Programa de Doctorado en Ingeniería Matemática, Estadística e Investigación Operativa

por la

Universidad Complutense de Madrid

y la

Universidad Politécnica de Madrid

Forecasting Methods under Shifting Conditions: Application to Electricity Markets

---

Tesis Doctoral

Carlos Sebastián Martínez-Cava

Directores

Carlos Eduardo González Guillén
Jesús Juan Ruiz

Año

2025

UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingenieros Industriales

Doctorado en Ingeniería Matemática, Estadística e Investigación Operativa

Forecasting Methods under Shifting Conditions: Application to Electricity Markets

TESIS DOCTORAL

Presentada para optar al título de Doctor por:

Carlos Sebastián Martínez-Cava
Máster en Tratamiento Estadístico Computacional de la Información

Bajo la dirección de:
Dr. Carlos Eduardo González Guillén
Dr. Jesús Juan Ruiz

Madrid, 2025

Título: Forecasting Methods under Shifting Conditions: Application to Electricity Markets

Autor: Carlos Sebastián Martínez-Cava

Programa de Doctorado: Ingeniería Matemática, Estadística e Investigación Operativa

Dirección de Tesis:

Dr. Carlos Eduardo González Guillén, profesor contratado doctor, Universidad Politécnica de Madrid (Director)

Dr. Jesús Juan Ruiz, Catedrático de Universidad, Universidad Politécnica de Madrid

Revisores Externos:

Tribunal de Tesis:

Fecha de Defensa de Tesis:

Esta tesis ha sido parcialmente financiada por la beca MIG-20211033 del Centro para el Desarrollo Tecnológico Industrial, Ministerio de Universidades y la Unión Europea-NextGenerationEU.

Prediction is very difficult, especially if it's about the future.

– Niels Bohr

Acknowledgements

La realización de una tesis puede ser un proceso muy solitario, pero en mi caso, he tenido la suerte de contar con el apoyo de muchas personas que han hecho el camino mucho más llevadero.

En primer lugar, quiero agradecer a mis directores, Carlos y Jesús. No solo han sido una guía fundamental, sino que, en incontables ocasiones, su paciencia ha sido clave para que esta tesis llegara a buen puerto.

A mi familia, que, aun sin entender del todo lo que estaba haciendo, me ha apoyado sin reservas a lo largo de todo el proceso.

Pero si alguien puede comprender realmente lo que implica escribir una tesis doctoral, son quienes han estado en la misma situación. María, Bibi, Paula, Antonio y Joao, vuestras historias, apoyo y compañía han transformado este viaje en algo mucho más tolerable y, sobre todo, memorable.

A todos, gracias.

Abstract

Electricity price forecasting in the Day-Ahead Market is critical for companies in the electricity sector, as it directly affects operational and strategic decisions. This process is even more relevant in the context of the energy transition, where decarbonization is transforming market dynamics. Renewable sources introduce high variability and uncertainty in prices due to their dependence on meteorological factors. In addition, socioeconomic and regulatory factors add further complexity to price prediction. Industry and academia have shown a growing interest in developing advanced models to address these particularities of the electricity market, which presents unique phenomena such as extreme price peaks or even negative prices, rarely observed in other commodity markets. In this constantly changing environment, it is crucial to develop predictive models capable of adapting to these new dynamics. The main objective of this thesis is to build a comprehensive decision-making scheme based on predictive models that address the complexities of the electricity market. To this end, three lines of research have been developed: feature selection in changing environments, point price prediction and probabilistic price prediction, each with its particular challenges and contributions.

Feature selection in shifting environments is a complex challenge that the literature has dealt only to a limited extent. In volatile electricity markets, the relationships between explanatory features, such as renewable production or demand, and the electricity price, are constantly evolving, making it difficult to identify the most relevant factors. This thesis proposes a methodology that outperforms state-of-the-art techniques, improving price prediction accuracy in high volatility scenarios. Moreover, the robustness of the approach is demonstrated in alternative contexts beyond electricity markets, underscoring is broad applicability.

Point forecasting is the traditional perspective in time series forecasting and remains essential for operational decision making. However, in the electricity market, the robustness of models to different market regimes is crucial, especially in times of high uncertainty. In this thesis we introduce a new point forecasting methodology validated across different European electricity markets and in different time periods, substantially improving performance in volatile contexts and, in particular, in the Spanish market.

Probabilistic forecasting is crucial to provide a measure of the uncertainty associated with the predictions, facilitating better decision making. This thesis presents a novel approach based on quantile regression and Conformal Prediction. The development of a novel conformal algorithm ensures prediction coverage, providing prediction intervals that allow market players to make decisions based on a pre-specified confidence level. Furthermore, this new approach is designed to facilitate its implementation in the industry, since it only requires having different point forecasts of the event of interest and the generated intervals have specific properties that make them suitable to serve as decision support tools. This methodology has been validated on a synthetic controlled example and also with data from the Spanish Day-Ahead Market.

Resumen

La predicción del precio de la electricidad en el Mercado Diario es fundamental para las empresas del sector eléctrico, ya que afecta directamente decisiones operativas y estratégicas. Este proceso es aún más relevante en el contexto de la transición energética, en el que la descarbonización está transformando la dinámica del mercado. Las fuentes de energía renovables introducen alta variabilidad e incertidumbre en los precios debido a su dependencia de factores meteorológicos. Además, factores socioeconómicos y regulatorios añaden una mayor complejidad a la predicción de precios. Industria y academia han mostrado un creciente interés en desarrollar modelos avanzados para abordar estas particularidades del mercado eléctrico, que presenta fenómenos únicos como picos de precios extremos o incluso precios negativos, raramente observados en otros mercados de commodities. En este entorno de cambio constante, es crucial desarrollar modelos predictivos capaces de adaptarse a estas nuevas dinámicas. El principal objetivo de esta tesis es construir un esquema integral de toma de decisiones basado en modelos predictivos que aborden las complejidades del mercado eléctrico. A tal fin, se han desarrollado tres líneas de investigación: la selección de variables en entornos cambiantes, la predicción puntual del precio y la predicción probabilística del precio, cada una con sus retos y contribuciones particulares.

La selección de variables en entornos cambiantes es un desafío complejo que la literatura ha tratado de forma limitada. En mercados eléctricos volátiles, las relaciones entre las variables explicativas, como la producción renovable o la demanda, y el precio de la electricidad, evolucionan constantemente, lo que dificulta identificar los factores más relevantes. Esta tesis propone una metodología que supera las técnicas estado del arte, mejorando la precisión en la predicción del precio en escenarios de alta volatilidad. Además, se demuestra la robustez de la aproximación en contextos alternativos independientes del mercado eléctrico, remarcando así la aplicabilidad en diferentes dominios.

La predicción puntual es la perspectiva tradicional en la predicción de series temporales y sigue siendo esencial para la toma de decisiones operativas. No obstante, en el mercado eléctrico, la robustez de los modelos ante los distintos regímenes de mercado es crucial, sobre todo en épocas de gran incertidumbre. En esta tesis se introduce una nueva metodología de predicción puntual validada a través de diversos mercados eléctricos europeos y en diferentes periodos temporales, mejorando sustancialmente el rendimiento en contextos volátiles y, en particular, en el mercado español.

La predicción probabilística es crucial para ofrecer una medida de la incertidumbre asociada a las predicciones, facilitando una mejor toma de decisiones. Esta tesis presenta un enfoque novedoso basado en regresión cuantílica y predicciones conformales. El desarrollo de un novedoso algoritmo conformal asegura cobertura de las predicciones, proporcionando intervalos de predicción que permiten a los actores del mercado tomar decisiones fundamentadas en un nivel de confianza preespecificado. Además, este nuevo enfoque está diseñado para facilitar su implementación en la industria, ya que solo requiere disponer de diferentes predicciones puntuales del evento de interés y los intervalos generados cuentan con propiedades específicas que los hacen idóneos para servir como herramientas de apoyo en la toma de decisiones. Esta

metodología ha sido validada sobre un ejemplo sintético controlado y también con datos del Mercado Diario español.

Contents

List of Figures

List of Tables

[6]The preprocessing phase was not applied to this dataset because it removed many variables. (Section 2.2)

Abbreviations and acronyms

ACI	Adaptive Conformal Inference
AIC	Akaike Information Criterion
APPA	Asociación de Empresas de Energías Renovables
ARIMA	AutoRegressive Integrated Moving Average
ARMA	AutoRegressive Moving Average
ARX	AutoRegressive with eXogenous variables
AS	Adaptive Standardization
aFRR	Automatic Frequency Restoration Reserve
BOE	Boletín Oficial del Estado
CDTI	Centro para el Desarrollo Tecnológico Industrial
CQR	Conformalized Quantile Regression
DA	Day-Ahead
DNN	Deep Neural Network
EPF	Electricity Price Forecasting
EPEX	European Power EXchange
ESIOS	Sistema de Información del Operador del Sistema
FE	Feature Extraction
FS	Feature Selection
GAMLSS	Generalized Additive Modelos for Location, Scale and Shape
GARCH	Generalized AutoRegressive Conditional Heteroskedasticity
HQR	Heteroscedastic Quantile Regression
IM	Intraday Market
LEAR	Lasso Estimated AutoRegressive
LASSO	Least Absolute Shrinkage and Selection Operator

MAE Mean Absolute Error

MIBGAS Mercado IBérico del Gas

MIC Continuous Intraday Market

ML Machine Learning

MSE Mean Squared Error

mFRR Manual Frequency Restoration Reserve

NP Nord Pool

OMIE Operador del Mercado Ibérico de Energía

PBC Base Matched Programme

PDVP Provisional Daily Viable Programme

PIMP Permutation IMPortance

PVB Virtual Balancing Point

QRA Quantile Regression Averaging

REE Red Eléctrica de España

rMAE Relative Mean Absolute Error

RMSE Root Mean Squared Error

RR Replacement Reserves

sMAPE Symmetric Mean Absolute Percentage Error

SHAP SHapley Additive exPlanations

WACI Width-Adaptive Conformal Inference

1 Introduction

1.1 Energy transition

Electricity is one of the most essential commodities in contemporary society and, like the rest of the energy landscape, is in an unprecedented state of transition. Mitigating the effects of climate change has become a global priority and among the measures adopted to achieve climate objectives highlights the growing role of non-fossil fuel based energies in the various energy mixes of each country. In 2023, installed renewable generation in Spain reached 61.3%, resulting in a real renewable production of 50.3% (REE, 2023). New factors have emerged, such as self-consumption, which in 2024 represented a production equivalent to 3.7% of national electricity demand (APPA, 2024), the development of batteries or green hydrogen. However, the incorporation of these renewable energy sources entails certain challenges, such as the problems of security and stability of the grid, derived from the intermittency of these sources, or curtailments that may occur[1].

This evolving landscape is bringing significant changes to electricity markets, requiring substantial investment. Accurate electricity price forecasting is essential for assessing the viability of these financial commitments and supporting the energy transition.

1.2 The Iberian electricity market

The Iberian short-term electricity market, on which this thesis focuses, is mainly composed of the Day-Ahead Market, the different sessions of the Intraday Market and the Imbalance Market. After the assignments in each of these markets, a process of technical restrictions is carried out, in order to achieve the security and viability of the programme.

1.2.1 Day-Ahead Market

The Day-Ahead (DA) Market, managed by the Iberian market operator (OMIE), has the purpose of carrying out electricity transactions for the next day. Energy purchase and sale offers for the 24 hours of the next day are presented on this market.

[1] Curtailments are caused when more energy is produced than demand exists at a given moment, wasting part of the energy generated.

To achieve this objective, the different generation and consumption offers for the day $D+1$ are presented no later than 12 noon on the current day, which will be denoted by day D, and, by 13:30, the Base Matched Programme (PBC) that establishes how this energy is to be generated must be known, although this only corresponds to an initial version. Figure 1.1 shows graphically the auction structure of the DA Market. Note that each hour has its own auction.

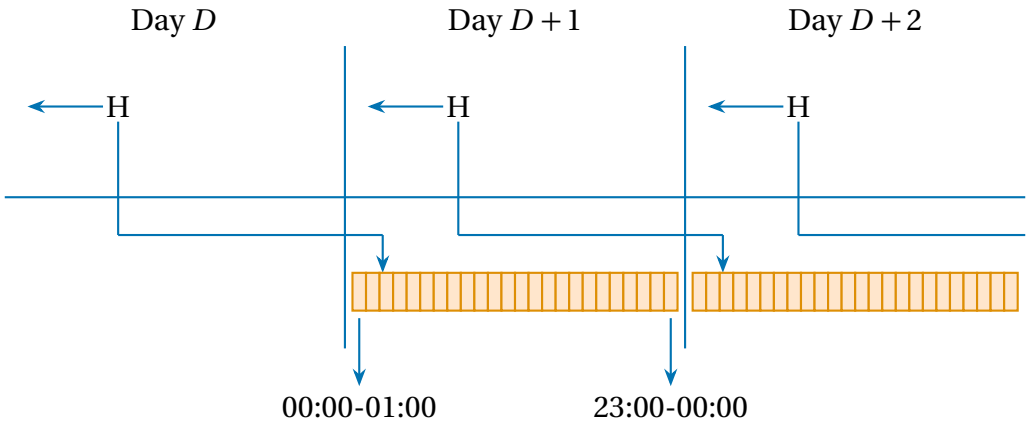

Figure 1.1: Daily structure of the Day-Ahead Market. In Europe the H-hour is usually 12:00 noon.

The determination of the PBC is made through the different bids of the market participants. For each hour, generation bids are ordered in ascending order and demand bids in descending order. The intersection of the supply and demand curves determines the electricity price (marginal price) for all matched bids. These include supply bids that were originally less than or equal to the matching price, as well as demand bids that were greater than or equal to it, provided they meet the specific complex conditions set by each market participant (Figure 1.2).

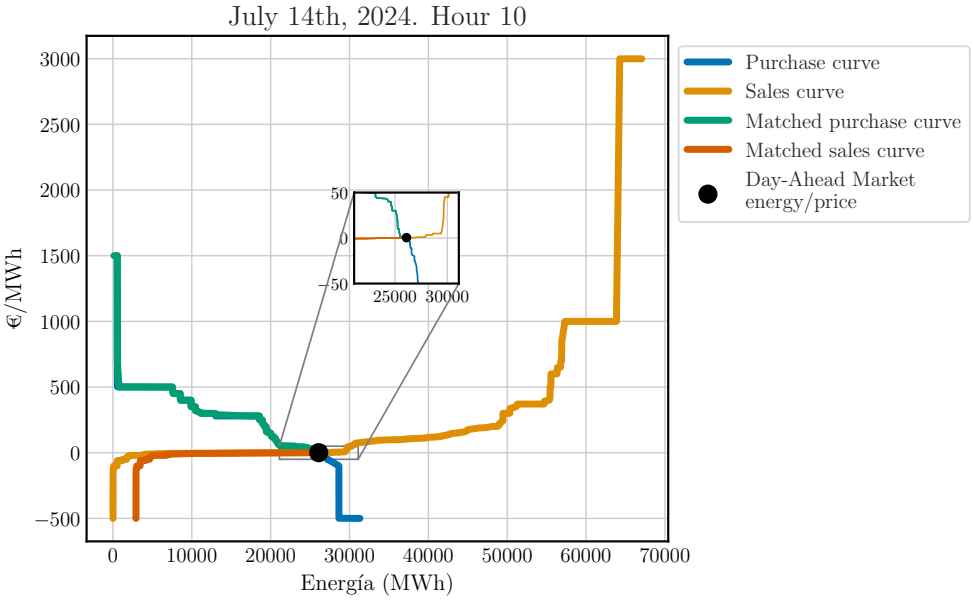

Figure 1.2: Purchase and sales curves for 14th of July 2024 in the period from 09:00 to 10:00.

After the generation of the PBC, there is a period of resolution of technical restrictions where the System Operator (Red Eléctrica de España) makes the modifications it considers appropriate to ensure that the generation of energy is safe and viable. This programme is called Provisional Daily Viable Programming (PDVP), which must be communicated to OMIE before 14:45 hours, concluding the Day-Ahead Market planning. In Figure 1.2 it can be seen the original bid staircases that would determine the PBC and the modification that is made that results in the PDVP.

1.2.2 Intraday Market

The Intraday Market, also managed by OMIE, consists of six energy auction sessions after the Day-Ahead Market. The procedure for energy matching is identical to the previous case, except that in each session different hourly periods are auctioned. The different sessions are described in the Table 1.1.

	Session 1	**Session 2**	**Session 3**	**Session 4**	**Session 5**	**Session 6**
Opening	17:00	21:00	1:00	4:00	8:00	12:00
Closing	18:45	21:45	1:45	4:45	8:45	12:45
Matching	19:30	22:30	2:30	5:30	9:30	13:30
Technical restrictions	20:10	23:10	3:10	5:10	10:10	14:10
Programme publication	20:20	23:20	3:20	6:20	10:20	14:20
Hourly periods	22(D)-24	1-24	5-24	8-24	12-24	16-24

Table 1.1: Different Intraday Market sessions

Since June 2024, the structure provided in Table 1.1 has been modified, maintaining only sessions 1, 3 and 6. This decision is due to the intention of aligning the market sessions of most European countries, facilitating the negotiation of interconnection between them. As the data used in this thesis is prior to this regulatory change, the old Intraday Market structure will be taken into account. Further developments in this direction are being implemented: in March 2025, intraday market trading was introduced with a 15-minute granularity, replacing the previous hourly framework. The same modification is expected to be adopted in the Day-Ahead Market by June 2025.

In addition, since 2018, intraday auctions coexist with the Continuous Intraday Market (MIC). In the MIC, electricity can be bought and sold from 15:00 on the day D, until one hour before the start of the delivery of electricity on the day $D+1$. Instead of following an auction structure, market participants can place bids to buy or sell power continuously 24 hours a day. However, prices are dynamic, as it is a pay-as-bid system. In other words, agents receive the price they themselves have requested, which can be very different from the outcome of the auction markets. Figure 1.3 shows the transactions executed at the MIC on a given day. The difference with the auction prices can be discerned.

Intraday markets are of vital importance in the electricity system. Firstly, they allow each agent's programmes to be adjusted according to variations in generation and consumption. This is a critical factor due to the growing incorporation of renewable energies into the system. They

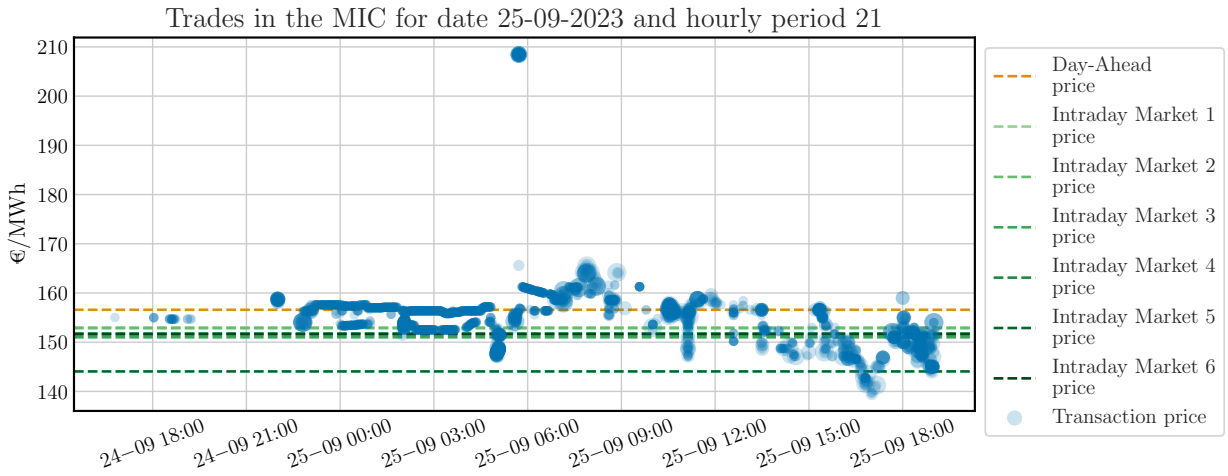

Figure 1.3: Price of the different transactions in the MIC over the trading period of the hour 21 of the 25th of September 2023. The size of the points is determined by the energy involved in the transaction. The colours represent different markets.

also allow for the correction of errors that may have been made in the Day-Ahead Market. In short, they facilitate the operational flexibility of agents in the market, while the electricity system benefits in terms of security and efficiency by favouring a balance between generation and demand.

1.2.3 Imbalance Market

The Imbalance Market is responsible for guaranteeing the balance between generation and demand in the system as a whole. There are three services for this purpose: RR balancing energy (replacement reserves), tertiary regulation (mFRR) and secondary regulation (aFRR), whose management is mainly based on demand and renewable forecasts. Each service exists in both directions, upwards or downwards, depending on whether there is a shortage of surplus of energy in the system.

Energies are activated in the order of RR, tertiary, and secondary. All are designed to free energy from balancing services operating on an activation horizon closer to real-time. A brief description of each follows.

RR balancing energy. It is a manually activated energy and has a quarter-hourly resolution period. If a RR energy service is activated, then it must be active for a minimum of 15 minutes and a maximum of 60 minutes for full quarter-hourly periods. Participation in this service is optional, both for generation and demand, and the criterion for activation of one unit or another is purely economic, following the order of bids, which can be made up to 30 minutes before the start of supply. Unlike other balancing energies, this is part of a European platform, making it a cross-border product.

Tertiary regulation. Tertiary regulation energy is still manually activated with a quarter-hourly resolution period, but if an agent participates in this service, then it is obliged to participate with all the energy it has available in each direction. Even if obliged to participate, the agent can bid that energy at any price it wishes. The allocation of bids follows the same criteria as above, by economic order. These bids may be updated up to 25 minutes before the start of supply. If a unit is assigned to the application of the service, the activation of the service has a maximum duration of 15 minutes.

Secondary regulation. This energy is self-triggered and is unique in that bids are made in energy bands, rather than as a single value. The offer is made the day before the service is to be offered, which is 5 to 7.5 minutes prior to the corresponding quarter hour. Activation is constant, with a duration of 100 seconds.

Each market participant has committed to consume or generate a certain amount of energy in each hour. Comparing actual consumption/generation with what is committed results in deviating energy. The sum of all deviations is the energy that the system operator has to cover through the operation of these markets. The system is designed to penalize the deviating energy of each market agent, which is paid at a specific price: the imbalance price.

In April 2022, a new system for calculating the imbalance price was established, distinguishing between two types of prices:

1. A dual price is given for all those hours when the sum of the volume of secondary and tertiary energies activated in the minority direction of the system requirement is greater than 2% of the sum of the volume of secondary and tertiary energies activated in the majority direction.
2. Otherwise, a single imbalance price is given.

When the price is dual, the rule of weighted prices is applied. In the event that a given unit has fallen short (from the demand point of view, it consumes more than the final schedule), the weighted average price of the imbalance energies activated to cover that direction is paid. In the event that a given unit has been long (from the demand point of view, less is consumed than set in the final schedule), the weighted average price of the activated imbalance energies in the opposite direction is paid. Alternatively, when the price is unique, if the energies activated are upwards, both the units that remain short and long pay the weighted average price of the energies to go up, while when the energies activated are downwards, in both cases the weighted average price of the energies to go down is paid.

1.3 Background

Since the appearance of competitive electricity markets, forecasting prices in different markets has been of vital importance in the decision-making process for companies in the electricity sector (Hong et al., 2020; Mayer and Trück, 2018).

But price forecasting is not a simple task. The current economic unviability of storage and the

necessary balance between generation and consumption lead to a very unique price behaviour. In addition, there are other causes that increase the complexity of the prediction task considerably: the dependence on meteorological variables, together with the increasingly frequent appearance of renewable energy sources, or the human behaviour itself, cause an increase in price volatility that is not present in other commodities. This leads to price spikes or even negative prices (Bunn, 2000; Weron, 2014). As a result, the industry is required to improve its forecasting systems and the academia is interested in the study of the resulting time series because of their great singularity. Figure 1.4 shows the evolution of prices over time and how the difficulty of prediction has increased in recent years. For example, the increase in volatility can be seen due to political-economic factors or the change in behaviour in hourly patterns due to the strong presence of renewable energies (see the evolution of prices during solar hours).

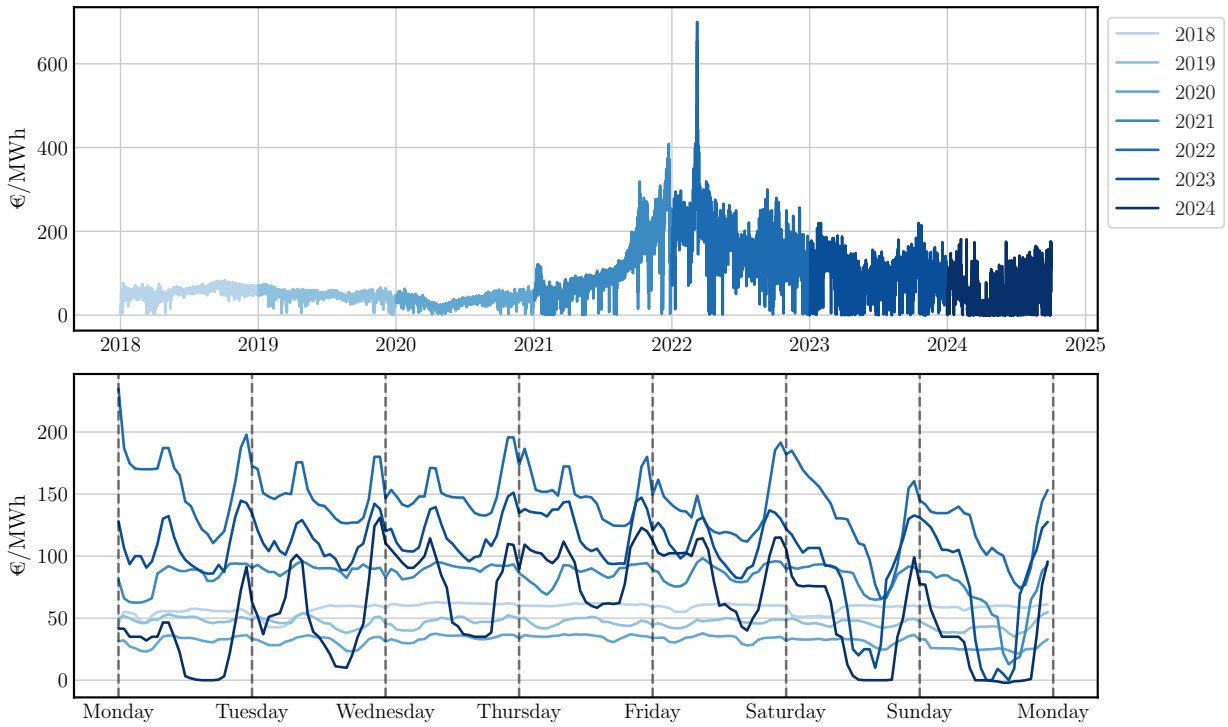

Figure 1.4: (Top) Evolution of DA prices from January 2014 to October 2024. **(Bottom)** Example of a typical weekly profile from 2018 to 2024

Moreover, in this context where prices are surrounded by high uncertainty, the prediction of the expected price alone does not provide enough information to make optimal decisions. The entire forecasting scenario needs to be evaluated. One way to achieve this objective is by making probabilistic forecasts that serve to quantify the uncertainty associated with the predictions. (Nowotarski and Weron, 2018).

It is important to note that the electricity market is characterized by a highly complex and dynamic environment, where socio-economic changes have a direct influence. This market operates in a context of constant adaptation, which generates continuous regulatory modifications that have a significant impact on its functioning. These factors are of crucial relevance for market participants, as they have a direct impact on their behaviour and, consequently, on price

formation. Therefore, any methodology used in the development of price forecasting models must take this dynamism into account to ensure their accuracy and effectiveness. Furthermore, the methodology must be computationally efficient, otherwise the use of the models within the industry will not be feasible.

1.4 The BrainEN project

In this scenario, Fortia Energía, a company operating in the electricity market, has undertaken a research project funded by the Centre for the Development of Industrial Technology (CDTI): "Experimental research in innovative technologies for an efficient and sustainable energy community (BrainEN)". Specifically, this thesis is part of this project.

One of the possible ways to facilitate the energy transition is through active demand response (Bakare et al., 2023). Fortia's main objective within the project is to highlight the value of industrial demand flexibility in order to achieve greater integration of renewable energy into the Spanish electricity system. One of the tasks to be carried out by Fortia is included as an *industrial* objective of the thesis: to predict the hours of lowest price in the peninsular electricity market, which correspond to the hours of highest supply of renewables, in order to adapt industrial consumption and avoid the potential curtailment in hours of overgeneration.

1.5 Electricity Price Forecasting

Predictive processes play a key role in many industries, and the energy sector is no exception. In industries such as electricity, gas, petroleum and coal, decision support systems based on predictive models allow informed decisions to be made, effective strategies to be designed and risks and losses to be minimized. This thesis focuses on the electricity sector, specifically on price forecasting. However, other factors are of great relevance for decision making in this field, such as load forecasting (Nti et al., 2020) or the generation of renewable energies, especially wind and solar photovoltaic (Hanifi et al., 2020; Antonanzas et al., 2016).

The literature on electricity price forecasting is extensive and diverse. Two fundamental questions help to delimit the type of study and the approach used: which market is to be predicted and which is the prediction horizon. Most studies, like this thesis, focus on the short-term Day-Ahead Market, usually with a horizon of one day. However, relevant contributions can be found in research addressing other markets and longer time horizons.

The choice of model, whether data-driven or fundamental (Weron, 2014), is a critical distinction that is typically linked to the forecasting time horizon. Data-driven models use historical information to unravel the relationships between variables, either through econometric approaches or through Machine Learning techniques, which have become increasingly popular. These models are often applied to short time horizons, although this is not always the case (for instance, Ghelasi and Ziel (2024)), as they tend not to extrapolate accurately to medium or long-term horizons, especially when market conditions evolve dramatically over time (Bello et al.,

2016). In contrast, fundamental models, which aim to replicate price formation mechanisms, are more effective for long-term forecasting. They are not suitable for hourly granularity, as they tend to group hours with similar electricity load characteristics, but they allow decisions to be made over longer time horizons, which is crucial in policy making or long-term energy contracting. Moreover, these models tend to be highly explainable, a key feature for decision making at such a level of anticipation.

In relation to the different markets, as mentioned above, the Day-Ahead Market is the most studied, but research on intraday and imbalance markets is crucial to establish a complete sequential decision-making scheme. One of the main challenges is that, while the DA Market is quite similar or even identical in most European countries, the structure of the intraday markets can vary significantly. In Section 1.2, it was introduced that, in the case of Spain, there are several auction-based sessions of the Intraday Market, which coexist with a Continuous Intraday Market. However, in other countries this is not the case. If we take the German market as an example, there is only one intraday auction that determines the opening prices of the MIC, which is the only market, together with the imbalance markets, available after the Day-Ahead auction. This implies that the studies are highly dependent on the country in question, which makes it difficult to generalize conclusions across markets. In fact, the products traded also differ from market to market: in Spain, before March 2025, only hourly products were traded, while in Germany and other European countries there were also quarterly products. In addition, the maturity of these markets is a key factor influencing the techniques and the results obtained. Methods that perform well in certain markets may fail in other due to the specific behaviour of market participants. The Intraday Market falls outside the scope of this work; however, further reading is provided for those who may find it relevant: Monteiro et al. (2016); Andrade et al. (2017); Kath and Ziel (2018); Janke and Steinke (2019); Maciejowska et al. (2019); Narajewski and Ziel (2020); Serafin et al. (2022); Cramer et al. (2023); Klein et al. (2023); Hirsch and Ziel (2024)

Gradually, regulations are being developed to unify intraday markets at European level (BOE, 2024). Such changes are also being implemented in the imbalance markets[2] and, although these markets have historically shown greater similarity, particularly in relation to aFRR, mFRR and RR energies (Section 1.2), the calculation of the imbalance price still shows significant differences between countries. There are very few studies on imbalance price forecasting due to the great complexity involved. This price is influenced by numerous real-time factors, many of which are impossible to anticipate, such as failures in generation units, and its volatility is considerably higher than in the DA or in the various IMs. Known studies include Klæboe et al. (2015); Dumas et al. (2019); Bunn et al. (2020); Lucas et al. (2020); Narajewski (2022) y O'Connor et al. (2024), where already the first one stated what has been confirmed in the following ones: it is practically impossible to obtain accurate results on this market by making predictions prior to the DA.

Short-term forecasts in the DA best reflect how the academic approach to the challenge of electricity price forecasting has evolved. Each of the chapters of this paper will provide a comprehensive review of the key areas covered in this thesis: feature selection, point forecasting and probabilistic forecasting. However, in order to properly understand the advances presented,

[2]Through the MARI, PICASSO and TERRE platforms.

a recapitulation of previous developments in this field is necessary.

One of the earliest models was introduced by Nogales et al. (2002), in which past prices and electricity load values are used as explanatory variables by using ARMA(X) models. Afterwards, in Contreras et al. (2003) ARIMA(X) models were employed, concluding that, while the results were reasonably good, they were inferior to those obtained by Nogales et al. (2002), probably due to differences in the way the load feature is included in each study. Even so, the results were remarkable in comparison with studies involving neural networks, such as that by Szkuta et al. (1999). Nevertheless, these works did not have a large historical data set, which made it difficult to properly evaluate the neural networks in these scenarios.

Classical statistical models continued to be developed, such as in the work of Misiorek and Weron (2005), where it is shown that modelling each hour separately, using 24 different models, produces better results than grouping all hours into a single model, since the correlation structure between hours varies significantly. Furthermore, the inclusion of explanatory features such as electricity load in this type of model produces improvements that are difficult to overcome (Weron and Misiorek, 2006). This approach was extended by García-Martos et al. (2007), who in addition to differentiating by hours, included a differentiation by type of weekday (working day or weekend).

While the ARIMA family models were being developed, applications of ARIMA(X)-GARCH models were also explored. However, findings on the effectiveness of these models are mixed. The most recent study on this topic, conducted by Janczura and Puć (2023), concludes that explicit modelling of volatility through GARCH models does not improve point forecasts of electricity prices.

Although studies based on statistical models have continued since then, improvements beyond autoregressive models with exogenous variables have been practically insignificant. In this context, the model that is considered state of the art within the classical statistical perspective is LEAR, presented by Uniejewski et al. (2016). This is a linear autoregressive model that incorporates automatic feature selection by means of a L1 penalty on the model parameters.

Over time, the availability of electricity price data has increased considerably, which has allowed a re-evaluation of models based on Machine Learning methodologies, in particular those based on neural networks. This new paradigm has resurfaced previously settled debates, such as the apparent superiority of separate 24-hour modelling versus a single model (Ziel and Weron, 2018), thanks to the new computational capabilities available. While there are several studies analysing these techniques (Ugurlu et al., 2018; O'Leary et al., 2021; Tschora et al., 2022), the model considered state of the art from this perspective is the neural network presented in Lago et al. (2018). Even though more models have been developed within the branch of neural networks with the focus on electricity price forecasting applications, such as Olivares et al. (2023), the gain with respect to the model of Lago et al. (2018) is not very remarkable and produces a significant increase in complexity. In Chapter 3, when the problem of point price forecasting is addressed, the models of Uniejewski et al. (2016) and Lago et al. (2018) will be discussed in more detail.

There have been other lines of work to predict DA electricity prices, such as Functional Data Analysis (Chen and Li, 2017; Jan et al., 2022) or working directly with sales and purchase curves

(Ziel and Steinert, 2016). Although interesting results have been achieved, in terms of performance they do not improve on ML models.

Whereas the model used has been the main focus of research within the electricity price forecasting methodology, there have been other key points of interest. Within feature selection, as mentioned above, L1 regularization is the most widely used approach, but other techniques have been successfully tested such as Orthogonal Matching Pursuit (Nickelsen and Müller, 2024). In any case, the most important factors generally agree in the different studies. Since the beginning of EPF research, the expected electricity load is one of the most important exogenous variables. Given their growing presence in electricity grids, renewable energy forecasting is essential for good forecasting. All these features are assumed to be related to the system for which the price is to be forecasted, but the inclusion of these variables from neighbouring systems could be considered, simulating the role of the interconnection. The price of natural gas, coal, petrol or CO2 emission rights have always had a great influence on prices, as they directly affect the generation costs of the different thermal power plants in the system in question. Figure 1.5 shows the price of the Spanish DA over the years and the evolution of some of these exogenous variables. For example, some peaks in demand (look for example at the beginning of 2021), correspond to peaks in prices. The effect of the COVID-19 pandemic can also be observed on this variable. With respect to renewable energy, it can be seen that the increase in renewable energy in turn produces an increase in price volatility, as was discussed in Section 1.3. The gas price crisis due to the Russian-Ukrainian conflict and its impact on prices is also noticeable.

The volume of data is substantial and continuous to grow over time. Thus, a key question is identifying the most suitable data transformation to enhance predictive performance. There has already been work addressing this question (Uniejewski et al., 2017) and it will be discussed in more detail in Chapter 3.

Another issue related to electricity prices that is becoming increasingly relevant is the forecasting of electricity price peaks. Increased price volatility has led large industrial consumers to optimize their production plans based on expected electricity prices. Periods of very low or even negative prices present significant risk for electricity generators. Therefore, knowing the price peaks in advance can be decisive for several important market participants. Thus, a whole literature is developing around this sub-area. The discussion in this topic of Sheybanivaziri et al. (2024) is recommended for more information on this subject.

All these developments have contributed to improve point price forecasting models. However, market participants need to be confident in their predictions in order to be able to make appropriate decisions. Quantifying that certainty is therefore vital. Thus, a whole branch of probabilistic forecasting has developed in the electricity market and in price forecasting. Probabilistic forecasting will be discussed in general in Chapter 4, but the main developments in the EPF field are presented here.

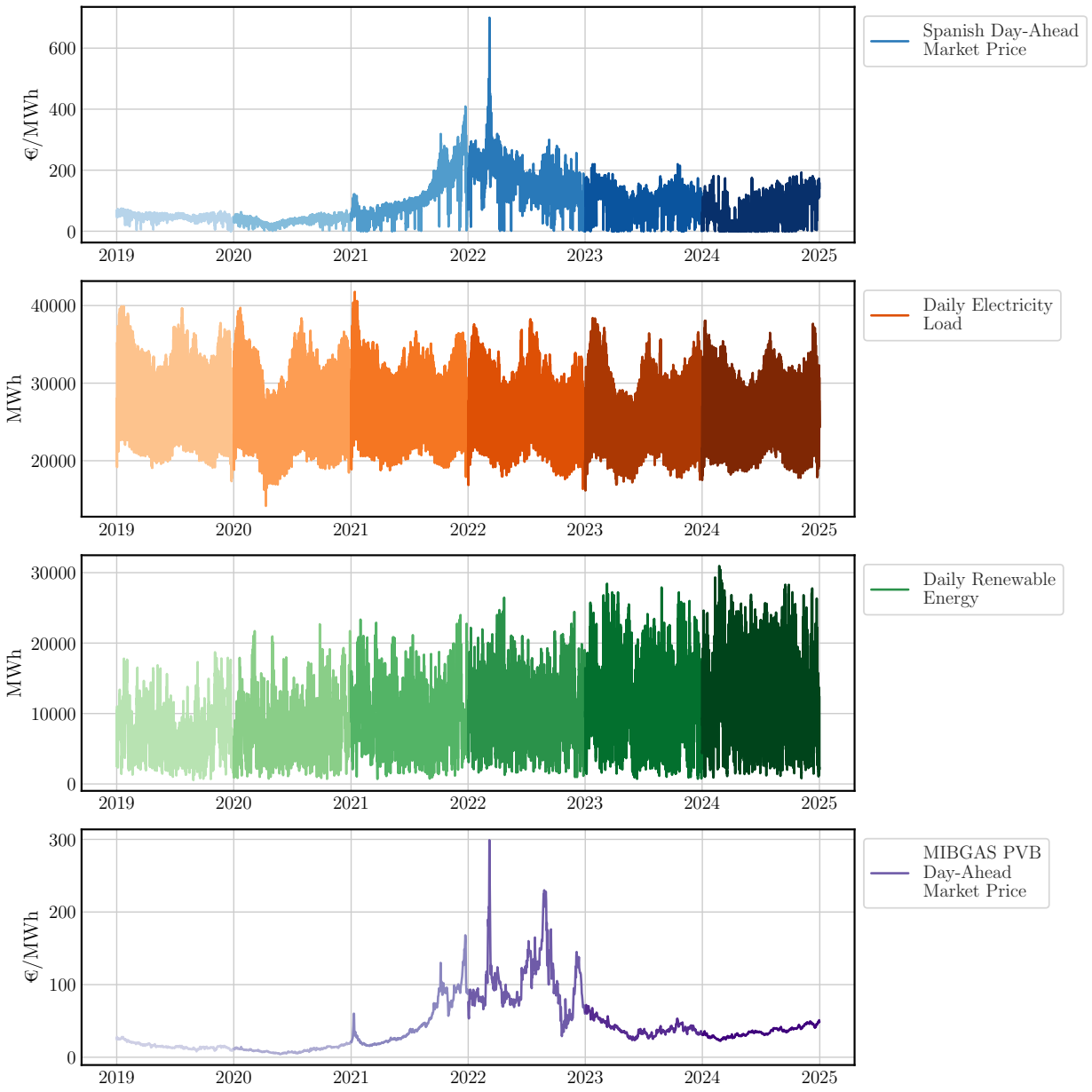

Figure 1.5: Evolution of DA electricity prices in Spain and of some explanatory features: electricity load forecast, renewable generation forecast and gas price in its Day-Ahead Market.

A satisfactory idea was introduced in Nowotarski and Weron (2015) named by Quantile Regression Averaging (QRA): applying quantile regression (Koenker and Bassett Jr, 1978) using the predictions obtained by point estimators as explanatory variables. Given the good results of the LEAR model in point estimation, Uniejewski and Weron (2021) proposed the use of quantile regression following the same philosophy, but applying L1 regularization on the loss function, so that an automatic selection of variables (point predictors) is carried out, improving the results. Fully connected neural networks were tested in Dudek (2016), producing probabilistic outputs assuming that the error distribution is Normal. Neural networks have also been tested

but with a Bayesian perspective in Brusaferri et al. (2019), assuming normal priors on the network weights. Complex recurrent or convolutional neural network structures in a probabilistic context are compared in Mashlakov et al. (2021) but focusing on the electricity market in general, not only on price. More classical techniques such as bootstrap on the residuals (Efron, 1992) to obtain probabilistic results have also been used for price prediction in the electricity market (Narajewski and Ziel, 2021). A more in-depth study on price appears in Marcjasz et al. (2023) through the use of distributional neural networks, used for the first time in this field. The particularity they present compared to conventional neural networks is the use of a last distributional layer, while trying to maximize a scoring rule such as the log-likelihood. They are applied using Normal or Johnson's SU distributions, the latter being the state of the art in this subdomain of EPF. It is important to note that the conformal prediction framework (Vovk et al., 2005), which will be of high importance in this thesis, has also been used previously in the context of EPF in the work of Kath and Ziel (2021), where valid intervals are obtained by improving some metrics with respect to the QRA.

1.6 Objectives and thesis outline

The predictive task discussed is influenced by many factors. In addition, a large history of values of both explanatory features and the target variable is available. This paradigm generates a highly favourable situation for the use of Machine Learning (ML) techniques, which more and more frequently demonstrate their effectiveness in highly complex tasks.

As has been mentioned, the thesis is framed within an industrial project. The importance of this fact is underlined in the *academic* objectives that are proposed:

- The explanatory features associated with the different electricity market prices have undergone major changes over time, as they are influenced by regulatory, social, economic or political situations. This has meant that their behaviour has varied over time, giving rise to the so-called *shift* situations.

  This changing paradigm, although it occurs frequently in a number of real problems, has not been studied in much depth. In particular, studies of *concept shift*, which occurs when the relationship between the explanatory features and the target variable varies over time, are the least studied. This type of behaviour can be observed in the electricity market. Thus, any development should take this fact into account by contributing to the literature in this field.

- As in any project related to Data Science, the phase prior to the application of learning algorithms is of vital importance. This includes feature selection. Traditionally, the features that most influence the target variable are the most important and those that should be selected to build a statistical learning model. However, when there are phenomena that cause the relationship between the different variables and the target variable to vary, the (historically) most important variables may begin to negatively influence the prediction of the variable of interest. Therefore, this work focuses on selecting features that maximize prediction performance and remain robust over time, avoiding the influence of variables

whose historical importance may distort model performance in shifting environments.

a method will be developed that looks only for features that have a positive influence on the prediction of the target variable, regardless of the importance of the features that may distort the most robust selection in shifting situations. Chapter 2 addresses this objective.

- Build the state of the art electricity price forecasting model in the Day-Ahead Market. In particular, to obtain the best possible performance in periods of higher volatility. In addition, the model must be relatively fast in both training and execution, as it must be used in a real industrial context. Chapter 3 describes how this task has been accomplished.
- Having established the importance of uncertainty quantification in the current market paradigm (Section 1.3), a methodology for creating prediction intervals with a special focus on industrial decision making and supported by theoretical foundations must be developed. Thus, a new probabilistic prediction methodology is detailed in Chapter 4.
- Most of the relevant research in the field of Electricity Price Forecasting (Section 1.5) has been performed using data sets prior to 2021, where the market is very stable. Nevertheless, it is since this date that market behaviour has changed considerably. Thus, developments should be tested over a period later than 2021, as their application needs to be valid in the current market that the industry is facing. At the same time, it will also serve as a validation of existing studies carried out in earlier periods.

2

Feature selection in concept shift scenarios

Note

Some of the contents of this chapter are published in the paper: Carlos Sebastián & Carlos E. González-Guillén. A feature selection method based on Shapley values robust for concept shift in regression. *Neural Computing and Applications*, 1-23, 2024.

Artificial Intelligence and Machine Learning in particular are becoming increasingly valuable nowadays. The understanding of the vast amount of data available being generated every day in a connected world, offers many possibilities like Natural Language Processing systems, recommendation engines, medical diagnosis or forecasting models like the ones that are going to be developed in this thesis. Thus, the use of historical data to build models capable of capturing relationships between different variables is becoming more and more frequent, either for explanatory or predictive purposes. However, different issues may arise when modelling a problem: the trade-off between the explainability and accuracy of different algorithms, computational complexity... A common problem to all fields is the selection of the correct information in high-dimensional contexts. Failure to perform this task properly can lead to problems of overfitting, that is, of poor generalization.

Dimension reduction is generally performed in two ways: feature extraction (FE) or feature selection (FS). FE consists of projecting the feature space in a high dimensional space to a smaller one, while FS chooses a subset of the original features (Venkatesh and Anuradha, 2019). In a large number of applications, transforming the original variables can make it challenging to analyse the results. This modification can be problematic when interpretability is needed, as the intuition behind the original features is lost. That is why, in such cases, feature selection is usually preferred (Li et al., 2017). Moreover, even when the dimension of the data is not so high, it is important to follow the principle of parsimony or Ockham's razor, as the models built will be more explainable and, in case they are used as predictive models, more general and robust.

When fitting a model with real data, a very common problem is the so-called dataset shift. This occurs when the statistical properties of the target variable change over time (Widmer and Kubat, 1996). This shift may occur for seemingly arbitrary reasons, but it may also be due to changes in the explanatory variables. If there is not sufficiently long history of data available since the change occurred and the affected variables are in the explanatory variables of the model, the relationship learned by the model through these features may be erroneous.

Feature selection algorithms do not detect this kind of problem, as they usually take as reference a measure of the degree of importance of the variable to determine whether it is relevant or not. However, they do not observe whether the effect that each variable has on the predictions is the desired one, regardless of the magnitude of the effect.

In this chapter a new method for feature selection in regression problems that is robust to the presence of characteristics whose behaviour has changed over time and have an undesirable effect on predictions in the current context in introduced. This selection is done using Shapley values, as previously considered in Marcílio and Eler (2020), Keany (2020), Calzolari (2020) or Verhaeghe et al. (2022). While all these works consider a global treatment of each feature to asses their importance, our method relates Shapley values to the errors of the predictions in order to pay special attention to the local effects of each variable for each prediction.

It is important to note that this is not a method to detect or characterize a dataset shift, but rather that in the event of any variable undergoing some sort of shift that negatively influences the performance of the model, this variable is a candidate for elimination, regardless of the overall level of influence it has on the behaviour of the model.

2.1 Literature review

Feature selection is a complicated task when there is no help from experts in the field. The selection in such cases has to be made directly from the data and optionally by using statistical models. Establishing the state of the art in variable selection methods is a difficult task, as it depends considerably on the task to be performed (classification or regression) and the dataset employed. However, there is a clear classification of the different types of existing methods: filters, wrappers and embedded methods.

Filters Filters make a selection of features using only the relationships that exist between the data, they do not rely on the use of any model. Thus, some criterion is used to establish a ranking among the variables and those that exceed a certain threshold in this ranking are chosen as relevant. Examples of different criteria are: correlation with the target variable, information gain, Fisher score, variance threshold or the chi-square test (Venkatesh and Anuradha, 2019). Since they do not rely on a model, they tend to be computationally very efficient methods. However, they require making assumptions about the data that are not always met, leading to a selection that is not necessarily optimal under the criteria used.

Wrappers Wrapper methods, popularized by Kohavi and John (1997), are those that interact with a predefined model to assess the quality of the selected feature set. The methodology is divided into two steps: finding a feature set and assessing the quality of the feature set (Li et al., 2017). Three common strategies when proposing a dataset are forward selection, backward selection and stepwise selection (Colaco et al., 2019). The use of metaheuristics, typically bio-inspired (Diao and Shen, 2015), is also common. There are methods that were specifically designed for certain models, such as the Boruta method (Kursa and Rudnicki, 2010) for Random

Forest models. The constant interaction with a model makes them computationally more expensive than filter methods, although the results are usually better (Venkatesh and Anuradha, 2019).

Embedded methods Embedded methods are those that incorporate variable selection into the learning phase of the model. In this way they are computationally efficient and take advantage of the benefits of interacting with a model. The most common are those that add some kind of regularization to the learning process, so that the coefficients of certain features are forced to be zero or close to zero (Li et al., 2017). Although there are complex methods, it is very common to use the already mentioned LASSO regression to select variables (Section 1.5), obtaining more than acceptable results on many occasions and in many subfields of time series forecasting (Petropoulos et al., 2022).

2.1.1 Shapley values in the context of feature selection

Shapley values (Shapley, 1953) are a game theory concept to explain the importance of an individual player in a collaborative team. The idea behind the notion is based on distributing the total gain among the players according to the relative importance of their contributions. Specifically, the Shapley value for player i is defined as the average marginal contribution over all possible coalitions. This is

$$\phi_i = \sum_{\mathscr{S} \subseteq \mathscr{N} \setminus \{i\}} \frac{|\mathscr{S}|!\,(|\mathscr{N}|-|\mathscr{S}|-1)!}{|\mathscr{N}|!} (v(\mathscr{S} \cup \{i\}) - v(\mathscr{S}))$$

where $\mathscr{N}$ is the set of players and $v : \mathscr{P}(\mathscr{N}) \longrightarrow \mathbb{R}$ a function that returns the gain of a coalition.

This idea can be transferred to the explainability of Machine Learning models (Štrumbelj and Kononenko, 2014). Specifically, the game would be the task of predicting an instance of the dataset, the gain would be the difference of the prediction with the average prediction of all instances, and the players would be the values of the different variables that collaborate to obtain the prediction. In this way, Shapley values allow us to measure the influence that each variable has on a prediction, distinguishing the characteristics that have a higher or lower impact on the predictions made by a model. However, for high dimensional data the calculation of the Shapley values in an exact way is not feasible, as the computational complexity is $\mathscr{O}(2^{|\mathscr{N}|})$.

The use of Shapley values as a local model-agnostic technique for the explainability of Machine Learning models became popular with the introduction of SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) and in particular with the appearance of TreeSHAP (Lundberg et al., 2018), which allows the efficient calculation of an approximation of Shapley values for models based on decision trees. Although this approximation presents problems estimating non-zero Shapley values in features that are not relevant but have a high correlation with variables that are influential (Sundararajan and Najmi, 2020; Janzing et al., 2020), it works well in practice.

For other types of models, such as neural networks, there are techniques such as the one presented in Castro et al. (2009) or Ancona et al. (2019) that allow the calculation of an approxi-

mation of the Shapley value in polynomial time. For instance, the first approximation is based on random sampling from the equivalent definition of Shapley value

$$\phi_i = \frac{1}{|\Pi(\mathcal{N})|} \sum_{\pi \in \Pi(\mathcal{N})} \left(v(\mathcal{P}_i^{\pi} \cup \{i\}) - v(\mathcal{P}_i^{\pi}) \right)$$

where $\Pi(\mathcal{N})$ denotes the set of permutations of $\mathcal{N}$ and $\mathcal{P}_i^{\pi} = \{j \in \mathcal{N} : \pi(j) < \pi(i)\}$ is the so-called predecessor set of the player i, that takes the position $\pi(i)$ in the permutation $\pi \in \Pi(\mathcal{N})$.

Taking into account that the Shapley value can be used to measure the influence of a characteristic on a prediction, an overall measure of the influence of a variable can be defined by taking the mean of the absolute values of the Shapley values of every observation for each feature. This measure can be used to filter by a minimum overall influence or to establish a ranking and select the k most important variables (Marcílio and Eler, 2020). In practice this method works well, although the idea of using Shapley values within the variable selection field can be refined considerably.

A first idea is the variation of the Boruta (Kursa and Rudnicki, 2010) method, Boruta-Shap (Keany, 2020). Boruta-Shap, like Boruta, is based on the use of shadow variables, which are copies of the original variables with the values randomly permuted. The general idea is that a variable is relevant if its importance measure is greater than that of the best shadow variable. Boruta-Shap modifies the algorithm by using the previously described global influence as a measure of importance and making technical modifications to speed up the process, establishing that not only better feature sets are obtained than through the original algorithm, but also in less time.

Another modification is Shapicant (Calzolari, 2020), which is inspired by the Permutation Importance (PIMP) method, (Altmann et al., 2010). In PIMP, a model is trained on the original data and an importance measure is obtained for each variable. Then, a permutation of the target variable is made and the model is re-trained to obtain another importance score for each of the variables. The process is repeated several times and if a variable is significantly more relevant on average over the original dataset than over the randomized set, then it is considered to be relevant. Shapicant uses the Shapley values as a measure of importance, but separating positive and negative values.

A new original method was proposed by Verhaeghe et al. (2022), Powershap. This algorithm is divided into two phases: in the *Explain* component, a uniform random variable is added to the dataset, a given model is trained and the overall influence measure is calculated based on Shapley values for each of the variables, including the random variable. This procedure is repeated for a fixed number of iterations but varying the random variable associated with the model to obtain different results. In the *Core* part, the performance of all the features is compared with that of the random variable and the most important variables are determined through a hypothesis test. Furthermore, it proposes an automatic method to optimize the hyperparameter associated with the number of iterations of the first component while keeping fixed the threshold for the p-value.

It is noteworthy that all these algorithms presented include the Shapley values through the mean of the absolute values for each variable as a measure of global influence.

2.1.2 Dataset shift

The concept of dataset shift was introduced in Quiñonero-Candela et al. (2008), where it is described as a phenomenon in which the joint distribution of the explanatory variables and the target variable is different between the training set and the test set used in the creation of a statistical learning model. More formally, given a time period $[0, t]$, a set of samples denoted $S_{0,t} = \{d_0, \ldots, d_t\}$, where $d_i = (X_i, y_i)$ with X_i the vector of explanatory variables and y_i the target variable. Let $F_{0,t}(X, y)$ be the distribution following $S_{0,t}$, analogously denote $S_{t+1,\infty}$ and $F_{t+1,\infty}(X, y)$ for a time period $[t+1, \infty)$. A dataset shift is said to occur at time $t+1$ if $F_{0,t}(X, y) \neq F_{t+1,\infty}(X, y)$, i.e. $\exists\, t$ such that $P_t(X, y) \neq P_{t+1}(X, y)$ (Lu et al., 2018).

Independently of the model used, a typical data analysis scheme assumes that the distribution of the data is static for the model to be valid. If there is a variation in the distribution, that change must be modelled. Such changes are very common in real problems, such as economic, political, social, regulatory, etc., reasons that affect the behaviour of many phenomena. In particular, this situation occurs very often in the electricity market. This is the reason why the dataset shift problem must be addressed.

Let $t+1$ be the instant at which the dataset shift occurs, there are three possible reasons why the joint probability distribution is different (Moreno-Torres et al., 2012):

1. *Covariate shift* which is probably the most studied type of shift and happens when $P_t(y|X) = P_{t+1}(y|X)$ but $P_t(X) \neq P_{t+1}(X)$.
2. *Prior probability shift* or *label shift*, when $P_t(X|y) = P_{t+1}(X|y)$ but $P_t(y) \neq P_{t+1}(y)$
3. *Concept shift*, which occurs when the relationship between the explanatory and target variables change, that is, when $P_t(y|X) \neq P_{t+1}(y|X)$ but $P_t(X) = P_{t+1}(X)$ or when $P_t(X|y) \neq P_{t+1}(X|y)$ but $P_t(y) = P_{t+1}(y)$. It is the most complex type of shift. An example in a classification problem is shown in Figure 2.1.

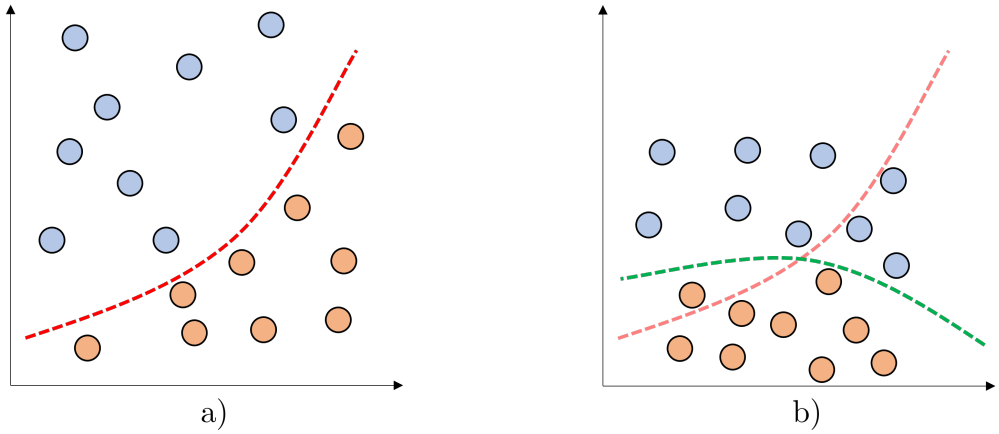

Figure 2.1: In **a)** the learned decision frontier is observed with respect to the training data. In **b)** it can be seen that the relationship between the variables has changed and that the decision frontier learned in training is not valid for the test data.

Much of the related research in this field focuses on shift detection (whether shift occurs or not),

on understanding why it occurs (when, how and where) and on shift adaptation (reacting to change). It is also often treated from the perspective of classification problems, while the field of regression has not been explored in great depth (Lima et al., 2022). Most strategies designed to react to the presence of the shift are based on retraining the models with more current data or with data similar to those occurring in the current context, although other strategies are possible (Lu et al., 2018). The implementation of such strategies can be complicated and requires constant maintenance. In addition, many algorithms require a large amount of data to perform satisfactorily, which severely limits the use of data related to the current paradigm.

In the particular case of a concept shift, where the relationship between the explanatory variables and the target variable changes, the situation where only one group of variables is responsible for the shift may arise. If the elimination of this group of variables still allows the construction of a good statistical model, the full use of the dataset and a much simpler model maintenance can be achieved, as the shift will no longer be present.

2.2 A new feature selection algorithm for regression problems

Generally, feature selection algorithms quantify the importance of a variable and, based on some criterion, determine a minimum threshold of importance to be considered. However, they do not take into account the effect that a variable has, i.e. a feature may have a large impact on the decisions a model makes, but it does not necessarily need to have the right influence. This should not be the case in a standard static situation where the learned behaviour directly corresponds to the actual behaviour, but under a concept shift setting this situation is quite common.

Here, a new variable selection algorithm that is robust to these situations is proposed. It is able to detect the features that cause the shift in case they are among the possible explanatory variables and, in situations where there is no shift, performs similar or better than the state of the art.

2.2.1 The intuitive idea

To achieve the goal of detecting the features that do not have the right influence on the predictions,the effect that each variable has on each prediction will be analyse at a more local level. This effect is going to be estimated using the Shapley value of the feature over the prediction. Then, depending on the error that is being made in each one of the observations (over or under predicted), a positive or negative effect is going to be consider. For instance, if an observation is under predicted, then a feature that decreases the prediction value is not ideal, because it is contributing to making that error higher. After that, the feature will be evaluated to determine whether it has more positive or negative effects. If the negative effects are higher, the feature will be assigned a “negative influence” and considered for removal in an iterative process.

At a high level, the algorithm classifies the observations based on whether their predictions result in under predicted, over predicted or well predicted. This categorization is achieved by

employing various quantiles across the distribution of prediction errors. Subsequently, both groups of wrongly predicted observations are analysed separately to investigate the impact of each feature on each group. In essence, the study focuses on identifying variables that contribute to the observed biases within each group.

As previously discussed, the consideration of Shapley values within feature selection methods is not novel; however, their use at a more local level and their association with prediction errors represents a distinctive approach.

Together with this local use of Shapley values, the groups of well classified over predicted and under predicted observations are crucial for the algorithm. Although quantiles were employed to establish these groups, it is important to note that this approach is not mandatory. There may exist other criteria that are equally suitable for creating this distinction among the observations. Alternative methods could be explored to effectively generate these groups, potentially providing additional insights for the analysis.

2.2.2 The detailed algorithm

The starting point of the algorithm is a given model and a set of training and validation data. All variables are considered and a backward selection strategy will be established, eliminating variables sequentially.

The first step is to train the model and obtain predictions on the validation set. Working individually with each of these predictions is not feasible, so three groups of observations are constructed based on the prediction error. For this purpose, two user-selected parameters are used, $q_{\text{low}}, q_{\text{high}} \in [0,1]$, which correspond to two quantiles.

Definition 1. Let $\mathbf{x}$ be the vector of explanatory variables of an observation $(\mathbf{x}, y)$ of the validation set, err(**x**,y) $= y - \widehat{y}(\mathbf{x})$ be the error of its prediction with the model considered, **err** be the vector of errors of all predictions, $Q_{\text{low}} = \text{Quantile}(\mathbf{err}, q_{\text{low}})$ and Q_{high} the analogue[3]. Let q^* be the quantile such that $\mathbb{P}(\mathbf{err} \leq 0) = q^*$, Q^* the value such that $\text{Quantile}(\mathbf{err}, q^*) = Q^*$ (note that Q^* is not necessarily 0). The following are defined as

$$Q^*_{\text{low}} = \begin{cases} Q_{\text{low}} & \text{if } 0 \in \left[Q_{\text{low}}, Q_{\text{high}}\right] \\ Q^* & \text{if } Q_{\text{high}} < 0 \\ Q_{\text{low}} - (Q_{\text{high}} - Q^*) & \text{if } Q_{\text{low}} > 0 \end{cases}$$

$$Q^*_{\text{high}} = \begin{cases} Q_{\text{high}} & \text{if } 0 \in \left[Q_{\text{low}}, Q_{\text{high}}\right] \\ Q_{\text{high}} - (Q_{\text{low}} - Q^*) & \text{if } Q_{\text{high}} < 0 \\ Q^* & \text{if } Q_{\text{low}} > 0 \end{cases}$$

It is said that:

- **x** is correctly predicted if err(**x**,y) $\in \left[Q^*_{\text{low}}, Q^*_{\text{high}}\right]$

[3]Lower case is used for the quantile and upper case for the value associated with the quantile.

- **x** is under predicted if err(**x**,y) $\in \left(Q^*_{\text{high}}, +\infty\right)$
- **x** is over predicted if err(**x**,y) $\in \left(-\infty, Q^*_{\text{low}}\right)$

In this way, it is considered that $\left[Q^*_{\text{low}}, Q^*_{\text{high}}\right]$ is the correctly predicted space of errors while the complementary is wrongly predicted.

Note that in the previous definition, in case that $0 \notin \left[Q_{\text{low}}, Q_{\text{high}}\right]$, the quantiles are simply being translated in case the model is highly biased. Thus, the bias of the model is included in the poorly predicted space but keeping the same width between boundaries as in the original proposal. In Figure 2.2 it is shown graphically over the distribution of the errors.

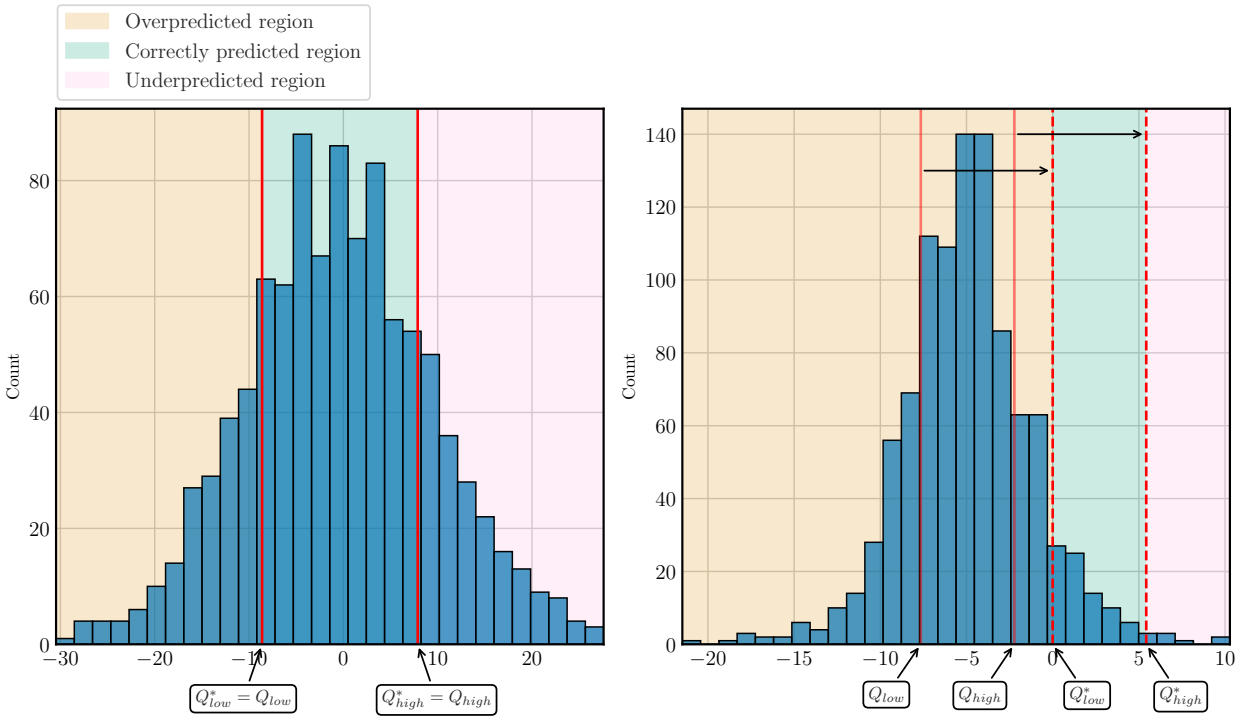

Figure 2.2: The left-hand side shows that no translation is necessary since, given the quantiles chosen, the model does not show any significant bias. On the right, the model tends to over predict, so this bias is penalized by shifting the quantiles to define the correctly predicted region.

Notation 1. Let $(\mathbf{x}, y)$ be an observation from the validation set, $\widehat{y}$ the model's prediction of that observation and SHAP_{var} a function that returns the Shapley value of a variable for an observation. The effect of a variable, var, on an observation is denoted as

$$\text{Effect}_{\text{var},\mathbf{x}} = \text{sgn}(\text{SHAP}_{\text{var}}(\mathbf{x}, y, \widehat{y})) \cdot \text{SHAP}_{\text{var}}(\mathbf{x}, y, \widehat{y})^2$$

Note that the Shapley value could actually be taken directly, but taking the square increases the effect of the most influential characteristics.

The effect of a variable on a group (correctly predicted, over predicted, under predicted) can be derived as:

$$\text{Effect}_{\text{var,group}} \equiv \text{Ef}_{\text{var,group}} = \sum_{\mathbf{x}\in\text{group}} \text{Effect}_{\text{var},\mathbf{x}}$$

Definition 2. Let var be a variable, **err** the vector of errors of the predictions on the validation set, $q_2(\mathbf{err})$ the median of the vector **err** and let C.P, O.P and U.P be the three groups described above. The negative influence of var, neg inf_{var} is defined as

$$\text{neg inf}_{\text{var}} = \begin{cases} +\infty & \text{if } \sum_{\text{group}} |\text{Ef}_{\text{var,group}}| = 0 \\ |\text{Ef}_{\text{var,O.P}}| - \left(|\text{Ef}_{\text{var,U.P}}| + |\text{Ef}_{\text{var,C.P}}|\right) & \text{if } q_2(\mathbf{err}) < 0,\ \text{Ef}_{\text{var,O.P}} > 0, \\ & \text{Ef}_{\text{var,U.P}} > 0, \\ & |\text{Ef}_{\text{var,O.P}}| > |\text{Ef}_{\text{var,U.P}}| + |\text{Ef}_{\text{var,C.P}}| \\ |\text{Ef}_{\text{var,U.P}}| - \left(|\text{Ef}_{\text{var,O.P}}| + |\text{Ef}_{\text{var,C.P}}|\right) & \text{if } q_2(\mathbf{err}) > 0,\ \text{Ef}_{\text{var,O.P}} > 0, \\ & \text{Ef}_{\text{var,U.P}} > 0, \\ & |\text{Ef}_{\text{var,U.P}}| > |\text{Ef}_{\text{var,O.P}}| + |\text{Ef}_{\text{var,C.P}}| \\ |\text{Ef}_{\text{var,U.P}}| + |\text{Ef}_{\text{var,O.P}}| - |\text{Ef}_{\text{var,C.P}}| & \text{if } \text{Ef}_{\text{var,O.P}} > 0,\ \text{Ef}_{\text{var,U.P}} < 0, \\ & |\text{Ef}_{\text{var,U.P}}| + |\text{Ef}_{\text{var,O.P}}| > |\text{Ef}_{\text{var,C.P}}| \\ 0 & \text{otherwise} \end{cases}$$

The general idea of the previous definition is to study the effect that each variable has on each group of observations. Each of the cases encapsulates the following corresponding idea:

1. If a variable has no effect on the predictions, its negative influence is defined as infinite, as a variable is being added to the model that does not contribute with any value and thus this variable is a candidate for overfitting.
2. In case the model is biased towards over predictions and the variable increases the value of over predictions (undesirable effect) and increases the value of under predictions (desirable effect), the negative influence is the difference of the absolute values of the undesirable minus the desirable effects. The correctly predicted group is always considered as a desired effect.
3. This is symmetrical to the previous case for the group of those under predicted.
4. In case the variable increases over predictions and decreases under predictions, the variable has an undesired effect for all but the correctly predicted ones.
5. In any other case, the variable has no negative influence.

Note the significance of segregating observations into different groups (C.P, O.P, and U.P). In each iteration, the model may be in a distinct bias position, consistently generating predictions

that are persistently higher or lower. The model's position, as determined through $q_2(\mathbf{err})$, influences whether a variable's effect can be negative or not. For instance, when $q_2(\mathbf{err}) < 0$, indicating a general trend of over predictions, a variable with a substantial effect on causing over predictions (large $|\mathrm{Ef}_{O.P}|$) becomes a candidate for having neg inf > 0. This indicates that the variable is one of the features contributing to the observed bias. On the contrary, when $q_2(\mathbf{err}) > 0$, indicating a tendency of under predictions, the same variable becomes a candidate to significantly aid the model, as $|\mathrm{Ef}_{O.P}|$ in this case has a positive effect that helps increase the value of predictions, thereby mitigating the current bias, and its neg inf will be 0.

With these concepts defined, the algorithm can be fully described (Algorithm 1)

Algorithm 1 New feature selection algorithm for regression problems

Input: $(\mathbf{X},\mathbf{y})_{\text{train}}$, $(\mathbf{X},\mathbf{y})_{\text{val}}$, model, n_iter_prev, q_{low}, q_{high}, metric
Output: Set of selected features

if n_iter_prev > 0 **then**
    Introduce a random variable to the training and validation sets
    for iteration in 1:n_iter_prev **do**
        Train a model with the new dataset
        Compute the global influence of each variable and store it for each iteration
    end for
    Calculate the average of the global influences of the previous n_iter_prev iterations
    Remove the random variable and the characteristics that have less overall influence than the random variable
end if
while len(features selected) > 0 **and** len(features to remove) > 0 **do**
    Train the model with the features selected up to now
    Compute Shapley values for all variables and predictions in the validation set
    Classify the observations of the validation set into correctly predicted, over predicted or under predicted
    Compute the effect of each variable on each observation of the validation set
    Compute the effect of each variable on each group of observations
    Establish the negative influence of each variable
    Compute the chosen metric on the validation set
    if There is a variable with infinite negative influence **then**
        Delete all variables with infinite negative influence
    else
        if There is a variable with a non-zero negative influence **then**
            Delete the variable with the greatest negative influence
        end if
    end if
end while

The algorithm is divided into two phases, a preliminary and optional phase and the main component. In the first preprocessing phase, a random variable is introduced into the dataset,

specifically, a permutation of the most influential variable according to the mean of the absolute values of the Shapley values between the original variables. Once included in the dataset, the global influence of each variable is recalculated a certain number of times specified by the user, where in each iteration the random seed that determines the learning process of the algorithm is modified to obtain different influences. At the end of the process, all those variables that have a global influence less than or equal to that of the new variable introduced are removed. Including this first phase seems to improve the results when the number of variables eliminated is not too large, since it deletes variables that only seem to introduce noise. In case a large number of variables are excluded in this phase, it is recommended not to apply it, since the importance or role of each variable in the decisions of the model may not be entirely clear.

The second component corresponds to the core part of the algorithm, where the effect of each variable is analysed when making predictions. This is a backward feature selection scheme, i.e., variables are sequentially removed until a criterion is no longer met. To eliminate a variable, first the predictions in the validation set are classified according to the error made and the q_{low} and q_{high} quantiles chosen (Definition 1). Then, according to the effect of each variable on each group (Notation 1), the negative influence is computed (Definition 2). If there are variables that have no effect on the predictions, that is, with infinite negative influence, in this iteration all those that meet this property are eliminated. If all the variables have some effect on the prediction, the variable with the greatest negative influence is eliminated. The process ends when there is no variable with non-zero negative influence or when there are no more features left. At the end of each iteration a metric (MAE, MSE, R^2, etc.) can be computed on the predictions of the validation set in an informative way, although it is not used to make decisions during the process. The feature set that obtained the best value of the metric is returned. Supposing that a value very similar to the best could be obtained with a notably smaller number of variables, which would also constitute a reasonable final feature set, the user could decide to consider it.

2.2.3 Remarks on the algorithm

As previously mentioned, the shift is not detected explicitly, but in the presence of a change with a negative influence on a variable, the proposed algorithm is able to detect its undesired effect and evaluate whether the preservation of the feature in the model is really beneficial.

Identifying variables that lead to changes in behaviour becomes challenging when dealing with correlated variables that jointly contribute to the observed change. The algorithm does not directly address this challenge due to the complexity involved in handling both groups of features and individualized variables automatically, especially without expert knowledge. However, the approach indirectly tackles this issue by selecting the final set of features based on the best metric from the validation set. When dealing with correlated features that may individually degrade the model but collectively enhance it, such scenarios can be identified by analysing increases in MAE variations between iterations in the validation set (as shown in Figure 2.9). Conversely, if correlated variables individually improve the model but collectively worsen it, these variables might be retained without considering the potential impact of different variable groups. To address this, employing a cross-validation strategy, commonly used in

Machine Learning tasks, can help detect and rectify such issues, facilitating the selection of a variable set expected to perform well on the test set. In any case, if the dependence between the features is a critical aspect, modifications can be adopted in the computation of the Shapley value that do take this fact into account. This way, a closer approximation of the true Shapley value is obtained while maintaining the good properties that it presents. See for example Aas et al. (2021).

Concerning the long-term performance of the algorithm, the main aspect to take into account, regardless of the number of concept shifts, is the amount of data available for a possible new behaviour in order to be able to make a proper comparison with the previous period. It should be noted that the proposed methodology makes use of a validation set that must contain a considerable part of the data corresponding to the new context. If this volume of data is not significant, the new context may go unnoticed and the algorithm will not behave correctly. If the behavioural changes of the phenomenon are very frequent, the methodology presented may not be adequate for the same reason, although part of the past data could be ignored for feature selection if the behaviour is clearly different. In any event, over extended periods of time, the algorithm should be run periodically to update the selection of variables that have a positive influence in the inference.

Regarding the computational complexity of the algorithm, four factors can influence it: the complexity associated with model training, the complexity of evaluating new observations, the efficiency in estimating the Shapley value, and the number of considered features. While these factors pose no issues for algorithms based on decision trees, linear regression models, or generalized linear models, they may make the algorithm impractical for neural networks. To analyse this case, consider, for example, the algorithm of Ancona et al. (2019). It gives an approximation of the Shapley values in $\mathcal{O}(M^2)$ network evaluations, where M is the number of features, which is usually dominated by the computational complexity associated to the network training. When dealing with complex tasks requiring networks with numerous parameters, conducting all the mentioned phases in each iteration of the algorithm can become infeasible, despite each step being computationally viable in polynomial time. Therefore, this limitation should be taken into consideration.

Considering that the model is given and thus one can consider its evaluation and training costs fixed. The complexity of the algorithm depends only on the number of features and the size of the dataset. If these are very large and time is a critical aspect of the task at hand, the following strategies can be followed to decrease the computing time. In the case that there is a big number of possible explanatory features one can handle a number of features in each iteration with the following procedure:

1. Instead of removing variables one by one, it is possible to remove them in blocks. That is, you can set a constant number of features or a percentage of them (which would limit, in part, the maximum number of iterations of the algorithm) to be deleted in each iteration. Thus, in the corresponding iteration, the determined M features with the greatest negative influence would be eliminated.

2. The strategy described above can be adapted progressively as the algorithm evolves. For example, when working with algorithms that make use of a learning rate, it is possible to

start with a high learning rate and, as the iterations go by, progressively reduce the learning rate until a fine-grained optimization is performed in the final iterations. This same idea can be adopted to the presented methodology: one can start with a high number of variables to be removed in the first iterations and this number is decreased over the iterations until they are again removed one at a time as in the original methodology. Naturally, how many features are allowed to be deleted in the first iteration, and how this number varies, depends entirely on the problem being addressed.

In case the computation time is related to a large number of samples, two aspects should be taken into account:

1. Depending on the model used, the training time can be long and costly. The user should evaluate whether all the data is necessary to build the final model. However, it is usual that performance improves as more data is available, at least if it belongs to the same behavioural context.
2. The estimation time of the Shapley value can be reduced by using a smaller number of samples. Castro et al. (2009) shows that from certain orders of magnitude the approximation of the Shapley value is practically equivalent, so that for the feature selection task to be performed, using all the samples is probably not necessary.

Outliers, noise in the data or different transformations such as data standardization or encoding of categorical variables should not be of concern over the feature selection process as such. The different strategies followed to make a correct modelling of the process should be maintained, because if performed correctly, they should improve the performance of the algorithm. By varying any part of the base learning algorithm in a broad sense, including the data transformation, the selection of features could change according to the modelling that has been performed, as there is the possibility of varying the Shapley values. However, if the transformations were adequate, there will be a better detection of those features with real negative influence, because the model will be more appropriate. Even so, if proper modelling has not been followed (e.g., outliers with a possible crucial effect have not been eliminated or mitigated), this may have an impact on the removed variables and the remaining ones may not be an optimal set. There are no measures designed and implemented in the algorithm to tackle this kind of problems, since it is considered that these cases must be resolved beforehand.

Let stress why this algorithm is designed to be robust to concept shift situations. The variable importance is not the only metric that is taken into account. In fact, the possible bias generated by a variable is the main focus of the algorithm. If a variable is constantly generating a bias in the model predictions, which is common in concept shift contexts, then that variable is actually discarded from the possible explanatory variables, even if it has a great impact on the model. It is crucial to emphasize that, similar to constant model retraining in real-world applications to adapt to changing conditions, the reintegration of, at least, these important features (based on a global metric like the mean of the absolute Shapley values across all observations) should be assessed. If the feature genuinely influences the target variable, even in the presence of concept shifts affecting that explanatory variable, there is a possibility that, with a substantial amount of new collected data related to the shift, it could once again have a positive effect on the predictions and, therefore, in model performance. In such instances, reintroducing it into

the model feature set should be considered. This type of mechanisms are related to a successful long-term performance of the algorithm. As previously mentioned, periodically re-running the algorithm is one of the best practices to check whether the positive and negative influence of the variables remain unchanged.

2.3 Experiments

To analyse the effectiveness of the algorithm, the proposed method was compared with other feature selection algorithms using different configurations. Specifically, the quantile configurations (0.25, 0.75), (0.2, 0.8), (0.15, 0.85), (0.1, 0.9), (0.05, 0.95), which correspond to considering from 50% of poorly predicted observations in the development of the algorithm to 10%, were analysed. In addition, 30 iterations of the preprocessing phase were applied and the MAE is used to select the final set of features.

The other algorithms evaluated are considered as references in the variable selection methods: Boruta, PIMP and LASSO as traditional methods and Boruta-Shap, Shapicant and Powershap based on Shapley values. In case they use validation and training sets to make the variable selection, the same are introduced in all cases, if only one dataset is used, validation and training are introduced as one. In the particular case of the LASSO regularization, the variables are previously normalized by the maximum and minimum, and four different values of the regularization constant are considered: 0.01; 0.001; 0.0001; 0.00001.

Following Verhaeghe et al. (2022), a CatBoost estimator (Prokhorenkova et al., 2018) with 250 iterations is applied on all datasets and on all variable selection algorithms, except LASSO regression to select features. On the test set, The MAE, RMSE, and R^2 are analysed over 50 different runs, with varying seeds to obtain different results and assess the stability of the outcomes with the selected variables. In contrast to the Verhaeghe's methodology, the 50 seeds are fixed in advance and the order of the variables is always entered alphabetically, otherwise different results could be obtained with the same set of features[4].

Two different situations are analysed: selection of variables with the presence of concept shift and selection of variables in standard situations.

2.3.1 Concept shift

Synthetic experiments

Studying the method on fully controlled toy datasets is considered necessary to test its effectiveness. To this end, two of the most typical dataset shift situations have been recreated: a sudden shift and an incremental shift (Lu et al., 2018), represented in Figure 2.3.

The following objective function will be fitted:

[4]To ensure the veracity and reproducibility of the results, the different databases, the algorithm code and the results can be found at https://github.com/CCaribe9/SHAPEffects

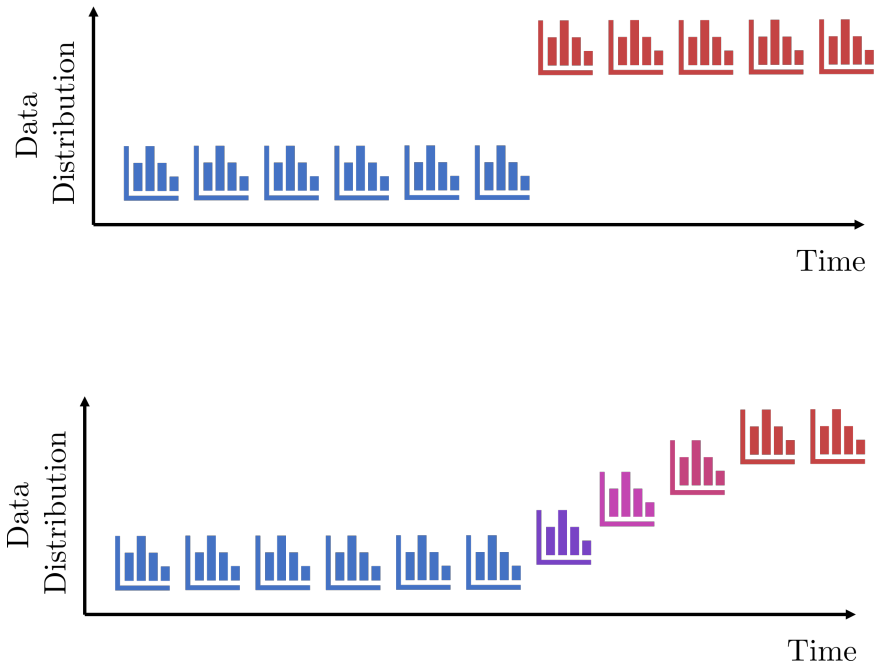

Figure 2.3: **(Top)** is a sudden shift situation. **(Bottom)** is an incremental shift situation.

$$
\begin{aligned}
f(\mathbf{x}_t) = {} & 2x_{1,t} + \lambda_1 x_{2,t}^2 + 3\sin(2\pi x_{3,t}) - 0.4x_{4,t} + \lambda_2 x_{5,t}^2 \\
& + 2x_{1,t-1} + \lambda_1 x_{2,t-1}^2 + 3\sin(2\pi x_{3,t-1}) - 0.4x_{4,t-1} + \lambda_2 x_{5,t-1}^2 \\
& + \varepsilon_t
\end{aligned}
$$

where $\mathbf{x}_t = (x_{1,t}, \ldots, x_{5,t}, \ldots, x_{10,t}) \in \mathbb{R}^{10}, \lambda_1, \lambda_2 \in \mathbb{R}$. Note that only the first five variables are informative. The scalars λ_1 and λ_2 are employed to generate diverse shift scenarios, while ε_t represents noise. This function has been chosen in a way that linearities and nonlinearites are both included. For nonlinearities standard simple functions has been selected. An autoregressive component has been considered, aligning with typical time series problems. Different coefficients have been chosen to visualize how the algorithm performs with various scales of the features.

30000 samples of $x_i \sim U(0,1)$, $i = 1, \ldots, 10$ are generated. The noise is modelled as a $N(0, 0.01)$. The first lag of the objective variable is also included as a possible explanatory variable .

Let $\lambda_1^a \in \{-10, -1, -0.1\}$, $\lambda_1^b \in \{-4, -0.4, -0.04\}$, $\lambda_2^a \in \{10, 1, 0.1\}$, $\lambda_2^b \in \{-25, -2.5, -0.25\}$. In order to recreate a sudden concept shift, it is defined

$$
\lambda_1 = \begin{cases} \lambda_1^a, & \text{for the first 20000 samples} \\ \lambda_1^b, & \text{for the last 10000 samples} \end{cases}
$$

$$
\lambda_2 = \begin{cases} \lambda_2^a, & \text{for the first 20000 samples} \\ \lambda_2^b, & \text{for the last 10000 samples} \end{cases}
$$

To recreate an incremental concept shift situation, if the sample number is called index, it is defined:

$$\lambda_1 = \begin{cases} \lambda_1^a, & \text{for the first 20000 samples} \\ \dfrac{(\lambda_1^b - \lambda_1^a)(\text{index} - 20000) + 10000\lambda_1^a}{10000}, & \text{if index } \in (20000, 25000) \\ \lambda_1^b, & \text{for the last 5000 samples} \end{cases}$$

$$\lambda_2 = \begin{cases} \lambda_2^a, & \text{for the first 20000 samples} \\ \dfrac{(\lambda_2^b - \lambda_2^a)(\text{index} - 20000) + 10000\lambda_2^a}{10000}, & \text{if index } \in (20000, 25000) \\ \lambda_2^b, & \text{for the last 5000 samples} \end{cases}$$

Every combination of $\lambda_1^a, \lambda_1^b, \lambda_2^a, \lambda_2^b$ is taken into account, resulting in a total of 81 potential scenarios for both types of shifts. This aspect holds particular interest as it enables the analysis of the system's behaviour when the impacted variables exhibit varying degrees of significance.

For both cases, the first 20000 samples are used for training, the next 5000 for validation and the last 5000 as test. The different sets are considered following a temporal order, as in time series problems. The shift occurs inside the validation set, which is the first moment in which the algorithms could detect the change of behaviour. The sudden shift is considered at the start of the validation set, so there is bigger contrast between the sudden shift and the incremental shift situation, which happens throughout the entire validation set. The difference between the mean MAE/RMSE/R^2 of the proposed algorithm with every other method is computed, although the standard deviation, the minimum value and the maximum value are also measured.

Results The proposed method is called SHAPEffects. The results are graphically described by the histogram of each one of these differences in Figures 2.4 and 2.5.

Only MAE outcomes are presented, as the results for RMSE and R^2 are equivalent. The obtained results are highly encouraging, given that the difference is predominantly negative (smaller MAE). In fact, when that difference is positive, it is practically negligible. These instances arise when the importance of one or both variables subjected to the shift is almost insignificant, indicated by comparable small coefficient values. The analysis reveals that the disparity between the two types of shifts is minimal. Moreover, the proposed algorithm consistently achieves superior results compared to other methods. The marginal discrepancies observed between the two shift types arise when the influence of the variables is relatively limited. In the case of incremental shifts, where a gradual change occurs, the proposed algorithm may not find beneficial to eliminate some of the variables that are undergoing the shift in the validation data, as at the beginning of the shift the variables may be contributing correctly to the outcomes. This behaviour aligns with the approach followed by the other algorithms, contributing to the small differences observed (see Tables 2.5 and 2.6).

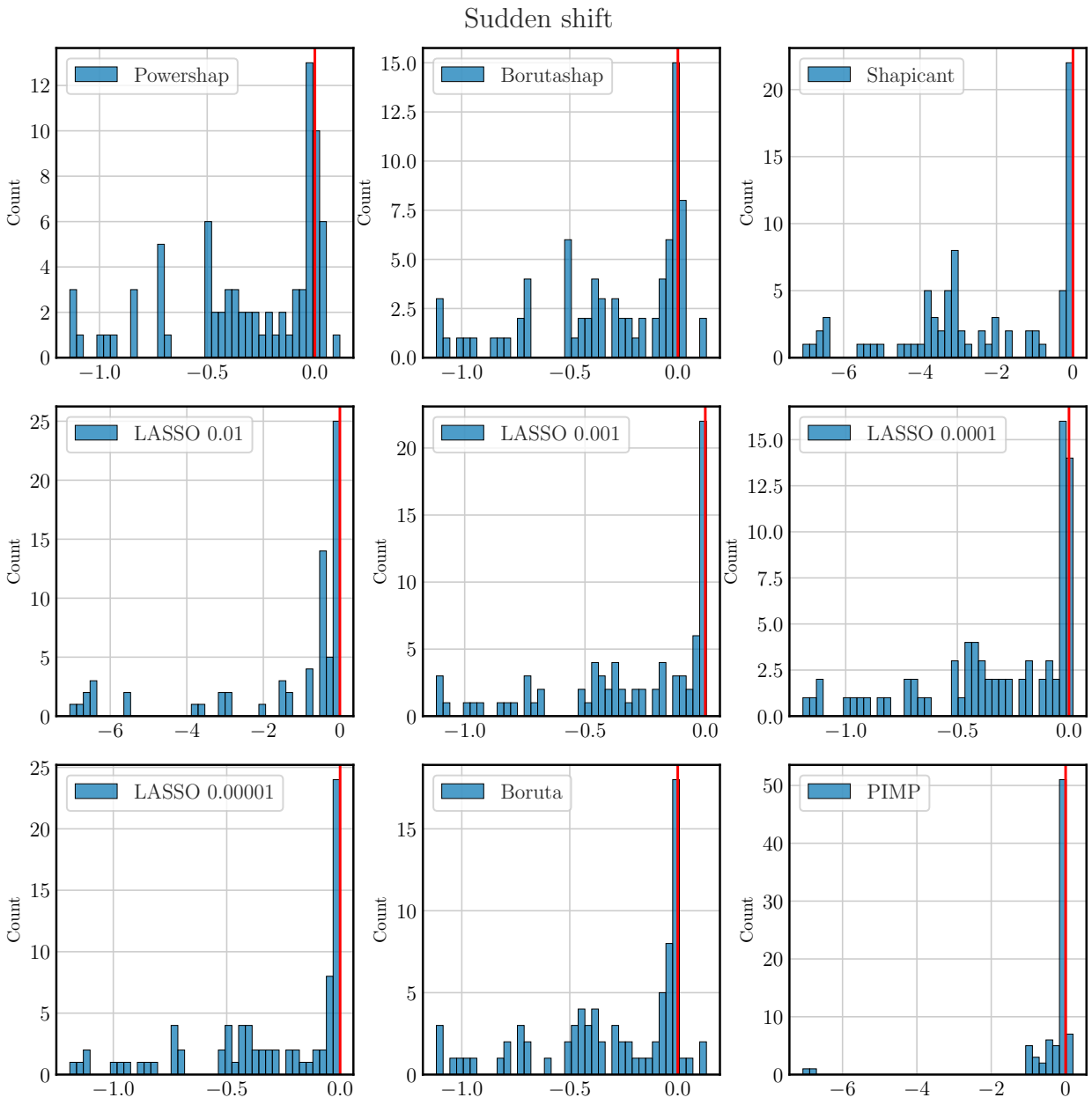

Figure 2.4: Histograms (for the 81 cases) of the difference of the mean MAE of the proposed method with every other algorithm for the sudden shift case

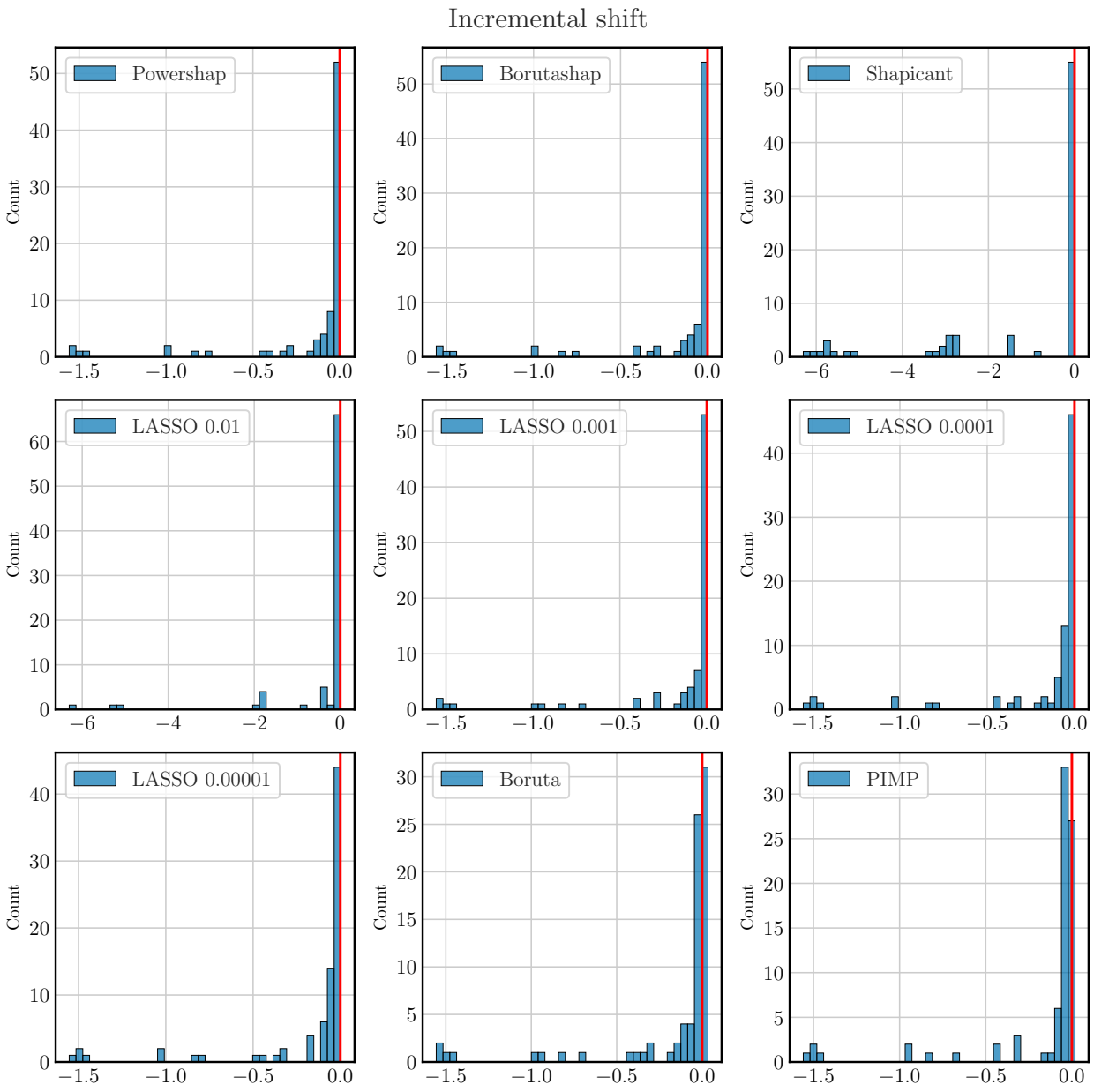

Figure 2.5: Histograms (for the 81 cases) of the difference of the mean MAE of the proposed method with every other algorithm for the incremental shift case

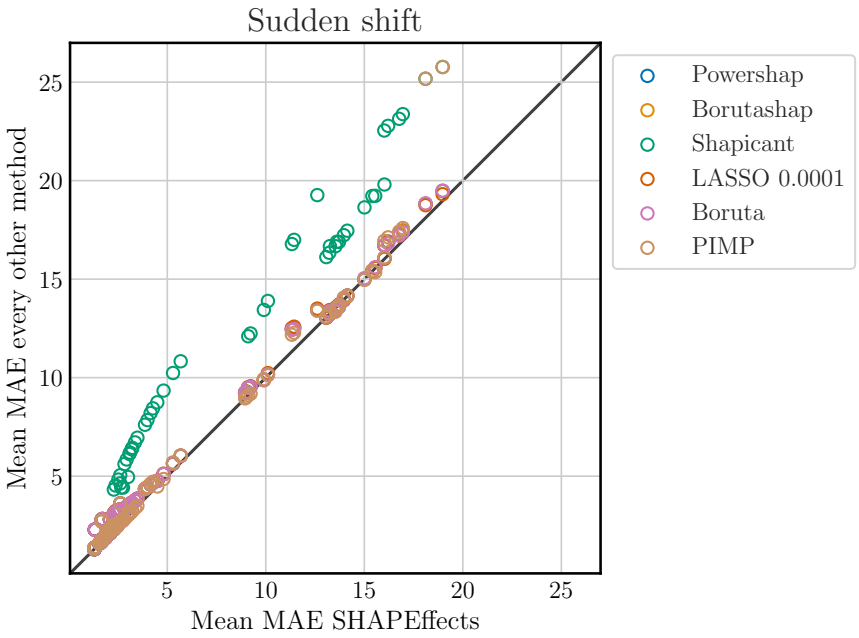

Figure 2.6: Mean MAE of the best SHAPEffects configuration vs mean MAE of every other method for the sudden shift case

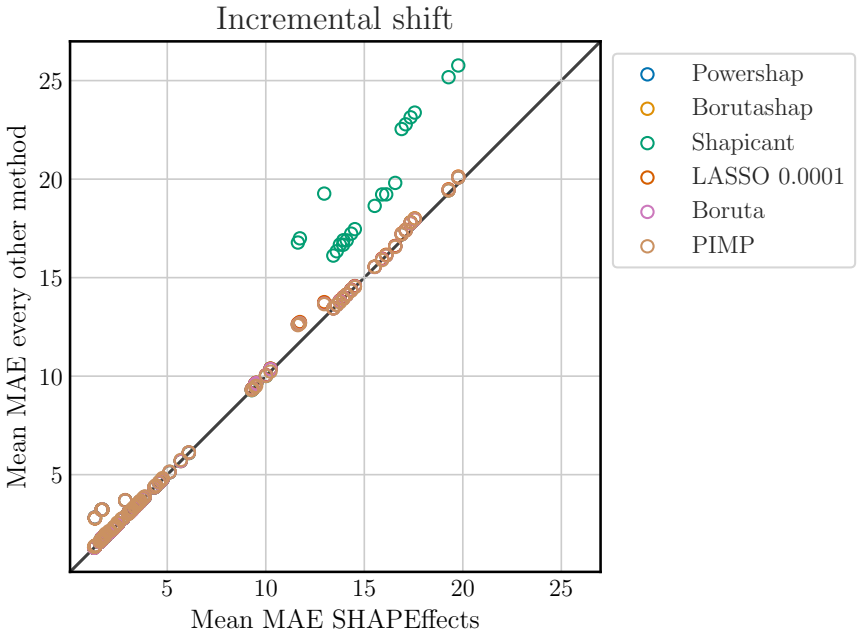

Figure 2.7: Mean MAE of the best SHAPEffects configuration vs mean MAE of every other method for the incremental shift case

For a more individualized evaluation rather than a collective comparison, the mean MAE of the optimal SHAPEffects configuration is compared with the mean MAE of each alternative method,

and subsequently visualized on a scatter plot (Figures 2.6 and 2.7). Similarly, in the case of the LASSO, the best configuration is selected, which from the previous figures is $\lambda = 0.0001$.

It can be seen that when the MAE is low (when the coefficients of the changing variables are very small) all methods exhibit similar performance, as previously stated out. However, as the error increases, which corresponds to a greater increase in the influence of the variables that undergo the change in behaviour, there are situations in which the MAE is notably greater than that obtained by the proposal for all the algorithms, which is evident when observing that no method exceeds the line $y = x$. These behaviours are more clear in the case of the sudden shift, although they are also visible in the incremental case.

It is important to note that the cases in which the proposed algorithm performs notably better than the other algorithms (when the points are separated from the diagonal), is when one of the two variables undergoing the change is removed. Specifically, when x_5 is eliminated, which is the most relevant one. The rest of the algorithms do not drop these variables when they take on a notable importance (Tables 2.1, 2.2, 2.3 and 2.4).

Furthermore, a detailed analysis is conducted on what are considered to be the three most representative cases:

1. $\lambda_1^a = -10, \lambda_1^b = -4, \lambda_2^a = 10, \lambda_2^b = -25$ (Table 2.1 and 2.2)
2. $\lambda_1^a = -1, \lambda_1^b = -0.4, \lambda_2^a = 1, \lambda_2^b = -2.5$ (Table 2.3 and 2.4)
3. $\lambda_1^a = -0.1, \lambda_1^b = -0.04, \lambda_2^a = 0.1, \lambda_2^b = -0.25$ (Table 2.5 and 2.6)

Algorithm	**MAE**				**RMSE**				**R^2**			
	Mean	**Std**	**Max**	**Min**	**Mean**	**Std**	**Max**	**Min**	**Mean**	**Std**	**Max**	**Min**
Powershap	13.44	3.99e-02	13.54	13.31	17.18	4.90e-02	17.30	17.05	-1.41	1.38e-02	-1.38	-1.45
Boruta-Shap	13.41	2.70e-02	13.47	13.36	17.14	3.95e-02	17.24	17.06	-1.40	1.11e-02	-1.38	-1.43
Shapicant	19.27	3.94e-03	19.27	19.26	22.30	1.02e-02	22.32	22.27	-3.07	3.73e-03	-3.06	-3.07
Boruta	13.41	2.70e-02	13.47	13.36	17.14	3.95e-02	17.24	17.06	-1.40	1.11e-02	-1.38	-1.43
PIMP	13.39	2.84e-02	13.46	13.33	17.11	3.64e-02	17.21	17.02	-1.39	1.02e-02	-1.37	-1.42
Best Lasso (0.001)	13.41	2.70e-02	13.47	13.36	17.14	0.04	17.24	17.06	-1.40	1.11e-02	-1.38	-1.43
SHAPEffects (0.25-075)	12.89	4.64e-02	12.97	12.78	15.89	5.14e-02	15.98	15.76	-1.06	1.33e-02	-1.03	-1.09
SHAPEffects (0.2-08)	12.73	4.00e-02	12.80	12.63	15.73	4.41e-02	15.81	15.62	-1.02	1.13e-02	-0.99	-1.04
SHAPEffects (0.15-085)	12.73	4.00e-02	12.80	12.63	15.73	4.41e-02	15.81	15.62	-1.02	1.13e-02	-0.99	-1.04
SHAPEffects (0.1-0.9)	**12.61**	2.90e-02	**12.68**	**12.52**	**15.60**	3.30e-02	**15.67**	**15.49**	**-0.99**	8.41e-03	**-0.96**	**-1.01**
SHAPEffects (0.05-095)	12.80	4.63e-02	12.91	12.69	15.82	5.07e-02	15.94	15.69	-1.05	1.31e-02	-1.01	-1.08

Table 2.1: Test results on the first case for the sudden shift context

Algorithm	MAE				RMSE				R^2			
	Mean	**Std**	**Max**	**Min**	**Mean**	**Std**	**Max**	**Min**	**Mean**	**Std**	**Max**	**Min**
Powershap	13.71	1.96e-02	13.75	13.66	17.52	0.03	17.58	17.48	-1.51	7.25e-03	-1.50	-1.53
Boruta-Shap	13.71	1.96e-02	13.75	13.66	17.52	0.03	17.58	17.48	-1.51	7.25e-03	-1.50	-1.53
Shapicant	19.27	4.07e-03	19.27	19.26	22.30	0.01	22.32	22.27	-3.07	3.83e-03	-3.06	-3.07
Boruta	13.68	2.07e-02	13.71	13.61	17.48	0.03	17.53	17.38	-1.50	7.83e-03	-1.47	-1.51
PIMP	13.65	1.84e-02	13.69	13.61	17.44	0.03	17.49	17.38	-1.49	7.24e-03	-1.47	-1.50
Best Lasso (0.001)	13.68	2.07e-02	13.71	13.61	17.48	0.03	17.53	17.38	-1.50	7.83e-03	-1.47	-1.51
SHAPEffects (0.25-075)	**12.97**	2.54e-02	**13.04**	**12.91**	**15.98**	0.03	**16.06**	**15.92**	**-1.09**	7.43e-03	**-1.07**	**-1.11**
SHAPEffects (0.2-08)	**12.97**	2.54e-02	**13.04**	**12.91**	**15.98**	0.03	**16.06**	**15.92**	**-1.09**	7.43e-03	**-1.07**	**-1.11**
SHAPEffects (0.15-085)	**12.97**	2.54e-02	**13.04**	**12.91**	**15.98**	0.03	**16.06**	**15.92**	**-1.09**	7.43e-03	**-1.07**	**-1.11**
SHAPEffects (0.1-0.9)	**12.97**	2.54e-02	**13.04**	**12.91**	**15.98**	0.03	**16.06**	**15.92**	**-1.09**	7.43e-03	**-1.07**	**-1.11**
SHAPEffects (0.05-095)	**12.97**	2.54e-02	**13.04**	**12.91**	**15.98**	0.03	**16.06**	**15.92**	**-1.09**	7.43e-03	**-1.07**	**-1.11**

Table 2.2: Test results on the first case for the incremental shift context

Algorithm	MAE				RMSE				R^2			
	Mean	**Std**	**Max**	**Min**	**Mean**	**Std**	**Max**	**Min**	**Mean**	**Std**	**Max**	**Min**
Powershap	1.79	2.63e-03	1.79	1.78	2.23	2.79e-03	2.23	2.22	0.53	1.18e-03	0.53	0.53
Boruta-Shap	1.79	2.63e-03	1.79	1.78	2.23	2.79e-03	2.23	2.22	0.53	1.18e-03	0.53	0.53
Shapicant	1.79	2.43e-03	1.80	1.78	2.23	2.33e-03	2.24	2.22	0.53	9.90e-04	0.53	0.52
Boruta	1.79	2.43e-03	1.80	1.78	2.23	2.33e-03	2.24	2.22	0.53	9.90e-04	0.53	0.52
PIMP	1.74	1.55e-03	1.74	1.73	2.16	1.60e-03	2.16	2.15	0.56	6.59e-04	0.56	0.55
Best Lasso (0.001)	1.79	2.63e-03	1.79	1.78	2.23	2.79e-03	2.23	2.22	0.53	1.18e-03	0.53	0.53
SHAPEffects (0.25-075)	**1.70**	2.60e-03	**1.71**	**1.69**	**2.11**	2.70e-03	**2.11**	**2.10**	**0.58**	1.08e-03	**0.58**	**0.58**
SHAPEffects (0.2-08)	**1.70**	2.25e-03	**1.70**	**1.69**	**2.10**	2.65e-03	**2.11**	**2.10**	**0.58**	1.06e-03	**0.58**	**0.58**
SHAPEffects (0.15-085)	**1.70**	2.56e-03	**1.70**	**1.69**	**2.10**	2.76e-03	**2.11**	**2.09**	**0.58**	1.10e-03	**0.58**	**0.58**
SHAPEffects (0.1-0.9)	**1.70**	2.83e-03	**1.70**	**1.69**	**2.11**	3.08e-03	**2.11**	**2.10**	**0.58**	1.23e-03	**0.58**	**0.58**
SHAPEffects (0.05-095)	**1.70**	2.56e-03	**1.70**	**1.69**	**2.10**	2.76e-03	**2.11**	**2.09**	**0.58**	1.10e-03	**0.58**	**0.58**

Table 2.3: Test results on the second case for the sudden shift context

Algorithm	MAE				RMSE				R^2			
	Mean	**Std**	**Max**	**Min**	**Mean**	**Std**	**Max**	**Min**	**Mean**	**Std**	**Max**	**Min**
Powershap	1.79	4.44e-03	1.80	1.78	2.23	3.76e-03	2.24	2.22	0.53	1.60e-03	0.53	0.52
Boruta-Shap	1.79	4.44e-03	1.80	1.78	2.23	3.76e-03	2.24	2.22	0.53	1.60e-03	0.53	0.52
Shapicant	1.79	4.10e-03	1.80	1.78	2.23	3.04e-03	2.24	2.23	0.53	1.29e-03	0.53	0.52
Boruta	1.79	4.10e-03	1.80	1.78	2.23	3.04e-03	2.24	2.23	0.53	1.29e-03	0.53	0.52
PIMP	1.74	1.49e-03	1.74	1.73	2.16	1.39e-03	2.16	2.15	0.56	5.71e-04	0.56	0.56
Best Lasso (0.001)	1.79	4.44e-03	1.80	1.78	2.23	3.76e-03	2.24	2.22	0.53	1.60e-03	0.53	0.52
SHAPEffects (0.25-075)	**1.70**	2.92e-03	1.71	**1.69**	2.11	2.87e-03	**2.11**	2.10	**0.58**	1.15e-03	**0.58**	**0.58**
SHAPEffects (0.2-08)	**1.70**	3.67e-03	1.71	**1.69**	**2.10**	3.27e-03	**2.11**	2.10	**0.58**	1.31e-03	**0.58**	0.57
SHAPEffects (0.15-085)	**1.70**	2.34e-03	**1.70**	**1.69**	**2.10**	2.35e-03	**2.11**	**2.09**	**0.58**	9.40e-04	**0.58**	**0.58**
SHAPEffects (0.1-0.9)	**1.70**	2.34e-03	**1.70**	**1.69**	**2.10**	2.35e-03	**2.11**	**2.09**	**0.58**	9.40e-04	**0.58**	**0.58**
SHAPEffects (0.05-095)	1.79	4.44e-03	1.80	1.78	2.23	3.76e-03	2.24	2.22	0.53	1.60e-03	0.53	0.52

Table 2.4: Test results on the second case for the incremental shift context

Algorithm	MAE				RMSE				R^2			
	Mean	**Std**	**Max**	**Min**	**Mean**	**Std**	**Max**	**Min**	**Mean**	**Std**	**Max**	**Min**
Powershap	**1.289**	0.002	**1.293**	**1.285**	**1.583**	0.002	**1.586**	**1.578**	**0.733**	6.503e-04	**0.734**	**0.732**
Boruta-Shap	**1.289**	0.002	**1.293**	**1.285**	**1.583**	0.002	**1.586**	**1.578**	**0.733**	6.503e-04	**0.734**	**0.732**
Shapicant	1.291	0.002	1.295	1.287	1.587	0.002	1.591	1.583	0.731	6.189e-04	0.733	0.730
Boruta	1.291	0.002	1.295	1.287	1.587	0.002	1.591	1.583	0.731	6.189e-04	0.733	0.730
PIMP	1.351	0.002	1.357	1.347	1.669	0.002	1.673	1.666	0.703	6.029e-04	0.704	0.702
Best Lasso (0.001)	**1.289**	0.002	**1.293**	**1.285**	**1.583**	0.002	**1.586**	**1.578**	**0.733**	6.503e-04	**0.734**	**0.732**
SHAPEffects (0.25-075)	1.294	0.002	1.300	1.290	1.591	0.002	1.595	1.587	0.730	6.990e-04	0.731	0.728
SHAPEffects (0.2-08)	1.295	0.002	1.301	1.291	1.593	0.002	1.598	1.589	0.729	7.447e-04	0.731	0.727
SHAPEffects (0.15-085)	1.291	0.002	1.295	1.287	1.587	0.002	1.591	1.583	0.731	6.189e-04	0.733	0.730
SHAPEffects (0.1-0.9)	1.294	0.002	1.298	1.290	1.589	0.002	1.593	1.585	0.731	7.544e-04	0.732	0.729
SHAPEffects (0.05-095)	1.293	0.002	1.297	1.286	1.587	0.002	1.591	1.581	0.731	6.873e-04	0.733	0.730

Table 2.5: Test results on the third case for the sudden shift context

Algorithm	MAE				RMSE				R^2			
	Mean	**Std**	**Max**	**Min**	**Mean**	**Std**	**Max**	**Min**	**Mean**	**Std**	**Max**	**Min**
Powershap	1.291	0.002	1.295	1.286	1.585	0.002	1.589	1.581	0.732	7.101e-04	0.733	0.730
Boruta-Shap	**1.289**	0.002	**1.294**	**1.284**	**1.583**	0.002	**1.588**	**1.578**	**0.733**	7.179e-04	**0.734**	**0.731**
Shapicant	1.291	0.002	1.295	1.288	1.587	0.002	1.592	1.584	0.731	6.103e-04	0.732	0.730
Boruta	1.291	0.002	1.295	1.288	1.587	0.002	1.592	1.584	0.731	6.103e-04	0.732	0.730
PIMP	1.351	0.002	1.355	1.347	1.669	0.002	1.672	1.665	0.703	5.940e-04	0.704	0.702
Best Lasso (0.001)	1.289	0.002	**1.294**	**1.284**	**1.583**	0.002	**1.588**	**1.578**	**0.733**	7.179e-04	**0.734**	**0.731**
SHAPEffects (0.25-075)	1.293	0.002	1.298	1.287	1.588	0.002	1.594	1.582	0.731	7.552e-04	0.733	0.729
SHAPEffects (0.2-08)	1.294	0.002	1.299	1.290	1.591	0.002	1.596	1.587	0.730	6.544e-04	0.731	0.728
SHAPEffects (0.15-085)	**1.289**	0.002	**1.294**	**1.284**	**1.583**	0.002	**1.588**	**1.578**	**0.733**	7.179e-04	**0.734**	**0.731**
SHAPEffects (0.1-0.9)	1.294	0.002	1.300	1.290	1.589	0.003	1.595	1.582	0.730	8.828e-04	0.733	0.728
SHAPEffects (0.05-095)	**1.289**	0.002	**1.294**	**1.284**	**1.583**	0.002	**1.588**	**1.578**	**0.733**	7.179e-04	**0.734**	**0.731**

Table 2.6: Test results on the third case for the incremental shift context

In scenarios characterized by substantial or moderate influence of the variables, the proposed methodology demonstrates a clear advantage, yielding superior results across all configurations. Conversely, when the influence of the variables is minimal, the algorithms behave similarly. Additionally, no significant difference is observed between the results in sudden shift cases and incremental shift cases, highlighting the robustness of the algorithm.

In this synthetic example, only a concept shift affecting the entire validation set has been examined. It is worth noting that if the concept shift were to be incorporated into the training set, the results across all methodologies could potentially improve, depending on the model's ability to effectively learn from this shift. Conversely, if the concept shift were to occur later and later in the validation data, at a certain point our algorithm would no longer eliminate the variables, resulting in comparable outcomes with the other methods, in a similar way to what happened in the previous examples with the incremental concept shift and small coefficient.

Electricity Price Forecasting

The electricity market is constantly subject to regulatory changes and is highly influenced by political and economic situations. Specifically, in the Iberian market, given the crisis related to the gas price in mid-2022, a regulatory measure related to the gas price subvention to thermal power plants was established on June 15th, 2022 (BOE, 2022), changing the dynamics of the electricity market. Predictive models typically use the price of different fuels as regressors (as discussed in Section 1.5), among them the price of gas (Shiri et al., 2015; Ortiz et al., 2016; Marcjasz et al., 2023), so this change of behaviour has created a situation of uncertainty around the models. Figure 2.8 show the relationship between the electricity prices and the natural gas price. Moreover, it is possible to see the change in behaviour from 15th June onwards: there are peaks in the gas price that do not correspond to peaks in the electricity price. At least not as

was the case prior to this date. It can also be seen that at the time of the introduction of the regulatory change, the relationship between the variables changes drastically, as can be seen through the calculation of the linear correlation[5]. This problem is interesting for two reasons: firstly, it is a situation that occurs in the electricity market and therefore is part of the problems of interest. But also, from a generic point of view on the study of behavioural change situations, this is a case in which it is known that concept shift occurs and when it occurs, data that is not usually available in real problems. Thus, it is possible to work with real data, while at the same time having a high degree of control over the problem.

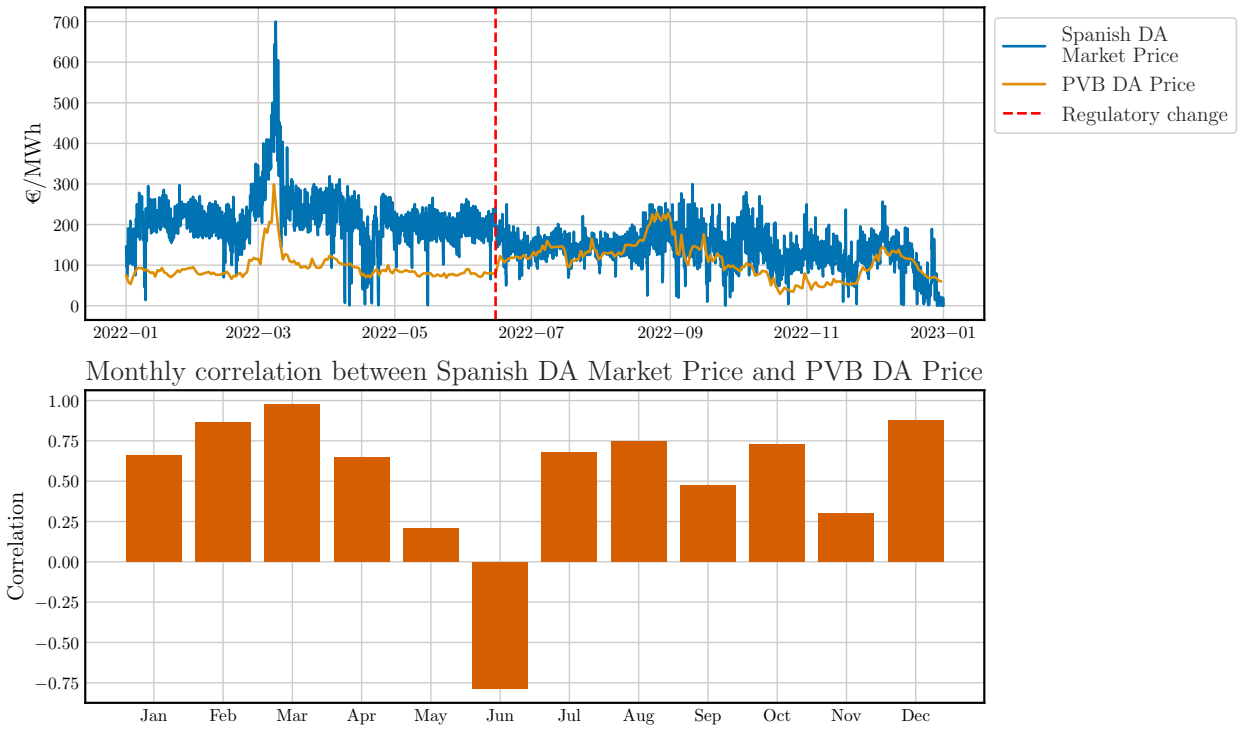

Figure 2.8: **(Top)** Evolution of Spanish electricity and gas prices during the year 2022. The instant of the regulatory change is indicated in red. **(Bottom)** Linear correlation calculated on a month-to-month basis between the daily series of average electricity prices and the gas price.

The question that arises in this case is whether keeping variables related with the gas price in the models is useful to predict the current situation. A dataset with 335 explanatory variables linked to the electricity market, several connected to the gas price, is available to predict the price of the DA Market. The data considered is publicly available, mostly obtained from the ESIOS portal, which is an information system developed by Red Eléctrica de España, and the data related to the gas price is obtained from the Iberian Gas Market (MIBGAS). The main variables considered are forecasts related to renewable energy generation, load forecasts, prices from the different markets, prices of previous days of the DA Market in France and lagged values of the series to be predicted. To describe the current behaviour of the market with respect to each of the characteristics, statistics of some of these same variables during the last week are

[5]It should be noted that a significant change in correlation does mark a significant change in the relationship between the two, but no change in correlation does not imply that there is no change in behaviour.

also included. These are the mean, standard deviation, median, minimum, maximum, first quartile, third quartile, skewness and kurtosis. The data start in January 2021 and end in August 2022, both included. The data is divided into: January 1st, 2021 to May 31st, 2022 for training, June 1st, 2022 to July 31st, 2022 as validation, and August 2022 as test.

Results Following the methodology described above, the results obtained are shown in Table 2.7.

Algorithm	MAE				RMSE				R^2				Number of variables
	Mean	Std.	Max	Min	Mean	Std.	Max	Min	Mean	Std.	Max	Min	
Powershap	23,13	2,23	31,88	19,28	29,37	2,39	38,76	25,25	0,09	0,16	0,33	-0,57	107
Boruta-Shap	23,2	2,06	30,57	19,65	29,58	2	36,14	25,79	0,08	0,13	0,31	-0,36	53
Shapicant	22,04	1,5	25,85	19,26	28,24	1,63	32,49	24,94	0,16	0,1	0,35	-0,1	11
Boruta	23,49	3,16	33,86	18,35	29,84	3,25	40,55	23,72	0,06	0,22	0,41	-0,72	68
PIMP	23,31	1,73	27,72	20,11	29,49	1,75	33,96	26,19	0,09	0,11	0,28	-0,2	56
Best Lasso (0.0001)	22,79	1,81	30,01	19,5	29,22	1,96	36,03	25,96	0,1	0,13	0,3	-0,36	47
SHAPEffects 0.25-0.75	18,27	0,99	20,27	16,47	23,36	1,13	26,48	21,19	0,43	0,06	0,53	0,27	122
SHAPEffects 0.2-0.8	**16,97**	0,77	**19,32**	**15,77**	**21,5**	0,86	**23,36**	**20,02**	0,52	**0,04**	**0,58**	**0,43**	111
SHAPEffects 0.15-0.85	19,01	1,69	23,88	16,4	23,78	1,83	28,83	21	0,41	0,09	0,54	0,13	136
SHAPEffects 0.1-0.9	20,57	1,88	27,72	16,7	25,77	2,1	33,51	21,51	0,3	0,12	0,52	-0,17	78
SHAPEffects 0.05-0.95	21	1,5	24,21	18,5	27,19	1,68	31,34	23,87	0,23	0,1	0,4	-0,03	133

Table 2.7: DA Market price prediction test results

The advantage of the proposed method is evident in all aspects. All the suggested configurations achieve better results than any other method analysed. The configurations $(0.25, 0.75)$ and $(0.2, 0.8)$ eliminate all the variables related to the gas price, obtaining the best results. Moreover, in these two configurations the worst-case iterations ("Max" column for MAE and RMSE and "Min" column for R^2) are better than the average of the other feature selection methods. In addition, not only better results are obtained, but they are also more stable: the standard deviation is noticeably smaller in general than with the other algorithms. In particular, the best configuration is the one with the lowest variance. Furthermore, the percentage of improvement between the novel proposed methodology and the other methodologies is around 20-25%, which is significative. The number of variables selected seems to be higher than with the other methods; however, the improvement in the quality of the predictions justifies this selection. This essentially highlights the absence of overfitting concerns due to a high number of features selected in alternative methodologies, as our algorithm makes an even more extensive selection. The phenomenon observed here is that the features chosen by our method exhibit higher predictive performance within the current test set context, which is the actual key in feature selection. Recall that the selected variables are those that obtain the best MAE in the validation (Figure 2.9).

Another possible selection, as discussed above, would be the use of a smaller number of variables

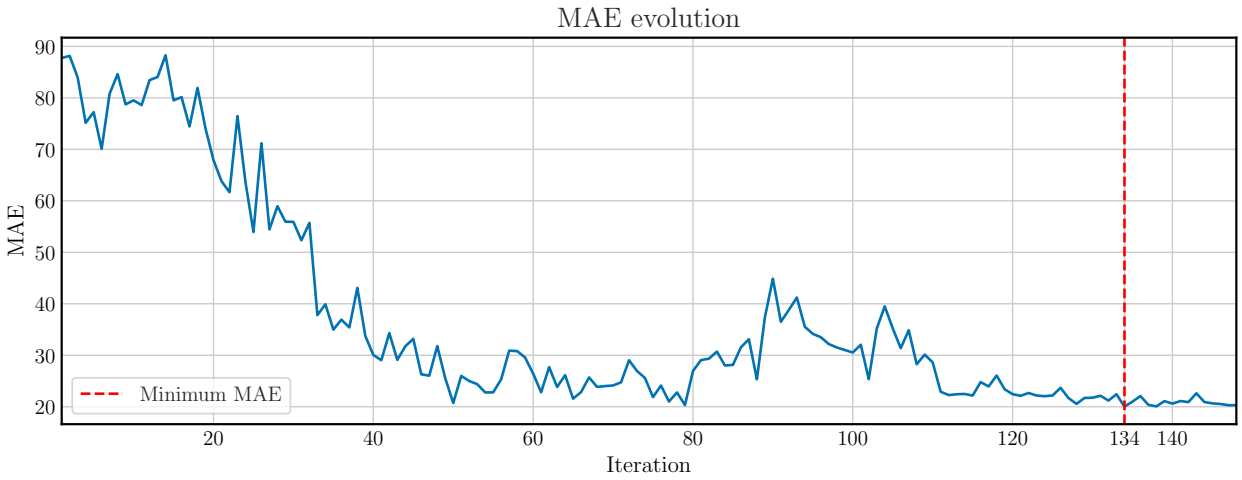

Figure 2.9: MAE evolution in the validation set over the iterations of the proposed variable selection procedure. The configuration shown is $(0.2, 0.8)$

with a slightly higher MAE, since it can be observed that the elimination of the last variables does not produce a significant increase in MAE. It is also possible to observe a considerable increase in the MAE several iterations before the optimal MAE is achieved. It could be studied which variables cause this increase and if their elimination is really advantageous.

Another real world example

In order to validate the results, one more real world data example is analysed: the Sberbank Russian Housing Market dataset (Matveev et al., 2017). It corresponds to data from a Kaggle competition with the objective of predicting the price of different houses. As the objective is not to obtain the best result in this dataset, only the training data have been used, which have been divided into three different sets chronologically, previously eliminating all the null data. As explanatory variables different variables related to information about the area in which each property is located are available. On the Kaggle competition website itself, a concept shift situation is described: '*Although the housing market is relatively stable in Russia, the country's volatile economy makes forecasting prices as a function of apartment characteristics a unique challenge*'. This perfectly describes a common situation in which it is known that it is a concept shift scenario, but the knowledge is limited regarding the presence of any potentially responsible feature, and, if such a feature exists, which specific one it may be.

Results The results are shown in Table 2.8.

Algorithm	MAE				RMSE				R^2				Number of variables
	Mean	Std.	Max	Min	Mean	Std.	Max	Min	Mean	Std.	Max	Min	
Powershap	2,972E+06	2,257E+04	3,013E+06	2,922E+06	4,899E+06	6,584E+04	5,002E+06	4,771E+06	0,53	1,300E-02	0,55	0,51	7
Boruta-Shap	2,977E+06	1,747E+04	3,008E+06	2,942E+06	4,907E+06	6,457E+04	5,055E+06	4,778E+06	0,53	1,300E-02	0,55	0,5	8
Shapicant	2,658E+06	3,425E+04	2,738E+06	2,571E+06	4,357E+06	7,911E+04	4,593E+06	4,232E+06	0,62	1,000E-02	0,65	0,58	15
Boruta	3,023E+06	2,024E+04	3,064E+06	2,983E+06	4,997E+06	5,792E+04	5,104E+06	4,836E+06	0,51	1,100E-02	0,54	0,49	6
PIMP	2,672E+06	3,595E+04	2,737E+06	2,580E+06	4,428E+06	8,025E+04	4,628E+06	4,265E+06	0,61	1,400E-02	0,64	0,58	26
Best Lasso (0.00001)	2,610E+06	2,831E+04	2,671E+06	2,560E+06	4,292E+06	7,824E+04	4,465E+06	4,127E+06	0,64	1,300E-02	0,66	0,61	177
SHAPEffects 0.25-0.75	2,615E+06	3,273E+04	2,677E+06	2,517E+06	4,311E+06	8,586E+04	4,505E+06	4,119E+06	0,63	1,500E-02	0,67	0,6	58
SHAPEffects 0.2-0.8	2,602E+06	2,612E+04	2,662E+06	2,557E+06	4,301E+06	7,870E+04	4,506E+06	4,137E+06	0,64	1,300E-02	0,66	0,6	39
SHAPEffects 0.15-0.85	2,593E+06	3,495E+04	2,692E+06	2,524E+06	4,268E+06	8,643E+04	4,472E+06	4,123E+06	0,64	1,500E-02	0,67	0,61	67
SHAPEffects 0.1-0.9	**2,567E+06**	2,219E+04	**2,604E+06**	**2,508E+06**	**4,229E+06**	7,021E+04	**4,374E+06**	**4,011E+06**	**0,65**	1,200E-02	**0,68**	**0,62**	35
SHAPEffects 0.05-0.95	2,639E+06	2,730E+04	2,718E+06	2,574E+06	4,405E+06	8,783E+04	4,575E+06	4,186E+06	0,62	1,500E-02	0,66	0,59	123

Table 2.8: Test results on the Sberbank Russian Housing Market dataset[6]

Once more, a notable disparity becomes apparent between each configuration of the proposed method and the remaining algorithms, with the former standing out prominently. The algorithm that achieves closer results is the LASSO, but uses five times more variables than the best provided configuration so, in this case, this extended selection is not justified. Again, in this instance, the worst results obtained in each metric by the best configuration of the proposed methodology is better than the best results obtained by most of the other algorithms. Moreover, in this dataset there is up to a 14% improvement, which is considerable but not as high as in the EPF example. In this particular instance, the specific variables that were excluded, resulting in the algorithm's superior performance, remain unidentified, but it would be interesting to analyse each one of the features that produces big decreases in MAE during the procedure. As mentioned earlier, it is important to note that the intention of the proposed algorithm is not to detect these features but rather to eliminate them if they are present.

2.3.2 Static scenarios

Other datasets presenting a more static situation between training, validation and test sets are now analysed. Studying the behaviour in this context is also relevant, as the aim is for the algorithm to perform optimally in all types of scenarios. They are all available in Kaggle or in the UCI Machine Learning repository (Dua and Graff, 2017):

- **CAT Scan Localization**[7]. This dataset consists of 384 features extracted from Computed Tomography (CT) images. The target variable denotes the relative location of the CT slice on the axial axis of the human body. The data was obtained from a set of 53500 CT images from 74 different patients. Each CT slice is described by two histograms in polar space. The first histogram describes the location of bony structures in the image, the second one the location of air inclusions inside the body. Both histograms are concatenated to form the final feature vector. Bins outside the image are marked with the value -0.25. The target variable (relative location of an image on the axial axis) was constructed by manually annotating up to 10 different landmarks in each CT volume with known location. The location of the slices between landmarks was interpolated. The division into training, validation and test is done randomly.

- **Appliances Energy Prediction**[8] (Candanedo et al., 2017). The aim in this dataset is to predict the energy consumption of household appliances in a low energy building. For this purpose, the consumption of the appliances every 10 minutes has been stored for 4 and a half months. There are 28 possible explanatory variables, including temperature, humidity and different weather conditions of nearby places of interest. There are also two random variables in the dataset. The division into training, validation and test is random.

[6]The preprocessing phase was not applied to this dataset because it removed many variables. (Section 2.2)
[7]https://archive.ics.uci.edu/ml/datasets/Relative+location+of+CT+slices+on+axial+axis
[8]https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction

- **Max Planck Weather Dataset** [9]. It is a dataset with 12 atmospheric characteristics over time taken every 10 minutes. The variable to predict is the wind speed after first filtering only by hourly values and eliminating data with negative wind speed. In order to have more variables, lags from one day to one week are considered for each of the variables, resulting in 80 possible explanatory variables, also considering month and time. The division is also done chronologically: from 2009 to 2015 for training, 2015 for validation and 2016 for testing.

Results All results are shown in the Tables 2.9 - 2.11.

In general, very close results can be observed among all methods. For the CAT Scan Localization dataset (Table 2.9), the selection of variables through LASSO regression seems to obtain the best results. The proposed method behaves similarly to the other algorithms, with very similar results for all five configurations. A similar situation occurs with the Appliances Energy Prediction dataset (Table 2.10), where Powershap, Boruta and Boruta-Shap obtain the best result. The proposed method is slightly worse than them and better than the other methods. In the remaining dataset (Table 2.11), the described methodology obtains the best results, selection via LASSO regression performs similarly but with a much smaller number of variables. Thus, the presented methodology got equivalent results to the other methods in the three datasets, with a number of variables in the upper range of those of the models analysed. In these instances, the variation in improvement between the different methodologies is only around 1%, so It can be concluded that the methods achieve equivalent results. Consequently, it can be asserted that the proposed methodology exhibits the capability to discern features that lack substantive impact on the target variable, thereby mitigating overfitting and the imposition of artificial relationships that are not actually present as well as the other methods.

Algorithm	**MAE**				**RMSE**				**R^2**				**Number of variables**
	Mean	**Std.**	**Max**	**Min**	**Mean**	**Std.**	**Max**	**Min**	**Mean**	**Std.**	**Max**	**Min**	
Powershap	4,42	0,09	4,62	4,2	7,12	0,17	7,46	6,78	**0,89**	5,00E-03	**0,9**	**0,88**	147
Boruta-Shap	4,37	0,09	4,61	4,16	7,11	0,17	7,5	6,72	**0,89**	5,00E-03	**0,9**	**0,88**	155
Shapicant	4,79	0,07	4,92	4,59	8,07	0,14	8,34	7,7	0,86	5,00E-03	0,87	0,85	63
Boruta	4,41	0,11	4,67	**4,12**	7,19	0,19	7,55	6,62	**0,89**	6,00E-03	**0,9**	0,87	182
PIMP	4,82	0,08	4,97	4,64	7,94	0,16	8,33	7,61	0,86	6,00E-03	0,87	0,85	33
Best Lasso (0.001)	**4,32**	0,08	**4,47**	**4,12**	**7,01**	0,15	**7,31**	**6,54**	**0,89**	5,00E-03	**0,9**	**0,88**	98
SHAPEffects 0.25-0.75	4,42	0,09	4,63	4,2	7,1	0,16	7,44	6,6	**0,89**	5,00E-03	**0,9**	**0,88**	174
SHAPEffects 0.2-0.8	4,45	0,08	4,62	4,26	7,32	0,15	7,61	6,96	0,88	4,00E-03	0,89	0,87	159
SHAPEffects 0.15-0.85	4,78	0,08	5,02	4,63	7,87	0,18	8,38	7,5	0,86	6,00E-03	0,88	0,84	105
SHAPEffects 0.1-0.9	4,45	0,11	4,72	4,2	7,23	0,18	7,78	6,66	0,88	5,00E-03	**0,9**	0,87	158
SHAPEffects 0.05-0.95	4,36	0,08	4,61	4,21	7,15	0,16	7,59	6,76	**0,89**	5,00E-03	**0,9**	0,87	168

Table 2.9: Test results on CAT Scan Localization dataset

[9] https://www.bgc-jena.mpg.de/wetter/

Algorithm	MAE				RMSE				R^2				Number of variables
	Mean	Std.	Max	Min	Mean	Std.	Max	Min	Mean	Std.	Max	Min	
Powershap	**42,24**	0,23	**42,74**	**41,7**	**81,84**	0,41	82,79	**81,02**	**0,39**	6,00E-03	**0,4**	0,37	24
Boruta-Shap	**42,24**	0,23	**42,74**	**41,7**	**81,84**	0,41	82,79	**81,02**	**0,39**	6,00E-03	**0,4**	0,37	24
Shapicant	45,04	0,22	45,51	44,66	84,74	0,33	85,57	84	0,34	5,00E-03	0,35	0,33	10
Boruta	**42,24**	0,23	**42,74**	**41,7**	**81,84**	0,41	82,79	**81,02**	**0,39**	6,00E-03	**0,4**	0,37	24
PIMP	45,15	0,21	45,52	44,56	85,53	0,33	86,22	84,28	0,33	5,00E-03	0,35	0,32	10
Best Lasso (0.00001)	42,7	0,25	43,23	42,21	82,33	0,36	83,28	81,55	0,38	5,00E-03	0,39	0,37	26
SHAPEffects 0.25-0.75	42,57	0,19	43,01	42,09	82,61	0,34	83,38	81,97	0,38	5,00E-03	0,39	0,36	23
SHAPEffects 0.2-0.8	42,57	0,19	43,01	42,09	82,61	0,34	83,38	81,97	0,38	5,00E-03	0,39	0,36	23
SHAPEffects 0.15-0.85	42,57	0,19	43,01	42,09	82,61	0,34	83,38	81,97	0,38	5,00E-03	0,39	0,36	23
SHAPEffects 0.1-0.9	42,36	0,21	42,78	41,8	81,94	0,35	**82,59**	81,2	**0,39**	5,00E-03	**0,4**	**0,38**	25
SHAPEffects 0.05-0.95	43,22	0,19	43,8	42,82	82,94	0,41	84,04	82	0,37	6,00E-03	0,38	0,35	19

Table 2.10: Test results on Appliances Energy Prediction dataset

Algorithm	MAE				RMSE				R^2				Number of variables
	Mean	Std.	Max	Min	Mean	Std.	Max	Min	Mean	Std.	Max	Min	
Powershap	1,044	0,003	1,053	1,037	1,402	0,004	1,41	1,395	0,168	0,004	0,176	0,158	79
Boruta-Shap	1,044	0,003	1,051	1,038	1,402	0,003	1,408	1,398	0,168	0,003	0,173	0,161	33
Shapicant	1,049	0,003	1,054	1,041	1,407	0,002	1,412	1,403	0,161	0,003	0,167	0,156	13
Boruta	1,044	0,003	1,051	1,037	1,401	0,003	1,407	1,395	0,169	0,003	0,176	0,162	48
PIMP	1,045	0,004	1,061	1,04	1,406	0,007	1,436	1,397	0,163	0,009	0,174	0,327	18
Best Lasso (0.001)	1,042	0,001	1,046	1,039	**1,394**	0,001	**1,397**	1,391	**0,177**	0,002	0,18	**0,174**	15
SHAPEffects 0.25-0.75	**1,039**	0,002	**1,045**	**1,034**	**1,394**	0,003	1,4	**1,389**	**0,177**	0,003	**0,184**	0,17	75
SHAPEffects 0.2-0.8	1,042	0,003	1,054	1,037	1,399	0,004	1,413	1,39	0,172	0,005	0,182	0,155	74
SHAPEffects 0.15-0.85	1,042	0,003	1,051	1,038	1,399	0,003	1,411	1,393	0,171	0,004	0,179	0,157	79
SHAPEffects 0.1-0.9	1,057	0,003	1,062	1,049	1,416	0,004	1,428	1,406	0,151	0,005	0,163	0,137	70
SHAPEffects 0.05-0.95	1,043	0,003	1,051	1,037	1,398	0,003	1,407	1,393	0,173	0,003	0,179	0,162	63

Table 2.11: Test results in the Max Planck Weather Dataset

2.4 Conclusions and future lines of research

A new feature selection method for regression problems has been developed. The algorithm is based on the idea of observing how the variables of a certain model influence when making predictions, independently of the global influence they have. Specifically, a relationship is established between the model errors and the Shapley values, allowing for a more local analysis than other feature selection algorithms. On the one hand, in situations where there is a concept shift, the algorithm is able to find the features that undergo this change of behaviour

in case they are available and have a negative influence on the predictions made by the model. Therefore, these variables are eliminated, leading to higher predictive performance and easier maintenance of the model after it has been put into production. On the other hand, in static situations, the model is able to detect the variables that cause overfitting obtaining comparable results with the state of the art. These variables create forced relationships in training, hence the effects they produce in the validation stage result in an undesired negative influence and, thus, are eliminated.

For the particular case of the EPF problem, it has been seen that the improvement provided by the presented method is more than remarkable and that its application is of clear value in the real industry. It has been put into practice through its use in a clear case of application: a regulatory change of great impact on the market. In fact, the proposed methodology has assisted in the development of the price prediction models in operation at Fortia, in particular during the months following the regulatory change where the previously existing models experienced periods of large errors and great uncertainty. Thus, it can also be concluded that the industrial impact of this development has been a success.

There are two clear problems to tackle as a continuation of this work. An interesting future line of research would be the omission of the parameters q_{low} and q_{high}. As observed, these parameters of the algorithm can change the development of the algorithm iterations, producing different results in each case. Automatic selection of such quantiles in advance or varying between iterations in an optimal way would facilitate the use of the method. From a broader perspective, the generalization of our methodology to classification problems would be a natural extension of this research, contributing to the different studies that already exist on concept shift and feature selection in this area but with a different perspective.

3

Point forecasting: An adaptive standardization methodology for Day-Ahead Electricity Price Forecasting

Note

Some of the contents of this chapter are accessible from a preprint version: Carlos Sebastián, Carlos E. González-Guillén & Jesús Juan. An adaptive standardisation methodology for Day-Ahead Electricity Price Forecasting, *arXiv preprint arXiv:2311.02610*, 2023.

In the previous chapter, a feature selection methodology was developed for contexts in which frequent behavioural changes occur, such as in the electricity market. This variable selection serves as input for predictive models that will be used as auxiliary tools for decision making. In fact, the main objective of this thesis is the development of a predictive model of the DA price in its completeness, that is, both in its point and probabilistic version. This chapter deals with the first approach: obtaining point forecasts, which is the traditional view.

3.1 Literature review

An extensive review of the various advances that have been made in this area over the years has already been provided in Section 1.5. However, most of the work covers old periods that do not correspond to the current market situation and, therefore, the adoption of these methodologies in a real context is challenging.

Furthermore, establishing state-of-the-art models in the field of EPF is not a simple task for various reasons, as pointed out in Lago et al. (2021):

- It is typical for different studies to come to contradictory conclusions, especially when comparing classical statistical methods with machine learning techniques.
- Similarly, comparisons are often drawn with basic models that fall considerably short of achieving the performance levels exhibited by the best known models. Furthermore,

studies often analyse datasets that have not been previously explored or documented in the existing literature.

- The test period is often relatively short, which introduces the possibility of overlooking special situations and failing to adequately assess the annual seasonality inherent in price dynamics. As a consequence, the obtained results may be influenced by the specific choice of the test window, potentially limiting their generalizability.
- Sometimes the results cannot be replicated, not allowing the results to be tested by other community members.

Given the strong agreement with these statements, the definition of state-of-the-art models is the same as the one presented in the review by Lago et al. (2021).

The Lasso Estimated AutoRegressive (LEAR), proposed in Uniejewski et al. (2016), is considered as the best statistically based model. It is an autoregressive model with exogenous variables (ARX) to which an L1 regularization is applied.

The DNN proposed in Lago et al. (2018) is considered one of the best machine learning based models. It has a simple structure consisting of a fully connected neural network with two hidden layers and an output layer comprising 24 neurons, each corresponding to a specific hour. So, instead of having one model for each hour as in the previous case, there is just one model for the whole series.

Both methods will be detailed in Section 3.3.1, as they will serve as the basis for the methodology to be developed in the different experiments (Section 3.3.5) that will be conducted.

In works based on electricity price forecasting, it is well known that the electricity market undergoes changes in its behaviour. However, practically all studies model this fact implicitly by using autoregressive models and calibration windows, i.e. using only the latest known data instead of all the data in the training to adjust for the latest behaviour. Nevertheless, if a fact is known, it should be modelled explicitly.

The only work that explicitly considers this fact and establishes a concrete methodology to solve this problem appears to be that of Nasiadka et al. (2022). To address this issue, the integration of change point methods is proposed, aiming to identify historical windows in the past that align with the current market conditions. These selected windows are then incorporated into the training set. However, it is important to recognize that these change point methods may not always accurately identify the relevant windows due to the complexity of the underlying market dynamics. The effectiveness of such methods relies on the assumption that the identified change points truly reflect the shifts in the market behaviour. In practice, there can be instances where the change point detection may fail to capture the subtle nuances or abrupt changes in the series, leading to potential inaccuracies in the selection of training data. Moreover, a significant portion of the available information in the historical data remains unused. Thus, while change point methods offer a valuable approach, their limitations should be considered.

3.2 Proposed methodology

A widely adopted methodology, regardless of whether statistical or machine learning-based methods are employed, is to transform non-stationary price series into a more stationary form before implementing the learning algorithm. This involves applying one of several available transformations that aim to achieve a constant mean and stabilize the variance (Uniejewski et al., 2017). These transformations are mainly based on some form of standardization of the data and are justified for several reasons. In the context of neural networks, which are commonly used in Electricity Price Forecasting (EPF) problems, employing standardized features enhances the stability of the numerical algorithms that solve the underlying optimization problems. When working with linear models, it has been observed that using some kind of regularizer improves the results. If the variables are not on the same scale, this technique becomes less effective, as the penalty imposed by the regularization does not uniformly affect all features. Regardless of the transformation used, they are computed in the training set for subsequent application in the validation and test sets. However, it is important to acknowledge that this methodology assumes the absence of any change in the joint distribution of the data, implying no occurrence of dataset shift (Section 2.1.2). Nonetheless, the presence of shifts in this series is beyond accepted.

The models presented in Section 3.1 implicitly assume that the series is stationary in both mean and variance. There have also been studies that assume that the variance of the process can vary over time, for example, by considering GARCH-type models. In Janczura and Puć (2023), it is shown that considering this fact alone does not produce improvements from the perspective of point forecasting. A further step should be taken in this direction in order to model prices correctly.

Looking at the behaviour of prices, it is reasonable to consider that consecutive intervals of varying length in the time series exhibit stationarity, or at least a constant mean and variance. This means that the time series is piecewise stationary in mean and variance. Therefore, in order to perform proper modelling, the change points in the behaviour of the series should be correctly detected and each stability period should be transformed using the correct estimation of the mean and standard deviation of each period. But, as mentioned above, detecting these regime switching points is an extremely difficult task, and even more so if it has to be done online on a daily basis. In view of this fact, let p_d^h be the price series corresponding to the hour h in the day d, the following model is proposed

$$\begin{cases} p_d^h & = \mu(X_d) + \sigma(X_d) u_d^h \equiv \mu_d + \sigma_d u_d^h \\ u_{d+1}^h & = f(X_d, u_d^h, u_{d-1}^h, \ldots) + \varepsilon_{d+1}^h,\ \mathbb{E}\left[\varepsilon_{d+1}^h\right] = 0 \end{cases} \tag{3.1}$$

The following parameter estimation methodology is proposed:

$$\begin{cases} \widehat{\mu}_d & = \dfrac{1}{24v} \displaystyle\sum_{k=d-v-1}^{d-1} \sum_{h=1}^{24} p_k^h, \text{ for each } d \\ \widehat{\sigma}_d & = \sqrt{\dfrac{1}{24v} \displaystyle\sum_{k=d-v-1}^{d-1} \sum_{h=1}^{24} \left(p_k^h - \widehat{\mu}_d\right)^2}, \text{ for each } d \end{cases} \tag{3.2}$$

where ν is the length of a rolling window in days, f is a learning algorithm and X_d is the set of explanatory features available before the end of the biding period for day d.

Although this approach to parameter estimation does not guarantee that the variance of ε_h^d remains constant across all d and h, it operates under the assumption that the series is piecewise stationary in terms of variance. Deviations from this assumption typically occur near behavioural change points, which are inherently unpredictable. If the chosen window ν is sufficiently small and entirely within a stable period, the model effectively estimates the mean and variance for that period using the latest $24 \times \nu$ data points. However, if the subsequent observation corresponds to a change point, the standardization process will be incorrect, introducing errors proportional to the magnitude of that change. Similarly, if the window spans multiple stability periods, the parameters are likely to be misestimated. The extent of this misestimation depends on the proportion of data from the newer period and the intensity of the change in behaviour. Consequently, if the forthcoming data point to be predicted represents a change point or belongs to a very recent regime, predicting its value accurately becomes challenging.

An important detail in the estimation of parameters (3.2) is that it can be affected by outliers. Thus, instead of working with the original price series p_d^h it is proposed to work with one where the outliers have been filtered out and replaced appropriately. That is, at the time of predicting the observation p_{d+1}^h, the price series in the training set will be the one given by

$$\tilde{p}_{d'}^h = \begin{cases} p_{d'}^h & \text{if } \widehat{\mu}_{d'} - \kappa \cdot \widehat{\sigma}_{d'} \leq p_{d'}^h \leq \widehat{\mu}_{d'} + \kappa \cdot \widehat{\sigma}_{d'} \\ \underset{\substack{t \in \{1,\dots,\nu\} \\ h' \in \{1,\dots,24\}}}{\text{Median}} \left\{ p_{d'-t}^{h'} \right\} & \text{in other case} \end{cases},$$

where $\widehat{\mu}_{d'}$ and $\widehat{\sigma}_{d'}$ are computed as in (3.2) for a window of ν days.

Figure 3.1 presents a comparison between the current price series in the Spanish Day-Ahead market and the series obtained using a median-arcsinh standardization scheme, which has been applied in the EPF literature (Uniejewski and Weron, 2018), and another series representing the u_d^h process resulting from our proposed methodology.

It is evident from the figure that the proposed methodology yields a significantly more stationary series, which aligns with the main objective of applying such transformations. The fundamental idea of the transformation is to encapsulate the effects of potential shifts in the μ_d and σ_d parameters, ensuring that the resulting process u_d^h is minimally affected by such situations. Additionally, implementing the adaptive standardization through a reasonably sized rolling window enables a rapid response to these shifts.

As this adaptive standardization mitigates shifts in the behaviour of the price series, the unmodified explanatory features are sub-optimal candidates for the prediction of transformed prices. They also need to be transformed. Furthermore, this step is crucial to avoid spurious regressions caused by non-stationary explanatory variables (Harris and Sollis, 2003). It is logical to apply the rolling standardization of length ν, with the parameters μ_d and σ_d as in (3.2) to each one of the variables for non-dummy features and leave dummy features unchanged.

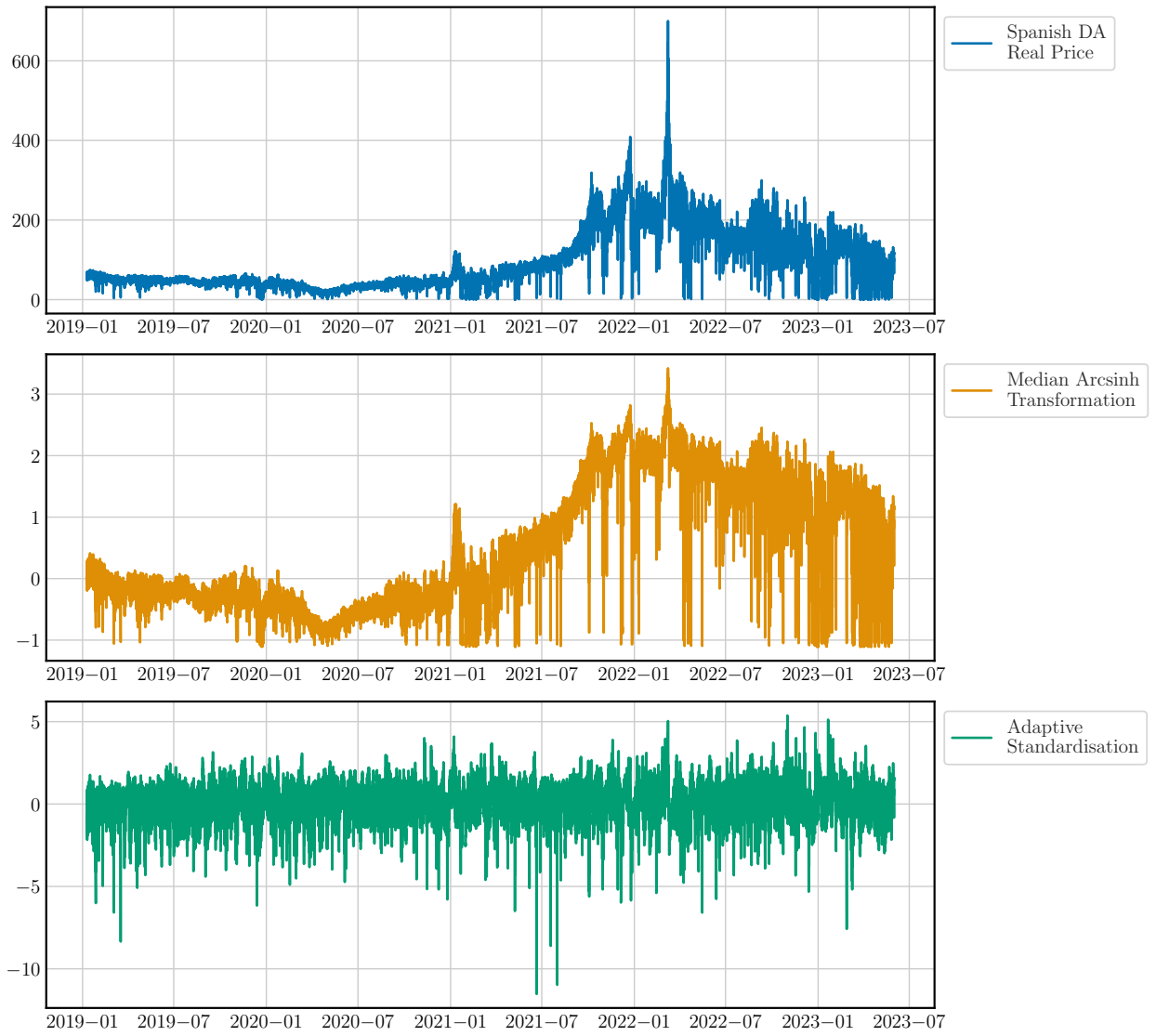

Figure 3.1: (Top) Spanish Day-Ahead electricity price. **(Middle)** Median-arcsinh transformation of the Spanish DA electricity price **(Bottom)** Adaptive standardization with $\nu = 7$ days, 168 hours, of the Spanish DA electricity price

3.3 Experiments

To evaluate the results, a thorough analysis of the proposed methodology is conducted using the established Python library `epftoolbox` (Lago et al., 2021). By comparing the use of a conventional transformation scheme in the EPF literature (median-arcsinh transformation, Uniejewski and Weron (2018)) with the one proposed here, the significance of our approach and its impact on the final model can be shown. The models are retrained each day, as one should do in a real industrial EPF situation. For the adaptive standardization, $\nu = 7$ days (168 hours) has been selected due to the seasonality of prices (caused by the electricity load). Although

other windows such as 14 days or 21 days could also be viable options, the smallest "weekly" window has been selected to allow for quick adaptation to new regimes. Logically, ν can be chosen by the user based on the need of the problem to be solved. In fact, the choice of this window could vary between the different datasets considered, with a larger window for more stable series and a smaller window for cases with strong changes in behaviour. However, for simplicity, only one value has been considered.

3.3.1 Learning algorithms

The LEAR model and the neural network both presented in Section 3.1 and available in the `epftoolbox` package are considered as potential learning algorithms f (Equation 3.1).

LEAR

The model specification is

$$\begin{aligned} u_d^h = &\sum_{i=1}^{24} \theta_{h,i} \cdot u_{d-1}^i + \sum_{i=1}^{24} \theta_{h,24+i} \cdot u_{d-2}^i + \sum_{i=1}^{24} \theta_{h,48+i} \cdot u_{d-3}^i + \sum_{i=1}^{24} \theta_{h,72+i} \cdot u_{d-7}^i \\ &+ \sum_{i=1}^{24} \theta_{h,96+i} \cdot x_{d,1}^i + \sum_{i=1}^{24} \theta_{h,120+i} \cdot x_{d,2}^i + \sum_{i=1}^{24} \theta_{h,144+i} \cdot x_{d-1,1}^i \\ &+ \sum_{i=1}^{24} \theta_{h,168+i} \cdot x_{d-1,2}^i + \sum_{i=1}^{24} \theta_{h,192+i} \cdot x_{d-7,1}^i + \sum_{i=1}^{24} \theta_{h,216+i} \cdot x_{d-7,2}^i \\ &+ \sum_{j=1}^{7} \theta_{h,240+j} \cdot z_d^j + \varepsilon_d^h \end{aligned},$$

where u_d^h is the standardized price of day d in hour h (p_d^h), $x_{d,1}^h$ and $x_{d,2}^h$ are two variables of interest associated with the market on day d in hour h, usually related to load, wind forecasting or solar forecasting, and z_d^j is a binary variable that assumes a value of 1 if day of the week j is the same as day d, and 0 otherwise. The values of $x_{d,1}^h, x_{d,2}^h$ may require a standardization similar to that of the price. The ε_d^h simply represents noise. Note that this is a model for each hour.

The coefficients of the model $\widehat{\theta}_h = (\widehat{\theta}_{h,1}, \ldots, \widehat{\theta}_{h,247})$ are calculated as

$$\widehat{\theta}_h = \underset{\theta_h}{\operatorname{argmin}} \sum_{d=1}^{N_d} (u_d^h - \widehat{u}_d^h)^2 + \lambda \sum_{i=1}^{247} |\theta_{h,i}|$$

with $\widehat{u}_d^h$ the forecast of day d at hour h, N_d the number of days in the training set and λ a regularization hyperparameter of the model that can be fitted across a multitude of schemes.

In Lago et al. (2021) the LARS method is used together with the AIC score to optimize the hyperparameter λ. Due to updates of the `Scikit-learn` library (Pedregosa et al., 2011) the

application of this method is not so easy for small calibration windows (where the number of observations is smaller than the number of variables). This is why cross-validation has been adopted as a method to determine this hyperparameter. This increases the computation time significantly, but facilitates the determination of λ for small calibration windows. In any case, for calibration windows where this problem is not encountered, experiments with the LARS-AIC scheme have been replicated to verify that the results differ only slightly (see A).

DNN

The DNN proposed in Lago et al. (2018) is considered as a representative of the machine learning based models. The model parameters are estimated using Adam optimization algorithm (Kingma and Ba, 2014). The same set of potential input features as employed in the LEAR model is employed. The selection of specific input features is determined as another hyperparameter of the network. To optimize these hyperparameters, a tree-structured Parzen estimator (Bergstra et al., 2011) is used. Notably, hyperparameters unrelated to the explanatory features encompass the following: the number of neurons per layer, the choice of activation function, the dropout rate, the learning rate, the inter-layer normalization scheme, preprocessing procedures prior to the input layer, weight initialization strategies, and the coefficient associated with the applied L1 regularization.

3.3.2 Datasets

Five datasets are analysed, two of which are novel contributions to the literature. These datasets are in the public domain and they are made available to the research community[10], as they offer valuable opportunities for studying and addressing market conditions that are more representative of the current state. The five markets to consider are (Figure 3.2)

- The OMIE-SP market, representing the Spanish electricity market. This market is not available in the Python package. Price data spans from January 1st, 2019 to May 31st, 2023. Two explanatory variables, namely the day-ahead load forecast and the forecast of renewable energy generation (solar photovoltaic, solar thermal and wind). The data has been obtained through the ESIOS platform[11].
- The EPEX-DE market, representing the German electricity market, is another dataset considered in this study. Data from January 1st, 2019 to May 31st, 2023 has been obtained from the ENTSO-E transparency platform[12]. Similar to the previous market, the same exogenous variables have been considered. However, in this case, the renewable generation forecast does not include the solar thermal generation forecast.
- The EPEX-FR market, which is the Day-Ahead electricity market in France. The dataset also encompasses the period from January 9th, 2011 to December 31st, 2016. It incorporates two exogenous variables: the day-ahead load and generation forecasts from France.

[10] https://github.com/CCaribe9/AdaptStdEPF
[11] https://www.esios.ree.es/
[12] https://transparency.entsoe.eu/

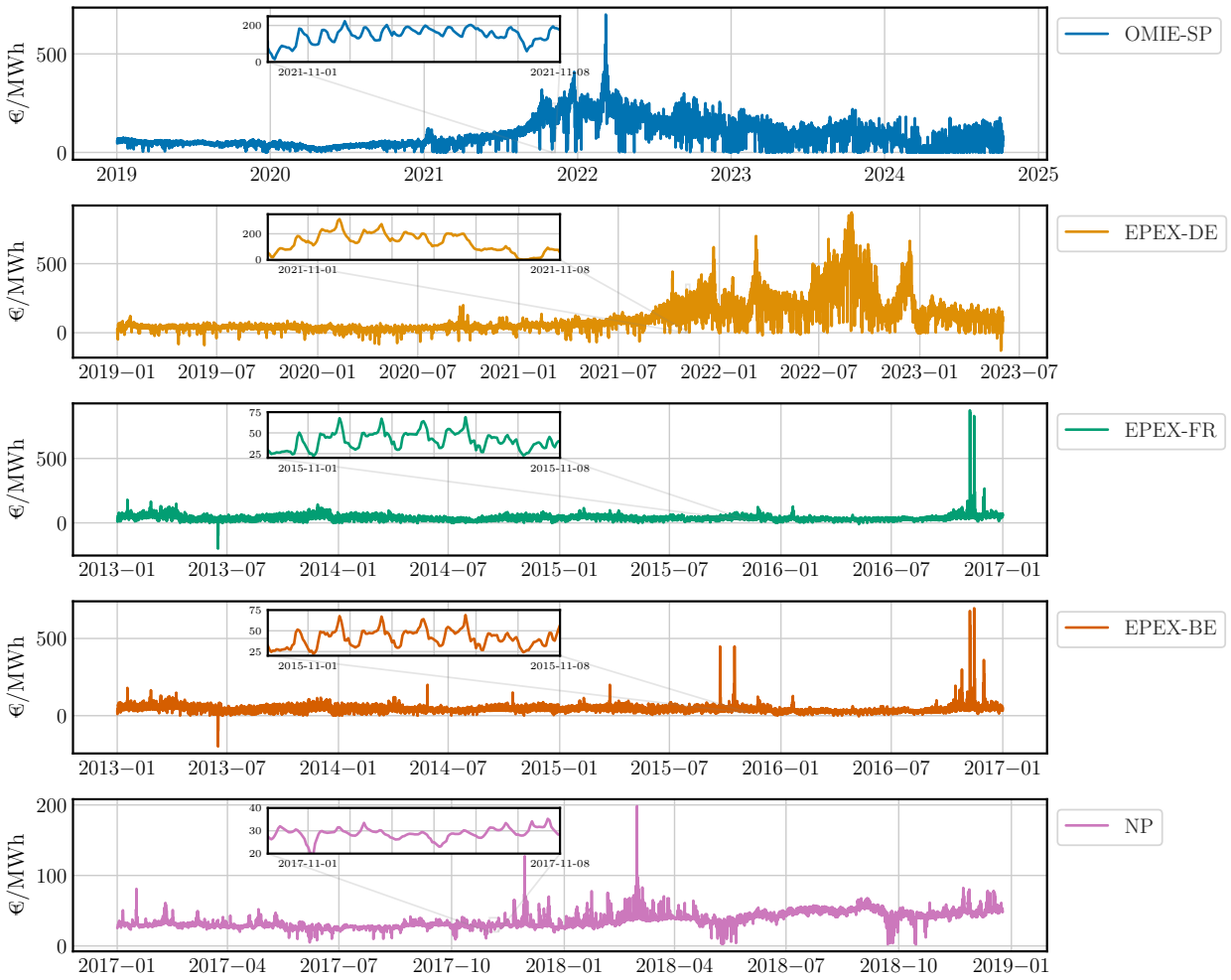

Figure 3.2: Markets considered. Observe the differences between the variability of current markets and previous markets.

The dataset is accessible from the `epftoolbox` package.

- The EPEX-BE market, which is the Day-Ahead electricity market in Belgium. The dataset encompasses the period from January 9th, 2011 to December 31st, 2016. It incorporates the same two exogenous variables from France, which, although surprising, are two of the best regressors for this market (Lago et al., 2018). The dataset is accessible from the `epftoolbox` package.
- The NP market, which is the European power market of the Nordic countries. Data spans from January 1st, 2013 to December 14th, 2018. The explanatory features considered are the day-ahead load forecast and the day-ahead wind generation forecast. The dataset is accessible from the `epftoolbox` package.

The test period of every market is detailed in Table 3.1.

The LEAR model will be evaluated using four different calibration windows: two long and two short. According to Lago et al. (2021), the calibration windows for the EPEX-BE, EPEX-FR,

Market	Test period
OMIE-SP	01.01.2022 - 31.05.2023
EPEX-DE	01.01.2022 - 31.05.2023
EPEX-BE	04.01.2015 - 31.12.2016
EPEX-FR	04.01.2015 - 31.12.2016
NP	27.12.2016 - 24.12.2018

Table 3.1: Test period of every dataset considered

and NP datasets will be 56, 84, 1092, and 1456 days, corresponding to 8 weeks, 12 weeks, 3 years, and 4 years, respectively. Given the higher volatility and more pronounced behavioural changes in the OMIE-SP and EPEX-DE datasets, shorter long windows of 56, 84, 364, and 728 days will be used. Furthermore, scenarios where the entire dataset is used for training will also be considered. This approach is particularly relevant for adaptive standardization, as the series seems to be comparable across all instances, which implies that the use of fixed calibration windows may be unnecessary if all changes are effectively captured by the parameters μ_d and σ_d. This study extends the work of Lago et al. (2021) by including this comparison to critically assess the necessity for calibration windows in traditional normalization schemes. For the DNN model, four different sets of hyperparameter optimizations are conducted, with the comparative results of these configurations examined. If adaptive standardization is not implemented, a calibration window of 4 years is employed for the EPEX-BE, EPEX-FR, and NP datasets, and 3 years for the OMIE-SP and EPEX-DE datasets. When adaptive standardization is applied, all available training data is used.

For the outlier filtering process, we consider $\kappa = 10$ as the intention is to filter out only prices that were really atypical and disproportionate to those of the last few days. For example, in Figure 3.2 it can be seen near July 2013 in EPEX-BE and EPEX-FR clearly negative and atypical prices. In these same series, at the end of 2016, prices above 500€/MWh can be seen which are contrary to the dynamics observed in the rest of the series. These are clear examples of outliers that need to be mitigated. Probably better results could be obtained with other values of κ, as it actually is an hyperparameter in the methodology.

3.3.3 Evaluation metrics

Four accepted metrics in the forecasting literature are used to measure the quality of outcomes:

- $\text{MAE} = \dfrac{1}{24N_d}\displaystyle\sum_{d=1}^{N_d}\sum_{h=1}^{24}|p_d^h - \widehat{p}_d^h|$
- $\text{RMSE} = \sqrt{\dfrac{1}{24N_d}\displaystyle\sum_{d=1}^{N_d}\sum_{h=1}^{24}\left(p_d^h - \widehat{p}_d^h\right)^2}$
- $\text{sMAPE} = \dfrac{1}{24N_d}\displaystyle\sum_{d=1}^{N_d}\sum_{h=1}^{24}2\dfrac{|p_d^h - \widehat{p}_d^h|}{|p_d^h| + |\widehat{p}_d^h|}$

- $$\text{rMAE} = \frac{\frac{1}{24N_d}\sum_{d=1}^{N_d}\sum_{h=1}^{24}|p_d^h - \widehat{p}_d^h|}{\frac{1}{24N_d}\sum_{d=1}^{N_d}\sum_{h=1}^{24}|p_d^h - p_d^{h,\text{naive}}|}$$

where N_d is the number of days in the test set and p_d^h, $\widehat{p}_d^h$, $p_d^{h,\text{naive}}$ are the current price, the prediction and the prediction by a naive model of the hour h on day d, respectively. In the EPF field, is common to consider $p_d^{h,\text{naive}} = p_{d-7}^h$, so weekly effects are captured.

3.3.4 Forecasting procedure

For each dataset and for each day $D+1$ of the test period to be forecasted the process is as follows:

1. Take as training data set the one given by the corresponding calibration window. If no calibration window is considered, take all data prior to the day to be predicted as training data.
2. If adaptive standardization is applied, transform the price series by filtering outliers as explained in Section 3.2.
3. Transform the price series and the explanatory variables using the selected normalization process. Either adaptive standardization or median-arcsinh.
4. Select and train the learning algorithm: LEAR or DNN.
5. Predict all hours of the day $D+1$.
6. Undo the selected normalization scheme.

When all days of the test period have been forecasted, the metrics presented in Section 4.4.1 are computed.

In addition, results have been computed without filtering outliers in the case of adaptive standardization to show the importance of this step (Appendix B). Also, to show that this step alone is not the one that produces considerable improvements, the results of the filtered series have been computed by applying the median-arcsinh transformation instead of applying the adaptive standardization (Appendix C).

3.3.5 Results

Tables 3.2 and 3.3 show the evaluation metrics for the mentioned datasets for the algorithms LEAR and DNN, respectively. The notation "AS" before each model name will be adopted to indicate that it incorporates the adaptive standardization approach.

Looking at the performance of the LEAR model it can be seen that larger windows produce better results. However, there is a point where this larger calibration window stops producing

		LEAR					ASLEAR				
Market	**Metrics**	**56**	**84**	**364/ 1092**	**728/ 1456**	**All**	**56**	**84**	**364/ 1092**	**728/ 1456**	**All**
OMIE SP	**MAE**	21,03	20,86	19,40	19,46	20,69	21,75	20,35	18,68	18,48	**18,27**
	RMSE	30,66	31,09	27,96	27,57	29,28	32,31	29,26	26,78	26,18	**25,93**
	sMAPE	0,23	0,23	0,22	0,22	0,22	0,23	0,22	**0,21**	**0,21**	**0,21**
	rMAE	0,55	0,54	0,51	0,51	0,54	0,57	0,53	0,49	**0,48**	**0,48**
EPEX DE	**MAE**	31,42	30,45	30,67	28,54	31,13	29,72	28,76	26,47	25,99	**25,65**
	RMSE	47,31	44,99	42,52	40,60	44,60	44,67	42,69	39,14	38,65	**38,11**
	sMAPE	0,25	0,25	0,25	0,23	0,24	0,24	0,23	0,22	0,22	**0,21**
	rMAE	0,44	0,42	0,43	0,40	0,43	0,41	0,40	0,37	**0,36**	**0,36**
EPEX BE	**MAE**	7,03	6,92	**6,45**	**6,45**	6,54	8,30	7,83	6,85	6,84	6,84
	RMSE	**16,16**	16,19	16,51	16,42	16,36	25,92	26,16	17,16	17,24	17,24
	sMAPE	0,17	0,16	0,16	0,16	0,16	0,17	0,17	0,16	0,16	**0,15**
	rMAE	0,69	0,68	**0,64**	**0,64**	**0,64**	0,82	0,77	0,67	0,67	0,67
EPEX FR	**MAE**	4,82	4,62	4,21	4,28	4,37	4,86	4,69	4,15	**4,14**	4,19
	RMSE	**10,83**	11,49	11,67	11,67	11,67	14,30	14,56	12,17	12,24	12,64
	sMAPE	0,14	0,13	0,13	0,13	0,14	0,13	0,13	**0,12**	**0,12**	**0,12**
	rMAE	0,66	0,63	**0,57**	0,58	0,60	0,66	0,64	**0,57**	**0,57**	**0,57**
NP	**MAE**	2,02	1,96	1,96	1,96	**1,92**	2,61	2,47	2,01	1,98	1,97
	RMSE	3,76	3,73	3,56	3,57	**3,53**	4,88	4,79	3,83	3,80	3,80
	sMAPE	**0,06**	**0,06**	**0,06**	**0,06**	**0,06**	0,08	0,07	**0,06**	**0,06**	**0,06**
	rMAE	0,49	0,47	0,47	0,47	**0,46**	0,63	0,60	0,49	0,48	0,48

Table 3.2: Evaluation metrics the LEAR and ASLEAR models for every dataset

improvements or even worsens the results, except for the NP market. This makes sense, since in view of this model not all observations are comparable to each other. The ASLEAR model does have this feature, always taking advantage of all the available information. It can be seen how the best results for this case are always considering all the data, except for the EPEX-FR market, where they are practically the same. In any case, the results with small calibration windows are not positive, being very distant from the rest of the configurations. Nevertheless, this is inline with the intended used of the methodology. Comparing LEAR and ASLEAR, the biggest differences appear in the OMIE-SP and EPEX-DE markets, the markets that deal with the least stable periods. For the former, comparing the best models, the predictions are 6% better in MAE, and for the latter dataset 10% better, very remarkable figures. For the rest of the datasets the results are very similar comparing the best models in each case. These datasets show little change in behaviour, so the adaptive standardization is actually producing very similar results to what a static standardization would be and, in fact, what is observed are the differences in performance of two different static standardizations. Anyway, the results are slightly better for LEAR in the case of EPEX-BE and NP and for ASLEAR in the case of EPEX-FR.

With regard to the DNN and ASDNN models[13], the first point to be made concerns the volatility of the results. Neural networks, due to their hyperparameter configuration and the initialization

[13]It is important to note that DNN and ASDNN are not comparable for a hyperparameter configuration labelled with the same number. It simply means four different configurations for each case.

		DNN				ASDNN			
Market	**Metrics**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
OMIE SP	**MAE**	22,38	19,24	23,37	20,79	18,19	18,10	17,87	**17,67**
	RMSE	31,82	27,46	32,96	30,15	26,01	26,12	25,31	**25,11**
	sMAPE	0,23	**0,21**	0,24	0,22	**0,21**	**0,21**	**0,21**	**0,21**
	rMAE	0,58	0,50	0,61	0,54	0,47	0,47	0,47	**0,46**
EPEX DE	**MAE**	27,21	25,61	26,18	26,83	26,28	25,92	**24,35**	26,26
	RMSE	39,17	37,32	38,14	38,69	38,75	38,37	**36,11**	38,70
	sMAPE	0,22	**0,21**	0,22	0,22	**0,21**	**0,21**	**0,21**	**0,21**
	rMAE	0,38	0,36	0,36	0,37	0,36	0,36	**0,34**	0,36
EPEX BE	**MAE**	6,29	**6,22**	6,46	6,42	6,39	6,58	6,65	6,44
	RMSE	16,13	16,24	16,63	16,55	**15,89**	16,48	16,55	16,21
	sMAPE	**0,15**	**0,15**	**0,15**	**0,15**	**0,15**	**0,15**	0,16	**0,15**
	rMAE	0,62	**0,61**	0,64	0,63	0,63	0,65	0,65	0,63
EPEX FR	**MAE**	4,25	4,24	4,23	4,12	**4,08**	4,11	4,16	4,17
	RMSE	11,94	12,14	11,89	12,01	**11,16**	12,04	11,90	11,71
	sMAPE	**0,12**	**0,12**	**0,12**	**0,12**	**0,12**	**0,12**	**0,12**	**0,12**
	rMAE	0,58	0,58	0,58	**0,56**	**0,56**	**0,56**	0,57	0,57
NP	**MAE**	2,11	1,83	1,96	1,84	1,74	1,76	**1,73**	1,77
	RMSE	3,95	3,47	3,67	3,50	**3,37**	3,48	3,46	3,58
	sMAPE	0,06	**0,05**	0,06	**0,05**	**0,05**	**0,05**	**0,05**	**0,05**
	rMAE	0,51	0,44	0,48	0,45	**0,42**	**0,42**	**0,42**	0,43

Table 3.3: Evaluation metrics the DNN and ASDNN models for every dataset

of the network weights, can give very different results for the same network architecture. In the case of the DNN this variation is very clear, producing highly diverse results for each dataset. The best examples are seen in the OMIE-SP, EPEX-DE and NP markets, which are also the most volatile markets, at least compared to the other two. For the ASDNN model this variation in results is minimal, seeing very stable results across all configurations for all datasets, perhaps with the exception of EPEX-DE. This property is important, as relatively good performance is ensured without the need for multiple runs or combining different networks. Regarding the performance on each dataset, the trend of the LEAR and ASLEAR models is somewhat observed. For OMIE-SP the difference is even more noticeable, where all ASDNN networks are better than the best DNN network. A significant difference is also seen in the case of EPEX-DE, as expected. The DNN model is still better than the ASDNN for EPEX-BE, although the results are still very close. For EPEX-FR the ASDNN model is better for all networks, except for one DNN network which performs at the same level. Finally, in the case of the NP market, it is observed a separation that was not present between the LEAR and ASLEAR, perhaps because of the now existing ability to capture non-linear behaviour. Under these conditions, the ASDNN model performs better for all networks.

An easy improvement, as shown in Lago et al. (2021), can be achieved by creating ensembles of the different predictions. Normally, the simplest schemes in this sense already obtain very good results that are hard to beat (Petropoulos et al., 2022). Thus, for each group of models (LEAR, ASLEAR, DNN and ASDNN) the mean of all the predictions of each group is considered as a new predictor. In the case of LEAR, the case in which all the observations available for training are considered is excluded, since it is not included in the state-of-the-art model of Lago et al. (2021) either. In addition, for ASLEAR two variants, 1 (2), are considered, where the 54 and 86 day calibration windows are not (are) considered because of their poor individual performance. Table 3.4 presents the results, comparing each ensemble with the best individual model of each group.

Market	Metrics	Ens. LEAR	Best LEAR	Ens.$_1$ ASLEAR	Ens.$_2$ ASLEAR	Best ASLEAR	Ens. DNN	Best DNN	Ens. ASDNN	Best ASDNN
OMIE SP	**MAE**	18,19	19,40	18,11	18,16	18,27	19,60	19,24	**17,29**	17,67
	RMSE	26,16	27,96	25,86	26,20	25,93	28,60	27,46	**24,70**	25,11
	sMAPE	0,20	0,22	0,21	**0,20**	0,21	0,21	0,21	**0,20**	0,21
	rMAE	0,47	0,51	0,47	0,47	0,48	0,51	0,50	**0,45**	0,46
EPEX DE	**MAE**	26,24	28,54	25,52	25,41	25,65	**23,69**	25,61	23,91	24,35
	RMSE	37,78	40,60	37,99	37,97	38,11	**34,61**	37,32	35,22	36,11
	sMAPE	0,22	0,23	0,21	0,21	0,21	**0,20**	0,21	**0,20**	0,21
	rMAE	0,36	0,40	0,35	0,35	0,36	**0,33**	0,36	**0,33**	0,34
EPEX BE	**MAE**	6,22	6,45	6,82	6,96	6,84	**6,06**	6,22	6,31	6,39
	RMSE	15,85	16,42	17,15	18,50	17,24	16,14	16,24	16,07	**15,89**
	sMAPE	0,15	0,16	0,15	0,15	0,15	**0,14**	0,15	0,15	0,15
	rMAE	0,61	0,64	0,67	0,69	0,67	**0,60**	0,61	0,62	0,63
EPEX FR	**MAE**	4,04	4,21	4,14	4,17	4,14	4,00	4,12	**3,94**	4,08
	RMSE	10,86	11,67	12,34	12,63	12,24	11,87	12,01	11,53	**11,16**
	sMAPE	0,12	0,13	**0,11**	**0,11**	0,12	**0,11**	0,12	**0,11**	0,12
	rMAE	0,55	0,57	0,57	0,57	0,57	0,55	0,56	**0,54**	0,56
NP	**MAE**	1,75	1,96	1,98	2,02	1,97	1,74	1,83	**1,64**	1,73
	RMSE	3,39	3,56	3,80	3,92	3,80	3,44	3,47	**3,33**	3,46
	sMAPE	**0,05**	0,06	0,06	0,06	0,06	**0,05**	**0,05**	**0,05**	**0,05**
	rMAE	0,42	0,47	0,48	0,49	0,48	0,42	0,44	**0,40**	0,42

Table 3.4: Evaluation metrics for the ensembles of the different models computed

When ensembles are computed, methods that make use of adaptive standardization do not improve as much as those that do not. The variance of the individual models when adaptive standardization is not applied is by itself a negative feature when a single model is to be used. Nevertheless, when several models are mixed together they take into account very diverse behaviour, leading to large improvements in the metrics considered. In particular, in the case of OMIE-SP, the ASLEAR and ASDNN ensembles are better than their respective non-adaptive versions. However, the improvement over the best single model is much larger for the non-adaptive schemes. For EPEX-DE the ASLEAR ensembles remain better, but the DNN ensemble slightly outperforms the ASDNN ensemble in MAE, although they maintain the same level of rMAE. For the EPEX-BE market there was already a better performance of the non-adaptive models and this trend is also maintained in this case. In the French case, EPEX-FR, the ASLEAR ensembles do not produce any improvement over the best individual model, while for LEAR

they do. Still, the ASDNN ensemble is the best model over the DNN combination. Finally, the same behaviour is observed for NP as for France, where ASDNN is again the best model. Ultimately, the ASDNN ensembles are generally the best models, but the improvement over the best model is not as great as for the non-adaptive versions. While this may seem like a negative feature for adaptive standardization, it actually indicates that very similar performance to the combination of non-adaptive models can be achieved using a single model to which adaptive standardization has been applied.

A good practice for evaluating models, which is commonly ignored, is the dynamic evaluation of some metric. For instance, the MAE per month of each of the models. Figure 3.3 compares the ensembles with each other for each of the datasets on a month-by-month basis. In the following, the ensemble of the ASLEAR model set never considers the 56 and 84 day windows.

It can now be appreciated where the advantages of each method come from: whether they are period-specific or consistent over time. In the case of OMIE-SP it can be seen that the adaptive methods perform better over the entire period evaluated. This could be suspected as the gain in this market is very significant compared to the other markets. In the EPEX-DE market there are mixed situations where the adaptive and non-adaptive versions exchange the best model position. There is also a significant jump in quality between using the (AS)LEAR model and the (AS)DNN model. The Belgian and French markets are very similar. The main differences are given by those MAE peaks, which are coincidentally periods where very atypical prices stand out. Probably with a better treatment of outliers (e.g. by not setting such a high κ) these problems would be alleviated. Other than that, performance is roughly equivalent, except for a period in late 2015 and early 2016 where in both markets the adaptive versions are slightly better. For NP, the performance of the ASLEAR ensemble was by far the worst of all (Table 3.4), as it was the only model that did not improve when averaging over the individual models. It can now be seen that this is mainly determined by a period from November 2017 to March 2018.

Another property that can be observed from these plots is that the adaptive and non-adaptive versions capture different behaviours. Just as it made sense to combine models from different calibration windows or different hyperparameter settings, it makes sense to combine the adaptive and non-adaptive versions. These combinations are denoted by LEAR-ASLEAR and DNN-ASDNN and are constructed by making the arithmetic mean of the two ensembles involved. Table 3.5 shows the results.

For the OMIE-SP market, the difference between adaptive and non-adaptive versions is so large that there is no improvement when combining neural networks. However, when combining the linear models the improvement is very significant, reaching the level of the ASDNN ensemble. For the rest of the markets, the best model is given by the DNN-ASDNNN combination and, generally, by a wide margin with respect to the second best proposal. Therefore, although the individual models of the adaptive versions perform well on their own, and although combinations of models from the same methodology provide excellent performance, the real gain comes from using the two methodologies together, as they take into account properties that are captured by both techniques.

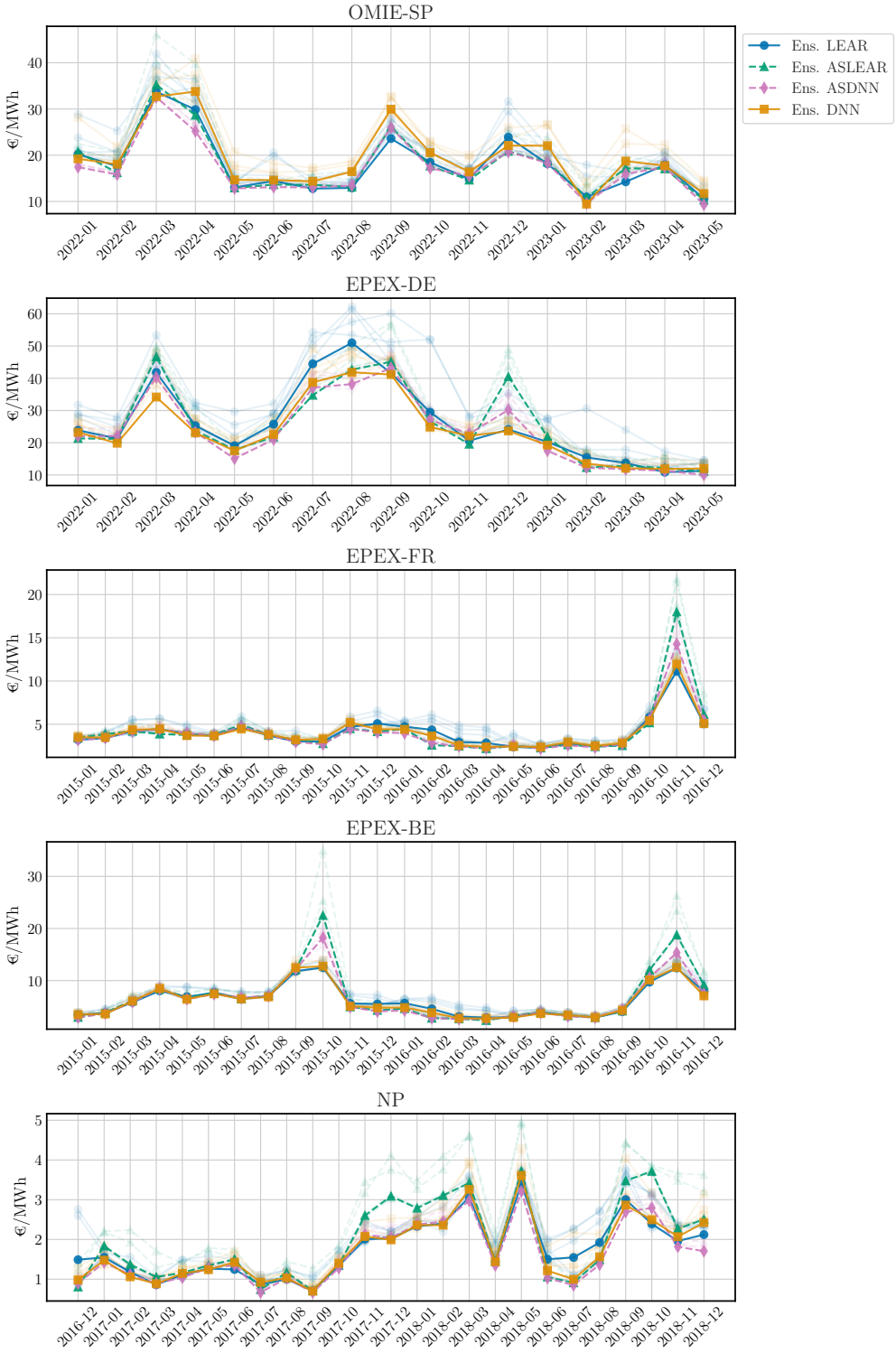

Figure 3.3: MAE per month and per dataset for each model set. Each individual model is shown in translucent form.

Market	Metrics	Ens. LEAR	Ens. ASLEAR	LEAR ASLEAR	Ens. DNN	Ens. ASDNN	DNN ASDNN
OMIE SP	MAE	18,19	18,11	17,30	19,60	**17,29**	17,33
	RMSE	26,16	25,86	24,76	28,60	**24,70**	25,23
	sMAPE	0,20	0,21	**0,20**	0,21	**0,20**	**0,20**
	rMAE	0,47	0,47	**0,45**	0,51	**0,45**	**0,45**
EPEX DE	MAE	26,24	25,52	23,84	23,69	23,91	**22,10**
	RMSE	37,78	37,99	34,89	34,61	35,22	**32,47**
	sMAPE	0,22	0,21	0,20	0,20	0,20	**0,19**
	rMAE	0,36	0,35	0,33	0,33	0,33	**0,31**
EPEX BE	MAE	6,22	6,82	6,23	6,06	6,31	**5,95**
	RMSE	15,85	17,15	15,85	16,14	16,07	**15,84**
	sMAPE	0,15	0,15	**0,14**	**0,14**	0,15	**0,14**
	rMAE	0,61	0,67	0,61	0,60	0,62	**0,59**
EPEX FR	MAE	4,04	4,14	3,90	4,00	3,94	**3,81**
	RMSE	10,86	12,34	**11,38**	11,87	11,53	11,54
	sMAPE	0,12	0,11	0,11	0,11	0,11	0,11
	rMAE	0,55	0,57	0,53	0,55	0,54	**0,52**
NP	MAE	1,75	1,98	1,74	1,74	1,64	**1,60**
	RMSE	3,39	3,80	3,40	3,44	3,33	**3,27**
	sMAPE	**0,05**	0,06	**0,05**	**0,05**	**0,05**	**0,05**
	rMAE	0,42	0,48	0,42	0,42	0,40	**0,39**

Table 3.5: Evaluation metrics for the ensembles and the combination of the adaptive methods with the non-adaptive ones

3.3.6 Statistical testing

It is important to analyse whether the difference between predictions from different models is statistically significant. In the context of EPF, the Diebold-Mariano test (Diebold and Mariano, 2002) is used for this purpose.

The Diebold-Mariano test evaluates the hypothesis that the expected value of $\Delta_{d,h}^{A,B}$ is zero, where $\Delta_{d,h}^{A,B} = L(\varepsilon_{d,h}^{A}) - L(\varepsilon_{d,h}^{B})$. Here, $\varepsilon_{d,h}^{Z} = p_d^h - \widehat{p}_d^h$ represents the prediction error of model Z for day d at hour h, and L denotes the loss function used in the analysis. For this study, the absolute loss is considered as the loss function.

The statistic $\text{DM} = \sqrt{N}\frac{\widehat{\mu}}{\widehat{\sigma}}$ is computed, where $\widehat{\mu}$ and $\widehat{\sigma}$ represent the mean and standard deviation of $\Delta_{d,h}^{A,B}$, respectively, and N is the number of observations in the test period. The test statistic asymptotically follows a standard normal distribution. In this way, the computation of

the p-value associated with the test

$$\begin{cases} H_0: & \mathbb{E}\left(\Delta_{d,h}^{A,B}\right) \leq 0 \\ H_1: & \mathbb{E}\left(\Delta_{d,h}^{A,B}\right) > 0 \end{cases}$$

can be obtained.

If the null hypothesis is rejected, it indicates that the forecasts of model B are statistically more accurate than those of model A.

It should be noted that the Diebold-Mariano test assumes that the observations are covariance stationary. Given that predictions are made for all hourly periods of day $D+1$, it is likely that this condition may not hold. To address this issue, there are two possible approaches. The first approach involves transforming the hourly time series into 24 daily series. This allows for separate tests to be performed for each hour of the day, resulting in 24 individual tests. This hourly univariate perspective enables a more detailed analysis of the model performance for each specific hour. Alternatively, a multivariate perspective can be adopted by considering 24-dimensional vectors representing the predictions for all hours of the day. In this case, a single test is conducted based on the norms of these vectors. This multivariate perspective provides a more concise and direct analysis, as the comparison is summarized in a single test statistic. This last version will be applied through the `epftoolbox`.

Results

A statistical analysis is conducted for each market to compare the predictive performance of different models using the multidimensional Diebold-Mariano test. The results of this analysis are presented in the form of a colored matrix heatmap, where each cell represents a p-value. It is important to note that rejecting the null hypothesis indicates that the model in the column of the matrix performs better than the model in the corresponding row. Figure 3.4 shows the results.

The statistical test confirms the results discussed in the previous section. The individual non-adaptive models do not perform particularly well compared to the other models. However, the ASDNN type models show very good performance generally, even outperforming ensembles in some situations. LEAR and ASLEAR models with small calibration windows (56 and 84) do not produce good results on their own and, in case of using a single model of this type larger windows should be considered. The Belgian market is the only one that does not seem to benefit from the use of adaptive versions, although it has been analysed in Figure 3.3 that this is probably related to an improved treatment of outliers. Ensembles significantly improve performance in general and the combination of DNN and ASDNN models stands out in all markets.

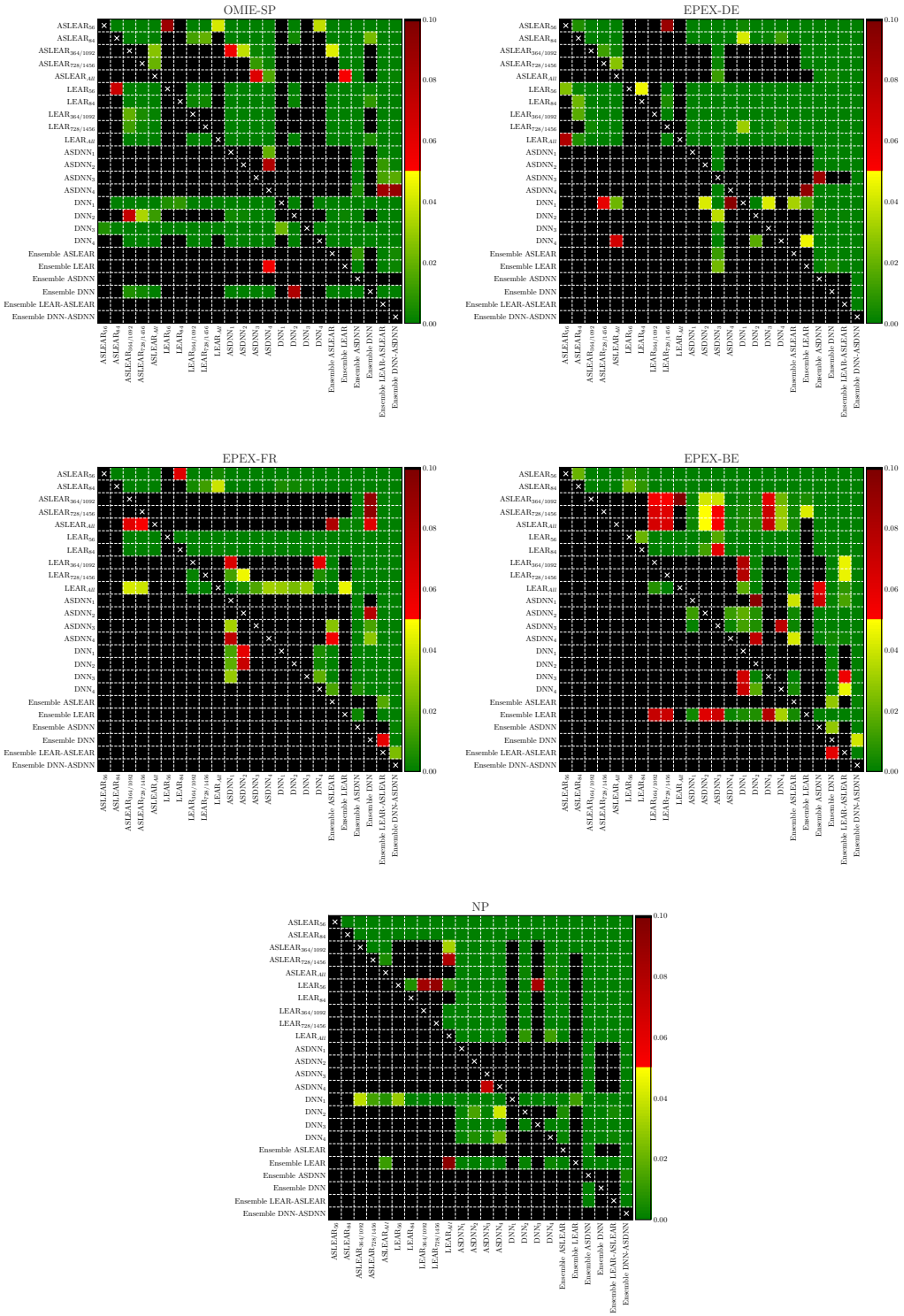

Figure 3.4: Multivariate DM test between each analysed model, including ensembles, for the every dataset considered

3.4 Conclusions and future lines of research

A new framework for a price prediction model in the Day-Ahead market based on the price dynamics has been proposed. This new approach has been thoroughly studied, demonstrating improved results across various metrics and showing statistical improvement in five different markets and two distinct market periods. In particular, the improvement in the Spanish market is truly remarkable. The outlier mitigation process has proven to be vital for achieving these results for the proposed models, although it could still be improved for better results. Additionally, two new and recent datasets have been made available to the community, aiming to explore new models on datasets that are closer to the current market situation. To get the best potential out of the models, combining them has proven to be key. In particular, combining the two methodologies evaluated produces the best results.

While the results may be further improved by using alternative learning algorithms for f in Equation (3.1), in some cases no statistically significant difference has been observed among the evaluated individual adaptive models in this study, which are the ones that have performed the best by themselves. One way to achieve better results could be through the use of additional explanatory variables in estimating μ_d and σ_d. The inclusion of new explanatory features, especially those related to variable generation costs such as gas prices, oil prices, etc., can lead to improved results (Section 1.5). And, as it has been shown in Chapter 2, the selection of variables must be done carefully. It is crucial to consider the potential impact of dataset shifts and the biases that can be generated in this context by different explanatory variables. Additionally, adaptive standardization has been used, but other adaptive transformations could also be considered, like adaptive median-arcsinh. Finally, it has already been indicated that the combination of models produces very good results. This combination has been done by simply using the mean, which usually produces results that generalize very well. However, following the study in Figure 3.3, it can be seen that in some cases the best model alternates and is not always the same. This is why a promising line of future work are online aggregation schemes such as those presented in Gaillard et al. (2014), Wintenberger (2017) or Adjakossa et al. (2023) in the context of EPF.

4

Probabilistic Forecasting: Heteroscedastic Quantile Regression and Width-Adaptive Conformal Inference

Note

Some of the contents of this chapter are under review in *International Journal of Forecasting* and accessible from a preprint version: Carlos Sebastián, Carlos E. González-Guillén & Jesús Juan. Enhancing reliability in prediction intervals using point forecasters: Heteroscedastic Quantile Regression and Width-Adaptive Conformal Inference, *arXiv preprint arXiv:2406.14904*, 2024.

4.1 Context of the problem

In the previous chapter, electricity price point forecasting was examined, where the goal was to predict the most likely future price at each time step. While point forecasts are valuable for decision-making, they provide limited insight into the uncertainty surrounding future prices.

This chapter introduces probabilistic forecasting, an extension of point forecasting that estimates the distribution of potential future outcomes at each instant. Given a time series $y_1, y_2, \ldots, y_T$, $y_t \in \mathbb{R}$, the objective is to forecast the next h steps, $y_{T+1}, y_{T+2}, \ldots, y_{T+h}$. Probabilistic forecasting aims to model the conditional density $\mathbb{P}\left(y_{T+i} \middle| y_1, y_2, \ldots, y_T, X_1, X_2, \ldots, X_{T+i}\right)$, for each of the next h steps, where $X_t = \left(x_{t,1}, x_{t,2}, \ldots, x_{t,k}\right)$ represents a vector of k regressors available at each time step t.

The importance of probabilistic forecasting is particularly evident in high-stakes applications. For example, this can be seen in the electricity market through electricity price forecasting or renewable energy prediction (Zhang et al., 2014; Nowotarski and Weron, 2018). In the first case, high price volatility requires the presence of risk-aware strategies. Underestimating future prices may result in missed opportunities for higher profits, while overestimating prices could

lead to losses from overcommitment or uncompetitive bidding. Probabilistic forecasts allow market participants to incorporate these cost asymmetries into their strategies. In renewable energy forecasting, the inherent unpredictability of weather conditions creates a pressing need for probabilistic methods. The power output of renewable sources like wind turbines and solar panels depends heavily on stochastic variables such as wind speed or solar irradiance. Probabilistic forecasts enable grid operators to account for this variability by scheduling backup generation and maintaining reserve capacity. They also help renewable energy producers avoid penalties by providing a clearer picture of the likelihood of over or under-delivery in their market bids.

Let's assume that only one-step predictions are made ($h = 1$) and that the conditional distribution to be modelled is unimodal. In this work, uncertainty is represented through prediction intervals, which provide a range within which future values are expected to fall with a specified probability. That is, given a miscoverage rate $\alpha \in (0,1)$ a prediction interval $\widehat{C}_\alpha(X_{T+1}) = \left[\widehat{l}_\alpha(X_{T+1}), \widehat{u}_\alpha(X_{T+1})\right] \subseteq \mathbb{R}$ is build such that

$$\mathbb{P}\left(y_{T+1} \in \widehat{C}_\alpha(X_{T+1})\right) \geq 1-\alpha. \tag{4.1}$$

An interval is said to be valid when property (4.1) is satisfied, i.e. when its marginal coverage is greater than or equal to the target coverage determined by the user. However, coverage alone does not guarantee optimal intervals. When building a prediction interval, the most efficient valid interval possible is desired (Shafer and Vovk, 2008). The efficiency of an interval is related to its length. When two prediction intervals achieve the same coverage level, the one with shorter length is preferred. This preference stems from the fact that narrower intervals provide more precise and actionable information, reducing the uncertainty range while maintaining reliability. Achieving the specified level of coverage with minimal interval length is crucial for constructing intervals that are not only statistically valid but also practically useful. The interval length of a prediction interval $\widehat{C}_\alpha(X_{T+1})$ is denoted by $\left|\widehat{C}_\alpha(X_{T+1})\right|$.

While these are crucial properties, this thesis asserts that they alone are not sufficient for ensuring practical utility. In addition to these considerations, two further essential properties that prediction intervals must satisfy to enhance their applicability in real-world decision-making are proposed.

1. **Adaptivity and correlation with prediction difficulty**: prediction intervals should adapt to different levels of uncertainty present in the data. Specifically, the length of the intervals should be correlated with the difficulty of the prediction, such that shorter intervals are associated with easier-to-predict situations and longer intervals with more challenging ones. This ensures that the intervals effectively reflect the underlying uncertainty in the forecasting process.
2. **Independence between coverage and interval length**: In practice, the assessment of a prediction interval's validity and efficiency requires a comprehensive evaluation over a period involving multiple prediction intervals. The empirical coverage is defined as

$$\frac{1}{N}\sum_{t=T}^{T+N} \mathbb{1}\left(y_{t+1} \in \widehat{C}_\alpha(X_{t+1})\right)$$

where $\mathbb{1}(\cdot)$ is the indicator function and $N \in \mathbb{N}$ is the number of predictions that have been made. The marginal coverage of $\widehat{C}_\alpha(X_{T+1})$ is approximated by this quantity. That is,

$$\widehat{\mathbb{P}}\big(y_t \in \widehat{C}_\alpha(X_t)\big) = \frac{1}{N}\sum_{t=T}^{T+N} \mathbb{1}\big(y_{t+1} \in \widehat{C}_\alpha(X_{t+1})\big).$$

Let $\mathscr{I}_\rho$ be the set of indices such that the length of the interval associated with that index is within $\delta > 0$ of $\rho \in \mathbb{R}$, i.e,

$$\mathscr{I}_\rho = \left\{ t+1 : \left|\widehat{C}_\alpha(X_{t+1}) - \rho\right| \leq \delta\,,\, \rho \in \mathbb{R} \right\}.$$

Let N_ρ be the number of elements of that set. For all ρ such that $N_\rho \neq 0$, the desired property that it is advocated is

$$\widehat{\mathbb{P}}\big(y_t \in \widehat{C}_\alpha(X_t) \,\big|\, \big|\widehat{C}_\alpha(X_t)\big| \approx \rho\big)^{14} = \frac{1}{N_\rho}\sum_{i \in \mathscr{I}_\rho} \mathbb{1}\big(y_i \in \widehat{C}_\alpha(X_i)\big) = 1-\alpha \tag{4.2}$$

for a specified miscoverage value $\alpha \in (0,1)$. This ensures that the coverage level is consistent and independent of the interval's length, maintaining its reliability across different prediction difficulties.

While the first of these properties has been recognised in numerous papers as a desirable characteristic (Angelopoulos and Bates, 2021), to our knowledge there is only one paper (Feldman et al., 2021) that has dealt with the second and with a different perspective. In Feldman et al. (2021), independence between the coverage indicator and the interval length is pursued to enhance the conditional coverage of a quantile regression process. This approach is grounded in the observation that, for the true quantiles of the conditional distribution, the coverage indicator and the interval length are orthogonal. In this work this approximation is extended to the case of time series, but with the similar objective of making a better approximation of the real quantiles, thus avoiding biases in situations of low or high uncertainty.

Moreover, these properties are intrinsically linked to aleatoric and epistemic uncertainty. Adaptability to more or less predictable situations is directly tied to aleatoric uncertainty, as it ensures that prediction intervals accurately reflect variations in inherent randomness. For instance, periods of higher or lower volatility can be effectively differentiated. In contrast, the independence between coverage and interval length is associated with epistemic uncertainty, as it indicates the absence of biases linked to prediction difficulty. A method that produces noticeably different coverage levels depending on the interval length exhibits signs of inadequate modelling or insufficient data to capture the true data-generating process.

Finally, this work addresses a practical scenario commonly encountered in the industry, where only M point forecasting models are available, and no additional information about the underlying data-generating process or the models themselves is accessible. Specifically, at time T,

[14] Here using $\big|\widehat{C}_\alpha(X_t)\big| \approx \rho$ means $\big|\widehat{C}_\alpha(X_{t+1}) - \rho\big| \leq \delta$, thus allowing a small difference δ around ρ. This way, you avoid requiring the interval length to match ρ exactly, which can be restrictive or even infeasible depending on the data.

the only available information consists of the M predictions $\widehat{\boldsymbol{y}}_{T+1} = \left(\widehat{y}_{T+1,1}, \widehat{y}_{T+1,2}, \ldots, \widehat{y}_{T+1,M}\right)$ for y_{T+1} along with their historical values. This situation is typical in contexts where organizations rely on external forecasting tools without detailed knowledge of their construction or assumptions. The primary goal of this work is to propose methods for generating reliable prediction intervals using only the outputs of these point forecasters, i.e, a prediction interval $\widehat{C}_\alpha(X_{T+1}) \equiv \widehat{C}_\alpha(\widehat{\boldsymbol{y}}_{T+1})$ is going to be built at the same time that the resulting intervals exhibit the two key properties introduced earlier, which are often overlooked in the literature. By addressing this limitation, the proposed approach ensures both theoretical soundness and practical applicability in industrial settings, making it a versatile solution for real-world forecasting challenges. To simplify the notation, in the rest of the chapter the constructed interval will be denoted by $\widehat{C}_{\alpha,T+1}$, although it should be noted that the source of information in the construction comes from the different predictions.

In this chapter two novel contributions are presented:

1. A quantile regression model is proposed, inspired by the philosophy of the Quantile Regression Averaging (QRA) model of Nowotarski and Weron (2015), but with modifications so that there is an increasing relationship between the length of the interval and the difficulty of the prediction. Due to the particular use of the standard deviation of the point predictors, the model is called Heteroscedastic Quantile Regression (HQR).

2. To provide theoretical coverage guarantees and to achieve uniformity of coverage regardless of the difficulty of the prediction, the Width-Adaptive Conformal Inference (WACI) method is proposed, which modifies the Adaptive Conformal Inference (ACI, Gibbs and Candès (2021)) method by solving the problems that the rest of the models in the literature may present in this regard.

The combination of HQR with WACI ensures strong results in terms of both validity and efficiency while simultaneously fulfilling the two desired properties.

4.2 Prior work on probabilistic forecasting

4.2.1 General overview

Bayesian methods, by their very nature, are clear candidates for probabilistic prediction. Through Bayes' theorem, a posterior distribution can be obtained by updating beliefs as new information is obtained. Assuming a parametric model dependent on weights on the target variable, a distribution over these weights can be adopted. This is the approach followed in Bayesian neural networks (Neal, 2012). One can also consider the Bayesian approach directly on the target variable in the variant known as evidential regression (Amini et al., 2020) or with a functional approach through Gaussian processes (Rasmussen, 2003). However, Bayesian

methods present problems such as the choice of the prior distribution or the computational complexity.

Assuming a specific distribution, one can try to estimate the distribution of y_{T+1} based on the information known at time T. This is done by methods such as NGBoost (Duan et al., 2020), GAMLSS (Stasinopoulos and Rigby, 2008) as well as distributional neural networks and mixture density networks (Bishop, 1994). But the constraint of selecting a particular distribution can be quite restrictive. Data behaviour often evolve over time, and a fixed distribution may fail to remain valid as these changes occur. Additionally, it is common to assume simple distributions (e.g., the normal distribution), which often fail to capture the complex characteristics exhibited by real-world data, such as skewness, heavy tails, or multimodality. Conversely, selecting an overly complex distribution can lead to issues such as overfitting or excessive computational demands.

From a non-parametric point of view, classical methods such as bootstrapping the residuals to generate prediction intervals can be applied (Efron, 1987). However, the generality of the method tends not to produce the most satisfactory results. The application of quantile regression (Koenker and Bassett Jr, 1978) is also very popular, either through a linear model or by extending the method to more complex approaches such as neural networks (Cannon, 2011). Quantile regression is explained in more detail in Section 4.2.2

All these methods can be easily extended to time series problems (for example by considering autoregressive effects, which is common practice) but none of them can assure the marginal coverage needed to provide valid prediction intervals. The Conformal Prediction framework (Vovk et al., 2005) ensures such marginal coverage in finite samples by assuming exchangeability between observations and without any assumptions about the probability distribution. In fact, Conformalized Quantile Regression (CQR) (Romano et al., 2019) extends quantile regression by providing the property of validity under exchangeability, while trying to fit the heteroscedasticity properties encountered in the data. Indeed, the properties that are to be found in the intervals to be constructed are closely tied to achieving conditional coverage, a topic of significant interest in the field of conformal predictions (Romano et al., 2019; Sesia and Romano, 2021; Chernozhukov et al., 2021; Han et al., 2022; Izbicki et al., 2022). However, the existing body of work on conditional coverage does not fully align with the framework considered here, as it typically assumes a static data distribution over time, a condition that is not met in the context presented. As the exchangeability property is very demanding in time series, a large branch of research has focused on maintaining the good properties of the conformal predictors without assuming it. See for example (Gibbs and Candès, 2021; Zaffran et al., 2022; Bhatnagar et al., 2023; Auer et al., 2023; Gibbs and Candès, 2022). More details related with Conformal Prediction can be found in Section 4.2.3

Regarding the context of the problem at hand, where only different predictors of the event to be forecasted are known, most methodologies can, in principle, be adapted by treating these predictors as explanatory variables. However, to our knowledge, there is only one work that

has approached it in such a way: the Quantile Regression Averaging (QRA) model proposed by Nowotarski and Weron (2015). Additionally, there are methods designed to combine predictors of the mean to enhance point forecasts, which have been adapted for probabilistic forecasting in an online setting, such as the approach proposed by Gaillard et al. (2016) using the algorithm in Gaillard et al. (2014). This work will focus on QRA, which will be examined in greater detail in Section 4.2.2.

4.2.2 Quantile regression

Let F_{T+1} be the cumulative distribution function of the random variable Y_{T+1}. For a given probability level $\beta \in (0,1)$, the quantile β of Y_{T+1} is defined as

$$q_\beta(Y_{T+1}) = \inf\{y \in \mathbb{R} : F_{T+1}(y) \geq \beta\}.$$

When constructing prediction intervals, prioritizing their validity is crucial. Valid intervals are essential for making informed decisions and managing risk, as they guarantee the desired level of coverage. If the intervals do not align with the expected confidence level, their reliability is compromised, leading to a loss of trust in the predictions and diminishing their value as decision-making tools. Therefore, the estimation of quantiles is a logical approximation to the problem. Let $\alpha \in (0,1)$ be the target miscoverage value and $\xi \in (0,\alpha)$. If $C_{\alpha,T+1} = [l_{\alpha,T+1}, u_{\alpha,T+1}] = [q_\xi(Y_{T+1}), q_{1-\alpha+\xi}(Y_{T+1})]$ is considered, then

$$\mathbb{P}(y_{T+1} \in C_{\alpha,T+1}) \geq 1-\alpha,$$

where y_{T+1} is the actual observed value at time $T+1$.

The typical choice for ξ is $\xi = \frac{\alpha}{2}$ and this is the approach that will be followed in this work. However, it should be noted that if the smallest intervals are desired, this option is not necessarily optimal. For example, when the distribution of Y_{T+1} is asymmetric, other choices of ξ can provide shorter intervals without sacrificing coverage.

Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ be the vector of n observations of the variable of interest and $\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,m})$ be the vector of m explanatory features for the observation i for all $i \in \{1,2,\ldots,n\}$. Let $y_i \sim Y$ and $\boldsymbol{x}_i \sim X$ for all $i = 1,2,\ldots,n$. Let $Y|X$ be distributed as F. Let's consider the model

$$q_\beta(Y|\boldsymbol{x}_i) = \lambda_0(\beta) + \lambda_1(\beta)x_{i,1} + \lambda_2(\beta)x_{i,2} + \cdots + \lambda_m(\beta)x_{i,m} + \varepsilon_i(\beta);\ \mathbb{E}[\varepsilon_i(\beta)] = 0$$

where $\boldsymbol{\lambda}(\beta) \equiv \boldsymbol{\lambda} = (\lambda_0(\beta), \lambda_1(\beta), \lambda_2(\beta), \ldots, \lambda_m(\beta))$ are the parameters of the model and $\varepsilon_i(\beta)$ represents noise.

Just as the mean squared error serves as the loss function optimized to estimate the conditional mean as a point estimator, the conditional quantile is estimated by minimizing the pinball loss

function (Koenker and Bassett Jr, 1978):

$$\ell_\beta(y_i, \widehat{y}_i) = \beta |y_i - \widehat{y}_i| \mathbb{1}\left\{y_i - \widehat{y}_i \geq 0\right\} + (1-\beta)|y_i - \widehat{y}_i| \mathbb{1}\left\{y_i - \widehat{y}_i \leq 0\right\}$$

The pinball loss function for different quantile values β is represented in Figure 4.1.

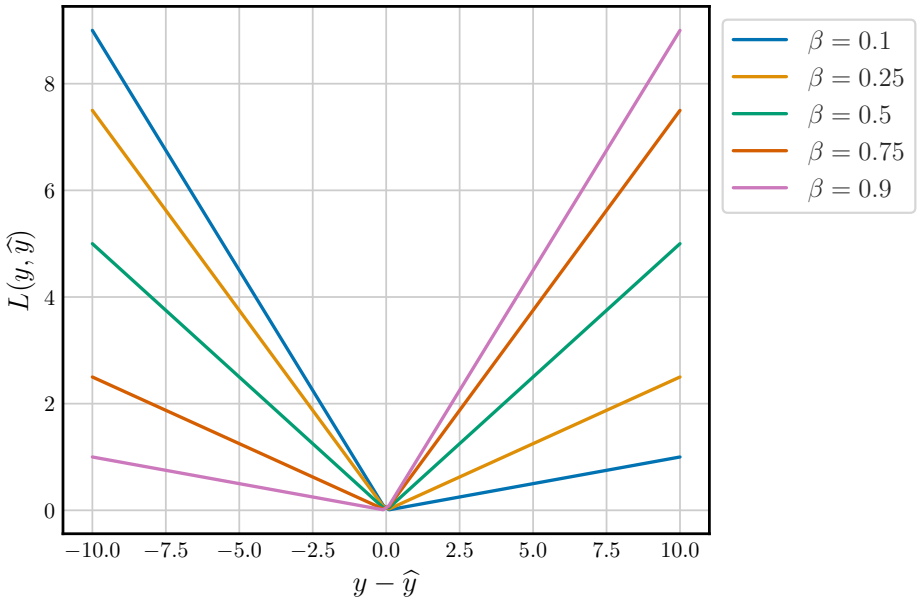

Figure 4.1: Pinball loss depending on the quantile β.

The parameters $\boldsymbol{\lambda}$ are estimated as

$$\widehat{\boldsymbol{\lambda}} = \min_{\boldsymbol{\lambda}} \left\{ \sum_{i=1}^{n} \ell_\beta(y_i, \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{x}_i) \right\}$$

and inference about a new observation $n+1$ is done through

$$\widehat{q}_\beta(Y|\boldsymbol{x}_{n+1}) = \widehat{\lambda}_0(\beta) + \widehat{\lambda}_1(\beta)\, x_{n+1,1} + \widehat{\lambda}_2(\beta)\, x_{n+1,2} + \cdots + \widehat{\lambda}_m(\beta)\, x_{n+1,m}$$

Quantile Regression Averaging

The core idea in Nowotarski and Weron (2015) is to estimate the quantiles by treating the individual point forecasts as independent variables. While the original work does not explicitly restrict itself to using only this information, the model presented there is well-suited to the current problem, which is to produce prediction intervals relying solely on different point predictors.

Although the model is presented in the context of Day-Ahead electricity price forecasting, it is perfectly generalizable to any regression problem. In particular, the model proposed for the

quantile β at time t is

$$q_\beta(Y_t|\boldsymbol{y}_t) = \lambda_0(\beta) + \lambda_1(\beta)\widehat{y}_{t,1} + \lambda_2(\beta)\widehat{y}_{t,2} + \cdots + \lambda_m(\beta)\widehat{y}_{t,M} + \varepsilon_t(\beta);\ \mathbb{E}\left[\varepsilon_t(\beta)\right] = 0, \tag{4.3}$$

where Y_t is the random variable associated to instant t and $\boldsymbol{y}_t = \left(\widehat{y}_{t,1}, \widehat{y}_{t,2}, \ldots, \widehat{y}_{t,M}\right)$ are predictions of the mean from M different models for the same time instant t. In particular, one-step ahead predictions would be obtained by:

$$\widehat{q}_\beta(Y_{T+1}|\boldsymbol{y}_{T+1}) = \widehat{\lambda}_0(\beta) + \widehat{\lambda}_1(\beta)\widehat{y}_{T+1,1} + \widehat{\lambda}_2(\beta)\widehat{y}_{T+1,2} + \cdots + \widehat{\lambda}_m(\beta)\widehat{y}_{T+1,M}. \tag{4.4}$$

To fit the models for time series problems, the use of a rolling window approach is proposed. Thus, the model described in equation (4.3) and (4.4) would be the particular model for one window. For another window, another estimation of the model parameters would be obtained. Figure 4.2 describes the process of a rolling window methodology. The window size in this procedure is chosen empirically.

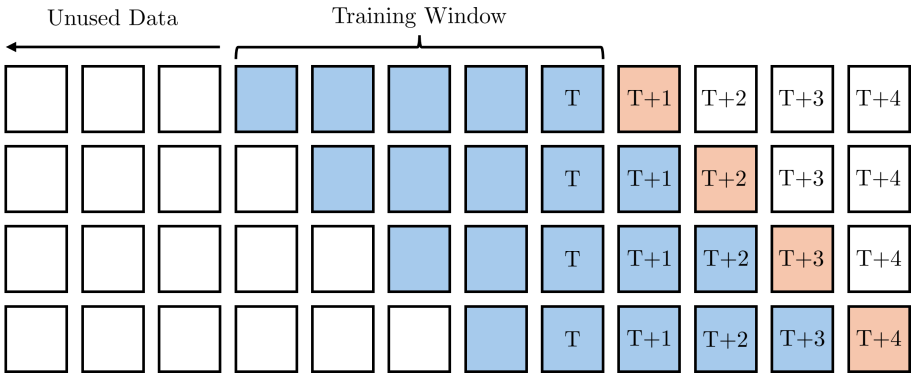

Figure 4.2: Rolling window mechanism with size equal to 5 time steps. To predict the next time step, only the data from the previous 5 time steps is used to estimate the model parameters.

Although quantile regression procedures based on the pinball loss produce asymptotically consistent estimators (Koenker and Bassett Jr, 1978), over a finite amount of data there is no theoretical guarantee of obtaining the desired marginal coverage. This is where the Conformal Prediction framework adds value.

4.2.3 Conformal Prediction

Conformal predictions were introduced in Vovk et al. (2005) to build prediction intervals (in the regression framework) that are valid with a finite number of data, without assumptions, except exchangeability, about any kind of distribution. It operates as a post-processing phase within an existing prediction pipeline, enabling the construction of valid prediction intervals without requiring any modifications to the existing forecasting process. Its fundamental base assures that intervals are valid regardless of the quality of the initial predictions, although their

efficiency remains influenced by the accuracy of the preceding forecasting phase.

Although the original approach, commonly referred to as Full Conformal Prediction, is not computationally feasible on a large scale, the approach known as Split Conformal Prediction (SCP, Lei et al. (2018); Papadopoulos et al. (2002)) solves such problems making its use more appealing in a multitude of situations. This work only focuses on the second approach.

Suppose there are n points $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^m \times \mathbb{R}$, $i = 1, \ldots, n$ and it is of interest to provide a prediction interval for the next observation y_{n+1} for which $\boldsymbol{x}_{n+1}$ is known. Conformalization in its simplest form consists in making a correction to a prediction of the mean. Let $\widehat{\mu}(\cdot)$ be that predictor. The steps to perform its conformalization for an objective miscoverage of α are described in Algorithm 2.

Algorithm 2 Conformalized Mean Regression trough SCP

Require: $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, significance level α, regression algorithm $\widehat{\mu}$

1: Randomly split the n known points into two disjoint sets: training Tr and calibration Cal.
2: Train the regression algorithm $\widehat{\mu}$ using the data from the training set Tr.
3: Compute the conformity scores for the calibration set Cal using the absolute error:

$$\mathscr{S} = \mathscr{S}_{\text{Cal}} \cup \{+\infty\},$$

where

$$\mathscr{S}_{\text{Cal}} = \{|y_i - \widehat{\mu}(\boldsymbol{x}_i)| : i \in \text{Cal}\}.$$

4: Compute the $(1-\alpha)$ quantile of the conformity scores, denoted as $Q_{1-\alpha}(\mathscr{S})$.
5: Construct the conformal prediction interval for observation $n+1$ as:

$$\widehat{C}_{\alpha,n+1} = \left[\widehat{\mu}(\boldsymbol{x}_{n+1}) - Q_{1-\alpha}(\mathscr{S}),\ \widehat{\mu}(\boldsymbol{x}_{n+1}) + Q_{1-\alpha}(\mathscr{S})\right].$$

Ensure: $\widehat{C}_{\alpha,n+1}$ prediction interval of level $1-\alpha$ for the observation $n+1$.

Theorem 1 (Lei et al. (2018))**.** Let $(\boldsymbol{x}_i, y_i)_{i=1}^{n+1}$ be exchangeable. The process of conformalizing a conditional mean predictor as described in Algorithm 2 produces a prediction interval for the observation $n+1$, $\widehat{C}_{\alpha,n+1}$, such that

$$\mathbb{P}\left(y_{n+1} \in \widehat{C}_{\alpha,n+1}\right) \geq 1-\alpha.$$

If, in addition, the scores $\mathscr{S}_{\text{Cal}}$ have a continuous joint distribution, then:

$$\mathbb{P}\left(y_{n+1} \in \widehat{C}_{\alpha,n+1}\right) \leq 1-\alpha+\frac{1}{\#\text{Cal}+1}.$$

Conformalized Quantile Regression (CQR)

While this methodology is useful, its simplicity does not take into account the possible heteroscedasticity depending on the covariates. That is, a stronger property that would be desirable

is conditional coverage:

$$\mathbb{P}\left(y_{n+1} \in C_{\alpha,n+1} | \boldsymbol{x}_{n+1} = \boldsymbol{x}\right) \geq 1-\alpha \quad \forall \boldsymbol{x} \in \mathbb{R}^m.$$

In Figure 4.3 the difference between marginal and conditional coverage is appreciated. Notice how in the case of conditional coverage, the prediction intervals are adjusted to the heteroscedasticity of the data as a function of X.

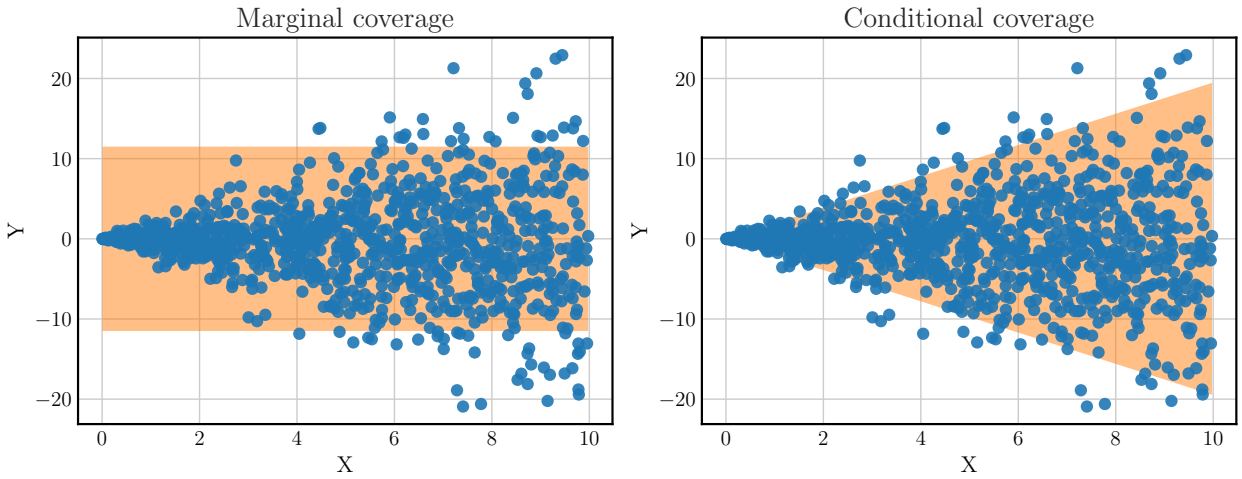

Figure 4.3: Difference between marginal coverage and conditional coverage in a toy dataset.

For any distribution-free approximation, not just Conformal Prediction, enforcing this property would require the intervals to be uninformative. That is, achieving this property in a practical way and without assuming any distribution is not possible (Vovk, 2012; Lei and Wasserman, 2014). Therefore, a variety of works have been developed to approximate it as best as possible. The most popular of these is probably the Conformalized Quantile Regression (CQR) procedure proposed in Romano et al. (2019). CQR follows the conformal methodology to correct the coverage obtained by estimating the quantiles through a quantile regression procedure. Since a quantile regression procedure cannot guarantee the desired coverage level in finite samples, CQR adjusts the interval bounds generated by this procedure. The intervals are enlarged or reduced if the empirical marginal coverage in a calibration set is found to be smaller or larger than the target level, respectively. Although there is no theoretical result related to conditional coverage, as the correction is performed on these estimated conditional quantiles, it is expected that heteroscedasticity is captured with much better quality than with the traditional conformal approach.

The CQR methodology for an objective miscoverage rate of α is describe in Algorithm 3.

Theorem 2 (Romano et al. (2019))**.** Let $\left(\boldsymbol{x}_i, y_i\right)_{i=1}^{n+1}$ be exchangeable. Applying CQR $\left(\boldsymbol{x}_i, y_i\right)_{i=1}^{n}$ produces a prediction interval $\widehat{C}_{\alpha,n+1}$ such that:

$$\mathbb{P}\left(y_{n+1} \in \widehat{C}_{\alpha,n+1}\right) \geq 1-\alpha.$$

Algorithm 3 Conformalized Quantile Regression

Require: $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, significance level α, quantile regression algorithm $\mathscr{A}$

1: Randomly split the n known points into two disjoint sets: training Tr and calibration Cal.
2: Train the quantile regression algorithm $\mathscr{A}$ using the data from the training set Tr and obtain a first approximation of $l_{\alpha,i}$ and $u_{\alpha,i}$, $\widehat{l}_{\alpha,i}$ and $\widehat{u}_{\alpha,i}$, $i \in \text{Cal} \cup \{n+1\}$
3: Compute the conformity scores $\mathscr{S}$

$$\mathscr{S} = \{\mathscr{S}_i : i \in \text{Cal}\} \cup \{+\infty\}$$

where $\mathscr{S}_i = \max\left\{y_i - \widehat{u}_{\alpha,i}, \widehat{l}_{\alpha,i} - y_i\right\}$
4: Compute the $(1-\alpha)$ quantile of the conformity scores, denoted as $Q_{1-\alpha}(\mathscr{S})$.
5: Construct the conformal prediction interval for observation $n+1$ as:

$$\widehat{C}_{\alpha,n+1} = \left[\widehat{l}_{\alpha,n+1} - Q_{1-\alpha}(\mathscr{S}), \widehat{u}_{\alpha,n+1} + Q_{1-\alpha}(\mathscr{S})\right]$$

Ensure: $\widehat{C}_{\alpha,n+1}$ prediction interval of level $1-\alpha$ for the observation $n+1$.

Moreover, if the conformity scores $\{S_i\}_{i\in\text{Cal}\cup\{+\infty\}}$ are almost surely distinct, then the prediction interval is nearly perfectly calibrated:

$$\mathbb{P}\left(y_{n+1} \in \widehat{C}_{\alpha,n+1}\right) \leq 1 - \alpha + \frac{1}{\#\text{Cal}+1}.$$

Given that the computational complexity is minimal once the quantile regression model is trained, and valid intervals are ensured on finite and exchangeable samples, CQR is considered the standard approach in Conformal Prediction for regression tasks. Therefore, it will serve as the foundation for the conformalizations performed through ACI and WACI (Sections 4.2.3 and 4.3.2). For more details of the CQR algorithm Romano et al. (2019) is recommended.

Adaptive Conformal Inference (ACI)

CQR or any other conformal algorithm following the presented scheme (Algorithm 3) depends on the condition of exchangeability among observations. In time series, which are the problems of interest, this condition is not fulfilled. Removing the condition of exchangeability while maintaining the validity property of the intervals has been one of the primary research objectives in the field. One such work is the Adaptive Conformal Inference (ACI) method proposed by Gibbs and Candès (2021).

The application of ACI over the CQR procedure with α^* as the objective miscoverage rate looks

as follows. Let $\alpha_1 = \alpha^*, \text{err}_1 = 0$ and $\gamma > 0$.

$$\begin{cases} \alpha_{t+1} & = \alpha_t + \gamma(\alpha^* - \text{err}_t) \\ \text{err}_t & = \begin{cases} 1 & \text{if } y_t \notin \widehat{C}_{\alpha^*,t} \\ 0 & \text{otherwise} \end{cases} \\ \widehat{C}_{\alpha^*,t+1} & = \left[\widehat{l}_{\alpha^*,t+1} - Q_{1-\alpha_{t+1}}(\mathscr{S}), \widehat{u}_{\alpha^*,t+1} + Q_{1-\alpha_{t+1}}(\mathscr{S})\right] \end{cases}$$

It is a CQR procedure where the quantile used to make the correction is not necessarily that of the target coverage. It is taken adaptive depending on whether too large or too small intervals are being considered. The speed of adaptation is determined by the parameter γ. The following result can be derived:

Theorem 3 (Gibbs and Candès (2021))**.** With probability one it follows that for all $T \in \mathbb{N}$,

$$\left| \frac{1}{T} \sum_{t=1}^{T} \text{err}_t - \alpha^* \right| \leq \frac{\max\{\alpha_1, 1-\alpha_1\} + \gamma}{T\gamma}.$$

In particular,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \text{err}_t = \alpha^*.$$

In other words, there is asymptotic marginal coverage.

For more details on the ACI algorithm Gibbs and Candès (2021) and Zaffran et al. (2022) are recommended.

4.3 Our proposal

Previous research has focused on the issue of providing valid and efficient prediction intervals for individual values. A comprehensive analysis of the coverage of prediction bands, as well as of the interval lengths associated, based on this individual approach reveals significant issues which were described in Section 4.1. The first is that the length of the intervals should vary depending on the difficulty of the observation to be predicted. That is, there should be an increasing relationship between the error of the point forecasting model and the length of the proposed interval.

One of the main works leveraging point predictors to build prediction intervals, as in the context of this study, is the QRA method described in Section 4.2.2. However, its approach does not account for the desired property of adapting to heteroscedasticity. In this thesis, a model inspired by QRA is proposed, but explicitly designed to incorporate the heteroscedasticity. This design enabled to simultaneously capture the aleatoric uncertainty associated with the event

and the epistemic uncertainty of the predictors, ensuring that the estimated uncertainty reflects the underlying complexity of the prediction task.

4.3.1 Heteroscedastic Quantile Regression (HQR)

The QRA model expresses the quantile of interest as a linear combination of point predictors of the mean. The effectiveness shown by this model manifests that the information given by different predictors of the event of interest provides information when quantifying the associated uncertainty.

It is clear that having different predictors of the expected value can provide information on the safety of the prediction: in very common situations for the model, i.e., in areas where the space of regressor variables is highly explored, all forecasters are likely to obtain very similar predictions. However, in the more unfamiliar situations, which generally correspond to unexplored areas where models have to extrapolate, the forecasts start to differ, and, in particular, the error of the models in such cases is generally larger (Figure 4.4). This reflects higher epistemic uncertainty, as it stems from a lack of knowledge or a lack of data in some of these regions. Additionally, situations of high aleatoric uncertainty, where inherent randomness in the data-generating process dominates, can also lead to divergence in the predictions of the mean. In these cases, even with well-trained models, the variability in the predictions reflects the fundamental unpredictability of the process. In other words, both epistemic and aleatoric uncertainty contribute to the observed differences in the forecasts, which is a factor that should be taken into account when building prediction intervals.

A good indicator of the level of exploration of the explanatory features space is a dispersion measure of the prediction of the different models. Suppose, following the model of Nowotarski and Weron (2015), that for each time t M different forecasts $\widehat{y}_{t,1}, \widehat{y}_{t,2}, \ldots, \widehat{y}_{t,M}$ intended to predict y_t are available. Thus, denoting by $\overline{\widehat{y}}_t = \frac{1}{M}\sum_{i=1}^{M}\widehat{y}_{t,i}$ and $s^2_{\widehat{y}_t} = \frac{1}{M}\sum_{i=1}^{M}\left(\widehat{y}_{t,i} - \overline{\widehat{y}}_t\right)^2$ the following quantile regression model is proposed:

$$q_\beta(Y_t|\widehat{\boldsymbol{y}}_t) = \lambda_0(\beta) + \lambda_1(\beta)\overline{\widehat{y}}_t + \lambda_2(\beta)s_{\widehat{y}_t} + \varepsilon_t(\beta),\ \mathbb{E}\left[\varepsilon_t(\beta)\right] = 0, \tag{4.5}$$

where the parameters are obtained by minimizing the pinball loss (equation (4.2.2)) and vary over time using a rolling window procedure in the same way as in equation (4.3). In particular, for the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantiles of interest,

$$\begin{cases} \widehat{q}_{\frac{\alpha}{2}}(Y_{T+1}|\widehat{\boldsymbol{y}}_{T+1}) & = \widehat{\lambda}_0(\frac{\alpha}{2}) + \widehat{\lambda}_1(\frac{\alpha}{2})\overline{\widehat{y}}_{T+1} + \widehat{\lambda}_2(\frac{\alpha}{2})s_{\widehat{y}_{T+1}} \\ \widehat{q}_{1-\frac{\alpha}{2}}(Y_{T+1}|\widehat{\boldsymbol{y}}_{T+1}) & = \widehat{\lambda}_0(1-\frac{\alpha}{2}) + \widehat{\lambda}_1(1-\frac{\alpha}{2})\overline{\widehat{y}}_{T+1} + \widehat{\lambda}_2(1-\frac{\alpha}{2})s_{\widehat{y}_{T+1}} \end{cases}$$

Intuitively, high values of the $\lambda_2(\beta)$ parameter would be expected for quantiles further away from the median with a positive sign for quantiles greater than 0.5 and a negative sign for

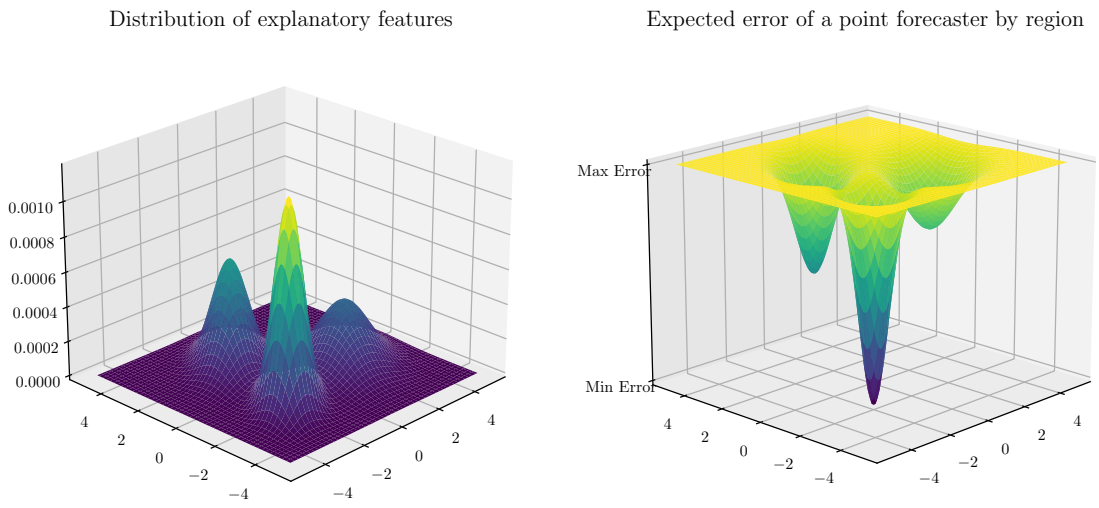

Figure 4.4: The joint distribution of two explanatory features is shown on the left. On the right, the expected error for a predictive model is plotted as a function of the two features. One would expect to have a higher error in the unexplored areas of the space, while a lower error would be expected in the very common areas. The plot is for guidance as the model could have good extrapolation properties in some situations.

quantiles less than 0.5. Similarly, smaller values of $\lambda_2(\beta)$ would be found for quantiles close to the median. If this behaviour occurs, then the relationship between the length of the interval and the error that is desired would be accomplished (Appendix D).

Note that equation (4.5) is actually an extension of the QRA model defined in (4.3). In the case of the QRA model, what is being done is to estimate the mean through a weighted average, which results in different values of the coefficients $\lambda_1,\ldots,\lambda_M$. That is, the QRA model is a model of the type

$$q_\beta(Y_t|\widehat{\boldsymbol{y}}_t)=\lambda_0(\beta)+\lambda_1(\beta)\overline{\widehat{y}}_t+\varepsilon_t(\beta),\ \mathbb{E}\left[\varepsilon_t(\beta)\right]=0,$$

where the estimation $\overline{\widehat{y}}_t$ is not done with equal weights. In that sense, the model is being extended with a further component that refers to a first assessment of the level of uncertainty that exists. Because this extension is directly related to the heteroscedasticity of the predictions, the model has been named Heteroscedastic Quantile Regression (HQR).

To evaluate the significance of mean estimation (whether with equal weights or not) in the experimental section, the model denoted as HQR-W (Weighted Heteroscedastic Quantile Regression),

whose expression is given by

$$\begin{aligned} \widehat{q}_{\beta}(Y_t|\boldsymbol{y}_t) &= \widehat{\lambda}_0(\beta) + \widehat{\lambda}_1(\beta)\widehat{y}_{t,1} + \widehat{\lambda}_2(\beta)\widehat{y}_{t,2} + \dots \\ &\quad + \widehat{\lambda}_M(\beta)\widehat{y}_{t,m} + \widehat{\lambda}_{M+1}(\beta)s_{\widehat{y}_t} + \varepsilon_t(\beta), \\ \mathbb{E}\big[\varepsilon_t(\beta)\big] &= 0, \end{aligned} \tag{4.6}$$

will also be considered. This assessment is particularly important because the number of variables included in the models can differ significantly depending on whether the approximation follows (4.5) or (4.6). While techniques such as L1 regularization (Uniejewski and Weron, 2021) or similar approaches could be employed to address this potential issue, it is essential to first determine whether such measures are necessary at all.

4.3.2 Width-Adaptive Conformal Inference

The second property to achieve is maintaining the same level of confidence in the coverage level regardless of the prediction's difficulty. In other words, the coverage should remain independent of the complexity of the situation, as is the case for the true quantiles (Feldman et al., 2021). When properly estimated, the interval length serves as an indicator of prediction difficulty, reflecting both aleatoric and epistemic uncertainty. In this case, the uncertainty is captured through a quantile regression process, such as HQR. Let's denote this interval at time $T+1$ by $\widehat{C}_{\alpha,T+1}$. To ensure the desired property, a second interval is built, $\widehat{C}^{c}_{\alpha,T+1}$, by modifying $\widehat{C}_{\alpha,T+1}$. This adjusted interval is designed to satisfy property (4.2). From now on, the first initial interval $\widehat{C}_{\alpha,T+1}$ is called the unconformalized interval and the second one $\widehat{C}^{c}_{\alpha,T+1}$ the conformalized interval.

Since the data in question is of a time series nature, the ACI method will be modified to apply a different α as a function of time, like the original method, and also as a function of the length of the interval. Given a range of possible interval lengths from the unconformalized interval, the objective is to partition this space into smaller sub-intervals, with each sub-interval receiving a distinct correction. In other words, there is a different correction depending on the unconformalized interval length. This approach enables the differentiation of varying levels of uncertainty in the data. For instance, in Electricity Price Forecasting, two distinct regimes often emerge: one in which prices are high and the share of renewables in the energy mix is relatively small, and another in which prices are low yet uncertainty grows due to a higher proportion of renewables. The quantile regression model generating the unconformalized interval may exhibit different behaviours in these two states or may not differentiate between them when it should. In any case, it is appropriate to apply different corrections for each state. This methodology allows for the unified treatment of these states within a single framework, while also being suitable for time series data.

Let $\mathscr{S}$ be the conformity scores (Section 4.2.3). Given a step $\delta \in \mathbb{R}^{+}$, the 1-d grid $\boldsymbol{L}$ is defined as $\boldsymbol{L} = (L_{\min}, L_{\min}+\delta, L_{\min}+2\delta, \dots, L_{\max})$ whose elements belong in $\mathbb{R}$. Let α^{*} be the objective

miscoverage rate. Let's denote the element in position i of a vector $\boldsymbol{v}$ by $\boldsymbol{v}[i]$, the p power of $\boldsymbol{v}$ as the p power of each one of the elements of $\boldsymbol{v}$ and the absolute value of $\boldsymbol{v}$ as the absolute values of each one of the elements of $\boldsymbol{v}$. The application of WACI (Width-Adaptive Conformal Inference) over the CQR procedure looks as follows. Let $\boldsymbol{\alpha}_1 = (\alpha^*, \alpha^*, \ldots, \alpha^*)$ with the same dimension as $\boldsymbol{L}$, $\text{err}_1 = 0, \gamma, \sigma > 0$.

$$
\begin{cases}
\boldsymbol{\alpha}_{t+1} & = \boldsymbol{\alpha}_t + \gamma \boldsymbol{w}_t(\alpha^* - \text{err}_t) \\
\text{err}_t & = \begin{cases} 1 & \text{if } y_t \notin \widehat{C}^c_{\alpha^*,t} \\ 0 & \text{otherwise} \end{cases} \\
\textbf{dist}_t & = |\boldsymbol{L} - |\widehat{C}_{\alpha,t}|| \\
\boldsymbol{w}_t & = \dfrac{\exp\left(\frac{-\textbf{dist}_t^2}{2\sigma^2}\right)}{\max\left\{\exp\left(\frac{-\textbf{dist}_t^2}{2\sigma^2}\right)\right\}} \\
\text{idx}_{t+1} & = \text{argmin}_i \left\{\boldsymbol{L}[i] - |\widehat{C}_{\alpha,t+1}|\right\} \\
\tilde{\alpha}_{t+1} & = \boldsymbol{\alpha}_{t+1}[\text{idx}_{t+1}] \\
\widehat{C}^c_{\alpha^*,t+1} & = \left[\widehat{l}^c_{\alpha_t^*}, \widehat{u}^c_{\alpha_t^*}\right] = \left[\widehat{l}_{\alpha^*,t+1} - Q_{1-\tilde{\alpha}_{t+1}}(\mathscr{S}), \widehat{u}_{\alpha^*,t+1} + Q_{1-\tilde{\alpha}_{t+1}}(\mathscr{S})\right]
\end{cases}
\tag{4.7}
$$

The first difference that can be seen with the ACI method is that in this case there is not a single α_t in each iteration, but a vector $\boldsymbol{\alpha}_t$. This is done in order to be able to differentiate the real scalar $\tilde{\alpha}_t$ that will actually be used in that iteration, which will depend on the length of the unconformalized interval. That is, each element of the vector is associated with a different length of the initial interval. The possible unconformalized interval lengths considered are set through the 1-d grid $\boldsymbol{L}$. Thus, $\boldsymbol{\alpha}_t[i]$ is the $\tilde{\alpha}$ to be used when the length of the initial interval of the observation at time t is $\boldsymbol{L}[i]$ (or $\boldsymbol{L}[i]$ is the closest of all those considered in $\boldsymbol{L}$). The update of $\boldsymbol{\alpha}_t$ is done in the same way as in ACI. However, as the conformal correction is being done as a function of interval length, only the positions associated with that interval length (and close to it) are updated. To do this, the weight vector $\boldsymbol{w}_t$ is constructed through a Gaussian kernel, so a new parameter σ related to the amplitude of the kernel effect is introduced. The difference between the ACI and WACI methods throughout iterations is shown in Figure 4.5.

The upper left graph in Figure 4.5 illustrates the first iteration of both methods, where they coincide as both start with the target α. In the next iteration (upper right graph), the methods begin to diverge, although very slightly. The ACI method adjusts α for all possible interval lengths, whereas the WACI method only modifies α for unconformalized interval lengths close to those of the previous observation. The ACI method correction will always remain constant, displaying a horizontal line since it does not differentiate between interval lengths. On the contrary, WACI exhibits variations, using significantly different alphas at "close" interval lengths. The lower graphs correspond to subsequent iterations. For example, in the bottom right graph, for the next iteration, if the unconformalized interval length is around 30 or 60, the correction applied is actually bigger (in the sense that $\tilde{\alpha} > \alpha^*$) compared to the standard correction that CQR would use. However, if the unconformalized interval length is around 40, the correction is smaller. Such distinctions cannot be made by ACI.

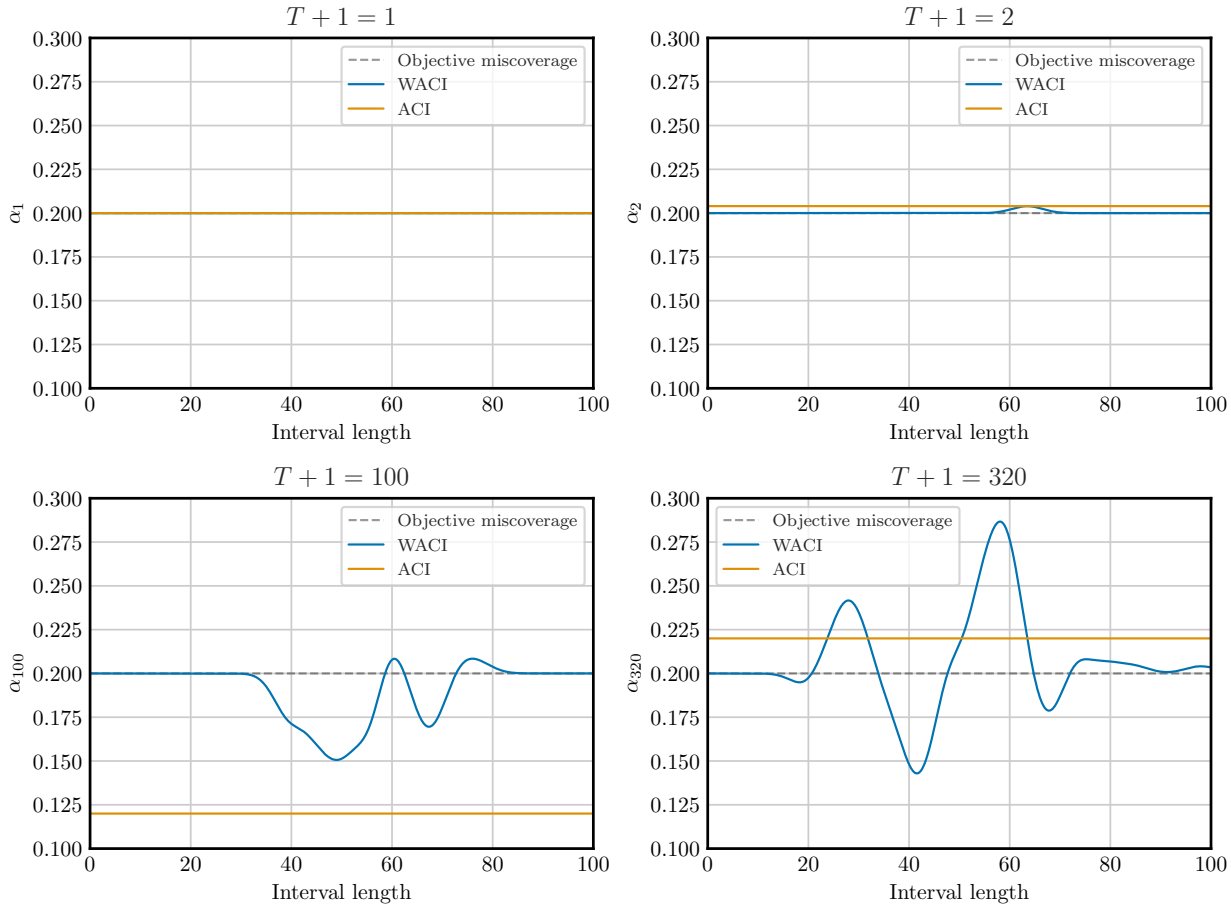

Figure 4.5: Evolution of α_t in the ACI (orange line) and WACI (blue line) methods. The α used in each iteration per interval length is shown.

Weighting Scheme Considerations

In the Algorithm (4.7), an exponential decay function of distance has been chosen for the weighting scheme. However, alternative weighting schemes could be considered. For example, a scheme with fixed weights based on the position in the vector could also be used, such as one where the weights of each interval follow a geometric progression.

That is, let $\boldsymbol{L} = (L_{\min}, L_{\min}+\delta, L_{\min}+2\delta, \ldots, L_{\max}) \in \mathbb{R}^*$ and $\mathscr{L}(i_t) = [\boldsymbol{L}[i_t], \boldsymbol{L}[i_t+1])$ with i_t the index of the interval length for the sample of instant t. Then,

$$\boldsymbol{w}_t[j] = \lambda^{|i_t - j|}, \quad |C_{\alpha,t}| \in \mathscr{L}(i_t) \tag{4.8}$$

Figure 4.6 shows the difference between the two proposed weighting schemes. Despite their

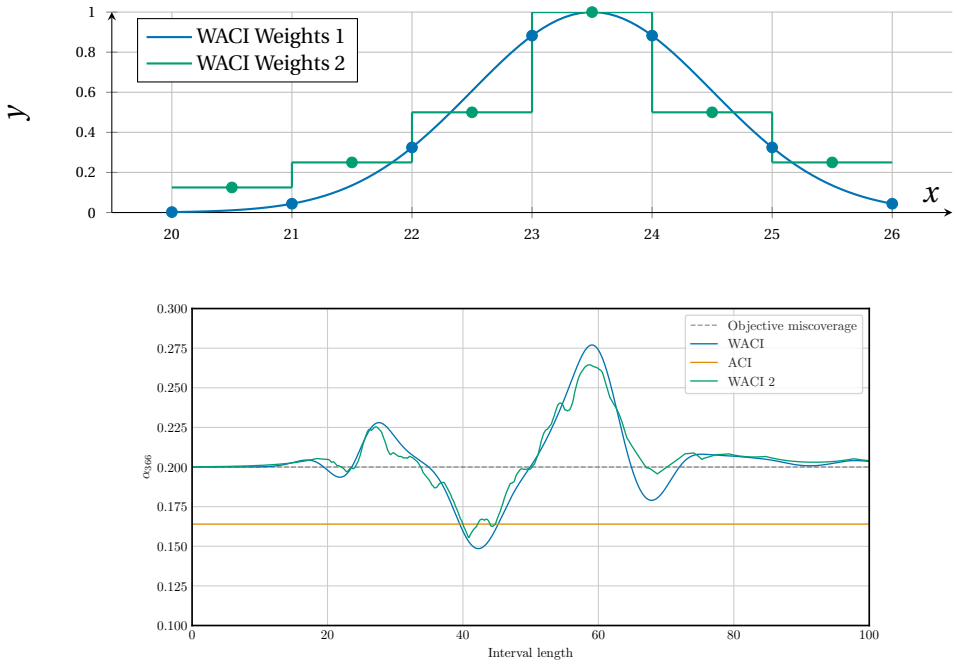

Figure 4.6: **(Top)** Comparing the different weight schemes for the WACI algorithm. The exponential decay weight is shown before scaling. **(Bottom)** The behaviour of the two schemes can be very similar in practice.

differences, by selecting the parameters σ and λ in a certain way, the behaviour of both is very similar.

If the weighting scheme (4.8) is considered, asymptotic conditional coverage can be proved with respect to each of the intervals considered in the grid **L**.

Theorem 4. Let's assume there exists $\nu \in \mathbb{N}$ such that $\boldsymbol{\alpha}_t[i] \in [-\nu, 1+\nu]$ for all $i = 1, \ldots, n$ and $t \in \mathbb{N}$. Let $i \in \{1, ..., n\}$ such that there is an infinite number of $t \in \mathscr{L}(i)$. If $T \longrightarrow \infty$ and the weighting scheme of (4.8) is considered, then

$$\mathbb{P}\left(y_{T+1} \in \widehat{C}^{c}_{\alpha^*, T+1} \,\middle|\, \left|\widehat{C}_{\alpha^*, T+1}\right| \in \mathscr{L}(i)\right) \underset{T \longrightarrow \infty}{\longrightarrow} 1-\alpha^*,$$

where α^* is the objective miscoverage rate and $|\widehat{C}_{\alpha^*, T+1}|$ is the length of the first interval produced at time step $T+1$.

Proof. The equation of the process is given by

$$\boldsymbol{\alpha}_{T+1} = \boldsymbol{\alpha}_T + \gamma \boldsymbol{w}_T(\alpha^* - \mathrm{err}_T).$$

Expanding the recursion,

$$\boldsymbol{\alpha}_{T+1} = \boldsymbol{\alpha}_1 + \sum_{t=1}^{T} \gamma \boldsymbol{w}_t(\alpha^* - \mathrm{err}_t).$$

In particular, for each position i

$$\boldsymbol{\alpha}_{T+1}[i]-\boldsymbol{\alpha}_1[i]=\sum_{t=1}^{T}\gamma \boldsymbol{w}_t[i](\alpha^*-\text{err}_t).$$

Consider the set of indices of instants in whose iteration the length of the unconformalized interval belonged to the interval grid j. That is,

$$\mathscr{I}_j=\left\{t : \left|\widehat{C}_{\alpha^*,t}\right| \in \mathscr{L}\left(j\right),\, t=1,\ldots,T\right\}.$$

Then, the previous expression can be decomposed based on the weight updated carried out during each iteration as

$$\boldsymbol{\alpha}_{T+1}[i]-\boldsymbol{\alpha}_1[i]=\sum_{j=1}^{n}\sum_{t\in\mathscr{I}_j}\gamma\lambda^{|i-j|}(\alpha^*-\text{err}_t)$$

Denoting by $b_k=\frac{\boldsymbol{\alpha}_{T+1}[k]-\boldsymbol{\alpha}_1[k]}{\gamma}$ and $c_k=\sum_{t\in\mathscr{I}_k}(\alpha^*-\text{err}_t)$ for $k=1,\ldots,n$;

$$b_i=c_i+\sum_{j\neq i}\lambda^{|i-j|}c_j, \text{ for } i=1,\ldots,n.$$

By construction, the following system of equations is obtained:

$$\underbrace{\begin{pmatrix} b_1\\ b_2\\ \vdots\\ b_n\end{pmatrix}}_{\boldsymbol{b}}=\underbrace{\begin{pmatrix} 1 & \lambda^{-1} & \lambda^{-2} & \ldots & \lambda^{-(n-1)}\\ \lambda^{-1} & 1 & \lambda^{-1} & \ldots & \lambda^{-(n-2)}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \lambda^{-(n-1)} & \lambda^{-(n-2)} & \lambda^{-(n-3)} & \ldots & 1\end{pmatrix}}_{\Lambda}\underbrace{\begin{pmatrix} c_1\\ c_2\\ \vdots\\ c_n\end{pmatrix}}_{\boldsymbol{c}}$$

The matrix Λ is a Toeplitz matrix equivalent to the correlation matrix of a Markov-1 signal. As discussed in Britanak et al. (2007), the inverse of Λ exists (and it is known) and, therefore,

$$\boldsymbol{c}=\Lambda^{-1}\boldsymbol{b}.$$

Let $i\in\{1,\ldots,n\}$ such that $\{t\in\mathscr{I}_i\cap\mathbb{N}\}$ has an infinite number of elements and let $T_i=\#\{t\in\mathscr{I}_i : t=1,\ldots,T\}$. Then, as Λ^{-1} and $\boldsymbol{b}$ are bounded,

$$\lim_{T\to\infty}\frac{1}{T_i}\|\boldsymbol{c}\|_2=\lim_{T\to\infty}\frac{1}{T_i}\|\Lambda^{-1}\boldsymbol{b}\|_2=0.$$

This implies

$$\lim_{T\to\infty}\frac{1}{T_i}\boldsymbol{c}=\boldsymbol{0}\Longrightarrow\lim_{T\to\infty}\frac{c_i}{T_i}=\lim_{T\to\infty}\frac{1}{T_i}\sum_{t\in\mathscr{I}_i}\text{err}_t-\alpha^*=0,$$

which gives the result

$$\mathbb{P}\left(y_{T+1} \in \widehat{C}^{c}_{\alpha^*,T+1} \,\middle|\, \left|\widehat{C}_{\alpha^*,T+1}\right| \in \mathscr{L}(i)\right) \underset{T\longrightarrow\infty}{\longrightarrow} 1-\alpha^*$$

□

In view of Theorem 4, asymptotic coverage conditional on the difficulty of the prediction is obtained, where that difficulty is measured by the length of the interval of the first quantile regression algorithm used. As a consequence, asymptotic marginal coverage is also achieved, as in the original ACI algorithm.

Although this is not exactly the condition represented by equation (4.2), coverage is still achieved depending on the complexity of the forecast. Therefore, it is important that the unconformalized interval shows the desired relationship between interval length and prediction difficulty (which it is achieved by applying HQR in the first step).

The only assumption made to obtain the result is that the value of $\boldsymbol{\alpha}$ is bounded for every position. Although this is not formally proven, it seems a reasonable feature of the algorithm. If the value of a certain position of $\boldsymbol{\alpha}$ exceeds 1 or falls below 0 during any iterations, it is forced to decrease or increase accordingly, thereby controlling the explosion of that value. The only scenario where no limits exists on a certain position of $\boldsymbol{\alpha}$ is when, after surpassing 1 (from above) or 0 (from below), that position is no longer frequently updated compared to others. In such cases, distant positions might continuously increase or decrease at a faster rate. This behaviour is irrational, as one would expect the algorithm to update different $\boldsymbol{\alpha}$ positions uniformly over iterations, adjusting the values both upward and downward.

Another implicit assumption is that δ is sufficiently small, meaning the distance between grid separator points is not excessively large. This is important because, if the grid intervals are too wide, a scenario may arise where all initial intervals fall within the same grid interval. In such cases, the procedure essentially reduces to applying ACI. When using a distance-based weighting scheme (as in equation (4.7)), choosing a very fine grid generally does not present a problem, aside from potentially increasing the algorithm's computational runtime. However, for position-based weighting schemes (as in equation (4.8)), the decay of the weights along positions must reasonably reflect the distances between observations. Specifically, the weight decay should adjust based on the choice of δ: if δ is small relative to the scale of the problem, the weight decay should be smoother, whereas a larger δ requires a steeper weight decay. In any case, it is recommended using as fine a grid as possible to ensure better granularity and accuracy in the results.

It is emphasized that while HQR and WACI can be applied independently, combining them is essential to ensure that the desired properties are achieved.

4.4 Computational experiments

To evaluate the effectiveness of the proposed WACI-HQR method, the different interval construction schemes are compared. The quantile regression models include the QRA model, the HQR model, and the HQR-W model. Additionally, the conformal post-processing methods, ACI and WACI, are applied to these models. This comparison evaluates the impact of each modelling step on the interval properties, from the initial quantile regression to the final adaptive conformal approach. The weights used for WACI are given in equation (4.7), with similar results observed using the weights in equation (4.8).

4.4.1 Evaluation metrics

Mean empirical coverage The mean empirical coverage is used to measure the validity property: if there are N prediction intervals for observations y_i, $i = 1, \ldots, n$, the empirical coverage on those predictions is defined as

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left(y_i \in \widehat{C}_{\alpha,i}\right).$$

For an objective miscoverage rate of α, the empirical coverage is sought to be as close to $1-\alpha$ as possible.

Mean interval length The efficiency is usually measured through the mean or median interval length. The ACI and WACI adaptive conformal procedures are not constrained in their definition of α, allowing for intervals that are either empty sets or cover all of $\mathbb{R}$. While empty intervals are rare, occurring only when the α value in the corresponding iteration exceeds 1, the case of infinite intervals is more frequent, arising when α is less than 0. To address these situations, it is common practice to use the median interval length, as it is unaffected by infinite intervals. In this work, the mean interval length is computed, replacing infinite intervals with a fixed interval defined by the largest upper bound and smallest lower bound of the base model observed in the training set. This approach effectively penalizes the production of excessive infinite intervals.

Winkler score The Winkler score (Winkler, 1972) is used to measure validity and efficiency together. For each time step t and for a miscoverage rate of α, it is defined as the length of the interval plus a penalty term proportional to how far the prediction is from being in the interval:

$$W_{\alpha,t} = \begin{cases} (\widehat{u}_{\alpha,t} - \widehat{l}_{\alpha,t}) + \frac{2}{\alpha}(\widehat{l}_{\alpha,t} - y_t) & \text{if } y_t < \widehat{l}_{\alpha,t} \\ (\widehat{u}_{\alpha,t} - \widehat{l}_{\alpha,t}) & \text{if } \widehat{l}_{\alpha,t} \leq y_t \leq \widehat{u}_{\alpha,t} \\ (\widehat{u}_{\alpha,t} - \widehat{l}_{\alpha,t}) + \frac{2}{\alpha}(y_t - \widehat{u}_{\alpha,t}) & \text{if } y_t > \widehat{u}_{\alpha,t} \end{cases}$$

Thus, better intervals will have smaller Winkler score. The Winkler score is actually a proper scoring rule (Gneiting and Raftery, 2007) and so, the mean Winkler score over every interval forecast will be also measured.

Pearson's correlation Following Feldman et al. (2021), conditional coverage is assessed by computing Pearson's correlation coefficient between the interval length and the coverage indicator function. The closer to 0, the better, as the indicator function of coverage should be independent of the length when true quantiles are considered.

ILS λ Coverage A variation of the metric ΔILS-Coverage presented in Feldman et al. (2021) is used to check the effectiveness of the post-processing methodologies. To evaluate whether the modifications made to the intervals produced by quantile regression algorithms are truly useful, consider a base quantile regression model with produces a prediction interval $\widehat{C}_i$ and a conformal procedure applied post hoc to this base model which produces the prediction interval $\widehat{C}_i^c$ for observation i, $i = 1,\dots,N$. Let Δ_i denote the difference in interval lengths proposed by the two algorithms for observation i, defined as

$$\Delta_i = \left| |\widehat{C}_i^c| - |\widehat{C}_i| \right|.$$

Now consider the $\lambda \cdot 100\%$ of samples most affected by the conformal procedure, i.e.,

$$\text{ILS} = \left\{ i : \Delta_i \geq q_\lambda \left(\{\Delta_i\}_{i=1}^N \right) \right\},$$

where q_λ denotes the λ-empirical quantile. The ILS λ Coverage metric is then defined as

$$\text{ILS } \lambda \text{ Coverage} = \frac{1}{|\text{ILS}|} \sum_{i \in \text{ILS}} \left| \mathbb{1}\left(y_i \in \widehat{C}_i^c \right) - (1-\alpha) \right|.$$

The objective is to determine whether the intervals that have been modified to a greater extent indeed achieve the desired level of coverage.[15] $\lambda = 0.10$ is used in this thesis.

Spearman's correlation The Spearman's correlation coefficient between the mean absolute error of the mean prediction of the point forecasts and the interval length will be computed to assess the strength of the relationship between prediction difficulty and interval length. Since this relationship does not need to be linear, Spearman's rank correlation coefficient is preferred over the traditional linear correlation coefficient.

Standard deviation of the interval length The standard deviation of the interval lengths generated by an algorithm is computed to assess the algorithm's ability to distinguish between varying uncertainty contexts. This metric is closely related to the previous one, as different levels of prediction difficulty are associated with corresponding variations in the interval lengths.

[15] The difference with Feldman et al. (2021) is that it evaluates only the intervals that have increased in size while ignoring those that have been reduced by the conformal procedure.It is proposed to consider both types of modifications, since reducing the size of the interval when it does not apply should also be penalized.

MCD λ Let $1-\alpha$ denote the target coverage level, and let's assume there are N observations for which a prediction interval $\left\{\widehat{C}_{\alpha,i}\right\}_{i=1}^{N}$ has been built. To assess the property described in equation (4.2), the mean coverage deviation (MCD) is defined as follows: the data is divided into $K=\frac{100}{\lambda}$ subsets $\{\mathscr{G}_k\}_{k=1}^{K}$, where each subset contains approximately λ% of the data. The subsets are defined by the empirical quantiles of $\left\{\left|\widehat{C}_{\alpha,i}\right|\right\}_{i=1}^{N}$, such that:

$$\mathscr{G}_k=\left\{i : q_{\frac{k-1}{K}}\left(\left|\widehat{C}_{\alpha,i}\right|\right)\leq\left|\widehat{C}_{\alpha,i}\right| < q_{\frac{k}{K}}\left(\left|\widehat{C}_{\alpha,i}\right|\right)\right\},\quad k=1,\ldots,K.$$

For each subset $\mathscr{G}_k$, the deviation between its mean empirical coverage and the objective coverage is computed

$$D_k=\left|\frac{1}{|\mathscr{G}_k|}\sum_{i\in\mathscr{G}_k}\mathbb{1}\left(y_i\in\widehat{C}_{\alpha,i}\right)-(1-\alpha)\right|.$$

Finally, the mean coverage deviation is defined as

$$\text{MCD}=\frac{1}{K}\sum_{k=1}^{K}D_k.$$

A smaller value of MCD indicates better alignment of empirical coverage with the target coverage, regardless of interval length. $\lambda=5$ is going to be used.

To the best of our knowledge two of these metrics have not been utilized in similar studies: Spearman's correlation coefficient and the MCD. These metrics provide valuable insights into the performance of the various methods.

4.4.2 A synthetic example

A synthetic example is designed to evaluate the effectiveness of WACI and provide insights into its behaviour. Consider a time series y_t generated from a normal distribution $N(\mu,\sigma_t)$, where the standard deviation σ_t alternates between two states representing different levels of uncertainty. Specifically, the process alternates between:

- A high-uncertainty state where $\sigma_t=\sigma_1=7$,
- A low-uncertainty state where $\sigma_t=\sigma_2=2$.

The transition between these two states is governed by a probabilistic mechanism. The process begins in the high-uncertainty state ($\sigma_t=\sigma_1$), and at each time step, the probability of transitioning to the other state increases incrementally by 0.0001. Once a transition occurs, the probability resets to zero, ensuring alternating states. A binary indicator variable δ_t is used to represent the current state: $\delta_t=0$ corresponds to the high-uncertainty state, while $\delta_t=1$ represents the low-uncertainty state.

For simplicity, the mean $\mu = 100$ is assumed to be known. To compute the length of the interval it is simulated that a sample of size 10 is drawn from the distribution at each time step t. However, instead of estimating the standard deviation from a sample, a deterministic relationship is used for $\widehat{\sigma}_t$, introducing smooth time-dependent fluctuations that reflect transitions between overcoverage and undercoverage. This is given by:

$$\widehat{\sigma}_t = \begin{cases} \sigma_1 + 2 \cdot \sin(0.001 \cdot t) & \text{if } \delta_t = 0, \\ \sigma_2 + \cos(0.005 \cdot t) & \text{if } \delta_t = 1. \end{cases}$$

and the unconformalized interval is computed based on the following relationship:

$$\widehat{l}_{\alpha,t} = \mu - T_{1-\frac{\alpha}{2},9} \cdot \widehat{\sigma}_t \cdot \sqrt{1 + \frac{1}{10}}, \quad \widehat{u}_{\alpha,t} = \mu + T_{1-\frac{\alpha}{2},9} \cdot \widehat{\sigma}_t \cdot \sqrt{1 + \frac{1}{10}},$$

where $T_{1-\frac{\alpha}{2},9}$ is the critical value of the t-distribution for a two-tailed test with significance level α and 9 degrees of freedom.

This design ensures smooth changes in coverage, making the example particularly well-suited to evaluating ACI-based conformalization methods.

Figure 4.7 illustrates 1000 time steps of this process, showing the generated observations, true intervals, and unconformalized intervals.

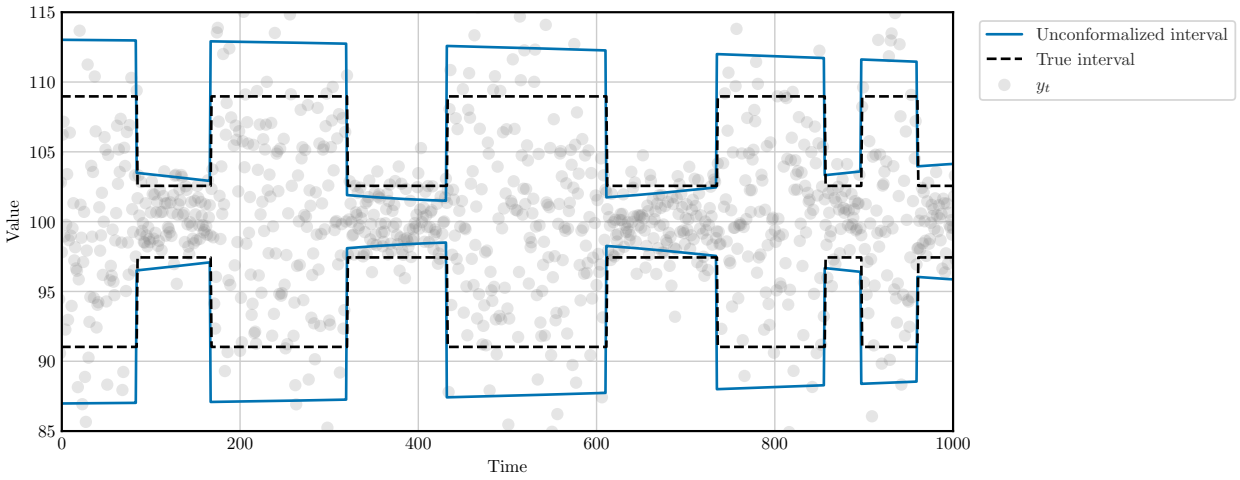

Figure 4.7: Simulated time series data, true intervals, and unconformalized intervals over 1000 time steps.

The goal is to conformalize the intervals such that they remain valid and efficient across both uncertainty states, while achieving optimal performance according to the evaluation metrics introduced in Section 4.4.1. For this experiment, hyperparameters were fixed ($\gamma = 0.01$ for ACI and WACI, $\sigma = 1$ for WACI), and no optimization was performed. To account for the randomness

inherent in the process, the experiment was repeated 100 times with different seeds, enabling standard deviation estimates for all metrics. Each run simulated a time series of length 10000 with $\alpha = 0.2$.

Results and discussion

Tables 4.1 and 4.2 present the values of the different evaluation metrics previously described for the two uncertainty states separately.

Method	Mean Empirical Coverage	Average Length	Winkler Score	Pearson Correlation	ILS 0.10	MCD 5
Initial	85.13 (0.68)	21.14 (0.23)	26.05 (0.22)	0.25 (0.01)	–	9.17 (0.52)
ACI	83.24 (0.51)	19.90 (0.17)	25.48 (0.24)	0.22 (0.01)	3.30 (0.55)	7.25 (0.50)
WACI	81.08 (0.23)	18.59 (0.23)	**24.89 (0.27)**	**0.15 (0.01)**	**1.47 (0.30)**	**4.35 (0.37)**

Table 4.1: Mean results of 100 runs of the synthetic experiment for the high uncertainty state samples. The standard deviation of the metrics is shown in brackets.

Method	Mean Empirical Coverage	Average Length	Winkler Score	Pearson Correlation	ILS 0.10	MCD 5
Initial	79.92 (1.19)	5.88 (0.12)	7.97 (0.10)	0.34 (0.02)	–	14.27 (0.83)
ACI	76.64 (0.59)	5.13 (0.09)	7.57 (0.10)	0.25 (0.02)	3.34 (0.70)	9.87 (0.59)
WACI	80.72 (0.13)	5.35 (0.07)	**7.18 (0.09)**	**0.10 (0.01)**	**0.92 (0.22)**	**4.57 (0.36)**

Table 4.2: Mean results of 100 runs of the synthetic experiment for the low uncertainty state samples. The standard deviation of the metrics is shown in brackets.

A quick comparison of the classical empirical coverage and mean interval length values reveals that WACI is the only method that effectively captures both states. This observation is further supported by the Winkler Score, which is the lowest in both cases. The other metrics provide deeper insights into why WACI is clearly the preferred choice. Notably, WACI does not exhibit the same dependence between interval length and coverage as the other two intervals, as shown in the values of the Pearson's correlation coefficient. Specifically, the large changes observed in ACI do not align closely with the desired coverage levels, unlike WACI. This difference is evident in the ILS 0.10 metric. The MCD metric highlights how WACI achieves the desired outcome by reducing the dependence between interval length and empirical coverage, resulting in intervals with superior overall characteristics. In particular, WACI allows for targeted adjustments at each uncertainty state without the need for incremental fine-tuning, as seen with ACI. This distinction is more clearly illustrated in Figure 4.8.

Figure 4.8 illustrates how, following a change in behaviour, ACI requires several iterations to detect and adjust to the necessary modifications for approximating the true interval. This behaviour is expected since ACI relies solely on information from the previous iteration to guide its corrections. In contrast, WACI adapts immediately to the change, as it is capable of distinguishing between the two behavioural states present in the system. This is because WACI

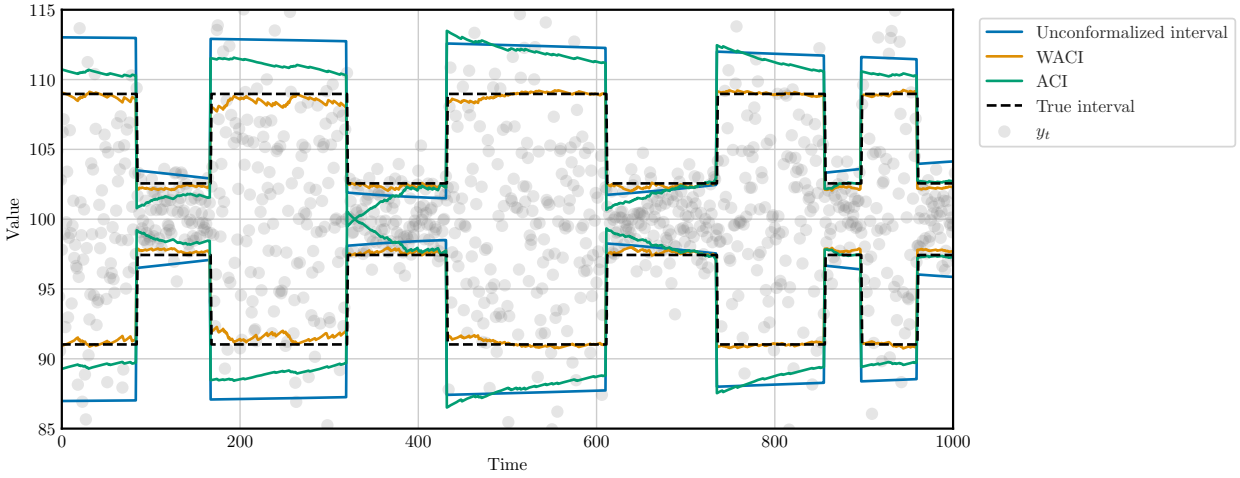

Figure 4.8: Prediction intervals produced by ACI (green) and WACI (orange) over the synthetic example.

references prior iterations where the initial interval length, before conformalization, was similar.

It is essential to analyse the two (or more) states separately when their existence is known. Table 4.3 presents the results of the evaluation metrics when all observations are considered together. At first glance, ACI appears to perform well, achieving results only slightly inferior to WACI in terms of mean empirical coverage and mean interval length. However, as shown in Tables 4.1 and 4.2, this overall performance obscures significant discrepancies: in some situations, ACI exhibits over-coverage, while in others, there is clear under-coverage, leading to the observed average outcomes. Additionally, the other metrics reinforce that these discrepancies are closely tied to the relationship between interval length and coverage. The only metric for which ACI performs well is ILS 0.10, but this result warrants closer examination, particularly to identify which observations are most significantly modified by ACI.

Method	Mean Empirical Coverage	Average Length	Winkler Score	Pearson Correlation	ILS 0.10	MCD 5
Initial	82.51 (0.75)	13.48 (0.44)	19.68 (0.52)	0.17 (0.02)	–	11.13 (0.55)
ACI	79.93 (0.02)	12.49 (0.41)	16.49 (0.52)	0.15 (0.01)	**0.09 (0.07)**	7.89 (0.42)
WACI	80.90 (0.11)	11.95 (0.40)	**16.01 (0.52)**	**0.04 (0.003)**	1.17 (0.17)	**3.68 (0.29)**

Table 4.3: Mean results of 100 runs of the synthetic experiment for every observation. The standard deviation of the metrics is shown in brackets.

4.4.3 Electricity Price Forecasting (EPF)

Electricity Price Forecasting serves as an ideal example for evaluating different techniques in the area of probabilistic forecasting. Producing prediction intervals in this context is of significant interest due to the intricate dynamics of electricity markets, which are increasingly influenced by stochastic factors, such as renewable energy integration. These factors directly impact the strategies of market participants and the formulation of their bids. As a result, the

area of probabilistic forecasting has gained traction, with numerous studies using the electricity market as a benchmark for evaluating their methodologies. For instance, Wisniewski et al. (2020); Kath and Ziel (2021); Zaffran et al. (2022).

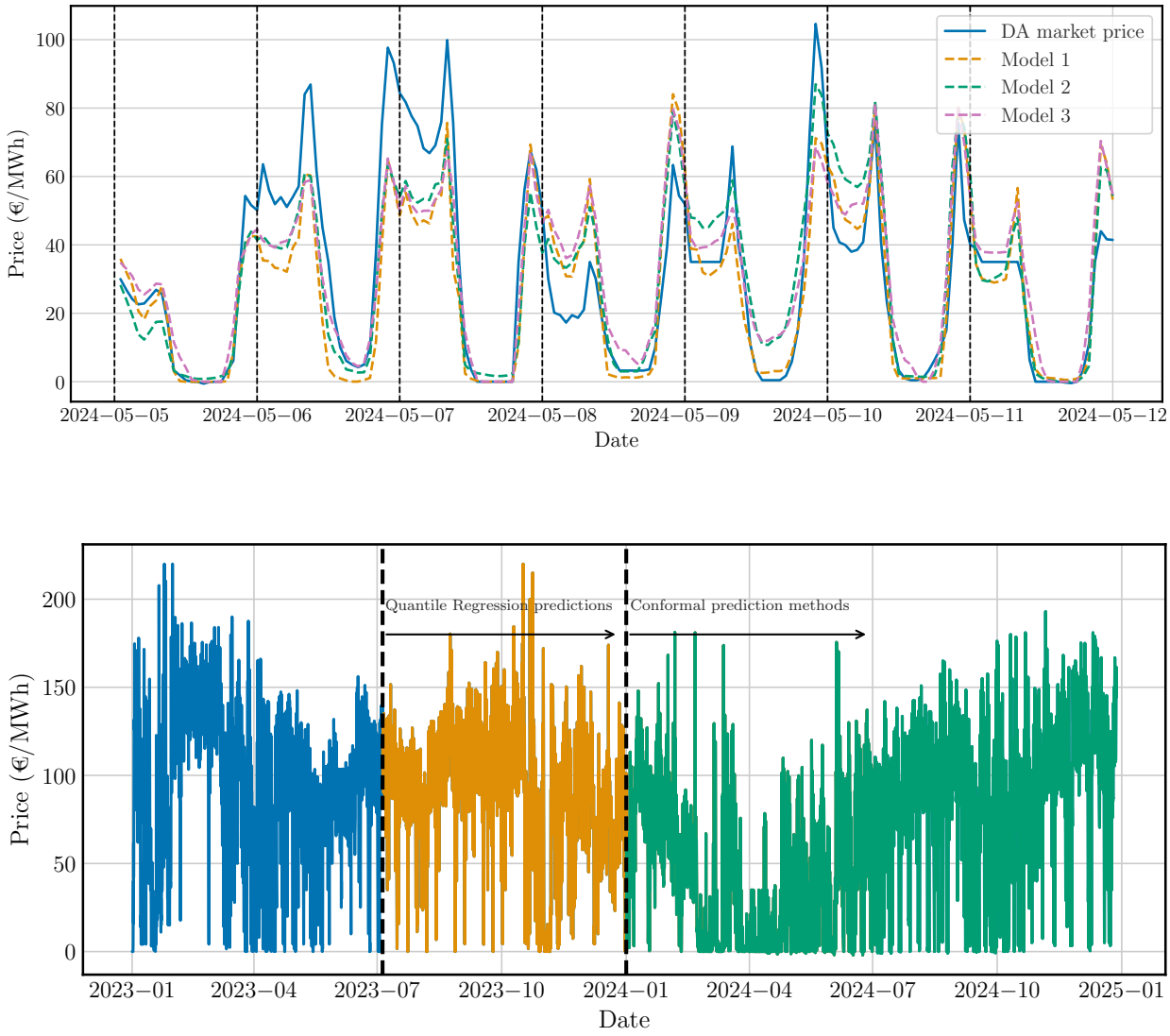

Figure 4.9: (Top) One week example of the point forecasters for the EPF example and **(Bottom)** time series of the Spanish Day-Ahead market.

Data

Data for the Day-Ahead market price in Spain and 3 different one day-ahead point forecasters ($M = 3$, Figure 4.9) is available from 1st of January 2023 to 28th of December 2024. The period from 1st of January 2024 to 28th of December 2024 is considered as test data. This is the window where conformal methods are applied. In order for these methods to be applied, quantile

regression forecasts must be available. These will be obtained from 5th of July 2023 (Figure 4.9). This period is ideal for testing this type of models because moments of great uncertainty can be observed at the same time as very steady phases. The point forecasting models are related to the ones presented in Lago et al. (2021) and the ones presented in the previous chapter.

Methodology

When forecasting the price one day in advance, it must be predicted for the 24 hours of the next day. Thus, there are two ways of proceeding with the quantile regression: consider 24 daily series (one per hour) and a quantile regression model for each one of these series or a single hourly frequency series and therefore a single quantile regression model for all hours. As not many data is available, only the second approximation is considered. In any case, both approximations are valid and it has been shown that, for point forecasting scenarios, results are quite similar (Ziel and Weron, 2018). A training rolling window of 180 days is considered. That is, at the time of predicting the day D, data from days $D-180$ to $D-1$ are considered to train the quantile regression models. For the day $D+1$, the window from $D-179$ to D is considered for training and so on. For those days for which the training window is shorter than 180 days due to data unavailability, all available data will be used. Conformalization is always carried out individually for each hour. For both ACI and WACI $\gamma = 0.02$ is taken, which seems reasonable in view of previous studies (Zaffran et al., 2022) and for the WACI approach $\sigma = 3$, which also seems appropriate given the price scale. For WACI, a grid $\boldsymbol{L}$ based on a 0.1€/MWh step is used. With all these distinctions, 9 possible methodologies will be compared over a test period of almost one year. It is crucial to include at least one year of data to ensure the generalizability of the results (Lago et al., 2021). A full year encompasses special situations such as holidays, demand seasonality, and varying meteorological conditions. This approach prevents the results from being biased or influenced by the exclusion of any particular situation, providing a comprehensive evaluation of the methods. Two possible values of α will be distinguished: $0.2, 0.1$, which correspond to coverage values of 80, 90%, respectively. The calibration window size, as well as the γ and σ values, are hyperparameters that could be optimized to improve performance. However, in this study, such optimization has not been conducted, indicating that better results may be achievable with further tuning. To better understand the impact of σ on the WACI methodology and the differences with ACI, an analysis can be found in Appendix E. In this example, estimations of the standard deviation for each metric have been obtained through the stationary bootstrap procedure from Politis and Romano (1994), considering 1000 bootstrapped samples of size 1000.

Results and discussion

Results for every evaluation metric are shown in Tables 4.4 and 4.5, for $\alpha = 0.20$ and $\alpha = 0.10$, respectively. Metric names are displayed with their abbreviations, for better readability of the tables. On the left side of the tables, the standard metrics of mean empirical coverage, mean interval length and Winkler score are displayed. On the right side, conditional coverage metrics

as well as those related to desired properties are shown.

Methodology	M.E.C.	M.I.L.	W.S.	P.C.	ILS 0.10	S.C.	I.L. Std	MCD 5
QRA	79.68	32.07	49.85	0.16	–	0.10	9.43	7.43
	(2.31)	(1.65)	(2.12)	(0.06)		(0.10)	(1.35)	(1.96)
HQR	80.24	31.15	47.48	**0.04**	–	0.32	12.54	4.83
	(2.04)	(2.06)	(2.67)	**(0.05)**		(0.10)	(0.96)	(1.39)
HQR-W	80.78	30.95	47.11	0.05	–	0.30	12.04	5.55
	(2.01)	(1.90)	(2.52)	(0.05)		(0.10)	(1.17)	(1.44)
QRA (ACI)	79.20	32.20	49.13	0.20	0.60	0.20	12.27	6.91
	(2.12)	(2.35)	(2.50)	(0.04)	(1.28)	(0.10)	(1.95)	(1.40)
HQR (ACI)	79.62	31.77	47.13	0.12	0.40	**0.35**	**14.86**	4.86
	(1.85)	(2.47)	(2.93)	(0.03)	(1.12)	**(0.10)**	**(1.50)**	(1.32)
HQR-W (ACI)	79.50	31.21	46.81	0.14	0.21	0.33	14.17	5.69
	(1.90)	(2.24)	(2.76)	(0.03)	(1.16)	(0.10)	(1.68)	(1.41)
QRA (WACI)	80.67	32.33	48.99	0.14	1.15	0.16	10.03	5.38
	(2.09)	(1.91)	(2.24)	(0.04)	(1.40)	(0.10)	(1.51)	(1.63)
HQR (WACI)	79.90	31.16	47.09	0.08	**0.08**	0.33	13.20	**3.84**
	(2.01)	(2.35)	(2.77)	(0.04)	**(1.21)**	(0.10)	(1.28)	**(1.25)**
HQR-W (WACI)	80.25	31.17	**46.80**	0.08	0.09	0.31	12.52	4.42
	(1.91)	(2.06)	**(2.63)**	(0.04)	(1.12)	(0.10)	(1.44)	(1.24)

Table 4.4: Evaluation metrics for the EPF example for $\alpha = 0.20$. Standard deviations of each metric are shown in brackets.

Methodology	M.E.C.	M.I.L.	W.S.	P.C.	ILS 0.10	S.C.	I.L. Std	MCD 5
QRA	89.10	44.86	63.62	0.14	–	0.08	11.45	4.65
	(1.51)	(1.99)	(2.41)	(0.04)		(0.10)	(1.77)	(1.11)
HQR	89.58	43.11	59.67	**0.06**	–	0.31	15.79	3.64
	(1.23)	(2.63)	(3.20)	**(0.04)**		(0.11)	(1.27)	(1.27)
HQR-W	89.61	42.45	59.75	**0.06**	–	0.29	14.36	3.25
	(1.27)	(2.20)	(2.82)	**(0.04)**		(0.10)	(1.55)	(1.00)
QRA (ACI)	89.50	46.36	63.04	0.16	0.41	0.18	16.50	4.54
	(1.36)	(2.85)	(2.81)	(0.04)	(1.38)	(0.10)	(2.29)	(2.29)
HQR (ACI)	89.84	43.81	59.38	0.11	0.08	**0.34**	**19.53**	3.32
	(1.08)	(3.14)	(3.71)	(0.03)	(1.10)	**(0.10)**	**(2.00)**	(0.85)
HQR-W (ACI)	89.61	44.04	59.59	0.11	0.24	0.31	18.78	3.80
	(1.20)	(2.94)	(3.28)	(0.04)	(1.16)	(0.10)	(2.34)	(0.85)
QRA (WACI)	90.31	46.65	62.98	0.13	0.38	0.13	12.97	3.72
	(1.26)	(2.11)	(2.52)	(0.03)	(1.27)	(0.10)	(1.71)	(1.05)
HQR (WACI)	90.14	43.52	**59.35**	0.07	**0.06**	0.32	17.60	**2.57**
	(1.21)	(2.99)	**(3.47)**	(0.04)	**(1.25)**	(0.11)	(1.70)	**(0.91)**
HQR-W (WACI)	90.03	43.90	59.65	0.07	0.13	0.30	16.90	2.66
	(1.16)	(2.48)	(3.02)	(0.04)	(1.11)	(0.10)	(1.73)	(0.81)

Table 4.5: Evaluation metrics for the EPF example for $\alpha = 0.10$. Standard deviations of each metric are shown in brackets.

In order to have a better understanding of the base behaviour of the models, a plot has been made of the mean empirical coverage against the mean interval length (Figure 4.10).

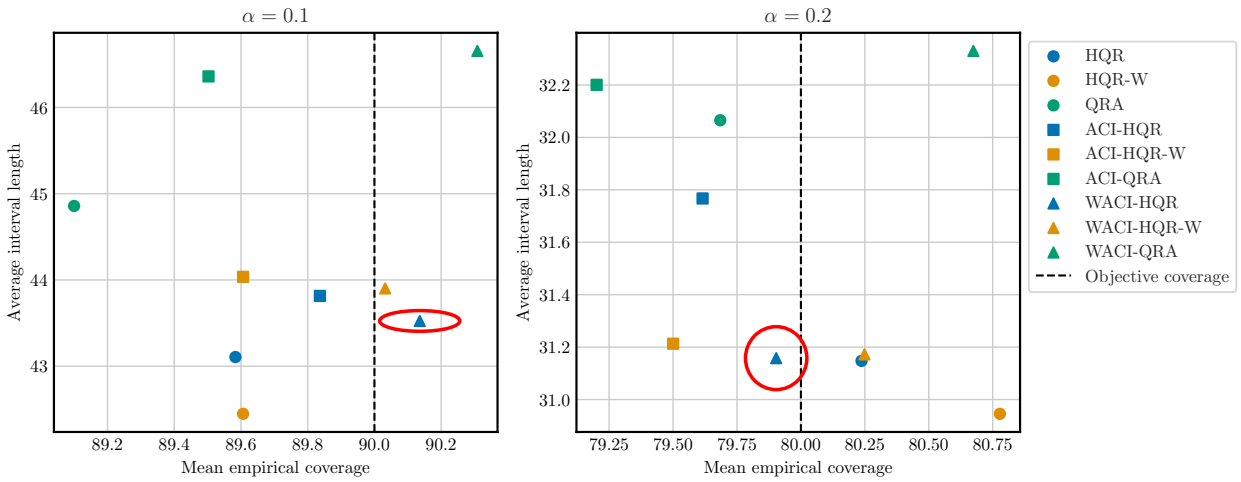

Figure 4.10: Mean empirical coverage vs. mean interval length for both levels of α, 0.10 and 0.20. The colour of each point is determined by the base quantile regression model used. The shape of each point is determined by the applied conformalization. The WACI-HQR combined methodology is highlighted with a red circle.

Comparison between quantile regression models At first glance, it is evident that models based on QRA perform significantly worse than those based on HQR(-W), regardless of the chosen α level. This is reflected in the Winkler score, where the difference is particularly pronounced, indicating poorer overall performance for QRA. Delving deeper into the metrics, several shortcomings of QRA become apparent. First, the Pearson correlation for QRA is notably higher, signalling a weaker level of conditional coverage compared to HQR(-W). Additionally, the relationship between interval length and prediction difficulty, as measured by the Spearman correlation, is considerably weaker for QRA. This suggests that QRA struggles to adjust interval lengths appropriately to reflect varying levels of uncertainty. Further evidence of this shortcoming is found in the lower standard deviation of interval lengths for QRA. This indicates an inability to differentiate between contexts of varying prediction difficulty, which complicates decision-making processes. Finally, while none of the methods are explicitly designed for this purpose, the superior approximation of quantiles achieved by HQR(-W) also benefits the MCD metric. In conclusion, QRA fails to match the quality achieved by HQR(-W) across all evaluated metrics.

Inclusion (or not) of varying weights in the HQR model The differences between HQR and HQR-W are minimal and not statistically significant, suggesting that both models exhibit nearly identical behaviour. This observation holds true even when analysing different conformalization methods applied to the models. Based on the principle of parsimony, this suggests that HQR should be preferred. By selecting the simpler model, one can avoid potential overfitting issues that might arise from the additional complexity introduced by HQR-W.

Differences between ACI and WACI When examining traditional metrics, ACI and WACI produce very similar results and are nearly indistinguishable, though two key observations

stand out. First, WACI consistently achieves better Winkler score values when comparing the same base models and, second, produces valid prediction intervals, except for the HQR (WACI) case with $\alpha = 0.20$ by a very small and reasonable margin. Pearson's correlation reveals a clear difference: when applied to the QRA model, WACI improves conditional coverage, whereas ACI worsens it. For the HQR(-W) base model, ACI considerably degrades this property (with the coefficient being over 0.1), while WACI maintains it at levels comparable to the original model. Further insights emerge from the ILS 0.10 metric, where WACI demonstrates clear superiority, achieving minimal values and bringing interval coverage closer to the target. This robustness gives users confidence, as the methodology performs well even with significant ex-post corrections. Regarding the Spearman correlation and the standard deviation of interval lengths, WACI and ACI preserve these characteristics from the base models, a positive outcome that aligns with the intent behind using HQR. Finally, on the MCD metric, WACI once again outperforms ACI with a notable difference, demonstrating greater independence between the coverage indicator and interval length, which is the idea behind WACI's design. In conclusion, while ACI and WACI often appear similar in their outcomes, WACI consistently demonstrates superior performance across desirable properties.

4.5 Conclusions and future lines of research

In this chapter the problem of obtaining prediction intervals that are built with the intention of assisting in decision making correctly is considered. It has been discussed how the classical measures of validity and efficiency of intervals are not sufficient to be able to use these intervals in an appropriate manner. It is important that the intervals are varied in a way that this variation is directly related to the difficulty of the prediction and that the coverage does not depend on this difficulty, as it is possible to make the mistake of taking decisions with a certainty that does not correspond to the real one. Thus, one forecasting pipeline combining two innovations has been introduced. The first part consist on applying the HQR model, which focuses on the length of the intervals having the appropriate relationship with the difficulty of the prediction, and the second part involves the conformalization of the intervals produced by the quantile regression model through the WACI adaptive conformal process, which seeks uniformity of safety regardless of the difficulty. This is all considered in the context that only different forecasters of the event of interest are known, as this is a typical situation for practitioners. The different improvements provided by these models have been evaluated with two examples: one synthetic example to showcase the potential of WACI and the differences with ACI, and one related with electricity price forecasting, which is not a simple task. The results show how each of the proposed stages produces the desired results, correcting flaws in established models in the literature. Also, the inclusion of individual weights for each predictor has been shown to provide no significant improvement in the estimation of quantiles. This supports the approach of first obtaining the best possible estimate of the mean through techniques like Gaillard et al. (2014); Wintenberger (2017); Adjakossa et al. (2023) and then using that estimate, along with the standard deviation of the predictors, to estimate the quantiles.

As futures lines of work, the HQR model uses only two explanatory variables, which are the

first two (estimated) moments of the distribution to be predicted. However, moments such as skewness or kurtosis may be of interest and could contribute to the estimation of the quantiles, following the ideas set out in Cornish and Fisher (1938). Estimating these moments with so few predictors is not feasible, but if a considerable number of them is available, assessing the improvement by considering higher order moments is of interest. In addition, the variance estimation has not taken into account the individual quality of each of the provided models or the correlation between them. A correct use of this information could lead to better results, although something similar to what happens when combining point prediction models could occur, where the simplest combinations such as the mean perform remarkably well (Wang et al., 2023).

5
Conclusions

The completion of this thesis represents both academia progress and practical contributions to the industry. With the electricity market as the central axis, it has contributed to three different areas of statistics and, specifically, to the prediction of time series in evolving contexts. The work bridges theoretical developments and operational deployment, aiming to address the pressing needs of modern forecasting systems under non-stationary conditions. By tackling both methodological and application-specific challenges, this thesis advances the state of the art while offering actionable insights for industry practitioners.

Firstly, a feature selection algorithm has been developed, being the first in the literature to focus on these varying situations and, in particular, on concept shift scenarios. The algorithm is based on observing how each variable influences predictions within a given model, independently of its overall importance. Through the use of Shapley values and their link to prediction errors, a more local analysis about the influence of each feature on the response variable has been performed. In scenarios affected by concept shift, the algorithm identifies features whose behaviour changes and negatively impacts model predictions. By removing these variables, it enhances predictive performance and simplifies model maintenance after deployment. In contrast, under stable conditions, the algorithm detects features that contribute to overfitting, delivering performance on par with state of the art approaches. Such variables induce spurious relationships during training, resulting in detrimental effects during validation, and are therefore excluded from the model.

This adaptive feature selection methodology brings particular value in dynamic environments where data distributions shift over time. By enabling automated detection and exclusion of unstable predictors, the algorithm reduces the need for manual feature engineering and frequent retraining, issues that are especially costly in production environments. Moreover, the methodology is model-agnostic, allowing it to be integrated into a wide variety of forecasting systems, from simple linear models to complex ensemble approaches.

In the specific context of the Electricity Price Forecasting problem, the improvement achieved by the proposed method is notably significant, demonstrating clear value for real world industrial applications. This methodology was instrumental in refining Fortia's price forecasting models during the challenging months after a major regulatory change, a time when previous models struggled with high error rates and unpredictability. The positive results from this application clearly highlight the method's practical impact and value to the industry.

Afterwards, a new framework for price forecasting in the Day Ahead market, grounded in price dynamics, has been introduced. Under the assumption of piecewise stationarity, a new price

forecasting model has been formulated and a computationally efficient parameter estimation has been proposed. Thus, efficient and effective learning algorithms from the literature have been reused, moving away from the current research trends of using increasingly complex learning algorithms whose implementation in real industrial situations is unreasonable due to their long implementation time.

This choice reflects a deliberate trade-off: sacrificing marginal gains from excessively complex state of the art black box models in favour of interpretability (due to the use of simpler models), reliability, and speed, factors that are often more valuable in operational settings.

An extensive evaluation of this novel approach has revealed significant improvements across multiple performance metrics, consistently demonstrating statistical gains in five distinct markets over two separate time frames. The generalization of the model's effectiveness across heterogeneous markets suggests a high degree of transferability, highlighting its potential for widespread adoption. In particular, the results achieved in the Spanish market are exceptionally noteworthy. Furthermore, in alignment with leading approaches reported in the literature, the combination of models produced the most superior results.

What stands out most is the model's ability to remain accurate and reliable even as the underlying data undergoes substantial behavioural changes. By leveraging the assumption of piecewise stationarity, the framework supports frequent and targeted recalibration, ensuring that the model adapts effectively to shifts in price dynamics. This recalibration capability is crucial in the electricity market, where abrupt structural changes can otherwise lead to sharp performance degradation. The success of this strategy highlights the value of designing forecasting systems that explicitly accommodate data evolution, rather than assuming static behaviour over time.

Lastly, it is necessary to quantify the reliability of the predictions. It has been argued that traditional measures of interval validity and efficiency are insufficient for ensuring their appropriate use in practice. It is essential that the variability of the intervals reflects the underlying difficulty of the prediction, while the coverage remains independent of this difficulty. Otherwise, there is a risk of making decisions with a perceived level of certainty that does not align with the true level of uncertainty. Accordingly, a forecasting pipeline integrating two key innovations has been proposed. The first component involves the application of the HQR model, which ensures that the length of the prediction intervals appropriately reflects the difficulty of the forecast. The second component applies the WACI adaptive conformalization process to the intervals produced by the quantile regression model, aiming to achieve uniform coverage regardless of prediction difficulty. This framework is developed under the realistic assumption that only different forecasters of the target variable are available, a common scenario in practical settings.

This work reframes probabilistic forecasting not only as a task of generating intervals but as a calibration problem in dynamically changing environments. Traditional methods often overlook how predictive uncertainty fluctuates with the underlying context, and they tend to deliver overly rigid or misleading intervals. In contrast, the proposed pipeline explicitly separates two goals: learning forecast difficulty (via HQR) and ensuring robust coverage (via WACI). This separation allows for more interpretable uncertainty estimates and better control over the risk associated with decisions made from forecasts.

The effectiveness of the proposed models has been assessed using two illustrative examples: a

synthetic case designed to highlight the advantages of WACI over ACI, and the recurring EPF task. It has been shown that traditional evaluation metrics are insufficient to fully capture the usability of prediction intervals. As a result, new metrics aligned with the practical utility of prediction intervals have been introduced and assessed, revealing a clear superiority of the WACI-HQR methodology.The results demonstrate that each stage of the proposed pipeline successfully addresses specific limitations of existing methods in the literature, delivering the intended improvements.

These experiments underscore a crucial point: in real world forecasting, accuracy is not enough. Users, whether traders, operators, or decision support systems, must also know how much trust to place in each forecast. The WACI-HQR framework provides a meaningful solution to this need.

In addition, Fortia is rewarded for the execution of industrial doctoral theses such as this one. Apart from the problems described and solved here, tasks have been carried out that have contributed to the development of Fortia itself: better management of existing data, automation of a large number of tasks and the production of numerous models that facilitate the company's operations. In particular, three factors have stood out above the rest:

- Decision-making has been optimized and automated, using statistical models as a basis for decision-making. Thus, the sometimes undesirable human bias has been minimized.
- There is less dependence on external suppliers, which saves on the available budget. In addition, in case external services are maintained, various sources of information can be compared and the quality of the information can be benchmarked against that developed internally at Fortia.
- Strategies have been developed or existing ones optimized, increasing the benefits by around 15.7% per year.

In sections 2.4, 3.4 and 4.5 different avenues for future work have been proposed to continue or improve on the developments already presented here. However, the electricity market is constantly developing and there are other topics common to it that are interesting and probably need further development. For example, the study of the future effect of technologies such as batteries or green hydrogen on prices is of vital importance. Anticipating the arrival of these technologies by knowing their impact in advance would provide a high quality competitive advantage. Alternatively, focusing on other markets and/or their relationship with the Day-Ahead Market is vital for the construction of a complete decision scheme. In particular, the Continuous Intraday Market is starting to gain a remarkable importance in the market and forecasting studies in countries such as Spain are very rare.

5.1 Scientific Production

During the PhD a total of three manuscripts have been produced, with one of them already accepted in a JCR impact factor journal, while the others are currently under revision:

1. Carlos Sebastián & Carlos E. González-Guillén. A feature selection method based on

Shapley values robust for concept shift in regression. *Neural Computing and Applications*, 1-23, 2024.

2. Carlos Sebastián, Carlos E. González-Guillén & Jesús Juan. An adaptive standardisation methodology for Day-Ahead Electricity Price Forecasting, *arXiv preprint arXiv:2311.02610*, 2023.

3. Carlos Sebastián, Carlos E. González-Guillén & Jesús Juan. Enhancing reliability in prediction intervals using point forecasters: Heteroscedastic Quantile Regression and Width-Adaptive Conformal Inference, *arXiv preprint arXiv:2406.14904*, 2024.

Moreover, a total of nine oral presentations were delivered at conferences and seminars during the PhD:

1. *Una nueva metodología de selección de variables para predicción de series temporales basada en los valores de Shapley*, FuzzyMad 2022.

2. *Una nueva metodología de selección de variables para predicción de series temporales basada en los valores de Shapley*, seminario GI-TACA diciembre 2022.

3. *A Feature Selection Method Based on Shapley Values Robust to Concept Shift in Regression*, ENBIS 2023.

4. *Un método de selección de variables robusto a situaciones de "concept shift" en problemas de regresión*, SEIO23.

5. *Modelos de predicción del precio de la electricidad*, III Workshop Interdisciplinar UNED, 2024.

6. *Análisis de datos en la empresa: buenas prácticas y caso de uso*, ETSII, 2024.

7. *A Feature Selection Method Based on Shapley Values Robust to Concept Shift in Regression*, ICMAT DataLab, 2024.

8. *Enhancing reliability in point forecasters: Cornish-Fisher Quantile Regression Averaging and Width-Adaptive Conformal Inference*, ISF 2024.

9. *Enhancing reliability in point forecasters: Heteroscedatic Quantile Regression and Width-Adaptive Conformal Inference*, INREC 2024.

Bibliography

[1] Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502, 2021.

[2] Eric Adjakossa, Yannig Goude, and Olivier Wintenberger. Kalman recursions aggregated online. *Statistical Papers*, pages 1–36, 2023.

[3] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

[4] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937, 2020.

[5] Marco Ancona, Cengiz Oztireli, and Markus Gross. Explaining deep neural networks with a polynomial time algorithm for shapley value approximation. In *International Conference on Machine Learning*, pages 272–281. PMLR, 2019.

[6] José R Andrade, Jorge Filipe, Marisa Reis, and Ricardo J. Bessa. Probabilistic price forecasting for day-ahead and intraday markets: Beyond the statistical model. *Sustainability*, 9(11):1990, 2017.

[7] Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

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

[9] APPA. Informe anual del autoconsumo fotovoltaico, 2024.

[10] Andreas Auer, Martin Gauch, Daniel Klotz, and Sepp Hochreiter. Conformal prediction for time series with modern hopfield networks. *Advances in Neural Information Processing Systems*, 36:56027–56074, 2023.

[11] Mutiu Shola Bakare, Abubakar Abdulkarim, Mohammad Zeeshan, and Aliyu Nuhu Shuaibu. A comprehensive overview on demand side energy management towards smart grids: challenges, solutions, and future direction. *Energy Informatics*, 6(1):4, 2023.

[12] Antonio Bello, Derek W. Bunn, Javier Reneses, and Antonio Muñoz. Medium-term probabilistic forecasting of electricity prices: A hybrid approach. *IEEE Transactions on Power Systems*, 32(1):334–343, 2016.

[13] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems*, 24, 2011.

[14] Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. Improved online conformal prediction via strongly adaptive online learning. *arXiv preprint arXiv:2302.07869*, 2023.

[15] Christopher M. Bishop. Mixture density networks. 1994.

[16] BOE. Real decreto-ley 10/2022, por el que se establece con carácter temporal un mecanismo de ajuste de costes de producción para la reducción del precio de la electricidad en el mercado mayorista. *Boletín Oficial del Estado*, (115):67146–67208, 2022.

[17] BOE. Resolución de 23 de mayo de 2024, de la comisión nacional de los mercados y la competencia, por la que aprueban las reglas de funcionamiento de los mercados diario e intradiario de electricidad para su adaptación a las subastas europeas intradiarias, 2024.

[18] Vladimir Britanak, Patrick C. Yip, and K.R. Rao. Chapter 3 - the karhunen–loéve transform and optimal decorrelation. In Vladimir Britanak, Patrick C. Yip, and K.R. Rao, editors, *Discrete Cosine and Sine Transforms*, pages 51–72. Academic Press, Oxford, 2007.

[19] Alessandro Brusaferri, Matteo Matteucci, Pietro Portolani, and Andrea Vitali. Bayesian deep learning based method for probabilistic forecast of day-ahead electricity prices. *Applied Energy*, 250:1158–1175, 2019.

[20] Derek W. Bunn. Forecasting loads and prices in competitive power markets. *Proceedings of the IEEE*, 88(2):163–169, 2000.

[21] Derek W. Bunn, John N Inekwe, and David MacGeehan. Analysis of the fundamental predictability of prices in the british balancing market. *IEEE Transactions on Power Systems*, 36(2):1309–1316, 2020.

[22] Manuel Calzolari. Shapicant, 2020. URL https://github.com/manuel-calzolari/shapicant.

[23] Luis M. Candanedo, Véronique Feldheim, and Dominique Deramaix. Data driven prediction models of energy use of appliances in a low-energy house. *Energy and Buildings*, 140:81–97, 2017.

[24] Alex J. Cannon. Quantile regression neural networks: Implementation in r and application to precipitation downscaling. *Computers & geosciences*, 37(9):1277–1284, 2011.

[25] Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009.

[26] Ying Chen and Bo Li. An adaptive functional autoregressive forecast model to predict electricity price curves. *Journal of Business & Economic Statistics*, 35(3):371–388, 2017.

[27] Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021.

[28] Savina Colaco, Sujit Kumar, Amrita Tamang, and Vinai George Biju. A review on feature selection algorithms. *Emerging research in computing, information, communication and applications*, pages 133–153, 2019.

[29] Javier Contreras, Rosario Espinola, Francisco J. Nogales, and Antonio J. Conejo. Arima models to predict next-day electricity prices. *IEEE Transactions on Power Systems*, 18(3): 1014–1020, 2003.

[30] Edmund A. Cornish and Ronald A. Fisher. Moments and cumulants in the specification of distributions. *Revue de l'Institut international de Statistique*, pages 307–320, 1938.

[31] Eike Cramer, Dirk Witthaut, Alexander Mitsos, and Manuel Dahmen. Multivariate probabilistic forecasting of intraday electricity prices using normalizing flows. *Applied Energy*, 346:121370, 2023.

[32] Ren Diao and Qiang Shen. Nature inspired feature selection meta-heuristics. *Artificial Intelligence Review*, 44(3):311–340, 2015.

[33] Francis X. Diebold and Robert S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 20(1):134–144, 2002.

[34] Dheeru Dua and Casey Graff. Uci machine learning repository. http://archive.ics.uci.edu/ml, 2017. Accessed: 2022-08-08.

[35] Tony Duan, Avati Anand, Daisy Yi Ding, Khanh K. Thai, Sanjay Basu, Andrew Ng, and Alejandro Schuler. Ngboost: Natural gradient boosting for probabilistic prediction. In *International Conference on Machine Learning*, pages 2690–2700. PMLR, 2020.

[36] Grzegorz Dudek. Multilayer perceptron for gefcom2014 probabilistic electricity price forecasting. *International Journal of Forecasting*, 32(3):1057–1060, 2016.

[37] Jonathan Dumas, Ioannis Boukas, Miguel Manuel de Villena, Sébastien Mathieu, and Bertrand Cornélusse. Probabilistic forecasting of imbalance prices in the belgian context. In *2019 16th International Conference on the European Energy Market (EEM)*, pages 1–7. IEEE, 2019.

[38] Bradley Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185, 1987.

[39] Bradley Efron. *Bootstrap methods: another look at the jackknife*. Springer, 1992.

[40] Shai Feldman, Stephen Bates, and Yaniv Romano. Improving conditional coverage via orthogonal quantile regression. *Advances in neural information processing systems*, 34: 2060–2071, 2021.

[41] Pierre Gaillard, Gilles Stoltz, and Tim Van Erven. A second-order bound with excess losses. In *Conference on Learning Theory*, pages 176–196. PMLR, 2014.

[42] Pierre Gaillard, Yannig Goude, and Raphaël Nedellec. Additive models and robust aggregation for gefcom2014 probabilistic electric load and electricity price forecasting. *International Journal of forecasting*, 32(3):1038–1050, 2016.

[43] Carolina García-Martos, Julio Rodríguez, and Mara Jess Sanchez. Mixed models for short-run forecasting of electricity prices: Application for the spanish market. *IEEE Transactions on Power Systems*, 22(2):544–552, 2007.

[44] Paul Ghelasi and Florian Ziel. Far beyond day-ahead with econometric models for electricity price forecasting. *arXiv preprint arXiv:2406.00326*, 2024.

[45] Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

[46] Isaac Gibbs and Emmanuel Candès. Conformal inference for online prediction with arbitrary distribution shifts. *arXiv preprint arXiv:2208.08401*, 2022.

[47] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[48] Xing Han, Ziyang Tang, Joydeep Ghosh, and Qiang Liu. Split localized conformal prediction. *arXiv preprint arXiv:2206.13092*, 2022.

[49] Shahram Hanifi, Xiaolei Liu, Zi Lin, and Saeid Lotfian. A critical review of wind power forecasting methods—past, present and future. *Energies*, 13(15):3764, 2020.

[50] Richard Harris and Robert Sollis. *Applied time series modelling and forecasting*. Wiley, 2003.

[51] Simon Hirsch and Florian Ziel. Multivariate simulation-based forecasting for intraday power markets: Modeling cross-product price effects. *Applied Stochastic Models in Business and Industry*, 2024.

[52] Tao Hong, Pierre Pinson, Yi Wang, Rafał Weron, Dazhi Yang, and Hamidreza Zareipour. Energy forecasting: A review and outlook. *IEEE Open Access Journal of Power and Energy*, 7:376–388, 2020.

[53] Rafael Izbicki, Gilson Shimizu, and Rafael B Stern. Cd-split and hpd-split: Efficient conformal regions in high dimensions. *Journal of Machine Learning Research*, 23(87): 1–32, 2022.

[54] Faheem Jan, Ismail Shah, and Sajid Ali. Short-term electricity prices forecasting using functional time series analysis. *Energies*, 15(9):3423, 2022.

[55] Joanna Janczura and Andrzej Puć. Arx-garch probabilistic price forecasts for diversification of trade in electricity markets—variance stabilizing transformation and financial risk-minimizing portfolio allocation. *Energies*, 16(2):807, 2023.

[56] Tim Janke and Florian Steinke. Forecasting the price distribution of continuous intraday electricity trading. *Energies*, 12(22):4262, 2019.

[57] Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on artificial intelligence and statistics*, pages 2907–2916. PMLR, 2020.

[58] Christopher Kath and Florian Ziel. The value of forecasts: Quantifying the economic gains of accurate quarter-hourly electricity price forecasts. *Energy Economics*, 76:411–423, 2018.

[59] Christopher Kath and Florian Ziel. Conformal prediction interval estimation and applications to day-ahead and intraday power markets. *International Journal of Forecasting*, 37(2):777–799, 2021.

[60] Eoghan Keany. Borutashap : A wrapper feature selection method which combines the boruta feature selection algorithm with shapley values., 2020. URL https://doi.org/10.5281/zenodo.4247618.

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

[62] Gro Klæboe, Anders Lund Eriksrud, and Stein-Erik Fleten. Benchmarking time series based forecasting models for electricity balancing market prices. *Energy Systems*, 6:43–61, 2015.

[63] Nadja Klein, Michael Stanley Smith, and David J. Nott. Deep distributional time series models and the probabilistic forecasting of intraday electricity prices. *Journal of Applied Econometrics*, 38(4):493–511, 2023.

[64] Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

[65] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273–324, 1997.

[66] Miron B. Kursa and Witold R. Rudnicki. Feature selection with the boruta package. *Journal of statistical software*, 36:1–13, 2010.

[67] Jesus Lago, Fjo De Ridder, and Bart De Schutter. Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. *Applied Energy*, 221:386–405, 2018.

[68] Jesus Lago, Grzegorz Marcjasz, Bart De Schutter, and Rafał Weron. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, 293:116983, 2021.

[69] Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1): 71–96, 2014.

[70] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[71] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50 (6):1–45, 2017.

[72] Marília Lima, Manoel Neto, Telmo Silva Filho, and A. de A. Roberta. Learning under concept drift for regression-a systematic literature review. *IEEE Access*, 2022.

[73] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, and Guangquan Zhang. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12): 2346–2363, 2018.

[74] Alexandre Lucas, Konstantinos Pegios, Evangelos Kotsakis, and Dan Clarke. Price forecasting for the balancing energy market using machine-learning regression. *Energies*, 13 (20):5420, 2020.

[75] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[76] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

[77] Katarzyna Maciejowska, Weronika Nitka, and Tomasz Weron. Day-ahead vs. intraday—forecasting the price spread to maximize economic benefits. *Energies*, 12(4):631, 2019.

[78] Wilson E. Marcílio and Danilo M. Eler. From explanations to feature selection: assessing shap values as feature selection mechanism. In *2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 340–347. Ieee, 2020.

[79] Grzegorz Marcjasz, Michał Narajewski, Rafał Weron, and Florian Ziel. Distributional neural networks for electricity price forecasting. *Energy Economics*, 125:106843, 2023.

[80] Aleksei Mashlakov, Toni Kuronen, Lasse Lensu, Arto Kaarna, and Samuli Honkapuro. Assessing the performance of deep learning models for multivariate probabilistic energy forecasting. *Applied Energy*, 285:116405, 2021.

[81] Alexey Matveev, Anastasia_Sidorova_50806198, and DataCanary. Sberbank russian housing market, 2017. URL https://kaggle.com/competitions/sberbank-russian-housing-market.

[82] Klaus Mayer and Stefan Trück. Electricity markets around the world. *Journal of Commodity Markets*, 9:77–100, 2018.

[83] Adam Misiorek and Rafał Weron. Forecasting spot electricity prices with time series models. In *Proceedings of the European Electricity Market EEM-05 Conference*, pages 133–141, 2005.

[84] Claudio Monteiro, Ignacio J. Ramírez-Rosado, L. Alfredo Fernández-Jiménez, and Pedro Conde. Short-term price forecasting models based on artificial neural networks for intraday sessions in the iberian electricity market. *Energies*, 9(9):721, 2016.

[85] Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45 (1):521–530, 2012.

[86] Michał Narajewski. Probabilistic forecasting of german electricity imbalance prices. *Energies*, 15(14):4976, 2022.

[87] Michał Narajewski and Florian Ziel. Ensemble forecasting for intraday electricity prices: Simulating trajectories. *Applied Energy*, 279:115801, 2020.

[88] Michał Narajewski and Florian Ziel. Optimal bidding on hourly and quarter-hourly day-ahead electricity price auctions: trading large volumes of power with market impact and transaction costs. *arXiv preprint arXiv:2104.14204*, 2021.

[89] Julia Nasiadka, Weronika Nitka, and Rafał Weron. Calibration window selection based on change-point detection for forecasting electricity prices. In *Computational Science–ICCS 2022: 22nd International Conference, London, UK, June 21–23, 2022, Proceedings, Part III*, pages 278–284. Springer, 2022.

[90] Radford M. Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[91] Daniel Nickelsen and Gernot Müller. Bayesian hierarchical probabilistic forecasting of intraday electricity prices. *arXiv preprint arXiv:2403.05441*, 2024.

[92] Francisco J. Nogales, Javier Contreras, Antonio J. Conejo, and Rosario Espínola. Forecasting next-day electricity prices by time series models. *IEEE Transactions on Power Systems*, 17(2):342–348, 2002.

[93] Jakub Nowotarski and Rafał Weron. Computing electricity spot price prediction intervals using quantile regression and forecast averaging. *Computational Statistics*, 30:791–803, 2015.

[94] Jakub Nowotarski and Rafał Weron. Recent advances in electricity price forecasting: A review of probabilistic forecasting. *Renewable and Sustainable Energy Reviews*, 81: 1548–1568, 2018.

[95] Isaac Kofi Nti, Moses Teimeh, Owusu Nyarko-Boateng, and Adebayo Felix Adekoya. Electricity load forecasting: a systematic review. *Journal of Electrical Systems and Information Technology*, 7:1–19, 2020.

[96] Christian O'Leary, Conor Lynch, Rose Bain, Gary Smith, and Diarmuid Grimes. A comparison of deep learning vs traditional machine learning for electricity price forecasting. In *2021 4th International Conference on Information and Computer Technologies (ICICT)*, pages 6–12. IEEE, 2021.

[97] Kin G. Olivares, Cristian Challu, Grzegorz Marcjasz, Rafał Weron, and Artur Dubrawski. Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with nbeatsx. *International Journal of Forecasting*, 39(2):884–900, 2023.

[98] María Ortiz, Olatz Ukar, Filipe Azevedo, and Arantza Múgica. Price forecasting and validation in the spanish electricity market using forecasts as input data. *International Journal of Electrical Power & Energy Systems*, 77:123–127, 2016.

[99] Ciaran O'Connor, Joseph Collins, Steven Prestwich, and Andrea Visentin. Electricity price forecasting in the irish balancing market. *Energy Strategy Reviews*, 54:101436, 2024.

[100] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive

confidence machines for regression. In *Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13*, pages 345–356. Springer, 2002.

[101] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[102] Fotios Petropoulos, Daniele Apiletti, Vassilios Assimakopoulos, Mohamed Zied Babai, Devon K. Barrow, Souhaib Ben Taieb, Christoph Bergmeir, Ricardo J. Bessa, Jakub Bijak, John E. Boylan, et al. Forecasting: theory and practice. *International Journal of Forecasting*, 2022.

[103] Dimitris N Politis and Joseph P Romano. The stationary bootstrap. *Journal of the American Statistical association*, 89(428):1303–1313, 1994.

[104] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, Red Hook, NY, 2018. Curran Associates, Inc.

[105] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset shift in machine learning*. Mit Press, Cambridge, Massachusetts, 2008.

[106] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer, 2003.

[107] REE. Informe del sistema eléctrico. informe resumen de energías renovables, 2023.

[108] Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32, 2019.

[109] Tomasz Serafin, Grzegorz Marcjasz, and Rafał Weron. Trading on short-term path forecasts of intraday electricity prices. *Energy Economics*, 112:106125, 2022.

[110] Matteo Sesia and Yaniv Romano. Conformal prediction using conditional histograms. *Advances in Neural Information Processing Systems*, 34:6304–6315, 2021.

[111] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

[112] L. S. Shapley. *A Value for n-Person Games*, pages 307–318. Princeton University Press, Princeton, New Jersey, 1953.

[113] Samaneh Sheybanivaziri, Jérôme Le Dréau, and Hussain Kazmi. Forecasting price spikes in day-ahead electricity markets: techniques, challenges, and the road ahead. *NHH Dept. of Business and Management Science Discussion Paper*, (2024/1), 2024.

[114] Ali Shiri, Mohammad Afshar, Ashkan Rahimi-Kian, and Behrouz Maham. Electricity price forecasting using support vector machines by considering oil and natural gas price impacts. In *2015 IEEE International Conference on Smart Energy Grid Engineering (SEGE)*, pages 1–5. IEEE, 2015.

[115] D. Mikis Stasinopoulos and Robert A. Rigby. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23:1–46, 2008.

[116] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41:647–665, 2014.

[117] Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *International conference on machine learning*, pages 9269–9278. PMLR, 2020.

[118] Boguslaw R. Szkuta, L. Augusto Sanabria, and Tharam S. Dillon. Electricity price short-term forecasting using artificial neural networks. *IEEE Transactions on Power Systems*, 14 (3):851–857, 1999.

[119] Léonard Tschora, Erwan Pierre, Marc Plantevit, and Céline Robardet. Electricity price forecasting on the day-ahead market using machine learning. *Applied Energy*, 313:118752, 2022.

[120] Umut Ugurlu, Ilkay Oksuz, and Oktay Tas. Electricity price forecasting using recurrent neural networks. *Energies*, 11(5):1255, 2018.

[121] Bartosz Uniejewski and Rafał Weron. Efficient forecasting of electricity spot prices with expert and lasso models. *Energies*, 11(8):2039, 2018.

[122] Bartosz Uniejewski and Rafał Weron. Regularized quantile regression averaging for probabilistic electricity price forecasting. *Energy Economics*, 95:105121, 2021.

[123] Bartosz Uniejewski, Jakub Nowotarski, and Rafał Weron. Automated variable selection and shrinkage for day-ahead electricity price forecasting. *Energies*, 9(8):621, 2016.

[124] Bartosz Uniejewski, Rafał Weron, and Florian Ziel. Variance stabilizing transformations for electricity spot price forecasting. *IEEE Transactions on Power Systems*, 33(2):2219–2229, 2017.

[125] B. Venkatesh and J. Anuradha. A review of feature selection and its methods. *Cybernetics and Information Technologies*, 19(1):3–26, 2019.

[126] Jarne Verhaeghe, Jeroen Van Der Donckt, Femke Ongenae, and Sofie Van Hoecke. Powershap: A power-full shapley feature selection method. In *ECML-PKDD 2022*. Springer, 2022.

[127] Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian Conference on Machine Learning*, pages 475–490. PMLR, 2012.

[128] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

[129] Xiaoqian Wang, Rob J. Hyndman, Feng Li, and Yanfei Kang. Forecast combinations: an over 50-year review. *International Journal of Forecasting*, 39(4):1518–1547, 2023.

[130] 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.

[131] Rafał Weron and Adam Misiorek. Short-term electricity price forecasting with time series models: A review and evaluation. *HSC Research Reports*, (HSC/06/01), 2006.

[132] Gerhard Widmer and Miroslav Kubat. Learning in the presence of concept drift and hidden contexts. *Machine learning*, 23(1):69–101, 1996.

[133] Robert L. Winkler. A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, pages 187–191, 1972.

[134] Olivier Wintenberger. Optimal learning with bernstein online aggregation. *Machine Learning*, 106:119–141, 2017.

[135] Wojciech Wisniewski, David Lindsay, and Sian Lindsay. Application of conformal prediction interval estimations to market makers' net positions. In *Conformal and probabilistic prediction and applications*, pages 285–301. PMLR, 2020.

[136] Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pages 25834–25866. PMLR, 2022.

[137] Yao Zhang, Jianxue Wang, and Xifan Wang. Review on probabilistic forecasting of wind power generation. *Renewable and Sustainable Energy Reviews*, 32:255–270, 2014.

[138] Florian Ziel and Rick Steinert. Electricity price forecasting using sale and purchase curves: The x-model. *Energy Economics*, 59:435–454, 2016.

[139] Florian Ziel and Rafał Weron. Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate modeling frameworks. *Energy Economics*, 70:396–420, 2018.

A

Results of the LEAR model with LARS-AIC method for hyperparameter tuning

As the use of the LARS-AIC method has been replaced by cross-validation to calibrate the hyperparameter associated with regularisation in the LEAR method, Table A.1 shows the results comparing runs between methodologies to show that they are practically equivalent. To directly compare the results with those of the proposed methodology, the results are presented as a function of the ratio $\frac{\text{Metric using LARS-AIC}}{\text{Metric using CV}}$. Thus, if the result is greater than 1, it means that using the LARS-AIC procedure is worse than the CV strategy (and by what percentage) and if it is less than 1, it means that it is better (and by what percentage). This ratio will be called as the performance ratio.

The results are in line with what is stated in (68): the results can be improved using a strategy based on cross-validation, but the computational cost is much higher. Moreover, this difference is minimal with the LARS-AIC procedure. The larger differences are observed in the most current datasets and in the NP market. The non-adaptive methodology is considerably improved in the case of using cross-validation. In the case of models based on the adaptive standardisation methodology, the EPEX-DE dataset is the only case where large improvements are obtained after hyperparameter determination. Overall, the results are generally similar and the methodologies are comparable.

Market	Metrics	ASLEAR			LEAR	
		1092/364	1456/728	All	1092/364	1456/728
OMIE SP	MAE	1,0018	0,9941	1,0080	1,0770	1,0695
	RMSE	1,0049	1,0012	1,0068	1,0651	1,0530
	sMAPE	0,9998	0,9873	1,0260	1,0644	1,0479
	rMAE	1,0047	0,9947	1,0059	1,0859	1,0627
EPEX DE	MAE	0,9713	0,9642	0,9692	1,0360	1,0628
	RMSE	0,9787	0,9698	0,9675	1,0296	1,0649
	sMAPE	1,0009	0,9728	0,9844	1,0244	1,0478
	rMAE	0,9794	0,9697	0,9827	1,0330	1,0598
EPEX BE	MAE	1,0110	1,0081	1,0150	0,9916	1,0106
	RMSE	1,0167	1,0041	1,0093	0,9926	0,9973
	sMAPE	1,0269	1,0315	1,0340	1,0247	1,0651
	rMAE	1,0079	1,0093	1,0104	0,9917	1,0078
EPEX FR	MAE	1,0152	1,0134	1,0237	1,0137	1,0263
	RMSE	1,0340	1,0243	1,0463	0,9966	0,9922
	sMAPE	1,0356	1,0400	1,0393	1,0243	1,0506
	rMAE	1,0077	1,0083	1,0322	1,0095	1,0284
NP	MAE	1,0071	1,0083	1,0153	1,0551	1,0441
	RMSE	1,0112	1,0126	1,0121	1,0449	1,0354
	sMAPE	1,0208	1,0289	1,0358	1,0759	1,0829
	rMAE	1,0101	1,0005	1,0075	1,0537	1,0530

Table A.1: Performance ratio when using the LARS-AIC methodology for hyperparameter selection

B

Results of the adaptive standardisation models without filtering outliers

To show the importance of the outlier filtering process in the models to which adaptive standardisation is applied, Table B.1 shows the results without this process for the ASLEAR model. The results are shown using the performance ratio $\frac{\text{Metric without filtering outliers}}{\text{Metric filtering outliers}}$ as in the previous appendix.

Market	Metrics	56	84	1092/ 364	1456/ 728	All
OMIE SP	MAE	1,0000	1,0000	1,0018	1,0017	1,0008
	RMSE	1,0000	1,0000	1,0022	1,0003	0,9999
	sMAPE	1,0000	1,0000	0,9998	0,9990	0,9979
	rMAE	1,0000	1,0000	1,0018	1,0017	1,0008
EPEX DE	MAE	1,0000	1,0000	1,0000	1,0007	1,0179
	RMSE	1,0000	1,0000	1,0000	1,0001	1,0199
	sMAPE	1,0000	1,0000	1,0000	1,0005	1,0082
	rMAE	1,0000	1,0000	1,0000	1,0007	1,0179
EPEX BE	MAE	1,2946	1,3579	1,1175	1,1252	1,1279
	RMSE	2,4396	2,5351	1,1505	1,1567	1,1152
	sMAPE	1,0986	1,1136	1,0913	1,0965	1,1090
	rMAE	1,2946	1,3579	1,1175	1,1252	1,1279
EPEX FR	MAE	1,2862	1,2535	1,0980	1,1107	1,1111
	RMSE	4,5893	3,7559	1,2325	1,0691	1,0609
	sMAPE	1,0497	1,0490	1,0599	1,0778	1,0802
	rMAE	1,2862	1,2535	1,0980	1,1107	1,1111
NP	MAE	1,0679	1,0677	1,0615	1,0547	1,0522
	RMSE	1,1259	1,2046	1,0561	1,0522	1,0544
	sMAPE	1,0576	1,0532	1,0574	1,0483	1,0447
	rMAE	1,0679	1,0677	1,0615	1,0547	1,0522

Table B.1: Performance ratio of not filtering outliers for the ASLEAR model

For the ASDNN the results are shown in Table B.2. As the hyperparameter configurations are not directly comparable, the best network without filtering outliers is compared to the best

network with filtering outliers (indicated by 1 in the table), the second best network without filtering outliers is compared to the second best network with filtering outliers (indicated by 2) and so on. This way the performance ratio can be maintained, which makes the two situations easy to compare.

Market	Metrics	1	2	3	4
OMIE SP	**MAE**	1,0185	1,0080	1,0068	1,0496
	RMSE	1,0212	1,0144	0,9888	1,0536
	sMAPE	0,9991	1,0024	1,0266	1,0101
	rMAE	1,0185	1,0080	1,0068	1,0496
EPEX DE	**MAE**	1,0149	0,9614	0,9541	0,9797
	RMSE	1,0100	0,9594	0,9480	1,0105
	sMAPE	1,0241	1,0112	0,9926	1,0348
	rMAE	1,0149	0,9614	0,9541	0,9797
EPEX BE	**MAE**	1,0476	1,0673	1,0453	1,0927
	RMSE	1,0255	1,0452	1,0079	1,0389
	sMAPE	1,0482	1,0802	1,0695	1,0912
	rMAE	1,0476	1,0673	1,0453	1,0927
EPEX FR	**MAE**	1,0759	1,0815	1,1041	1,1293
	RMSE	1,0264	1,0303	1,0990	1,1591
	sMAPE	1,0605	1,0333	1,0296	1,0635
	rMAE	1,0759	1,0815	1,1041	1,1293
NP	**MAE**	1,0322	1,0313	1,0539	1,1493
	RMSE	1,0139	1,0550	1,0756	3,3125
	sMAPE	1,0098	1,0189	1,0773	1,0439
	rMAE	1,0322	1,0313	1,0539	1,1493

Table B.2: Performance ratio of not filtering outliers for the ASDNN model

In the case of ASLEAR, it can be seen that the OMIE-SP and EPEX-DE datasets are practically unaffected as they have very few outliers. However, the rest of the datasets are affected, especially EPEX-DE and EPEX-FR, which show very significant outliers. It can also be seen that the metric that increases the most is the RMSE, which is the least robust to outliers, so the change in performance is mainly due to outliers. In this case, outlier filtering is a good measure.

The same conclusion can be drawn for the ASDNN, although in this case the measure is not positive for EPEX-DE. The best network is obtained by filtering outliers, but the rest of the networks in this market are not better when filtering is done. Except for this case, in the rest of the markets the neural network also performs better when such preprocessing is done, also improving the results by a considerable percentage.

C

Results of the median-arcsinh transformation when filtering outliers

To show that the improvements come from the adaptive standardisation process and not from the outlier filtering, the experiments have been reproduced on the filtered series but with the median-arcsinh standardisation. The results are shown using the performance ratio $\frac{\text{Metric without filtering outliers}}{\text{Metric filtering outliers}}$ as in the previous appendices for both the LEAR model (Table C.1) and the DNN model (Table C.2).

Market	Metrics	56	84	1092/ 364	1456/ 728	All
OMIE SP	MAE	1,0000	1,0000	0,9999	0,9993	0,9995
	RMSE	1,0000	1,0000	0,9997	0,9998	0,9993
	sMAPE	1,0000	1,0000	0,9998	0,9999	1,0000
	rMAE	1,0000	1,0000	0,9999	0,9993	1,0000
EPEX DE	MAE	1,0000	1,0000	1,0000	1,0004	0,9981
	RMSE	1,0000	1,0000	1,0000	1,0003	0,9969
	sMAPE	1,0000	1,0000	1,0000	1,0003	1,0000
	rMAE	1,0000	1,0000	1,0000	1,0004	1,0000
EPEX BE	MAE	1,0086	1,0063	0,9977	0,9977	0,9969
	RMSE	1,0751	1,0651	1,0118	1,0127	1,0141
	sMAPE	0,9985	0,9991	0,9915	0,9929	1,0000
	rMAE	1,0086	1,0063	0,9977	0,9977	1,0000
EPEX FR	MAE	1,0044	1,0045	1,0057	1,0055	1,0069
	RMSE	1,1561	1,0944	1,0424	1,0468	1,0566
	sMAPE	0,9997	1,0006	0,9970	0,9973	1,0000
	rMAE	1,0044	1,0045	1,0057	1,0055	1,0000
NP	MAE	1,0040	0,9992	0,9991	1,0012	0,9948
	RMSE	1,0094	1,0131	1,0086	1,0132	1,0085
	sMAPE	1,0016	0,9970	0,9941	1,0004	0,9091
	rMAE	1,0040	0,9992	0,9991	1,0012	1,0000

Table C.1: Performance ratio of not filtering outliers for the LEAR model

For the LEAR model there is no noticeable difference between filtering or not filtering outliers

Market	Metrics	1	2	3	4
OMIE SP	MAE	0,9926	0,9375	0,9481	0,9172
	RMSE	0,9859	0,9076	0,9651	0,9172
	sMAPE	0,9955	0,9782	0,9508	0,9364
	rMAE	0,9926	0,9375	0,9481	0,9172
EPEX DE	MAE	1,0175	1,0399	1,1356	1,1860
	RMSE	1,0011	1,0265	1,2452	1,2203
	sMAPE	1,0192	1,0038	1,1069	1,0718
	rMAE	1,0175	1,0399	1,1356	1,1860
EPEX BE	MAE	1,0102	1,0132	0,9963	1,0494
	RMSE	0,9722	1,0257	0,9828	1,0196
	sMAPE	1,0263	1,0368	1,0115	1,0683
	rMAE	1,0102	1,0132	0,9963	1,0494
EPEX FR	MAE	0,9981	0,9924	1,0084	1,0214
	RMSE	0,9842	0,9876	1,0008	0,9978
	sMAPE	1,0012	1,0220	1,0202	1,0209
	rMAE	0,9981	0,9924	1,0084	1,0214
NP	MAE	0,9665	0,9870	0,9769	1,0457
	RMSE	1,0042	1,0102	0,9823	1,0089
	sMAPE	0,9577	0,9976	0,9749	1,0339
	rMAE	0,9665	0,9870	0,9769	1,0457

Table C.2: Performance ratio of not filtering outliers for the DNN model

using a “traditional” standardisation. For the DNN model there really are no major differences either. The EPEX-BE, EPEX-FR and NP datasets show slight differences, while for OMIE-SP and EPEX-DE they seem somewhat larger. The best networks in these two datasets behave equivalently and the differences are most noticeable in the worst hyperparameter settings. However, this could all be due to the randomness of neural networks, since in these two datasets the filtering of outliers is minimal compared to the others where the differences are negligible.

D

Intuitive behaviour of the HQR model

The intuitive idea about the behaviour of the coefficients $\widehat{\lambda}_2(\frac{\alpha}{2})$ and $\widehat{\lambda}_2(1-\frac{\alpha}{2})$ is tested. These have been analysed in the example related to EPF (Section 4.4.3) and the results can be seen in Figure D.1.

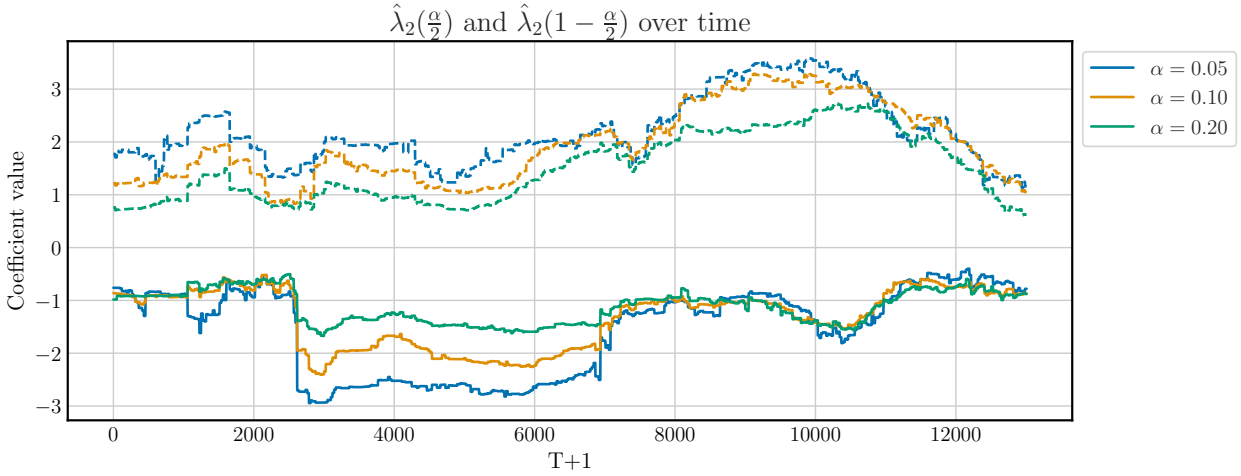

Figure D.1: Value of the coefficients $\widehat{\lambda}_2(\frac{\alpha}{2})$ (continuous line) and $\widehat{\lambda}_2(1-\frac{\alpha}{2})$ (discontinuous line) for different values of α for the EPF example.

As expected, the coefficients associated with the upper extremes are greater than 0, while those associated with the lower extremes are less than 0. In general, the further away the value of α from 0.5, the larger the absolute value of the coefficient. In small periods of time this is not the case, which is probably related to the estimation of the other coefficients of the model. Anyway, it can be said that the intuitive idea about the expected behaviour of the model holds.

E

Analysis of the impact of σ on the WACI methodology

The impact of σ on the evaluation metrics (mean empirical coverage, mean interval length, ILS 0.10, MCD, and Pearson's correlation) has been analysed using the EPF example. For this analysis, γ was fixed at 0.02, and the HQR model was conformalized using the WACI method with σ values ranging from 0.1 to 200 in increments of 0.1. The only value of α analysed is 0.20. The results are presented in Figure E.1, where the result for the ACI-HQR methodology are represented with a black star. This would be equivalent to consider $\sigma = \infty$.

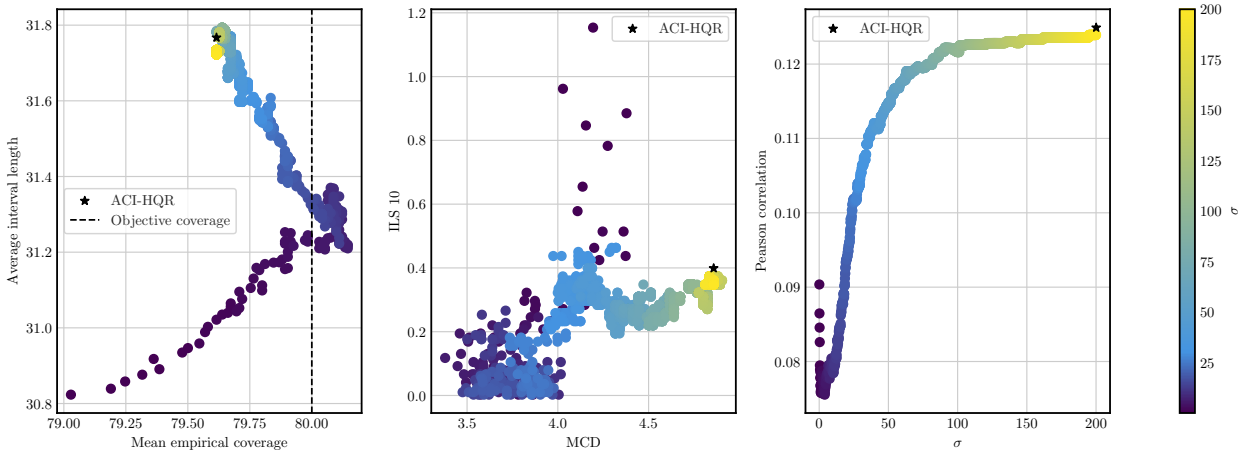

Figure E.1: Mean empirical coverage vs. mean interval length, MCD vs ILS 0.10 and σ vs Pearson's correlation plots

The analysis reveals two distinct σ scales: from 0.1 to 25 and from 25.1 to 200. Very small σ values (less than 1) are not the best choice, as they result in negligible influence on neighbouring observations and there is not enough data to influence all interval lengths. Within the first scale, it is observed that all metrics are interrelated. Specifically, at the σ values that yield efficient intervals, the smallest values for ILS 0.10, MCD, and Pearson's correlation are also achieved. Conversely, when the intervals become inefficient, the other metrics deteriorate significantly. Among invalid intervals, smaller σ values outperform larger ones, achieving similar coverage with shorter intervals. Pearson's correlation also improves with smaller σ values, except at the very smallest values, where the relationship weakens. Interestingly, there appears to be a monotonically increasing relationship between σ and Pearson's correlation,

with larger σ values amplifying the relationship between the coverage indicator and interval length. However, large σ values heavily penalize metrics like MCD, while ILS 0.10 is less affected.