2.96	3.03
Elastic net	3.00	2.85	3.87	6.08	5.25	4.22	3.61	9.00	4.78	4.74	2.96	3.03
SVR	*2.70*	*2.52*	4.04	5.30	*2.36*	2.36	1.33	6.26	1.74	3.46	3.19	2.89
GP	2.91	2.88	4.11	4.89	2.48	***1.29***	*1.21*	***2.96***	*1.18*	*3.07*	3.21	2.88
KNN	6.25	4.96	7.42	11.34	4.37	2.82	1.81	3.80	2.12	6.54	7.76	5.13
Random forest	3.59	3.21	4.54	7.64	4.24	2.30	1.59	2.96	1.78	4.27	3.82	4.12
Torus	3.36	3.35	4.07	5.28	3.42	3.39	1.57	3.86	1.44	3.38	3.03	2.83
ANN	2.96	2.73	4.34	5.48	2.49	1.94	1.46	4.31	1.77	3.96	3.32	*2.82*
Simple average	2.81	2.70	3.88	5.49	2.47	2.16	1.37	4.53	1.94	3.23	2.88	2.72
Weighted average	2.51	2.64	3.69	5.00	2.31	1.84	1.17	3.15	1.11	3.01	2.72	2.47
Subset average	**2.28**	**2.43**	3.69	**4.73**	**2.09**	1.73	1.18	3.68	1.22	**2.79**	**2.70**	**2.30**
SVR aggregation	2.63	2.64	**3.61**	4.79	2.21	1.76	**1.01**	3.05	**1.07**	3.12	2.79	2.52

Table 2.3. Monthly MAPEs [%] on Residential Gas Demand test sets 2015-2018: best performers are in **boldface blue**, while the best among base models are in *italics*.

can mitigate the errors committed by individual models, thus increasing the overall accuracy and robustness.

The improvement due to aggregation was particularly evident for RGD (Table 2.4) and IGD (Table 2.5), where the best base model was outperformed by the best ensemble model by 0.25 MSCM or 11% and 0.07 MSCM or 12% respectively. The gap between base and ensemble models was smaller for TGD (Table 2.6), with subset average outperforming SVR by 0.09 MSCM or 2.3%. Finally, SVR is worse than subset average by 0.29 MSCM or 5.7% for the global Italian GD (Table 2.7). Such improvements may not appear significant, but in a competitive environment such as energy markets, lowering the error by a few percentage points may lead to important financial gains and to notable reduction in unused pipe capacity.

The forecasts for 2018 and the corresponding residuals provided by subset average, the best ensemble predictor, are displayed in Fig. 2.22 and Fig. 2.23, respectively. From the time series of the residuals, it is possible to appreciate the difference between RGD and TGD or IGD. While the latter have residuals of constant magnitude during the year, the former shows larger residuals in the cold season. This is in agreement with the analysis of Section 2.5 and underlines the importance of an accurate model of the relationship between weather and residential gas demand.

To the best of our knowledge, the only term of comparison for the task is the forecasts of the Italian GD issued by SNAM Rete Gas, the Italian Transmission System Operator (TSO) [79]. In 2017 and 2018, the improvement achieved by our method is neat: the out-of-sample MAE of SNAM predictions was 9.62 MSCM in 2017 and 8.30 MSCM in 2018, while our best model (subset average) scored 5.16 MSCM in 2017 and 5.46 MSCM in 2018 (Table 2.7).

2.9 Effect of Weather Forecasting

We now focus on RGD. In Section 2.5 it was shown that temperature is the most important driver of RGD and thus one of the most useful inputs for a predictive model. Unfortunately, the actual temperature cannot be used in forecasting tasks: only weather forecasts are available and they are affected by a small yet non-negligible error which also impacts the performance of the gas demand prediction.

An interesting and relevant problem is thus assessing the influence of the temperature error on the accuracy of RGD models. For this purpose, we resort to an idealized

Model	2015	2016	2017	2018	Average
Ridge	3.39	3.10	3.01	3.49	3.25
Lasso	3.38	3.10	3.01	3.49	3.25
Elastic net	3.38	3.10	3.01	3.49	3.25
SVR	2.84	2.62	2.38	2.93	2.69
GP	2.60	2.48	2.51	2.61	2.55
KNN	4.57	5.51	5.08	5.52	5.17
Random forest	3.04	3.36	3.50	3.48	3.35
Torus	3.18	2.66	2.54	3.13	2.88
ANN	2.76	2.68	2.43	3.10	2.74
Simple average	2.66	2.57	2.45	2.91	2.65
Weighted average	2.59	2.33	**2.06**	2.64	2.40
Subset average	**2.41**	**2.17**	**2.06**	**2.56**	**2.30**
SVR aggregation	2.58	2.30	2.19	2.67	2.44

Table 2.4. Out-of-sample MAEs for Residential Gas Demand: best performers are in **boldface blue**.

Model	2015	2016	2017	2018	Average
Ridge	0.75	0.75	0.74	0.77	0.75
Lasso	0.75	0.75	0.74	0.77	0.75
Elastic Net	0.75	0.75	0.74	0.77	0.75
SVR	0.57	0.58	0.7	0.75	0.65
GP	0.61	0.61	0.68	0.70	0.65
KNN	1.46	1.25	1.95	1.23	1.47
Random Forest	0.78	0.86	0.95	0.83	0.86
Torus	0.96	0.97	1.05	1.10	1.02
ANN	0.66	0.80	0.57	0.74	0.69
Simple average	0.60	0.62	0.69	0.66	0.64
Weighted average	**0.55**	**0.55**	0.65	0.70	0.61
Subset average	0.56	0.56	0.58	**0.61**	**0.58**
SVR aggregation	0.57	0.79	**0.57**	0.81	0.68

Table 2.5. Out-of-sample MAEs for Industrial Gas Demand: best performers highlighted in **boldface blue**.

Model	2015	2016	2017	2018	Average
Ridge	3.73	4.15	4.26	4.48	4.15
Lasso	3.73	4.15	4.26	4.49	4.16
Elastic Net	3.73	4.15	4.26	4.49	4.16
SVR	3.41	3.64	4.33	4.33	3.93
GP	3.49	3.70	4.39	4.34	3.98
KNN	6.13	5.22	5.83	5.54	5.68
Random Forest	4.66	4.43	4.87	4.84	4.70
Torus	3.98	4.48	4.96	4.94	4.59
ANN	3.40	3.97	4.32	4.41	4.03
Simple average	3.50	3.75	4.21	4.36	3.96
Weighted average	3.35	3.71	4.31	4.31	3.92
Subset average	**3.26**	3.65	**4.17**	**4.26**	**3.84**
SVR aggregation	3.38	**3.62**	4.28	4.37	3.91

Table 2.6. Out-of-sample MAEs for Thermoelectric Gas Demand: best performers highlighted in **boldface blue**.

Model	2015	2016	2017	2018	Average
Ridge	6.32	6.34	5.80	6.57	6.26
Lasso	6.32	6.34	5.81	6.57	6.26
Elastic Net	6.32	6.35	5.81	6.57	6.26
SVR	5.23	5.05	5.55	5.85	5.42
GP	5.33	5.23	5.88	5.82	5.57
KNN	9.04	9.31	9.97	9.83	9.54
Random Forest	6.58	6.45	7.15	7.11	6.82
Torus	6.56	6.47	6.40	7.00	6.61
ANN	5.43	5.50	5.47	6.08	5.62
Simple average	5.53	5.40	5.56	5.98	5.61
Weighted average	5.27	5.01	5.34	5.55	5.29
Subset average	**5.02**	**4.80**	**5.23**	**5.46**	**5.13**
SVR aggregation	5.19	4.91	5.29	5.79	5.30
SNAM forecast	n.a.	n.a.	9.62	8.30	n.a.

Table 2.7. Out-of-sample MAEs for Italian gas demand: best performers highlighted in **boldface blue**.

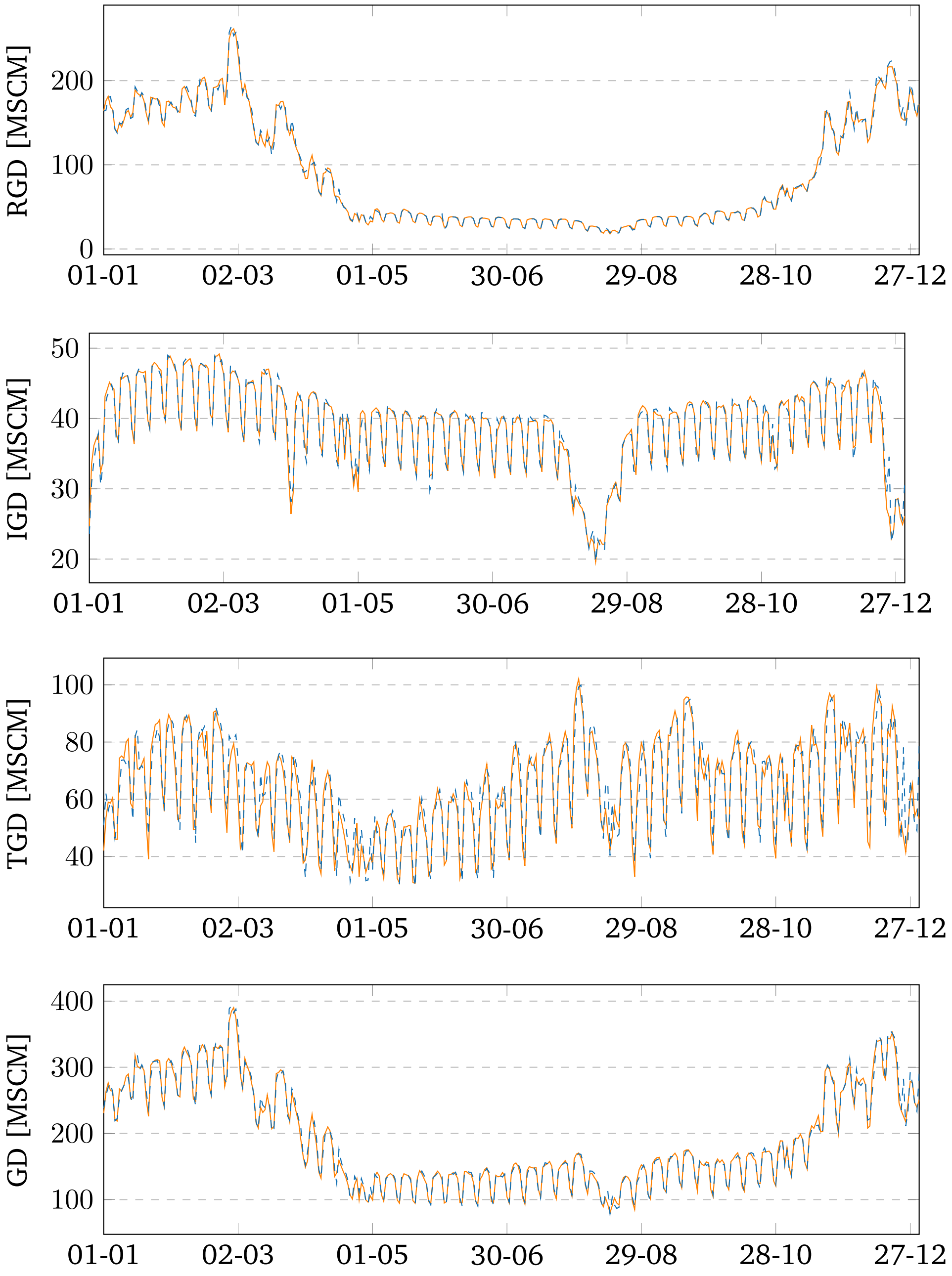


Figure 2.22. Subset average ensemble model: predicted gas demands in 2018. From top to bottom: residential gas demand (RGD), industrial gas demand (IGD), thermoelectic gas demand (TGD), overall gas demand (GD).

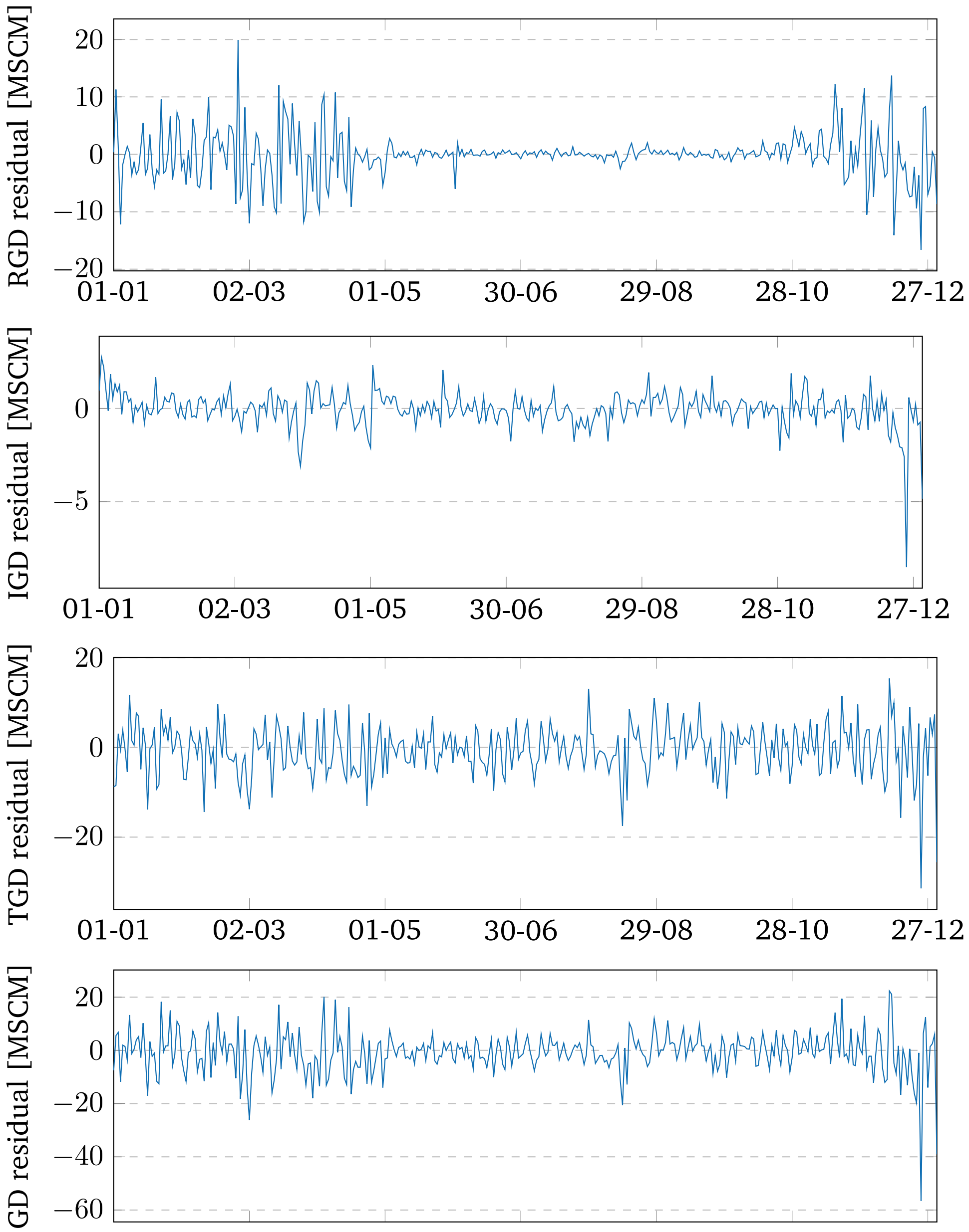


Figure 2.23. Subset average ensemble model: residuals $y_d - \hat{y}_d$ of one-day-ahead prediction in 2018. From top to bottom: residential gas demand (RGD), industrial gas demand (IGD), thermoelectic gas demand (TGD), overall gas demand (GD).

error propagation model that, despite its simplicity, provides an accurate description, as confirmed by the subsequent experimental validation.

2.9.1 Error Propagation

Let RGD be a deterministic function g of the true temperature $\tilde{t}$ and other factors $\boldsymbol{x} = (x_1, x_2, \ldots)$: $\text{RGD} = g\left(\tilde{t}, \boldsymbol{x}\right)$. In view of the analysis and the charts presented in Section 2.5, a first-order approximation of the relationship between RGD and $\tilde{t}$ is a linear function of HDD (Definition 2), while the dependence on the other factors can be represented as an additive term $\bar{g}(\boldsymbol{x})$:

$$\text{RGD} = g\left(\tilde{t}, \boldsymbol{x}\right) = \bar{g}\left(\boldsymbol{x}\right) + \alpha\text{HDD}(\tilde{t}), \tag{2.37}$$

where α is the sensitivity of the gas demand to HDD.

The formula (2.37) applies to each region where domestic heating is the main driver of RGD. Notably, it is not restricted to country-level models and applies also to more restricted areas. Indeed, α depends on the size of the pool of end users and can be estimated from historical data.

Consider now the ideal case when α and $\bar{g}$ are perfectly known, yet, only a forecast t of the correct temperature $\tilde{t}$ is available. The uncertainty in the weather forecast is modelled by a zero-mean additive error ϵ with variance σ_ϵ^2:

$$t = \tilde{t} + \epsilon \tag{2.38}$$

The optimal forecast $\hat{\text{RGD}}$, given t, is therefore

$$\hat{\text{RGD}} = \bar{g}\left(\boldsymbol{x}\right) + \alpha\text{HDD}(t) \tag{2.39}$$

In order to obtain the mean squared error of $\hat{\text{RGD}}$, we first compute its conditional variance:

$$\text{Var}\left[\hat{\text{RGD}} \mid \tilde{t} \geq t_h\right] = \text{Var}\left[\bar{g}\left(\boldsymbol{x}\right) + \alpha \cdot 0\right] = 0 \tag{2.40a}$$

$$\text{Var}\left[\hat{\text{RGD}} \mid \tilde{t} < t_h\right] = \text{Var}\left[\bar{g}\left(\boldsymbol{x}\right) + \alpha\left(t_h - t\right)\right] = \alpha^2\sigma_\epsilon^2. \tag{2.40b}$$

Since $\mathrm{E}[\epsilon] = 0$, it follows that $\mathrm{E}[\mathrm{R\hat{G}D}] = \mathrm{RGD}$. Thus:

$$\begin{aligned}\mathrm{E}\left[\left(\mathrm{R\hat{G}D} - \mathrm{RGD}\right)^2\right] &= \mathrm{E}\left[\left(\mathrm{R\hat{G}D} - \mathrm{RGD}\right)^2 \mid \tilde{t} \geq t_h\right] P\left(\tilde{t} \geq t_h\right) + \\ &+ \mathrm{E}\left[\left(\mathrm{R\hat{G}D} - \mathrm{RGD}\right)^2 \mid \tilde{t} < t_h\right] P\left(\tilde{t} < t_h\right) = \\ &= 0 + P\left(\tilde{t} < t_h\right) \mathrm{Var}\left[\mathrm{R\hat{G}D} \mid \tilde{t} < t_h\right] = \\ &= P\left(\tilde{t} < t_h\right) \alpha^2 \sigma_\epsilon^2. \end{aligned} \tag{2.41}$$

This last equation provides an estimate of the mean squared error in RGD forecasting ascribable to the inaccuracy in temperature forecasting. Since it was derived under an ideal setting, i.e. α and $\bar{g}$ perfectly known, it provides a lower limit to the precision that can be achieved by the best possible forecaster. Indeed, in a real setting, α must be estimated from the available data and $\bar{g}$ is not known and must be identified by statistical learning methods.

The arguments entering the bound can be estimated from a dataset like the one presented in Section 2.4 with the following procedure:

1. Estimate $P\left(\tilde{t} < t_h\right)$ by computing the ratio between the number of samples such that $\tilde{t} < t_h$ and the total number of available data.
2. Compute α through a least square fit of RGD over $\tilde{t}$.
3. Estimate σ_ϵ^2 as the sample variance of $t - \tilde{t}$.

It must be noted that points 1, 2, and 3 are intended as a-posteriori analyses, thus the real temperature $\tilde{t}$ is available. Considering the Italian RGD and adopting $t_h = 18$ °C as in Section 2.5, in the period from 2015 to 2017 the probability $P\left(\tilde{t} < t_h\right)$ ranges between 54% and 67%, while σ_ϵ^2 between 0.05 and 0.09 °C, and α between 9.85 and 10.96 MSCM/°C. Considering the period from 2015 to 2017 inclusive altogether, we have $P\left(\tilde{t} < t_h\right) = 63\%$, $\sigma_\epsilon^2 = 0.063$ °C, $\alpha = 10.56$ MSCM/°C, corresponding to a best achievable root mean squared error

$$\mathrm{RMSE}^* = \sqrt{P\left(\tilde{t} < t_h\right) \alpha^2 \sigma_\epsilon^2} = 2.22 \ \ \mathrm{MSCM}. \tag{2.42}$$

Finally, we consider the more realistic case in which $\bar{g}$ is unknown and its estimate $\hat{g}$ is affected by error. We model this situation by considering $\hat{g}$ an unbiased estimator of $\bar{g}$ - i.e. $\mathrm{E}[\hat{g}(\boldsymbol{x})] = \bar{g}(\boldsymbol{x})$, with variance

$$\mathrm{Var}\left[\hat{g}\left(\boldsymbol{x}\right)\right] = \sigma_0^2 > 0. \tag{2.43}$$

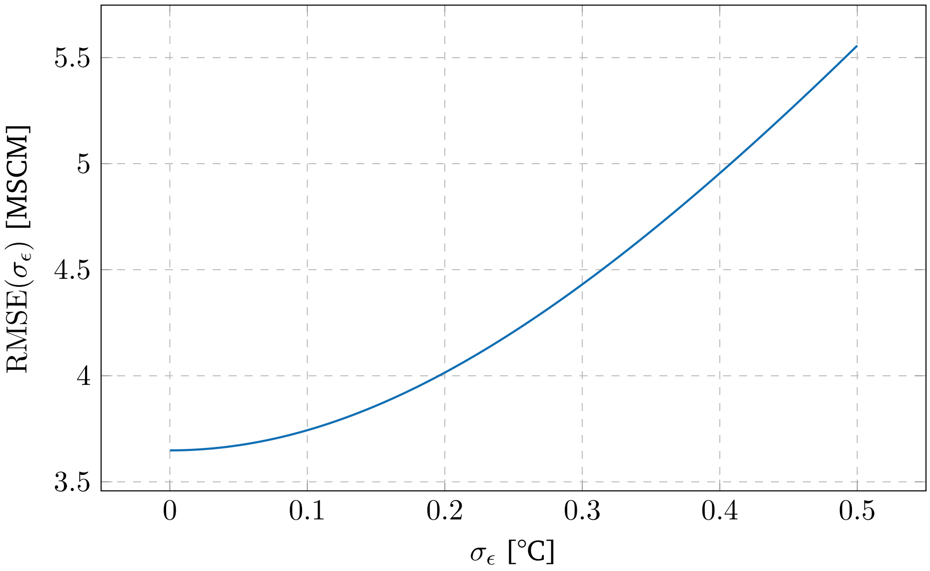


Figure 2.24. RMSE of the forecast for residential gas demand $\mathrm{RMSE}(\sigma_\epsilon^2)$ as a function of the RMSE on temperature forecasts σ_ϵ - equation (2.44).

Then, assuming that ϵ and $\hat{g}$ are statistically independent, it is possible to obtain the RMSE on the prediction $\mathrm{R\hat{G}D}$ as a function of σ_ϵ^2:

$$\mathrm{RMSE}(\sigma_\epsilon) = \sqrt{\mathrm{Var}\left[\mathrm{R\hat{G}D}\right]} = \sqrt{\mathrm{Var}\left[\hat{g}(\boldsymbol{x})\right] + (\mathrm{RMSE}^*)^2}$$
$$= \sqrt{\sigma_0^2 + P\left(\tilde{t} < t_h\right)\alpha^2\sigma_\epsilon^2}. \tag{2.44}$$

In Fig. 2.24 this relationship is displayed assuming $P\left(\tilde{t} < t_h\right) = 63\%$, $\sigma_\epsilon^2 = 0.063$, $\sigma_0^2 = 13.31$ - this last value is the test MSE achieved by the ANN forecaster trained with true temperature data on $\mathcal{T}_{2017}$. Notably, the sensitivity of the gas forecasting error tends to increase as the temperature forecast error grows. In particular, if we define the threshold

$$\bar{\sigma}_\epsilon^2 = \frac{\sigma_0^2}{P\left(\tilde{t} < t_h\right)\alpha^2}, \tag{2.45}$$

the influence of temperature errors is negligible as far as $\sigma_\epsilon^2 \ll \bar{\sigma}_\epsilon^2$, while the temperature errors have a linear influence on the gas RMSE for $\sigma_\epsilon^2 \gg \bar{\sigma}_\epsilon^2$.

2.9.2 Experiments

In order to validate our model for the propagation of temperature errors, we selected five of the nine base models described in Section 2.7, namely ridge regression, Torus, ANN, GP, and KNN and performed two sessions.

Year	2015	2016	2017	2015-2017
Performance limit	2.15	2.02	1.98	2.05
Ridge	4.75	4.58	4.57	4.63
Torus	4.73	4.69	4.20	4.55
ANN	4.45	4.13	3.97	4.19
GP	4.37	4.20	4.15	4.24
KNN	7.60	8.73	8.61	8.33

Table 2.8. Predicted performance on the test sets when temperature forecasts with $\sigma_\epsilon^2 = 0.063$ are used: yearly RMSE [MSCM] of the five models.

In the first one, the models were trained and tested using historical records of true temperatures, assuming that the one-day-ahead exact temperature is available as a feature. Then, we used equation (2.44) in order to predict how much the RMSE of the prediction of RGD would increase in the realistic scenario where temperature forecasts are employed in place of the actual values.

In the second session, the models were trained and tested using historical records of weather forecasts. The same organization of train and test sets described in Section 2.8 was adopted. A perfect match between the RMSE in the two sessions would validate our error propagation model.

The results of the first session are reported in Table 2.8. In the first line, the theoretical performance limits computed with (2.42) are displayed. These values were combined with the RMSE of the predictions based on the actual temperature, using (2.44), to obtain an estimate of the RMSE to be achieved in the real-world situation where weather forecasts are used.

In the second session, the estimates of Table 2.8 were validated by comparing them with the RMSE on RGD forecasting achieved by models fed with weather forecasts. As it can be seen in Table 2.9, the actual errors are in good agreement with their estimates. This can also be appreciated in Fig. 2.25, where the values of Table 2.9 are on the horizontal axis while values from Table 2.8 are on the vertical axis. The closer the points are to the 45° line, the more accurate the error propagation model described by (2.44) is.

2.10 Industrial Applications

All the models proposed in Section 2.7 were implemented and integrated into the IT infrastructure of A2A, with the development of extract, transform and load (ETL)

Year	2015	2016	2017	2015-2017
Ridge	4.68	4.28	4.28	4.42
Torus	5.40	4.33	3.96	4.60
ANN	4.34	4.10	3.64	4.04
GP	4.25	4.12	4.07	4.15
KNN	7.35	8.55	8.37	8.11

Table 2.9. Performance on the test sets: yearly RMSE [MSCM] of the five models trained and tested using the one-day-ahead weather forecast.

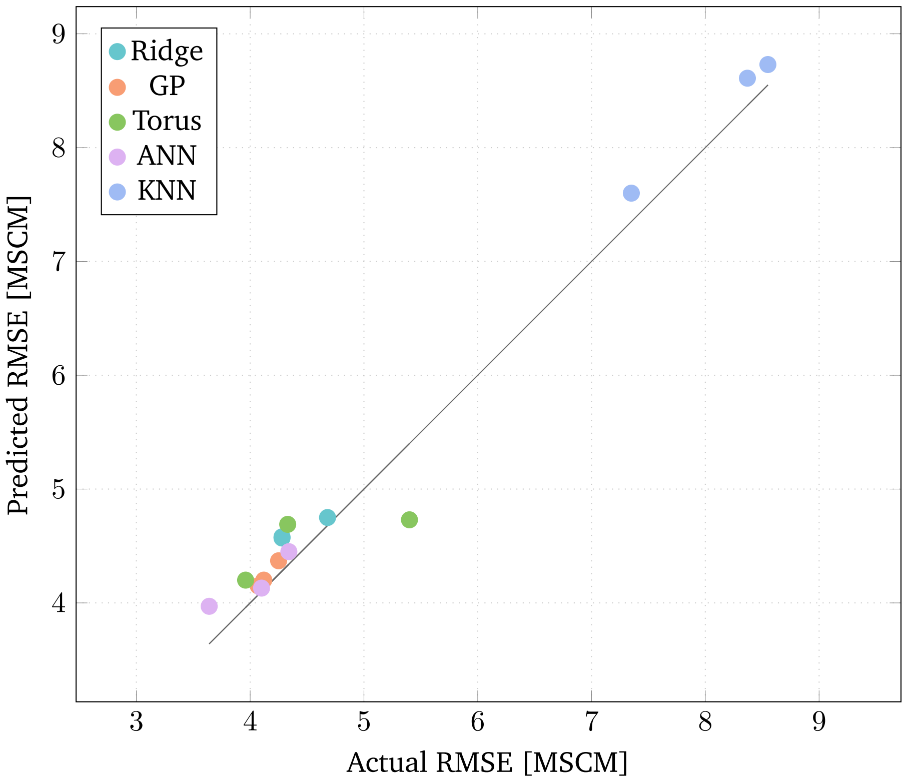


Figure 2.25. Validation of the model predicting effects on gas forecast of temperature forecast errors. Root Mean Square Error on the prediction of RGD: actual values vs theoretical predictions.

jobs to collect, clean and normalize the dataset. The subset average, which resulted the best model in Section 2.8, is currently used to support the daily operations of the utility and its actions on the market. The impact of the improvement in accuracy on the financial Key Performance Indicators (KPIs) of relevant business units was deemed "significant" by the company management, although actual figures cannot be disclosed.

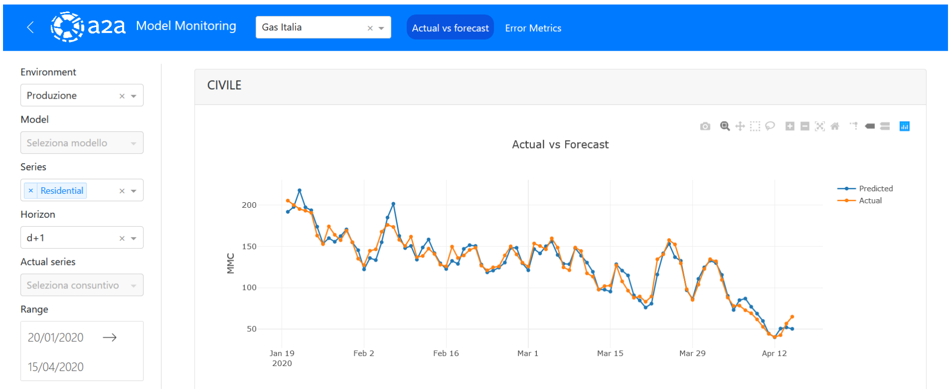
In addition to the pipelines for data extraction, data preprocessing, and model training and prediction, a support monitoring system was developed – a screenshot is displayed in Fig. 2.26. Different from the models implemented only for research purpose, the machine learning systems deployed in an industrial context need to be constantly overseen. Covariate shifts, i.e. drifts in the inputs, or the influence of unmodelled phenomena can affect the predictive performance of the models and result in financial losses, while issues in the IT infrastructure may delay or prevent the production of the forecasts.

A possible solution is to have operators periodically perform manual checks of the data and the model performance, but this practice results in high costs. Therefore, following the guideline proposed by the Google Machine Learning team [80], we implemented a set of automatic checks and a dashboard to easily and reliably control the models.

The system was put to the test in March 2020, when shelter-in-place orders were issued by national authorities to fight the spread of COVID-19. RGD, IGD and TDG were all impacted and the sudden drift in demand was apparent from the model monitoring dashboard. However, due to the reliance on day-before and week-before data, the performace of our predictive models only slightly deteriorated. An example of the performance of a model during the lockdown period is displayed in Fig. 2.26.

Moreover, the results presented in Section 2.9 were considered by A2A in the procurement of weather data in order to attribute an approximate monetary value to the accuracy of weather forecasts.

Finally, the lessons learnt in gas forecasting were applied in the development of new forecasting models for power load, in collaboration with another major Italian utility. The outcome of the year-long project was the creation of predictors for residential and industrial power load at a regional level. In the case of large industrial facilities, the demand of individual plants was also forecasted. Moreover, the system took advantage from the ongoing installation of smart meters, capable of communicating the measurements with a delay of one or two days. For comparison, consumptive



Environment: Produzione

Model: Seleziona modello

Series: Residential

Horizon: d+1

Actual series: Seleziona consuntivo

Range: 20/01/2020 → 15/04/2020

Figure 2.26. A screenshot of the model monitoring system representing actual demand vs day-ahead forecast for RGD between January and April 2020. The country-wide lockdown started on 10th March and ended on 27th April.

data for customers with traditional meters are communicated one to two months after the measurement.

Including in the dataset samples only a few days old yielded drastic improvements and allowed also to enhance the prediction for the customers currently lacking smart meters. On the other hand, ad-hoc data cleaning and missing data management was required to deal with the poor quality of the samples recorded by smart meters. The new system achieved a reduction in MAPE between 20% and 30% with respect to legacy predictors, and contains several elements of interest. Unfortunately, its details could not be published and cannot be discussed in the present work due to commercial confidentiality agreements.

3 Identification of Power Distribution Networks

3.1 Structure of the Power Grid

While natural gas remains an important energy source for Italy and other nations, major effort is being produced to boost the adoption of electric power in applications typically covered by fossil fuels, like transportation and heating. The spread of electric vehicles and heat pumps may further reduce carbon emissions, provided that the growth in demand is met by renewable sources.

However, the rising load and the penetration of distributed generation and storage are projected to pose serious challenges to the power grid. In order to better understand such issues, we will start by presenting the typical structure of the grid.

Fig. 3.1 depicts the most relevant components of the grid and their interconnection. All the lines carry AC current, as it was proven easier and more convenient to transport. Power is typically generated by large plants far from towns or industrial facilities and then transported by high-voltage three-phase transmission lines close to the final users. High-voltage lines are convenient in reducing the losses caused by Joule effect, but cannot be extended too close to cities for safety concerns.

While rare large-scale plants, like steel mills, can be directly connected to the transmission lines, the majority of the industrial facilities and entirety of the household consumer are served by low-voltage local distribution networks. Such grids usually cover a limited geographic area - from a single village to a neighborhood of a large city - and can feature both suspended and underground cables. An intermediate infrastructure between the high-voltage transmission network and the low-voltage distribution network is the medium-voltage grid, which usually covers entire regions and serves several distribution networks. The interface between high-, medium-, and low-voltage networks is provided by substations, which feature large transformers and banks of capacitors.

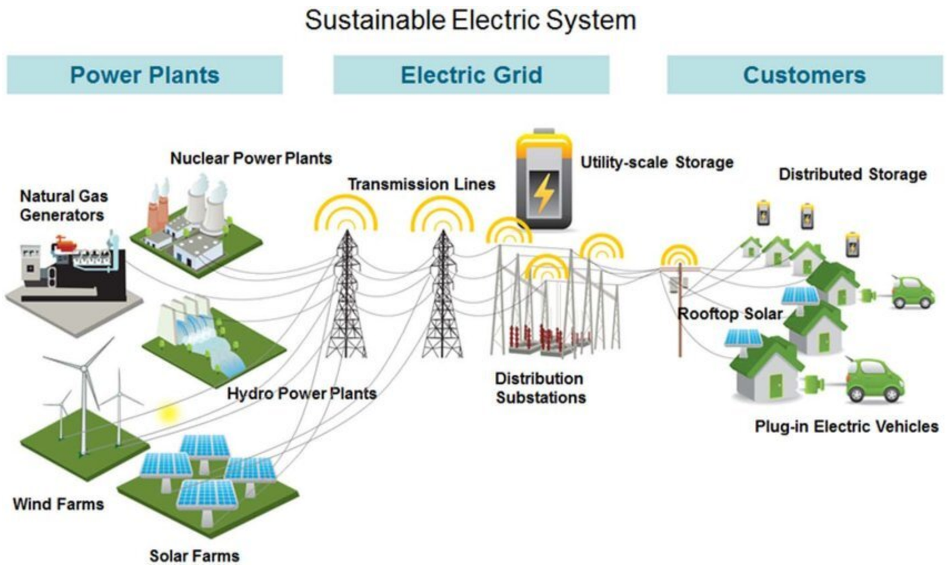


Figure 3.1. Typical structure of a power grid [81].

Until the late 2000s, research focused on the transmission grid, paying little attention to distribution networks. Indeed, careful planning and operation of the transmission system is critical to achieve profitable and reliable transportation of energy. Faults or unbalance in the transmission lines can result in cascading blackouts, with heavy economic consequences. Conversely, issues on distribution networks only affect limited areas and can be easily fixed within a short time.

Due to their importance, transmission lines are usually well known and well monitored. Power transmission system operators, organizations similar in purpose to the TSOs of gas networks, maintain maps of the existing connections and collect data from phasor measurement units (PMUs), typically installed on each high-voltage line. PMUs sample the waveforms of voltage and current at frequencies up to 120 Hz and transmit the data to central control rooms through an appropriate infrastructure.

Protection against failures in the transmission networks is also guaranteed by redundancy. The transmission grid is meshed: in case of a sudden outage of a line, power can be redirected on other branches. Being less critical, distribution networks do not follow the same design and are usually radial, with the substation being the root of the tree and the final customers the leaves.

The situation, however, is changing with the spreading of microgrids. Designed to be autonomously controllable, microgrids are local power networks which can operate both when islanded or connected to the main grid. Microgrids contain distributed energy resources (DERs), including loads (e.g. houses, charging stations), generators (e.g. domestic photovoltaic panels) and storage devices. In order to function as independent systems, they feature their own local controllers. Usually, a hierarchical structure is adopted, with decentralized low-level loops for voltage stability and centralized high-level controllers for economic optimization [82, 83]. To guarantee the delivery of power to all the connected devices even when operating in islanded mode, microgrids are usually meshed and reconfigurable, i.e. they contain switches that can be operated in real-time by the controllers.

3.2 Importance of Distribution Network Identification

By relying on local operation and control, microgrids enjoy favourable properties. The lack of long transmission lines between generators and loads ensure low losses and high efficiency and the capability for reconfiguration guarantees robustness to faults. Moreover, advanced control algorithms can improve security against malicious attacks [84].

However, despite offering many advantages, DERs can compromise grid reliability due to the intermittent operation and the creation of reverse power flows. In order to ensure a safe and resilient activity of distribution systems, and to avoid network violations, comprehensive monitoring and efficient control algorithms are necessary [85]. Nevertheless, any meaningful grid optimization and monitoring task entails the knowledge of grid topology and line parameters.

Unfortunately, different from the transmission grid, such information is not often available for distribution networks. Due to economic reasons, line parameters are seldom recorded and maps of the connections are often outdated if ever existing.

In the last decade, inexpensive and high-fidelity phasor measurement units (termed micro-PMUs or μPMUs) were invented and installed in distribution grids across America, Asia, and Europe [86]. In the coming years, as the shift towards distributed generation and smart grid progresses, the penetration of such metering devices in the distribution grid is expected to steadily increase [87]. Methods capable of recovering the topology and the line parameters of a time-varying electrical network from μPMU data are thus critical in order to enable an economically viable application of theoretical frameworks for microgrid control.

3.3 State of the Art

Until the 2010s, most works on the identification of electric networks focused on topology verification, assuming a known initial structure and aiming at detecting sparse changes, such as a line trip or switch activation [88]. More recently, attention shifted to the estimation of network topology and line parameters without any apriori information. Research tackling the grid identification problem can broadly be classified into two main branches.

3.3.1 Statistical Models

On the one hand, some works proposed learning algorithms which draw on the statistical properties of nodal measurements to determine the operational structure and line impedances [89, 90, 91]. This approach has the major advantage of accounting for buses with no available measurements (hidden nodes) [90]; although restrictive assumptions are required, e.g. hidden nodes must not be adjacent to each other. Moreover, methods based on second-order statistics either make assumptions on the covariance of nodal injections [89] or assume its foreknowledge [90, 91], and apply only to radial feeders. The latter restriction was dropped after recent developments, but only for the purpose of topology estimation [91]. In a realistic setting, these assumptions might not be satisfied; more so due to the rise of distributed generation and smart grids leading to meshed network structures.

3.3.2 Regression Models

On the other hand, network identification was often cast into the problem of learning the admittance matrix, where the position of non-zero elements provides topological information while the values of these are related to the electrical parameters of the lines [92, 93, 94, 95]. This approach requires voltage, current or power measurements at each bus but it can be applied to both radial and meshed structures. In particular, Lasso and its variants were widely adopted as they enforce sparsity of the admittance matrix.

For instance, compressive sensing theory was used to justify a Lasso formulation to recover the connections of each bus [93]. Similarly, a probabilistic graphical model motivated the adoption of Lasso to identify the non-zero elements of the admittance matrix [94]. Due to the symmetric structure of the admittance matrix, each edge was estimated twice, and deterministic rules were adopted to combine the estimates. While both approaches focused on topology, neither considered the estimation of the electrical parameters of the lines. Finally, topology and line parameters were obtained at once owing to learning the admittance matrix using Adaptive Lasso. In addition, a procedure to cope with collinearity in measurements was also proposed [95].

3.3.3 Active Learning

Different from previous schemes banking on passively recorded data, active data collection paradigms were also explored in recent publications [96, 97, 98, 99]. In the active learning framework, the inputs of a regression problem can be chosen in order to maximize some measure of the informative content of the output. Such framework well applies to power network identification, as microgrids contain generators and batteries whose output can be controlled, within appropriate constraints.

Inverter probing, for instance, was used to help the estimation of grid topology and parameters [97, 98]. However such studies, besides assuming a resistive radial network and employing approximate linearized power-flow equations, lack a comprehensive framework for the optimal design of probing injections.

To the best of our knowledge, Du et al. proposed the only study featuring a systematic procedure for maximizing the information content of data samples [99]. The authors obtained active power setpoints for generator nodes with an online design-of-experiment procedure. Nonetheless, the proposed identification algorithm assumes the availability of line power flows, and neglects the structural constraints of the admittance matrix.

3.3.4 Open Points

Our review of the literature shows the need for an online procedure, entirely based on nodal measurements, for estimating the admittance matrix. Indeed, most of the proposed methods rely on a batch of samples: such approach, however, may be inadequate for networks with temporally varying structures, like microgrids. Conversely, an online method enables on-the-fly update of topology and fault detection in modern distribution networks.

Moreover, a gap is spotted to improve existing active learning procedures and introduce a comprehensive framework to design inputs which are optimal for identification purposes while complying with the safe operation of the network.

3.4 Problem Statement

In order to state the problem of the identification of distribution grids, we first introduce a standard way of modelling the network as an undirected graph. Such model holds irrespective of the nature and the features of the network. Then, we specialize the model for distribution networks and introduce the required data and assumptions for our method.

3.4.1 Distribution Network Modelling

The electric distribution network is modeled as a connected undirected and weighted graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{W})$, where nodes represent buses, that is either generators or loads, and edges represent power lines, each connecting two distinct buses and modeled after the standard lumped $\pi-$model [100]. In such model, the resistance and the inductance of the line are concentrated in an impedance made by the series of a resistor and an inductor, placed at the middle of the cable, while the capacitance is equally split among two capacitor, connecting each end of the impedance with the ground. To each edge $(i, j) \in \mathcal{E}$ we associate a complex weight $w_{ij} \in \mathcal{W}$ equal to the line admittance $y_{ij} = g_{ij} + \mathbf{j}b_{ij}$, where $g_{ij} > 0$ are the line conductances and $b_{ij} \in \mathbb{R}$ the line susceptances. The network is then represented by the admittance matrix $Y \in \mathbb{C}^{|\mathcal{V}|\times|\mathcal{V}|}$, with elements $Y_{ij} = -y_{ij}$ and $Y_{ii} = \sum_{i=1,i\neq j}^{|\mathcal{V}|} y_{ij} + y_{s,i}$, where $y_{s,i}$ is the shunt element at the i^{th} bus.

We consider a phase-balanced power network operating in sinusoidal regime. To each bus $h \in \mathcal{V}$, we associate a phasor voltage $v_h e^{\mathbf{j}\theta_h} \in \mathbb{C}$, where $v_h > 0$ is the voltage magnitude and $\theta_h \in \mathbb{R}$ the voltage angle, a phasor current $i_h e^{\mathbf{j}\phi_h} \in \mathbb{C}$, and a complex apparent power $s_h = p_h + \mathbf{j}q_h$ with $p_h, q_h \in \mathbb{R}$. As standard in distribution networks, we assume the point of common coupling (PCC) with the main grid to be the slack bus with fixed v_0 and $\theta_0 = 0$. The remaining buses are classified as generators $\mathcal{S}$ and loads $\mathcal{L}$, such that $\mathcal{V} = \mathcal{S} \cup \mathcal{L} \cup \{0\}$. For notational simplicity we set $|\mathcal{V}| = n$, $|\mathcal{S}| = g$, and $|\mathcal{L}| = l$, where $g, l \geq 1$. In active distribution networks, generators are DERs generally interfaced with inverters equipped with voltage and/or power control [101]. The current-voltage relation descending directly from Kirchhoff's and Ohm's laws is given by

$$\boldsymbol{i} = Y\boldsymbol{v}, \tag{3.1}$$

where $\boldsymbol{i} \in \mathbb{C}^n$ is the vector of nodal currents, and $\boldsymbol{v} \in \mathbb{C}^n$ the vector of nodal voltages [102]. Similarly, one can deduce the relation between the vectors of nodal complex power injections $\boldsymbol{s}$ and nodal voltages $\boldsymbol{v}$ as

$$\boldsymbol{s} = [\boldsymbol{v}](\overline{Y\boldsymbol{v}}). \tag{3.2}$$

Equation (3.2) represents the basic form of the AC power flow, a system of non-linear equation which represents the balance of power in an electrical network. A thorough introduction to the topic can be found in the survey by Frank and Rebennak [103].

3.4.2 Identification of AC distribution networks

The identification problem for AC distribution networks aims at reconstructing the admittance matrix from a sequence of voltage and current phasor measurements corresponding to different steady states of the system [92, 95]. In line with similar studies in the literature, we assume that the network is fully observable, i.e., phasor voltage and current measurements are available at each node. Such assumption is quite demanding at the present time, as most of the distribution grid is not monitored: yet, as explained in Section 3.2, the diffusion of micro-PMU and the shift towards smart grids will boost the penetration of metering devices [86, 87] in the near future.

Let t be the the total number of measurements collected up to a certain time instant. We denote by $\boldsymbol{v}_k$ and $\boldsymbol{i}_k$ the n-dimensional vectors of current and voltage measurements for $k = 1 \dots t$. Using (3.1), one can obtain

$$I_t = YV_t, \tag{3.3}$$

where $V_t = [\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_t]$, and $I_t = [\boldsymbol{i}_1, \boldsymbol{i}_2, \dots, \boldsymbol{i}_t]$ are complex matrices.

The admittance matrix Y, encoding both line parameters and topological information, is typically sparse as each bus is not connected to all the remaining nodes. An accurate grid identification, despite the sparsity of Y, entails estimating n^2 parameters, the majority of which are expected to be zero. We highlight that Y is symmetric if phase-shifting transformers are absent in the network, and power lines are not compensated by series capacitors. Moreover, Y is a Laplacian matrix for networks wherein shunt elements $y_{s,i}$ are negligible [104].

Since phase-shifting transformers are usually employed in transmission networks, and shunt admittances are negligible for medium-sized networks - with line lengths less than 60 km, it is safe to assume that the admittance matrix Y corresponding to a standard distribution network is Laplacian [105]. Besides being symmetric, the Laplacian matrix has each diagonal element equal to the negative sum of the remaining elements of the corresponding row, thus implying the property $Y\mathbf{1}_n = \mathbf{0}_n$. Such assumption greatly reduces the number of parameters to be estimated.

3.5 Iterative Identification

Given that current and voltage measurements are not affected by errors, the identification of Y reduces to solving a system of linear equations (3.3) once enough samples are collected. Unfortunately, μPMUs and other metering devices introduce an error commonly modeled as white noise [95, 90]. In the following, it is assumed that the measurement error, acting on the currents only, is distributed as a Gaussian random vector $\mathcal{N}(\mathbf{0}_n, \sigma^2\mathbb{I}_n)$, thus implying that the error at each bus has the same variance. As it will be clear in the sequel, extensions to more complex structures of the covariance matrix are immediate. The effect of the measurement error on voltages is discussed empirically in Section 3.8.

Regression methods can be used to get a least squares estimate of the admittance matrix. Vectorizing either side of equation (3.3) yields

$$\operatorname{vec}(I_t) = \operatorname{vec}(YV_t) = \left(V_t^\top \otimes \mathbb{I}_n\right) \operatorname{vec}(Y). \tag{3.4}$$

Note that when Y is symmetric, the number of free parameters becomes $n(n+1)/2$, and, if Y is Laplacian, it is further reduced to $n(n-1)/2$. In order to prevent over-parametrization, it is thus critical to choose the most convenient set of parameters.

If shunt admittances are relevant, and Y is symmetric but not Laplacian, the half-vectorization $\operatorname{vech}(Y)$ can be adopted for the learning procedure, and the full vectorization can be recovered by the linear map defined by the duplication matrix [106], i.e., the unique matrix D such that

$$\operatorname{vec}(Y) = D \operatorname{vech}(Y). \tag{3.5}$$

It is worth noting that D is a deterministic function of n, thus it does not require to be estimated.

However, if Y is Laplacian, the half-vectorization is still redundant, as the diagonal elements can be derived from $Y\mathbf{1}_n = \mathbf{0}_n$. We thus introduce a novel non-redundant vectorization $\mathrm{ve}(Y) \in \mathbb{C}^{n(n-1)/2}$, obtained by removing diagonal elements from $-\operatorname{vech}(Y)$ as

$$\operatorname{vech}(Y) = T\,\mathrm{ve}(Y), \tag{3.6}$$

where T is the unique $(n(n+1)/2, n(n-1/2))$ transformation matrix. Indeed, one can recover the full vectorization of Y using

$$\operatorname{vec}(Y) = D\operatorname{vech}(Y) = DT\,\mathrm{ve}(Y). \tag{3.7}$$

A proof of the existence and uniqueness of T as well as formulae to construct T given n can be found in Appendix A. Python and MATLAB implementations of these formulae are publicly available on GitHub [107].

If the Y matrix is not Laplacian, one can still utilize our algorithm - set forth in what follows - to identify the network admittance matrix. In such a scenario, one needs to simply choose an apposite parametrization of Y, that is, $\operatorname{vech}(Y)$ for symmetric admittance matrices and $\operatorname{vec}(Y)$ for generic ones.

Hereafter, for ease of presentation, we consider only the case where Y is Laplacian and $\operatorname{vec}(Y) = DT\,\mathrm{ve}(Y)$: generalizations to the case where the admittance matrix is not Laplacian can be readily derived and are tested in Section 3.8. By combining (3.4) and (3.7) we get

$$\operatorname{vec}(I_t) = \left(V_t^\top \otimes \mathbb{I}_n\right) DT\,\mathrm{ve}(Y), \tag{3.8}$$

where $\otimes$ denotes the Kronecker product. Introducing the following matrices and vectors

$$A_t = \left(\boldsymbol{v}_t^\top \otimes \mathbb{I}_n\right) DT, \tag{3.9a}$$

$$\underline{A_t} := \left(V_t^\top \otimes \mathbb{I}_n\right) DT, \tag{3.9b}$$

$$\boldsymbol{b}_t := \operatorname{vec}(I_t), \text{ and} \tag{3.9c}$$

$$\boldsymbol{x} := \mathrm{ve}(Y), \tag{3.9d}$$

the least squares estimation problem at time t writes as

$$\hat{\boldsymbol{x}}_t = \arg\min_{\boldsymbol{x}} \left\|\boldsymbol{b}_t - \underline{A_t}\boldsymbol{x}\right\|_2^2. \tag{3.10}$$

The formulation in (3.10) equally weights samples at any time instant, which can be detrimental for time-varying distribution networks and smart grids [95]. We thus introduce a forgetting factor $\lambda \in (0, 1]$ and reformulate the estimation problem as

$$\hat{\boldsymbol{x}}_t = \arg\min_{\boldsymbol{x}} \sum_{i=1}^{t} \lambda^{t-i} \left\| \boldsymbol{i}_i - A_i \boldsymbol{x} \right\|_2^2 . \tag{3.11}$$

Given an initial guess of the parameter vector $\hat{\boldsymbol{x}}_0$ and the matrix $Z_0 := \sigma^{-2} \operatorname{Cov}[\hat{\boldsymbol{x}}_0]$, estimates of $\hat{\boldsymbol{x}}_t$ and $Z_t := \sigma^{-2} \operatorname{Cov}[\hat{\boldsymbol{x}}_t]$ can be obtained by the recursive least squares algorithm [108, p. 541]:

$$\hat{\boldsymbol{x}}_t = \hat{\boldsymbol{x}}_{t-1} + Z_t A_t^{\mathsf{H}} \left(\boldsymbol{i}_t - A_t \hat{\boldsymbol{x}}_{t-1} \right) \tag{3.12a}$$

$$Z_t = (\lambda Z_{t-1}^{-1} + A_t^{\mathsf{H}} A_t)^{-1} \tag{3.12b}$$

$$= \lambda^{-1} (Z_{t-1} - Z_{t-1} A_t^{\mathsf{H}} \left(\lambda \mathbb{I}_n + A_t Z_{t-1} A_t^{\mathsf{H}} \right)^{-1} A_t Z_{t-1}). \tag{3.12c}$$

From $\hat{\boldsymbol{x}}_t$, one can derive the estimated admittance matrix $\hat{Y}_t = DT\hat{\boldsymbol{x}}_t$. In a real scenario, existing information or batch data can be used to improve the initial guess $\boldsymbol{x}_0$ and Z_0.

The RLS algorithm with constant or bounded forgetting factor is known to have notable stability and convergence properties [109, 110]. For noisy measurements, RLS with constant forgetting factor is consistent under some excitation conditions only when the forgetting factor is 1 [109]. Otherwise, RLS has limited memory, preventing it from achieving consistency, which is generally traded off with the ability to follow changes in the parameters. In order to establish a basic degree of competency for the RLS estimator (3.12), we consider the case of a static network with noise-free measurements. In Section 3.8 we present numerical simulations to show how the identification method can tolerate noise and can adapt its estimation to changes in network topology.

Classical works establish that, when data are not affected by noise, the error on the parameters is bounded, and its projection onto the subspace for which persistent excitation holds – see [110] for a definition – converges to zero as the number of samples approaches infinity. Still, the arguments in [110] consider only real-valued, single-input-single-output setting. In Appendix A we provide convergence results pertaining to our case which involves complex inputs, outputs and parameters, and a multivariate output at each iteration.

Recursive least squares assumes that the matrix V_t is full-rank. If not, one can still apply the method to learn part of the admittance matrix [95].

Recursive least squares can also be applied to three-phase unbalanced networks. As detailed in [95], the variables to be measured are line-to-ground voltages and current injections for each phase of the nodes, while the admittance matrix to be estimated shares the properties described in Section 3.4.1.

3.6 Optimal Design of Experiment

Unlike learning problems where models can only capitalize on measured inputs and outputs, like the one discussed in Chapter 2, identification algorithms appropriately probing controllable generators can improve the estimation of the admittance matrix. In this work, each generator is assumed to be equipped with a voltage controller - necessary for networks with high photovoltaic integration [101]. Targeting these controllers, we henceforth propose a modified version of the recursive estimation algorithm where, at each iteration, generator voltages are set according to an optimal design of experiment.

However, before presenting the application of optimal design of experiment to grid identification, we briefly introduce and discuss the relevant theory. Design of experiment is a well-established field of research and a complete discussion of its many branches is outside the scope of this section. The interested reader can refer to the textbook by Atkinson at al [111].

An experiment aimed at estimating the parameters of a system involves the application of known inputs and the recording of the corresponding outputs. In a realistic context, the time and resources available to the experimenter are finite and thus a limited number of tests can be performed. Design of experiment concerns the choice of the inputs that can produce the maximum amount of information about the parameters to estimate. The proposed approaches can be broadly divided into two main categories: combinatorial design and optimal design.

Combinatorial design applies theory from geometry and combinatorial mathematics to choose inputs so that certain symmetry and balance principles are respected in the feature space. The field was pioneered by Ronald Fisher and was the only viable option before the advent of computer-aided numerical optimization.

The application of combinatorial design to regression problems led to the creation of factorial experiments. In full factorial designs, continuous variables are discretized into levels and all the possible combination of levels are tested. Such approach was proven more efficient, in terms of information gained per number of experiments, than varying variables one at a time.

Unfortunately, as the dimensionality of the input increases, the number of experiments required for a full factorial design explodes rapidly. Fractional factorial experiments may then be applied, where a share of the possible combinations are omitted.

Combinatorial design enjoyed considerable success in the XX century. However, the advent of cheap and fast computers and progresses in numerical optimization allowed for the adoption of optimal designs. Originally theorized by Kirstine Smith in 1918, optimal design of experiments aims at solving a formal optimization problem to maximize some measure of the information content of the samples generated by the experiments, using the inputs as control variables.

Let us consider, for instance, a linear model

$$\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{\epsilon}, \tag{3.13}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the vector of the outputs, $A \in \mathbb{R}^{n \times p}$ is the matrix of the inputs, $\boldsymbol{x} \in \mathbb{R}^p$ is the vector of parameters, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbb{I}_n)$ is a Gaussian noise corrupting output measurements. A measure of the information provided by the data $\{\boldsymbol{y}, A\}$ on the parameters $\boldsymbol{x}$ is the Fisher information matrix, defined by

$$F_{ij} = \mathrm{E}\left[\left(\frac{\partial}{\partial x_i} \log f(A, \boldsymbol{x})\right)\left(\frac{\partial}{\partial x_j} \log f(A, \boldsymbol{x})\right)\right], \tag{3.14}$$

where the likelihood $f(X, \boldsymbol{x})$ represents the probability $p(\{\boldsymbol{y}, A\}|\boldsymbol{x})$. With reference to the model (3.13), it can be shown that

$$F = (\mathrm{Cov}(\hat{\boldsymbol{x}}))^{-1} = \frac{1}{\sigma^2} A^\top A, \tag{3.15}$$

where

$$\hat{\boldsymbol{x}} = (A^\top A)^{-1} A^\top \boldsymbol{y} \tag{3.16}$$

is the least square estimator of $\boldsymbol{x}$. Notably, F depends only on the inputs A and has a geometrical interpretation based on confidence ellipsoids. Indeed, the confidence ellipsoid of level α for $\hat{\boldsymbol{x}}$ is

$$\mathcal{I} = \{\boldsymbol{z} \in \mathbb{R}^p \mid (\boldsymbol{z} - \hat{\boldsymbol{x}}) F (\boldsymbol{z} - \hat{\boldsymbol{x}}) < k_\alpha\} \tag{3.17}$$

where k_α is a constant depending on α.

In order to maximize the informative power of the data, an experimental design should choose A so that F is maximized. Several summary statistics $l : \mathcal{S}_p \to \mathbb{R}$, $\mathcal{S}_p$ begin the space of positive semidefinite symmetric matrices, were proposed to rank information matrices, each with its own motivation and interpretation. The general optimal design problem reads

$$A^* = \arg\min_A l(F), \tag{3.18}$$

and constraints can be included in order to account for the budget, the maximum allowed number of tests, or other limitations.

Different choices of l lead to different formulation of the problem (3.18): the following ones are of particular interest.

A-optimal design $l(F) = \mathrm{tr}(F^{-1})$. It minimizes the sum of the variances of the elements of $\hat{\boldsymbol{x}}$, disregarding the covariance. This solution is appealing because it does not require the computation of the entire matrix F but only of its diagonal. Geometrically, it minimizes the total length of the axes of $\mathcal{I}$.

D-optimal design $l(F) = -\log\det(F)$. It minimizes the determinant of the covariance matrix F^{-1}, or, equivalently, it maximizes the determinant of the information matrix F. A geometrical interpretation builds on the equivalence between the determinant of F and the volume of $\mathcal{I}$.

T-optimal design $l(F) = \frac{1}{\mathrm{tr}(F)}$. It is similar to the A-optimal design, but it works directly on the information matrix. Studies showed that it performs poorly if certain condition on the data matrix A are met [112].

E-optimal design $l(F) = \|F^{-1}\|_1$. It minimized the maximum eigenvalue of the covariance matrix F^{-1}. The geometrical interpretation is that it minimized the diameter of $\mathcal{I}$, defined as the longest axis of the confidence ellipsoid.

3.7 Application of Design of Experiment

We apply a D-optimal design to the grid identification problem. With reference to the least squares problem (3.11), the Fisher information matrix [111] at time t is

$$F_t = (\mathrm{Cov}(\boldsymbol{x}_t))^{-1}. \tag{3.19}$$

As the measurement noise is assumed to be distributed as a Gaussian vector $\mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbb{I}_n)$, we have

$$F_t = \sigma^{-2} Z_t^{-1} = \sigma^{-2}(\lambda Z_{t-1}^{-1} + A_t^{\mathsf{H}} A_t). \tag{3.20}$$

We note that A_t depends on the nodal voltages $\boldsymbol{v}_t$; see (3.9a). The D-optimal design is the result of the optimization problem

$$\boldsymbol{v}_t^* = \arg\max_{\boldsymbol{v}_t} \det(F_t). \tag{3.21}$$

We observe that σ does not influence the optimum and can thus be neglected. Moreover, upon applying the logarithm to the target function - a common practice for improving numerical properties [111, Chap. 10], we get

$$\boldsymbol{v}_t^* = \arg\min_{\boldsymbol{v}_t} - \log\det(\lambda Z_{t-1}^{-1} + A_t^{\mathsf{H}} A_t). \tag{3.22}$$

While formulating the design-of-experiment problem, we need to take into account voltage limits for all nodes, as well as the active and reactive power dispatched by generators. Furthermore, the power requirements of loads, expressed by the power flow equations (3.2), must be satisfied.

By adding these constraints, we get the optimization problem

$$
\begin{aligned}
(\boldsymbol{v}_t^*, \boldsymbol{p}_t^*) = \arg\min_{\boldsymbol{v}_t, \boldsymbol{p}_t} - \log\det(\lambda Z_{t-1}^{-1} + A_t^{\mathsf{H}} A_t) && && \text{(3.23a)}\\
\text{subject to:} \quad \boldsymbol{s}_t = [\boldsymbol{v}_t]\overline{(\hat{Y}_{t-1}\boldsymbol{v}_t)} && && \text{(3.23b)}\\
v_i^{min} \le v_{t,i} \le v_i^{max} && \forall i \in \mathcal{V} && \text{(3.23c)}\\
\theta_i^{min} \le \theta_{t,i} \le \theta_i^{max} && \forall i \in \mathcal{V} && \text{(3.23d)}\\
p_j^{min} \le p_{t,j} \le p_j^{max} && \forall j \in \mathcal{S} && \text{(3.23e)}\\
q_j^{min} \le q_{t,j} \le q_j^{max} && \forall j \in \mathcal{S}, && \text{(3.23f)}
\end{aligned}
$$

where A_t depends on $\boldsymbol{v}_t$ as in (3.9a).

It is worth noting that the computation of Z_{t-1}^{-1} in (3.23a) does not require the inversion of Z_{t-1}: from (3.12b), one has $Z_t^{-1} = \lambda Z_{t-1}^{-1} + A_t^{\mathsf{H}} A_t$, which allows for a recursive update of Z_t^{-1}.

Constraint (3.23b) depends on the estimated admittance matrix $\hat{Y}_{t-1}$, and thus produces suboptimal results with respect to a procedure based on the real Y. However, a sequential design of experiment, resulting from an online estimation, allows for the best available estimate of Y to be adopted at each iteration. The experiments described in Section 3.8 show that the impact on the estimation is small.

The results on convergence mentioned in Section 3.5 also apply when inputs are chosen by the design of experiment. We note that the proposed design-of-experiment procedure helps achieve the persistent excitation, which can intuitively be equated to the information matrix of the parameters being full rank at each iteration. Since the design of experiment aims at maximizing the determinant of the information matrix, its objective is in contrast with a loss of rank.

The design of experiment formulation (3.23) is flexible: one can append more constraints to the optimization problem to cope with technical limitations. For example, the voltage of some generators may be fixed, or power limitations for certain lines can be introduced. The solution of problem (3.23) is the vector of all nodal voltages; however, voltage references are provided only to distributed generators as loads cannot generally be controlled.

The design of experiment program (3.23) outputs both voltage magnitude and active power for each generating unit. In this work, we assume that the former is directly used as a control reference, however, the latter can be equivalently adopted in case of power-controlled generators. When excited with the power reference signal, the generating units cause voltage variations in the network [97, 98]: the resulting current-voltage data can then be utilized in (3.12) for the admittance matrix estimation.

To summarize, given an initial guess of $\hat{\boldsymbol{x}}_0$ and Z_0, a value of λ, and active and reactive power demands for loads, the recursive estimation enhanced with design of experiment can be described by the following steps repeated at each time t.

1. Solve the design-of-experiment problem (3.23) for the nodal voltages $\boldsymbol{v}_t^*$, using the current estimation $\hat{\boldsymbol{x}}_{t-1}$ and Z_{t-1}.
2. Provide the voltage magnitude set-point $v_{j,t}^*$ to the distributed generators $j \in \mathcal{S}$.
3. Collect measurements of current and voltage phasors from each bus $i \in \mathcal{V}$.
4. Update the estimates of $\hat{\boldsymbol{x}}_t$ and Z_t using the recursive least squares algorithm (3.12).

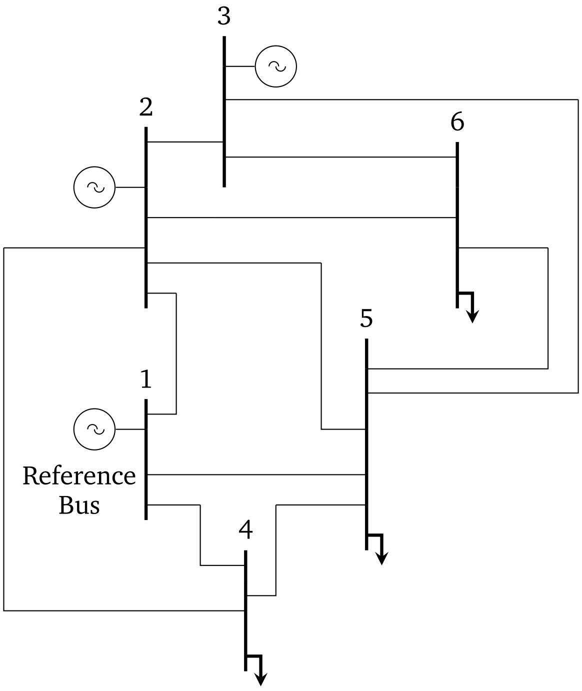


Figure 3.2. A representative diagram of grid T, the 6-bus transmission network [100]. Buses 1, 2, and 3 are generators, while 4, 5, and 6 are loads.

3.8 Experiments

In order to validate our algorithms, we set up simulations with standard testbeds. As discussed in Section 3.1, identification is usually an issue only for distribution networks, while transmission networks are known and constantly monitored. However, to prove the generality of our method, we adopted an example of both a transmission and a distribution network.

3.8.1 Experimental Setup

We considered two grids: the 6-bus transmission network by Wood and Wollenberg (grid T) [100, p. 104] and a modified version of the IEEE 13-bus radial feeder (grid D) [113]. While the method could scale to much larger networks in theory, in practice collinearity, although mitigated by the design of experiment, still leads to numerical instability for large networks.

In order to test the proposed method on a meshed network, we added two lines to grid D, one connecting bus 1 with 6, and the other bus 7 with 10. As all the lines have negligible capacitance, the admittance matrix of grid D is Laplacian.

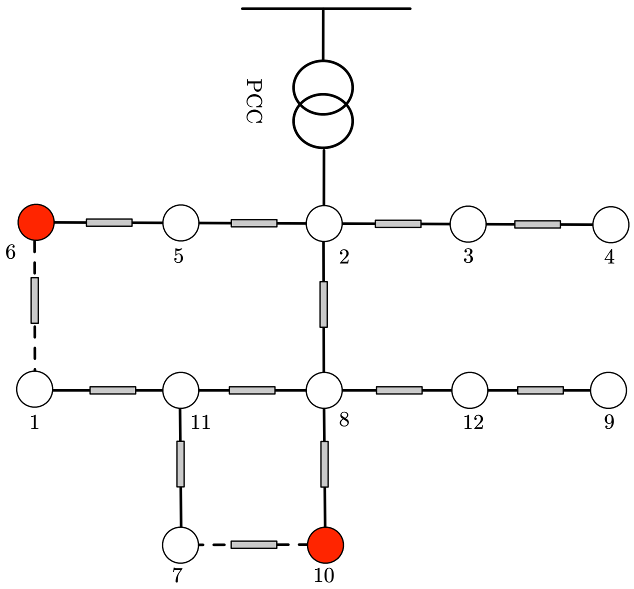


Figure 3.3. A representative diagram of grid D, the modified IEEE 13-bus feeder [113]. Buses ● and ○ represent generators and loads, respectively.

Conversely, in grid T shunt capacitances are not negligible, resulting in a symmetric, yet non-Laplacian, admittance matrix.

The presence of controllable generators is a requirement for the application of design of experiment. While grid T features 3 power sources, in grid D distributed generation is introduced through the addition of controllable power sources to buses 6 and 10. Grid D is represented in Fig. 3.3, while grid T is displayed in Fig. 3.2.

In grid T, load profiles were generated with incorrelated Gaussian active and reactive load fluctuations, centered on the nominal values. This procedure is justified by the observation that, over short periods of time, active and reactive power demands of loads can be modeled as Gaussian random variables [114, 93]. In grid D, a more realistic setup was adopted: load profiles with one-minute granularity were extracted from the public Pecan Street dataset [115]. Since this dataset did not include reactive power, a random lagging power factor between 0.85 and 0.95 was considered. Following the procedure adopted in [95], we connected a random number of customers between 5 and 15 to each node. For both grid T and D, we used the AC power flow solver MATPOWER to derive nodal current and voltage phasors [116].

For each grid, we considered two scenarios to asses the performance of our method in providing an accurate estimate of the admittance matrix. Scenario 1 looks at a network whose topology does not change over time and allows for a comparison between our online algorithm and batch methods like ordinary least squares (OLS) and adaptive lasso [95], whereas scenario 2 considers a time-varying configuration. More specifically, scenario 2 simulates a fault leading to tripping of a line. In grid T, the fault happens on the line connecting bus 2 with 6, while in grid D it impacts the line between bus 7 and 10. Batch algorithms cannot be applied to temporally varying networks and are thus excluded by the tests on scenario 2. In this respect, scenario 2 illustrates the main value of online methods over offline approaches.

We considered three different online estimation methods:

- RLS1, solely imposing the symmetric structure of Y by adopting the parametrization $\boldsymbol{x} = \mathrm{vech}(Y)$;
- RLS2, forcing a Laplacian structure of Y by adopting the parametrization $\boldsymbol{x} = \mathrm{ve}(Y)$, as in (3.9) and (3.12);
- DoE, where the generator voltages, excluding the slack bus, are set according to the design-of-experiment procedure presented in Section 3.7. The generated inputs and the corresponding outputs are fed to RLS1 if the admittance matrix of the network under consideration if symmetric, and to RLS2 if it is Laplacian.

RLS2 was not tested on grid T, as not suitable to the non-Laplacian structure of the admittance matrix of that network. The solution of the design-of-experiment problem (3.23) was computed using an interior-point non-convex solver.

In order to assess the identification performance, we used the error metrics

$$m_F := \|Y - \hat{Y}\|_{\mathrm{F}}, \tag{3.24a}$$

$$m_{\max} := \|Y - \hat{Y}\|_{\max}, \tag{3.24b}$$

$$m_R := \|Y - \hat{Y}\|_{\mathrm{F}}/\|Y\|_{\mathrm{F}}, \tag{3.24c}$$

where subscripts F and $\max$ denote the Frobenius norm and the max norm, respectively. The metric m_{F} assesses the overall goodness of the estimation, $m_{\max}$ is intended to capture possible issues in the identification of single elements, while m_{R} provides a relative measure of the identification error.

In all the experiments, we introduced a Gaussian measurement error $\mathcal{N}(\mathbf{0}_n, \sigma^2\mathbb{I}_n)$ on both the real and the imaginary part of the measurements. In both grid T and D, we chose σ so that the accuracy 3σ was 0.1% of the average magnitude of the measurement, a figure compatible with the characteristics of real metering devices [97]. This led to $\sigma = 10^{-5}$ in grid D and $\sigma = 10^{-4}$ for grid T. The recursive estimation algorithms were initialized with $\hat{\boldsymbol{x}}_0 = \delta\mathbf{1}$, $\delta = 10^{-4}$ and $Z_0 = K\mathbb{I}$, $K = 10^4$, where $\mathbf{1}$ and $\mathbb{I}$ have suitable dimensions. The forgetting factor was set to $\lambda = 0.8$.

	m_F [$\times 10^{-2}$]	$m_{\max}$ [$\times 10^{-2}$]	m_R
OLS (batch)	5.44	1.69	0.055%
Adaptive Lasso (batch)	**2.58**	**0.87**	**0.026%**
RLS1	9.55	3.84	0.095%
RLS2	7.97	3.26	0.080%
DoE	4.74	1.27	0.047%

Table 3.1. Error metrics grid D, scenario 1, after 100 samples. The best performer is in **boldface blue**.

3.8.2 Experimental Results

For sake of completeness, we present the results on both grids D and T and scenarios 1 and 2. There are little substantial differences, as the following sections show.

Grid D

For scenario 1, Table 3.1 shows the comparison with benchmarks after 100 iterations, when the estimates provided by all online algorithms no longer improve. The error metrics can be noticed to be of the same order of magnitude for all methods; although RLS1 and RLS2 achieve poorer performance than OLS and Lasso. This is expected as both OLS and Lasso are batch estimators making use of simultaneous use of all the collected data. We also note that DoE outperforms all other methods, except for Lasso.

In both scenarios 1 and 2, DoE achieves faster convergence as well as better accuracy than other iterative methods; see Fig. 3.5. The downside is the stress on generator voltages, which are subjected to frequent changes (Fig. 3.6). Nevertheless, due to constraints in the formulation of the design problem (3.23), both voltage set-points and realized voltages stay within the prescribed interval, which is $[0.95, 1.05]$ p.u. In both scenarios, $m_{\max}$ follows the same trend as m_F until convergence to a low value, thus ruling out issues about the estimation of specific elements of Y.

In the context of scenario 2, the error in the estimation of $y_{7,10}$ (See Fig. 3.7) is worth a few comments. Note that $|y_{7,10}| = 9.8$ up to $t = 100$, and subsequently drops to zero as a consequence of the simulated fault. All our recursive implementations are able to quickly adapt to a change in topology, thus proving the usefulness of online estimation. After mere two iterations ($t = 102$), the absolute value of the estimated line admittance is 2.21 for RLS1, 2.11 for RLS2, and 1.1 for DoE. Moreover, after 7 iterations, the estimation is lower than 1 for all the online algorithms.

	m_{F} $[\times 10^{-2}]$	$m_{\max}$ $[\times 10^{-2}]$	m_{R}
OLS (batch)	3.93	1.78	0.079%
Adaptive Lasso (batch)	3.40	1.62	0.068%
RLS1	4.84	2.41	0.097%
DoE	**1.34**	**0.55**	**0.027%**

Table 3.2. Error metrics for grid T, scenario 1 after 50 samples. The best performer is in **boldface blue**.

Grid T

Results on grid T are aligned with the ones reported for grid D.

The comparison with benchmarks (Table 3.2) shows that, after 50 iterations, RLS1 achieves poorer performance than both OLS and Lasso. However, DoE outperforms the batch methods, proving the value of optimal voltage excitations. The difference with grid D may be explained by the higher share of generator in grid T, which enables an higher effectiveness of the design of experiment. The visual comparison between the actual and the estimated admittance matrix shows that all the elements are well estimated. In particular, it is worth noting that the maximum error is 2 orders of magnitudes lower than the smallest element in the admittance matrix. Therefore, inferring the topology of the network from the estimated admittance matrix is trivial. A similar analysis yields the same conclusions on grid D.

DoE achieves faster convergence than RLS1 in both scenarios 1 and 2, as well as better accuracy after 50 iterations - see Fig. 3.8. The stress posed on generators is comparable to grid D but, coherently with the other test case, voltage set-points and realized voltages never violate the limits, set to $[0.95, 1.05]$ p.u. for bus 2 and $[0.93, 1.07]$ p.u. for bus 3 (Fig. 3.9).

In the context of scenario 2, it is worth analyzing the error on the estimation of $y_{2,6}$, whose real value becomes zero at time $t = 50$ as a consequence of the simulated fault (Fig. 3.10). After 7 iterations, at $t = 57$, the absolute value of the estimation with DoE is 0.48, while it is 2.55 with RLS1. Hence, DoE is again faster in updating the admittance matrix after localized changes.

Sensitivity to Voltage Noise

In real applications, measurement noise affects both currents and voltages. Although a systematic discussion of this scenario is outside the scope of this section, we assess the deterioration in performance experienced by the proposed algorithms when a zero-mean Gaussian noise with covariance matrix $\sigma_v^2\mathbb{I}$ is applied to both the real and the imaginary part of voltage measurements. As displayed in Fig. 3.11, all methods suffer from input noise; however, DoE is less affected than other methods, and achieves an acceptable performance even when the noise on voltages is of the same order of magnitude as that on currents.

Effect of the Design of Experiment Formulation

As noted in Section 3.7, the design-of-experiment formulation (3.23) has to rely on estimated admittance matrix $\hat{Y}_{t-1}$, instead of the unknown real admittance matrix Y. In order to show the effect of such an approximation on the identification algorithm, we run DoE on scenario 1 by setting $\hat{Y}_{t-1} = Y$ in (3.23b). The results in Fig. 3.12, produced for grid D, show that the procedure based on the real model of the network performs better; but the difference is marginal. The analysis for grid T yields the same conclusions and it is therefore not reported.

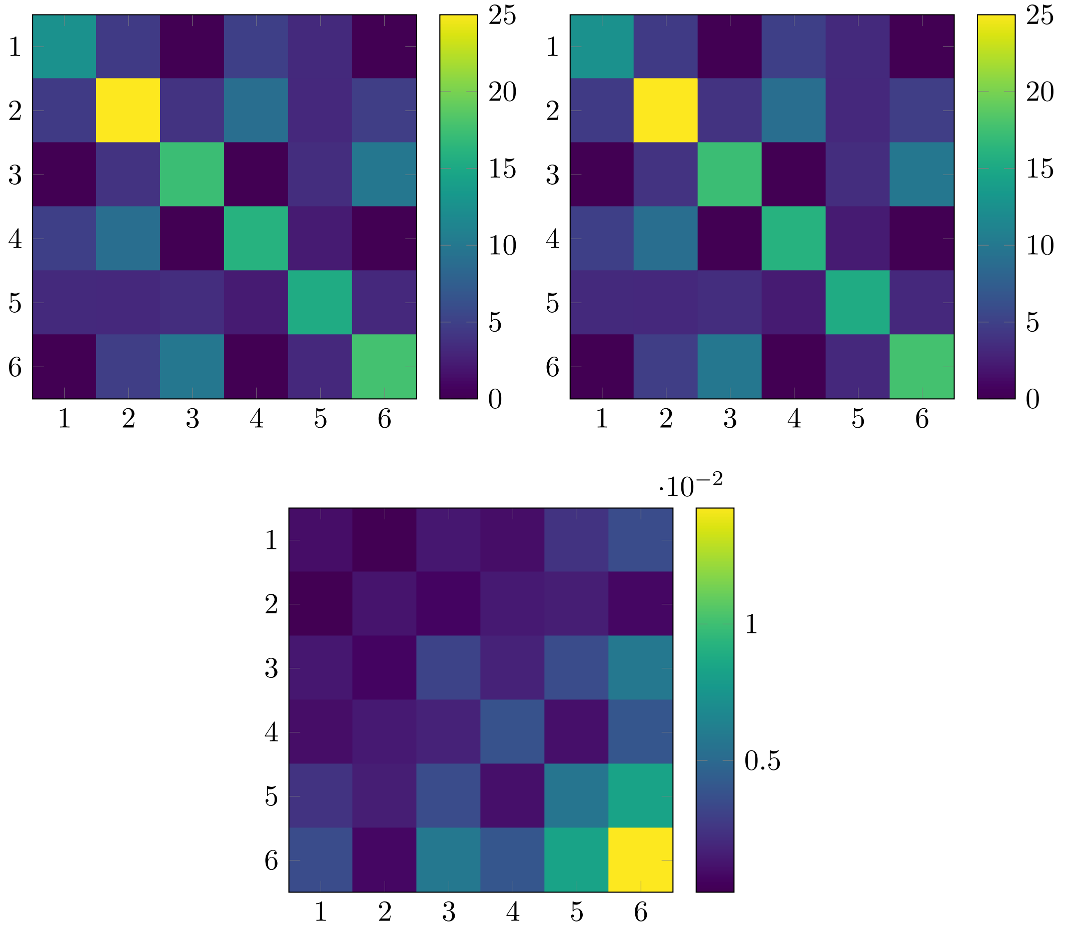
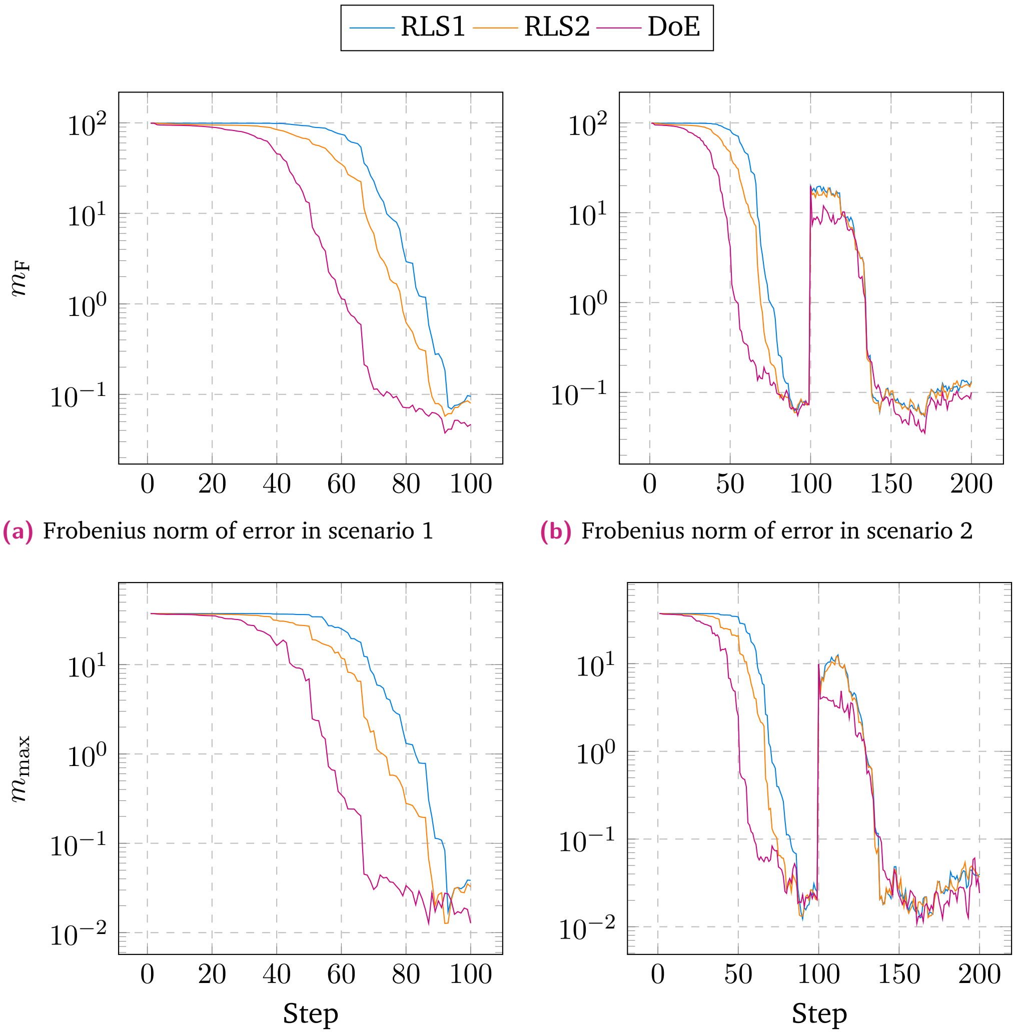


Figure 3.4. Absolute value of the actual and estimated admittance matrix, and of the estimation error after 50 time steps for grid T, scenario 1, using the DoE method. Top left panel: actual admittance matrix $|Y|$, top right panel: estimated admittance matrix $|\hat{Y}_{50}|$, bottom panel: estimation error $|Y - \hat{Y}_{50}|$ – note the difference in scale with respect to the top panels.



(a) Frobenius norm of error in scenario 1

(b) Frobenius norm of error in scenario 2

(c) Max norm of error in scenario 1

(d) Max norm of error in scenario 2

Figure 3.5. Error metrics in grid D, scenarios 1 and 2

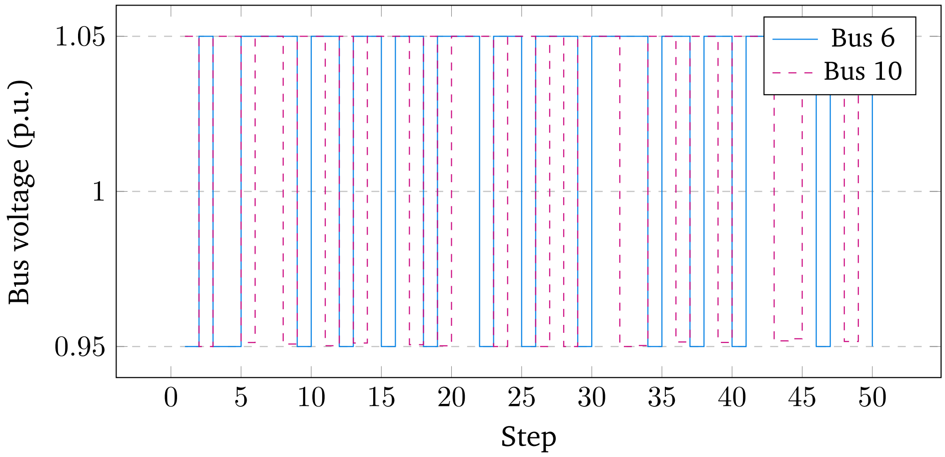


Figure 3.6. Generator voltages produced by DoE for the first 50 iterations in grid D, scenario 1.

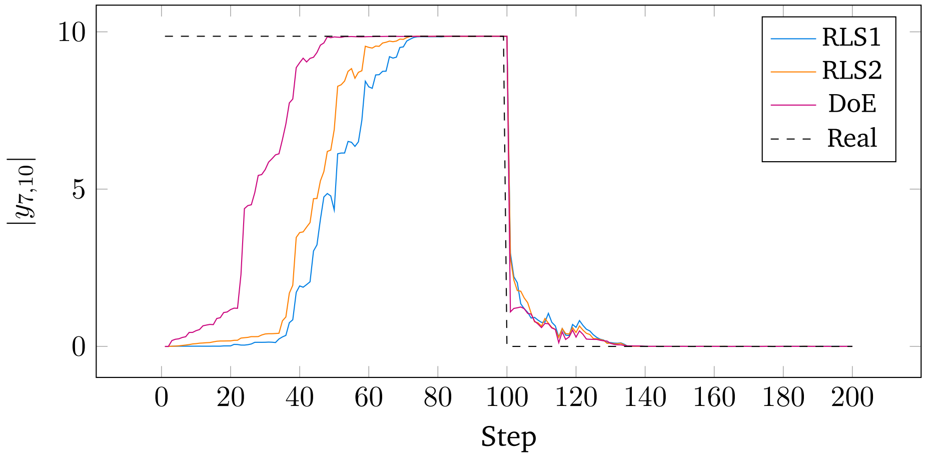
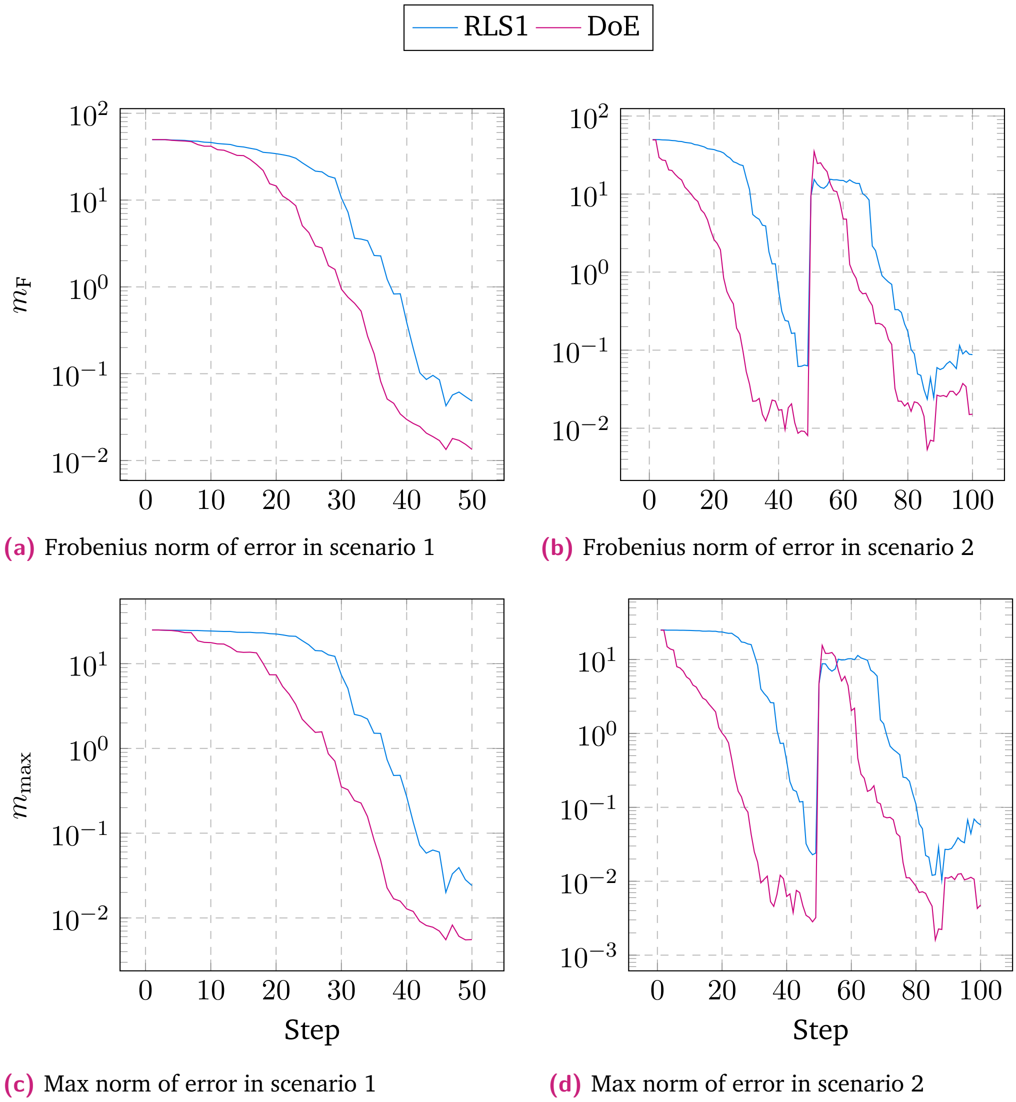


Figure 3.7. Estimation of element $y_{7,10}$ in grid D, scenario 2.



(a) Frobenius norm of error in scenario 1

(b) Frobenius norm of error in scenario 2

(c) Max norm of error in scenario 1

(d) Max norm of error in scenario 2

Figure 3.8. Error metrics in grid T, scenarios 1 and 2

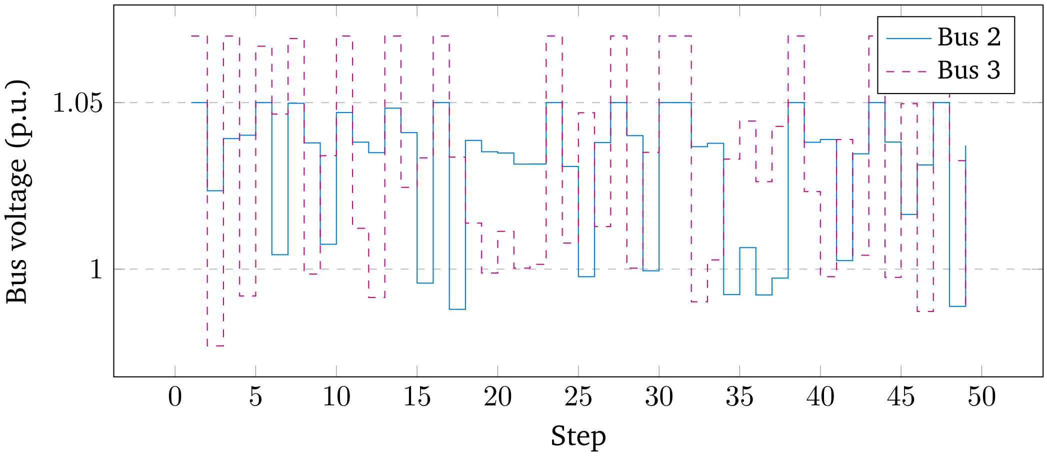


Figure 3.9. Generator voltages produced by DoE in grid T, scenario 1.

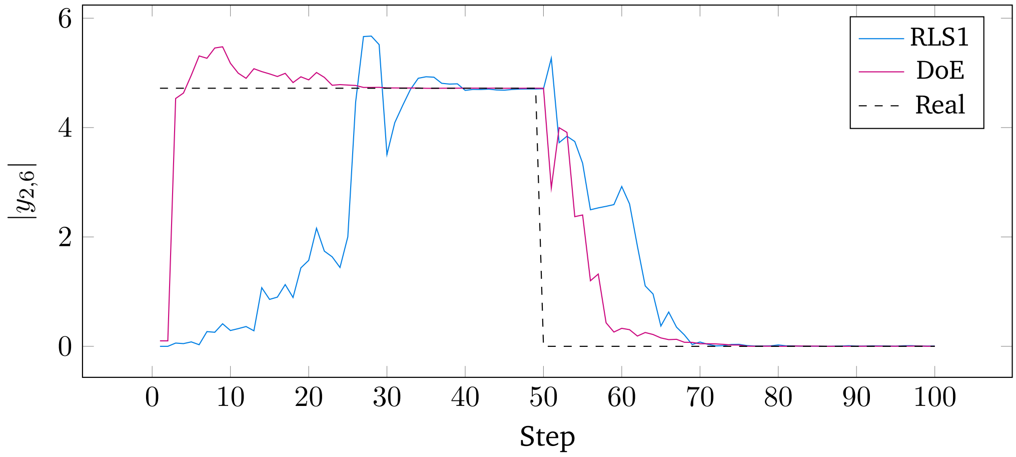


Figure 3.10. Estimation of element $y_{2,6}$ in grid T, scenario 2.

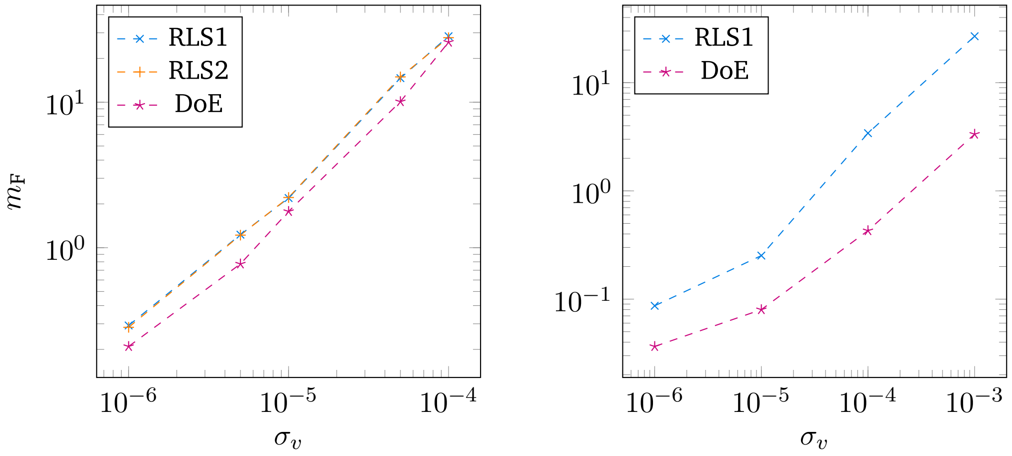


Figure 3.11. Frobenius norm of estimation error for different levels of noise on voltage measurements in scenario 1. Left panel: grid D, right panel: grid T.

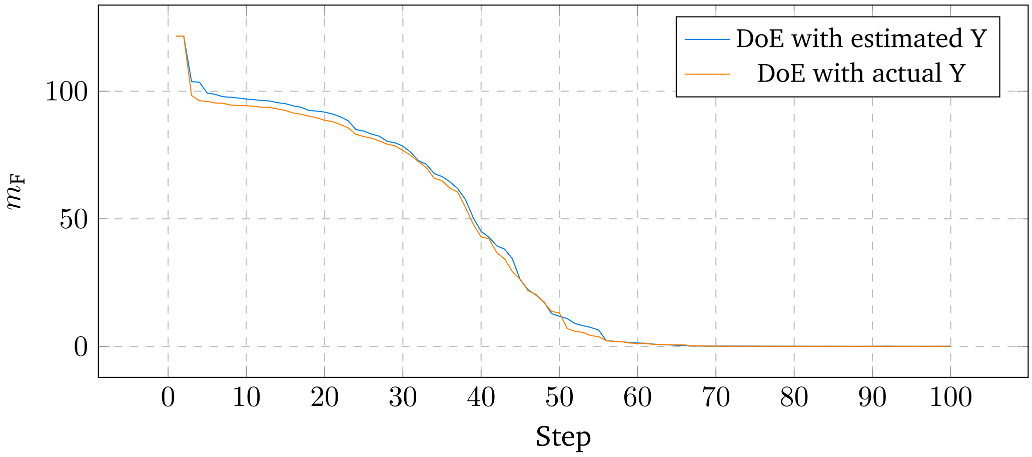


Figure 3.12. Comparison between DoE with estimated and actual admittance matrix Y.

Conclusion 4

4.1 Final Remarks

This work is meant to investigate applications of machine learning to energy, with particular attention to the challenges posed by the ongoing transition towards sustainable generation, delivery, and management.

The study is split in two main components.

The first part deals with the forecasting of the Italian gas consumption. In collaboration with A2A, an Italian utility, we carried out a detailed analysis of the demand, investigating the effects of seasonality and temperature. The insight we got may help in better plan the construction and management of storage facilities and may provide guidance for future energy policies.

After, we developed forecasting models, outperforming the only available benchmark. In particular, we proposed ensembling methods to boost the performance of base models. While allowing for a more efficient usage of natural gas, and a more effective exploitation of the existent infrastructure, such models have an added benefit for utilities: the improved predictions reduce the imbalance penalties, thus unlocking resources which can be invested in new facilities or alternative energy vectors.

The research presented in this work resulted not just in publications, but was also the basis for the models currently used in A2A. A similar argument applies to the novel probabilistic model we developed to link inaccuracies in weather forecast with errors in the prediction of natural gas. In this respect, an increased awareness about the differences in the requirements and the methods of the academia and the industry, shared in a talk delivered in an international conference, was an unusual side effect.

The second part of the study is devoted to the development of online methods for the identification of distribution grids. In collaboration with the DECODE laboratory of EPFL Lausanne, we proposed a new procedure which combines recursive identification and optimal design of experiments.

While not yet tested on the filed, but only by means of numerical simulations, our method proved successful in mitigating – yet not in solving – the issue of collinearity, which is known to affect voltage measurements in distribution grids. A separate research group has already built on a procedure similar to ours in order to achieve a balance between optimal economic operation and quick identification of the grid.

None of the two topics addressed in this work can be considered exhausted. The discussion in the next section is but a small fraction of the possible future developments. Similarly, the two subjects chosen for this thesis are just a tiny portion of all the fields where the application of machine learning to energy can deliver real value and lead to concrete improvements.

4.2 Future Work

The discussion on the forecasting of natural gas demand presented in this work is complemented by the thesis by Andrea Marziali, head of modelling and pricing in A2A and former executive PhD at the University of Pavia [75].

Considering both works, as well as the state of the art in the literature, we can suggest two main directions for future developments. On the one hand, hybrid models, which have recently gained traction, were not considered in our studies. While more complex to design, implement, and maintain, hybrid models may be capable of outperforming both base and ensemble models in sheer accuracy.

On the other hand, it may be of interest to test our approach on both different geographic areas and time scales. Our models take advantage of the smoothing effect that the aggregation of a large number of independent users has on the demand at country level. Disaggregated data, collected on individual residential or industrial units, may lack the patterns we detected and discussed for the national demand. For instance, relevant differences may arise between older and newer residential buildings, due to the improvements in thermal efficiency and insulation. Yet, accurate short-term forecasting on disaggregated data may enable effective anomaly detection, thus preventing dangerous accidents due to leaks or other failures.

If the chapter about gas demand forecasting is not closed, the one about the identification of power grids has just been opened. Microgrids and smart grids are not pervasive technologies yet, but they are projected to rise quickly in diffusion, complexity, and impact on the power infrastructure. While low-level controller can be plug-and-play, state-of-the-art designs for high-level supervisory layers require perfect knowledge of the network. Therefore, it is easy to predict that more and more advanced methods will be proposed to tackle identification.

Our ongoing research focuses on creating a more realistic and comprehensive model for the measurement noise which affects micro-PMUs. Actual metering device measure current and voltage on polar coordinates and introduce measurement errors on both the electrical variables. Therefore, the common assumption of a Gaussian noise in Cartesian coordinates on currents only is but a useful simplification. Building on the novel noise model, both online and batch algorithms can be developed. Moreover, the method proposed in this work does not take into account any prior knowledge about the network. In a real context, it is likely that some line is known to exist and some other is known not to be present. Bayesian approaches may result valuable in taking advantage of such information to reduce the amount of samples required to achieve a satisfactory estimation.

Bibliography

[1] Emanuele Fabbiani, Andrea Marziali, and Giuseppe De Nicolao. Fast calibration of two-factor models for energy option pricing. *Applied Stochastic Models in Business and Industry*, n/a(n/a), 2021.

[2] Emanuele Fabbiani, Andrea Marziali, and Giuseppe De Nicolao. vanilla-option-pricing: Pricing and market calibration for options on energy commodities. *Software Impacts*, 6:100043, 2020.

[3] Emanuele Fabbiani, Andrea Marziali, and Giuseppe De Nicolao. Forecasting residential gas demand: machine learning approaches and seasonal role of temperature forecasts. *International Journal of Oil, Gas and Coal Technology*, 26(2):202–224, 2021.

[4] Andrea Marziali, Emanuele Fabbiani, and Giuseppe De Nicolao. Ensembling methods for countrywide short term forecasting of gas demand. *International Journal of Oil, Gas and Coal Technology*, 26(2):184–201, 2021.

[5] Maurizio Polano, Emanuele Fabbiani, Eva Adreuzzi, Federica Di Cintio, Luca Bedon, Davide Gentilini, Maurizio Mongiat, Tamara Ius, Mauro Arcicasa, Miran Skrap, et al. A new epigenetic model to stratify glioma patients according to their immunosuppressive state. *Cells*, 10(3):576, 2021.

[6] Emanuele Fabbiani, Pulkit Nahata, Giuseppe De Nicolao, and Giancarlo Ferrari-Trecate. Identification of ac networks via online learning. *arXiv preprint arXiv:2003.06210*, 2020.

[7] Emanuele Fabbiani. Fast calibration of two-factor models for energy option pricing. https://www.slideshare.net/EmanueleFabbiani/fast-calibration-of-twofactor-models-for-energy-option-pricing, February 2018. [Online; accessed 1 August 2020].

[8] Emanuele Fabbiani. Short-term forecasting of italian gas demand. https://www.slideshare.net/EmanueleFabbiani/shortterm-forecasting-of-italian-gas-demand-with-machine-learning-models-130345500, February 2019. [Online; accessed 1 August 2020].

[9] Emanuele Fabbiani. Academic and buisness research in ai for energy. https://www.slideshare.net/EmanueleFabbiani/academic-and-business-research-in-ai-for-energy, January 2020. [Online; accessed 1 August 2020].

[10] Nichola Groom. California governor commits to 100 percent clean energy. https://www.reuters.com/article/us-usa-california-cleanenergy/california-governor-commits-to-100-percent-clean-energy-idUSKCN1LQ28J, September 2018. [Online; accessed 1 August 2020].

[11] California Energy Commission. California energy commission tracking progress. https://www.energy.ca.gov/sites/default/files/2019-12/renewable_ada.pdf, February 2020. [Online; accessed 1 August 2020].

[12] California Independent System Operator. What the duck curve tells us about managing a green grid. *Calif. ISO, Shap. a Renewed Future*, pages 1–4, 2012.

[13] Paul Denholm, R Margolis, and J Milford. Production cost modeling for high levels of photovoltaics penetration. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2008.

[14] Paul Denholm, Matthew O'Connell, Gregory Brinkman, and Jennie Jorgenson. Overgeneration from solar energy in california. a field guide to the duck chart. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2015.

[15] California Independent System Operator. Managing oversupply. http://www.caiso.com/informed/Pages/ManagingOversupply.aspx, 2020. [Online; accessed 1 August 2020].

[16] Jeff St. John. Southern california edison contracts huge storage portfolio to replace gas plants. https://www.greentechmedia.com/articles/read/southern-california-edison-picks-770mw-of-energy-storage-projects-to-be-built-by-next-year, May 2020. [Online; accessed 1 August 2020].

[17] WA Reardon. Electrical load forecasting. a review. Technical report, Bettelle Pacific Northwest Labs., Richland, Washington, 1972.

[18] Amir Mosavi, Mohsen Salimi, Sina Faizollahzadeh Ardabili, Timon Rabczuk, Shahaboddin Shamshirband, and Annamaria R Varkonyi-Koczy. State of the art of machine learning models in energy systems, a systematic review. *Energies*, 12(7):1301, 2019.

[19] Hesham K Alfares and Mohammad Nazeeruddin. Electric load forecasting: literature survey and classification of methods. *International journal of systems science*, 33(1):23–34, 2002.

[20] Juan Huo, Tingting Shi, and Jing Chang. Comparison of random forest and svm for electrical short-term load forecast with different data sources. In *2016 7th IEEE International conference on software engineering and service science (ICSESS)*, pages 1077–1080. IEEE, 2016.

[21] Kasun Amarasinghe, Daniel L Marino, and Milos Manic. Deep neural networks for energy load forecasting. In *2017 IEEE 26th International Symposium on Industrial Electronics (ISIE)*, pages 1483–1488. IEEE, 2017.

[22] Jakub Nowotarski, Bidong Liu, Rafał Weron, and Tao Hong. Improving short term load forecast accuracy via combining sister forecasts. *Energy*, 98:40–49, 2016.

[23] Lulu Wen, Kaile Zhou, Shanlin Yang, and Xinhui Lu. Optimal load dispatch of community microgrid with deep learning based solar power and load forecasting. *Energy*, 171:1053–1065, 2019.

[24] Souhaib Ben Taieb, James W Taylor, and Rob J Hyndman. Hierarchical probabilistic forecasting of electricity demand with smart meter data. *Journal of the American Statistical Association*, pages 1–17, 2020.

[25] Aoife M Foley, Paul G Leahy, Antonino Marvuglia, and Eamon J McKeogh. Current methods and advances in forecasting of wind power generation. *Renewable Energy*, 37(1):1–8, 2012.

[26] Javier Antonanzas, Natalia Osorio, Rodrigo Escobar, Ruben Urraca, Francisco J Martinez-de Pison, and Fernando Antonanzas-Torres. Review of photovoltaic power forecasting. *Solar Energy*, 136:78–111, 2016.

[27] Ruiguo Yu, Jie Gao, Mei Yu, Wenhuan Lu, Tianyi Xu, Mankun Zhao, Jie Zhang, Ruixuan Zhang, and Zhuo Zhang. Lstm-efg for wind power forecasting based on sequential correlation features. *Future Generation Computer Systems*, 93:33–42, 2019.

[28] Mohamed Abdel-Nasser and Karar Mahmoud. Accurate photovoltaic power forecasting models using deep lstm-rnn. *Neural Computing and Applications*, 31(7):2727–2740, 2019.

[29] Rafał Weron. Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International journal of forecasting*, 30(4):1030–1081, 2014.

[30] Lorenzo Baldacci, Matteo Golfarelli, Davide Lombardi, and Franco Sami. Natural gas consumption forecasting for anomaly detection. *Expert systems with applications*, 62:190–201, 2016.

[31] Marco De Nadai and Maarten van Someren. Short-term anomaly detection in gas consumption through arima and artificial neural network forecast. In *2015 IEEE Workshop on Environmental, Energy, and Structural Monitoring Systems (EESMS) Proceedings*, pages 250–255. IEEE, 2015.

[32] Hermine N Akouemo and Richard J Povinelli. Probabilistic anomaly detection in natural gas time series data. *International Journal of Forecasting*, 32(3):948–956, 2016.

[33] Wenqiang Cui and Hao Wang. A new anomaly detection system for school electricity consumption data. *Information*, 8(4):151, 2017.

[34] Anthony Faustine, Nerey Henry Mvungi, Shubi Kaijage, and Kisangiri Michael. A survey on non-intrusive load monitoring methodies and techniques for energy disaggregation problem. *arXiv preprint arXiv:1703.00785*, 2017.

[35] Min Xia, Ke Wang, Xu Zhang, Yiqing Xu, et al. Non-intrusive load disaggregation based on deep dilated residual network. *Electric Power Systems Research*, 170:277–285, 2019.

[36] Andrea Pozzi, Gabriele Ciaramella, Stefan Volkwein, and Davide M Raimondo. Optimal design of experiments for a lithium-ion cell: parameters identification of an isothermal single particle model with electrolyte dynamics. *Industrial & Engineering Chemistry Research*, 58(3):1286–1299, 2018.

[37] Zidong Zhang, Dongxia Zhang, and Robert C Qiu. Deep reinforcement learning for power system applications: An overview. *CSEE Journal of Power and Energy Systems*, 6(1):213–225, 2019.

[38] Italian ministry for economic development (MISE). National plan for energy development. https://www.mise.gov.it/index.php/it/energia/energia-e-clima-2030. [Online; accessed 1 August 2020].

[39] Italian natural gas demand report. https://www.snam.it/en/Natural-gas/the_context/, 2018. [Online, Accessed: 2020-08-01].

[40] Italian gas network. https://www.snam.it/en/Natural-gas/snam-infrastructures/the-transportation-network/index.html, 2019. [Online, Accessed: 2020-08-01].

[41] Božidar Soldo. Forecasting natural gas consumption. *Applied Energy*, 92:26–37, 2012.

[42] Dario Šebalj, Josip Mesarić, and Davor Dujak. Predicting natural gas consumption–a literature review. In *28th International Conference" Central European Conference on Information and Intelligent Systems*", 2017.

[43] Lifeng Wu, Sifeng Liu, Haijun Chen, and Na Zhang. Using a novel grey system model to forecast natural gas consumption in china. *Mathematical Problems in Engineering*, 2015, 2015.

[44] Xin Ma and Zhibin Liu. Application of a novel time-delayed polynomial grey model to predict the natural gas consumption in china. *Journal of Computational and Applied Mathematics*, 324:17–24, 2017.

[45] Xin Ma, Xie Mei, Wenqing Wu, Xinxing Wu, and Bo Zeng. A novel fractional time delayed grey model with grey wolf optimizer and its applications in forecasting the natural gas and coal consumption in chongqing china. *Energy*, 178:487–507, 2019.

[46] H Sarak and A Satman. The degree-day method to estimate the residential heating natural gas consumption in turkey: a case study. *Energy*, 28(9):929–939, 2003.

[47] Salvador Gil and J Deferrari. Generalized model of prediction of natural gas consumption. *J. Energy Resour. Technol.*, 126(2):90–98, 2004.

[48] Zia Wadud, Himadri S Dey, Md Ashfanoor Kabir, and Shahidul I Khan. Modeling and forecasting natural gas demand in bangladesh. *Energy Policy*, 39(11):7372–7380, 2011.

[49] Yusuf Karadede, Gultekin Ozdemir, and Erdal Aydemir. Breeder hybrid algorithm approach for natural gas demand forecasting model. *Energy*, 141:1269–1284, 2017.

[50] Syed Aziz Ur Rehman, Yanpeng Cai, Rizwan Fazal, Gordhan Das Walasai, and Nayyar Hussain Mirjat. An integrated modeling approach for forecasting long-term energy demand in pakistan. *Energies*, 10(11):1868, 2017.

[51] A Azadeh, SM Asadzadeh, M Saberi, V Nadimi, A Tajvidi, and M Sheikalishahi. A neuro-fuzzy-stochastic frontier analysis approach for long-term natural gas consumption forecasting and behavior analysis: the cases of bahrain, saudi arabia, syria, and uae. *Applied Energy*, 88(11):3850–3859, 2011.

[52] Sasan Barak and S Saeedeh Sadegh. Forecasting energy consumption using ensemble arima–anfis hybrid algorithm. *International Journal of Electrical Power & Energy Systems*, 82:92–104, 2016.

[53] Lixing Zhu, MS Li, QH Wu, and L Jiang. Short-term natural gas demand prediction based on support vector regression with false neighbours filtered. *Energy*, 80:428–436, 2015.

[54] Nan Wei, Changjun Li, Chan Li, Hanyu Xie, Zhongwei Du, Qiushi Zhang, and Fanhua Zeng. Short-term forecasting of natural gas consumption using factor selection algorithm and optimized support vector regression. *Journal of Energy Resources Technology*, 141(3), 2019.

[55] A Azadeh, SM Asadzadeh, and A Ghanbari. An adaptive network-based fuzzy inference system for short-term natural gas demand estimation: Uncertain and complex environments. *Energy Policy*, 38(3):1529–1536, 2010.

[56] Ioannis P Panapakidis and Athanasios S Dagoumas. Day-ahead natural gas demand forecasting based on the combination of wavelet transform and anfis/genetic algorithm/neural network model. *Energy*, 118:231–245, 2017.

[57] Fatih Taşpınar, Numan Celebi, and Nedim Tutkun. Forecasting of daily natural gas consumption on regional basis in turkey using various computational methods. *Energy and Buildings*, 56:23–31, 2013.

[58] Ömer Fahrettin Demirel, Selim Zaim, Ahmet Çalişkan, and Pinar Özuyar. Forecasting natural gas consumption in istanbul using neural networks and multivariate time series methods. *Turkish Journal of Electrical Engineering & Computer Sciences*, 20(5):695–711, 2012.

[59] Jolanta Szoplik. Forecasting of natural gas consumption with artificial neural networks. *Energy*, 85:208–220, 2015.

[60] Božidar Soldo, Primož Potočnik, Goran Šimunović, Tomislav Šarić, and Edvard Govekar. Improving the residential natural gas consumption forecasting models by using solar radiation. *Energy and buildings*, 69:498–506, 2014.

[61] Nan Wei, Changjun Li, Jiehao Duan, Jinyuan Liu, and Fanhua Zeng. Daily natural gas load forecasting based on a hybrid deep learning model. *Energies*, 12(2):218, 2019.

[62] Nan Wei, Changjun Li, Xiaolong Peng, Yang Li, and Fanhua Zeng. Daily natural gas consumption forecasting via the application of a novel hybrid model. *Applied Energy*, 250:358–368, 2019.

[63] Primož Potočnik, Marko Thaler, Edvard Govekar, Igor Grabec, and Alojz Poredoš. Forecasting risks of natural gas consumption in slovenia. *Energy policy*, 35(8):4271–4282, 2007.

[64] Vincenzo Bianco, Federico Scarpa, and Luca A Tagliafico. Analysis and future outlook of natural gas consumption in the italian residential sector. *Energy Conversion and Management*, 87:754–764, 2014.

[65] Vincenzo Bianco, Federico Scarpa, and Luca A Tagliafico. Scenario analysis of nonresidential natural gas consumption in italy. *Applied Energy*, 113:392–403, 2014.

[66] Federico Scarpa and Vincenzo Bianco. Assessing the quality of natural gas consumption forecasting: An application to the italian residential sector. *Energies*, 10(11):1879, 2017.

[67] George EP Box, Gwilym M Jenkins, and Gregory C Reinsel. *Time series analysis: forecasting and control*, volume 734. John Wiley & Sons, 2011.

[68] Rafal Weron. *Modeling and forecasting electricity loads and prices: A statistical approach*, volume 403. John Wiley & Sons, 2007.

[69] Alice Guerini and Giuseppe De Nicolao. Long-term electric load forecasting: A torus-based approach. In *2015 European Control Conference (ECC)*, pages 2768–2773. IEEE, 2015.

[70] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[71] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[72] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

[73] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[74] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[75] Andrea Marziali. *Machine learning models applied to energy time-series forecasting*. PhD thesis, University of Pavia, 2020.

[76] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[77] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[78] Véronique Genre, Geoff Kenny, Aidan Meyler, and Allan Timmermann. Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, 29(1):108–121, 2013.

[79] SNAM Rete Gas. Snam daily gas consumption forecast. http://www.snam.it/it/trasporto/dati-operativi-business/8_dati_operativi_bilanciamento_sistema. [Online; accessed 1 August 2020].

[80] Eric Breck, Shanqing Cai, Eric Nielsen, Michael Salib, and D. Sculley. Whats your ml test score? a rubric for ml production systems. In *Reliable Machine Learning in the Wild - NIPS 2016 Workshop*, 2016.

[81] Vahid Madani, Ratan Das, Farrokh Aminifar, John McDonald, SS Venkata, Damir Novosel, Anjan Bose, and Mohammad Shahidehpour. Distribution automation strategies challenges and opportunities in a changing landscape. *IEEE Transactions on Smart Grid*, 6(4):2157–2165, 2015.

[82] Pulkit Nahata, Raffaele Soloperto, Michele Tucci, Andrea Martinelli, and Giancarlo Ferrari-Trecate. A passivity-based approach to voltage stabilization in dc microgrids with zip loads. *Automatica*, 113:108770, 2020.

[83] Pulkit Nahata, Alessio La Bella, Riccardo Scattolini, and Giancarlo Ferrari-Trecate. Hierarchical control in islanded dc microgrids with flexible structures. *arXiv preprint arXiv:1910.05107*, 2019.

[84] Alexander J Gallo, Mustafa S Turan, Pulkit Nahata, Francesca Boem, Thomas Parisini, and Giancarlo Ferrari-Trecate. Distributed cyber-attack detection in the secondary control of dc microgrids. In *2018 European Control Conference (ECC)*, pages 344–349. IEEE, 2018.

[85] Alessio La Bella, Pulkit Nahata, and Giancarlo Ferrari-Trecate. A supervisory control structure for voltage-controlled islanded dc microgrids. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 6566–6571. IEEE, 2019.

[86] Alexandra Von Meier, Emma Stewart, Alex McEachern, Michael Andersen, and Laura Mehrmanesh. Precision micro-synchrophasors for distribution systems: A summary of applications. *IEEE Transactions on Smart Grid*, 8(6):2926–2936, 2017.

[87] Deepa S Kumar, JS Savier, and SS Biju. Micro-synchrophasor based special protection scheme for distribution system automation in a smart city. *Protection and Control of Modern Power Systems*, 5(1):1–14, 2020.

[88] Yue Zhao, Jianshu Chen, and H Vincent Poor. Efficient neural network architecture for topology identification in smart grid. In *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 811–815. IEEE, 2016.

[89] Saverio Bolognani, Nicoletta Bof, Davide Michelotti, Riccardo Muraro, and Luca Schenato. Identification of power distribution network topology via voltage correlation analysis. In *52nd IEEE Conference on Decision and Control*, pages 1659–1664. IEEE, 2013.

[90] Deepjyoti Deka, Michael Chertkov, and Scott Backhaus. Joint estimation of topology and injection statistics in distribution grids with missing nodes. *IEEE Transactions on Control of Network Systems*, 2020.

[91] Deepjyoti Deka, Saurav Talukdar, Michael Chertkov, and Murti Salapaka. Graphical models in meshed distribution grids: Topology estimation, change detection & limitations. *IEEE Transactions on Smart Grid*, 2020.

[92] Ye Yuan, Steven Low, Omid Ardakanian, and Claire Tomlin. Inverse power flow problem. *arXiv preprint arXiv:1610.06631*, 2016.

[93] Mohammad Babakmehr, Marcelo G Simões, Michael B Wakin, and Farnaz Harirchi. Compressive sensing-based topology identification for smart grids. *IEEE Transactions on Industrial Informatics*, 12(2):532–543, 2016.

[94] Yizheng Liao, Yang Weng, Guangyi Liu, and Ram Rajagopal. Urban MV and LV distribution grid topology estimation via group lasso. *IEEE Transactions on Power Systems*, 34(1):12–27, 2018.

[95] Omid Ardakanian, Vincent WS Wong, Roel Dobbe, Steven H Low, Alexandra von Meier, Claire J Tomlin, and Ye Yuan. On identification of distribution grids. *IEEE Transactions on Control of Network Systems*, 6(3):950–960, 2019.

[96] Marko Angjelichinoski, Čedomir Stefanović, Petar Popovski, Anna Scaglione, and Frede Blaabjerg. Topology identification for multiple-bus dc microgrids via primary control perturbations. In *2017 IEEE Second International Conference on DC Microgrids (ICDCM)*, pages 202–206. IEEE, 2017.

[97] Guido Cavraro and Vassilis Kekatos. Graph algorithms for topology identification using power grid probing. *IEEE control systems letters*, 2(4):689–694, 2018.

[98] Guido Cavraro and Vassilis Kekatos. Inverter probing for power distribution network topology processing. *IEEE Transactions on Control of Network Systems*, 6(3):980–992, 2019.

[99] Xu Du, Alexander Engelmann, Yuning Jiang, Timm Faulwasser, and Boris Houska. Optimal experiment design for ac power systems admittance estimation. *arXiv preprint arXiv:1912.09017*, 2019.

[100] Allen J Wood, Bruce F Wollenberg, and Gerald B Sheblé. *Power generation, operation, and control*. John Wiley & Sons, 1996.

[101] Daniel K Molzahn, Florian Dörfler, Henrik Sandberg, Steven H Low, Sambuddha Chakrabarti, Ross Baldick, and Javad Lavaei. A survey of distributed optimization and control algorithms for electric power systems. *IEEE Transactions on Smart Grid*, 8(6):2941–2962, 2017.

[102] Florian Dörfler, John W Simpson-Porco, and Francesco Bullo. Electrical networks and algebraic graph theory: Models, properties, and applications. *Proceedings of the IEEE*, 106(5):977–1005, 2018.

[103] Stephen Frank and Steffen Rebennack. An introduction to optimal power flow: Theory, formulation, and examples. *IIE Transactions*, 48(12):1172–1197, 2016.

[104] Prabha Kundur, Neal J Balu, and Mark G Lauby. *Power system stability and control*, volume 7. McGraw-hill New York, 1994.

[105] Maamar Taleb, Mohamed Jassim Ditto, and Tahar Bouthiba. Performance of short transmission lines models. In *2006 IEEE GCC Conference (GCC)*, pages 1–7. IEEE, 2006.

[106] Jan R Magnus and H Neudecker. The elimination matrix: some lemmas and applications. *SIAM Journal on Algebraic Discrete Methods*, 1(4):422–449, 1980.

[107] E. Fabbiani and P. Nahata. Transformation matrix for non-redundant parametrization of laplacian matrices. Python: https://git.io/JfzHg, MATLAB: https://git.io/JfzH0, 2020. Accessed February 28th, 2020.

[108] Monson H Hayes. *Statistical digital signal processing and modeling*. John Wiley & Sons, 2009.

[109] Yanjun Liu and Feng Ding. Convergence properties of the least squares estimation algorithm for multivariable systems. *Applied Mathematical Modelling*, 37(1-2):476–483, 2013.

[110] Sergio Bittanti and Paolo Bolzern. Recursive least-squares identification algorithms with incomplete excitation: Convergence analysis. *IEEE Transactions on Automatic Control*, 35(12), 1990.

[111] Anthony Atkinson, Alexander Donev, Randall Tobias, et al. *Optimum experimental designs, with SAS*. Oxford University Press, 2007.

[112] Friedrich Pukelsheim. *Optimal design of experiments*. SIAM, 2006.

[113] KP Schneider, BA Mather, BC Pal, C-W Ten, GJ Shirek, Hao Zhu, JC Fuller, José Luiz Rezende Pereira, Luis Fernando Ochoa, Leandro Ramos de Araujo, et al. Analytic considerations and design basis for the ieee distribution test feeders. *IEEE Transactions on power systems*, 33(3):3181–3188, 2017.

[114] Hanie Sedghi and Edmond Jonckheere. Statistical structure learning to ensure data integrity in smart grid. *IEEE Transactions on Smart Grid*, 6(4):1924–1933, 2015.

[115] Pecan street inc. https://www.pecanstreet.org/dataport/, accessed May 2020.

[116] Ray Daniel Zimmerman, Carlos Edmundo Murillo-Sánchez, and Robert John Thomas. Matpower: Steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Transactions on power systems*, 26(1):12–19, 2010.

Appendix A

A.1 Existence and Uniqueness of the Transformation Matrix T

We show the existence and uniqueness of the transformation matrix T introduced in Section 3.5 and defined as follows.

Definition 5. *Given a Laplacian matrix $A \in \mathbb{C}^{n\times n}$, the transformation matrix T is such that*

$$\mathrm{vech}(A) = T\,\mathrm{ve}(A) \tag{A.1}$$

.

Lemma 1. *Given $n \in \mathbb{N}_+$, there exist a unique $n(n+1)/2 \times n(n-1)/2$ matrix T verifying Definition 5 for each Laplacian matrix $A \in \mathbb{C}^{n\times n}$.*

Proof. Each element of $\mathrm{vech}(A)$ is a linear combination of elements in $\mathrm{ve}(A)$ and this is sufficient to guarantee the existence of a linear map transforming $\mathrm{ve}(A)$ into $\mathrm{vech}(A)$. The uniqueness can be shown by contradiction. Assume there exists $\tilde{T} \neq T$ such that $\mathrm{vech}(A) = \tilde{T}\,\mathrm{ve}(A) = T\,\mathrm{ve}(A)$. Then, $\mathbf{0}_{n(n+1)/2} = (\tilde{T} - T)\,\mathrm{ve}(A), \forall A$. Thus, it has to be $\tilde{T} = T$. □

A.2 Derivation of the Transformation Matrix T

The construction of T is best understood starting with an example. Let $n = 4$ and $A \in \mathbb{C}^{4\times 4}$ be the following Laplacian matrix:

$$A = \begin{bmatrix} a_1 + a_2 + a_3 & -a_1 & -a_2 & -a_3 \\ -a_1 & a_1 + a_4 + a_5 & -a_4 & -a_5 \\ -a_2 & -a_4 & a_2 + a_4 + a_5 & -a_6 \\ -a_3 & -a_5 & -a_6 & a_3 + a_5 + a_6 \end{bmatrix} \tag{A.2}$$

By definition, the half-vectorization $\mathrm{vech}(A)$ and the non-redundant vectorization $\mathrm{ve}(A)$ are:

$$\mathrm{vech}(A) = \begin{bmatrix} a_1 + a_2 + a_3 \\ -a_1 \\ -a_2 \\ -a_3 \\ -a_1 + a_4 + a_5 \\ -a_4 \\ -a_5 \\ a_2 + a_4 + a_5 \\ -a_6 \\ a_3 + a_5 + a_6 \end{bmatrix}, \quad \mathrm{ve}(A) = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{bmatrix} \tag{A.3}$$

From the implicit Definition 5, it is immediate to check that the transformation matrix is:

$$T = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \tag{A.4}$$

In order to develop a construction procedure for T, it is convenient to divide it into n submatrices of different dimensions T_z, $z = 1...n$, with $T_z \in \mathbb{R}^{n+1-z \times n(n-1)/2}$ such that:

$$T = \begin{bmatrix} T_1 \\ \vdots \\ T_z \\ \vdots \\ T_n \end{bmatrix} \tag{A.5}$$

Applying the split to the T matrix in the example, we get:

$$T = \begin{bmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \end{bmatrix} = \left[\begin{array}{cccccc} 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ \hdashline 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ \hdashline 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ \hdashline 0 & 0 & 1 & 0 & 1 & 1 \end{array}\right] \tag{A.6}$$

Each T_z has a similar structure: when multiplied by $\mathrm{ve}(A)$, the first row yields a diagonal element of A, while the other $n - z$ rows adjust the signs of the off-diagonal elements.

We thus focus on a generic T_z. Its structure can be further divided into four submatrices: the first row is denoted T_{za}, while the reminder of T_i can be split into two zero matrices T_{zb} and T_{zd} and a negative identity matrix T_{zc}. The sizes of T_{zb}, T_{zc}, and T_{zd} change with the submatrix index z. We show the split with T_2 in the example:

$$T_2 = \left[\begin{array}{c:c:c} \multicolumn{3}{c}{T_{2a}} \\ \hdashline T_{2b} & T_{2c} & T_{2d} \end{array}\right] = \left[\begin{array}{ccc:cc:c} 1 & 0 & 0 & 1 & 1 & 0 \\ \hdashline 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \end{array}\right] \tag{A.7}$$

Due to the structure of $\mathrm{vech}(A)$ and $\mathrm{ve}(A)$, one has:

$$T_{ib} = \mathbb{O}_{n-z \times n(n-z)-z(z-1)/2} \tag{A.8}$$

$$T_{ic} = -\mathbb{I}_{n-z} \tag{A.9}$$

$$T_{id} = \mathbb{O}_{n-z \times (n(n-1)+z(z-1))/2-nz+z} \tag{A.10}$$

To justify the expressions, one can observe that every block T_z maps $\mathrm{ve}(A)$ into $n + 1 - z$ elements of $\mathrm{vech}(A)$, the first being a diagonal element of A. The negative identity matrix T_{zc} has size equal to the number of mapped elements of $\mathrm{vec}(A)$ which are not diagonal element of A, namely $n + 1 - z - 1 = n - z$, while the zero matrix T_{zb} has a number of columns equal to the number of off-diagonal elements of A mapped by T_w with $w < z$, namely $\sum_{k=1}^{z-1} n - k = n(n - z) - z(z - 1)/2$. The structure of T_{zd} follows from the size of T and the previous considerations.

The structure of the first row is more complex and reads:

$$T_{za} = \sum_{k=1}^{n-z} \boldsymbol{e}^{\top}_{k+n(z-1)-z(z-1)/2} + \sum_{k=1}^{z-1} \boldsymbol{e}^{\top}_{z-1+(n-1)(k-1)+k(k-1)/2} \tag{A.11}$$

where $e \in \mathbb{R}^{n(n-1)/2}$. The matrices T_{za} map the elements of $\mathrm{ve}(A)$ into the diagonal elements of A. By the properties of the Laplacian matrix A, its diagonal elements can be expressed as:

$$a_{ii} = -\sum_{j=1, j\neq i}^{n} a_{ij} = -\left(\sum_{j=1}^{i-1} a_{ij} + \sum_{j=i+1}^{n} a_{ij}\right) = -\left(\sum_{k=1}^{i-1} a_{ki} + \sum_{j=i+1}^{n} a_{ij}\right) \tag{A.12}$$

The first sum in (A.11) accounts for the terms in the second sum in (A.12) while the second sum in (A.11) identifies the terms in the first sum of (A.12).

It is worth noting that the construction method described for T is general and holds irrespective of the dimension of A. Python and MATLAB implementation of the construction formulae are publicly available on Github [107].

A.3 Convergence of Recursive Least Squares

We show here that the results about the convergence of recursive least squares claimed in the work by Bittanti and Bolzern [110] also hold in the context presented in Section 3.5, featuring complex inputs, outputs, and multivariate output at each iteration.

Lemma 2. *Consider the recursive least square algorithm* (3.12)*. Assume that Y is constant in time – therefore, $\boldsymbol{x}_t = \boldsymbol{x}$, V_t is full-rank and measurement are not affected by noise. Define the error on the parameters $\tilde{\boldsymbol{x}}_t := \hat{\boldsymbol{x}}_t - \boldsymbol{x}$. For any $\hat{\boldsymbol{x}}_0$ and $Z_0 = Z_0^{\mathsf{H}} \succ 0$, (i) the norm of the error $\|\tilde{\boldsymbol{x}}_t\|$ is bounded, and (ii) the projection of $\tilde{\boldsymbol{x}}_t$ on the excitation subspace converges to zero as t approaches infinity.*

Proof. In order to show (i), we start by substituting $\tilde{\boldsymbol{x}}_t = \hat{\boldsymbol{x}}_t - \boldsymbol{x}$ in (3.12) and considering that, in the noise-free case, $\boldsymbol{i}_t = A_t \boldsymbol{x}$. Then, we get the recursive formula

$$\tilde{\boldsymbol{x}}_t = \hat{\boldsymbol{x}}_{t-1} - Z_t A_t^{\mathsf{H}} A_t \tilde{\boldsymbol{x}}_{t-1}. \tag{A.13}$$

For convenience, we define $\boldsymbol{\epsilon}_t := A_t \tilde{\boldsymbol{x}}_{t-1} = A_t \hat{\boldsymbol{x}}_{t-1} - \boldsymbol{i}_t$. Next, we introduce the Lyapunov-like function $W_t := \tilde{x}^{\mathsf{H}} Z_t^{-1} \tilde{x}_t$. Note that W is a real-valued function, as Z_t

is Hermitian and so is Z_t^{-1}. By combining the definition of W_t with equations (A.13) and (3.12b), we derive

$$W_t = \lambda W_{t-1} - \boldsymbol{\epsilon}_t^{\mathsf{H}}(\mathbb{I}_n - A_t Z_t A_t^{\mathsf{H}})\boldsymbol{\epsilon}_t. \tag{A.14}$$

It can be shown that $\mathbb{I}_n - A_t Z_t A_t^{\mathsf{H}} \succeq 0$; see e.g. Lemma 1 in [109]. Consider now the quantity $\boldsymbol{\epsilon}_t^{\mathsf{H}}(\mathbb{I}_n - A_t Z_t A_t^{\mathsf{H}})\boldsymbol{\epsilon}_t$: it is real and non-negative because $\mathbb{I}_n - A_t Z_t A_t^{\mathsf{H}}$ is Hermitian and positive semidefinite. From (A.14), we obtain the inequality

$$W_t \leq \lambda W_{t-1}, \tag{A.15}$$

which one can recursively apply at each t to obtain

$$W_t \leq \lambda^t W_0. \tag{A.16}$$

Recalling the definition of W and (3.12b), we write:

$$\begin{aligned} \lambda^t W_0 \geq W_t &= \tilde{\boldsymbol{x}}_t^{\mathsf{H}} Z_t^{-1} \tilde{\boldsymbol{x}}_t \\ &= \tilde{\boldsymbol{x}}_t^{\mathsf{H}} \left(\lambda Z_{t-1}^{-1} + A_t^{\mathsf{H}} A_t\right) \tilde{\boldsymbol{x}}_t \\ &= \tilde{\boldsymbol{x}}_t^{\mathsf{H}} \left(\lambda^t Z_0^{-1} + \sum_{i=1}^{t} \lambda^{t-i} A_i^{\mathsf{H}} A_i\right) \tilde{\boldsymbol{x}}_t \\ &\geq \lambda^t \tilde{\boldsymbol{x}}_t^{\mathsf{H}} \left(Z_0^{-1} + \sum_{i=1}^{t} A_i^{\mathsf{H}} A_i\right) \tilde{\boldsymbol{x}}_t. \end{aligned} \tag{A.17}$$

Therefore, we conclude that

$$\tilde{\boldsymbol{x}}_t^{\mathsf{H}}(Z_0^{-1} + \sum_{i=1}^{t} A_i^{\mathsf{H}} A_i)\tilde{\boldsymbol{x}}_t \leq W_0. \tag{A.18}$$

As both $Z_0^{-1} \succ 0$ and $\sum_{i=1}^{t} A_i^{\mathsf{H}} A_i \succeq 0$, we have

$$\tilde{\boldsymbol{x}}_t^{\mathsf{H}} Z_0^{-1} \tilde{\boldsymbol{x}}_t \leq W_0 \tag{A.19}$$

and

$$\tilde{\boldsymbol{x}}_t^{\mathsf{H}} \left(\sum_{i=1}^{t} A_i^{\mathsf{H}} A_i\right) \tilde{\boldsymbol{x}}_t \leq W_0. \tag{A.20}$$

Since $\tilde{\boldsymbol{x}}_t^{\mathsf{H}} Z_0^{-1} \tilde{\boldsymbol{x}}_t \geq \operatorname{mineig}(Z_0^{-1}) \|\tilde{\boldsymbol{x}}_t\|^2$, equation (A.19) yields

$$\operatorname{mineig}(Z_0^{-1}) \|\tilde{\boldsymbol{x}}_t\|^2 \leq W_0, \tag{A.21}$$

where $\mathrm{mineig}(X)$ is the minimal (real) eigenvalue of a Hermitian matrix X. Therefore, $\|\tilde{\boldsymbol{x}}_t\|$ is bounded.

In order to show (ii), let G_t be the square root of $\sum_{i=1}^{t} A_i^{\mathsf{H}} A_i$, and $\tilde{\boldsymbol{x}}_t^{(e)}$ and $\tilde{\boldsymbol{x}}_t^{(u)}$ the projections of $\tilde{\boldsymbol{x}}_t$ onto the subspaces where persistent excitation holds and does not hold, respectively. Then, (A.20) can be written as

$$\left\| G_t \tilde{\boldsymbol{x}}_t^{(e)} + G_t \tilde{\boldsymbol{x}}_t^{(u)} \right\| \leq W_0^{1/2}. \tag{A.22}$$

By the reverse triangular inequality, we have:

$$\underbrace{\left\| \frac{G_t \tilde{\boldsymbol{x}}_t^{(e)}}{\left\| \tilde{\boldsymbol{x}}_t^{(e)} \right\|} \left\| \tilde{\boldsymbol{x}}_t^{(e)} \right\| \right\|}_{a} - \underbrace{\left\| \frac{G_t \tilde{\boldsymbol{x}}_t^{(u)}}{\left\| \tilde{\boldsymbol{x}}_t^{(u)} \right\|} \left\| \tilde{\boldsymbol{x}}_t^{(u)} \right\| \right\|}_{b} \leq W_0^{1/2}. \tag{A.23}$$

We can now apply to (A.23) the same argument proposed in [110, Proof of Theorem 1]. Due to part (i) of the proof, the norms of $\tilde{\boldsymbol{x}}_t^{(e)}$ and $\tilde{\boldsymbol{x}}_t^{(u)}$ are bounded. Moreover, as $\tilde{\boldsymbol{x}}_t^{(u)}$ is the projection of $\tilde{\boldsymbol{x}}_t$ onto the subspace where persistent excitation does not hold, the term (b) of (A.23) is bounded. For the inequality (A.23) to hold, (a) must also be bounded. Yet, the lemma in the appendix of [110] shows that $G_t \tilde{\boldsymbol{x}}_t^{(e)} / \left\| \tilde{\boldsymbol{x}}_t^{(e)} \right\|$ is unbounded. Therefore, $\tilde{\boldsymbol{x}}_t^{(e)}$ must converge to zero for (A.23) to be verified, proving part (ii). It is worth noting that the lemma in [110] involves sequences of real positive semidefinite matrices, but the proof holds without modification for complex hermitian positive semidefinite matrices. □