Politechnika Wrocławska

FIELD OF SCIENCE: Social Sciences

DISCIPLINE OF SCIENCE: Management and Quality Studies

DOCTORAL DISSERTATION

Decision support for electricity market participants: point and probabilistic forecasting using resampling methods and statistical learning

Weronika Nitka

Supervisor/Supervisors:
prof. dr hab. inż. Rafał Weron

Assistant supervisor:
dr hab. Katarzyna Maciejowska

Keywords: Electricity price forecasting, Energy markets, Decision support systems, Statistical learning, Trading strategies

WROCŁAW 2024

Abstract

This doctoral thesis aims to develop data-driven forecasting methods designed to support the decision-making processes of small and medium companies participating in electricity markets. The methods are designed with low to medium computational complexity and high level of automation. The thesis comprises five research papers which fulfill four objectives, addressing different research gaps related to this main goal: (i) conduct a critical study of calibration sample selection for automation of electricity price forecasting; (ii) use resampling methods to generate predictive distributions of electricity prices and better assess uncertainty; (iii) utilize renewable generation and load forecasts to design trading strategies in day-ahead and intraday markets; (iv) develop decision support methods for day-ahead bidding that use combinations of predictive distributions. The proposed approaches are evaluated from statistical and financial points of view, offering both scientific novelty and practical applicability.

Streszczenie

Ta rozprawa doktorska ma na celu opracowanie metod prognostycznych opartych na danych, wspierających procesy decyzyjne małych i średnich firm operujących na rynkach energii elektrycznej. Przedstawione metody są zaprojektowane tak, aby mieć niską lub średnią złożoność obliczeniową i wysoką automatyzację. Rozprawa doktorska jest cyklem pięciu publikacji, które realizują cztery zadania adresujące luki badawcze związane z tym głównym celem: (i) przeprowadzenie krytycznego badania metod wyboru próbki kalibracji do automatyzacji prognozowania cen energii elektrycznej; (ii) wykorzystanie metod ponownego próbkowania do tworzenia rozkładów prognostycznych cen energii elektrycznej i oceny niepewności; (iii) użycia prognoz produkcji energii ze źródeł odnawialnych i zapotrzebowania na elektryczność do planowania strategii handlu na rynkach dnia następnego i intraday; (iv) opracowanie metod wspomagania podejmowania decyzji do krótkoterminowego handlu wykorzystujących uśrednienia rozkładów prognostycznych. Zaproponowane w tej pracy rozwiązania są oceniane zarówno ze statystycznego jak i finansowego punktu widzenia, niosąc ze sobą wkład w rozwój dyscypliny i potencjał w praktycznych zastosowaniach.

Contents

Chapter 1

Introduction

1.1 Forecasting as a decision support tool in electricity markets

Within the last few decades, electricity markets have been moving in the direction of decentralization and deregulation. Aiming to increase competition and efficiency, energy supply previously monopolized by state-owned utilities has undergone fragmentation and privatization. A large share of developed countries introduced electrical power exchanges, where companies may freely buy and sell energy. While operating in these markets is subject to extensive legal regulations, most of the time electricity is traded according to the law of supply and demand, under free market conditions.

The participants of electricity markets make a multitude of operational decisions on a day-to-day basis. While long-term contracts are popular among large companies, the spot market is widely considered the most important avenue for trading electricity (URE, 2023). It is the main reference point for risk management and the spot price is used as the underlying of derivative products (Burger et al., 2014). In continental Europe the spot market typically includes the day-ahead market, where trading is conducted within a uniform-price auction for all hourly or block (e.g. peak or off-peak hours) contracts of the next day; and the intraday market, where trading is possible until minutes before physical delivery for hourly or shorter load periods (Mayer and Trück, 2018; TGE, 2024). They are supplemented by technical markets, e.g. balancing markets, which hold a predominantly technical role of ensuring safe and continuous functioning of the electrical grid.

Trading in the spot market requires the participating companies to submit their bids and offers ahead of physical delivery. They need to carefully choose the delivery time and price at which they will bid in the market, taking into account the uncertainty of consumption and generation volume of renewable energy sources (RES), such as wind or photovoltaic (PV) farms. Underestimating or overestimating RES generation necessitates making additional, likely unfavorable, bids closer to delivery time. On a system-wide scale, misjudging the market events may lead to a number of negative consequences:

among them, price surges of a large magnitude, financial losses of companies or local black- or brownouts, i.e. energy shortages (Morales et al., 2014).

The companies make these decisions using their knowledge of the market and forecasts of prices and fundamental variables, such as electricity demand and generation. As explained by Waddell and Sohal (1994), "*Forecasting is generally used to predict or describe what will happen (for example, to sales demand, cash flow, or employment levels) given a set of circumstances or assumptions. Planning, on the other hand, involves the use of forecasts to help in making good decisions about the most attractive alternatives for the organization. [...] Generally speaking, forecasting and forecasts are inputs to the planning process.*" A similar perspective is presented by Petropoulos et al. (2022), stating the purpose of forecasting as "*to improve decision making in the face of uncertainty. To achieve this, forecasts should provide an unbiased guess at what is most likely to happen (...), along with a measure of uncertainty (...). Such information will facilitate appropriate decisions and actions.*" Types of forecasts used in management "*(...) include point estimates as well as expressions of uncertainty of such estimates in terms of probabilistic forecasts, prediction intervals, or path forecasts*" (Petropoulos et al., 2024).

With the majority of managers in the energy sector recognizing business analytics as a crucial area of development, automation and data-driven forecasts are becoming commonplace among energy market participants. According to an international survey of utility companies by SAS, over 70% of managers recognize analytics as a core part of their operations, while 60% of the respondents cite business analytics as a transformative factor in the way their business is conducted (SAS, 2017). Moreover, they point to energy forecasting as the highest priority activity, more so than asset management or customer segmentation analytics. These responses underscore a widespread practical relevance of business analytics and forecasting, while practical case studies, such as the ones of Hong (2016) and Fabbiani (2024), provide tangible evidence that reducing demand and/or price forecast errors allows energy companies to save hundreds of thousands USD or EUR yearly.

Typically, forecasting in organizations is a multi-step process with multiple participants involved. However, the actual methods used in practice do not realize the potential seen in the scientific literature (Petropoulos et al., 2022). Indicators of uncertainty, such as probabilistic forecasts, provide valuable and more complete information, but they may be misinterpreted or ignored by the users in business contexts (Goodwin, 2014; Vukovic, 2023). Additional value can be gained from simulations and case studies evaluating performance of forecasts applied in decision-making, since the connection between statistical accuracy and business value is usually not straightforward (Kolassa, 2023; Robette, 2023).

There exists a wide disparity of available computational resources and data between established, major market players and smaller utilities. While large companies are able to devote significant amounts of money and computational power to develop in-house forecasting models or procure proprietary data, smaller market participants typically do not have such capabilities. Therefore, accessible forecasts basing on openly available data and produced with interpretable models address needs of a growing number of firms.

In a recent review, Maciejowska et al. (2023) identify three current trends in electricity price forecasting:

- considering not only point but also probabilistic (interval, density) or path (also called ensemble) forecasts,
- using statistical and machine learning approaches instead of parsimonious but less accurate econometric models, and
- evaluating forecasts not only in terms of statistical error measures, but also by comparing profits from trading strategies.

The research articles that form the core part of this thesis cover all three directions, while keeping in mind accessibility to the forecast users. They describe automated methods with low to medium computational complexity, which can be used to complement or replace simple forecasting models. Thus, they balance novelty with accessibility, providing both scientific and practical value.

1.2 Aim and objectives

This thesis addresses the needs of small and medium enterprises participating in electricity markets. It aims to develop automated and data-driven forecasting methods with low to medium computational complexity, designed to support the decision-making processes. With this in mind, four objectives are set:

1. Conduct a critical study of calibration sample selection for automation of electricity price forecasting (**Paper 1**, **Paper 2**).
2. Use resampling methods to generate predictive distributions of electricity prices and better assess uncertainty (**Paper 3**).
3. Utilize renewable generation and load forecasts to design trading strategies in day-ahead and intraday markets (**Paper 3**, **Paper 4**).
4. Develop decision support methods for day-ahead bidding that use combinations of predictive distributions (**Paper 5**).

Objective 1 addresses a research gap of selecting appropriate calibration samples for estimating forecasting model parameters. Usually, electricity price forecasting literature is focused on model specification, while the choice of the calibration sample is commonly performed in an ad-hoc, arbitrary manner, or dictated by data availability. However, some studies are showing a major impact of this choice on forecasting accuracy. Optimal calibration samples usually change over time and are difficult or impossible to predict a priori, necessitating extensive backtesting to make informed decisions. As part of Objective 1, automated methods of calibration sample selection are proposed, aiming to help the users effectively calibrate their chosen models without additional expert knowledge.

Probabilistic forecasts, i.e. forecasts which give information about the distribution of the predicted variable rather than just its expected value, are highly valuable when used for decision support. They enable managers to assess the likelihood of positive or negative events and plan accordingly. At the same time, probabilistic forecasting methods can be seen as daunting and overly complex when implemented in practice. Objective 2 addresses this concern by generating predictive distributions using resampling methods, which can be an extension of an existing point forecasting model.

For a company which produces or trades renewable energy in the spot market, electricity prices are not the only relevant variable. Fundamental variables, such as system-wide load (i.e. demand) and generation volume, hold high importance in operational decisions. They also have a significant impact on future market prices. Objective 3 focuses on designing trading strategies for renewable energy generators that utilize forecasts of multiple variables in the decision-making process. This allows for a more comprehensive assessment and can lead to better decisions.

Finally, a simple method to improve forecast accuracy and reduce model selection risk at the same time is to combine forecasts from several different models into a single one. Objective 4 aims to research whether the reduction in forecast errors from combining can be leveraged for improving a company's profits while trading. Additionally, an automatic, data-driven method of assigning combination weights is compared to a naive averaging approach.

Chapter 2

Electricity markets and market participants

2.1 Major market players

While the thesis focuses on generators and traders, this chapter aims to place them in a wider context. Examples are drawn primarily from the German and Polish electricity markets. The largest group of players in these markets are consumers. However, the majority of them, such as households and small businesses, do not interact with the market directly. Instead, they are served by utilities. However, large consumers such as heavy industrial plants often bid directly or form personalized contracts with suppliers. Their main interest while trading lies in achieving a reliable electricity supply.

The generators provide electricity supply by operating conventional (e.g. fossil fuel) or renewable power plants. Large generation companies are likely to have a variety of power plants in their portfolio and participate in the markets by bidding directly. Smaller companies have fewer assets and may use services of intermediaries such as traders and aggregators. While they differ with regards to amounts and types of power plants owned, typically generation companies aim to earn revenue by selling produced energy in accordance with market demand. The private companies participating in electricity markets are overseen by the state-owned system regulators and operators, who are responsible for maintaining the distribution grids and ensuring proper functioning of the system.

The process of decentralization of energy markets has been accelerated by constantly increasing propagation of renewable energy solutions. With the increased availability of small-scale wind or solar power installations, smaller companies and even private households (i.e. prosumers) may invest in energy generation, oftentimes producing electricity for their personal needs and selling surplus back to the grid. While the majority of energy is still produced by a number of large utilities, their joint market shares are decreasing in favor of smaller generators. For example, Figure 2.1 shows the electricity market landscape in year 2023 in Poland. It can be seen that there are only four companies with

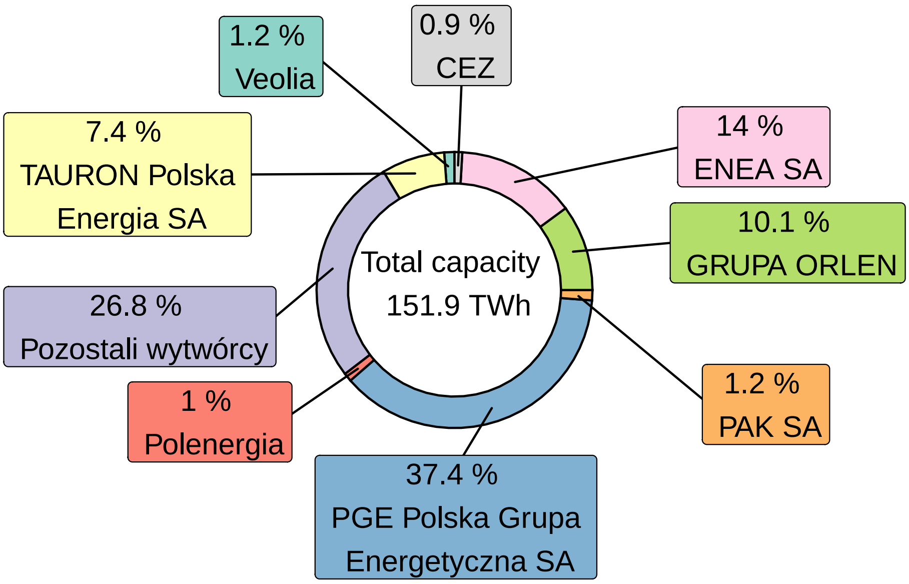


Figure 2.1: Shares of electricity generation supplied by market participants in Poland in 2023. Data source: URE (2024).

market share larger than 5%, and their combined output constitutes less than 70% of the total energy generation. Meanwhile, more than a quarter of total energy output is covered by small companies, many of which are renewable energy generators (URE, 2024). A similar trend can be seen in Germany, where in 2022 the combined market share of five biggest companies equals 63%, with only three of them covering more than 5% of total energy generation (BNetzA, 2023).

While prosumers do not independently participate in electricity markets and are thus counted in the shares of larger companies, their impact on the market landscape cannot be overlooked. In Poland, there are more than 1.4 million registered prosumers, 95% of which are private households. They contribute approximately 80% of total Polish photovoltaic energy generation (URE, 2024). Across Europe, affluent areas observe emergence of so-called renewable energy communities. They couple and connect multiple producers and consumers of energy within a small geographical area, using internet of things technologies to flexibly manage supply and demand (Neska and Kowalska-Pyzalska, 2022).

For electricity market participants, free market trading offers both opportunities and risks. Electrical grids require the influx and outflux of energy to be balanced at all times in order to provide consistent voltage to customers. When the energy supply is centrally regulated, the majority of uncertainty comes from the consumption. However, increasing the number of companies participating in electricity trading makes supply unpredictable as well. The volatility is further increased by the growing penetration of renewable energy sources in many countries' energy mixes (Paraschiv et al., 2014; Mwampashi et al., 2021).

2.2 Characteristics and challenges of renewable energy sources

Due to limited global supply and negative environmental impact, there is a worldwide trend to move on from conventional power plants, which burn fossil fuels, to renewable energy sources (Papież et al., 2019; Cullen and Reynolds, 2023). As their name implies, they produce electrical energy from practically inexhaustible sources such as the sun, wind or water. In 2023, the share of renewables in the global power generation mix reached 29% and it was forecasted to rise by further 6 percentage points until 2025. In Europe the penetration of renewables is even higher, with 35% across the whole continent and in excess of 60% of total energy demand in countries such as Germany or Denmark (IEA, 2023).

Aside from sustainability, the biggest advantage offered by renewable energy sources is their low cost of operation. Due to the constant buying and consumption of non-renewable fuels, conventional power plants have significantly higher marginal costs (i.e. variable costs, primarily fuel and CO_2 emissions) than RES generators, whose marginal costs are often close to zero. When considering the levelized cost of generating electricity, i.e. including fixed costs such as establishing and maintenance of generation facilities, the difference is smaller, but still substantial. In 2020, the levelized costs were approximately twice as large for coal and gas-powered plants than for industrial-scale onshore wind and photovoltaic plants (IEA, 2020). Since offers in electricity markets are typically accepted starting from the lowest marginal costs, this leads to a so-called merit-order effect, which shifts the price curve, lowering the average cost of electricity per 1 MWh (see Fig. 2.2; Cludius et al., 2014; Hagfors et al., 2016b; Kremer et al., 2021). In turn, increasing the share of electricity consumption from renewable energy sources is likely to positively affect the economic growth (Papież et al., 2019). Additionally, renewable energy investments improve the energy security of countries historically relying on imported fossil fuels (Papież et al., 2018).

While the benefits of renewable energy sources in terms of affordability and greenhouse gas emissions are substantial, they are not without drawbacks. The most common renewable power plant types include hydroelectric turbines, wind turbines and photovoltaic panels. The latter two and some types of hydroelectric plants are intermittent, non-dispatchable energy sources, meaning that their energy output is difficult or nearly impossible to accurately predict and control. Power supplied by such a utility can significantly vary between zero and nominal capacity. Their generation depends predominantly on weather conditions and peak output does not usually coincide with periods of the most intensive energy consumption. This behavior is different than conventional power generators, which can adjust their power output in a controlled and deterministic way, although

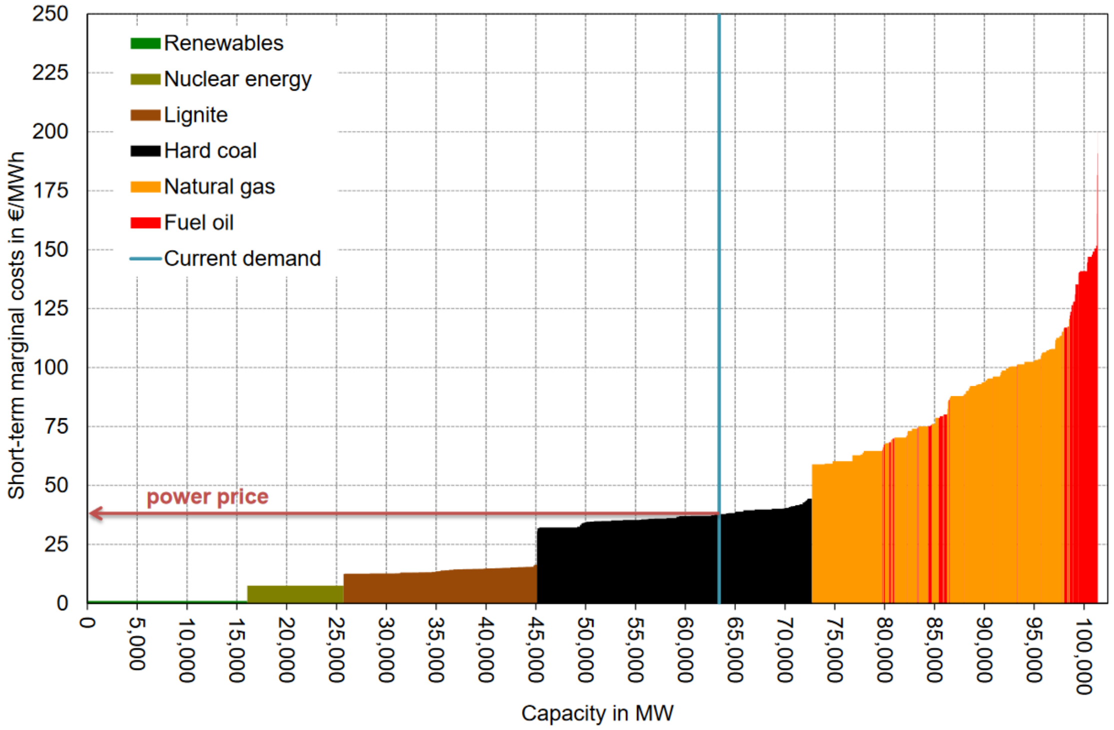


Figure 2.2: Stylized merit-order curve in German electricity market, settling the final spot price at the intersection of demand volume and supply offers from least to most expensive. Source: Cludius et al. (2014).

not without costs or difficulties.

Electricity is highly distinct from most other commodities, which correspond to physical assets, often suitable for long-term storage and logistically difficult to transport. On the contrary, the capacities for electricity storage are minuscule compared to the total demand. The demand for electricity is largely considered to be inflexible, with its average levels varying depending on the economic activity (higher during weekdays and lower during weekends and holidays) and weather (rising with the necessity for heating or air conditioning) (Paraschiv, 2013). Although there is a growing movement to implement flexible demand response solutions, which would allow to control and adjust the load to available supply, there are many technical and legislative barriers preventing a more widespread use of these technologies (O'Connell et al., 2014; Negash and Westgaard, 2018; Parrish et al., 2019).

Uncertainty of renewable energy production can be alleviated by energy storage solutions (see **Paper 5** for an example of trading strategy). However, currently these technologies face a number of limitations. They either require particular geographical conditions (such is the case of e.g. pumped-storage hydroelectric power plants, currently one of primary sources of flexibility (Bento et al., 2023) or have highly limited capacity coupled with prohibitive costs (which applies to lithium-ion batteries). More efficient dispatchable energy generation and storage solutions are currently being developed, with some examples being hydrogen power storage systems and sodium-ion batteries. Even with rapid evolution of such technologies, their widespread adoption may not happen before the international climate energy agreement deadlines, such as the European Green Deal's climate neutrality by 2050.

Introducing large volumes of renewable energy into generation mix can lead to strain on the transmission systems, due to congestion or instability of voltage (Intini and Waterson, 2023). This, in turn, necessitates curtailing the renewable energy production below its nominal effectiveness (Bird et al., 2016). Large conventional power plants have substantial ramp-up and ramp-down costs, i.e. costs in terms of time and money of turning generators on and off to adjust power output, which increases the difficulty of flexible adjustment of energy supply (Finnah et al., 2022; Cullen and Reynolds, 2023). On the other hand, profitability calculations are not straightforward when taking into account flexibility requirements of the entire system. Companies may receive financial incentives to reserve additional capacity in the event of excessive electricity demand, while those contributing significantly to imbalance may receive penalties or lose subsidies (Zugno et al., 2013; Laur et al., 2020; Intini and Waterson, 2023). This puts intermittent renewable energy generators at financial risk.

A third issue connected to the rise in popularity of renewable electricity generation is its high susceptibility to extreme weather events. While all energy sources are, to some extent, able to be influenced by external factors – e.g. geopolitical, such as the Russian-Ukrainian war – conventional fuels can be stockpiled in case of emergency, while most renewable energy sources have no such capacity (Śmiech et al., 2021; Będowska-Sójka et al., 2022). This issue can be further aggravated by the global climate changes and the rise in occurrences of extreme weather events. An example of such a catastrophic phenomenon are severe droughts in 2022, which in some countries caused a more than 20% drop in hydropower generation compared to previous five-year average (IEA, 2023). Due to such events, European electricity prices in 2022 and onwards have seen an unprecedented rise. With magnitudes larger values and volatility of the prices, market participants need accurate forecasts to manage their financial risks (Fałdziński et al., 2021).

Within this landscape, renewable energy utilities participate in electricity markets without precise knowledge about their actual generation. With the majority of trading happening a day or more ahead of physical delivery, it may necessitate the companies to submit multiple bids, both in day-ahead and real-time markets (see Section 2.3). While the intended goal of intraday trading is to correct unforeseen deviations from day-ahead schedules, it can also be used strategically by the participants (Pape et al., 2016). Larger companies with sufficient market power may withhold or overestimate generation in order to utilize arbitrage opportunities, especially with a diversified portfolio of renewable and conventional power plants (Fabra and Imelda, 2023). Alternatively, wind power plants

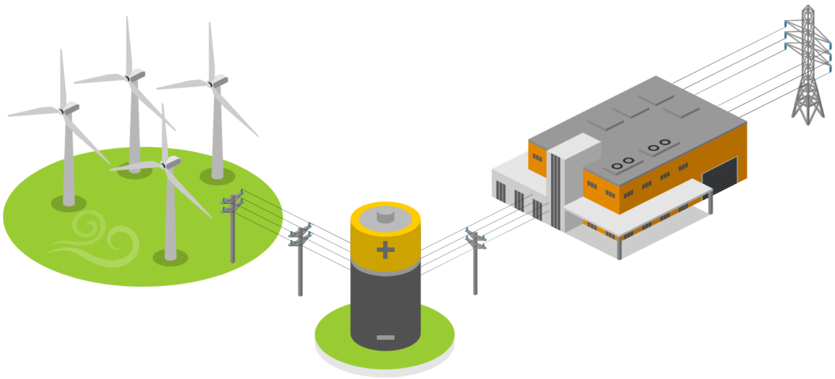


Figure 2.3: Graphical representation of the fictional company considered in Section 2.3.

have been observed to overestimate their generation in order to benefit from curtailment regulations under British law (Intini and Waterson, 2023). Companies with smaller market power cannot significantly impact market prices, but they can use forecasts of market prices and generation to inform their bidding decisions, as shown e.g. in **Papers 3–5**.

2.3 Trading in electricity markets

In order to illustrate the challenges and decisions faced by the electrical energy market participants, let us introduce a fictional company visualized in Fig. 2.3. This company owns a small power plant consisting of intermittent renewable energy sources, e.g. wind turbines, up to several megawatt hours of nominal capacity. They also own a battery storage system with a smaller capacity. Such a utility operates commercially, with the goal of turning a profit rather than producing for their own needs and only selling excess energy. We can assume that this company is small enough not to meaningfully affect the market prices through their actions.

At the beginning of the decision-making process, the company needs to estimate the volume of energy they should buy or sell in the next day. The maximum supplied amount is equal to the sum energy produced by the wind turbines and the planned state of charge of the battery. They can also purchase an amount of energy equal to the discharged battery capacity. When calculating these values they need to take into account the fact that wind turbines usually produce energy in amounts lower than its nominal capacity, strongly depending on atmospheric conditions. Looking at statistics on a national scale, they can expect to generate around a fourth of their nominal capacity: in year 2023 in Poland, installed wind farms with capacity of 9.63 GW produced approximately 2.52 GW per hour, while for solar power plants this fraction was an even lower – 1.51 GW compared to nominal capacity of 14.28 GW (Burger, 2024).

The primary source of revenue for this company is selling the energy generated by the

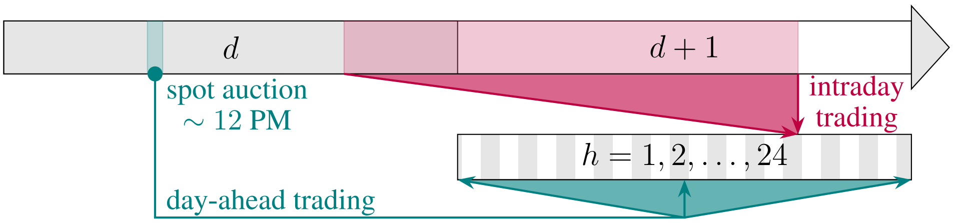


Figure 2.4: Timeline of energy market auctions.

wind turbines in the spot market. The initial offers have to be placed before the day-ahead auctions close, which is around noon (in Germany) or 1 PM (in Poland) on each working day (EEX, 2024; TGE, 2024). This is illustrated in Fig. 2.4 for a simplified timeline, where the green color marks the auction time and delivery period for the day-ahead auction, while the red color represents trading for an arbitrary contract in the intraday market. Separately for all hours in the subsequent day, the company has to place offers with set volume and price. The offers can be limited (executed at the designated or better price) or unlimited (also known as volume or market orders, executed unconditionally at best possible price) (Madlener and Kaufmann, 2002; Mayer and Trück, 2018). Additionally, in case of an unaccepted bid or unwillingness to trade, the company can temporarily curtail their production by turning off turbines.

Several hours after the day-ahead auction closes, e.g. at around 3 PM in Germany, the intraday (real-time) market opens for trading (EEX, 2024). The market participants are encouraged to self-balance their bids in this market (Pape et al., 2016; Lehna et al., 2022). In other words, they may use intraday trading to settle the difference between their day-ahead bids and the actual generation. This may be done by selling, if the actual generation is higher than the initial offer, or buying energy in the opposite case. Since intraday prices tend to be on average more volatile than day-ahead prices, the latter scenario may be costly (Maciejowska et al., 2019).

Although the company can also participate in the balancing market, its main role is to ensure a reliable functioning of the grid rather than turning profits. Therefore, it is usually not taken into account in trading case studies. In many countries there exist additional mechanisms intended to help increase the system's flexibility and reduce imbalances, such as financial incentives or imbalance fees (Hu et al., 2018; Laur et al., 2020). They can impact the company's decisions, but are difficult to consider in research intended for broad audience due to narrowing the range of likely applications.

During the decision-making process, the company has to be aware that electricity prices frequently fluctuate. The price spikes have large magnitudes, driven both by seasonality and by unexpected events, such as changes in supply or demand (Hagfors et al., 2016a). Even relatively small changes in supply and demand can lead to positive or negative price spikes, as illustrated in Figure 2.5.

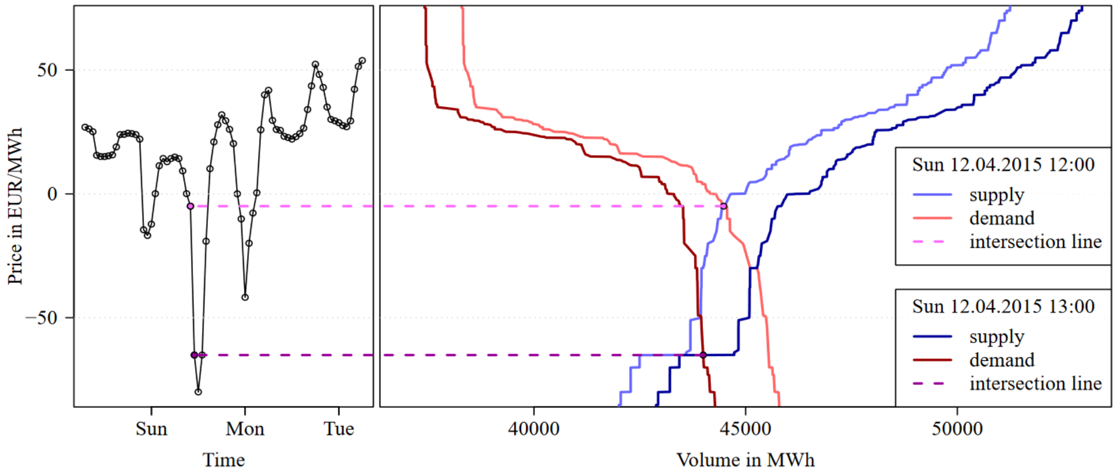
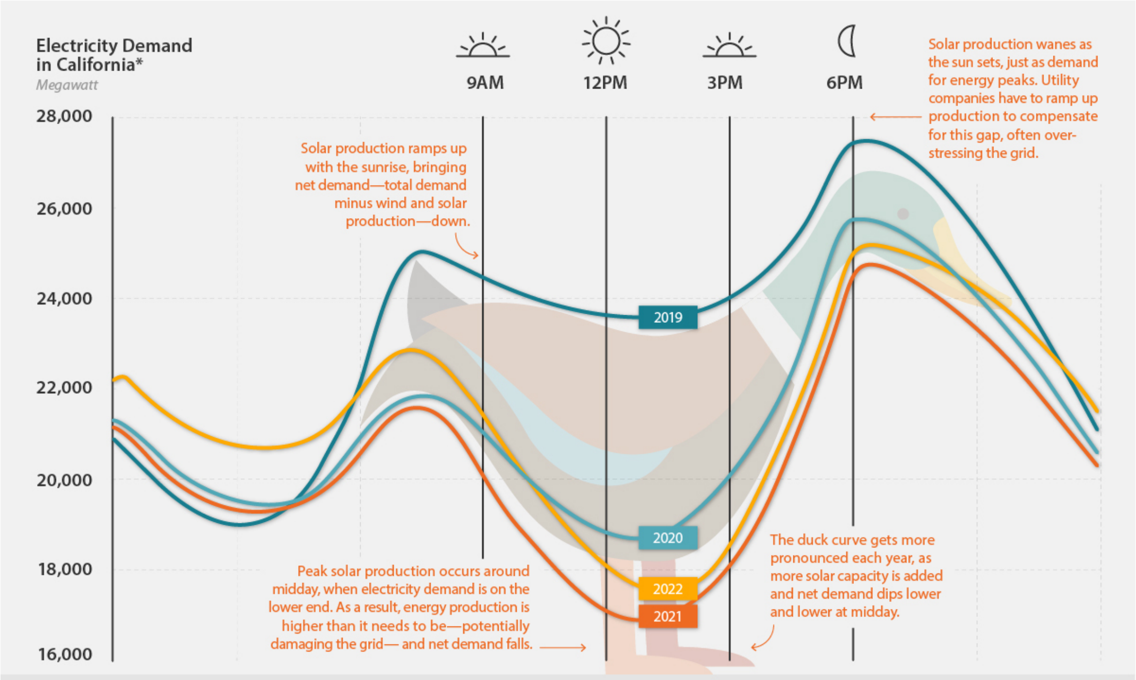


Figure 2.5: Illustration of market price formation at the intersection of supply and demand curves. Source: Ziel and Steinert (2016).

Additionally, the highest volumes of renewable energy are produced in hours outside of peak consumption periods. Wind energy is largely independent from the time of day, while photovoltaic energy generated throughout the day forms an approximate bell curve with a peak around midday. With the peak energy consumption times being early morning and late afternoon, this leads to a phenomenon dubbed "duck curve" by Denholm et al. (2015). The name playfully describes the shape taken by residual demand, i.e. the share of demand not covered by renewable generation, when photovoltaic energy sources are prevalent in the generation mix. The residual demand drops during the night and midday and rises in the mornings and evenings, resembling a duck's silhouette (see Fig. 2.6 for illustration). Despite the name, the phenomenon itself is not harmless, due to adding ramp-up and ramp-down strain on conventional power plants and increasing need for demand flexibility. From the perspective of managing the company, the mismatch between peak demand and generation of their power plants would expose them to the risk of low or even negative market prices for their product.

An example of the duck curve and complex interactions, highlighting the importance of forecasts in electricity markets can be seen in Figure 2.7. The plot depicts energy generation and demand overlaid with market prices for a week in October of 2023. Sunny and windy weather during that period led to unusually high renewable energy generation, especially towards the end of the week. Because most major conventional power plants need significant time and funds to ramp up and down, the generation cannot be elastically adjusted in such situations, leading to overproduction of energy. Even with increased cross-border export, this situation resulted in negative electricity market prices for brief periods of time, which could lead to financial losses of small producers like the considered company. These losses could potentially be mitigated with an accurate forecast of price or demand.



Solar production ramps up with the sunrise, bringing net demand—total demand minus wind and solar production—down.

Solar production wanes as the sun sets, just as demand for energy peaks. Utility companies have to ramp up production to compensate for this gap, often over-stressing the grid.

Peak solar production occurs around midday, when electricity demand is on the lower end. As a result, energy production is higher than it needs to be—potentially damaging the grid— and net demand falls.

The duck curve gets more pronounced each year, as more solar capacity is added and net demand dips lower and lower at midday.

Figure 2.6: Infographic illustrating the appearance and process of formation of the "duck curve". Source: Wallach (2022).

Ownership of an energy storage solution such as a lithium-ion battery system opens up more opportunities for the company. Treated independently from the wind turbines, it can be used to take advantage of electricity price variations while trading. Using a forecast of hourly prices for the subsequent day, the battery can be charged with energy purchased at the lowest possible price. Then, it can be discharged to sell the energy back to the market when prices are high, e.g. during the evening peak. Since the battery's capacity is a known quantity, the company needs to decide when to enter the market and at which price. This type of strategy is illustrated in Fig. 3.8 in Section 3.4.2.

In this thesis, strategies related to different types of power installations (wind turbines, batteries etc.) are considered as stand-alone. Each research paper discusses a single type of utility. This simplification of analyzed scenarios makes them relevant for a broader variety of small and medium companies. It also allows to evaluate the advantages of different forecasts more clearly than multi-step, complex scenarios. Since the business value of a forecast is not a direct function of its accuracy, adding additional factors to the decision-making processes can make interpretation of results even more difficult. However, a company owning multiple types of infrastructure would naturally use them together to mitigate risks. For example, instead of curtailing production when facing insufficient demand, the unsold energy could charge the battery and then be sold at a more convenient moment.

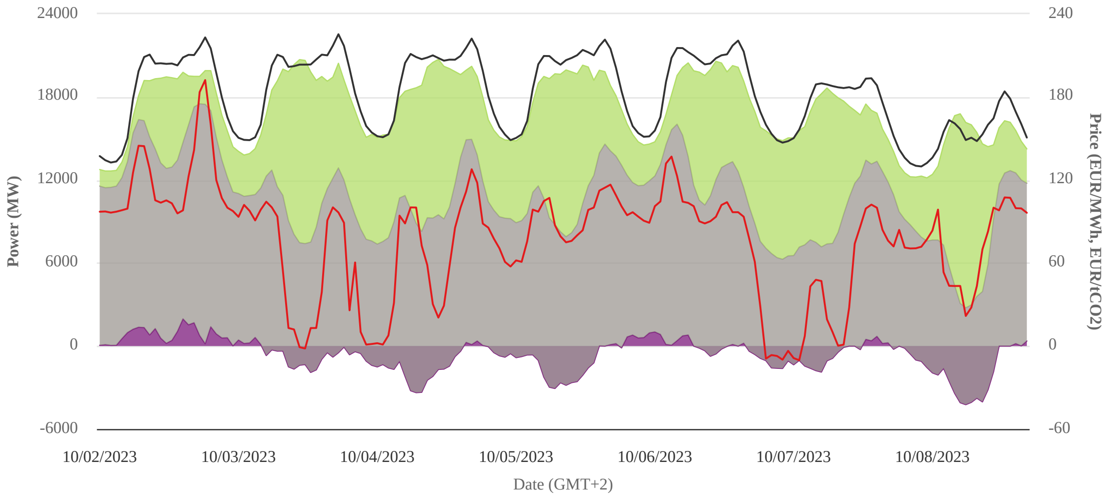


Figure 2.7: Electricity price (red line), demand (black line), and generation (renewable: green, conventional: gray, cross-border trading: purple) for a selected week in Poland. Source: Burger (2024).

Chapter 3

Summary of results

3.1 A critical study of calibration sample selection for automation of electricity price forecasting

3.1.1 Problem statement

In the context of forecasting, the accuracy of the predictions depends on both the model and the data used for estimation of its parameters. While the former aspect is widely considered in the electricity price forecasting literature, the latter is a much less popular topic. Typically, simpler models such as linear regression assume that the forecasted time series has a constant variance and has no linear relationship between independent variables (Greene, 2012). However, these assumptions are frequently violated in empirical applications, especially in response to external events, e.g. geopolitical situation or legal regulation changes.

The issue of selecting calibration sample has been addressed by econometric literature, see e.g. Zeileis et al. (2003); Tian and Anderson (2014), although it has only recently gained attention in electricity price forecasting (Marcjasz et al., 2018; Hubicka et al., 2019). In the presence of structural breaks, a natural approach would be to restrict the model calibration sample to observations more recent than the last change-point. While intuitive, this decision can increase the variance of estimators due a reduction of the number of observations used for estimating parameters. Pesaran and Timmermann (2007) have shown that it is often beneficial to include older data from different regimes. The resulting rise in bias does not necessarily negate the reduction of variance gained from a larger sample size. In order to mitigate the estimation difficulties, the authors propose using combination methods or cross validation to find optimal calibration windows.

Choosing an optimal calibration window length ex-ante is usually not straightforward. While most research papers follow the trend of choosing the maximum sample length available for the data and forecasting setup they are working with, the optimal choice can be vastly different (see e.g. Hubicka et al., 2019; Maciejowska et al., 2020). To further

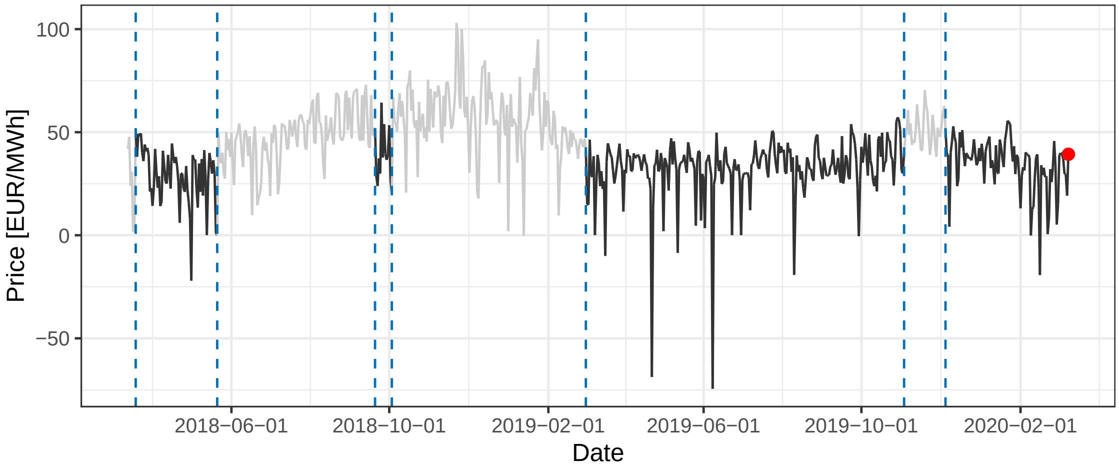


Figure 3.1: A sample run of the NOT-based calibration sample selection algorithm. The red dot is the target day. Vertical dashed lines indicate the located change-points separating periods with different statistical properties. The discarded prices are in gray. Source: **Paper 2**.

complicate the decision, electricity markets tend to exhibit heavy seasonality with several frequencies, as well as recurrently switching regimes (Avci et al., 2018).

Some long-term changes in the level or variance of data related to electricity markets can be explained with expert knowledge, such as regulatory laws or changes in the generation mix (opening new power plants). Others, e.g. weather related disturbances or changes related to geopolitical events, can be more difficult to identify. Using expert judgment can be a challenge in such situations, and nearly impossible if the forecast user does not have any access to analyses. This can be a case for small companies or prosumers. To that end, automated methods which rely on data rather than domain knowledge fill an important research gap.

In order to maximize forecasting power, the forecaster should ideally estimate model parameters on samples relevant to the current market situation. In other words, the observations in the calibration sample should belong to the same regime as the forecasted data point (de Marcos et al., 2020). An example of such segmentation based on identifying change-points can be seen in Figure 3.1. Electricity prices in the period leading up to the forecasted day are characterized by low mean level and frequent negative spikes. This situation has been observed several times in the past, interspersed with periods of higher prices. Estimating model parameters on the entire sample could lead to overestimating the price level.

There are different approaches that may be used to automatically identify past periods with similar statistical properties, i.e. belonging to the same regime. One may focus on the forecasted variable itself, e.g. electricity price, and aim to find continuous consecutive periods with particular properties. In electricity markets, such concepts have been addressed by regime switching models (e.g. Paraschiv et al., 2015) or change-point detection (de Marcos et al., 2020). Change-point detection methods typically assume that data can be modeled with a linear regression model, which has piecewise constant parameters on a set number of segments (Zeileis et al., 2003; Truong et al., 2020). Then, after identifying regimes, the data needs to be filtered in order to select subsamples belonging to the same regime.

Another idea for calibration sample selection would be to focus on the current market conditions, represented, for example, by explanatory variables. Such a filtering would be used to pick a set of individual data points rather than consecutive observations across a period of time. Similarity can be judged with clustering algorithms, which have been successfully used in electricity price and generation forecasting (Chaudhury et al., 2020; Yesilbudak et al., 2017).

3.1.2 Forecasting Electricity Prices: Autoregressive Hybrid Nearest Neighbors (ARHNN) Method (Paper 1)

As shown in Hubicka et al. (2019), the data used for model calibration has a major impact on forecast accuracy. The authors also noted that forecast quality may be further improved by averaging predictions obtained from estimating the same model on several calibration windows with varying lengths. However, the presented method has two main limitations. Firstly, the number and lengths of averaged calibration windows are selected based on expert domain knowledge and may not be appropriate in more general applications. Secondly, all considered calibration windows comprise consecutive most recent observations available at the time of forecasting.

Paper 1 aims to address both of these issues by introducing a hybrid model combining autoregressive models with exogenous variables (ARX models) with a machine learning algorithm: k-nearest neighbors. The k-nearest neighbors algorithm is used to find similarities between the independent variables of the ARX model for the forecasted data point and those within the historical observations, basing on Euclidean distance between vectors. In other words, it compares the values of past prices (autoregressive component) and fundamentals (forecasted load and RES generation) for the target observation with the corresponding sets of values from the past, aiming to find k most similar data points. This allows to estimate the model parameters on a subset of historical data which is most relevant to the current situation, through an automatic process which does not rely on user's judgment. An illustration of the process of sample selection is shown in Figure 3.2.

A downside of this approach is, yet again, its dependence on an arbitrarily chosen hyperparameter k, i.e. the size of the subset of most similar data points. In order to further automate the process and, at the same time, leverage the forecast combination technique, we introduce another step to the procedure, namely a validation window for

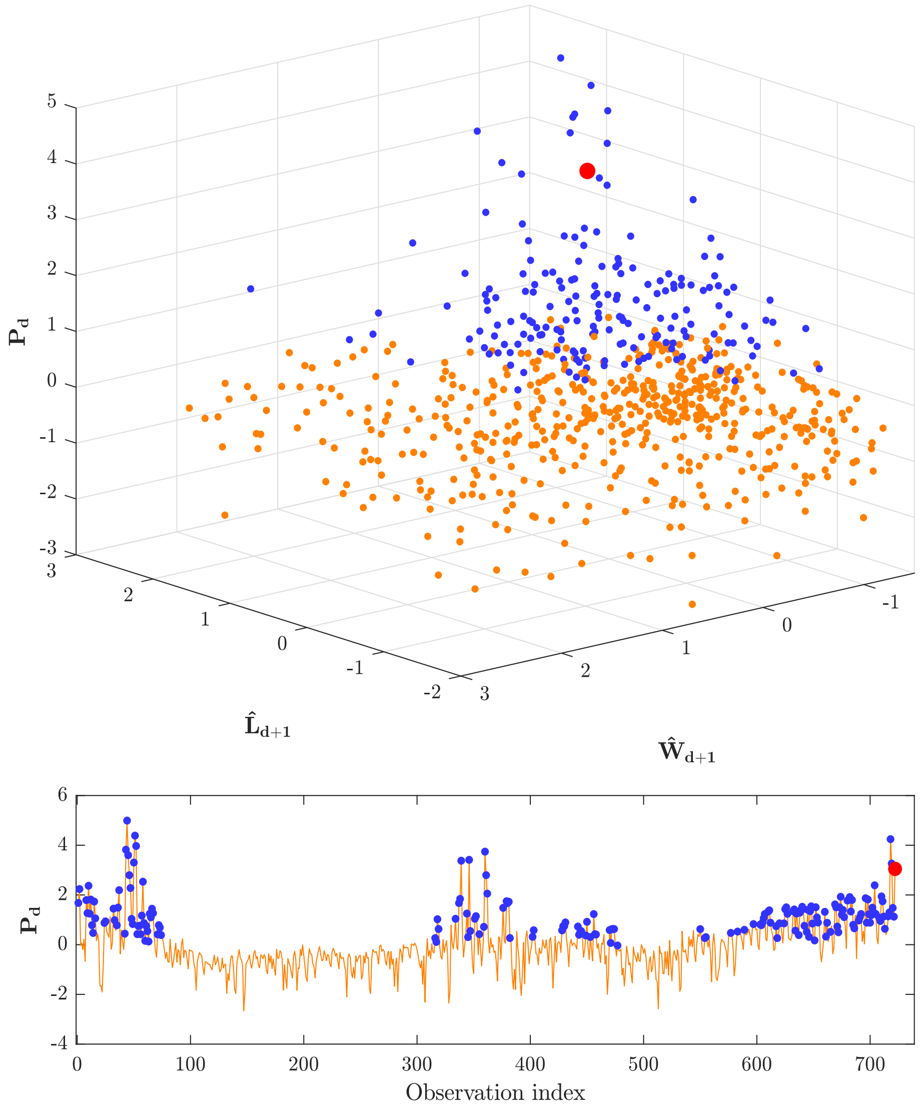


Figure 3.2: An example of optimal selection of the calibration sample. The upper panel illustrates the sample selection, presented on three key variables, i.e. preceding day's price, forecasts of load and wind generation; while the lower panel depicts the corresponding selection in the time dimension. The most recent observation is marked with a red dot, while the observations selected for the model calibration are depicted with blue points. Source: **Paper 1**.

exploring the impact of k on forecast accuracy. For each data point within that validation window, the k-nearest neighbors algorithm is applied for all possible values of k. Then, parameter k producing the most accurate prediction for that day is identified ex-post and saved. Finally, the set of optimal values of k from the validation window is used to produce an equally sized set of predictions for the target day, which are then averaged to obtain the final forecast.

The introduced procedure, called Autoregressive Hybrid Nearest Neighbors (ARHNN) is compared to several literature benchmarks using a day-ahead electricity price dataset from the German EPEX SPOT market. Aside from spot prices, the data includes official transmission system operator's day-ahead forecasts of load and renewable energy generation. The data covers six years, from January 2015 to December 2020, with hourly resolution. The data is tested on the final two years of the sample, with the remainder used for model estimation and validation.

In order to assess the impact of calibration sample selection on forecast accuracy, all methods use the same ARX model specification, differing in the size of the calibration sample and usage of forecast combination techniques. The structure of the model, based on expert knowledge, is as follows:

$$\begin{aligned} P_{d,h} = \underbrace{\alpha_h D_d}_{\text{Dummies}} + \underbrace{\sum_{p\in\{1,2,7\}} \beta_{h,p} P_{d-p,h}}_{\text{AR component}} + \underbrace{\theta_{h,1} P_{d-1,min} + \theta_{h,2} P_{d-1,max}}_{\text{Yesterday's price range}} + \\ + \underbrace{\theta_{h,3} P_{d-1,24}}_{\text{Last known price}} + \underbrace{\theta_{h,4} \hat{L}_{t,h} + \theta_{h,5} \hat{W}_{t,h} + \theta_{h,6} \hat{S}_{t,h}}_{\text{Exogenous variables}} + \varepsilon_{d,h}. \end{aligned} \tag{3.1}$$

The procedure is applied in a standard rolling window scheme. In other words, the models are retrained from scratch for each day in the test period. The forecasts are performed independently for all 24 hours, treating them as separate market products. For every subsequent day, the training data is updated to include the true observed values of the previously forecasted price and fundamentals, and remove one day of the oldest observations. This ensures that the training sample size stays constant. Such a setup reflects how forecasts may be performed in practice and is used in all research papers in this thesis.

Comparison in terms of the root mean squared error (RMSE) shows that the ARHNN algorithm outperforms the considered benchmarks. Additionally it can be observed that within the test period, the majority of optimal values of k produces very short or very long calibration samples, validating the expert approach proposed by Hubicka et al. (2019). The ARHNN algorithm is a novel algorithm combining well-established methods to better utilize historical data compared to an ARX-type model. It may be implemented in organizations which already use ARX-model generated point forecasts. ARHNN can also be easily modified to use other point forecasting methods, such as neural networks.

Aside from improving forecast accuracy, it can be used to gain insight into data, such as investigating optimal calibration sample sizes.

Publication details:

- Authors: Nitka, W., Serafin, T., Sotiros, D.
- Conference proceedings: Computational Science – ICCS 2021
- DOI: 10.1007/978-3-030-77970-2_24
- Publication year: 2021
- MNiSW: 140 pts, CORE A
- Contribution: 33%, model co-implementation and verification, co-writing, co-editing
- Citations according to Scopus: 3 (3 excluding self citations of all authors)

3.1.3 Calibration Window Selection Based on Change-Point Detection for Forecasting Electricity Prices (Paper 2)

The second paper approaches the topic of calibration sample selection from another angle, which is change-point detection. The motivation behind this approach is that changing geopolitical or economic situation in the electricity markets is reflected by structural changes in the price time series. With many time series models incorporating the assumption of stationarity and homoscedasticity in the entire sample, the existence of structural breaks may violate these assumptions, negatively impacting the models' suitability for this application.

Structural breaks analysis aims to mitigate this issue by identifying subperiods of time series with constant mean and variance, separated by change-points. This information can be used to select a relevant calibration sample. Prior to **Paper 2**, this idea has been rarely used in electricity price forecasting (de Marcos et al., 2020; Kaszuba, 2020).

While identifying a singular change-point is relatively straightforward, locating an unknown number of structural breaks is a non-trivial issue due to exponentially increasing computational complexity. Baranowski et al. (2019) propose an algorithm for generalized change-point detection and call it the narrowest-over-threshold (NOT). The idea behind NOT is to randomly draw subsamples from data, and use likelihood theory to locate a change-point in each subsample. Then, among subsamples where likelihood exceeds a certain user-defined threshold, the shortest one is chosen – rationalized as being the most likely to contain exactly one change-point. Those operations are then repeated recursively to locate a number of structural breaks up to a user-defined maximum. A major advantage of NOT compared to methods used by de Marcos et al. (2020) and Kaszuba (2020) is the possibility to modify the features used for change-point detection (e.g. only mean or mean and variance). This allows to adjust the algorithm to user's needs.

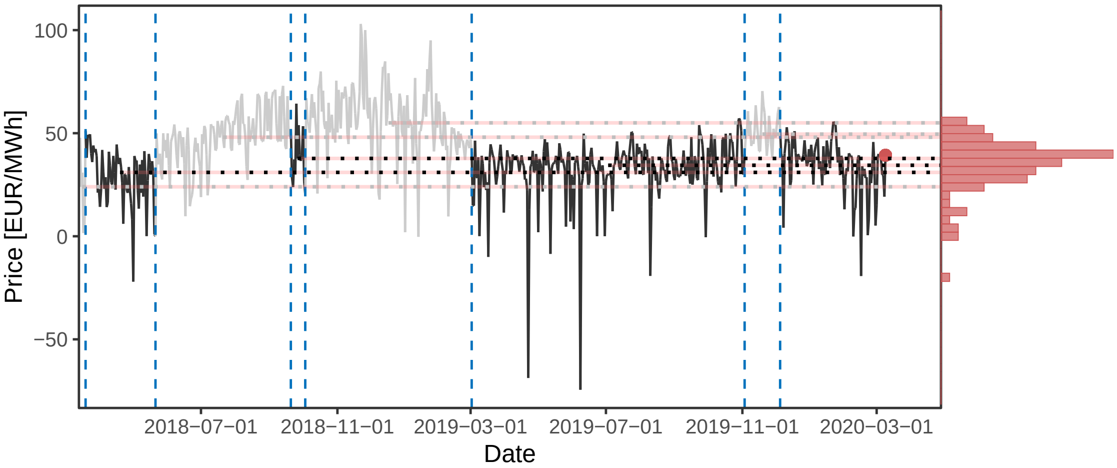


Figure 3.3: Graphical illustration of the NOT algorithm. The plot shows a sample forecasted day (red dot) and past two years of prices (continuous line). The detected change-points are marked with vertical dashed lines. The histogram on the right side of the plot approximates the distribution of prices in the "current period", while horizontal dotted lines mark the medians of each previous subperiod. Those falling in the center of the distribution (black lines) are included in the final calibration sample, while those on the extremes (gray lines) are discarded.

The flexibility of the NOT algorithm and its relatively low (near-linear) computational complexity are features attractive for applications in forecasting. **Paper 2** is the first research paper published in the energy forecasting literature which uses this approach. It proposes a procedure for forecasting electricity prices with an ARX model estimated on a filtered calibration sample. In the procedure, the NOT algorithm is applied to the initial calibration window of consecutive, most recent observations of electricity price, i.e. the dependent variable. If any change-points are identified, the period between the most recent change-point and the forecasted day is assumed to be the "current" market situation. Within this period, a symmetrical interval between extreme quantiles of the observed prices is computed. For all other subsamples between two consecutive change-points, their respective medians are computed. The final calibration sample consists of the most recent subsample and those among the remaining ones which are similar, defined as those where their median falls within the set quantile interval (in **Paper 2**: the middle 95% of the distribution). The procedure is illustrated in Fig. 3.3.

This procedure is applied to German day-ahead electricity prices from the EPEX SPOT market, with the data, the rolling window setup and the underlying autoregressive model, see Eq. (3.1), identical to **Paper 1**. The results are compared to several benchmarks, including ARHNN and combined forecasts from different windows. Additionally, the results are compared for raw data and data normalized through variance stabilizing *asinh* transformation (Uniejewski et al., 2018). The results show that using NOT is highly effective compared to a single-length calibration window, with the filtered calibration sample resulting in forecast accuracy higher than any single window of consecutive observations, even selected ex-post. However, when using more sophisticated approaches involving variance stabilizing transformations or forecast combination, the improvements from optimizing calibration sample are smaller and not statistically significant. An interesting pattern can be seen in the results of the procedure, where atypical, turbulent market conditions typically lead to choosing a very short calibration sample, while ordinary price behavior is correlated with much less selectiveness and longer calibration samples. Therefore, a user who needs point forecasts for irregular data, such as electricity prices, can achieve a higher forecast accuracy by automatically selecting relevant data. Additionally, this approach can offer additional insights for the user by revealing past subperiods with similar characteristics.

Publication details:

- Authors: Nasiadka, J., Nitka, W., Weron, R.
- Conference proceedings: Computational Science – ICCS 2022
- DOI: 10.1007/978-3-031-08757-8_24
- Publication year: 2022
- MNiSW: 140 pts, CORE A
- Contribution: 50%, including co-conceptualization, model co-implementation, visualization, co-writing
- Citations according to Scopus: 1 (1 excluding self citations of all authors)

3.2 Using resampling methods to generate predictive distributions of electricity prices and better assess uncertainty

3.2.1 Problem statement

The relative frequency of extreme events (e.g. positive and negative price spikes or supply shortages) in electricity markets requires participants to effectively manage their operational risk. In practice, the majority of trading decisions of electricity utilities are made on the basis of point forecasts (McGrath and Jonker, 2024; LEAG, 2024). While forecasting the expected values of variables of interest is the most common approach, such forecasts have limitations. They cannot convey uncertainty or the range of possible events. At the same time, the extreme price spikes are the main source of trading risks. Hence, forecasting them is highly valuable for electricity market participants (Hagfors et al., 2016a).

Especially small companies, with limited safety nets of resources and less know-how, are especially susceptible to risk when trading in power exchanges (Kraft et al., 2023).

Communicating uncertainty of forecasts can have a number of positive effects for planning and risk management. Studies comparing decision makers' effectiveness in managing extreme weather event risks showed that providing prediction intervals for forecasts results in higher decisiveness and smaller financial losses (Ramos et al., 2013; Savelli and Joslyn, 2013; Scoblic and Tetlock, 2021). However, trust in the forecast's quality and usefulness is a significant factor in contributing to these advantages (Goodwin, 2014). Costly and difficult to understand forecasts are less likely to be used in practice, diminishing their usefulness to support decisions (Yardley and Petropoulos, 2021).

The most commonly used methods for probabilistic forecasting produce marginal distributions of variables, which do not convey interactions between different quantities. However, for the purposes of decision-making processes, it is important to consider interdependencies of the set of forecasts, e.g. spatial relationship between different wind turbine locations (Pinson, 2013). Researchers in the field of electricity price forecasting have successfully developed trading strategies for electricity generators basing on multivariate distributions of prices and fundamentals (Lee et al., 2018; Toubeau et al., 2022). However, these methods rely on parametric multivariate Gaussian distributions or copulas, which may underperform when the observed data does not match the assumed distributions. For this reason, in real-world applications, non-parametric methods may prove more suitable.

One type of such a non-parametric approach is based on the concept of resampling. Resampling methods allow the forecaster to estimate uncertainty of a prediction with large flexibility with regards to the underlying model and data assumptions (Efron and Gong, 1983). The main idea of these methods is to divide the training sample into two or more subsets, which are then used respectively to make predictions and evaluate out-of-sample forecast accuracy. Resampling approaches include conformal prediction, bootstrap, jackknife and cross validation, among others (Barber et al., 2021; Kath and Ziel, 2021). However, these methods are mostly used in univariate scenarios. Multivariate probabilistic forecasts in energy forecasting are typically created using parametric approaches such as vector autoregressive (VAR) models (Maciejowska, 2022) or copulas (Manner et al., 2019; Toubeau et al., 2022). Hence, there exists a research gap in the area of non-parametric multivariate probabilistic forecasts.

3.2.2 Multiple Split approach – multidimensional probabilistic forecasting of electricity markets (Paper 3)

Price spikes, especially negative, are a major source of risk for a small or medium renewable energy generator. Information about the likelihood of extreme events given by probabilistic forecasts of prices may reduce uncertainty, thus aiding them in risk management. However, for a utility with uncertain generation, such as a wind farm, decisions made using only price forecasts may not be optimal. Probabilistic forecasts of demand or generation level, considered together with predictions of prices, are relevant to successful trading.

Paper 3 aims to support the decision-making process of such a renewable energy utility by proposing a data-based approach for informed market bidding. The goal of this algorithm is to mitigate financial risks by optimizing the amount of energy offered in the day-ahead market. With this topic being of great practical interest, has been increasingly explored in the recent scientific literature (see e.g. Maciejowska et al., 2019; Finnah et al., 2022; Kraft et al., 2023). Nevertheless, the vast majority of research focuses on point forecasts or univariate probabilistic forecasts, with multivariate probabilistic forecasting being still underrepresented.

In this study we propose and apply a novel Multiple Split method to produce joint probabilistic forecasts of day-ahead and intraday market prices, and renewable energy generation. The approach builds upon and extends previous literature related to resampling methods such as conformal prediction (Kath and Ziel, 2021) and jackknife+ (Barber et al., 2021). It is interpretable and versatile with regards to the underlying forecasting model used, which makes it suitable for a practical implementation.

In the Multiple Split method the training data is randomly split into two disjoint subsets: the estimation sample and the calibration sample. The estimation sample is used to estimate parameters of a point forecasting model, e.g. an autoregressive model, simultaneously for all variables of interest. The parameters are then used to forecast both the unknown future observation as well as the remaining training data, i.e. the calibration sample. Since the observations from the calibration sample are known to the forecaster yet have not been used to estimate parameters, the prediction errors of these data points can be used as an estimate of the uncertainty of the forecast itself. The probabilistic forecast then takes the form of an ensemble of simulated realizations. A schematic representation of the algorithm is shown in Figure 3.4.

The construction of the Multiple Split method offers several advantages over other probabilistic forecasting methods. Firstly, while the study in **Paper 3** uses linear ARX (autoregressive with exogenous variables) models to produce the underlying point forecasts, most other point forecasting models can be used in their place. This property, shared with other resampling methods, allows potential users to easily incorporate the method into already existing processes. Secondly, joint estimation of uncertainty for multiple variables in a single split preserves the correlation structure of their prediction errors. As a result, the final ensemble represents a multivariate distribution of the forecasted vector of variables rather than their marginal distributions. This property can be leveraged

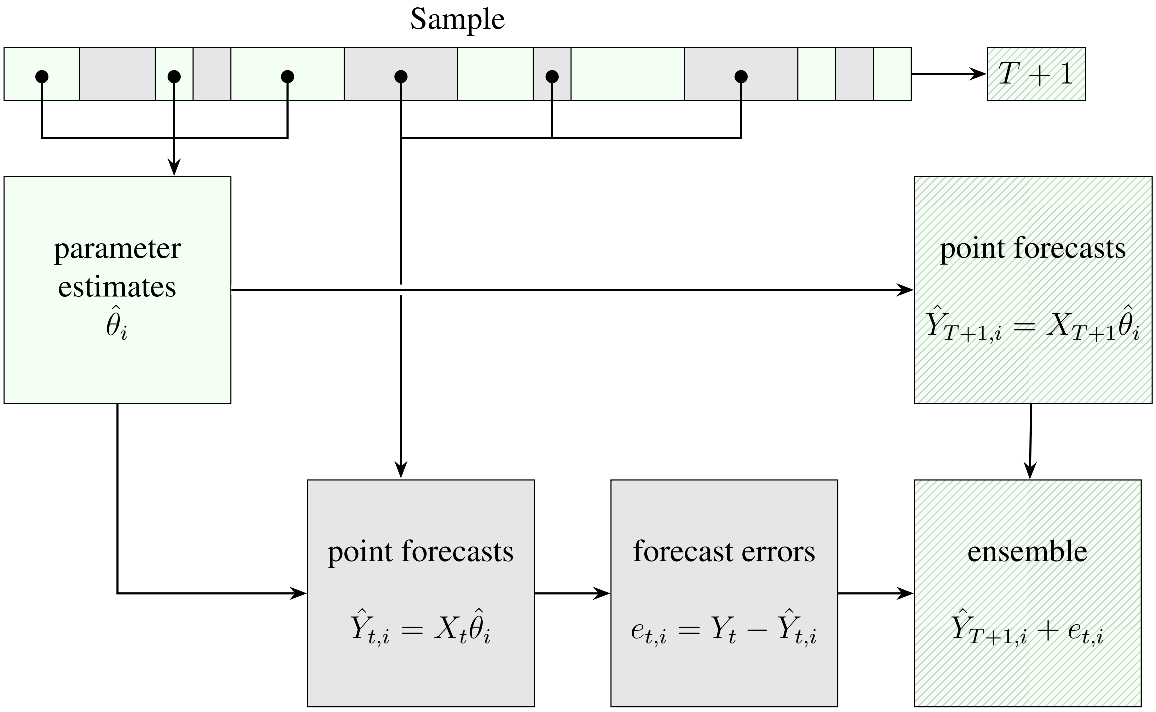

Sample
$T+1$
parameter estimates $\hat{\theta}_i$
point forecasts $\hat{Y}_{T+1,i} = X_{T+1}\hat{\theta}_i$
point forecasts $\hat{Y}_{t,i} = X_t\hat{\theta}_i$
forecast errors $e_{t,i} = Y_t - \hat{Y}_{t,i}$
ensemble $\hat{Y}_{T+1,i} + e_{t,i}$

Figure 3.4: Schematic illustration of the Multiple Split algorithm. Source: **Paper 3**.

to produce forecasts of functions of original variables, e.g. their linear combinations, for more sophisticated uses. A proposed application of this property is described in Section 3.3.3.

The data used in this study comes from the German EPEX SPOT market and covers a four-year-long period between October 2015 and September 2019. Due to a limited access to intraday market prices necessary in this case study, the analyzed period is shorter than in **Papers 1** and **2**. The ARX model for day-ahead prices is similar to the one shown in Eq. 3.1, with additional variables included to leverage more information. Specifically, the model is extended with lagged prices from 3 to 6 days ago, as well as fuel (coal and gas) index prices. The last known price is replaced with the entire previous day's average. The models for intraday prices, price spread and load are similarly specified, while the forecasts of wind generation use a much simpler model due to lack of seasonality or dependence on human activity.

Paper 3 assesses the performance of the Multiple Split method compared to other probabilistic forecasting methods: historical simulation and quantile regression. The accuracy measures evaluate three aspects of proabailistic forecasts:

- the coverage of the prediction intervals via distance to the nominal coverage and the Kupiec (1995) test,
- sharpness and calibration of the predictive distributions via the continuous ranked probability score (CRPS Gneiting and Raftery, 2007)),

- multivariate calibration of ensembles via the reliability index (Gneiting et al., 2008).

Results show that the Multiple Split method is better calibrated than its competitors with respect to the coverage of wide prediction intervals and the reliability index, while falling behind in terms of the CRPS. This indicates its suitability to predict extreme events. Additionally, the focus on correlation of errors allows the method to outperform other approaches when forecasting functions of several variables.

Publication details:

- Authors: Maciejowska, K., Nitka, W.
- Journal: preprint (arXiv)
- DOI: 10.48550/arXiv.2407.07795
- Publication year: 2024
- Contribution: 50%, including model co-implementation and verification, visualization, co-writing, co-editing

3.3 Utilizing renewable generation and load forecasts to design trading strategies in day-ahead and intraday markets

3.3.1 Problem statement

As described in more detail in Section 2.2, renewable energy is an important part of the generation mix in many European countries, and its prevalence continuously grows. Renewables are an important factor in price formation in electrical energy markets (Hagfors et al., 2016b; Maciejowska, 2020; Rai and Nunn, 2020). With that in mind, it should not be surprising that renewable energy generation is commonly included as an explanatory variable in statistical electricity price forecasting models (Gianfreda et al., 2020).

A company whose primary business activity is electricity trading can base their strategies solely on price forecasts. They can be only day-ahead market prices or include prices from real-time markets (see e.g. Maciejowska et al., 2019; Serafin et al., 2022). However, the majority of electricity market participants are generating or consuming energy. For these companies, additional variables are highly relevant, such as electricity demand, total generation or renewable generation (Zugno et al., 2013; Alipour et al., 2019). Including them in the decision-making process can improve the company's financial results.

Considering fundamental variables in decision-making processes is especially important for renewable energy generators, who need to manage their output uncertainty. Failure to do so accurately may lead to significant financial consequences, either due to price changes in real-time markets, or due to imbalance penalties from regulatory organs. Conversely, reducing imbalance of supply and demand may lead to substantial profits, thanks to market mechanisms such as flexibility premiums. Hence, energy utilities that cannot hedge or diversify their outputs have a need for optimized bidding strategies, aimed to reduce their financial risks (Zugno et al., 2013). Due to large volatility of electricity prices, risk management methods designed for other commodities are likely to not be suitable for their purposes (Westgaard et al., 2019). Therefore, designing trading strategies tailored to the needs of electricity market participants is a problem of high practical importance.

When evaluating forecasts used in trading, it has been shown that their statistical accuracy does not necessarily coincide with their economic or financial benefits (Maciejowska et al., 2019; Yardley and Petropoulos, 2021). When comparing several different forecasts, improving accuracy tends to increase business value. However, forecasts with the highest accuracy may not lead to the best operational decisions (Robette, 2023). In the light of finance-related goals of renewable energy producers, oftentimes optimal bids significantly differ from point predictions of their energy output. With different economic costs of over- and underestimating forecasts, quantiles have been shown to work well as point forecasts (Gneiting, 2011). Designing trading strategies and evaluating them on historical data can help forecast users choose methods suitable to their decision-making processes.

The tangible financial benefits resulting from the use of data-driven forecasts and analytics methods make them more attractive to managers and other end users, allowing for their real-life implementations (Hong, 2016). On the other hand, statistical accuracy measures can be considered to be more objective, compared to economic benefits which are highly dependent on the given assumptions and forecast user's needs (Kolassa, 2023). Therefore, it is best to consider and present both aspects of forecast evaluation in research papers.

3.3.2 Enhancing load, wind and solar generation for day-ahead forecasting of electricity prices (Paper 4)

The research presented in **Paper 4** has been motivated by earlier work of Maciejowska et al. (2019). It tackles designing a trading strategy for a small renewable energy generator, such as a wind farm. It is assumed that this company is too small to participate in the power exchange directly, and instead is managed by an intermediary, e.g. a larger trading company. However, they can autonomously decide on the amount of energy offered in the market, as well as the type of auction they participate in: day-ahead or intraday trading. Such a problem setup is increasingly relevant with the rise in popularity of prosumer installations or renewable energy communities (LEAG, 2024).

Clearly, at the time of the day-ahead bidding, the actual values of energy generation for the traded time periods are yet unknown. While large market participants usually have access to proprietary data about electricity production, a small producer will likely rely on publicly reported information. Transmission system operators (TSO) often publish official forecasts of electricity demand and renewable generation for entire countries or bidding zones, with a day-ahead or longer forecast horizon.

While evaluating historical data from the German EPEX SPOT market it was found that the TSO forecasts for electricity demand, wind energy generation and solar energy generation (i.e. fundamental variables) are systematically biased and hence exhibit large prediction errors. This observation contrasts with the most common literature approach, which assumes that TSO forecasts are accurate. Therefore, the main objectives of **Paper 4** are twofold: firstly to improve upon the TSO forecasts with the use of statistical models; and secondly, to investigate whether using such enhanced predictions leads to higher accuracy of point forecasts.

Simple ARX-type models, defined as in Eq. (3.1) with an additional independent variable representing previous day's average price, are used for forecasting both the fundamental variables and day-ahead and intraday market prices. The dataset used in this research is identical to that of **Paper 3**. In the case of fundamental variables, all obtained forecasts are more accurate than the original TSO predictions. The most significant improvement can be seen in the case of load forecasts, indicating that the TSO may not effectively utilize all information available in the market. The decrease in forecast error measured by MAE and RMSE is smaller for renewable energy generation forecasts, where more sophisticated models or additional information may be necessary.

The results of price forecasting experiment reveal an interesting insight into price formation mechanisms. The forecast accuracy for both markets is evaluated in three scenarios: with the exogenous variables being the original TSO forecasts, improved forecasts or "crystal ball" predictions, i.e. real values. In the case of intraday market prices, any improvement of the TSO forecast quality results in a more accurate price forecast. In other words, the intraday prices are established shortly before actual delivery basing on realized load and renewable generation. However, this is not the case for the day-ahead market price. In the latter case, having a perfect prediction of fundamentals does not help in predicting day-ahead prices, as all market participants make their decisions using forecasted values.

In order to evaluate the economic value of the enhanced forecasts, we consider the decision process of a company that consistently generates 1 MWh of electricity. On the day preceding physical delivery, before the settlement of the day-ahead auction, they choose whether to sell the electricity in the day-ahead or intraday market. The decision is based on the predicted sign of the price spread between these two markets, i.e. the market chosen is the one with higher expected price. This approach is compared to the naive strategy of always selling 1 MWh in the day-ahead market. With the statistically improved forecasts of fundamentals, the company can increase their yearly revenue by 13%. With perfect information, the maximum possible profit from optimizing market choice was approximately 300% higher than the benchmark, highlighting the importance of improving forecast accuracy, especially in the intraday market.

Publication details:

- Authors: Maciejowska, K., Nitka, W., Weron, T.
- Journal: Energy Economics 99, 105273
- DOI: 10.1016/j.eneco.2021.105273
- Publication year: 2021
- JCR Impact Factor: 9.252
- JCR classification(s): Economics
- MNiSW: 200 pts, assigned to the Management and Quality Studies (NZJ) discipline
- Contribution: 33%, including model co-implementation and verification, visualization, co-writing, co-editing
- Citations according to Scopus: 41 (38 excluding self citations of all authors)

3.3.3 Multiple Split approach – multidimensional probabilistic forecasting of electricity markets (Paper 3 cont.)

Section 3.2.2 introduced the Multiple Split method proposed in **Paper 3**. The Multiple Split method allows for non-parametric joint probabilistic forecasting of multiple variables, yielding a forecast in the form of an ensemble of points. The obtained multivariate forecast preserves the correlation structure between variables, allowing for the forecast to convey additional information. For example, an accurate forecast of a linear combination of several variables (e.g. price spread or residual load) can be derived, which is not possible with only marginal distribution forecasts.

In **Paper 3**, we leverage this property to construct several market strategies based on joint probabilistic forecasts of electricity market prices and wind energy generation. The forecasting experiment considers a utility that owns several wind farms spread across Germany. It is assumed that its generation is proportional to the total market production. This assumption is made due to data availability, and to generalize results for a broader potential audience.

The utility generates an unknown amount of energy every hour and sells it to the market in its entirety. They need to decide how large of a share of forecasted generation (q) should be offered in the day-ahead market. It is assumed that the remaining amount and potential forecast error are balanced (sold or purchased if needed) in the intraday market. The amount q of energy sold, defined as a percentage of predicted total generation, is

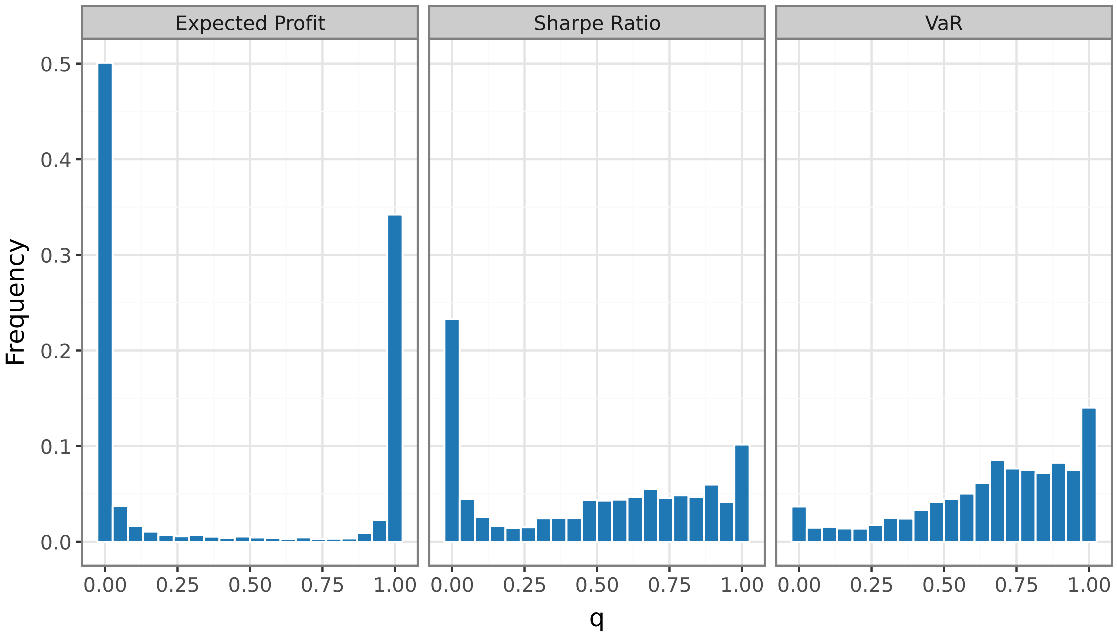


Figure 3.5: Histogram of optimal values of q for different data-driven strategies, where q is the fraction of predicted generation sold in the day-ahead market (thus $q = 1$ is the "naive" benchmark, $q = 1$ – selling entire generation in the intraday market). It can be seen that aiming to maximize profits typically leads to preferring the market with higher predicted price, while optimizing value-at-risk leads to hedging the bids $(0 < q < 1)$. Source: **Paper 3**.

decided one day before the delivery. It is chosen to optimize the median, value-at-risk (VaR) or Sharpe ratio of the utility's forecasted profit. Probabilistic forecasts of profit can be based on the joint distribution of market prices and generation. The distribution of the optimized bids q for different strategies is shown in Fig. 3.5. The strategies are compared to "naive" benchmarks, which assume that the utility does not adjust their bids contitional on market situation, rather choosing to bid the entire forecasted generation in the day-ahead market. Although this topic is highly relevant for practitioners, strategies basing on probabilistic forecasts of profits have not been previously explored in the electricity price forecasting literature.

Another layer of the decision-making process is inspired by a phenomenon observed in reality, where under certain circumstances renewable energy producers may be forced to withdraw from market participation (Bird et al., 2016). The curtailment may be caused by weather conditions, such as too light or heavy wind, as well as market events, e.g. negative price levels. With that in mind, the decision-making process of the considered company includes a possibility to turn off the generators (e.g. wind turbines) for a certain amount of time. This decision is based on the predicted distribution of income. More precisely, according to the utility's risk appetite, a certain likelihood of incurring losses leads to shutting down the production.

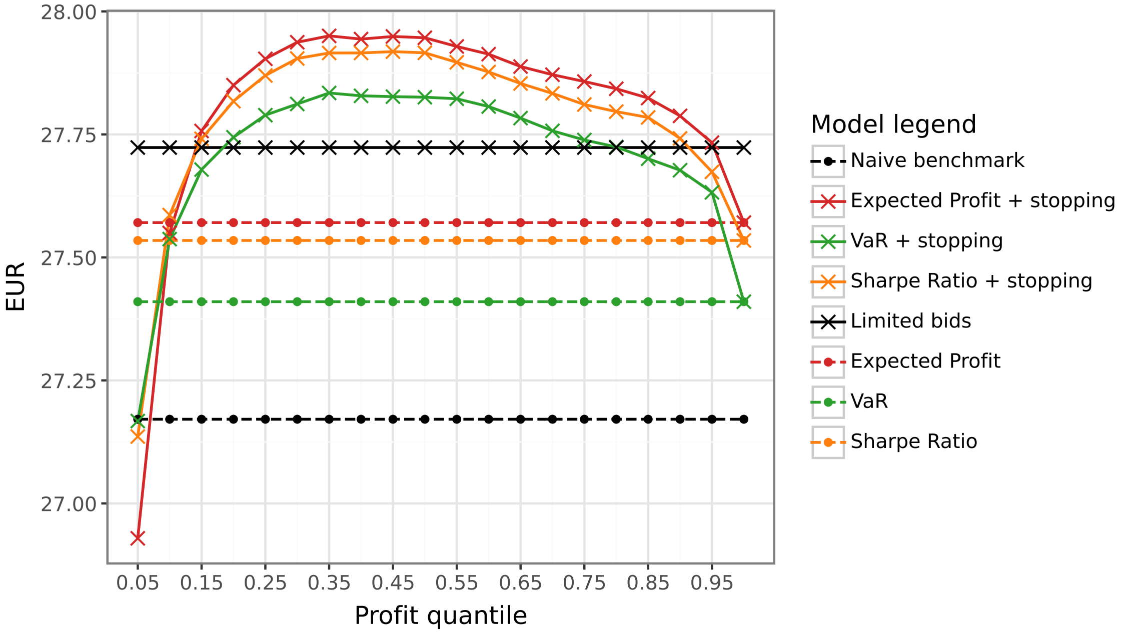


Figure 3.6: Average profit per 1 MWh of generation for two types of strategies: with production curtailment (marked with crosses) and without (marked with dots); depending on risk aversion level. The risk aversion is defined as percentile of income distribution. Production is curtailed if the considered quantile of predicted income is negative. Operation & Management costs are 10 Euro/MWh. Source: **Paper 3**.

We apply the discussed strategies for a model company operating in the EPEX SPOT market in order to evaluate the financial gains in terms of expected profit and value at risk. We show that the proposed methods can be successfully used to make data-driven business decisions, improving over strategies basing on point forecasts. Moreover, we illustrate that in this scenario there exists a trade-off between the total profit and financial risk. However, a moderate risk appetite leads to the maximal profit per 1 MWh (see Fig. 3.6).

3.4 Development of decision support methods for day-ahead bidding that use combinations of predictive distributions

3.4.1 Problem statement

Frequently, decision makers may have access to not one, but multiple forecasts of the same variable. They may come from human expert predictions, internal models or third-party providers (LEAG, 2024). Regardless of their source, the task of selecting the best single forecast ex-ante is a daunting one. Changing market situation may be reflected in varying performance over time, sometimes making historically successful forecasts less relevant.

Even with a single model specification, the output forecast can differ between forecasters. In case of machine learning models, this may be caused by stochasticity in hyperparameter optimization runs (Lago et al., 2021). Alternatively, as discussed in greater detail in Section 3.1.1, both traditional statistical models and machine learning methods produce different predictions depending on the observations used for calibration. As noted by Hubicka et al. (2019), predictions from long calibration samples are typically more robust, while short calibration samples allow for quick adaptation to evolving market conditions.

A possible solution to the problem is using forecast combinations, or ensemble forecasts. According to Timmermann (2006), diversification of information offered by ensemble forecasts is, in practice, often preferable to selecting a single forecast, unless it offers definite and significant advantages over its competitors.

The idea of combining forecasts motivated by analogies to portfolio diversification dates at least to the 1960's (Bates and Granger, 1969). It has recently been gaining popularity in the field of energy price forecasting, especially when it comes to probabilistic forecast combinations (Hong et al., 2020). Compared to point forecasts, combining probabilistic forecasts requires multiple times as many parameters. At each time point, every individual forecast has a range of values rather than only the point prediction, which may be combined with identical or diverse weights. The direction of averaging – e.g. for quantile forecasts, whether to combine quantiles or probabilities – has to be chosen as well (Uniejewski et al., 2019).

Within the scientific literature, many approaches to forecast combinations have been developed. Typically, "naive" approaches of assigning equal weights to each forecast are difficult to outperform due to estimation uncertainty (Blanc and Setzer, 2020). However, literature proposes ways to modify weights according to past performance of the forecasts. An exhaustive review of ensemble forecasting can be seen in Wang et al. (2022).

3.4.2 Combining predictive distributions of electricity prices. Does minimizing the CRPS lead to optimal decisions in day-ahead bidding? (Paper 5)

Forecast combination has been shown to be a simple but highly effective method to improve forecast accuracy. This technique finds increasingly common uses in electricity price forecasting, e.g. in **Papers 1** and **2**. However, in electricity price forecasting literature there is still a lack of conclusive recommendations about the sizes of forecast pools used for averaging or the optimal weighting methods. The existing research suggests that small ensembles of well-performing individual forecasts (i.e. experts) are preferred (Wang et al., 2022), and that optimizing weights does not offer additional benefits unless

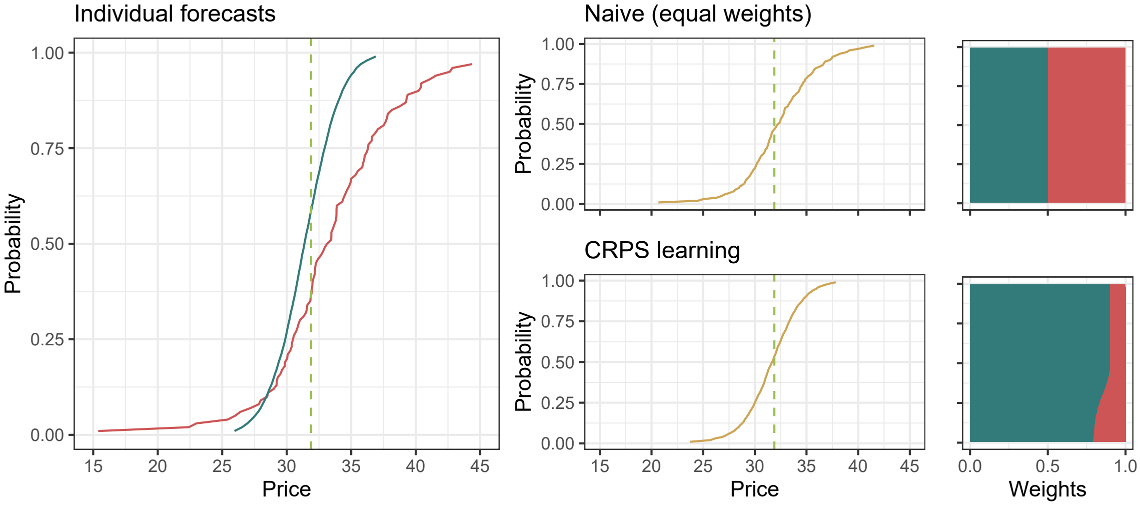


Figure 3.7: Illustration of the two weighting schemes. The left panel shows predictive distributions obtained from the DDNN_JSU_1 (teal color) and LEAR_QRA (red color) models for a selected hour and day. The center panels present the resulting ensemble forecasts obtained by estimating weights with naive (top) and CRPS learning (bottom) methods. The right panels illustrate the relative weights for each quantile; these are horizontal stacked bar plots with the length of the bar representing the weight of the forecast in corresponding color and all weights summing up to 1. The dashed vertical line marks the actual price. Source: **Paper 5**.

the ensemble is sufficiently diverse (Blanc and Setzer, 2020).

Paper 5 aims to address this research gap with a case study using German day-ahead electricity prices from the EPEX SPOT market. The main objective is investigating the impact of different methods of constructing ensemble forecasts on two aspects of forecast goodness: the accuracy of the combined forecasts and their economic benefits in a simulated trading scenario. To this end, a pool of probabilistic quantile forecasts is used (Marcjasz et al., 2023). Out of twelve forecasts in total, eight of them are produced by state-of-the-art distributional deep neural network (DDNN) models with different hyperparameter setups, while the remaining four are quantile regression-type models, well established in the literature and used in the aforementioned paper as benchmarks.

With such a pool of expert forecasts, two main aspects of forecast combination are tackled: selecting the pool of experts and optimizing weights. With the large amount of possible combinations of individual forecasts, some limitations are introduced to the process of selection. Firstly, following recommendations from (Marcjasz et al., 2023), every final ensemble includes a set of four DDNN forecasts. They may be supplemented by a set of two quantile regression forecasts, making the final ensembles contain four to six individual forecasts.

For the task of assigning weights, two approaches are considered. The first and most natural is the "naive" weighting, which assigns equal weights for all experts across all

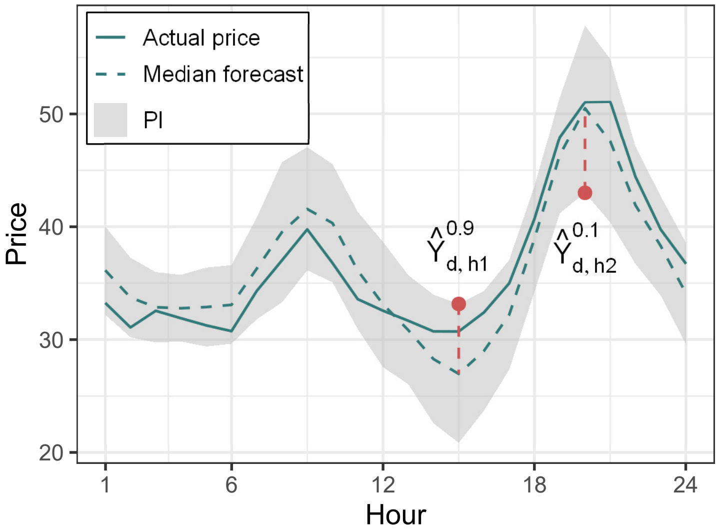


Figure 3.8: Illustration of the trading strategy with limit orders defined by the 80% PIs, corresponding to a risk appetite of 0.8. Red dots indicate the price limits for the selected hours. Source: **Paper 5**.

quantiles and time periods. The alternative weighting is achieved through applying CRPS learning method (Berrisch and Ziel, 2023). It relies on calculating weights basing on each expert's past performance measured in terms of the CRPS (see Sec. 3.2.2). The procedure allows for assigning different weights to forecasts of different quantiles, as well as updating them over time. This ensures that the experts which are more effective in predicting certain parts of distributions are assigned higher weights than those less accurate. A toy example comparing the results of both weighting approaches is shown in Figure 3.7.

The obtained ensemble forecasts are evaluated using a number of error measures. The CRPS is used to compare entire distributions, while the point forecasting performance is measured using MAE and RMSE. The results confirm that among the created combination forecasts, even the worst performing ensemble is more accurate than the best individual forecast. Furthermore, the choice of averaged forecasts is a more important factor than weighting. However, using CRPS learning to optimize combination weights allows to obtain small but statistically significant improvements.

The economic benefits of using ensemble forecasts for decision support are assessed by applying a simulated trading scenario using a battery, first proposed by Uniejewski (2024). In this scenario, the user aims to gain profits by buying and subsequently selling 1 MWh of energy every day in the day-ahead market (see Figure 3.8). The user selects the appropriate hours for their transactions basing on the medians of forecasted prices, while the tails of the distribution are used for risk management: setting limits on the bids when the transaction is expected to be profitable, and foregoing trading when incurring a loss is likely. The results of this simulation indicate that greater forecast accuracy is generally correlated with achieving higher profits. However, small improvements of accuracy gained from using CRPS learning do not necessarily improve profits even further.

Publication details:

- Authors: Nitka, W., Weron, R.
- Journal: Operations Research and Decisions 33(3), 105–118
- DOI: 10.37190/ord230307
- Publication year: 2023
- JCR Impact Factor: 0.7
- JCR classification(s): Operations Research & Management Science
- MNiSW: 200 pts, assigned to the Management and Quality Studies (NZJ) discipline
- Contribution: 75%, including model implementation, visualization, co-writing
- Citations according to Scopus: 3 (3 excluding self citations of all authors)

Chapter 4

Conclusions

Within the evolving landscape of electrical energy markets, one major change that has been observed is the increasing number of small companies and prosumers participating in trading. While such traders typically have simpler decision-making processes than large market players, they also tend to have less experience and resources at their disposal. At the same time, their financial safety reserves tend to be smaller, making market risks more dangerous.

The objectives set in this thesis aim to support decision-making in small and medium sized enterprises (SMEs) through data-driven forecasting. The thesis achieved this goal by proposing novel forecasting methods and evaluating them from the point of view of practical applications. To meet the needs of such decision makers, the methods had to fulfill several criteria: high level of automation, ease of implementation and interpretation, as well as relatively low computational complexity. Their viability is evaluated using historical data and practical examples.

The thesis comprises five research articles. **Paper 4** explores the idea of improving electricity price forecasts from simple linear models by enhancing predictions of the fundamental variables used as inputs to the models. **Papers 1** and **2** propose novel methods that increase forecasting accuracy through automatic selection of calibration samples: Autoregressive Hybrid Nearest Neighbors (ARHNN) and an algorithm based on Narrowest-Over-Threshold (NOT) change-point detection. Finally, **Papers 3** and **5** deal with post-processing of forecasts, by resampling forecast errors or averaging forecasts from several models. More specifically, **Paper 3** proposes a novel Multiple Split procedure for multivariate probabilistic forecasting that takes into account correlations between variables, while **Paper 5** uses CRPS learning, a cutting-edge method for combining probabilistic forecasts based on their past performance.

These five papers constitute a significant addition to the operations research literature by demonstrating how relatively simple statistical models can be used efficiently for electricity price forecasting. While often outperformed by more sophisticated forecasting methods, they offer low computational complexity and are well known among practitioners. The novel approaches proposed in this thesis are meant to extend and accompany such simple models. They allow to improve forecast accuracy and business value at a low cost in terms of money and user expertise. Thus, the thesis achieves its main goal of providing SMEs with forecasting methods designed to support their decision-making processes.

Finally, in accordance with current guidelines in operations research, the majority of the research papers in this thesis include an evaluation of the economic benefits from using the generated forecasts. By describing their potential use in trading strategies and assessing the financial benefits, they demonstrate the tangible value of this research to interested users.

Bibliography

Alipour, M., Zare, K., Zareipour, H., Seyedi, H., 2019. Hedging strategies for heat and electricity consumers in the presence of real-time demand response programs. IEEE Transactions on Sustainable Energy 10, 1262–1270.

Avci, E., Ketter, W., van Heck, E., 2018. Managing electricity price modeling risk via ensemble forecasting: The case of Turkey. Energy Policy 123, 390–403.

Baranowski, R., Chen, Y., Fryzlewicz, P., 2019. Narrowest-over-threshold detection of multiple change points and change-point-like features. Journal of the Royal Statistical Society. Series B: Statistical Methodology 81, 649–672. arXiv:1609.00293.

Barber, R.F., Candès, E.J., Ramdas, A., Tibshirani, R.J., 2021. Predictive inference with the jackknife+. The Annals of Statistics 49, 486–507.

Bates, J.M., Granger, C.W.J., 1969. The combination of forecasts. Journal of the Operational Research Society 20, 451–468.

Bento, P., Mariano, S., Carvalho, P., Calado, M.d.R., Pombo, J., 2023. Soaring electricity prices in the day-ahead Iberian market: policy insights, regulatory challenges and lack of system flexibility. International Journal of Energy Sector Management 18, 312–333.

Berrisch, J., Ziel, F., 2023. CRPS learning. Journal of Econometrics 237, 105221.

Bird, L., Lew, D., Milligan, M., Carlini, E.M., Estanqueiro, A., et al., 2016. Wind and solar energy curtailment: A review of international experience. Renewable and Sustainable Energy Reviews 65, 577–586.

Będowska-Sójka, B., Demir, E., Zaremba, A., 2022. Hedging geopolitical risks with different asset classes: A focus on the Russian invasion of Ukraine. Finance Research Letters 50, 103192.

Blanc, S.M., Setzer, T., 2020. Bias–variance trade-off and shrinkage of weights in forecast combination. Management Science 66, 5720–5737.

BNetzA, 2023. Bundesnetzagentur: Monitoring report 2023 in accordance with Energy Act. Technical Report. BNetzA.

Burger, M., Schindlmayr, G., Graeber, B., 2014. Managing energy risk: a practical guide for risk management in power, gas and other energy markets (2nd ed.). Wiley Finance Series. second edition ed., Wiley, Chichester, England.

Burger, P.D.B., 2024. Energy-Charts. https://www.energy-charts.info/index.html.

Chaudhury, P., Tyagi, A., Shanmugam, P.K., 2020. Comparison of various machine learning algorithms for predicting energy price in open electricity market, in: 2020 International Conference and Utility Exhibition on Energy, Environment and Climate Change (ICUE), pp. 1–7.

Cludius, J., Hermann, H., Matthes, F.C., Graichen, V., 2014. The merit order effect of wind and photovoltaic electricity generation in Germany 2008–2016: Estimation and distributional implications. Energy Economics 44, 302–313.

Cullen, J.A., Reynolds, S.S., 2023. Market dynamics and investment in the electricity sector. International Journal of Industrial Organization 89, 102954.

Denholm, P., O'Connell, M., Brinkman, G., Jorgenson, J., 2015. Overgeneration from solar energy in California. A field guide to the duck chart. Technical Report NREL/TP–6A20-65023, 1226167. NREL.

EEX, 2024. Trading Products: EPEX SPOT. https://www.epexspot.com/-en/tradingproducts.

Efron, B., Gong, G., 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. The American Statistician 37, 36–48.

Fabbiani, E., 2024. Simplifying forecasts with teamwork: The power of ensemble models. https://medium.com/@donlelef/simplifying-forecasts-with-teamwork-the-power-of-ensemble-models-cc6f9210b3de.

Fabra, N., Imelda, 2023. Market power and price exposure: Learning from changes in renewable energy regulation. American Economic Journal: Economic Policy 15, 323–358.

Fałdziński, M., Fiszeder, P., Orzeszko, W., 2021. Forecasting volatility of energy commodities: Comparison of GARCH models with Support Vector Regression. Energies 14, 6.

Finnah, B., Gönsch, J., Ziel, F., 2022. Integrated day-ahead and intraday self-schedule bidding for energy storage systems using approximate dynamic programming. European Journal of Operational Research 301, 726–746.

Gianfreda, A., Ravazzolo, F., Rossini, L., 2020. Comparing the forecasting performances of linear models for electricity prices with high RES penetration. International Journal of Forecasting 36, 974–986.

Gneiting, T., 2011. Quantiles as optimal point forecasts. International Journal of Forecasting 27, 197–207.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102, 359–378.

Gneiting, T., Stanberry, L.I., Grimit, E.P., Held, L., Johnson, N.A., 2008. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. TEST 17, 211–235.

Goodwin, P., 2014. Getting real about uncertainty. Foresight: The International Journal of Applied Forecasting , 4–7.

Greene, W.H., 2012. Econometric Analysis. 7th ed ed., Prentice Hall, Boston.

Hagfors, L.I., Kamperud, H.H., Paraschiv, F., Prokopczuk, M., Sator, A., Westgaard, S., 2016a. Prediction of extreme price occurrences in the German day-ahead electricity market. Quantitative Finance 16, 1929–1948.

Hagfors, L.I., Paraschiv, F., Molnar, P., Westgaard, S., 2016b. Using quantile regression to analyze the effect of renewables on EEX price formation. Renewable Energy and Environmental Sustainability 1, 32.

Hong, T., 2016. Crystal ball lessons in predictive analytics. Energy Central .

Hong, T., Pinson, P., Wang, Y., Weron, R., Yang, D., Zareipour, H., 2020. Energy forecasting: A review and outlook. IEEE Open Access Journal of Power and Energy 7, 376–388.

Hu, J., Harmsen, R., Crijns-Graus, W., Worrell, E., van den Broek, M., 2018. Identifying barriers to large-scale integration of variable renewable electricity into the electricity market: A literature review of market design. Renewable and Sustainable Energy Reviews 81, 2181–2195.

Hubicka, K., Marcjasz, G., Weron, R., 2019. A note on averaging day-ahead electricity price forecasts across calibration windows. IEEE Transactions on Sustainable Energy 10, 321–323.

IEA, 2020. Projected costs of generating electricity 2020. Technical Report. International Energy Agency.

IEA, 2023. Electricity Market Report 2023. Technical Report. Gas, Coal and Power Markets Division of the International Energy Agency (IEA).

Intini, M., Waterson, M., 2023. Strategic behaviour by wind generators: An empirical investigation. International Journal of Industrial Organization 89, 102947.

Kaszuba, A., 2020. Using local autoregressive (LAR) model for forecasting day-ahead electricity prices. Master's thesis. Wrocław University of Science and Technology.

Kath, C., Ziel, F., 2021. Conformal prediction interval estimation and applications to day-ahead and intraday power markets. International Journal of Forecasting 37, 777–799.

Kolassa, S., 2023. Using simulation to determine when forecast accuracy matters. Foresight: The International Journal of Applied Forecasting 68, 20–24.

Kraft, E., Russo, M., Keles, D., Bertsch, V., 2023. Stochastic optimization of trading strategies in sequential electricity markets. European Journal of Operational Research 308, 400–421.

Kremer, M., Kiesel, R., Paraschiv, F., 2021. An econometric model for intraday electricity trading. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 379, 20190624.

Kupiec, P., 1995. Techniques for Verifying the Accuracy of Risk Measurement Models.

Lago, J., Marcjasz, G., De Schutter, B., Weron, R., 2021. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. Applied Energy 293, 116983.

Laur, A., Nieto-Martin, J., Bunn, D.W., Vicente-Pastor, A., 2020. Optimal procurement of flexibility services within electricity distribution networks. European Journal of Operational Research 285, 34–47.

LEAG, 2024. Private communication.

Lee, D., Shin, H., Baldick, R., 2018. Bivariate probabilistic wind power and real-time price forecasting and their applications to wind power bidding strategy development. IEEE Transactions on Power Systems 33, 6087–6097.

Lehna, M., Hoppmann, B., Scholz, C., Heinrich, R., 2022. A Reinforcement Learning approach for the continuous electricity market of Germany: Trading from the perspective of a wind park operator. Energy and AI 8, 100139.

Maciejowska, K., 2020. Assessing the impact of renewable energy sources on the electricity price level and variability – A quantile regression approach. Energy Economics 85, 104532.

Maciejowska, K., 2022. Portfolio management of a small RES utility with a structural vector autoregressive model of electricity markets in Germany. Operations Research and Decisions 32.

Maciejowska, K., Nitka, W., Weron, T., 2019. Day-ahead vs. intraday—Forecasting the price spread to maximize economic benefits. Energies 12, 631.

Maciejowska, K., Uniejewski, B., Serafin, T., 2020. PCA forecast averaging—predicting day-ahead and intraday electricity prices. Energies 13, 3530.

Maciejowska, K., Uniejewski, B., Weron, R., 2023. Forecasting Electricity Prices. Oxford Research Encyclopedia of Economics and Finance .

Madlener, R., Kaufmann, M., 2002. Power exchange spot market trading in Europe: theoretical considerations and empirical evidence. Technical Report Deliverable 5.1b. OSCOGEN.

Manner, H., Alavi Fard, F., Pourkhanali, A., Tafakori, L., 2019. Forecasting the joint distribution of Australian electricity prices using dynamic vine copulae. Energy Economics 78, 143–164.

Marcjasz, G., Narajewski, M., Weron, R., Ziel, F., 2023. Distributional neural networks for electricity price forecasting. Energy Economics , 106843.

Marcjasz, G., Serafin, T., Weron, R., 2018. Selection of Calibration Windows for Day-Ahead Electricity Price Forecasting. Energies 11, 2364.

de Marcos, R.A., Bunn, D.W., Bello, A., Reneses, J., 2020. Short-Term electricity price forecasting with recurrent regimes and structural breaks. Energies 13, 5452.

Mayer, K., Trück, S., 2018. Electricity markets around the world. Journal of Commodity Markets 9, 77–100.

McGrath, A., Jonker, A., 2024. IBM: What is load forecasting?

Morales, J.M., Conejo, A.J., Madsen, H., Pinson, P., Zugno, M., 2014. Integrating Renewables in Electricity Markets: Operational Problems. volume 205 of *International Series in Operations Research & Management Science*. Springer US, Boston, MA.

Mwampashi, M.M., Nikitopoulos, C.S., Konstandatos, O., Rai, A., 2021. Wind generation and the dynamics of electricity prices in Australia. Energy Economics 103, 105547.

Negash, A.I., Westgaard, S., 2018. Evaluating optimal cost-effectiveness of demand response in wholesale markets, in: 2018 IEEE Power & Energy Society General Meeting (PESGM), pp. 1–5.

Neska, E., Kowalska-Pyzalska, A., 2022. Conceptual design of energy market topologies for communities and their practical applications in EU: A comparison of three case studies. Renewable and Sustainable Energy Reviews 169, 112921.

O'Connell, N., Pinson, P., Madsen, H., O'Malley, M., 2014. Benefits and challenges of electrical demand response: A critical review. Renewable and Sustainable Energy Reviews 39, 686–699.

Pape, C., Hagemann, S., Weber, C., 2016. Are fundamentals enough? Explaining price variations in the German day-ahead and intraday power market. Energy Economics 54, 376–387.

Papież, M., Śmiech, S., Frodyma, K., 2018. Determinants of renewable energy development in the EU countries. A 20-year perspective. Renewable and Sustainable Energy Reviews 91, 918–934.

Papież, M., Śmiech, S., Frodyma, K., 2019. Effects of renewable energy sector development on electricity consumption – Growth nexus in the European Union. Renewable and Sustainable Energy Reviews 113, 109276.

Paraschiv, F., 2013. Price Dynamics in Electricity Markets, in: Kovacevic, R.M., Pflug, G.C., Vespucci, M.T. (Eds.), Handbook of Risk Management in Energy Production and Trading. Springer US, Boston, MA, pp. 47–69.

Paraschiv, F., Erni, D., Pietsch, R., 2014. The impact of renewable energies on EEX day-ahead electricity prices. Energy Policy 73, 196–210.

Paraschiv, F., Fleten, S.E., Schürle, M., 2015. A spot-forward model for electricity prices with regime shifts. Energy Economics 47, 142–153.

Parrish, B., Gross, R., Heptonstall, P., 2019. On demand: Can demand response live up to expectations in managing electricity systems? Energy Research and Social Science 51, 107–118.

Pesaran, M.H., Timmermann, A., 2007. Selection of estimation window in the presence of breaks. Journal of Econometrics 137, 134–161.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M.Z., Barrow, D.K., et al., 2022. Forecasting: theory and practice. International Journal of Forecasting 38, 705–871.

Petropoulos, F., Laporte, G., Aktas, E., Alumur, S.A., Archetti, C., et al., 2024. Operational research: methods and applications. Journal of the Operational Research Society 75, 423–617.

Pinson, P., 2013. Wind energy: forecasting challenges for its operational management. Statistical Science 28, 564–585.

Rai, A., Nunn, O., 2020. On the impact of increasing penetration of variable renewables on electricity spot price extremes in Australia. Economic Analysis and Policy 67, 67–86.

Ramos, M.H., Van Andel, S.J., Pappenberger, F., 2013. Do probabilistic forecasts lead to better decisions? Hydrology and Earth System Sciences 17, 2219–2232.

Robette, J., 2023. Does improved forecast accuracy translate to business value? Foresight: The International Journal of Applied Forecasting 68, 12–19.

SAS, 2017. Utility analytics in 2017: Aligning data and analytics with business strategy. Technical Report 108902_G52948.061. SAS.

Savelli, S., Joslyn, S., 2013. The advantages of predictive interval forecasts for non-expert users and the impact of visualizations. Applied Cognitive Psychology 27, 527–541.

Scoblic, J.P., Tetlock, P.E., 2021. A Better Crystal Ball: The Right Way to Think About the Future. Foresight: The International Journal of Applied Forecasting 62, 27–32.

Serafin, T., Marcjasz, G., Weron, R., 2022. Trading on short-term path forecasts of intraday electricity prices. Energy Economics 112, 106125.

Śmiech, S., Papież, M., Rubaszek, M., Snarska, M., 2021. The role of oil price uncertainty shocks on oil-exporting countries. Energy Economics 93, 105028.

TGE, 2024. Rynek Dnia Następnego. https://tge.pl/energia-elektryczna-rdn.

Tian, J., Anderson, H.M., 2014. Forecast combinations under structural break uncertainty. International Journal of Forecasting 30, 161–175.

Timmermann, A., 2006. Chapter 4: Forecast Combinations, in: Elliott, G., Granger, C.W.J., Timmermann, A. (Eds.), Handbook of Economic Forecasting. Elsevier. volume 1, pp. 135–196.

Toubeau, J.F., Nguyen, T.H., Khaloie, H., Wang, Y., Vallée, F., 2022. Forecast-driven stochastic scheduling of a virtual power plant in energy and reserve markets. IEEE Systems Journal 16, 5212–5223.

Truong, C., Oudre, L., Vayatis, N., 2020. Selective review of offline change point detection methods. Signal Processing 167, 107299.

Uniejewski, B., 2024. Electricity price forecasting with Smoothing Quantile Regression Averaging: Quantifying economic benefits of probabilistic forecasts. `arXiv:2302.00411`.

Uniejewski, B., Marcjasz, G., Weron, R., 2019. On the importance of the long-term seasonal component in day-ahead electricity price forecasting: Part II — Probabilistic forecasting. Energy Economics 79, 171–182.

Uniejewski, B., Weron, R., Ziel, F., 2018. Variance Stabilizing Transformations for Electricity Spot Price Forecasting. IEEE Transactions on Power Systems 33, 2219–2229.

URE, 2023. Urząd Regulacji Energetyki: Charakterystyka rynku energii elektrycznej 2022. https://www.ure.gov.pl/pl/energia-elektryczna/charakterystyka-rynku/11089,2022.html.

URE, 2024. Sprawozdanie z działalności Prezesa Urzędu Regulacji Energetyki 2023. Technical Report. Urząd Regulacji Energetyki.

Vukovic, S., 2023. Accuracy and usefulness in applied forecasting. Foresight: The International Journal of Applied Forecasting 68, 45–46.

Waddell, D., Sohal, A.S., 1994. Forecasting: the key to managerial decision making. Management Decision 32, 41–49.

Wallach, O., 2022. The Solar Power Duck Curve Explained. https://elements.visualcapitalist.com/the-solar-power-duck-curve-explained/.

Wang, X., Hyndman, R., Li, F., Kang, Y., 2022. Forecast combinations: An over 50-year review. International Journal of Forecasting 39, 1518–1547.

Westgaard, S., Århus, G.H., Frydenberg, M., 2019. Value-at-risk in the European energy market: a comparison of parametric, historical simulation and quantile regression value-at-risk. Journal of Risk Model Validation .

Yardley, L., Petropoulos, F., 2021. Beyond error measures to the utility and cost of the forecasts. Foresight: the International Journal of Applied Forecasting 63, 36–45.

Yesilbudak, M., Sagiroglu, S., Colak, I., 2017. A novel implementation of kNN classifier based on multi-tupled meteorological input data for wind power prediction. Energy Conversion and Management 135, 434–444.

Zeileis, A., Kleiber, C., Krämer, W., Hornik, K., 2003. Testing and dating of structural changes in practice. Computational Statistics & Data Analysis 44, 109–123.

Ziel, F., Steinert, R., 2016. Electricity price forecasting using sale and purchase curves: The X-Model. Energy Economics 59, 435–454.

Zugno, M., Jónsson, T., Pinson, P., 2013. Trading wind energy on the basis of probabilistic forecasts both of wind generation and of market quantities. Wind Energy 16, 909–926.

Paper 1

Forecasting electricity prices: Autoregressive Hybrid Nearest Neighbors (ARHNN) method

Weronika Nitka, Tomasz Serafin, Dimitrios Sotiros

Forecasting Electricity Prices: Autoregressive Hybrid Nearest Neighbors (ARHNN) Method

Weronika Nitka, Tomasz Serafin, and Dimitrios Sotiros( )

Department of Operations Research and Business Intelligence,
Faculty of Computer Science and Management,
Wrocław University of Science and Technology, 50-370 Wrocław, Poland
dimitrios.sotiros@pwr.edu.pl

Abstract. The ongoing reshape of electricity markets has significantly stimulated electricity trading. Limitations in storing electricity as well as on-the-fly changes in demand and supply dynamics, have led price forecasts to be a fundamental aspect of traders' economic stability and growth. In this perspective, there is a broad literature that focuses on developing methods and techniques to forecast electricity prices. In this paper, we develop a new hybrid method, called ARHNN, for electricity price forecasting (EPF) in day-ahead markets. A well performing autoregressive model, with exogenous variables, is the main forecasting instrument in our method. Contrarily to the traditional statistical approaches, in which the calibration sample consists of the most recent and successive observations, we employ the k-nearest neighbors (k-NN) instance-based learning algorithm and we select the calibration sample based on a similarity (distance) measure over a subset of the autoregressive model's variables. The optimal levels of the k-NN parameter are identified during the validation period in a way that the forecasting error is minimized. We apply our method in the EPEX SPOT market in Germany. The results reveal a significant improvement in accuracy compared to commonly used approaches.

Keywords: Electricity price forecasting · Day-ahead market · ARX · k-nearest neighbors

1 Introduction

Electricity markets have witnessed significant changes over the last decades. Their deregulation, followed by the emergence of electrical power exchanges such as EPEX SPOT, OMIE and Nord Pool in Europe, or PJM in the USA, allowed for competitive electricity trading [23]. Electrical power exchanges usually consist of several markets. The market with the biggest volume of trade is the day-ahead (spot) market, which allows the traders to place bids and offers the day before the

The authors' names are ordered alphabetically.

M. Paszynski et al. (Eds.): ICCS 2021, LNCS 12745, pp. 312–325, 2021.
https://doi.org/10.1007/978-3-030-77970-2_24

physical delivery of electricity. The day-ahead market is usually supplemented by intraday and balancing markets, which allow trading until a few minutes before delivery and target at providing more accurate offers. However, this is often associated with paying significant balancing fees.

Notably, electricity market clearing prices, defined by the supply and the demand curve, are characterized by high volatility. The cost of storing electricity at large scales as well as the transition of power generation from conventional to renewable sources, permeated with uncertainty in the production levels, lead to fluctuations in the supply. On the other hand, demand may vary on an hourly (peak and off-peak hours) and daily (weekends, weekdays and festivities) basis. These factors, along with the requirement of supply and demand to be precisely balanced in the power grid, lead to highly volatile prices in the electricity markets which can undergo extreme changes within a span of a single day.

Traders, ideally, aim to maximize their profit as well as to minimize the financial risk by selecting the most appropriate strategy in an imperfect market, where there is incomplete information. Given the high level of price volatility and the limitations in storing electricity, the selection of a wrong strategy, based on price misinformation, may lead to economic losses or even bankruptcy. On the contrary, utilizing accurate forecasts may increase profits or reduce the risk of economic losses [9,14].

In this line of thought, there is a wide literature which focuses on providing accurate day-ahead electricity price forecasts. Extended literature reviews are provided in [1,21,23]. Two of the prominent broad classes of methods provided in the literature rely either on statistical approaches or on machine learning techniques. Statistical approaches utilize linear regression models or linear autoregressive models based on a set of variables related to observed prices and other exogenous variables (load, wind, solar, temperature) that may affect price levels. Differences in the implementation of the autoregressive models can be also identified in terms of the calibration window length, which can be predefined or estimated via more advanced econometric techniques [4,11,12]. However, in these cases, the calibration sample consists of the most recent and successive observations. Methods that rely on machine learning employ a variety of techniques such as artificial neural networks [24], support vector machines [27], clustering algorithms [22] or a combination of them [17]. It is worthy to mention that a hybrid approach that employs statistical and machine learning techniques has been also proposed in the literature. Specifically, in [18] three clustering algorithms and an autoregressive lag model were employed to predict consumers' energy consumption in a simulation suite. However, this approach was tested on simulated data.

In this paper we build on the bridge between the two aforementioned classes of methods and we propose a new hybrid method, called *autoregressive hybrid nearest neighbors* (ARHNN), for forecasting spot electricity prices. We generate one-day-ahead forecasts using a linear ARX (autoregressive with exogenous variables) model with parameters calibrated on samples selected with the k-nearest neighbors (k-NN) algorithm. ARX models are well-established in electricity price forecasting (EPF), as noted in [10,23]. The k-nearest neighbors algorithm has been found to be successful in the field of electricity market forecasts, mainly in forecasting electricity price and load [2,3,8,13,19,26] and renewable energy sources (RES) generation [25]. The proposed method is applied to the EPEX SPOT market in Germany. The results show a significant improvement in accuracy compared to commonly used benchmark approaches, while low increase in the computational load is ensured.

The rest of this paper unfolds as follows. Section 2 describes the most important features of the data used in this analysis. Section 3 provides an in-depth explanation of the proposed method. Section 4 illustrates the results of the proposed method applied to the EPEX SPOT data and provides comparison with commonly used benchmark models. Finally, conclusions are drawn in Sect. 5.

2 Data

To illustrate our method, we use data describing the day-ahead electricity prices in the EPEX SPOT market in Germany. As described in the Introduction, the day-ahead market is the most important market in terms of traded volume. The dataset, published by the transmission system operator (TSO), comprises four variables: the electricity price in EUR/MWh and the corresponding official TSO forecasts of total electrical load, wind energy generation and photovoltaic energy generation, expressed in GWh.

The dataset spans six full years, from January 2015 until December 2020, with hourly data (see Fig. 1). To evaluate the performance of the proposed algorithm, the data is divided into three periods with lengths of approximately two years each. The first 728-day period is reserved for the initial calibration window. Then, the middle period, of the same length, is utilized for validation and tuning the hyperparameters of the model as described in Subsect. 3.2. Finally, the procedure is tested on the last period with length equal to 736 days.

The time series of the price and the load forecasts, as well as the division into calibration, validation and testing periods, are depicted in Fig. 1. It can be seen that the spot prices are indeed highly volatile, with frequent upward and downward spikes multiple times greater in magnitude than the average price range. However, load is relatively predictable, exhibiting both weekly and yearly seasonality, which needs to be addressed by the predictive model.

3 Methods and Algorithms

As shown by numerous studies in the EPF [7,14], the selection of the calibration sample impacts the overall forecasting accuracy of the autoregressive model. While the majority of authors consider the longest possible portion of data for the model calibration, averaging predictions obtained from calibration samples of different lengths [16] or utilizing more sophisticated statistical methods [15] allows for the significant reduction of forecasting errors. In this paper, we propose a new method for the selection of the calibration sample, based on the k-nearest

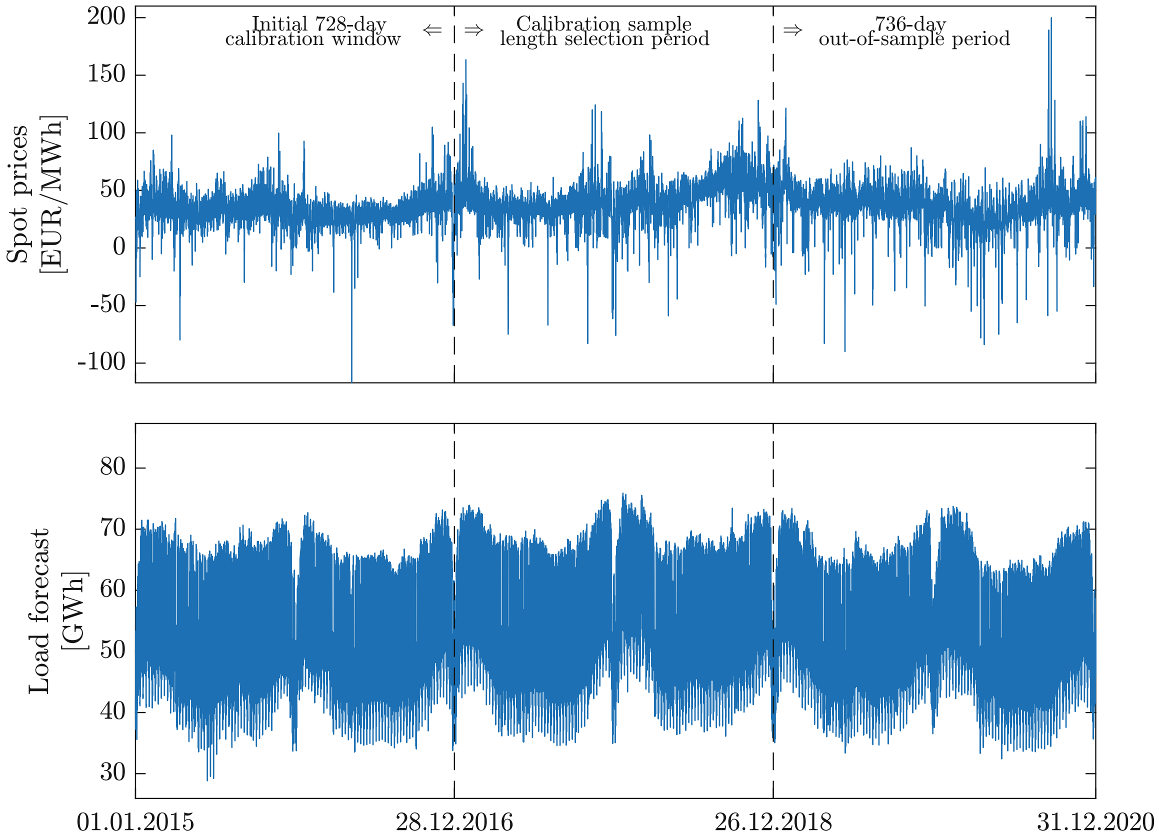


Fig. 1. Time series plot of the electricity spot prices (upper panel) and TSO load forecast (lower panel) from the EPEX SPOT market. Dashed lines indicate the split into calibration, validation and testing periods.

neighbors algorithm. The aforementioned methods rely on the time dimension to select the calibration sample, i.e. the most recent successive observations compose the calibration sample. On the contrary, in our method we define the calibration sample on the basis of a similarity measure over a set of features.

3.1 Predictive Model

To predict the spot prices in hour h of day $d+1$ we use an expert ARX model with a specification well-established in the electricity price forecasting literature [7,20]. Due to the idiosyncratic nature of the electricity market, every hour of the day is treated as a distinct market product and separate forecasts are implemented for each hour, i.e. predicting the prices for the entire day $d+1$ requires estimating 24 independent parameter sets. The models for every hour have an identical specification, incorporating an autoregressive structure with lags corresponding to two preceding days and a week, notated as $P_{d+1-p,h}$ where $p \in \{1, 2, 7\}$. The price dynamics are further captured by including the minimal and the maximal price from the previous day (respectively $P_{d,min}$ and $P_{d,max}$) as well as that day's price in hour 24 ($P_{d,24}$) – the previous day's last known price.

Finally, the model incorporates the publicly available forecasts of three exogenous variables relevant to the price levels: total electrical load ($\hat{L}$), wind energy generation ($\hat{W}$) and photovoltaic energy generation ($\hat{S}$). The complete model takes the form

$$\begin{aligned} P_{d+1,h} = \alpha_h D_{d+1} + \underbrace{\sum_{p\in\{1,2,7\}} \beta_{h,p} P_{d+1-p,h}}_{\text{AR component}} + \underbrace{\theta_{h,1} P_{d,min} + \theta_{h,2} P_{d,max}}_{\text{Daily statistics}} \\ + \underbrace{\theta_{h,3} P_{d,24}}_{\text{Last known price}} + \underbrace{\theta_{h,4} \hat{L}_{d+1,h} + \theta_{h,5} \hat{W}_{d+1,h} + \theta_{h,6} \hat{S}_{d+1,h}}_{\text{Exogenous variables}} + \varepsilon_{d+1,h}, \end{aligned} \tag{1}$$

where D_{d+1} is the 1×7 vector of dummy variables representing days of the week and $\varepsilon_{d+1,h}$ is the Gaussian white noise. Henceforth, by $\hat{P}_{d+1,h}(\tau)$ we denote the prediction obtained from model (1) calibrated on a sample containing the τ most recent observations.

3.2 ARHNN Calibration Sample Selection

The k-Nearest Neighbors is an instance-based learning algorithm that can be used either for classification or regression. In the former case, an observation is assigned to the most common class label shared by its k-nearest neighbors. In the second case, the property value for an observation derives from the average of the k-nearest neighbors' values. In both cases, a neighbor weighting function can be employed [6].

To explain the applicability of the k-NN algorithm in our case study and the differentiation of our method, suppose that at day d we want to forecast the electricity price for the day ahead (day $d+1$). We denote by x_d the vector of the explanatory variables from model (1) for a given day d, after omitting dummy variables, 2-day and 7-day lagged prices and random noise. Within the matrix $X_{d+1} = (x_{d-726}; \ldots; x_{d+1})$, it is evident that the most recent information we possess, x_{d+1}, provides the most accurate outlook at the current market state, i.e. prices from previous days as well as forecasts for day $d+1$. Notably, the proposed statistical methods in the literature, rely on this assumption and they further extend it. Specifically, they assume that the most recent observations will provide the most accurate forecast and thus, they should compose the calibration sample. However, in case structural breaks exist among the last observations, the selected calibration sample will lead to a decreased forecasting accuracy. In addition, this approach relies exclusively on the last observations (in terms of time) and does not exploit information from other past data.

The main idea of our method is to identify past observations that resemble x_{d+1} as closely as possible and use them to estimate the parameters of the forecasting model. To this end, we employ the k-NN algorithm to select a calibration sample for the ARX model (1) consisting of the k-nearest neighbors of the point x_{d+1} (see Fig. 2), in terms of the Euclidean distance. In a sense, we invert the rationale of the k-NN method - instead of classifying the latest observation based on its neighboring points, we assume that the closest neighbors (in terms of the distance, not time) of x_{d+1} belong to the same market "regime".

Analogously to the notation in Sect. 3.1, we denote the price prediction for day $d+1$ and hour h, obtained by calibrating the forecasting model (1) on the sample consisting of k closest observations, by $\hat{P}^{*}_{d+1,h}(k)$. Note that for the clarity of notation, forecasts corresponding to the ARHNN method are marked with an asterisk.

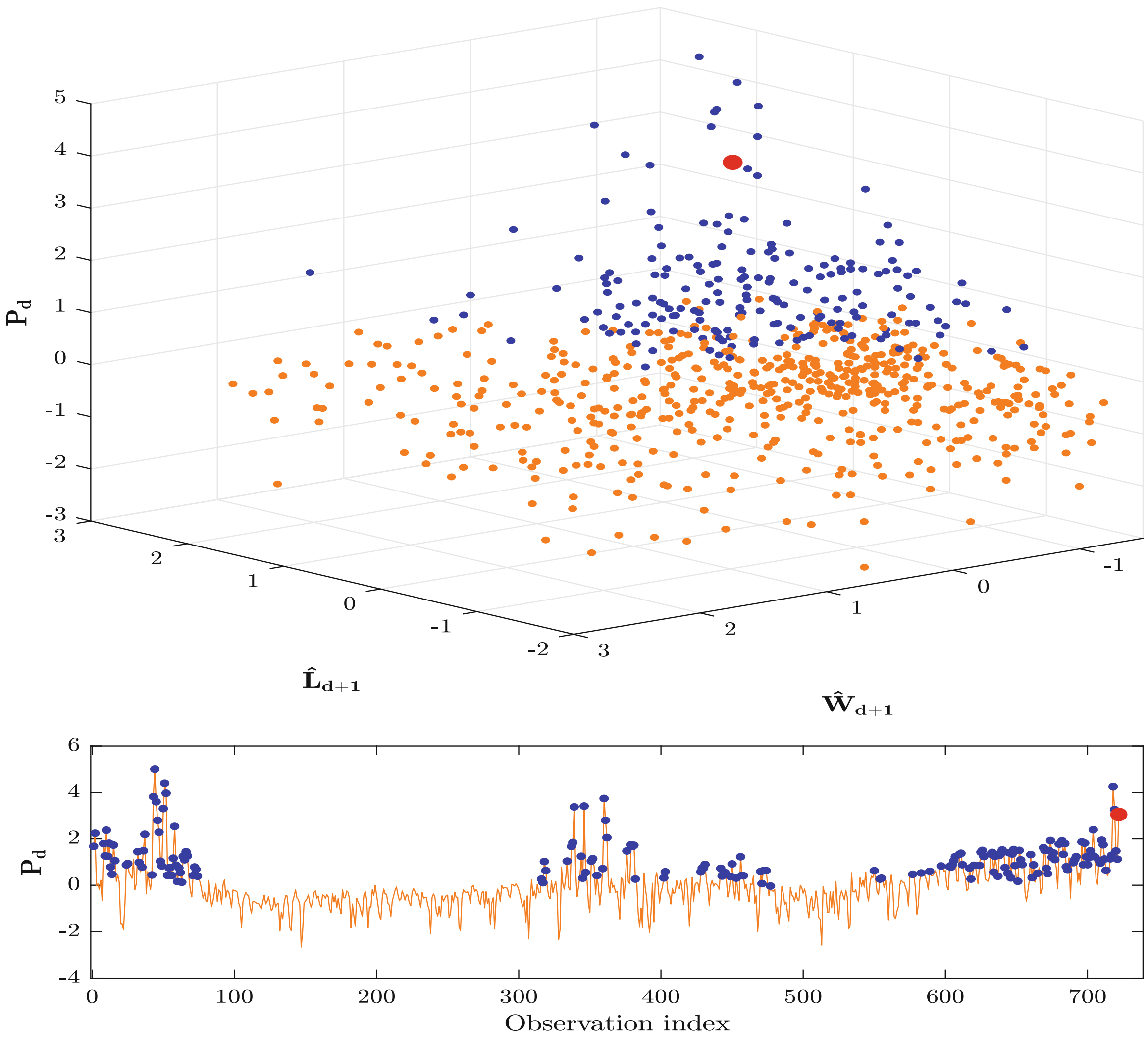


Fig. 2. The optimal (i.e. producing the lowest absolute prediction error) selection of the calibration sample ($\bar{k}_i = 181$) based on the matrix X_{d+1} for a specific day $(d+1)$. The upper panel illustrates the sample selection, presented on three key variables, i.e. preceding day's price as well as forecasts of load and wind generation; while the lower panel depicts the corresponding selection in the time dimension. The most recent observation is marked with a red dot, while the observations selected for the model calibration are depicted with blue points. (Color figure online)

Obviously, the choice of the parameter k has a direct impact on the forecasting accuracy of the model. Disentangling its effects, is one of the main challenges that we address in the paper. As discussed in Sect. 2, in the validation period, we use the 728-day rolling window to identify the optimal values of the k parameter, which is responsible for the number of observations in the calibration sample. For each of the 728 days in the validation period, the procedure identifies (ex-post) the optimal value (i.e. the one that produced the lowest absolute prediction error for a certain day; see Figs. 2, 3) of the parameter, $\bar{k}_i$, $i = 1, \ldots, 728$. Next, in the evaluation (testing) procedure, instead of selecting only one value of k for each day, we consider 728 calibration samples, based on the set of past optimal values $(\bar{k}_1, \ldots, \bar{k}_{728})$. In such way, we obtain 728 price predictions for day $d+1$, i.e. $\left(\hat{P}^*_{d+1,h}(\bar{k}_1), \ldots, \hat{P}^*_{d+1,h}(\bar{k}_{728})\right)$. Eventually, inspired by [16], we obtain the final price prediction for day $d+1$ and hour h from the average of these forecasts:

$$\hat{P}_{d+1,h} = \frac{1}{728} \sum_{i=1}^{728} \hat{P}^*_{d+1,h}(\bar{k}_i). \tag{2}$$

Notably, there may be cases where the values of $\bar{k}_i$, $i = 1, \ldots, 728$ coincide, i.e. $\bar{k}_i = \bar{k}_j$ for $i \neq j$. Therefore, the above expression is translated to the weighted average of forecasts calibrated to different samples, where the weight corresponding to a certain prediction $\hat{P}^*_{d+1,h}(\bar{k}_i)$ depicts the relative frequency of $\bar{k}_i$ in $(\bar{k}_1, \ldots, \bar{k}_{728})$. This can be interpreted as a weighting function which reflects the "relative significance" of the $\bar{k}_i$ values.

3.3 Benchmark Approaches

We evaluate the effectiveness of selecting the calibration period with the proposed ARHNN procedure by comparing it to a number of literature benchmarks. While all of them use Model (1) for computing the forecasts themselves, they differ in the selection of the calibration sample and in the forecasts post-processing. The first group of benchmark approaches provides forecasts obtained by using a single calibration window length throughout the entire test period. The calibration windows include from 56 to 728 days of the most recent data up to the moment of forecasting. The second group utilizes two additional approaches following [7]: the arithmetic mean of all the forecasts within the first group (673 predictions obtained from calibration windows of different lengths), and the average of forecasts from six hand-picked calibration windows: three short ones (56, 84, 112 days) and three long ones (714, 721, 728 days).

We assume the following convention to notate the aforementioned benchmark methods: the single-length windows with length τ are denoted as Win(τ). The forecast averages are named using the MATLAB sequence convention, respectively becoming Av(56:728) and Av(56:28:112, 714:7:728).

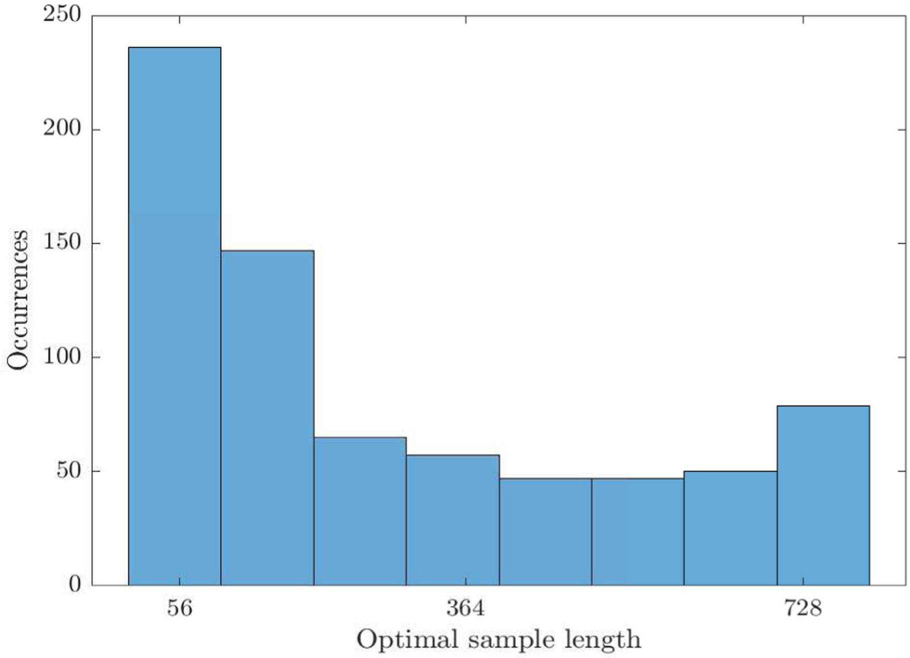


Fig. 3. Histogram of the optimal calibration sample lengths within the validation period (728 calibration sample lengths in total) for hour 18.

4 Results

We evaluate the accuracy of the forecasts obtained from different approaches with the use of the *root mean squared error* (RMSE). The reported error is calculated across all hours and days of the 736-day out-of-sample period. The results are presented in Fig. 4 and Table 1. The performance of single calibration window benchmarks (i.e. models trained on samples comprising a fixed amount of most recent observations) is presented with gray dots. In this approach, although the average error generally diminishes with the increase of the calibration window length τ and the longest window turns out to be the best choice, the decrease is not monotonic as we may expect. As shown by [16], for certain datasets, the error may even increase alongside with the calibration window length.

Table 1. The RMSE values of the selected benchmarks and the ARHNN method.

Method	RMSE
Win(364)	8.4443
Win(728)	8.2860
Av(56:728)	8.0584
Av(56:28:112, 714:7:728)	8.0286
ARHNN	7.8604

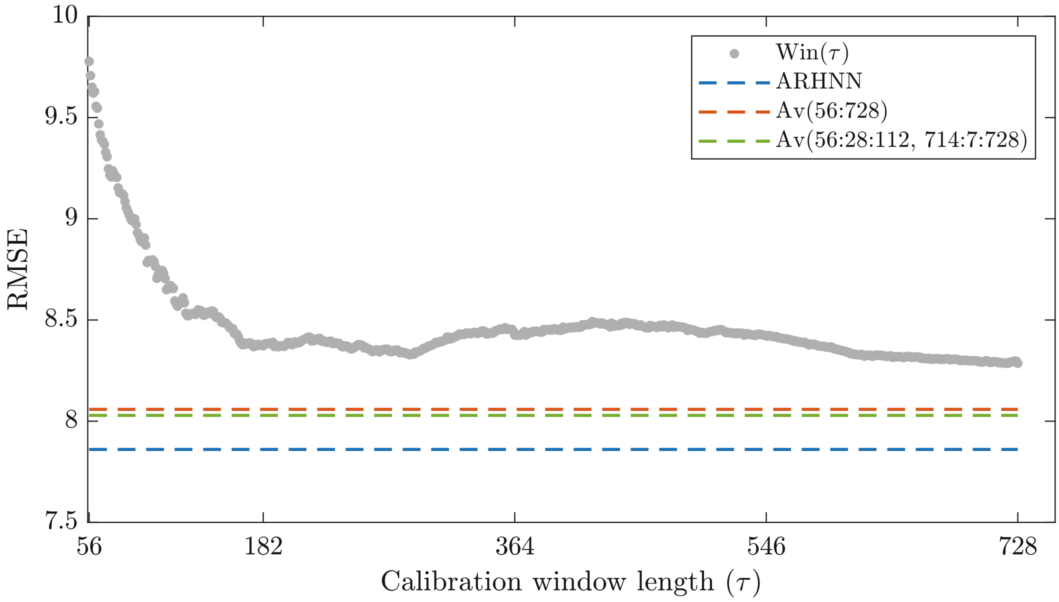


Fig. 4. The RMSE values as a function of calibration window length for the benchmark approaches and the ARHNN method.

As can be seen from Fig. 4, the ARHNN method as well as the averaging schemes outperform every approach based on a single, fixed calibration window length in terms of RMSE. Methods based on forecasts averaging, Av(56:728) and Av(56:28:112, 714:7:728), managed to outperform the predictive accuracy of the longest, 728-day calibration window, approximately by 3%. The forecasts obtained from the introduced ARHNN method exhibit over 5% lower error comparing to the best performing single calibration window length. The method also gains over 2% in terms of the forecasting accuracy compared to the well-performing literature benchmarks Av(56:728) and Av(56:28:112, 714:7:728). Since these results are not sufficient for determining the statistical significance of the difference between forecasts obtained from different approaches, we decided to use the Diebold and Mariano (DM) [5] test. First, for each pair of methods X and Y, we create a vector of errors for each day of the out-of-sample period. Here we consider two different perspectives - univariate and multivariate as classified by [28]. In the first one (multivariate), we consider 24-dimensional error vectors for each day:

$$\Delta_{X,Y,d} = ||\bar{\boldsymbol{\varepsilon}}_{X,d}|| - ||\bar{\boldsymbol{\varepsilon}}_{Y,d}||, \tag{3}$$

where $\bar{\boldsymbol{\varepsilon}}_{X,d} = \sqrt{\frac{1}{24}\sum_{h=1}^{24}\varepsilon_{X,d,h}^2}$ and $\varepsilon_{X,d,h}$ is the error of forecasts obtained with method X for day d and hour h. In the second approach (univariate), instead of considering 24 h jointly, we are looking at each of them separately. More precisely:

$$\Delta_{X,Y,d,h} = |\varepsilon_{X,d,h}| - |\varepsilon_{X,d,h}|. \tag{4}$$

For each pair of approaches, we compute the p-value of the DM test with null hypothesis H_0: $\mathbb{E}(\Delta_{X,Y,d}) \leq 0$ (or H_0: $\mathbb{E}(\Delta_{X,Y,d,h}) \leq 0$ in case of the univariate approach) and additionally perform a complementary test with the reverse null hypothesis, H_0^R: $\mathbb{E}(\Delta_{X,Y,d}) \geq 0$ (or H_0^R: $\mathbb{E}(\Delta_{X,Y,d,h}) \geq 0$).

In Fig. 5 and Fig. 6, we present the p-values of the test. We use a heatmap to indicate the span of p-values. The closer they are to zero (dark green), the more significant is the difference between forecasts obtained with the approach from X-axis (superior) and predictions from the method in the Y-axis (inferior) [7,15,16]. The "chessboard" in Fig. 5 corresponds to the results of the multivariate approach, considering 24-dimensional error vectors (see Eq. 3). It turns out, that forecasts from the ARHNN method were able to significantly outperform predictions from nearly all benchmarks. The well-performing averaging scheme Av(56:28:112, 714:7:728) was neither significantly worse nor better than the proposed approach. Two "chessboards" in Fig. 6, correspond to the results of the univariate DM test for two exemplary hours. The selected Hour 9 and Hour 15 correspond to the worst and the best performance of the ARHNN method across all hours, respectively. For Hour 9, the forecasts based on the ARHNN approach were not able to statistically outperform predictions from any other method. Additionally, they are outperformed by the forecasts based on the Av(56:728) averaging scheme. When it comes to the results for Hour 15, the predictions from the proposed ARHNN method significantly outperform forecasts from all benchmarks, with p-values of the DM test close to zero. In general, the performance of the ARHNN approach across 24 h of the day is shown in Table 2. The columns, corresponding to 24 h are associated with six different performance classes, each of them representing a certain result of the DM test:

- **Class 1** (Hours 2, 6, 13, 14, 15, 16, 17) - forecasts from the ARHNN method significantly outperform predictions from all benchmarks and are not outperformed by any of them,
- **Class 2** (Hours 1, 3, 4, 5) - forecasts from the ARHNN method significantly outperform predictions from three out of four benchmarks and are not outperformed by any of them,
- **Class 3** (Hours 7, 8, 10, 11, 12, 18) - forecasts from the ARHNN method significantly outperform predictions from two out of four benchmarks and are not outperformed by any of them,
- **Class 4** (Hours 22, 23, 24) - forecasts from the ARHNN method significantly outperform predictions from two out of four benchmarks and are outperformed by one of them,
- **Class 5** (Hour 19) - forecasts from the ARHNN method do not significantly outperform predictions from any benchmark and are not outperformed by any of them,
- **Class 6** (Hours 9, 20, 21) - forecasts from the ARHNN method do not significantly outperform predictions from any benchmark and are outperformed by one of them.

Looking at the results of the Diebold-Mariano test it can be observed that forecasts from the ARHNN approach exhibit very satisfactory predictive accuracy compared to forecasts from the selected benchmarks. For eleven hours,

Table 2. Results of the statistical significance test between forecasts from the ARHNN approach and the selected benchmarks for all 24 h. Each class represents a certain result of the DM test.

Class	2	1	2	2	2	1	3	3	6	3	3	3	1	1	1	1	1	3	5	6	6	4	4	4
Hour	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24

ARHNN forecasts were able to significantly outperform predictions from at least three out of four benchmarks and, for twenty hours, at least two out of four. Although the forecasts exhibit the worst performance for hours 9, 19, 20 and 21, they were significantly outperformed by at most one benchmark approach and, in the remaining twenty one hours of the day, by none of them.

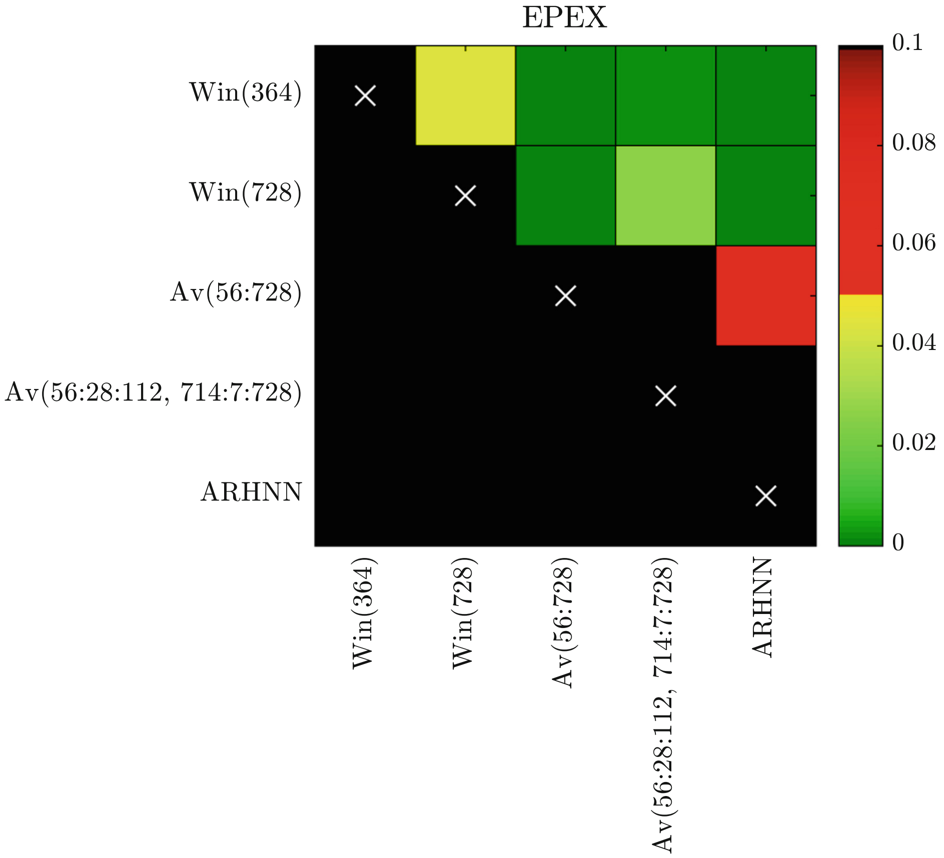


Fig. 5. Results of the multivariate approach to the pairwise Diebold-Mariano test between ARHNN method and the selected benchmarks. We illustrate the range of p-values using a heatmap: green squares indicate a statistically significant superiority of the forecasts from the method on the X-axis over the ones from the method on the Y-axis. (Color figure online)

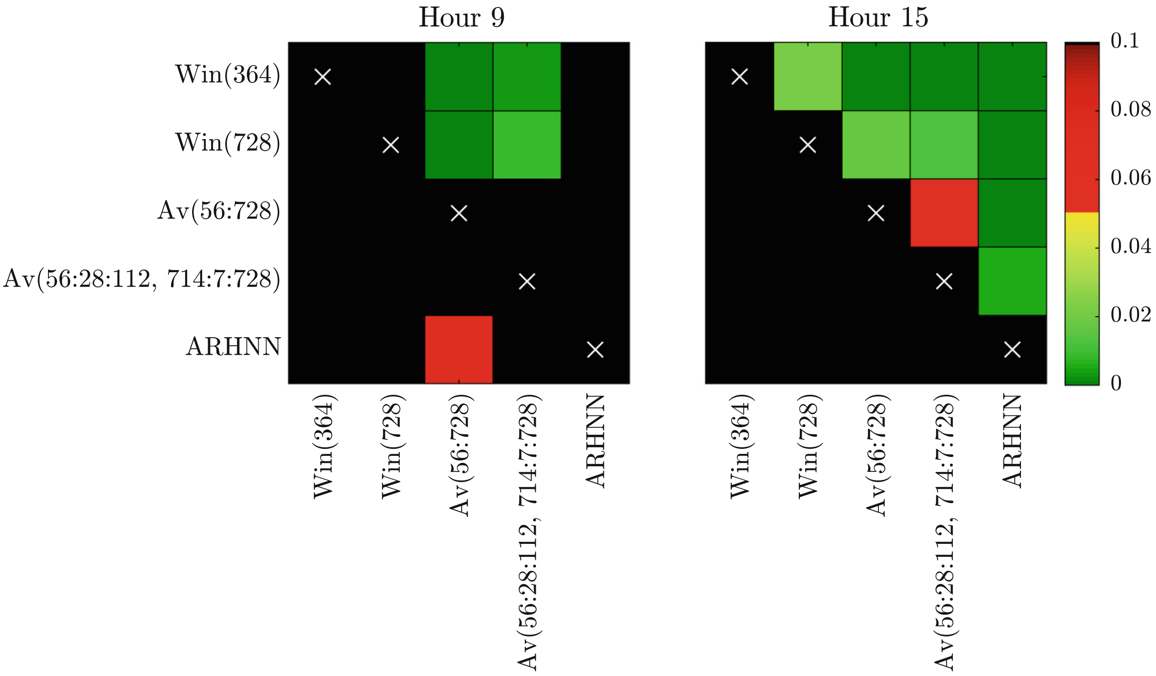


Fig. 6. Sample results of the univariate approach to the pairwise Diebold-Mariano test between ARHNN model and the selected benchmarks. We illustrate the range of p-values using a heatmap: green squares indicate a statistically significant superiority of the forecasts from the method on the X-axis over the ones from the method on the Y-axis. (Color figure online)

5 Conclusions and Discussion

In this paper we introduced a hybrid method for electricity price forecasting in day ahead markets. We employed a linear autoregressive model, with exogenous variables (total electrical load, wind energy generation and photovoltaic energy generation), as the underlying instrument for forecasts. Our novelty lies in the selection of the calibration sample which is achieved via a machine learning algorithm. Specifically, we utilized the k-NN instance-based learning algorithm to select the calibration sample based on a similarity (distance) measure between the most recent information and past observations, over a subset of the autoregressive model's variables. Our aim was to identify past observations that belong to the same "regime" with the latest available information.

The advantage of our method is therefore twofold. The selection of the calibration sample relies on a similarity measure over a set of variables rather than on the time dimension (i.e. to include only the most recent observations). With this type of selection, homogeneity within the calibration sample is secured and structural breaks are avoided. In addition, information from past observations is exploited and consequently, the selected calibration sample is expected to provide more accurate forecasts.

We applied our method on the EPEX SPOT market and we provided comparison with commonly used literature benchmarks. The results show that our proposed method achieves a statistically significant reduction in the forecasting error compared to the rest of the approaches, while remaining highly interpretable and meaningful. The accuracy of the proposed method in other markets, the adoption of other machine learning techniques as well as comparison with other methods relying exclusively on them, are subjects for future research. Nevertheless, our findings signify the importance and benefits of interdisciplinary research in this field.

Acknowledgments. This work is partially supported by the National Science Center (NCN, Poland) through MAESTRO grant no. 2018/30/A/HS4/00444 (D.S.). Also, it is partially supported by the Ministry of Science and Higher Education (MNiSW, Poland) through Diamond Grant no. 0027/DIA/2020/49 (W.N.) and Diamond Grant no. 0009/DIA/2020/49 (T.S.).

References

1. Aggarwal, S.K., Saini, L.M., Kumar, A.: Electricity price forecasting in deregulated markets: a review and evaluation. Int. J. Electr. Power Energy Syst. **31**(1), 13–22 (2009)
2. Ashfaq, T., Javaid, N.: Short-term electricity load and price forecasting using enhanced KNN. In: 2019 International Conference on Frontiers of Information Technology (FIT), pp. 266–2665 (2019)
3. Chaudhury, P., Tyagi, A., Shanmugam, P.K.: Comparison of various machine learning algorithms for predicting energy price in open electricity market. In: 2020 International Conference and Utility Exhibition on Energy, Environment and Climate Change (ICUE), pp. 1–7 (2020)
4. Chow, G.C.: Tests of equality between sets of coefficients in two linear regressions. Econometrica **28**(3), 591–605 (1960)
5. Diebold, F.X., Mariano, R.S.: Comparing predictive accuracy. J. Bus. Econ. Stat. **20**(1), 134–144 (2002)
6. Dudani, S.A.: The distance-weighted k-nearest-neighbor rule. IEEE Trans. Syst. Man Cybern. **SMC–6**(4), 325–327 (1976)
7. Hubicka, K., Marcjasz, G., Weron, R.: A note on averaging day-ahead electricity price forecasts across calibration windows. IEEE Trans. Sustain. Energy **10**(1), 321–323 (2019)
8. Jawad, M., et al.: Machine learning based cost effective electricity load forecasting model using correlated meteorological parameters. IEEE Access **8**, 146847–146864 (2020)
9. Kath, C., Nitka, W., Serafin, T., Weron, T., Zaleski, P., Weron, R.: Balancing generation from renewable energy sources: profitability of an energy trader. Energies **13**(1), 205 (2020)
10. Kiesel, R., Paraschiv, F.: Econometric analysis of 15-minute intraday electricity prices. Energy Econ. **64**, 77–90 (2017)
11. Killick, R., Fearnhead, P., Eckley, I.A.: Optimal detection of changepoints with a linear computational cost. J. Am. Stat. Assoc. **107**(500), 1590–1598 (2012)
12. Lavielle, M.: Using penalized contrasts for the change-point problem. Signal Process. **85**(8), 1501–1510 (2005)
13. Li, W., Kong, D., Wu, J.: A novel hybrid model based on extreme learning machine, k-nearest neighbor regression and wavelet denoising applied to short-term electric load forecasting. Energies **10**(5), 694 (2017)

14. Maciejowska, K., Nitka, W., Weron, T.: Day-ahead vs. intraday-forecasting the price spread to maximize economic benefits. Energies **12**(4), 631 (2019)
15. Maciejowska, K., Uniejewski, B., Serafin, T.: PCA forecast averaging—predicting day-ahead and intraday electricity prices. Energies **13**(14), 3530 (2020)
16. Marcjasz, G., Serafin, T., Weron, R.: Selection of calibration windows for day-ahead electricity price forecasting. Energies **11**(9), 2364 (2018)
17. de Marcos, R.A., Bunn, D.W., Bello, A., Reneses, J.: Short-term electricity price forecasting with recurrent regimes and structural breaks. Energies **13**(20), 5452 (2020)
18. Natividad, F., Folk, R.Y., Yeoh, W., Cao, H.: On the use of off-the-shelf machine learning techniques to predict energy demands of power TAC consumers. In: Ceppi, S., David, E., Hajaj, C., Robu, V., Vetsikas, I.A. (eds.) AMEC/TADA 2015-2016. LNBIP, vol. 271, pp. 112–126. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54229-4_8
19. Nawaz, M., et al.: An approximate forecasting of electricity load and price of a smart home using nearest neighbor. In: Barolli, L., Hussain, F.K., Ikeda, M. (eds.) CISIS 2019. AISC, vol. 993, pp. 521–533. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-22354-0_46
20. Nowotarski, J., Raviv, E., Trück, S., Weron, R.: An empirical comparison of alternative schemes for combining electricity spot price forecasts. Energy Econ. **46**, 395–412 (2014)
21. Nowotarski, J., Weron, R.: Recent advances in electricity price forecasting: a review of probabilistic forecasting. Renew. Sustain. Energy Rev. **81**, 1548–1568 (2018)
22. Rocha, H.R.O., Honorato, I.H., Fiorotti, R., Celeste, W.C., Silvestre, L.J., Silva, J.A.L.: An Artificial Intelligence based scheduling algorithm for demand-side energy management in Smart Homes. Appl. Energy **282** (2021)
23. Weron, R.: Electricity price forecasting: a review of the state-of-the-art with a look into the future. Int. J. Forecast. **30**(4), 1030–1081 (2014)
24. Yamin, H.Y., Shahidehpour, S.M., Li, Z.: Adaptive short-term electricity price forecasting using artificial neural networks in the restructured power markets. Int. J. Electr. Power Energy Syst. **26**(8), 571–581 (2004)
25. Yesilbudak, M., Sagiroglu, S., Colak, I.: A novel implementation of kNN classifier based on multi-tupled meteorological input data for wind power prediction. Energy Convers. Manage. **135**, 434–444 (2017)
26. Zhang, R., Xu, Y., Dong, Z.Y., Kong, W., Wong, K.P.: A composite k-nearest neighbor model for day-ahead load forecasting with limited temperature forecasts. In: 2016 IEEE Power and Energy Society General Meeting (PESGM), pp. 1–5 (2016)
27. Zhao, J.H., Dong, Z.Y., Xu, Z., Wong, K.P.: A statistical approach for interval forecasting of the electricity price. IEEE Trans. Power Syst. **23**(2), 267–276 (2008)
28. Ziel, F., Weron, R.: Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate modeling frameworks. Energy Econ. **70**, 396–420 (2018)

Paper 2

Calibration window selection based on change-point detection for forecasting electricity prices

Julia Nasiadka, Weronika Nitka, Rafał Weron

Calibration Window Selection Based on Change-Point Detection for Forecasting Electricity Prices

Julia Nasiadka, Weronika Nitka(✉), and Rafał Weron

Department of Operations Research and Business Intelligence, Wrocław University of Science and Technology, 50-370 Wrocław, Poland
weronika.nitka@pwr.edu.pl

Abstract. We employ a recently proposed change-point detection algorithm, the Narrowest-Over-Threshold (NOT) method, to select subperiods of past observations that are similar to the currently recorded values. Then, contrarily to the traditional time series approach in which the most recent τ observations are taken as the calibration sample, we estimate autoregressive models only for data in these subperiods. We illustrate our approach using a challenging dataset – day-ahead electricity prices in the German EPEX SPOT market – and observe a significant improvement in forecasting accuracy compared to commonly used approaches, including the Autoregressive Hybrid Nearest Neighbors (ARHNN) method.

Keywords: Change-point detection · Narrowest-Over-Threshold method · Electricity price forecasting · Autoregressive model · Calibration window

1 Introduction

Electricity price forecasting (EPF) is an extremely challenging task. A number of methods have been developed for this purpose, ranging from linear regression to hybrid deep learning architectures utilizing long-short term memory and/or convolutional neural networks. While most studies focus on improving model structures, selecting input features with more predictive power or implementing more efficient algorithms [3,5,6], the issue of the optimal calibration window is generally overlooked [4].

This work is inspired by a recent article [2], which utilized a relatively simple change-point detection method [9] to split the time series into segments with the 'same' price level, and an ICCS 2021 paper [7], which employed the k-nearest neighbors (k-NN) algorithm to select the calibration sample based on

Supported by the Ministry of Education & Science (MEiN, Poland) through Grant No. 0027/DIA/2020/49 (to WN) and the National Science Center (NCN, Poland) through Grant No. 2018/30/A/HS4/00444 (to RW).

D. Groen et al. (Eds.): ICCS 2022, LNCS 13352, pp. 278–284, 2022.
https://doi.org/10.1007/978-3-031-08757-8_24

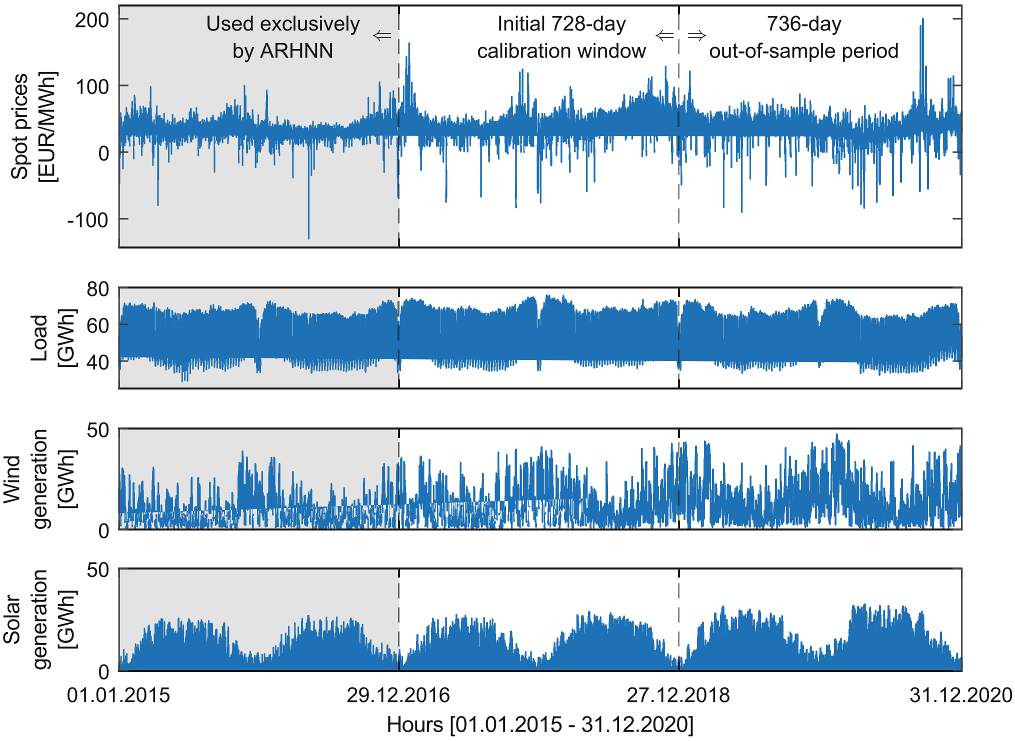


Fig. 1. Electricity spot prices and day-ahead load, wind and solar power generation forecasts in Germany. The last 736 days constitute the test period.

similarity over a subset of explanatory variables. Here, we utilize a recently proposed change-point detection algorithm – the Narrowest-Over-Threshold (NOT) method [1] – to construct an automatic method for detecting subperiods exhibiting different temporal dynamics. Once identified, those not resembling the current behavior are discarded when estimating the predictive model. In what follows, we provide empirical evidence that significant improvement in forecasting accuracy can be achieved compared to commonly used EPF approaches.

The remainder of the paper is structured as follows. In Sect. 2 we present the dataset and the transformation, which is used to standardize the data. In Sect. 3 we briefly describe the NOT method and introduce our approach to selecting subperiods for model calibration. Next, in Sect. 4 we present the forecasting models and in Sect. 5 the empirical results. Finally, in Sect. 6 we conclude and discuss future research directions.

2 The Data

For comparison purposes, we use the same dataset as in [7]. It spans six years (2015–2020) at hourly resolution and includes four series from the German EPEX SPOT market: electricity spot prices $P_{d,h}$ (more precisely: prices set in the day-ahead auction on day $d-1$ for the 24 h of day d) and day-ahead load $\hat{L}_{d,h}$, wind $\hat{W}_{d,h}$ and solar power generation $\hat{S}_{d,h}$ forecasts, see Fig. 1. The first two years are exclusively used for estimating the Autoregressive Hybrid Nearest Neighbors

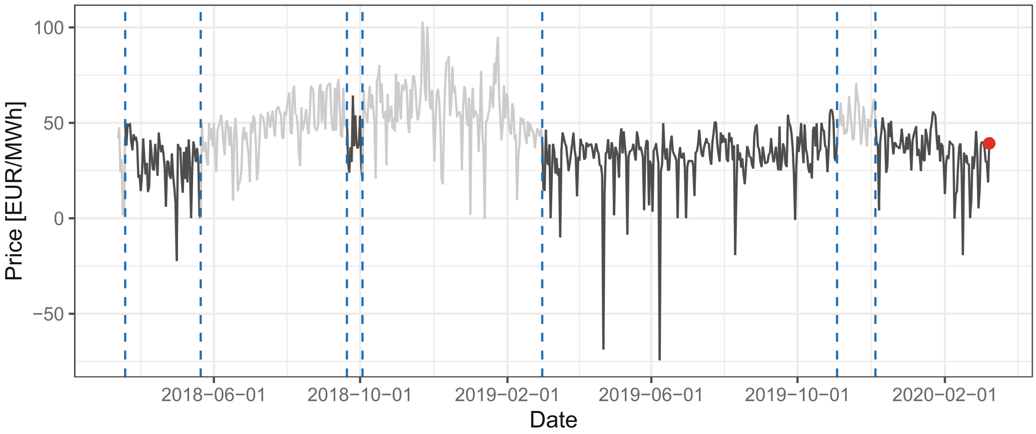


Fig. 2. A sample run of the algorithm introduced in Sect. 3. The red dot is the target day. Vertical dashed lines indicate the located change-points separating periods with different statistical properties. The discarded prices are in gray.

(ARHNN) method [7]; the remaining methods require less data for calibration. The last 736 days constitute the out-of-sample test period.

The most distinct feature of the German power market are frequent spikes and negative prices. Similarly volatile is wind energy generation, while load and solar generation are more predictable. Following [8,10], to cope with this extreme volatility, we transform the electricity prices using the area hyperbolic sine: $Y_{d,h} = \text{asinh}\left(\frac{1}{b}\{P_{d,h} - a\}\right)$ with $\text{asinh}(x) = \log\{x + (x^2+1)^{0.5}\}$, where a is the median of $P_{d,h}$ in the calibration window and b is median absolute deviation. The price forecasts are then obtained by the inverse transformation.

Note that in [7] a different transformation was used. All series, not just $P_{d,h}$, were normalized by subtracting the mean and dividing by the standard deviation in each calibration window. We denote models utilizing *asinh*-transformed data with subscript H; the remaining ones use the standard normalization, as in [7].

3 Calibration Window Selection Using NOT

The Narrowest-Over-Threshold (NOT) method [1] can detect an unknown number of change-points at unknown locations in one-dimensional time series data. The key feature is its focus on the smallest local sections of the data on which the existence of a change-point is suspected. A change-point is said to occur when the behavior of the series changes significantly [2]. See www.changepoint.info for an excellent review site and software repository on this topic. Said differently, change-points split the data into stationary subseries, see Fig. 2. This is what makes them interesting for model calibration and forecasting.

Our algorithm for calibration window selection, i.e., identifying periods with similar time series dynamics to the currently observed, is as follows:

1. Set the maximum number N_c^{max} of change-points to be identified.
2. Use the NOT method to identify $N_c \in [0, N_c^{max}]$ change-points c_i, $i = 1, \ldots, N_c$, in the initial calibration window $\boldsymbol{C}_0$ of length τ. Additionally, denote the first observation in $\boldsymbol{C}_0$ by c_0.

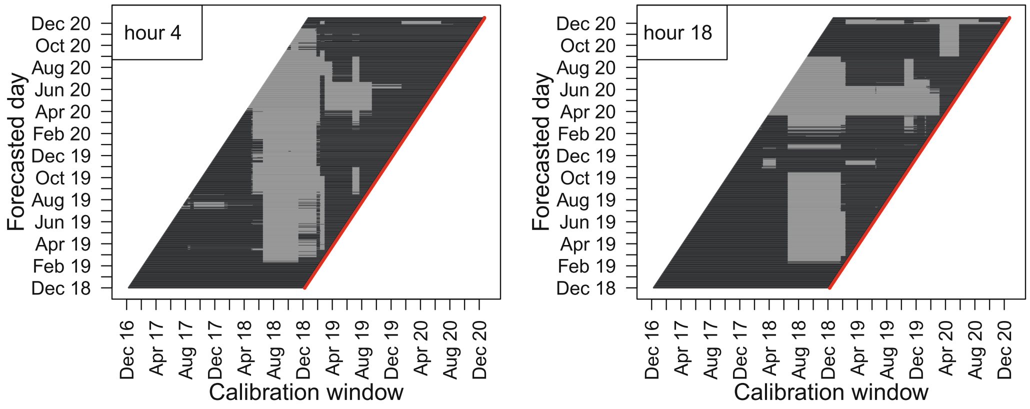


Fig. 3. Overview of NOT-selected (black) and discarded (gray) periods in the two-year calibration window (dates on the x axis) for hours 4 (*left panel*) and 18 (*right panel*). The red line indicates the forecasted day (date on the y axis).

3. If $N_c = 0$ return calibration sample $\boldsymbol{C} = \boldsymbol{C}_0$. Otherwise, compute the empirical quantiles q_{low} and q_{high} of the observations within the period between the most recent change-point found and the last observation in $\boldsymbol{C}_0$.
4. For every interval between two subsequent change-points c_{i-1} and c_i compute the median m_i of its observations, $i = 1, \ldots, N_c - 1$.
5. Set $\boldsymbol{C} = \bigcup_{m_i \in (q_{low}, q_{high})}^{i} [c_{i-1}, c_i]$.

Based on a limited simulation study, we set the maximum number of change-points $N_c^{max} = 12$ and the order of quantiles $(q_{low}, q_{high}) = (q_{0.025}, q_{0.975})$. We also use the least constraining form of NOT, i.e., we assume that the data have piecewise continuous variance and piecewise continuous mean. Any deviations from this are treated as a breach of stationarity.

A sample run of the algorithm is presented in Fig. 2. In the plot, the period closest to the forecasted day (red dot) is characterized by relatively stable, low prices with a small variance. The algorithm discards the light gray subperiods, when either the prices or their variance are significantly higher. In Fig. 3 we illustrate the results for two sample hours and the whole test window. We use a rolling scheme, i.e., once forecasts for the 24 h of the first day in the test sample are computed, the calibration window is moved forward by one day and forecasts for the 2nd day in the test sample are calculated. A clear pattern of vertical gray stripes emerges, meaning that for a range of windows the change-points are consistently detected on the same or neighboring days. Comparing these plots with the price trajectory in Fig. 1, we can observe that much fewer observations are selected by NOT when the prices tend to be more spiky, as can be seen in Spring 2020 (Apr 20 – Jun 20 on the y axis, esp. for hour 18).

4 Forecasting Models

For comparison purposes, the underlying model we use is the same as in [7]. It is an autoregressive structure with exogenous variables dubbed ARX. Since the prices $P_{d,h}$ are set in the day-ahead auction on day $d-1$ independently for the 24h of day d, it is customary in the EPF literature [4,6] to treat every hour as a separate time series. Hence, we consider 24 ARX models of the form:

$$\begin{aligned} P_{d,h} = \underbrace{\boldsymbol{\alpha}_h \boldsymbol{D}_d}_{\text{Dummies}} + \underbrace{\sum_{p\in\{1,2,7\}} \beta_{h,p} P_{d-p,h}}_{\text{AR component}} + \underbrace{\theta_{h,1} P_{d-1,min} + \theta_{h,2} P_{d-1,max}}_{\text{Yesterday's price range}} \\ + \underbrace{\theta_{h,3} P_{d-1,24}}_{\text{Last known price}} + \underbrace{\theta_{h,4} \hat{L}_{d,h} + \theta_{h,5} \hat{W}_{d,h} + \theta_{h,6} \hat{S}_{d,h}}_{\text{Exogenous variables}} + \underbrace{\varepsilon_{d,h}}_{\text{Noise}} . \end{aligned} \qquad (1)$$

The autoregressive (AR) dynamics are captured by the lagged prices from the same hour yesterday, two and seven days ago. Following [10], yesterday's minimum $P_{d-1,min}$, maximum $P_{d-1,max}$ and the last known price $P_{d-1,24}$, as well as day-ahead predictions of the three exogenous variables are included. Finally, a 1×7 vector of dummy variables $\boldsymbol{D}_d$ is used to represent the weekly seasonality and the uncertainty is represented by white noise.

Overall, we compare seven types of approaches that all use ARX as the underlying model. The first three are the same as in [7]: (i) **Win(τ)** – the ARX model estimated using a window of τ days, with $\tau \in [56, 57, ..., 728]$, (ii) **Av(Win)** – the arithmetic average of six forecasts of the ARX model for three short ($\tau = 56, 84, 112$) and three long windows ($\tau = 714, 721, 728$), and (iii) the **ARHNN** model. The next four include: (iv) **Win$_\mathbf{H}$(τ)** – the same as Win(τ) but calibrated to *asinh*-transformed prices, (v) **NOT$_\mathbf{H}$(728)** – the ARX model calibrated to *asinh*-transformed prices in NOT-selected subperiods from the 728-day window, (vi) **Av(Win$_\mathbf{H}$)** – the same as Av(Win) but calibrated to *asinh*-transformed prices, and (vii) **Av(NOT$_\mathbf{H}$)** – the same as Av(Win$_\mathrm{H}$) but with the forecasts for the three long windows ($\tau = 714, 721, 728$) replaced by NOT$_\mathrm{H}$(728). The rationale behind the latter averaging scheme is that NOT$_\mathrm{H}$(τ) performs best for long calibration windows and offers little or even no gain for $\tau < 1$ year.

5 Results

We evaluate the forecasting performance of the seven approaches presented in Sect. 4 in terms of the *root mean squared error* (RMSE; results for the mean absolute error are similar and available from the authors upon request). The RMSE values reported in Fig. 4 are aggregated (averaged) across all hours in the 736-day test sample, see Fig. 1. Additionally, to test the significance of differences in forecasting accuracy, for each pair of models we employ the multivariate variant of the Diebold-Mariano (DM) test, as proposed in [10].

Several conclusions can be drawn. Firstly, changing the preprocessing method from normalization [7] to *asinh* transformation [8] generally reduces the RMSE. Even the worst performing out of the latter approaches, Win$_\mathrm{H}$(728), improves on ARHNN, the most accurate method in [7]. Secondly, NOT-selection yields further improvement, although not statistically significant if considered on its own. Compare NOT$_\mathrm{H}$(728) with Win$_\mathrm{H}$(728) and Av(NOT$_\mathrm{H}$) with Av(Win$_\mathrm{H}$).

Method	RMSE
Win(728)	8.2860
Av(Win)	8.0286
ARHNN	7.8605
$\text{Win}_\text{H}(728)$	7.7286
$\text{NOT}_\text{H}(728)$	7.5994
$\text{Av}(\text{Win}_\text{H})$	7.0968
$\text{Av}(\text{NOT}_\text{H})$	7.0831

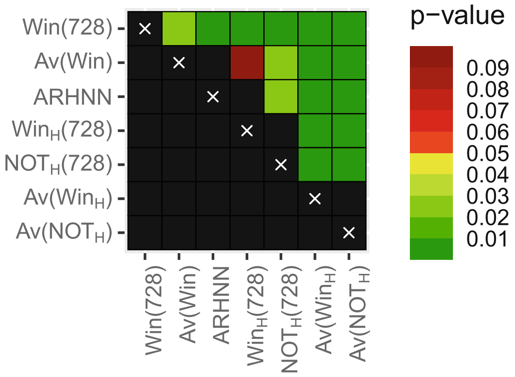


Fig. 4. RMSE errors in the out-of-sample test period (*left panel*). A heatmap of the p-values for the multivariate Diebold-Mariano test [10] for each pair of methods (*right panel*). The smaller the p-values, the more significant is the difference between the forecasts of a model on the x-axis (better) and the forecasts of a model on the y-axis (worse). Black color indicates p-values in excess of 0.1.

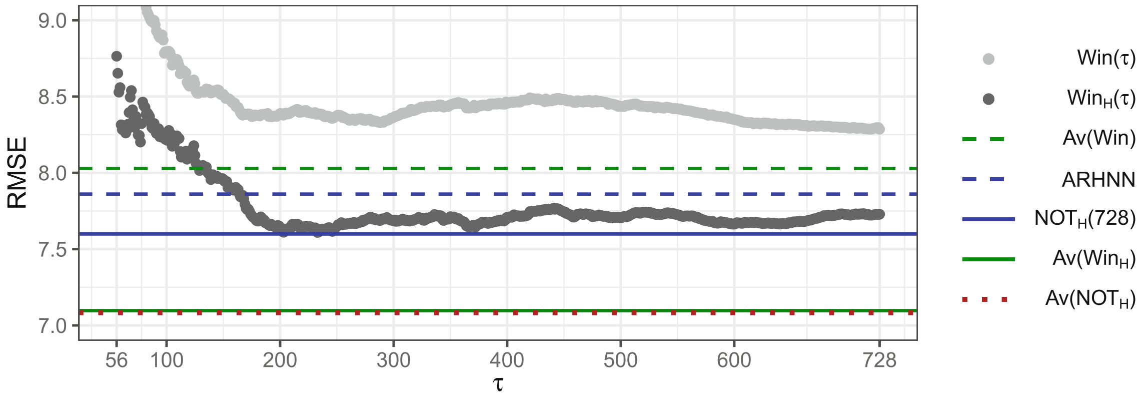


Fig. 5. RMSE values for all considered models; τ is the calibration window length.

The RMSE values for all considered approaches are presented graphically in Fig. 5. It clearly shows the significant improvement from using the *asinh* – compare between $\text{Win}(\tau)$ with $\text{Win}_\text{H}(\tau)$ for all τ's. While $\text{Win}_\text{H}(728)$ is not the best performing of all $\text{Win}_\text{H}(\tau)$ models, the differences in performance are relatively minor for $\tau \geq 200$ days. Even the best ex-post known model, $\text{Win}_\text{H}(233)$, is slightly worse than $\text{NOT}_\text{H}(728)$. The averaged forecasts $\text{Av}(\text{Win}_\text{H})$ and $\text{Av}(\text{NOT}_\text{H})$ are further able to improve on the accuracy, although the differences between them are not significant.

6 Conclusions and Discussion

In this paper we propose a novel method for selecting calibration subperiods based on Narrowest-Over-Threshold (NOT) change-point detection [1]. Contrarily to the traditional time series approach in which the most recent observations are taken as the calibration sample, we propose to estimate the predictive models only using data in the selected subperiods. We evaluate our approach using

German electricity market data and seven variants of autoregressive models tailored for electricity price forecasting (EPF). We provide empirical evidence that significant improvement in forecasting accuracy can be achieved compared to commonly used EPF approaches, including the recently proposed ARHNN [7]. In addition to calibration sample selection, our results also emphasize the importance of using transformations like the *asinh*, in line with [8,10].

The roughly sixfold increase in computational time of the NOT-based methods – 4.97s for 24 h forecasts using $\text{NOT}_\text{H}(728)$ vs. 0.83s using $\text{Win}_\text{H}(728)$, running R ver. 3.6.3 on an i7-9750H processor – can be seen as a drawback, especially compared to less complex ways of improving forecast accuracy, like calibration window averaging [4]. However, the automation of the forecasting process may make the trade-off worthwhile. If this is the case for more complex models than the autoregressive ones considered here or the shallow neural network in [2], e.g., LASSO-estimated AR (LEAR) and deep neural networks [6], is left for future work.

References

1. Baranowski, R., Chen, Y., Fryzlewicz, P.: Narrowest-over-threshold detection of multiple change-points and change-point-like features. J. R. Stat. Soc. **81**(3), 649–672 (2019)
2. De Marcos, R., Bunn, D., Bello, A., Reneses, J.: Short-term electricity price forecasting with recurrent regimes and structural breaks. Energies **13**(20), 5452 (2020)
3. Heijden, T., Lago, J., Palensky, P., Abraham, E.: Electricity price forecasting in European day ahead markets: a greedy consideration of market integration. IEEE Access **9**, 119954–119966 (2021)
4. Hubicka, K., Marcjasz, G., Weron, R.: A note on averaging day-ahead electricity price forecasts across calibration windows. IEEE Trans. Sustain. Energy **10**(1), 321–323 (2019)
5. Jahangir, H., Tayarani, H., Baghali, S., et al.: A novel electricity price forecasting approach based on dimension reduction strategy and rough artificial neural networks. IEEE Trans. Ind. Inform. **16**(4), 2369–2381 (2020)
6. Lago, J., Marcjasz, G., Schutter, B.D., Weron, R.: Forecasting day-ahead electricity prices: a review of state-of-the-art algorithms, best practices and an open-access benchmark. Appl. Energy **293**, 116983 (2021)
7. Nitka, W., Serafin, T., Sotiros, D.: Forecasting electricity prices: autoregressive hybrid nearest neighbors (ARHNN) method. In: Paszynski, M., Kranzlmüller, D., Krzhizhanovskaya, V.V., Dongarra, J.J., Sloot, P.M.A. (eds.) ICCS 2021. LNCS, vol. 12745, pp. 312–325. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-77970-2_24
8. Uniejewski, B., Weron, R., Ziel, F.: Variance stabilizing transformations for electricity spot price forecasting. IEEE Trans. Power Syst. **33**(2), 2219–2229 (2018)
9. Zeileis, A., Kleiber, C., Walter, K., Hornik, K.: Testing and dating of structural changes in practice. Comput. Stat. Data Anal. **44**, 109–123 (2003)
10. Ziel, F., Weron, R.: Day-ahead electricity price forecasting with high-dimensional structures: univariate vs. multivariate modeling frameworks. Energy Econ. **70**, 396–420 (2018)

Paper 3

Multiple split approach – multidimensional probabilistic forecasting of electricity markets

Katarzyna Maciejowska, Weronika Nitka

Multiple split approach – multidimensional probabilistic forecasting of electricity markets

Katarzyna Maciejowska[a], Weronika Nitka[a,*]

[a] *Department of Operations Research and Business Intelligence, Wrocław University of Science and Technology, Wybrzeże Wyspiańskiego 27, 50-370, Wrocław, Poland*

Keywords: probabilistic forecasting, multivariate forecasting, resampling, trading strategy, renewable energy

Abstract

In this article, a *multiple split* method is proposed that enables construction of multidimensional probabilistic forecasts of a selected set of variables. The method uses repeated resampling to estimate uncertainty of simultaneous multivariate predictions. This nonparametric approach links the gap between point and probabilistic predictions and can be combined with different point forecasting methods. The performance of the method is evaluated with data describing the German short-term electricity market. The results show that the proposed approach provides highly accurate predictions. The gains from multidimensional forecasting are the largest when functions of variables, such as price spread or residual load, are considered.

Finally, the method is used to support a decision process of a moderate generation utility that produces electricity from wind energy and sells it on either a day-ahead or an intraday market. The company makes decisions under high uncertainty because it knows neither the future production level nor the prices. We show that joint forecasting of both market prices and fundamentals can be used to predict the distribution of a profit, and hence helps to design a strategy that balances a level of income and a trading risk.

1. Introduction

Electrical energy markets play a crucial role in modern economies. Reliable and cheap energy supply is believed to be essential for the daily lives of citizens and business operations. With the evolution and changing needs of economies, a large number of developed countries abandoned monopolistic and government-controlled power systems in favor of decentralized market structures. Today, trade is conducted in multiple ways: through bilateral contracts, short-term spot markets organized in the form of energy exchanges, and futures markets. Spot markets typically take the form of day-ahead (DA) markets in which offers are placed for 24 hours of the next day. In many countries, they are complemented by an intraday market that allows trade up to a few minutes prior to delivery and helps to dynamically

*Corresponding author

Email addresses: `katarzyna.maciejowska@pwr.edu.pl` (Katarzyna Maciejowska), `weronika.nitka@pwr.edu.pl` (Weronika Nitka)

balance supply and demand. DA prices serve as a reference price for other types of contracts and therefore have a huge impact on the whole market.

Together with changes of the market structure, one could also observe dynamic development of new electricity generation technologies, in particular those utilizing renewable energy sources (RES). As RES generation is intermittent and depends on changing weather conditions, it introduces a lot of uncertainty into the trade. Varying generation combined with the constant need to balance demand and supply, and limited storage opportunities, leads to high volatility of electricity prices. Their average level changes according to the weather and demand. In unfavorable circumstances, they also exhibit spiky behavior, with both positive and negative jumps.

The growing complexity of electricity markets and the increasing exposure of market participants to various trading risks raise the need for reliable forecasts of both electricity prices and market fundamentals, such as the demand level or RES generation. When reviewing the literature on electricity price forecasting (EPF), it is evident that the primary focus is often on point forecasting, i.e. forecasting the expected value of prices [1]. Various methods have been proposed and examined in that context, starting with simple linear regression and autoregression types of models [2], through nonlinear models [see e.g. 3; 4], different estimation methods [5] up to artificial intelligence (AI) approaches [6].

In contrast to point predictions, probabilistic forecasts assess the entire distribution, thus allowing for assessing both the variable's level and the prediction uncertainty. Although more difficult to calculate, probabilistic predictions have gained popularity in the EPF literature [7; 8]. The available methods can be classified into three main categories based on the way they express the uncertainty of the forecast: quantile prediction, density estimations, and ensemble predictions. As stated in [9], the distinction between different types of forecast is to some extent artificial. For instance, quantiles can be derived from an ensemble, and an ensemble can be generated using a fine approximation of the distribution. The literature discusses various methods for constructing probabilistic forecasts. Popular approaches are based on the analysis of forecast errors that come from a model used for point forecasting [10; 11] or direct modeling of a density of data with quantile regression [QR, 12]. It should be mentioned that to obtain a fine approximation of a continuous distribution, a substantial number of quantiles needs to be modeled, which makes an application of QR computationally burdensome. Moreover, the concept of quantiles describes primarily a univariate distribution and does not have a direct and intuitive extension to a multiple output case [13]. Despite these obstacles, a very good QR prediction performance has been verified by a number of experiments, not only in the EPF area [10; 14; 15; 16].

Whether in the context of point or probabilistic forecasting, typical techniques are designed to predict a single random variable. However, in many applications, the dependencies between events are important and should be carefully modeled [9; 17]. For example, a wind farm manager can explore a correlation between wind generation and electricity price forecast errors to increase income [18]. An unexpected rise of RES generation is known to lead to a fall of DA and ID prices, which should be reflected in the offer curve. Toubeau et al. [19] analyze a performance of a *virtual power plant* (VPP) and shows how a multidimensional forecast of non-shiftable load, renewable generation, and electricity prices can be used in the management of the VPP. Finally, joint forecasting of electricity prices can be employed for designing a trading strategy that allows one to place offers on multiple electricity markets, similar to [20], [11], and [21].

Therefore, in this research paper, instead of forecasting electricity prices and fundamental variables separately, we consider multidimensional probabilistic forecasting. There are two popular approaches that enable joint predicting of a set of time series. The first aggregates all endogenous variables into a vector and uses a single model to describe its behavior. A well-known example of such multidimensional models is *vector autoregression* [VAR, 22]. The approach is used in [11], where behavior of electricity prices, total load, and RES generation is modeled with the structural VAR method. An alternative approach combines distributions coming from univariate models via, for example, copulas [see 23; 24]. In such a case, the modeling procedure is divided into two steps. Firstly, time-series or AI models are applied to the individual variables. In the second step, the dependence between fitted errors is described by the copula. In the EPF literature, this approach has been successfully applied by [25; 19; 26; 27].

In this study, we propose an alternative *multiple split* procedure that is based on resampling methods. Similarly to [25], we use univariate models to predict the expected values of individual variables and a multidimensional distribution of forecast errors to calculate probabilistic forecasts. The method can be particularly useful for practitioners, as it may be combined with any unidimensional point-forecasting scheme. In this article, we explore time series models. However, it could also be merged with other structural or AI based techniques.

The proposed method combines previous work of Lei et al. [28] and Barber et al. [29] and integrates *multiple split conformal* predictions with a *jackknife+* approach. Several names have been assigned to approaches that are based on the division of the sample into disjoint subsets: split method, leave-k-out (LKO), d delete jackknife, or cross-validation (CV) [30; 31; 29]. In the split approach [28; 32], the sample is randomly divided into disjoint sets. The subsets resulting from the split are next used to estimate the parameters and calculate the forecast errors. In this article, following [29], the outcome of each split is the probabilistic forecast of the variable of interest not just the distribution of errors. Hence, the method captures both the uncertainty that arises from estimation of parameters and the stochastic nature of the data. The *multiple split* described in this work enhances the existing literature in various directions:

(i) It extends the analysis from a univariate to a multivariate framework. The multidimensional property is particularly important when complex decision problems are considered that require approximation of a distribution of a function of random variables, e.g. their linear combination. In this case, the approach allows researchers to account for various sources of uncertainty that may cross-depend on each other.

(ii) Unlike in previous articles such as [32], random splitting is repeated multiple times to decrease the variability of the outcomes and reduce the dependence of the results on a particular division of the data. However, contrary to jackknife+ approach of Barber et al. [29], we do not consider all possible divisions of the original sample and hence reduce the computation time.

(iii) The results obtained through the splits are stored as an ensemble rather then the distributions (parametric or a set of quantiles) as in [28] or [32]. Therefore, the approach does not require aggregation of outcomes obtained from individual splits with probabilistic forecast combination methods such as Bonferroni averaging [28]. Moreover, the ensemble forecasts can be easily used to generate a prediction of any linear or non-linear function of the original data. In such a case, a new ensemble is constructed by applying the function to the set of multidimensional predictions.

The performance of the proposed method is evaluated with a dataset describing a German electricity market. First, it is applied for forecasting individual time series: day-ahead (DA) and intraday (ID) prices, total load, and RES generations. Next, it is used to predict the distribution of their function: the price spread and residual load (computed as the difference between load and RES). The accuracy of the forecast is assessed using various measures: the coverage of the prediction intervals, the *continuous ranked probability score* (CRPS), and the reliability index. They show how well the probabilistic predictions resemble the out-of-sample distribution. Furthermore, given a correct calibration, CRPS assesses the sharpness of the distribution obtained. The results are compared to the well-established benchmarks: quantile regression and historical simulations. The second benchmark is extended to multidimensional forecasting to evaluate the potential gains from multiple splitting of the data. Finally, the method is used to support the decision process of moderate wind farms. The utility is assumed to decide about the quantity offered in the day-ahead market on the day preceding the delivery. The choice is made under uncertainty because at the time it is taken, the generator does not know neither its production level nor electricity prices. We show that joint forecasting of both market prices and fundamentals can be used to predict the distribution of future profit, and hence helps to design a trading strategy that balances the level of income and the risk.

This article is structured as follows. Section 2 presents the main characteristics of the dataset explored in the study. Next, the forecasting methods are presented in Section 3. Section 4 describes the trading strategy of a small wind farm. Finally, the results of the empirical study are presented in Section 5. Section 6 provides conclusions.

2. Data

In this article, we use a dataset describing the EPEX SPOT market in Germany. The data span four calendar years, between October 1, 2015 and September 30, 2019, with an hourly resolution. They are roughly equally divided into training and test periods, with the first 728 days reserved for the initial calibration window and the remaining observations used for evaluation of the forecasts. The variables include day-ahead market prices (DA) and ID-3 intraday prices (ID), chosen as representative indicator of the average level of energy prices during continuous intraday trading; as well as actual and forecasted values of fundamental variables and, finally, closing prices of fuel futures contracts. The detailed description of all data used with units, sources, and notation is presented in Table 1.

The average daily values of the variables most relevant for the analysis, which are the prices in both the short-term markets and the generation structure, are plotted in Fig. 2 and Fig. 2, respectively. The price time series exhibits multiple features that differentiate the electrical energy market from most other commodity markets. They follow cyclical fluctuations that result from seasonal patterns of demand and yearly changes in weather. When prices on short-term markets are considered, it can be observed that they are typically highly volatile and strongly correlated. The average values across markets are similar. For the period analyzed, they are equal to 36.11 EUR/MWh for the day-ahead market and 36.24 EUR/MWh for the intraday market. However, the ID market is characterized by a higher variance, with a standard deviation of 18.31 EUR/MWh compared to 16.93 EUR/MWh on the DA market. Finally, unlike in the case of other commodities, electricity prices are subjected to spikes. Although in most days the average price stays between 10-30 EUR/MWh, sudden jumps occur when the price increases over 50 EUR/MWh or falls below zero.

Table 1: Data sources and units.

Data	Notation	Units	Source
Day-ahead prices	DA	EUR/MWh	`http://www.epexspot.com`
Intraday prices	ID	EUR/MWh	`http://www.epexspot.com`
Load	L	GWh	`https://transparency.entsoe.eu`
Wind generation	W	GWh	`https://transparency.entsoe.eu`
PV generation	S	GWh	`https://transparency.entsoe.eu`
Forecasted load	FL	GWh	`https://transparency.entsoe.eu`
Forecasted wind generation	FW	GWh	`https://transparency.entsoe.eu`
Forecasted PV generation	FS	GWh	`https://transparency.entsoe.eu`
API2 Coal futures price	C	EUR	`finance.yahoo.com`
TTF Gas futures price	G	EUR	`www.eex.com`

Figure 2 complements the information on the short-term electricity market and presents the load level (*top panel*, Fig. 2), the RES generation (*middle panel*, Fig. 2) and the wind generation (*bottom panel*, Fig. 2). It can be observed that the load behavior is regular and exhibits strong weekly and yearly seasonality. In the case of RES generation, the level of production varies mainly due to short-term intermittency and changing weather conditions. These fluctuations have a significant effect on the wholesale market and lead to an increase of the price variability. The impact is the stronger, the larger becomes the share of RES in the generation mix. Within the considered time period, the installed wind energy capacity in Germany increased by approximately 36%, from 44.58 GW in 2015 to 60.75 GW in 2019[1]. The actual average monthly generation varied between approximately 10% during summers (June to August) and 40% during winters (December to February). The share of RES in the generation mix increased even more over the period due to investment in solar generation. In 2019, RES has been estimated to account for 44.8% of the total energy generation in Germany.

All time series were pre-processed to account for time zone changes, as in [1]. The missing values (corresponding to the transition from winter to summer) were replaced by the arithmetic averages of the two nearest values. The doubled values (corresponding to the change from summer to winter) are replaced by their arithmetic mean.

3. Forecasting methods

In this paper, electricity prices and variables that describe the generation structure in consecutive hours are interpreted as separate time series (products). Their point forecasts are calculated with univariate models. Although the structure of the models remains unchanged throughout the day, the values of the parameters are estimated independently for each hour. Finally, in order to resemble the true trading problem, it is assumed that all the computations are performed in the morning at 11:00, hence only the information available at this time is used.

[1] Source: energy-charts.info

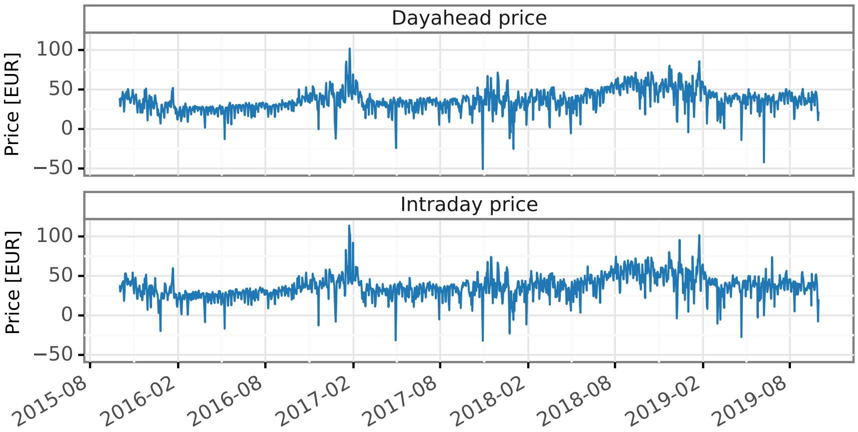


Figure 1: Germany, plots of market prices.

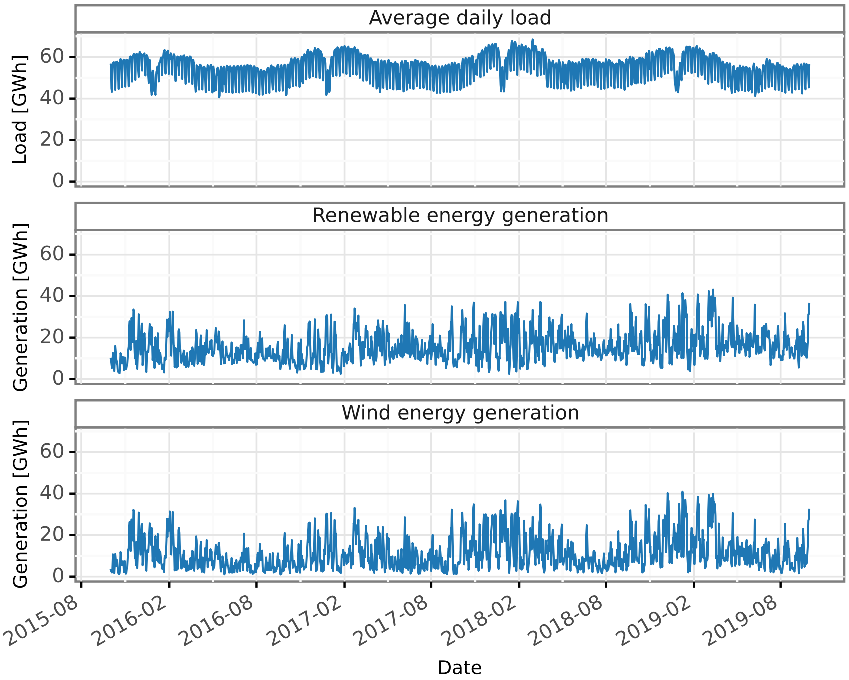


Figure 2: Germany, plots of total load and renewable energy generation.

3.1. ARX models

Autoregressive models with exogenous variables (ARX) are commonly used in the EPF literature [7]. In this type of model, an expected value of an endogenous variable is described as a linear function of its past realizations (an autoregressive component) and a set of explanatory variables. When the structure of the model is predetermined and is not subjected to statistical selection and verification, it can be viewed as an expert model. Here, we adopt specifications used for forecasting of electricity prices, RES generation, and the total load described by Maciejowska et al. [33]. Additionally, two models are proposed that describe a residual load, defined as the difference between total load and the renewable energy generation ($RL_{t,h} = L_{t,h} - RES_{t,h}$), and a price spread ($SP_{t,h} = DA_{t,h} - ID_{t,h}$).

Let us first consider models that explain the generation structure: total load, RES and wind generation, and residual load. Following [33], the load level is described by an equation below:

$$\begin{aligned} L_{t,h} =& \alpha_h + \underbrace{\theta_{h,1} L^*_{t-1,h} + \sum_{p\in\{2,7\}} \theta_{h,p} L_{t-p,h}}_{\text{AR component}} + \underbrace{\beta_{h,1} FL_{t,h} + \beta_{h,2} FRES_{t,h}}_{\text{Forecasts of fundamentals}} \\ &+ \underbrace{\beta_{h,3} FL_{t,ave} + \beta_{h,4} FL_{t,max} + \beta_{h,5} FL_{t,min}}_{\text{Daily statistics}} + \varepsilon_{t,h}, \end{aligned} \tag{1}$$

The model consists of three main components: an autoregressive component with lags $p \in \{1, 2, 7\}$ that captures short-term dependencies and weekly seasonality, TSO forecasts of fundamental variables ($FL_{t,h}$ and $FRES_{t,h} = FW_{t,h} + FS_{t,h}$), as well as the average, minimum, and maximum forecasted load within day t. Because forecasts are performed in the morning at 11:00, information on the load level for hours later than 10:00 is not available. In this case, $L_{t-1,h}$ is replaced by its TSO forecast. Therefore, in the regression (1), a variable $L_{t-1,h}$ is replaced by $L^*_{t-1,h}$, which is constructed as follows:

$$L^*_{t,h} = \begin{cases} L_{t,h} & \text{if } h \leq 10, \\ FL_{t,h} & \text{if } h > 10. \end{cases} \tag{2}$$

When RES and wind generation are considered, the model structure is slightly different. First, the AR component is shortened and contains only one lag, $p = 1$. This is motivated by the fact that neither wind nor solar generation exhibits weekly seasonality. TSO predictions comprise not only the forecasts for hour h, but also for the previous and next hour, respectively. It can be noticed that for hours $h = 1$ and 24 some of the information is not available. In such a case, the corresponding variables: $FRES_{t,h-1}$, $FW_{t,h-1}$ or $FRES_{t,h+1}$, $FW_{t,h+1}$, are not included in the regressions. The models take the following form:

$$W_{t,h} = \alpha_h + \theta_h W^*_{t-1,h} + \beta_{h,1} FW_{t,h-1} + \beta_{h,2} FW_{t,h} + \beta_{h,3} FW_{t,h+1} + \varepsilon_{t,h}, \tag{3}$$

$$RES_{t,h} = \alpha_h + \theta_h RES^*_{t-1,h} + \beta_{h,1} FRES_{t,h-1} + \beta_{h,2} FRES_{t,h} + \beta_{h,3} FRES_{t,h+1} + \varepsilon_{t,h}. \tag{4}$$

In order to prevent data leakage, for hours $h > 10$ the renewable energy generation is replaced by its forecast:

$$W^*_{t,h} = \begin{cases} W_{t,h} & \text{if } h \leq 10, \\ FW_{t,h} & \text{if } h > 10, \end{cases}$$

$$RES^*_{t,h} = \begin{cases} RES_{t,h} & \text{if } h \leq 10, \\ FRES_{t,h} & \text{if } h > 10. \end{cases}$$

A model describing the residual load comprises the above model specifications. Similarly to the model (1), it contains three lags of load level, the predicted load, together with information on the daily statistics of the forecasted load within the day t. Like in the model (4), it includes the information on the RES level on the previous day and the predicted RES generation in three consecutive hours: $h-1, h, h+1$, whenever available.

$$\begin{aligned} RL_{t,h} =& \alpha_h + \theta^L_{h,1} L^*_{t-1,h} + \sum_{p\in\{2,7\}} \theta^L_{h,p} L_{t-p,h} + \beta^L_{h,1} FL_{t,h} + \beta^L_{h,2} FL_{t,ave} + \beta^L_{h,3} FL_{t,max} \\ & + \beta^L_{h,4} FL_{t,min} + \theta^R_{h,1} RES^*_{t-1,h} + \sum_{i=-1}^{1} \beta^R_{h,2+i} FRES_{t,h+i} + \varepsilon_{t,h}, \end{aligned} \tag{5}$$

Finally, the last three models describe the behavior of market prices: DA, ID and their difference SP. These regressions are closely related to each other. They include seven deterministic dummy variables D_t, which capture weekly seasonality, an autoregressive component with lags $p \in \{1, 2, ..., 7\}$, TSO forecasts of fundamental variables ($FL_{t,h}$ and $FRES_{t,h}$), as well as the average, minimum and maximum levels of DA prices from the previous day, $t-1$. Furthermore, the model takes into account generation costs and includes fuel prices: coal (C_t) and gas (G_t).

$$\begin{aligned} DA_{t,h} = \alpha_h D_t + & \underbrace{\sum_{p\in\{1,...,7\}} \theta_{h,p} DA_{t-p,h}}_{\text{AR component}} + \underbrace{\beta_{h,1} DA_{t-1,ave} + \beta_{h,2} DA_{t-1,min} + \beta_{h,3} DA_{t-1,max}}_{\text{Daily quantities}} \\ & + \underbrace{\beta_{h,4} FL_{t,h} + \beta_{h,5} FRES_{t,h}}_{\text{Forecasts of fundamentals}} + \underbrace{\beta_{h,6} C_{t-1} + \beta_{h,7} G_{t-1}}_{\text{Fuel prices}} + \varepsilon_{t,h}. \end{aligned} \tag{6}$$

$$\begin{aligned} ID_{t,h} = \alpha_h D_t + \theta_{h,1} ID^*_{t-1,h} + & \sum_{p\in\{2,...,7\}} \theta_{h,p} ID_{t-p,h} + \beta_{h,1} DA_{t-1,ave} + \beta_{h,2} DA_{t-1,min} \\ & + \beta_{h,3} DA_{t-1,max} + \beta_{h,4} FL_{t,h} + \beta_{h,5} FR_{t,h} + \beta_{h,6} C_{t-1,h} + \beta_{h,7} G_{t-1,h} + \varepsilon_{t,h}. \end{aligned} \tag{7}$$

While the day-ahead prices for day t are all known at the time of forecasting, the intraday prices are represented by a variable ID^*, defined analogously to (2):

$$ID^*_{t,h} = \begin{cases} ID_{t,h} & \text{if } h \leq 10, \\ DA_{t,h} & \text{if } h > 10. \end{cases}$$

The model of price spread has an analogous structure to eq. (7). The only difference comes from the AR component, which includes lagged values of price spreads.

$$\begin{aligned} SP_{t,h} = \alpha_h D_t + \theta_{h,1} SP^*_{t-1,h} + & \sum_{p\in\{2,...,7\}} \theta_{h,p} SP_{t-p,h} + \beta_{h,1} DA_{t-1,ave} + \beta_{h,2} DA_{t-1,min} \\ & + \beta_{h,3} DA_{t-1,max} + \beta_{h,4} FL_{t,h} + \beta_{h,5} FR_{t,h} + \beta_{h,6} C_{t-1,h} + \beta_{h,7} G_{t-1,h} + \varepsilon_{t,h}. \end{aligned} \tag{8}$$

The intraday prices for hours after 10:00 are not known and therefore the $SP^*_{t,h}$ takes the following form

$$SP^*_{t,h} = \begin{cases} SP_{t,h} & \text{if } h \leq 10, \\ DA_{t,h} & \text{if } h > 10. \end{cases}$$

The use of ARX models can be motivated by their very low computational complexity and high interpretability coupled with a well-established performance in the literature. While the model specifications presented in this section are used for further algorithm steps in this paper, they may in practice be replaced with any point forecasting method.

3.2. Quantile regression

Quantile regression (QR) is a modeling approach that allows linking a selected quantile of a distribution, τ, of an endogenous variable Y_t with a vector of exogenous variables X_t. Since quantiles are well defined only for univariate processes, QR requires separate modeling of the variables of interest: electricity prices, load, and RES. Let us denote by $Q_\tau(Y_t)$ the τ quantile of Y_t. Then QR assumes that

$$Q_\tau(Y_t) = X_t \theta_\tau, \tag{9}$$

where X_t is a $(1 \times K)$ vector of explanatory variables and θ_τ is a $(K \times 1)$ vector of parameters. Notice that θ_τ depends on τ and changes with the analyzed quantiles. To keep the results comparable with other presented approaches, the variables included in the model (9) correspond to the ARX specifications from the previous Section 3.1.

The coefficients of (9) can be estimated by minimizing the sum of *pinball scores* (PS) throughout the calibration window. A PS_t for a given period t is defined as:

$$PS_t(\tau) = \begin{cases} (1-\tau)(Q_\tau(Y_t) - Y_t) & \text{for } Y_t < Q_\tau(Y_t), \\ \tau(Y_t - Q_\tau(Y_t)) & \text{for } Y_t \geq Q_\tau(Y_t). \end{cases} \tag{10}$$

The estimation process can be carried out for 99 percentiles: $\tau = 0.01, ..., 0.99$ and therefore QR can be used to approximate the entire distribution of Y_t. Here, QR is employed for the construction of prediction intervals (PI). A PI with a nominal coverage $1-\alpha$ can be estimated as

$$PI^{QR}_{1-\alpha} = [\ Q_{\alpha/2}(Y_{T+1}), \quad Q_{(1-\alpha/2)}(Y_{T+1})\ ],$$

where $Q_\tau(Y_{T+1})$ is a predicted quantile of Y_t for the period $T+1$.

3.3. Historical simulations

Historical simulation is a direct method of constructing probabilistic forecasts that is widely studied in the literature [32]. In order to obtain a multidimensional ensemble of forecasts, the method must be applied to all variables of interest at the same time. Let us denote by $Y_t \in R^K$ a K-dimensional vector of endogenous variables. In the case of electricity markets, the vector may include information about electricity prices and generation structure, for example $Y_t = [DA_{t,h}, ID_{t,h}, L_{t,h}, RES_{t,h}]$. Suppose that we want to calculate a forecast for a period $T+1$. Let us define a training set, S_{train}, as a window that covers periods preceding $T+1$, which is used to construct probabilistic predictions. The algorithm consists of the following steps:

1. Calculation of point forecasts $\hat{Y}_t$ for $t \in S_{train}$ with a moving window approach.

2. Estimation of forecast errors: $e_t = Y_t - \hat{Y}_t$.
3. Construction of the ensemble of predictions

$$\Psi = \{y \in R^K : y = \hat{Y}_{T+1} + e_t\}$$

In this research, point predictions used in steps $1 - 3$ are based on ARX models described in Section 3.1. It can be noticed here that although the variable Y_t is a vector, its individual elements are described by different equations. However, thanks to a simultaneous computation of the predictions, the residuals e_t maintain the correlation structure of the true forecast errors.

The collection of forecasts, Ψ, is next used to construct prediction intervals. In a classical form, a PI with a nominal coverage $1 - \alpha$ can be estimated as

$$PI_{1-\alpha}^{hist} = [\ Q_{\alpha/2}(\Psi), \quad Q_{(1-\alpha/2)}(\Psi)\ ], \tag{11}$$

where $Q_\tau(\Psi)$ is the τ quantile of the pool Ψ.

3.4. Multiple split method

In this research, we propose a *multiple split* forecasting method. It is an extension of an approach known in the literature under the name *split conformal prediction* or *inductive conformal inference* described by [28; 31; 29; 10]. The main idea of this forecasting scheme is to use a random split of the data to construct probabilistic predictions. First, the training data is divided into two disjoint windows: estimation and calibration. The first (estimation) subset is used to estimate the model parameters, which are then applied to calculate point predictions of the observations both within the training window and out-of-sample (i.e. target point prediction). The forecast errors are then estimated by computing the difference between the actual observations and their predictions in the calibration subset. Finally, the errors are used to approximate the distribution of the dependent variable. The way the are explored depends on the adopted methodology. When classical resampling methods are considered, they are used to directly approximate the quantiles of the distribution. In the context of conformal predictions, their absolute values are used to construct prediction intervals. Since our interest goes beyond prediction intervals, the first approach is adopted in this paper.

The *multiple split* approach extends previous works in various directions. Firstly, the random split is conducted multiple times in order to improve the forecast accuracy and decrease the variability of the outcomes. However, unlike in *leave-k-out* or *delete-d jackknife* approaches [see 34], we do not consider all possible divisions of the sample. This would significantly increase the computational complexity, but add little to the prediction quality. Second, forecast errors from the training window are used to directly construct an ensemble of forecasts rather than to estimate the distribution quantile. Therefore, the final ensemble is constructed by adding the results from individual splits, and there is no need for an intermediary step of averaging the quantiles or prediction intervals as in Lei et al. [28]. Finally, the approach is applied to predict a multidimensional random variable. We believe that resampling methods are of particular use in this context. They do not require parametric modeling of the multidimensional distribution, and at the same time maintain the correlation structure of forecasts and forecast errors.

In the *multiple split* method, similar to the historical approach, all variables are forecasted jointly and hence are collected in a K dimensional vector Y_t. Suppose that we observe a sample S that includes periods $t = 1, ..., T$ and we want to calculate a forecast for a period

$T+1$. Each iteration of the proposed algorithm consists of N independent splits. The procedure for a single split, $i = 1, ..., N$, is presented in Fig.3. On the graph, the estimation subset and associated steps are marked in green. Gray shading represents calibration periods. Striped boxes show the final forecasts. The placement of boxes indicates the order of operations (from left to right and from top to bottom), while the arrows show the dependency on the results from previous steps. The ith split consists of the following steps:

1. The sample is randomly divided into estimation ($S^{(i)}_{estim}$, green colour Fig. 3) and calibration ($S^{(i)}_{calib}$, grey colour Fig. 3) subsets such that $S = S^{(i)}_{estim} \cup S^{(i)}_{calib}$ and $S^{(i)}_{estim} \cap S^{(i)}_{calib} = \emptyset$.
2. The data in the estimation window, $S^{(i)}_{estim}$, is used to estimate the parameters of models used for forecasting, $\hat{\theta}_i$, which are next employed to calculate predictions for periods $t \in \{S^{(i)}_{calib}, T+1\}$: $\hat{Y}_{t,i} = X_t\hat{\theta}_i$
3. For each observation in the calibration window, $t \in S^{(i)}_{calib}$, the forecast error is calculated as $e_{t,i} = Y_t - \hat{Y}_{t,i}$
4. The ensemble of predictions is constructed as

$$\Psi_i = \{y \in R^K : y = \hat{Y}_{T+1,i} + e_{t,i}, t \in S^{(i)}_{calib}\}$$

Steps $1-4$ are repeated N times, and new sets of forecasts are added to the pool

$$\Psi = \bigcup_{i=1}^{N} \Psi_i. \tag{12}$$

Finally, analogously to the historical method, the ensemble of predictions is used to construct prediction intervals

$$PI^{MS}_{1-\alpha} = [\ Q_{\alpha/2}(\Psi), \quad Q_{(1-\alpha/2)}(\Psi)\ ], \tag{13}$$

where $1-\alpha$ is the nominal coverage level of PI.

As a result, the algorithm provides a set of forecasts that are derived from different splits of the data. In the proposed approach, we aggregate information rather than average prediction intervals as in [28]. This method has several advantages. First, by aggregating the ensembles instead of the distributions, we do not need to decide on the averaging method for probabilistic (e.g. quantile) forecasts [see 35; 36; 37]. Second, the Bonferroni averaging method applied by Lei et al. [28] to construct prediction intervals limits the number of splits, N, that can be used, as it requires estimation of a $\alpha/2N$ and $1-\alpha/2N$ quantiles of data. Finally, it allows us to use a wider variety of measures for assessing the forecast accuracy.

3.5. Forecast evaluation

Probabilistic forecasts discussed in previous sections represent two popular types of prediction: quantile forecasts and ensemble forecasts. Although it is possible to estimate quantiles from a set of predictions, it is more difficult to generate a diversified ensemble from a set of quantiles. Therefore, in this article, we use two separate approaches to assess the forecast accuracy that are dedicated to one of these types of predictions.

First, using QR, historical simulations, or the multiple split method, we approximate the distribution of the variables of interest with 99 quantiles and construct prediction intervals of nominal coverage 80%, 90%, 95% and 98%. The precision of PIs is evaluated with the *PI*

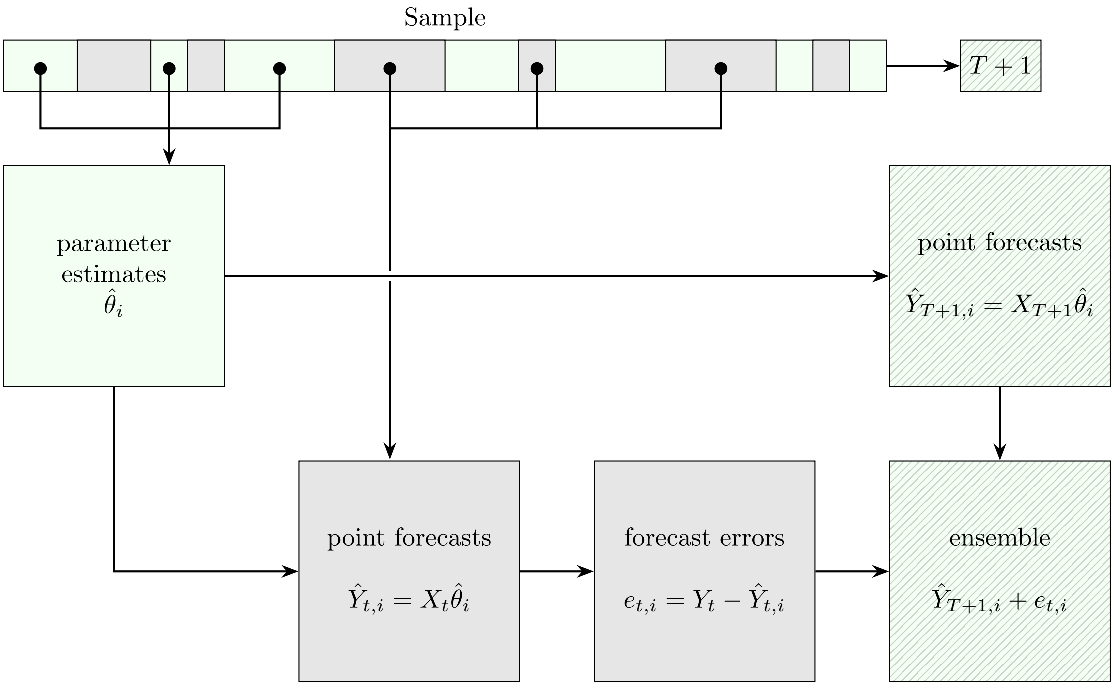

Sample
$T+1$
parameter estimates $\hat{\theta}_i$
point forecasts $\hat{Y}_{T+1,i} = X_{T+1}\hat{\theta}_i$
point forecasts $\hat{Y}_{t,i} = X_t\hat{\theta}_i$
forecast errors $e_{t,i} = Y_t - \hat{Y}_{t,i}$
ensemble $\hat{Y}_{T+1,i} + e_{t,i}$

Figure 3: Schematic illustration of the algorithm.

coverage probability (PICP). The measure is based on the average number of observations in the testing window that fall into the PI. For an hour h, it is calculated as

$$PICP_h = \frac{1}{T}\sum_t \mathbb{I}(Y_{t,h} \in PI_{t,h}), \tag{14}$$

where $\mathbb{I}(Y_{t,h} \in PI_{t,h})$ is an indicator function that takes value 1 when the variable $Y_{t,h}$ falls to the prediction interval $PI_{t,h}$ and zero otherwise. To obtain the final value, the $PICP$s for individual hours are averaged

$$PICP = \frac{1}{24}\sum_h PICP_h. \tag{15}$$

From the definition, this empirical coverage should be as close as possible to the nominal one. To evaluate whether $PICP_h$ is sufficiently close to the nominal coverage, we perform a Kupiec test [38] for each hour of the day separately. The null hypothesis says that the empirical coverage equals the nominal level, whereas under the alternative the $PICP_h$ differs significantly from $1-\alpha$. In this article, we report the percentages of hours for which we were unable to reject the null at the significance level 5%. Therefore, the closer the measure is to one, the more successful the prediction method is in providing PIs of a predefined probability level.

Finally, to assess the quality of the estimated 99 quantiles, we use CRPS described by Gneiting et al. [39] that evaluates both the calibration and the sharpness of the distribution. The measure is defined as an integral of the the pinball score defined in eq. 10 over the entire predictive distribution, and can be approximated for quantile forecasts as the arithmetic mean of pinball scores. For each of the 99 percentiles $\tau = 0.01, ..., 0.99$, we calculate $PS_{t,h}(\tau)$. The $CRPS$ of a given observation can then be calculated as

$$CRPS_{t,h} = \frac{1}{99} \sum_{\tau} PS_{t,h}(\tau). \tag{16}$$

The measure for the whole out-of-sample period is an average over all observations:

$$CRPS = \frac{1}{T} \frac{1}{24} \sum_{t} \sum_{h} CRPS_{t,h}. \tag{17}$$

The lower the value of $CRPS$, the better the approximation of the distribution. It should be noted here that the sharpness of the distribution has a significant impact on the measure. However, sharpness is a criterion that should be analyzed only when the calibration is correct. Furthermore, since the pinball score is a loss function used to estimate the QR, CRPS may favor the results that arise from this forecasting method. Thus, we believe that CRPS should be analyzed together with other measures, such as PICP.

In this article, we also use evaluation methods that are dedicated only to ensemble forecasts. In particular, we apply the reliability index [9], which assesses whether the ranks of actual observations within the pool of predictions have a distribution close to uniform.

In the case of a univariate variable, $Y_{t,h}$, we denote by $r_{t,h}$ a proportion of ensemble forecasts smaller than or equal to $Y_{t,h}$. Since $r_{t,h}$ resembles a cumulative distribution function, it should have a uniform distribution. Let us divide an interval $[0, 1]$ into M equal bins: $B_1, ..., B_M$, and denote by $f_{j,h}$ the frequency of $r_{t,h}$ falling into bin j in the out-of-sample period.

$$f_{j,h} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}(r_{t,h} \in B_j).$$

Then the reliability index can be calculated as follows

$$\Delta_h = \sum_{j=1}^{M} |f_{j,h} - \frac{1}{M}|. \tag{18}$$

The lower the value of Δ_h, the closer the empirical distribution of $r_{t,h}$ is to the uniform one. The final measure of the discrepancy is calculated as an average of Δ_h over 24 hours.

$$\Delta = \frac{1}{24} \sum_{h=1}^{24} \Delta_h \tag{19}$$

To assess the accuracy of the multivariate probabilistic forecast, we use the idea of a multivariate rank histogram described by [9]. Given the ensemble of forecasts, Ψ, and the verifying observation, $Y_0 \in R^K$, the procedure consists of the following steps:

1. First, the *pre-ranks* are assigned, such that

$$\rho_j = \sum_{i=0}^{M} \mathbb{I}(Y_i \preceq Y_j),$$

where $Y_i \in \Psi$ for $i = 1, ..., M$. Moreover, $Y_i \preceq Y_j$ if and only if $Y_{i,k} \leq Y_{j,k}$ for all $k = 1, ..., K$.

2. Next, *the multivariate rank* is calculated as the rank of the pre-ranks. When two observations have the same pre-rank, the final rank is assigned randomly. Let us adopt the following notation:

$$r_1 = \frac{1}{M}\sum_{j=0}^{M} \mathbb{I}(\rho_j < \rho_0),$$

$$r_2 = \frac{1}{M}\sum_{j=0}^{M} \mathbb{I}(\rho_j < \rho_0).$$

   Then the final rank, r, is chosen randomly from the set $\{r_1 + 1, ..., r_1 + r_2\}$.
3. Similarly to the univariate case, to calculate the reliability index, divide an interval $[0, 1]$ into M equal bins and denote by $f_{j,h}$ the frequency of $r_{t,h}$ falling into bin j. Then the discrepancy indices can be calculated according to (18) - (19).

The multivariate rank histogram shows whether the multidimensional distribution is well calibrated to the data. It depends on the quality of the marginal distributions as well as on the ability to approximate the correlation between the variables.

4. Support of the decision process of a wind farm

In this research, we show how the joint prediction of different market fundamentals can be used to support the decision process in the electricity market. We analyze the trading decisions of a company that owns wind farms spread throughout Germany. The amount of energy produced is assumed to be small enough not to influence market prices. Each day, wind turbines generate electricity, which is sold in the power energy exchange. The company has access to the day-ahead (DA) and intraday (ID) markets and does not speculate; i.e., it aims to sell the entire generated energy in either of these markets. Although purchasing energy on the intraday market may be necessary in the case of an overestimated generation forecast, this operation is never done intentionally.

The day before delivery, the company places bids on the DA market. The decision on how much to offer there has a significant impact on the trading outcome. Due to the intermittent nature of the wind, the company faces uncertainty about the level of generation that hampers the decision-making process. If the company sells less electricity in the DA market than it produces, the remaining generation needs to be offered in the ID market. On the contrary, if it offers more than it generates, it needs to buy the missing production in the ID market. To construct the offer, the company may use predictions of both its generation and market conditions. Similarly to [11], decision is described by a parameter $q \in [0, 1]$ that represents the fraction of the forecasted wind generation $\hat{W}$ offered in the day-ahead market.

In this paper, it is assumed that the company faces some operational and maintenance costs $C_{O\&M}$, which influence the level of marginal profits. For simplicity, these costs are kept constant and expressed in EUR per 1 MWh of production. Other types of cost are interpreted as fixed costs and therefore do not influence the decision process. Then the profit of the company can be calculated as follows

$$\Pi(q) = \underbrace{q\hat{W}_{t,h}DA_{t,h}}_{\text{DA market}} + \underbrace{(W_{t,h} - q\hat{W}_{t,h})ID_{t,h}}_{\text{ID market}} - \underbrace{W_{t,h}C_{O\&M}}_{\text{O\&M costs}}. \tag{20}$$

Hence, the profit per 1MWh of generation is

$$\pi(q) = q\hat{w}_{t,h}DA_{t,h} + (1 - q\hat{w}_{t,h})ID_{t,h} - C_{O\&M}, \tag{21}$$

where $\hat{w}_{t,h}$ shows how close is the predicted generation level to the actual one: $\hat{w}_{t,h} = \hat{W}_{t,h}/W_{t,h}$. When $W_{t,h} = 0$ then the average profit is also set to zero.

It should be noted here that the level and distribution of income depend on three main sources of uncertainty associated with the unknown level of generation, $W_{t,h}$, and market prices: $DA_{t,h}$, $ID_{t,h}$. Since profit is a nonlinear function of these variables, its distribution is nontrivial. Resampling and simulation methods, such as the historical approach or the multiple split method, are of great help in estimating its probabilistic forecasts. The set of joint predictions of different market fundamentals can be used to construct an ensemble of future profits, which in turn can be used to approximate the income distribution.

4.1. Bidding strategies

In this research, various trading strategies are considered. The benchmark **naive strategy** assumes that the entire predicted generation is sold on the DA market. The profitability of this approach depends primarily on the accuracy of wind generation forecasts. It is assumed that the company places an unlimited bid for the quantity $\hat{W}_{t,h}$ on the DA market and balances the position on the ID market. Hence, the parameter q is fixed and equal to $q = 1$. This implies that the strategy responds only to the fluctuation of the generation but does not adjust to the market situation.

Next, three data-driven strategies are proposed that explore probabilistic profit forecasts to design the optimal trading strategy. As mentioned above, the strategy is adjusted with a parameter q. For a grid of different values of q, the ensemble of predictions of $\pi(q)$ is constructed. The set of future profit values, $\Psi(q)$, is used to derive the optimization criteria and select the best value of q.

The first and most natural approach to choosing q is to maximize the expected value of the profits, henceforth called the **expected profit strategy**. It is based on a point forecast of income, calculated here as the median of the ensemble $\Psi(q)$. The value of q is chosen as the one for which the median is the highest. This strategy is expected to bring high income at the cost of increased risk.

The second **VaR strategy** is relevant for more risk-averse traders. It focuses on minimizing risk, commonly represented in business applications by Value-at-Risk (VaR). Here, we calculate the Value-at-Risk as the 5th percentile of the ensemble. Note that VaR is often expected to be negative and therefore presented as the absolute value of the quantiles of predicted or historical returns, in which case the quantity should be minimized. In this application, it is possible that VaR is positive (i.e., even the worst case scenario would not lead to losses), and therefore we do not apply absolute value, leading to higher VaR being more desirable.

The final strategy, called **Sharpe Ratio strategy**, aims to find a balance between maximizing profits and minimizing risk. To do this, the Sharpe ratio (expected profit divided by its standard deviation, [40]) is calculated for different values of q. The optimal q is chosen so that the ratio is maximized. The Sharpe ratio strategy, contrary to two previous approaches, explores information about the entire predictive distribution instead of its selected quantile.

The three data-driven strategies can be summarized as follows:

1. **Expected profit strategy** ($E\pi$): $q = \text{argmax}_{q\in[0,1]} Q_{50\%}(\Psi(q))$,

2. **VaR strategy** (VaR): $q = \text{argmax}_{q\in[0,1]} \text{Q}_{5\%}(\Psi(q))$,
3. **Sharpe ratio strategy** (SR): $q = \text{argmax}_{q\in[0,1]} \text{SR}(q) = \text{argmax}_{q\in[0,1]} \frac{\bar{\pi}(q)}{\sigma_\pi(q)}$, where $\bar{\pi}(q)$ and $\sigma_\pi(q)$ denotes the average value and the standard deviation of the predicted profits in the ensemble $\Psi(q)$, respectively.

It can be noted that in the above specifications, the value of the profit per 1 MWh of generation is used instead of the total profit. In addition, in the remaining parts of the article, the profit per 1 MWh of generation is used to evaluate the results of the experiment because π does not depend on the scale of the company (as long as it does not have market power to impact electricity prices).

4.2. Stopping rules

It is well documented in the literature and is visible in Fig. 2 that electricity prices can fall below zero. The existence of very low or negative prices implies that even an optimal bidding strategy can cause losses. To avoid such a situation, we assume that the company is allowed to curtail the production. It means that it either stops the generation (the turbines are turned off) or stores electricity. The second solution becomes an attractive alternative as more and more investments are made in the development of energy storage systems.

In our research, the production curtailment implies that no electricity is sold in any of the markets, which simulates turning off the turbines without storing the unsold generation. In the case of naive strategies, the trader may place a limited bid instead of an unlimited bid on the DA market and set a lower bound for a price at zero. We call this a **naive limited bid strategy**. In case of data driven strategies, more complex solutions are available that account for a risk aversion of company owners. We assume that the generation is stopped when a selected quantile of the profit distribution is less than zero. It means that the company is engaged in the trade when $Q_\tau(\Psi(q^*)) \geq 0$ and curtails the production for $Q_\tau(\Psi(q^*)) < 0$. The fraction of predicted generation offered on the DA market, q^*, is selected with one of the three approaches: $E\pi$, VaR or SP discussed in the previous Section 4.1 .

It can be noticed that a risk-neutral trader would likely base the decision on the center of the distribution (for example, a median), whereas a risk-averse trader would place bids depending on pessimistic scenarios (low quantiles). A trader who is reluctant to stop production may consider high quantiles that exceed 50%. Finally, as the aversion to curtailment increases, the selected quantile may converge to one, which will represent the strategy that assumes daily trade (i.e. without stopping generation).

4.3. Evaluation of trading strategies

The performance of the presented trading strategies is evaluated according to the level of average profit, the trading risk, and the frequency of generation curtailment. Hence, we provide the company with a broad perspective, which encompasses potential preferences and aversions of the trader.

When the income level is considered, two quantities are reported: the average profit and the average profit per trade. The main difference between these measures results from the approach towards days when the production is curtailed. The average profit includes the whole testing period and is calculated as the weighted mean of individual hourly profits:

$$\bar{\pi} = \frac{1}{24T} \sum_{t=1}^{T} \sum_{h=1}^{24} \pi_{t,h}(q^*), \tag{22}$$

where q^* is an optimal value of the parameter selected with the analyzed strategy (q^* changes over days and hours). On the contrary, the profit per trade, $\tilde{\pi}$, is calculated only for periods when the trade occurs, and therefore observations when the generation is curtailed are disregarded. These two measures, $\bar{\pi}$ and $\tilde{\pi}$, are different only for strategies with a stopping rule.

To better understand the significance of the stopping rule, the frequency of trade is computed. It shows how often the stopping rule is not violated. We expect that a frequent curtailment of generation may result in a rise of the average profit per trade and a fall in the average total profit.

Finally, the trading risk is evaluated with the VaR measure calculated as the 5% quantile of the average profit $\pi_{t,h}$. In the calculation of VaR only the days when the trade occurs are taken into account. Otherwise, VaR for many strategies with stopping rules will simply be zero, reflecting the profit on days when the generation is curtailed. However, in these periods, the company is not exposed to a trading risk.

5. Results

The results of the research are divided into two parts. First, Section 5.1 focuses on evaluating the statistical accuracy of probabilistic forecasts obtained with the methods discussed above. Next, in Section 5.2, the performance of the trading strategies based on the multiple split approach is assessed and the economic value of the predictions is measured. In the analysis, we adopt the following specifications:

- The evaluation period consists of 730 observations.
- Two different window sizes are used to calibrate the parameters: 365 and 730 days, which correspond to one and two years of observations, respectively.
- In the multiple split approach, the sample is evenly split between the estimation and calibration parts. When the window of 365 days is considered, the calibration window consists of 183 observations.
- Two values of the number of splits are evaluated: $N = \{1, 20\}$. The results of these specifications are denoted $MS(1)$ and $MS(20)$, respectively.

The forecasting experiment is based on a rolling window scheme, in which each hour of the entire two-year validation period is predicted separately.

5.1. Statistical accuracy of probabilistic forecasts

To evaluate the accuracy of forecasts, we first consider the four fundamental variables that describe the electricity markets. DA price, ID price, total load, and RES generation. Next, the ability to predict a linear combination of these variables is evaluated: price spread (a difference between the DA and ID prices) and residual load (a difference between load and RES). In the case of QR models, the residual load regression is specified as eq. (5). When the price spread is predicted, the exogenous variables included in the model are the same as in eq. (6). For ensemble forecasts based on historical simulations or multiple split methods, the set of predictions is constructed as a function of individual elements (univariate forecasts) from the pool. Hence, there is no need to specify separate models for forecasting the linear combination of fundamental variables.

5.1.1. Forecasting of market fundamentals

First, let us consider the probabilistic forecasts of the individual fundamental variables. The results presented in Table 2 show the PI coverage probabilities for four levels of PIs: 80%, 90%, 95%, and 98%, together with the average frequencies of the Kupiec test indicating a correct coverage and the levels of the CRPS measure. Forecasts are calculated using two different sizes of calibration windows: 365 and 730 days. The outcomes show that all the forecasting methods provide PIs that are too narrow and hence have PICP below the nominal level. However, the PIs of the ensemble forecasts seem to be better calibrated to the data. In particular, for $T = 365$, PICP of QR are much below the nominal coverage level, while for MS (20) they are close to $1 - \alpha$. This finding is confirmed by the results of the Kupiec test. In case of electricity prices (columns DA and ID, Table 2), the test indicates that the coverage level of the QR method is correct in less than 2% cases for T=365. At the same time, for MS(20) forecasting method, the PICPs are not statistically different from the nominal level in more than 90%. When the longer calibration window is considered, the numbers are 43.33% and 76.67% for the ID prices, respectively. For load and RES, the differences between QR and MS are less substantial but still visible. The proportion of cases for which the Kupiec test cannot reject the null is greater for MS (20) than QR by almost 63 and 20 percentage points for load and RES, respectively.

When ensemble forecasts are analyzed, it can be observed that historical simulations have PICPs closer to the nominal level than QR and MS(1) but worse than MS(20). Moreover, the results for different MS specifications show that an increase of the number of splits improves the calibration of the quantiles. There are also differences in behavior of the approaches for different sizes of calibration windows. MS methods work relatively better for shorter windows with $T = 365$, while historical simulations have empirical coverage closer to the nominal level for $T = 730$. In this respect, the QR performs similarly to the historical simulations.

Finally, the forecasting methods can be compared on the basis of CRPS. As the measure is based on the pinball score, which is minimized when estimating QR, it is not surprising that QR outperforms other approaches in this context. However, the differences between QR and MS(20) are moderate, and MS(20) has the lowest CRPS of all the methods for Load. Unlike in case of PICP, the historical simulations are outperformed by both QR and MS(20). Also, the comparison of different window sizes leads to different conclusions than PICP showing that a shorter window is preferable for all the models.

Next, let us take a deeper look at the ensemble methods. Table 3 presents the reliability index (RI) for both univariate and multivariate analysis. It can be noted that the historical simulation can be interpreted as the multivariate approach because it provides residuals that maintain the correlation structure of the forecast errors. When MS approaches are considered, we show first the results for cases where MS is applied independently to all variables (called *Uncorr.* in Table 3). Second, the outcomes of the simultaneous forecasting of the fundamentals are presented as MS *Corr.* Joint modeling allows to approximate the correlation of the residuals. Similarly to the evaluation of quantile predictions, the reliability index is shown separately for short and long calibration window sizes.

The results of the marginal distributions show that the MS(20) approach provides the lowest index value for all variables, except for the DA, $T = 730$. Moreover, it is demonstrated that from the perspective of an individual variable, there are no differences between joint and separate applications of the MS method. Therefore, there are no gains of multidimensional modeling. Finally, similar to PICP, the reliability index indicates that MS(1) is inferior to MS(20) and the historical approach. When the forecasts of the multidimensional distributions are evaluated (last column of Table 3), the results confirm the superiority of the MS(20) method. It provides the lowest value among all specifications $RI = 0.3027$. Furthermore, substantial differences could be observed between approaches that account for the correlation of forecast errors or not. For example, for $T = 365$ and MS(20), the reliability index is 0.3267 for uncorrelated residuals and 0.3027 for correlated ones. This indicates substantial gains of using multidimensional modeling in joint forecasting of market fundamentals.

5.1.2. Forecasting of linear combinations of market fundamentals

Since the analysis of the prediction accuracy of the fundamentals of the market indicates that MS(1) is inferior to MS(20), only the results for the latter are presented in the following sections. Next, to assess the gains from multidimensional forecasting, we analyze separately the outcomes for MS approach computed separately (*Uncorr.*) or jointly (*Corr.*) for all four fundamental variables. In case of the QR, the multidimensional modeling is not available, and therefore the price spread and the residual load need to be forecasted directly.

Table 4 shows three measures that describe the accuracy of quantile predictions: PICP, proportions of the Kupiec test indicating a correct empirical coverage, and the average value of CRPS. When the PICP is considered, the outcomes confirm that QR provides prediction intervals which are too narrow and therefore exhibit coverage much lower than the nominal levels. This property is particularly well visible when the rejection of the Kupiec test is analyzed. In the case of QR, the test does not allow one to reject the null of a correct coverage in only 4.15% and 12.5% of the cases for Price Spread and Residual Load, respectively ($T = 365$). As the calibration window increases, the frequencies increase accordingly to 62.50% and 24.17%, but still remain much lower than for historical simulations and MS(20) with correlated errors. Similarly to previous outcomes, it can be observed that the accuracy of QR predictions increases as the sample size rises from 365 to 730 days.

The empirical coverage measured by PICP of both ensemble methods is much closer to the nominal level than in the case of QR. For a shorter calibration window, the results of the Kupic test show that the coverage of the analyzed PIs is not statistically different from the nominal level in 73.33% and 92.5% cases for the MS (20) approach (MS, *Corr.*). The frequencies are slightly lower for the historical simulation method and are approximately 20.00% and 46.67%.

When the results of the CRPS measure are examined, the outcomes support previous findings, which show that the QR method provides a forecast with the lowest CRPS level. This indicates that the probabilistic forecasts of QR are sharper than those of other methods.

Finally, as multidimensional models are compared with independent forecasting of fundamental variables, the results indicate the superiority of the first approach. The MS(20) method that does not account for the correlation of forecast errors (MS, *Uncorr.*), provide too wide intervals with the empirical coverage far from the nominal level. The problem is particularly severe for the Price Spread, for which the PICP of 80% PI exceeds 92%. Furthermore, the performance of the method does not improve when a longer calibration window is used. In this case, the Kupiec measure falls below 6% and 50% for Price Spread and Residual Load, respectively.

Table 2: Accuracy measures of probabilistic forecasts based on quantile predictions: market fundamentals

Model	T=365				T=730			
Data	DA	ID	Load	RES	DA	ID	Load	RES
PICP	QR							
80%	71.66%	72.99%	76.39%	78.22%	75.52%	77.52%	76.04%	79.09%
90%	82.28%	83.54%	86.78%	88.28%	86.06%	87.10%	87.18%	89.45%
95%	87.71%	89.46%	92.49%	93.41%	91.41%	92.06%	93.23%	94.32%
98%	90.72%	92.06%	96.00%	96.72%	95.41%	96.11%	96.54%	97.29%
Kupiec (%)	1.67%	0.00%	25.83%	74.17%	17.50 %	43.33%	35.83%	90.83%
CRPS	1.8692	2.5276	1.9691	1.8576	2.0067	2.6453	2.2387	1.9021
PICP	Historical							
80%	75.73%	76.28%	77.32%	77.55%	75.47%	76.50%	77.45%	79.02%
90%	86.31%	86.33%	87.13%	87.50%	87.57%	88.11%	88.56%	89.27%
95%	92.55%	92.80%	93.09%	93.50%	93.78%	93.87%	93.76%	94.70%
98%	96.68%	96.72%	97.03%	97.35%	97.65%	97.48%	97.35%	97.72%
Kupiec (%)	20.83%	33.33%	54.17%	74.17%	45.00%	55.00%	84.17%	98.33%
CRPS	2.0154	2.7723	2.0188	1.9258	2.1130	2.7387	2.2521	1.9440
PICP	MS(1)							
80%	73.10%	74.08%	77.00%	77.19%	72.60%	74.83%	75.05%	77.34%
90%	84.10%	85.36%	87.75%	87.42%	84.69%	85.51%	87.03%	88.07%
95%	91.36%	92.24%	94.07%	93.32%	91.51%	92.04%	93.15%	94.02%
98%	96.61%	96.92%	97.66%	97.15%	96.55%	96.86%	96.87%	97.41%
Kupiec (%)	10.83%	13.33%	70.83%	64.17%	8.33%	20.00%	28.33%	73.33%
CRPS	2.0749	2.7761	2.0129	1.9077	2.1627	2.7905	2.2774	1.9404
PICP	MS(20)							
80%	80.07%	80.76%	78.83%	78.61%	76.99%	78.50%	76.19%	78.12%
90%	90.13%	90.33%	89.75%	88.53%	87.95%	88.34%	88.14%	88.62%
95%	94.78%	94.95%	95.03%	93.81%	93.25%	93.63%	93.68%	94.25%
98%	97.67%	97.92%	97.90%	97.33%	97.51%	97.49%	97.05%	97.58%
Kupiec (%)	90.00%	98.33%	98.33%	92.50%	60.00%	76.67%	51.67%	81.67%
CRPS	1.9090	2.5653	1.9689	1.8843	2.0780	2.6873	2.2542	1.9284

Table 3: Reliability index: univariate and multivariate analysis: market fundamentals

		DA	ID	Load	RES	All
Model		T = 365				
Historical		0.3386	0.3460	0.3363	0.3400	0.3350
MS(1)	Uncorr.	0.4053	0.3797	0.3415	0.3380	0.3660
	Corr.	0.4034	0.3769	0.3432	0.3393	0.3551
MS(20)	Uncorr.	0.3315	0.3092	0.2923	0.2927	0.3267
	Corr.	0.3375	0.3133	0.3049	0.2889	0.3027
		T = 730				
Historical		0.3245	0.3185	0.3172	0.2904	0.3123
MS(1)	Uncorr.	0.3631	0.3401	0.3363	0.3129	0.3444
	Corr.	0.3598	0.3407	0.3308	0.3140	0.3327
MS(20)	Uncorr.	0.3219	0.3112	0.3106	0.2941	0.3359
	Corr.	0.3292	0.3112	0.3044	0.2893	0.3202

Table 4: Accuracy measures of probabilistic forecasts based on quantile predictions: linear combination of market fundamentals

	QR		Historical		MS(20)			
					Corr.		Uncorr.	
	Spread	RL	Spread	RL	Spread	RL	Spread	RL
PICP	T=365							
80%	74.71%	75.88%	75.83%	76.98%	82.61%	78.76%	92.74%	76.03%
90%	85.13%	86.23%	86.04%	86.92%	91.30%	89.16%	96.15%	87.01%
95%	90.67%	91.82%	92.00%	92.96%	95.55%	94.19%	97.87%	92.99%
98%	92.89%	95.02%	96.56%	97.12%	97.61%	97.22%	99.02%	97.15%
Kupiec (%)	4.17%	12.50%	20.00%	46.67%	73.33%	92.50%	18.33%	65.00%
CRPS	1.9176	1.9182	2.1840	2.9201	1.9870	2.8387	2.0984	3.2046
PICP	T=730							
80%	78.85%	75.74%	77.77%	76.86%	80.39%	76.90%	92.73%	76.59%
90%	88.23%	86.82%	88.09%	87.08%	89.90%	87.40%	96.48%	87.41%
95%	93.24%	92.77%	93.92%	93.13%	94.55%	93.18%	98.04%	93.33%
98%	96.31%	96.22%	97.35%	96.85%	97.61%	96.90%	99.16%	97.12%
Kupiec (%)	62.50%	24.17%	72.50%	44.17%	91.67%	45.83%	5.83%	46.67%
CRPS	1.9264	2.2044	2.2268	3.2397	1.9918	3.1097	2.1293	3.1101

5.2. Economic value of forecasts

To assess the economic value of the newly proposed method, we use probabilistic predictions to support the decision process of a generation utility. As described in Section 4, the company owns wind farms that are spread throughout Germany and needs to decide on the share of predicted production, q, offered on the DA market. The remaining part of the generation is sold on the ID market. In the case of wind generation, operational and maintenance costs account for about 25%-35% of the *levelized energy cost* (LCOE) [41; 42]. It implies that O&M costs vary between USD 10-30 per MWh [43]. In the article [44], the costs are estimated at USD 11 per MWh for onshore installations and are expected to decrease in the future. Since the exchange rate between USD and EUR has oscillated between 0.9-1 USD/EUR, in this research we assume that $C_{O\&M} = 10$.

First, let us analyze the average total income earned from different strategies. The results are presented in Fig. 4, in which lines marked with dots show the outcomes of strategies without production curtailment, and the lines marked with crosses present the strategies that restrict trade. The naive benchmark, which assumes that all predicted generation is sold on the DA market, $q = 1$, is depicted to facilitate comparison. It earns on average 27.27 EUR/MWh. The results show that the adoption of data-driven strategies is profitable even when no curtailment is allowed. The highest revenue is achieved for the approach that aims to maximize the expected profit. A strategy optimizing the Sharpe ratio is only slightly worse, and brings 27.55 EUR/MWh instead of 27.56 EUR/MWh on average. When the results are compared with the benchmark, it can be seen that they bring 1.39% and 1.46% more, respectively.

The application of strategies that enable production curtailment requires the selection of a profit quantile that is used as a threshold in the stopping rule. Therefore, in the case of three data-driven approaches, revenue depends on the quantile τ. For low values of τ, generation is often reduced and, therefore, the company loses potential income. As the threshold quantile increases, the profit rises and exceeds both the naive benchmark and the limited bid. The strategies obtain their maximum of 27.95, 27.92, and 27.83 EUR/MWh for strategies based on $E\pi$, Sharpe ratio and VaR. Finally, as seen in Fig. 4, income starts to decrease and falls to the level of the corresponding strategies without stopping the generation. The results indicate that the curtailment of production leads to a substantial increase of income. Compared to the naive benchmark, the limited bid strategy rises profits by 2.03%. In the case of data-driven approaches, the increase is even greater and reaches 2.87% for $E\pi$, 2.77% for Sharpe ratio and 2.42% for VaR strategies. This means that the stopping rule increases the profits by additional 2%.

When the profit per trade is considered, it can be noticed that for strategies without curtailment, the average profit and the profit per trade are the same. Therefore, Fig. 5 presents only the results of the strategies with stoppage of production. The profits of the naive benchmark are added to the plot for comparison. The results show that the risk averse strategies that are based on low quantiles of profits lead to an increase of the profit per trade to 31 EUR per MWh. This implies that it exceeds the income of the naive benchmark strategy by 15.34%. As τ increases, the frequency of trade also rises, and the income per trade converges to the average total income presented in Fig. 4. When the limited bid is considered, the profit per trade reaches 28.22 EUR/MWh and surpasses the naive benchmark by 3.85%. However, income remains lower than for data-driven approaches for most thresholds, $\tau < 70\%$. Only when the company becomes reluctant to curtailing the generation, the outcome of the limited offer exceed those of other strategies. Finally, it can be observed that among the data-driven approaches, the one that maximizes the expected value of profits is the most profitable. Similarly to previous results, it is followed by the Sharpe ratio strategy. The lowest income is earned by the VaR approach.

The performance of different strategies is summarized in Table 5, which shows the frequency of trade, average profits and average profits per trade relative to the naive benchmark. The outcomes of approaches that allow for stoppage of productions depend on the quantile of profits used as the threshold. When τ is set equal to one (last column, Table 5), the strategy is equivalent to an approach without the stopping rule. The results indicate that data-driven methods lead to a reduction in the generation of 0.00%–14.22%. The greatest improvement is obtained for $\tau \in [0.3, 0.5]$, when the decrease in the generation frequency by less than 5% brings an increase in profits of more than 2.5% in the case of average profits and 5-7% in the case of average profit per trade. When the limited bid approach is considered, it can be seen that it leads to production curtailment in less than 2% cases. At the same time, it brings an additional 2% of the average profits. The strategies without stopping rule, shown in the last column of the table, confirm previous findings and demonstrate that they lead to a moderate increase in profits (between 0.88% and 1.46%).

Economic evaluation is not complete without risk analysis. Here, 5% VaR is used to measure profit in the case of a pessimistic scenario. The VaR for different strategies is presented in Fig. 6. The plot shows the results of the strategies without (dot markers) and with stopping rules (cross markers). First, it can be noticed that the naive benchmark is characterized with VaR slightly above zero. Other approaches guarantee a positive profit even at the bottom 5% of the scenarios. When the income of strategies without stopping rule is considered, then the increase of VaR is moderate. Significantly lower risk is incurred by approaches that allow for generation curtailment. The limited bid approach provides VaR at 4.92 EUR/MWh. The results of data-driven strategies depend on the adopted threshold quantile, τ, and vary between 2.05 and 13.88 EUR/MWh. Finally, although the plot resembles the income per trade figure, it can be observed that the ordering of data-driven approaches is different. The least risky is the VaR strategy, while the $E\pi$ approach is characterized by the lowest value of VaR. Hence, more profitable strategies are at the same time more hazardous.

5.2.1. Portfolio analysis

The data-driven approaches presented above are characterized by different levels of profit and associated risks. Similarly to other commodities, contracts that bring greater income are associated with a higher level of uncertainty. Hence, the selection of the strategy depends on the risk appetite of the utility owners. To make a proper decision, it could also be beneficial for the company to understand where the difference comes from. Fig. 7 shows the histograms of the decision variable, $q*$, for strategies without generation curtailment: $E\pi$ (*left panel*, Fig. 7), SR (*middle panel*, Fig. 7) and VaR (*right panel*, Fig. 7). It could be seen that they differ significantly in terms of the given recommendations. The strategy that maximizes expected profit selects in most cases extreme values of q. Hence, it offers the entire forecasted generation in the DA market or leaves the whole production for the ID market. Moreover, it chooses the ID market in 52% of the cases, compared to only 36% of the times when DA is selected. In contrast, the VaR strategy provides mainly diversified portfolios of contracts and recommends selling electricity in both markets. It selects $q = 0$ or $q = 1$ in 5% and 17% of the cases, respectively. Furthermore, most of the time more than half of the predicted generation is offered on the DA market ($q \geq 0.5$). In 40% of the periods, it recommends choosing $q \geq 0.8$. As observed previously, the strategy SR stays between $E\pi$ and VaR . It selects intermediate values of q in more than 60% of cases. However, similar

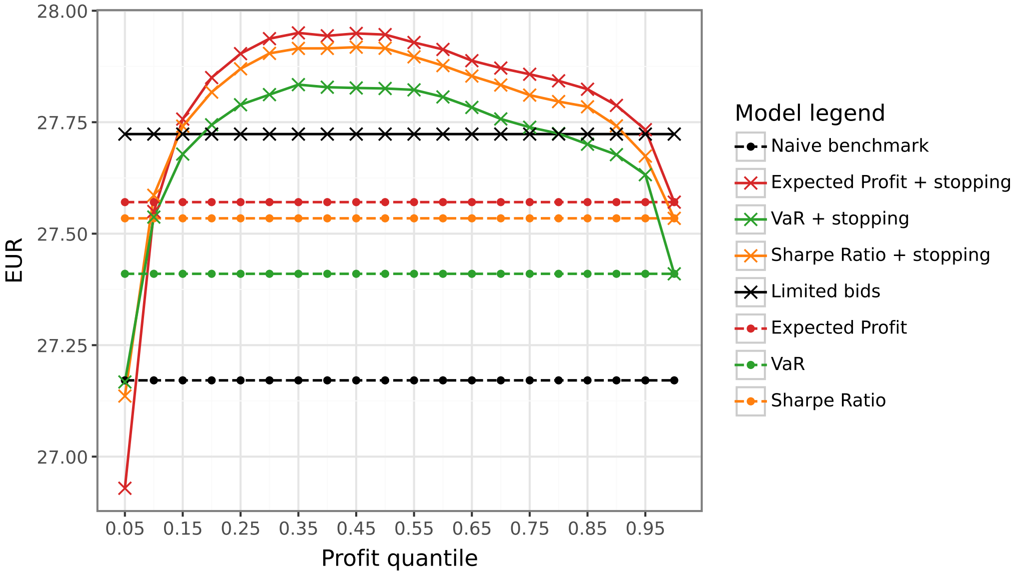


Figure 4: Average profit per 1 MWh of generation for two types of strategies: with production curtailment (marked with crosses) and without (marked with dots); depending on risk aversion level. Operation & Management costs are 10 Euro/MWh.

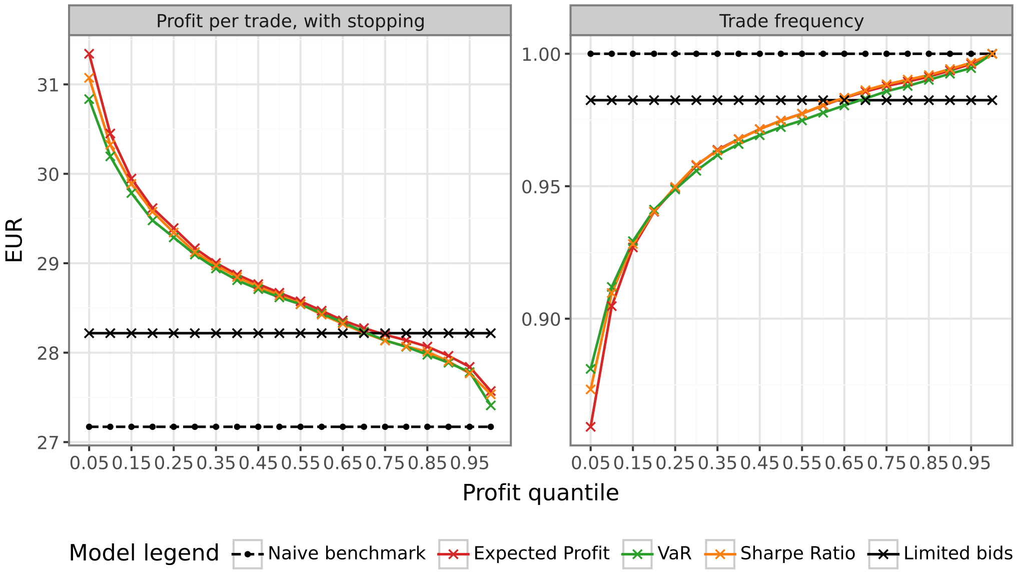


Figure 5: Performance of strategies with production curtailment: average profit per trade (left panel) and frequency of trade (right panel); Operation & Management costs are 10 Euro/MWh.

Table 5: Performance of trading strategies, relative to the naive benchmark

Strategy	Quantile of profits, τ					
	0.05	0.30	0.50	0.70	0.95	1.00
	Frequency of trade					
$E\pi$	85.88%	95.75%	97.44%	98.61%	99.57%	100%
VaR	88.05%	95.64%	97.23%	98.30%	99.45%	100%
SR	87.28%	95.80%	97.43%	98.65%	99.64%	100%
Limited bid	0.9825					
	Average profit					
$E\pi$	-0.95%	2.80%	2.87%	2.57%	2.11%	1.46%
VaR	-0.02%	2.37%	2.42%	2.22%	1.69%	0.88%
SR	-0.17%	2.71%	2.77%	2.03%	1.39%	1.39%
Limited bid	2.03%					
	Average profit per trade					
$E\pi$	15.34%	7.37%	5.58%	4.02%	2.55%	1.46%
VaR	13.54%	7.03%	5.34%	3.98%	2.25%	0.88%
SR	14.38%	7.20%	5.48%	3.87%	2.28%	1.39%
Limited bid	3.85%					

Remark: Average profit and Average profit per trade are presented as the percentage difference between the selected strategy and the naive benchmark approach.

to the $E\pi$ approach, it is relatively often recommended to leave the entire generation for the ID market. It occurs in almost 26% of the cases.

In conclusion, data-driven selection of the market (setting $q = 0$ or $q = 1$) is profitable, but exposes the utility to a higher risk than choosing a balanced portfolio. Moreover, strategies that offer a greater proportion of forecasted generation in the DA market bring lower income, but at the same time reduce the trade uncertainty.

6. Conclusions and discussion

In this article, we propose a *multiple split* method to construct probabilistic forecasts of both one- and multidimensional random variables. The approach splits the training sample into two disjoint sets: estimation and calibration. The first subset is explored to estimate the model parameters, whereas the second is used to calculate the forecast errors. Point forecasts based on calibrated parameters and prediction errors are used to construct an ensemble of forecasts. The splitting is repeated multiple times, and through merging the splits a final ensemble is constructed.

This work enhances the current body of literature on multidimensional probabilistic forecasting. The proposed approach combines and extends two methods: jackknife+ and split conformal predictions. In the MS forecasting scheme, the random division of the sample is performed multiple times to improve the accuracy of the forecast and decrease

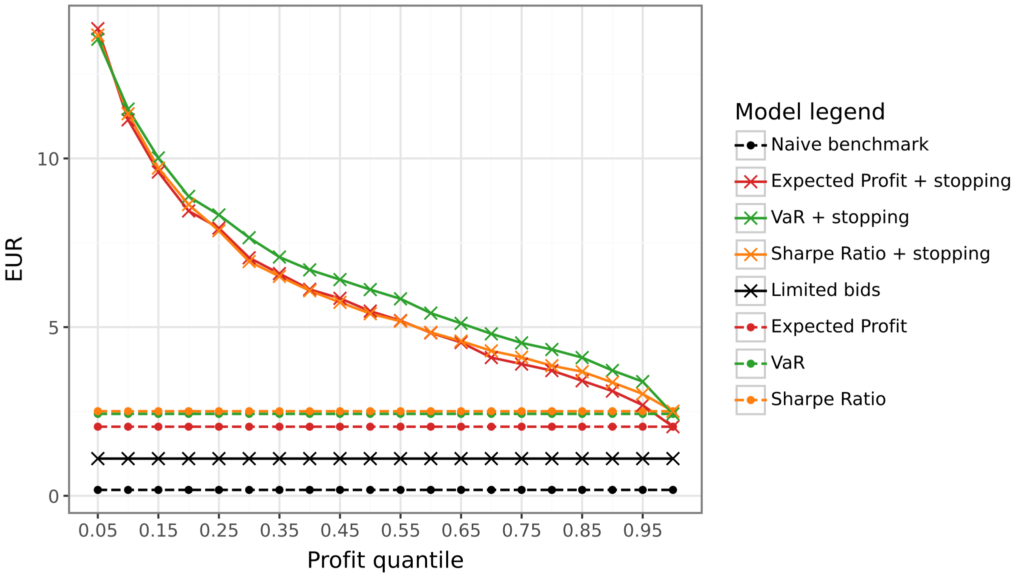


Figure 6: Value at Risk for two types of strategies: with production curtailment (marked with crosses) and without (marked with dots). Values depend on risk aversion level and are calculated only for the hours when the trade occurs. Operation & Management costs are 10 Euro/MWh.

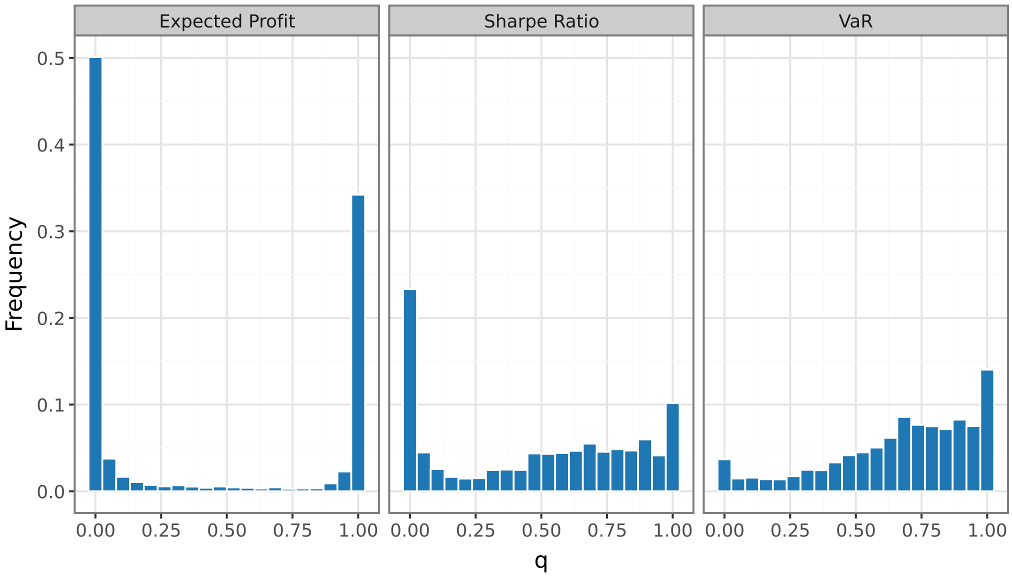


Figure 7: Histogram of optimal values of q for different data-driven strategies: $E\pi$, SR and VaR

the variability of the results. However, unlike other methods such as delete-d jackknife, jackknife+ or leave-k-out , MS uses a small number of splits: here, the number varies between 1 and 20. This significantly decreases the computational complexity. Second, forecast errors from the training window are used to construct an ensemble of forecasts rather than to estimate the distribution quantiles. Hence, to obtain the final probabilistic forecasts, it is not necessary to average quantiles or prediction intervals, as in [28]. It is sufficient to aggregate the outcomes of individual splits into a final ensemble and then use it, for example, to estimate the selected quantiles. Finally, when the split is performed simultaneously on all the variables analyzed, the obtained residuals maintain the correlation structure of the forecast errors and, therefore, are suitable for approximating the multidimensional distribution. We believe that resampling methods are of particular use in this context. They do not require parametric modeling of the multidimensional distribution and are able to approximate well the complex relationship between variables. The multivariate distribution of several variables can then be leveraged in decision-making processes by simultaneously assessing multiple correlated sources of uncertainty or computing functions of the original variables.

The accuracy of the multiple split forecasting approach is evaluated with data describing the German electricity market: DA and ID electricity prices, total load, and RES generation. It is compared with two well-known approaches: quantile regression and historical simulation. The out-of-sample period consists of 730 days, which corresponds to two years of observations. The forecast performance is evaluated with measures that focus on the fit of selected quantiles and explore the distribution of the ensemble. The results lead to the following conclusions:

- QR provides prediction intervals that are less accurate than PIs from ensemble methods (historical simulations and MS). However, the probabilistic forecasts obtained with QR are relatively sharper than those based on other methods, resulting in lower values of the CRPS measure.

- The multiple split method allows one to construct prediction intervals that have an accurate empirical coverage. The Kupiec test was unable to reject the null of a correct PI calibration in more than 90% of the analyzed cases for $T = 365$.

- When different specifications of the MS approach are compared, the model with 20 splits, MS(20), outperforms the MS(1) method.

- Ensemble methods, MS in particular, perform very well when the linear combination of fundamental variables is forecasted. In the case of Price Spread and Residual Load, a multivariate MS(20) approach provides PIs with empirical coverage very close to the nominal one. The Kupiec test confirms the correct empirical coverage in 73.33% and 92.5%, respectively.

- When the multidimensional distribution is considered, the ensemble methods that allow for the joint forecasting of the fundamentals provide the best calibration of the probabilistic predictions. Furthermore, the reliability index indicates that the MS(20), *Corr.* approach produces the most accurate fit to the multidimensional distribution.

Next, the economic value is evaluated with an example of a generation utility that owns wind farms spread throughout Germany. The company needs to make bids on the DA market. In particular, it decides on the share of predicted production, q, that is offered on the DA market. To balance the position, the remaining part of generation is either sold or purchased on the ID market. The company acts under uncertainty: it knows neither the next-day production nor the future electricity prices. In this research, we consider two benchmark strategies: a naive strategy, which assumes that all predicted generation is offered on the DA market, and a limited bid strategy, which also has $q = 1$, but allows for production curtailment when the DA prices fall below zero. These two benchmarks are compared with data-driven approaches that explore the probabilistic prediction of profits. They select the optimal value of q by maximizing the expected profit, the VaR, or Sharpe ratio. Additionally, it is assumed that the company may curtail the generation, as in the limited-bid case. The stopping rule ensures that the utility sells electricity only when a selected quantile, τ, of profits is positive. The results can be summarized as follows:

- The naive benchmark approach is outperformed by all strategies, both in terms of level of profits and risk.
- Adopting data-driven strategies without generation curtailment leads to an increase in profits of 1-1.5% and a decrease in risk.
- Strategies that allow for stoppage of production bring on average higher profits than those without any stopping rule, particularly when profit per trade is considered.
- The Limited bid strategy is dominated by data-driven methods for many values of τ. When the quantile of profits used in the stopping rule varies between 0.25 and 0.6 then the Limited bid provides a lower average profit and, at the same time, is characterized by a higher risk than any of the data-driven strategies.
- The selection among data-driven strategies depends on the approach to risk. The strategy that aims to maximize expected profits yields the highest income, but at the same time has a lower value of VaR. The opposite can be observed for the method based on VaR. As expected, the approach using the Sharpe ratio balances revenue level and risk.

The results show the potential of the MS method in forecasting a multidimensional distribution of variables describing the electricity market and designing trading strategies. They encourage further research on properties and applications that go beyond the presented analysis. First, it would be interesting to develop a rule to choose the number and proportion of splits. Here, we show the results of only two scenarios MS(1) and MS(20) in which these quantities were arbitrarily selected. A more comprehensive analysis in this area is recommended. Next, since the size of the prediction intervals does not change much over time, one could consider applying locally weighted methods to condition the length of the PI on the fluctuating market situation. This can increase the sharpness of the distribution and improve the statistical performance of the method. Finally, MS can be combined with point forecasting methods other than AR, for example machine learning. Due to a simple construction and flexibility in selecting the number of splits, it can be used together with methods that are computationally burdensome.

Author contributions

Conceptualization: K.M., W.N.; Investigation: K.M., W.N.; Software: K.M., W.N.; Supervision: K.M.; Validation, K.M., W.N. ; Writing - original draft: K.M., W.N. ; Writing - review & editing: W.N.

Acknowledgments

This work was supported by the Ministry of Education and Science (MEiN, Poland) through Diamond Grant No. 0027/DIA/2020/49 (to W.N.) and the National Science Center (NCN, Poland) through SONATA BIS grant No. 2019/34/E/HS4/00060 (to K.M).

References

[1] R. Weron. *Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach*. John Wiley & Sons, Chichester, 2006.

[2] R. 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.

[3] K. Wang, C. Xu, Y. Zhang, S. Guo, and A. Y. Zomaya. Robust Big Data Analytics for Electricity Price Forecasting in the Smart Grid. *IEEE Transactions on Big Data*, 5(1):34–45, March 2019. ISSN 2332-7790. doi: 10.1109/TBDATA.2017.2723563.

[4] K. Maciejowska, B. Uniejewski, and T. Serafin. PCA forecast averaging—predicting day-ahead and intraday electricity prices. *Energies*, 13(14):3530, 2020.

[5] F. Ziel. Forecasting electricity spot prices using LASSO: On capturing the autoregressive intraday structure. *IEEE Transactions on Power Systems*, 31(6):4977–4987, 2016.

[6] J. Lago, G. Marcjasz, B. De Schutter, and R. 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.

[7] J. Nowotarski and R. Weron. Recent advances in electricity price forecasting: A review of probabilistic forecasting. *Renewable and Sustainable Energy Reviews*, 81:1548–1568, January 2018. ISSN 1364-0321. doi: 10.1016/j.rser.2017.05.234.

[8] F. Petropoulos, D. Apiletti, V. Assimakopoulos, and *et al.* Forecasting: theory and practice. *International Journal of Forecasting*, 38(3):705–871, 2022.

[9] T. Gneiting, L. I. Stanberry, E. P. Grimit, L. Held, and N. A. Johnson. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST*, 17(2):211–235, 2008.

[10] C. Kath and F. Ziel. Conformal prediction interval estimation and applications to day-ahead and intraday power markets. *International Journal of Forecasting*, 37(2): 777–799, April 2021. ISSN 0169-2070. doi: 10.1016/j.ijforecast.2020.09.006.

[11] K. Maciejowska. Portfolio management of a small RES utility with a structural vector autoregressive model of electricity markets in Germany. *Operations Research and Decisions*, 32(4):75–90, 2022.

[12] R. W. Koenker. *Quantile Regression*. Cambridge University Press, 2005.

[13] J. Breckling and R. Chambers. M-quantiles. *Biometrika*, 75(4):761–771, 1988.

[14] B. Uniejewski and R. Weron. Regularized quantile regression averaging for probabilistic electricity price forecasting. *Energy Economics*, 95:105121, 2021. ISSN 0140-9883.

[15] K. Maciejowska and J. Nowotarski. A hybrid model for GEFCom2014 probabilistic electricity price forecasting. *International Journal of Forecasting*, 32(3):1051–1056, 2016.

[16] P. Gaillard, Y. Goude, and R. Nedellec. Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. *International Journal of Forecasting*, 32(3):1038–1050, 2016.

[17] P. Pinson. Wind energy: Forecasting challenges for its operational management. *Statistical Science*, 28(4):564–585, 2013.

[18] D. Lee, H. Shin, and R. Baldick. Bivariate probabilistic wind power and real-time price forecasting and their applications to wind power bidding strategy development. *IEEE Transactions on Power Systems*, 33(6):6087–6097, 2018. ISSN 0885-8950. doi: 10.1109/TPWRS.2018.2830785.

[19] J.-F. Toubeau, T.-H. Nguyen, H. Khaloie, Y. Wang, and F. Vallée. Forecast-Driven Stochastic Scheduling of a Virtual Power Plant in Energy and Reserve Markets. *IEEE Systems Journal*, 16(4):5212–5223, December 2022. ISSN 1937-9234. doi: 10.1109/JSYST.2021.3114445.

[20] N. Kumbartzky, M. Schacht, K. Schulz, and B. Werners. Optimal operation of a CHP plant participating in the German electricity balancing and day-ahead spot market. *European Journal of Operational Research*, 261(1):390–404, August 2017. ISSN 0377-2217. doi: 10.1016/j.ejor.2017.02.006.

[21] J. Janczura and E. Wójcik. Dynamic short-term risk management strategies for the choice of electricity market based on probabilistic forecasts of profit and risk measures. the german and the polish market case study. *Energy Economics*, 110:106015, 2022.

[22] H. Lütkepohl. *New Introduction to Multiple Time Series Analysis.* Springer-Verlag, Berlin, 2005.

[23] F. Durante and C. Sempi. *Principles of copula theory.* CRC press, 2015.

[24] T. Gneiting and M. Katzfuss. Probabilistic forecasting. *The Annual Review of Statistics and Its Application*, 1:125–151, 2014.

[25] F. Durante, A. Gianfreda, F. Ravazzolo, and L. Rossini. A multivariate dependence analysis for electricity prices, demand and renewable energy sources. *Information Sciences*, 590:74–89, 2022.

[26] H. Manner, F. A. Fard, A. Pourkhanali, and L. Tafakori. Forecasting the joint distribution of Australian electricity prices using dynamic vine copulae. *Energy Economics*, 78:143–164, 2019.

[27] K. Ignatieva and S. Trück. Modeling spot price dependence in Australian electricity markets with applications to risk management. *Computers and Operations Research*, 66:415–433, 2016.

[28] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018.

[29] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486 – 507, 2021. doi: 10.1214/20-AOS1965. URL `https://doi.org/10.1214/20-AOS1965`.

[30] B. Efron and G. Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1):36–48, 1983.

[31] V. Vovk, I. Nouretdinov, V. Manokhin, and A. Gammerman. Cross-conformal predictive distributions. In A. Gammerman, V. Vovk, Z. Luo, E. Smirnov, and R. Peeters, editors, *Proceedings of the Seventh Workshop on Conformal and Probabilistic Prediction and Applications*, volume 91 of *Proceedings of Machine Learning Research*, pages 37–51. PMLR, 11–13 Jun 2018.

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

[33] K. Maciejowska, W. Nitka, and T. Weron. Enhancing load, wind and solar generation for day-ahead forecasting of electricity prices. *Energy Economics*, 99:105273, July 2021. ISSN 0140-9883. doi: 10.1016/j.eneco.2021.105273.

[34] J. Shao and C. F. J. Wu. A General Theory for Jackknife Variance Estimation. *The Annals of Statistics*, 17(3), September 1989. ISSN 0090-5364. doi: 10.1214/aos/1176347263.

[35] K. C. Lichtendahl, Y. Grushka-Cockayne, and R. L. Winkler. Is it better to average probabilities or quantiles? *Management Science*, 59(7):1594–1611, 2013.

[36] J. Berrisch, F., and Ziel. Crps learning. *Journal of Econometrics*, (105221), 2021. doi: 10.1016/j.jeconom.2021.11.008.

[37] W. Nitka and R. Weron. Combining predictive distributions of electricity prices.Does minimizing the CRPS lead to optimal decisions in day-ahead bidding? *Operations Research and Decisions*, 33(3):105–118, 2023. doi: DOI:10.37190/ord230307.

[38] P. H. Kupiec. Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives*, 3(2):73–84, 1995.

[39] T. Gneiting, F. Balabdaoui, and A. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society B*, 69:243–268, 2007.

[40] W. F. Sharpe. The Sharpe ratio. *The Journal of Portfolio Management*, 21(1):49–58, 1994.

[41] Z. Ren, A. S. Verma, Y. Li, J. J. Teuwen, and Z. Jiang. Offshore wind turbine operations and maintenance: A state-of-the-art review. *Renewable and Sustainable Energy Reviews*, 144:110886, 2021.

[42] A. R. Nejad, J. Keller, Y. Guo, S. Sheng, H. Polinder, S. Watson, J. Dong, Z. Qin, A. Ebrahimi, R. Schelenz, F. Gutiérrez Guzmán, D. Cornel, R. Golafshan, G. Jacobs, B. Blockmans, J. Bosmans, B. Pluymers, J. Carroll, S. Koukoura, E. Hart, A. McDonald, A. Natarajan, J. Torsvik, F. K. Moghadam, P.-J. Daems, T. Verstraeten, C. Peeters, and J. Helsen. Wind turbine drivetrains: state-of-the-art technologies and future development trends. *Wind Energy Science*, 7(1):387–411, 2022.

[43] A. M. Costa, J. A. Orosa, D. Vergara, and P. Fernández-Arias. New tendencies in wind energy operation and maintenance. *Applied Sciences*, 11(4), 2021.

[44] R. Wiser, M. Bolinger, and E. Lantz. Assessing wind power operating costs in the united states: Results from a survey of wind industry experts. *Renewable Energy Focus*, 30: 46–57, 2019.

Paper 4

Enhancing load, wind and solar generation for day-ahead forecasting of electricity prices

Katarzyna Maciejowska, Weronika Nitka, Tomasz Weron

Energy Economics 99 (2021) 105273

Contents lists available at ScienceDirect

Energy Economics

journal homepage: www.elsevier.com/locate/eneeco

Enhancing load, wind and solar generation for day-ahead forecasting of electricity prices

Katarzyna Maciejowska [a,*], Weronika Nitka [a], Tomasz Weron [b]

[a] Department of Operations Research and Business Intelligence, Faculty of Computer Science and Management, Wrocław University of Science and Technology, 50-370 Wrocław, Poland
[b] Faculty of Pure and Applied Mathematics, Wrocław University of Science and Technology, 50-370 Wrocław, Poland

ARTICLE INFO

Article history:
Received 9 December 2019
Received in revised form 3 April 2021
Accepted 5 April 2021
Available online 21 April 2021

Keywords:
Renewables
Electricity prices
Day-ahead market
Intraday market
Forecasting

ABSTRACT

In recent years, a rapid development of renewable energy sources (RES) has been observed across the world. Intermittent energy sources, which depend strongly on weather conditions, induce additional uncertainty to the system and impact the level and variability of electricity prices. Predictions of RES, together with the level of demand, have been recognized as one of the most important determinants of future electricity prices. In this research, it is shown that forecasts of these fundamental variables, which are published by Transmission System Operators (TSO), are biased and could be improved with simple regression models. Enhanced predictions are next used for forecasting of spot and intraday prices in Germany. The results indicate that improving the forecasts of fundamentals leads to more accurate predictions of both, the spot and the intraday prices. Finally, it is demonstrated that utilization of enhanced forecasts is helpful in a day-ahead choice of a market (spot or intraday), and results in a substantial increase of revenues.

1. Introduction

In the recent decades, electricity markets across the world have undergone reforms, which have resulted in a deep market deregulation. Electricity power exchanges have been created, such as Nord Pool or EEX in Europe, PJM in the USA and NEM in Australia, which allow for a competitive electricity trade. Nowadays, a large share of the transactions is done in the day-ahead markets, where offers are placed around noon on the day preceding the delivery. The day-ahead prices, which clear the markets, are often called 'spot prices'. In order to allow for an adoption of trading positions to unplanned events, spot markets have been complemented by intraday and balancing markets. The intraday markets, typically organized by power exchanges, take the form of auctions (e.g., in Spain) or continuous trading (e.g., in Germany), and allow to trade the electricity throughout the whole day, up to a few minutes before the physical delivery. The final balancing of the demand and the supply is achieved via the balancing markets, which are controlled by the Transmission System Operators (TSO) and aim at securing the system stability. A more detailed explanations of the European electricity markets can be found in Gianfreda et al. (2016) and Koch and Hirth (2019). It is worth noting that trading in the day-ahead or intraday markets is usually not mandatory, neither for generators nor for consumption units.

The market participants are now facing new trade opportunities and can, to some extent, choose between different markets and contract types. In particular, RES utilities do not know their exact generation and therefore, are encouraged to self-balance their position in the intraday market (Pape et al., 2016; Kiesel and Paraschiv, 2017; Gianfreda et al., 2016). As the result, managers can offer the majority of their predicted generation in the day-ahead market and leave a part for flexible trade in order to manage the risk and revenue. Finally, it should be underlined here, that the core business of RES utilities is electricity generation and therefore, it focuses on a real trade rather than speculation.

The literature indicates that the choice of the trading strategy could result in a profit increase (Maciejowska et al., 2019) or risk reduction (Kath and Ziel, 2018). In order to support the decision process, accurate day-ahead predictions of spot and intraday prices are needed. The literature is rich in publications focusing on modelling and forecasting of spot prices (see Weron, 2014; Nowotarski and Weron, 2018, for a comprehensive review). Many papers indicate that the predicted RES generation and electricity demand are one of the main drivers of the day-ahead prices (Paraschiv et al., 2014; Woo et al., 2016; Gürtler and Paulsen, 2018; Pape et al., 2016) and hence should be included in the modelling scheme (Uniejewski and Weron, 2018; Ziel and Steinert, 2018; Gianfreda et al., 2020).

At the same time, not much attention has been placed on modelling intraday markets. There are a few articles which analyze the intraday markets in Europe (Kath and Ziel, 2018; Kiesel and Paraschiv, 2017; Monteiro et al., 2016) and the US (Woo et al., 2016). Most of them focus on a very short term – a few hours ahead – forecast, as in

[*] Corresponding author.
E-mail addresses: katarzyna.maciejowska@pwr.edu.pl (K. Maciejowska), weronika.nitka@pwr.edu.pl (W. Nitka), tomasz.weron@pwr.edu.pl (T. Weron).

https://doi.org/10.1016/j.eneco.2021.105273
0140-9883/© 2021 Published by Elsevier B.V.

Uniejewski et al. (2019b) and hence, assume the knowledge of spot prices. This type of models could not be directly used by utilities when making operational decisions, such as a choice between the spot and intraday market (see Maciejowska et al., 2019). In such case, new models of intraday prices, which only use the information available at the time of the decision, need to be developed.

This article extends the literature in various directions. First, it shows that the TSO forecasts of total load, wind and solar generations, which are crucial for electricity markets, may be systematically biased and could be improved with autoregressive types of models. Although the literature considers TSO predictions as the most efficient ones, practitioners seek more accurate ways of forecasting the demand and the generation structure. This article follows this idea and corrects the TSO predictions with information available day-ahead, ensuring that the outcomes could be used by utilities while placing final orders in the spot market.

Second, it is investigated if the corrected forecasts could enhance the predictions of the spot and intraday prices. Here, a novel approach for forecasting of intraday electricity prices is adopted, which explores the difference between the enhanced and TSO predictions. One could see this model as a day-ahead counterpart of an approach used by Kiesel and Paraschiv (2017). The results indicate that predictions of both day-ahead and intraday prices could be significantly improved with the use of enhanced fundamental variables forecasts.

Finally, the possibility of price spread forecast is examined. While a related topic of forecasting load imbalance volumes has gained some interest from researchers (Lisi and Edoli, 2018; Bunn et al., 2018), this issue, although of a great practical importance, has not been studied much in the literature. The results confirm previous findings of Maciejowska et al. (2019) and show that the forecasted sign of the difference between intraday and spot prices could be used in the decision process and may lead to an increase of utility revenue. Moreover, usage of the enhanced predictions of fundamentals in a decision process substantially raises the additional revenues.

The article is structured as follows. First, in Section 2, we present and discuss the data. Then, in Section 3, we introduce and describe the models. Next, in Section 4, we show the results and finally, in Section 5, we conclude the study.

2. Data

This article analyzes the German electricity market, which is known for its high RES penetration. In the first two quarters of the year 2019, the RES share in the total electricity production exceeded 47% (see https://www.energy-charts.de). The data used in this research is hourly and spans the period from 1 October 2015 to 30 September 2019. The sample is divided into four years. The first year, 1 October 2015–30 September 2016, is utilized for calibrating models used for forecasting fundamental variables. In the second one, 1 October 2016–30 September 2017, the predictions of the fundamentals are collected, evaluated and next used as an input to price models. Finally, in the last two years, 1 October 2017–30 September 2019, the performance of price forecasts is assessed and financial gains from the proposed approach are computed. The notation and sample division are summarized in Table 1.

Table 1
Sample division and notation.

Notation	Start date	End date
2015	1 October 2015	30 September 2016
2016	1 October 2016	30 September 2017
2017	1 October 2017	30 September 2018
2018	1 October 2018	30 September 2019

The data set comprises day-ahead ($DA_{h,t}$) and intraday ($ID_{h,t}$) market prices for corresponding bidding zones: Austria + Germany + Luxembourg before 1 October 2018, Germany + Luxembourg after 1 October 2018. The intraday prices used in this research are ID3 indexes (volume weighted prices from the last 3 hours of trade). They are complemented by actual levels and system forecasts of fundamental variables: the total load ($L_{h,t}$), which can be treated as a proxy for the demand, and the RES (wind – $W_{h,t}$ and solar – $S_{h,t}$) generation. Fundamental variables are collected for Germany. They are supplemented by the forecasted temperatures for two German cities: Hamburg and Munich ($FT_{h,t}$). In the remaining part of the paper, the index h stands for an hour and t for a day number. Data sources, units and notation are summarized in Table 2.

The time paths of the day-ahead and intraday prices are presented in Fig. 1. For illustrative purposes two hours, $h = 4, 18$, have been chosen, representing the peak and the off-peak periods of a day. It can be noticed that peak prices are higher than off-peak ones, both for the day-ahead and intraday markets. Their variability changes in time and exhibits a tendency for clustering. Finally, prices in different markets co-move together. The occurrence of positive and negative spikes is synchronized in both markets, but their magnitude is more pronounced for the intraday prices.

When the fundamental variables are considered, it could be observed that load, wind and solar generations have different statistical properties. Daily averages of the variables together with their TSO forecast errors are presented in Fig. 2. The plots indicate that load depends strongly on a day of the week and follows a yearly seasonality. In Germany, the electricity consumption is the highest during the winter, when the energy is used for heating, and falls when the temperature increases. Additionally, one could observe a slight increase in the demand during summer months, when air-conditioning is used. Unlike the load, the RES generation does not exhibit a weekly pattern because it does not depend on the electricity consumption. The wind generation rises in the winter and drops slightly in the summer. Moreover, wind shows a lot of variation and can change drastically within a few days. At the same time, solar generation has an opposite yearly pattern with a peak in summertime. It falls almost to zero during winter, when days are short and the sunlight is insufficient.

In Fig. 2, fundamental variables are accompanied by their forecast errors, computed as the difference between their actual values and corresponding TSO predictions. Unlike fundamentals, forecast errors follow neither weekly nor yearly seasonality. Only solar errors seem to be larger in spring, when the PV (photovoltaic) generation increases. The basic statistical properties of TSO forecast errors are presented in Table 3, which includes information on their means and mean standard deviations across peak (9–20) and off-peak (1–8, 21–24) hours. It is clearly visible that there is a substantial bias in the TSO forecasts for all three fundamental variables. The biggest bias is observed for load,

Table 2
Data sources and units.

Data	Notation	Units	Source
Day-ahead prices	*DA*	EUR/MWh	EPEX SPOT, http://www.epexspot.com
Intraday prices	*ID*	EUR/MWh	EPEX SPOT, http://www.epexspot.com
Load	*L*	GWh	https://transparency.entsoe.eu
Wind generation	*W*	GWh	https://transparency.entsoe.eu
PV generation	*S*	GWh	https://transparency.entsoe.eu
Forecasted load	*FL*	GWh	https://transparency.entsoe.eu
Forecasted wind generation	*FW*	GWh	https://transparency.entsoe.eu
Forecasted PV generation	*FS*	GWh	https://transparency.entsoe.eu
Forecasted temperature	*FT*	°C	https://api.meteo.pl

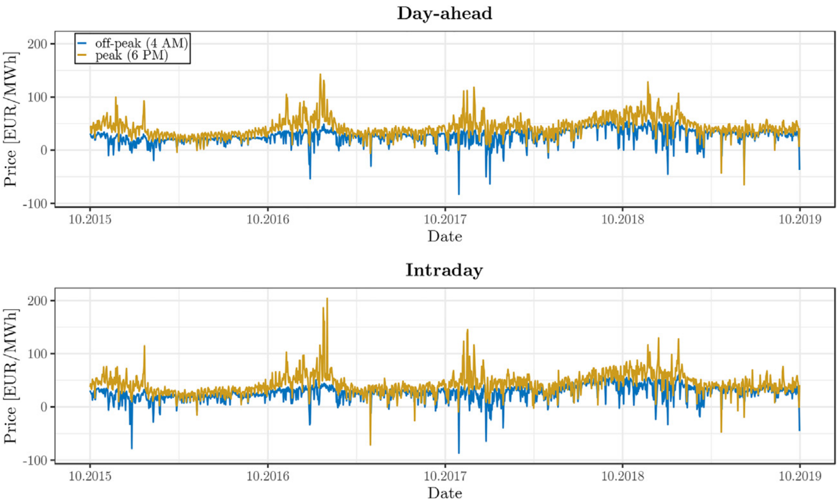


Fig. 1. Time plots of the day-ahead and intraday prices for two illustrative hours (4 a.m. and 6 p.m.).

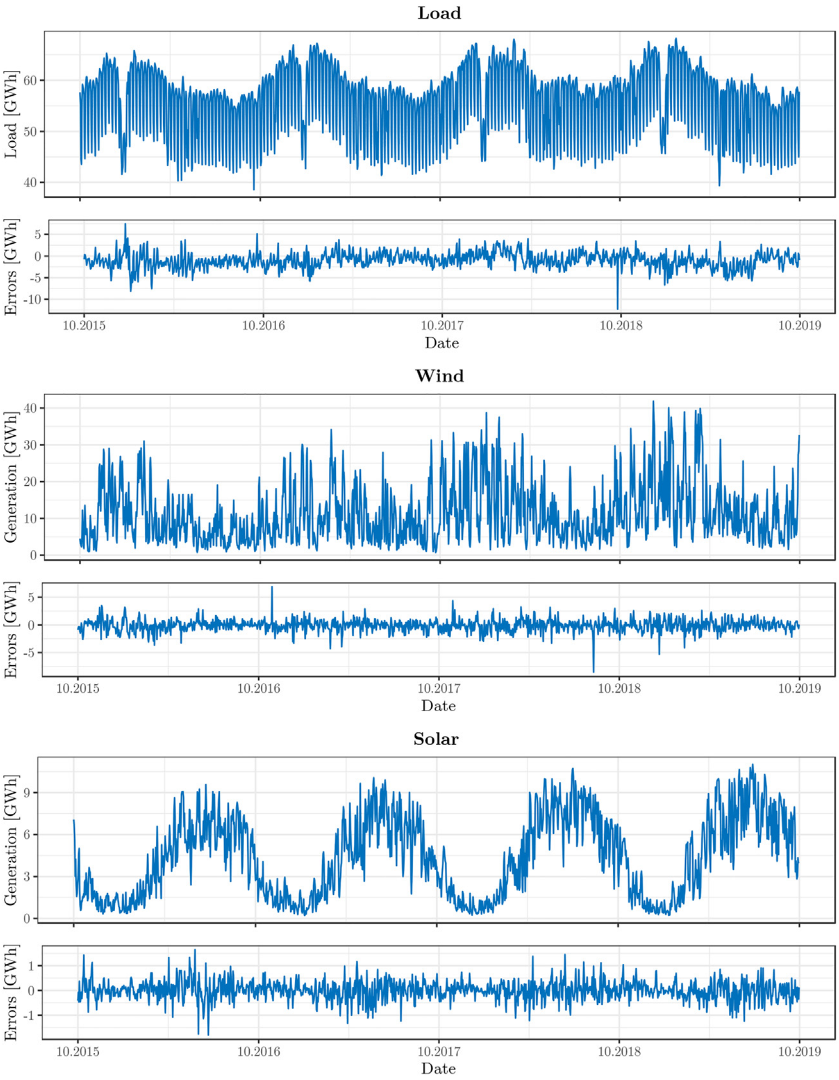


Fig. 2. Time plots of daily averages of fundamental variables and the TSO forecasts errors.

Table 3
Descriptive statistics of TSO forecast errors.

Statistics	Peak			Off-peak		
	Load	Wind	Solar	Load	Wind	Solar
Mean	−1.022 (0.017)	−0.041 (0.012)	0.009 (0.008)	−0.915 (0.014)	−0.139 (0.012)	0.002 (0.001)
LB test	12	11	12	12	12	5

Note: standard deviations of the mean estimators are stated in brackets; Ljung-Box (LB) test reports a number of hours from the twelve-hour-long peak/off-peak blocks, where the null is rejected at 5% significance level.

which, with mean error of −1.022 GWh during peak hours and −0.915 GWh in off-peak hours, is systematically underestimated. The standard deviations of the mean of load, wind and solar generation are presented in brackets. They indicate that biases are significantly different from zero for both peak and off-peak hours, and all the fundamental variables. Next, the autocorrelation of residuals is verified with the Ljung-Box (LB) test statistic, for each hour separately. The last row of Table 3 shows the number of hours from the twelve-hour-long peak/off-peak blocks, where the LB test rejects the null hypothesis at 5% significance level. The results indicate a strong autocorrelation of forecast errors. It can be observed that only for seven hours of solar generation and one hour of wind generation the null hypothesis cannot be rejected. Most of these hours coincide with times of very low or no photovoltaic generation. To sum up, the initial analysis suggests that TSO forecasts are not only systematically underestimated but also could be improved with autoregressive type of models, which explore the autocorrelation structure of TSO forecast errors.

3. Methods

3.1. Problem setup

This research consists of three major parts: (i) a calibration of ARX-type models of load, wind and solar generation, and a calculation of their day-ahead forecasts, (ii) an assessment of prediction accuracy of fundamental variables, (iii) an evaluation of financial gains resulting from enhancement of fundamental forecasts. The gains are measured in two ways. First, it is verified whether more accurate predictions of load, wind and solar improve electricity price forecasts. Second, additional revenues from the price-driven choice of the trading strategy are calculated and compared with those based on TSO information only.

Like the majority of energy studies, we consider a rolling window scheme with a limited memory and model separately each hour of a day. The algorithm consists of three steps. First, the model parameters are estimated using the data from a calibration window of a fixed length. Next, one day ahead forecasts of fundamental variables or electricity prices are calculated. Finally, the window is moved one step ahead. All the steps are repeated until the forecasts of the last observation in the evaluation window are computed.

Two forecasting methods are used, conditional on the modeled variable. In case of fundamentals, we follow recent papers of Hubicka et al. (2019) and Marcjasz et al. (2018), which show that averaging across different calibration windows yields better results than selecting ex-ante a single window length. Here, we combine forecasts based on three short and three long estimation windows. A similar choice of window lengths was previously used by Marcjasz et al. (2018) and Serafin et al. (2019). When we consider the electricity prices, the models are calibrated with one year of observations. We restrict the analysis to a single window length in order to capture the direct impact of enhanced fundamental forecasts on the price predictions.

Finally, it should be emphasised here that the article focuses on day-ahead forecasts because the market participants need to place their orders in the morning of the day preceding the delivery. Moreover, the forecasts are calculated before the decisions are made, which is assumed to happen at 11 am. As a consequence, any new information arriving after 10 am, for example an actual level of generation and its structure or the level of intraday prices, are excluded from the information set and not used for fundamental variables or price predictions.

3.2. Forecasting fundamentals

In order to forecast fundamental variables, ARX type of models are adopted, which utilize both the information on system forecasts and actual past realizations of these variables. In this research, three different model specifications are adopted. First, the total load is modeled as follows

$$L_{t,h}= \alpha_h^L D_t^L + \underbrace{\theta_{h,1}^L L_{t-1,h}^{*} + \sum_{p\in\{2,7\}} \theta_{h,p}^L L_{t-p,h}}_{\text{AR component}} + \underbrace{\beta_{h,1}^L FL_{t,h} + \beta_{h,2}^L FW_{t,h} + \beta_{h,3}^L FS_{t,h}}_{\text{Forecasts off undamentals}}$$
$$+\underbrace{\beta_{h,4}^L FL_{t,ave} + \beta_{h,5}^L FL_{t,max} + \beta_{h,6}^L FL_{t,min}}_{\text{Daily statistics}} + \underbrace{\beta_{h,7}^L FT_{t,h}}_{\text{Weather forecasts}} + \varepsilon_{t,h}^L, \quad (1)$$

where D_t^L is a (4×1) vector of deterministic variables consisting of a constant and three dummy variables for Mondays, Saturdays and Sundays/Holidays. $FL_{t,h}$, $FW_{t,h}$ and $FS_{t,h}$ are the TSO forecasts of all three fundamental variables for the current day and hour, as defined in Table 2. $FL_{t,ave}$, $FL_{t,max}$ and $FL_{t,min}$ are daily statistics computed as the mean, maximum and minimum of the TSO load forecast for the day t over 24 hours. The weather forecast vector, $FT_{t,h}$, includes predicted temperature for two cities: Hamburg and Munich). In the AR part, lags $p \in \{1,2,7\}$ are chosen, which corresponds to lags used in price forecasting studies (see Nowotarski et al., 2014; Uniejewski et al., 2016, 2019a; Ziel, 2016 among others). This lag structure captures both the short run dependence and the weekly seasonality. It should be mentioned here that for some hours, $h > 10$, there is no information on the actual generation available at the time of forecasts. Therefore we define a variable:

$$L_{t,h}^{*} = \begin{cases} L_{t,h} & \text{if } h \le 10, \\ FL_{t,h} & \text{if } h > 10, \end{cases} \quad (2)$$

which replaces the missing observations with their TSO forecasts.

The model for wind generation is simpler than (1) and is given by

$$W_{t,h} = \alpha_h^W D_t^W + \theta_h^W W_{t-1,h}^{*} + \beta_{h,1}^W FW_{t,h} + \beta_{h,2}^W FW_{t,h-1} + \beta_{h,3}^W FW_{t,h+1} + \varepsilon_{t,h}^W. \quad (3)$$

The deterministic variable, D_t^W, includes only a constant because wind does not follow a weekly seasonality. The variable $W_{t-1,h}^{*}$ controls for missing information and is defined as (2). Finally, Eq. (3) includes also information on predicted wind generation in two neighbouring hours: $h-1$ and $h+1$ (when such information is available). It is assumed that wind generation does not depend on the other fundamental variables: load or solar, and its AR structure consists of only one lag.

A similar model is adopted for solar generation. It is described by the following equation

$$S_{t,h} = \alpha_h^S D_t^S + \theta_h^S S_{t-1,h}^{*} + \beta_{h,1}^S FS_{t,h} + \beta_{h,2}^W FS_{t,h-1} + \beta_{h,3}^W FS_{t,h+1} + \varepsilon_{t,h}^S, \quad (4)$$

where D_t^S consists of an intercept and the number of sun hours within a day, which approximates the yearly seasonality. Analogously to (3), we use a simple autoregressive structure with one lag.

The models (1), (3) and (4) are estimated for different lengths, τ, of calibration windows. Following Marcjasz et al. (2018); Serafin et al. (2019), we choose three short windows with $\tau \in \{56,84,112\}$, which correspond to 8, 12 and 16 weeks of observations, and three long windows with $\tau \in \{351,358,365\}$, which balance the short term effect. The predictions are next computed as a simple average over individual forecasts.

It should be noticed here that the model (4) can be estimated only for hours in which at least some TSO forecasts are different from zero for each of the calibration windows. The proportion of nonzero TSO solar predictions ($FS_{t,h}$) in consecutive calibration windows (56-day long) are presented in Fig. 3, Appendix. The plots indicate that only hours 8–17 satisfy the nonzero condition and therefore the prediction of solar generation and its usage in the analysis is limited to hours 8–17.

3.3. Forecasting electricity prices

In order to compute the day-ahead forecasts of electricity prices, we use autoregressive models with exogenous variables. The $DA_{t,h}$ price of the day t and hour h is given by

$$\begin{aligned} DA_{t,h} = \alpha_h D_t + \underbrace{\sum_{p\in\{1,2,7\}} \theta_{h,p} DA_{t-p,h}}_{\text{AR component}} \\ + \underbrace{\beta_{h,4} DA_{t-1,ave} + \beta_{h,4} DA_{t-1,min} + \beta_{h,5} DA_{t-1,max}}_{\text{Daily quantities}} \\ + \underbrace{\beta_{h,6} DA_{t-1,24}}_{\text{Last known price}} + \underbrace{\theta_h \widehat{X}_{t,h}}_{\text{Fundamentals}} + \varepsilon_{t,h}, \end{aligned} \tag{5}$$

where $DA_{t-1,ave}$, $DA_{t-1,min}$ and $DA_{t-1,max}$ are the average, the minimum and the maximum of prices from the preceding day, $DA_{t-1,24}$ is the last known price and D_t is a (4×1) vector of deterministic variables: a constant and dummies for Mondays, Saturdays and Sundays/Holidays. Finally, $\widehat{X}_{t,h} = \left(\widehat{L}_{t,h}, \widehat{W}_{t,h}, \widehat{S}_{t,h}\right)'$ is a vector of forecasts of fundamental variables, which are based either on TSO predictions (then $\widehat{X}_{t,h} = (FL_{t,h}, FW_{t,h}, FS_{t,h})'$) or results from models described in the previous section. Note that the predictions of solar generation are included in the model only for hours $8-17$, when the solar radiation is substantial.

In this research, two different models of intraday prices are considered. The first model is based on the approach adopted for forecasting of day-ahead prices and takes the following form:

$$\begin{aligned} ID_{t,h} = \alpha_h D_t + \theta_{h,1} ID^*_{t-1,h} + \underbrace{\sum_{p\in\{2,7\}} \theta_{h,p} ID_{t-p,h}}_{\text{AR component}} \\ + \underbrace{\beta_{h,4} DA_{t-1,ave} + \beta_{h,4} DA_{t-1,min} + \beta_{h,5} DA_{t-1,max}}_{\text{Daily day-ahead quantities}} \\ + \underbrace{\beta_{h,6} DA_{t-1,24}}_{\text{Last known price}} + \underbrace{\theta_h \widehat{X}_{t,h}}_{\text{Fundamentals}} + \varepsilon_{t,h}. \end{aligned} \tag{6}$$

It should be noticed that for hours $h > 10$, the intraday prices – $ID_{t,h}$ – are not known. In such case, they are replaced by their day-ahead counterparts. Hence

$$ID^*_{t,h} = \begin{cases} ID_{t,h} & \text{if } h \leq 10, \\ DA_{t,h} & \text{if } h > 10. \end{cases} \tag{7}$$

Moreover, due to the lack of sufficient information, we can use neither average, minimum, maximum nor last intraday price from the day $t-1$. Therefore the model (6) uses the corresponding quantities from the day-ahead market.

The second model is similar to the approach proposed by Kiesel and Paraschiv (2017), in which the intraday prices are conditioned on the day-ahead prices and fundamental forecast errors. In this research, the model includes additionally lagged prices and the current predictions of fundamentals. Since the forecast errors of fundamentals are not known, they are approximated by the difference between the model based and TSO forecasts. The final form of the model is given by the following equation

$$\begin{aligned} ID_{t,h} = \alpha_h D_t + \beta_{h,1} \widehat{DA}_{t,h} + \beta_{h,2} ID^*_{t-1,h} + \theta_{h,1} \widehat{X}_{t,h} \\ + \underbrace{\theta_{h,2}\left(\widehat{X}_{t,h} - FX_{t,h}\right) + \theta_{h,3}\left(X^*_{t-1,h} - FX_{t-1,h}\right)}_{\text{forecast errors of fundamentals}} + \varepsilon_{t,h}, \end{aligned} \tag{8}$$

where $FX_{t,h} = (FL_{t,h}, FW_{t,h}, FS_{t,h})'$ is a (3×1) vector of summarizing TSO forecasts and the difference $\left(\widehat{X}_{t,h} - FX_{t,h}\right)$ approximates the forecast error of fundamental variables. The variable $X^*_{t,h}$ is defined similar to (2), with $X^*_{t,h} = \widehat{X}_{t,h}$ for $h > 10$. Notice that when the model (8) is estimated using only the information provided by TSO, then the forecast error, $\left(\widehat{X}_{t,h} - FX_{t,h}\right)$, equals to zero and hence it needs to be removed from the equation. Finally, while calculating $ID_{t,h}$, there is no information on $DA_{t,h}$ available. Therefore, instead of actual level of day-ahead prices, the model utilizes their forecasts, $\widehat{DA}_{t,h}$, obtained with model (5).

3.4. Forecasting the sign of the price spread

When modelling the market choice, we follow the methodology established in Maciejowska et al. (2019). A binary decision variable $Y_{t,h}$ is defined, which equals to one when the generator decides to sell the electricity produced for day t hour h in the intraday market, and zero otherwise. Following Maciejowska et al. (2019), a benchmark, called a naïve day-ahead strategy is considered. It assumes that all generated electricity is sold in the day-ahead market and hence $Y_{t,h} = 0$ for all t and h. This benchmark strategy is compared with a data-driven approach, which links the decision to the relationship between the day-ahead and the intraday price:

$$Y_{t,h} = \begin{cases} 1 & \text{if } ID_{t,h} - DA_{t,h} > 0, \\ 0 & \text{if } ID_{t,h} - DA_{t,h} \leq 0. \end{cases} \tag{9}$$

As the utility has to make its decision before the actual price difference $\Delta P_{t,h} = ID_{t,h} - DA_{t,h}$ is known, it has to be based on the forecasted spread $\Delta\widehat{P}_{t,h} = \widehat{ID}_{t,h} - \widehat{DA}_{t,h}$. Hence, the generator sells electricity in the intraday market, if the predicted spread is positive and in the day-ahead market otherwise.

4. Results

In this research, three types of results are analyzed. First, the possibility of an improvement of fundamental predictions over their TSO forecasts is considered. The outcomes are compared using MAE and RMSE forecast accuracy measures. Second, gains from enhancement of fundamentals predictions in forecasting of electricity price is analyzed. Finally, improved price forecasts are used in a decision process. The resulting revenues are calculated and compared with the TSO based strategies.

4.1. Enhancing the forecasts of fundamentals

The forecasts computed with models (1), (3) and (4) are compared with those published by TSO using the data from the last three years: 2016–2018. The results are presented in Table 4, which shows MAE and RMSE for the three considered fundamentals and three analyzed years. The significance of the forecast accuracy change is statistically verified with the Diebold-Mariano (DM) test (Diebold and Mariano, 1995). In order to compare the model performance across all 24 hours, we follow a vectorized DM approach described by Ziel and

Table 4
Forecast accuracy of fundamental variables.

Variable	Load			Wind			Solar		
Year	2016	2017	2018	2016	2017	2018	2016	2017	2018
MAE									
TSO	1.529	1.538	1.940	0.999	1.180	1.160	0.740	0.679	0.718
Enhanced	1.045	1.125	1.204	1.013	1.166	1.147	0.740	0.677	0.707
(p-val)	(0.000)	(0.000)	(0.000)	(0.017)	(0.795)	(0.249)	(0.383)	(0.352)	(0.832)
RMSE									
TSO	1.961	2.024	2.475	1.538	1.765	1.580	1.077	1.003	1.014
Enhanced	1.496	1.631	1.624	1.537	1.746	1.549	1.076	0.999	1.008
(p-val)	(0.000)	(0.000)	(0.000)	(0.960)	(0.978)	(0.929)	(0.789)	(0.292)	(0.852)

Note: the difference of the forecast accuracy is tested with Diebold-Mariano test with autocorrelation of order 7 and 2 for load and RES variables, respectively; *p*-values of the DM test are presented in brackets.

Weron (2018). We apply the DM test to the multivariate loss differential series between compared models *X*, *Y* defined as

$$\Delta_{X,Y,d} = \|\widehat{\varepsilon}_{X,d}\| - \|\widehat{\varepsilon}_{Y,d}\|, \qquad (10)$$

where $\widehat{\varepsilon}_{X,d}$ and $\widehat{\varepsilon}_{Y,d}$ are the 24-dimensional vectors of out-of-sample errors for models *X* and *Y* respectively. The norm utilized for calculation of the loss differential depends on the forecast accuracy measure. We use Euclidean, $\|.\|_2$, norm for RMSE and $\|.\|_1$ for MAE. Since not all market information is available when calculating the predictions, we allow for the autocorrelation of forecast errors while computing DM test statistics.

The outcomes indicate that load and wind forecasts could be significantly improved with ARX types of models. The results show that MAE of load is reduced by 31.6%, 26.8% and 37.9% in years 2016 to 2018, respectively. Also RMSE measure decreases considerably and falls by more than 20% in all years. This indicates that TSO load forecasts are strongly biased and could be substantially improved by the employment of statistical models. When wind forecast is considered, it should be noticed that gains from prediction enhancement are much lower than in load case. The MAE of wind forecast rises slightly in 2016 and falls by 1.1% in years 2017 and 2018. Similar, RMSE decreases by less than 2%. The changes of wind forecast accuracy in years 2017 and 2018 are not statistically significant.

The solar predictions seem to be the most difficult to improve. The proposed models are not able to significantly reduce the MAE and RMSE. We believe that the proposed linear model is not sufficient enough to capture the dynamic structure of solar generation.

Table 5
MAE and RMSE of price forecasts (DA and ID), in years 2017–2018.

Variable	Model	Measure	Fundamentals			
			TSO	None	Enhanced	Real
DA	(5)	MAE	6.004	7.195	5.920	5.912
		(p-val)		(0.000)	(0.033)	(0.229)
		RMSE	8.526	10.885	8.434	8.621
		(p-val)		(0.000)	(0.037)	(0.370)
ID	(6)	MAE	7.701	8.638	7.621	7.176
		(p-val)		(0.000)	(0.014)	(0.000)
		RMSE	11.183	12.994	11.066	10.523
		(p-val)		(0.000)	(0.002)	(0.000)
	(8)	MAE	7.899	8.935	7.719	7.058
		(p-val)		(0.005)	(0.001)	(0.000)
		RMSE	11.315	13.470	11.117	10.363
		(p-val)		(0.920)	(0.094)	(0.000)

Note: the difference of the forecast accuracy is tested with Diebold-Mariano test with autocorrelation of order 7; p-values of the DM test are presented in brackets.

4.2. Price forecasts

The enhanced predictions of load, wind and solar generation are next used for electricity price forecasting. In order to evaluate the impact of fundamentals on price predictions, four model setups are considered. In the first one, later called a benchmark, the TSO forecasts are included in models (5), (6) and (8). The benchmarks are compared with models, in which fundamentals are excluded from mentioned regressions. It should be noticed that in such case, the model (8) is simplified substantially and includes only the predicted level of day-ahead prices and lagged intraday prices. Next, we assess the performance of models utilizing the predictions obtained with models (1), (3) and (4). Finally, we consider a case, in which perfect forecasts of fundamental variables are available for researchers. This means that the real values of load, wind and solar generations are known before the price forecasts are calculated.

Two measures of price forecast accuracy, MAE and RMSE, are presented in Table 5 jointly for years 2017–2018. They are complemented by DM tests, which compare the forecast performance of a particular model with a benchmark (TSO) model. Since we propose two different model specifications for intraday prices, their outcomes are evaluated separately.

The results indicate that fundamental variables contain information, which can significantly improve price forecast accuracy. MAE of models, which do not include fundamentals, is larger by 19.8%, 12.2% and 13.1% respectively than MAE of benchmark models.

When day-ahead prices are considered the results are mixed. MAE measure indicates that improving the fundamental forecasts (Enhanced and Real columns) results in more accurate price forecasts. On the other hand, when RMSE is analyzed, forecasts utilizing real generation structure are not significantly different from predictions computed with the benchmark model. Hence, the exact knowledge of future levels of load, wind and solar generation does not help in predicting the day-ahead prices.

The results for intraday prices depend on the model specification. It could be noticed that in the case of perfect forecasts of fundamentals, model (8) gives more accurate predictions than the other (6). This implies that day-ahead prices are main driver of intraday prices, which

Table 6
Share of correct market classifications and additional revenues, in years 2017–2018.

Models		Correct classifications, *p* (%)			Revenues, *π* (EUR)		
DA	ID	TSO	Enhanced	Real	TSO	Enhanced	Real
(5)	(6)	52.0%	52.1%	52.9%	5705	5908	7577
(5)	(8)	49.6%	50.8%	58.9%	4075	4627	16,333

confirms previous findings of Kiesel and Paraschiv (2017) and Ziel (2017). Secondly, the outcomes indicate that models using the enhanced forecasts of fundamentals are significantly better than the TSO based predictions. The differences between models are more pronounced for model (8), when the MAE falls from 7.899 to 7.719. At the same time, using the real observations of load, wind and solar generation decreases MAE by 10.6% to 7058. This shows the range of potential gains from the enhancement of forecasts of fundamentals.

4.3. Market choice

In this research, the market choice in based on the sign of the price spread. The decision variable, $Y_{t,h} \in \{0,1\}$, is defined by (9). It takes value 1, when the intraday price is higher than the day-ahead price, $ID_{t,h} > DA_{t,h}$, and zero otherwise. Since the actual prices are not known, the sign of the spread is predicted using the day-ahead price forecasts. In such case, $\widehat{Y}_{t,h} = 1$ when $\widehat{ID}_{t,h} > \widehat{DA}_{t,h}$ and $\widehat{Y}_{t,h} = 0$, when $\widehat{ID}_{t,h} \leq \widehat{DA}_{t,h}$. Similar to Maciejowska et al. (2019), the accuracy of forecasts, $\widehat{Y}_{t,h}$, is evaluated using two measures: a ratio of correct predictions (p) and an additional revenues (π). The ratio of correct predictions is computed as follows

$$p = \frac{\#\left(Y_{t,h} = \widehat{Y}_{t,h}\right)}{\#Y_{t,h}}, \qquad (11)$$

where $\#\left(Y_{t,h} = \widehat{Y}_{t,h}\right)$ is the number of correctly predicted spread signs and $\#Y_{t,h}$ is the size of the evaluation sample.

The additional revenues are calculated as income from selling 1 MWh according to the predicted decision variable, $\widehat{Y}_{t,h}$, over the day-ahead benchmark. This implies that the total additional daily revenue, π_t, becomes

$$\pi_t = \sum_{h=1}^{24} \left(\widehat{Y}_{t,h} ID_{t,h} + \left(1 - \widehat{Y}_{t,h}\right) DA_{t,h} - DA_{t,h}\right) = \sum_{h=1}^{24} \widehat{Y}_{t,h} \Delta P_{t,h}. \qquad (12)$$

The two-year additional revenue, π, is computed as a sum of daily revenues from 1 October 2017 to 30 September 2019. As shown by the literature (Kath and Ziel, 2018; Gianfreda et al., 2018; Maciejowska et al., 2019), financial measures are of particular importance in evaluation of forecast accuracy and does not necessary coincide with classical statistical measures, for example p.

The results are presented in Table 6. First, the ratio of correct market classification is evaluated. The TSO based models are able to predict correctly, which market offers a higher price, in 49.6%–52.0% of cases. The ratios for models using the enhanced forecasts range from 50.8% to 52.1% and reach the higher level with the intraday model (6). At the same time, an access to perfect forecasts of fundamentals enables to choose the market correctly in 52.9%–58.9% of cases. In this case utilizing the intraday model (8) is favorable, showing that this model (8) has more potential, should the forecast be improved further.

Next, additional revenues from the price driven market choice are compared. It can be noticed that the additional revenues will reach more than 16,000 EUR from 1 October 2017 to 30 September 2019, if the true values of fundamentals are known at the time of the decision. It approximates the upper bound of the presented methodology. At the same time, revenues from the choice utilizing the TSO predictions vary between 4075 and 5705 EUR, which is about 30% of the perfect information result. Finally, decision based on enhanced fundamentals leads to an increase of the revenue by 4627–5908.

Finally, when the results of specifications using enhanced fundamental forecasts are analyzed, one could observe gains over the TSO based approaches in both cases. When the model (8) is adopted for forecasting intraday prices, the revenues reach 4627 EUR, which is 552 EUR more than the benchmark. Although the additional revenues increase slightly, the full information case shows that there is a plenty of room for the further improvement, and that the main idea itself is reasonable.

5. Conclusions

This article analyzes the German electricity market and evaluates the system forecasts of the fundamental variables: load, wind and solar generation. The research consists of three parts: calibration of time series models of fundamentals and assessment of their prediction accuracy, utilization of the enhanced predictions in day-ahead forecasting of the day-ahead and intraday electricity prices and finally, an employment of the calculated forecasts in the utility decision process.

First, ARX types of models are employed for forecasting fundamental variables. Since we do not know the optimal length of a calibration window, we follow an approach proposed by Hubicka et al. (2019) and combine forecasts computed from a few short and a few long window sizes. The obtained results show that the load forecast can be significantly improved over the TSO benchmark. This indicates that the system operator does not utilize all the information available in the market at the time of the forecast publication. On the other hand, the ARX types of models are able to improve the TSO wind and solar predictions only slightly. Hence, these two turn out to be more demanding variables, which need a particular modelling approach (nonlinear or/and including additional exogenous variables).

Second, the enhanced predictions of fundamentals are used for forecasting day-ahead and intraday prices. Both variables are predicted day-ahead, as they are later used for choosing an optimal market for selling the energy. Two types of intraday models are analyzed: one using the ARX specification similar to the day-ahead market and second, exploring the dependence on the day-ahead prices and fundamentals forecasts errors. The results reveal that the knowledge of real levels of generation and its structure does not help to forecast day-ahead prices. On the other hand, enhanced fundamental forecasts reduce both MAE and RMSE of day-ahead prices, as compared to the TSO benchmark. This indicates that market participants use the information available at the time of taking decision to place their offers. When the intraday prices are considered, the results clearly demonstrate that any improvement of accuracy of fundamental predictions leads to better price forecasts, both in terms of MAE and RMSE measures.

Finally, the improved forecasts of fundamentals and electricity prices are utilized in the decision process. In this article, a utility needs to choose where to sell 1 MWh of electricity: in the day-ahead or in the intraday market. The decision is data driven and is compared with a benchmark (selling 1 MWh in the day-ahead market). The gains from prediction enhancement are measured by an additional yearly revenue. It turns out that the correction of fundamental forecasts results in a significant income increase. The extra revenue coming from the market choice rises by 13.5% from 4075 EUR to 4627 EUR a year per 1 MWh. Moreover, the results show that if the actual values of fundamentals are known, the revenue could reach 16,333 EUR, which encourages further research in the field.

Acknowledgments

This work is partially supported by the National Science Center (NCN, Poland) through grant SONATA BIS no. 2019/34/E/HS4/00060 (K.M.); the German Research Foundation (DFG, Germany) and the National Science Center (NCN, Poland) through BEETHOVEN grant no. 2016/23/ G/HS4/01005 (T.W.); the Ministry of Science and Higher Education (MNiSW, Poland) through Diamond Grant no. DI2019 002749 (W.N.).

Appendix A

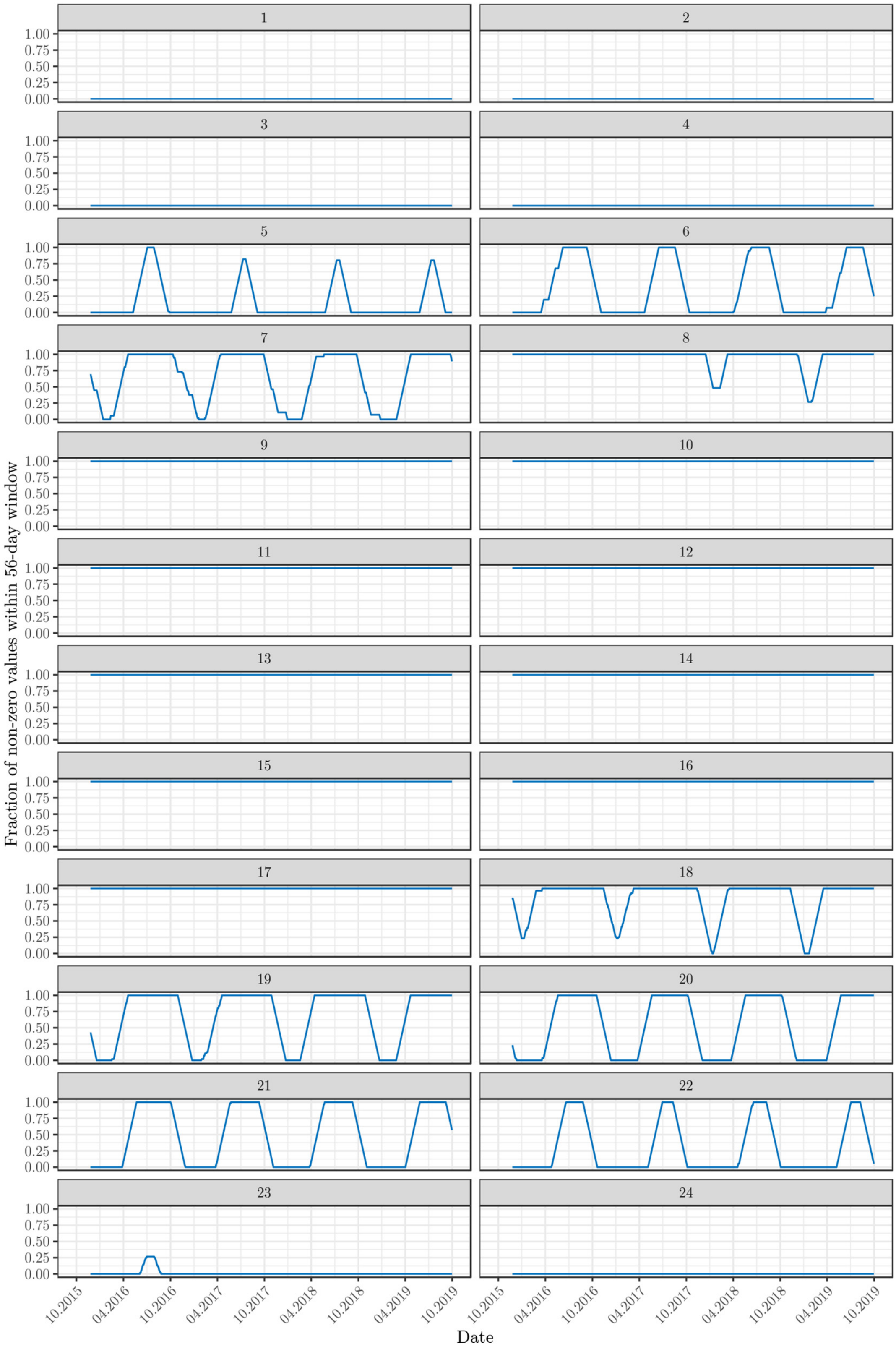


Fig. 3. Fraction of days within a 56-day rolling window with forecasted solar energy values greater than 0 MWh for every hour of the day.

References

Bunn, D.W., Gianfreda, A., Kermer, S., 2018. A trading-based evaluation of density forecasts in a real-time electricity market. Energies 11.

Diebold, F., Mariano, R., 1995. Comparing predictive accuracy. J. Bus. Econ. Stat. 13, 253–263.

Gianfreda, A., Parisio, L., Pelagatti, M., 2016. The impact of RES in the Italian day–ahead and balancing markets. Energy J. 37.

Gianfreda, A., Parisio, L., Pelagatti, M., 2018. A review of balancing costs in Italy before and after res introduction. Renew. Sust. Energ. Rev. 91, 549–563.

Gianfreda, A., Ravazzolo, F., Rossini, L., 2020. Comparing the forecasting performances of linear models for electricity prices with high RES penetration. Int. J. Forecast. 36, 974–986.

Gürtler, M., Paulsen, T., 2018. The effect of wind and solar power forecasts on day-ahead and intraday electricity prices in Germany. Energy Econ. 75, 150–162.

Hubicka, K., Marcjasz, G., Weron, R., 2019. A note on averaging day-ahead electricity price forecasts across calibration windows. IEEE Trans. Sustain. Energy 10, 321–323.

Kath, C., Ziel, F., 2018. The value of forecasts: quantifying the economic gains of accurate quarter-hourly electricity price forecasts. Energy Econ. 76, 411–423.

Kiesel, R., Paraschiv, F., 2017. Econometric analysis of 15-minute intraday electricity prices. Energy Econ. 64, 77–90.

Koch, C., Hirth, L., 2019. Short-term electricity trading for system balancing: an empirical analysis of the role of intraday trading in balancing Germany's electricity system. Renew. Sust. Energ. Rev. 113.

Lisi, F., Edoli, E., 2018. Analyzing and forecasting zonal imbalance signs in the Italian electricity market. Energy J. 39.

Maciejowska, K., Nitka, W., Weron, T., 2019. Day-ahead vs. intraday forecasting the price spread to maximize economic benefits. Energies 12, 631.

Marcjasz, G., Serafin, T., Weron, R., 2018. Selection of calibration windows for day-ahead electricity price forecasting. Energies 11, 2364.

Monteiro, C., Ramirez-Rosado, I.J., Fernandez-Jimenez, L.A., Conde, P., 2016. Short-term price forecasting models based on artificial neutral networks for intraday sessions in the Iberian electricity markets. Energies 9, 1–24.

Nowotarski, J., Weron, R., 2018. Recent advances in electricity price forecasting: a review of probabilistic forecasting. Renew. Sust. Energ. Rev. 81, 1548–1568.

Nowotarski, J., Raviv, E., Trück, S., Weron, R., 2014. An empirical comparison of alternate schemes for combining electricity spot price forecasts. Energy Econ. 46, 395–412.

Pape, C., Hagemann, S., Weber, C., 2016. Are fundamentals enough? Explaining price variations in the German day-ahead and intraday power markets. Energy Econ. 54, 376–387.

Paraschiv, F., Erni, D., Pietsch, R., 2014. The impact of renewable energies on EEX day-ahead electricity prices. Energy Policy 73, 196–210.

Serafin, T., Uniejewski, B., Weron, R., 2019. Averaging predictive distributions across calibration windows for day-ahead electricity price forecasting. Energies 12, 2561.

Uniejewski, B., Weron, R., 2018. Efficient forecasting of electricity spot prices with expert and LASSO models. Energies 11, 2039.

Uniejewski, B., Nowotarski, J., Weron, R., 2016. Automated variable selection and shrinkage for day-ahead electricity price forecasting. Energies 9, 621.

Uniejewski, B., Marcjasz, G., Weron, R., 2019a. On the importance of the long-term seasonal component in day-ahead electricity price forecasting. Part II - probabilistic forecasting. Energy Econ. 79, 171–182.

Uniejewski, B., Marcjasz, G., Weron, R., 2019b. Understanding intraday electricity markets: variable selection and very short-term price forecasting using LASSO. Int. J. Forecast. 35, 1533–1547.

Weron, R., 2014. Electricity price forecasting: a review of the state-of-the-art with a look into the future. Int. J. Forecast. 30, 1030–1081.

Woo, C.K., Moore, J., Schneiderman, B., Ho, T., Olson, A., Alagappan, L., Chwala, K., Toyama, N., Zarnikau, J., 2016. Merit-order effects of renewable energy and price divergence in California's day-ahead and real-time electricity markets. Energy Policy 92, 299–312.

Ziel, F., 2016. Forecasting electricity spot prices using LASSO: on capturing the autoregressive intraday structure. IEEE Trans. Power Syst. 31, 4977–4987.

Ziel, F., 2017. Modeling the impact of wind and solar power forecasting errors on intraday electricity prices. IEEE Conference Proceedings - EEM17 https://doi.org/10.1109/EEM.2017.7981900.

Ziel, F., Steinert, R., 2018. Probabilistic mid- and long-term electricity price forecasting. Renew. Sust. Energ. Rev. 94, 251–266.

Ziel, F., Weron, R., 2018. Day-ahead electricity price forecasting with high-dimensional structures: univariate vs. multivariate modeling frameworks. Energy Econ. 70, 396–420.

Paper 5

Combining predictive distributions of electricity prices. Does minimizing the CRPS lead to optimal decisions in day-ahead bidding?

Weronika Nitka, Rafał Weron

Vol. 33, No. 3 (2023) | DOI: 10.37190/ord230307

OPEN ACCESS

Operations Research and Decisions

www.ord.pwr.edu.pl

OPERATIONS RESEARCH AND DECISIONS QUARTERLY

Combining predictive distributions of electricity prices. Does minimizing the CRPS lead to optimal decisions in day-ahead bidding?

Weronika Nitka[1]* Rafał Weron[1]

[1] *Department of Operations Research and Business Intelligence, Wrocław University of Science and Technology, Wrocław, Poland*

* *Corresponding author, email address: weronika.nitka@pwr.edu.pl*

Abstract

Probabilistic price forecasting has recently gained attention in power trading because decisions based on such predictions can yield significantly higher profits than those made with point forecasts alone. At the same time, methods are being developed to combine predictive distributions, since no model is perfect and averaging generally improves forecasting performance. In this article, we address the question of whether using CRPS learning, a novel weighting technique minimizing the continuous ranked probability score (CRPS), leads to optimal decisions in day-ahead bidding. To this end, we conduct an empirical study using hourly day-ahead electricity prices from the German EPEX market. We find that increasing the diversity of an ensemble can have a positive impact on accuracy. At the same time, the higher computational cost of using CRPS learning compared to an equal-weighted aggregation of distributions is not offset by higher profits, despite significantly more accurate predictions.

Keywords: *decision support, day-ahead electricity bidding, predictive distribution, combining forecasts, CRPS learning*

1. Introduction

To mitigate risks or increase profits from trading in day-ahead power markets, market participants use data-driven decision support techniques [12, 16, 17, 28]. For years, these have relied on point forecasts of the major variables of interest: loads (or demand for electricity), generation from renewable energy sources (RES), and electricity prices [10, 30]. However, as recently shown by Uniejewski and Weron [27], decisions based on probabilistic price forecasts, i.e., quantiles, prediction intervals, or whole predictive distributions, can yield significantly higher profits. For the quantile-based bidding strategies considered in the Polish day-ahead power market, the profit obtained was from 5% to 19% higher than for the strategy based on point forecasts alone.

Point forecasts are far more popular in the electricity price forecasting (EPF) literature, not only in a decision support context. As reported by Maciejowska et al. [18], probabilistic EPF was not a part of

Received 27 August 2023, accepted 6 September 2023, published online 21 September 2023
ISSN 2391-6060 (Online)/© 2023 Authors

The costs of publishing this issue have been cofinansed by the program *Development of Academic Journals* of the Polish Ministry of Education and Science under agreement No. RCN/SP/0241/2021/1

the mainstream literature until the Global Energy Forecasting Competition in 2014 [9], and even now, no more than 15% of the Scopus-indexed articles concern it. Although business analysts have begun to recognize their importance in the planning and operation of energy systems (see, e.g., [13]), it is not easy to generate accurate probabilistic predictions. Combining forecasts obtained from different model specifications [19] or calibration sample lengths [11, 26] one can significantly increase accuracy without sacrificing computational complexity or interpretability. Compared to selecting a single best-performing forecast, combining forecasts from multiple models offers several advantages such as increased resilience against model uncertainty or misspecification, and better adaptability in the event of structural breaks [29].

Forecast combinations (also called ensemble forecasts) involve assigning weights to the individual predictions (or experts). While naive, i.e., equal, weighting is a straightforward – and surprisingly robust – way of averaging point forecasts, in the case of predictive distributions, a choice must be made about what to combine. Two natural approaches are vertical averaging of probabilities and horizontal averaging of quantiles [15, 20] but the authors do not agree on which is better. Berrisch and Ziel [1] have recently proposed a cutting-edge weighting technique, called CRPS learning that accounts for variations in predictive performance over time and across quantiles of the distribution. It optimizes weights with respect to the continuous ranked probability score (CRPS), the standard error metric for probabilistic forecasts [8, 18].

In this article, we address the question of whether forecast combinations obtained by minimizing the CRPS lead to optimal decisions in day-ahead bidding. To this end, we conduct a comprehensive empirical study involving:

- six years of hourly day-ahead electricity prices from the German EPEX market,
- state-of-the-art probabilistic forecasts generated by Marcjasz et al. [20] using distributional deep neural networks (DDNN), as well as deep neural networks (DNN) and LASSO-estimated autoregressive (LEAR) models combined with quantile regression (QR),
- two approaches to combining predictive distributions – horizontal averaging of quantiles [15] and CRPS learning [1].

Since statistical measures of forecast accuracy do not assess the utility of a forecast to its potential end users [10, 13, 18, 31], we calculate the profits of a day-ahead bidding strategy [20, 25]. The latter aims to find the most financially beneficial hours of the next day to buy electricity and charge a battery, then discharge it and sell electricity. To minimize the risk of losses, limit orders are submitted to the power exchange with the limits determined by selected quantiles of the predictive distributions.

The remainder of the article is organized as follows. In Section 2, the dataset and assumptions for the forecasting problem are introduced. Section 3 describes the details of ensemble construction. The results are presented in Section 4, with the forecast accuracy being the focus of Section 4.1, the trading simulation described in 4.2, and its financial results in Section 4.3. Finally, Section 5 wraps up the results and concludes.

2. Preliminaries and data sources

We assume a standard short-term forecast horizon of 1 day, performed in a rolling window scheme [30]. More precisely, forecasts of all 24 hourly prices on the day d are calculated at the same time in the morning of day $d-1$, i.e., before the day-ahead market for day d closes, and that the model parameters are estimated using a calibration sample of D most recent past observations. In our case, the underlying data are hourly day-ahead electricity prices from the German EPEX market spanning the period from 1 January 2015 to 31 December 2020. The prices, day-ahead predictions of the loads, and RES generation are publicly available from the ENTSO-E Transparency platform (`https://transparency.entsoe.eu`). The full dataset, including emission allowances and fuel prices, is also available from `https://github.com/gmarcjasz/distributionalnn`, a GitHub repository that accompanies [20].

The first $D = 1456$ days of the dataset are used as the initial calibration sample for all models, and the additional 182 days are needed to calculate the quantile regression forecasts. The remaining 554 days from 27 June 2019 until 31 December 2020 constitute the out-of-sample test period. Note that the latter includes a major drop in the level of prices associated with a decrease in the demand for electricity during the initial stage of the COVID-19 pandemic.

Since this study focuses on the evaluation of combination schemes for probabilistic forecasts and not on the computation of predictive distributions themselves, we work directly with a pool of readily available state-of-the-art forecasts generated by Marcjasz et al. [20]. The latter takes the form of 99 predicted percentiles for each day and hour, which approximate the predictive distribution quite well. They are generated by twelve different models, eight of which are distributional deep neural networks (DDNN) with the output layer returning fitted parameters of the normal or Johnson's SU (JSU) distributions. Since the quantile functions have no closed-form representations, the percentile forecasts are obtained as empirical quantiles of a 10000-element random sample generated from the output normal or JSU distribution. These forecasts are denoted further in the text as `DDNN_N_{1-4}` and `DDNN_JSU_{1-4}`, respectively, with the numbers representing the hyperparameter set used for tuning the DDNNs.

The remaining models directly predict 99 percentiles with the use of quantile regression averaging [QRA; 22] or quantile regression machine [QRM; 21] methods, applied to point forecasts of two well--performing benchmarks – LASSO-estimated autoregressive models (LEAR) and deep neural networks (DNN) [14]. The combinations of these techniques make up the final four forecasts used in the ensembles: `LEAR_QRA`, `LEAR_QRM`, `DNN_QRA` and `DNN_QRM`. Note that in the LEAR models, the prices for each of the 24 hourly load periods are treated as separate time series and estimated independently, whereas in the DNN and DDNN neural networks, the 24 prices or 24 distributions are estimated jointly.

3. Methods

Combining forecasts has become a well-established method to increase predictive accuracy. The advantages of using ensembles of experts in place of individual models include diversification of used information and increasing robustness against model misspecification and structural breaks in the data [24]. While the literature generally recommends combining forecasts, many questions still remain open regarding the construction of ensembles. Across a multitude of possible specifications, the forecaster must decide on how many predictions to combine, how to perform forecast selection, and how to choose weights for each expert. Combining probabilistic forecasts is even more tricky, as the assigned weights may change not just across experts and time, but also across quantile levels [29].

3.1. Equal weighting

In point forecasting, the use of naive, i.e., equal, weights is often found to outperform more sophisticated weighting schemes because the latter introduce excessive estimation bias [4]. In the case of predictive distributions, however, a choice must be made about what to combine. Two natural approaches are vertical averaging of probabilities, which boils down to computing a mixture distribution, and horizontal averaging of quantiles, where each quantile of the ensemble forecast is a weighted average of the corresponding quantiles of all individual experts [15]. While the literature does not agree on which approach is better, Marcjasz et al. [20] emphasize that horizontal averaging is more robust and results in a sharper, i.e., more concentrated, unimodal distribution. On the other hand, vertical averaging may lead to increased variance and multimodality. For this reason, as well as potential information loss due to interpolation needed to perform vertical averaging, only horizontal averaging of quantiles is considered in this paper. For consistency with other EPF studies, we denote it in the text by `qEns`.

3.2. CRPS learning

Berrisch and Ziel [1, 2] have recently proposed a cutting-edge weighting technique that accounts for variations in predictive performance over time and across quantiles; it is freely available in the *profoc* package for R ([3], https://cran.r-project.org/web/packages/profoc). The authors called it CRPS learning since it optimizes weights with respect to the continuous ranked probability score (CRPS). The latter is a proper scoring rule and the standard error metric for probabilistic forecasts [7, 8]. It is defined as:

$$\mathrm{CRPS}(F,\,x) = -\int\limits_{-\infty}^{\infty} \left(F(y) - \mathbb{1}_{\{y\geq x\}}\right)^2 dy \tag{1}$$

where F is the cumulative distribution function of the evaluated probabilistic forecast. It can equivalently be represented as a scaled integral of the quantile loss, which for an equidistant grid can be approximated by:

$$\mathrm{CRPS}(F,\,x) \approx \frac{2}{M}\sum_{i=1}^{M} \mathrm{QL}_{p_i}\left(F^{-1}(p_i),\,x\right) \tag{2}$$

where $(p_1,\,\ldots,\,p_M)$ is an equidistant monotonically increasing dense grid of probabilities and $\mathrm{QL}_p(q,x) = \left(\mathbb{1}_{\{x<q\}} - p\right)(q-x)$ is the quantile loss for a quantile forecast q of true value x for probability $p \in (0,\ 1)$, also known as the pinball score [1, 18]. In practice, the scaling factor of 2 in eq. (2) is typically omitted; this is also the case here.

The CRPS learning algorithm aims to combine probabilistic forecasts by selecting optimal weights for averaging across quantiles to minimize the CRPS of the resulting ensemble. The weight functions are subject to online updating throughout the forecasting period and are chosen pointwise, i.e., for each quantile of the distribution separately, depending on each expert's performance. The framework additionally includes smoothing procedures that reduce estimation noise of the weights [2].

In this study, the CRPS learning framework was applied once per ensemble, with the following arbitrarily chosen set of parameters: Bernstein online aggregation (BOA) for updating weights, penalized probabilistic smoothing with $\lambda = 2^{(-5,\,\cdots,\,5)}$ updated based on past performance, and no forgetting past regret. The remaining options were set to the *profoc* package defaults. Such an approach is denoted in the text by CRPS. Finally, note that the time required to compute forecasts of a single CRPS learning ensemble for the entire test period is ca. 500 times longer than that for the naive qEns weighting. However, it does not exceed 20 s on a laptop equipped with a 9th-generation Intel Core i7-9750H processor.

3.3. Comparison of the two weighting schemes

The general idea of averaging across quantiles, as well as differences between the two weighting schemes, are shown in Figure 1. The illustration shows a toy example of a two-forecast ensemble. Among the two experts, the DDNN_JSU_1 forecast (teal color) is sharper, i.e., more concentrated, predicting prices between 26 and 37 €, and has a smoother cumulative distribution function (CDF), while the LEAR_QRA predictive distribution (red color) is less sharp (with prices between 15 and 44 €) and more rugged. Medians of both experts are relatively close to the actual observed price (31.89 €, vertical line), albeit leaving room for improvement, i.e., with absolute errors of 0.52 and 1.24 €, respectively.

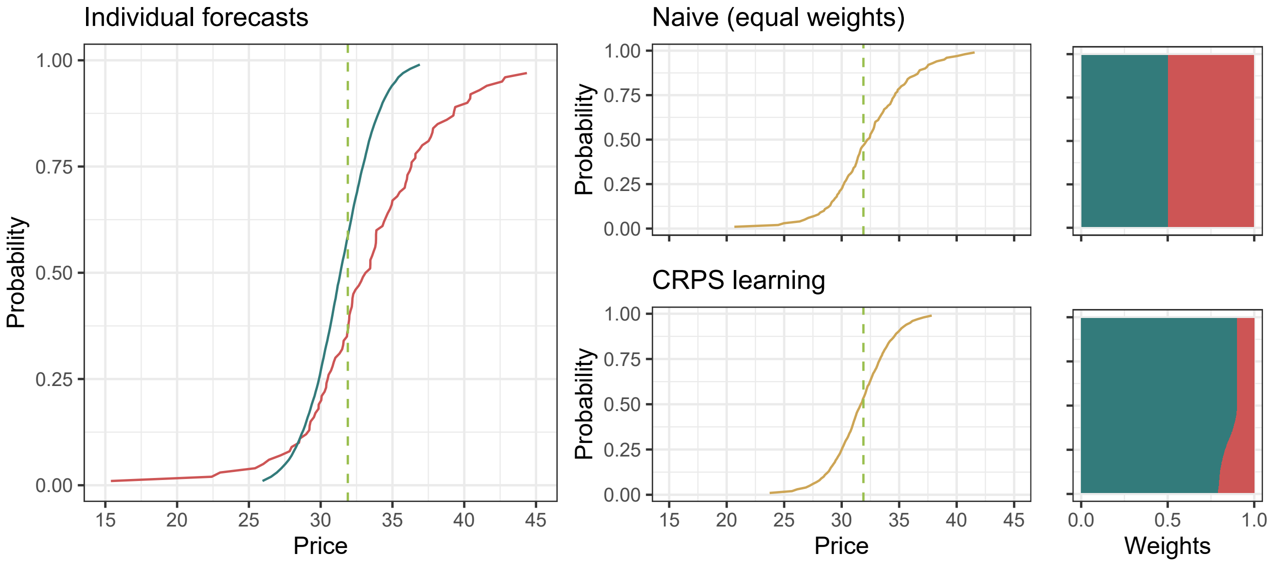


Figure 1. Illustration of the two weighting schemes. The left panel shows predictive distributions obtained from the DDNN_JSU_1 (teal color) and LEAR_QRA (red color) models for a selected hour and day. The center panels present the resulting ensemble forecasts obtained by estimating weights with naive (top) and CRPS learning (bottom) methods. The right panels illustrate the relative weights for each quantile; these are horizontally stacked bar plots with the length of the bar representing the weight of the forecast in the corresponding color and all weights summing up to 1. The dashed vertical line marks the actual price

The two individual forecasts are combined using the two weighting approaches, with the resulting CDFs and the assigned weights shown in the panels to the right. It can be seen that while the qEns approach, by definition, assigns equal weights to all models and quantiles, CRPS learning assigns larger weights to the DDNN_JSU_1 forecast, based on its better past performance (not shown in the plot). The share of the DDNN_JSU_1 forecast is smaller for the lowest 25 percentiles, but nevertheless, it still dominates the CRPS ensemble, leading to its higher sharpness (price range of [24, 38] €) and smoothness compared to the equally weighted ensemble (with values in the range [21, 42] €). However, both forecast combinations provide a more accurate median forecast than the individual experts, with absolute errors of 0.36 € for qEns ensemble and 0.22 € for CRPS learning.

3.4. Selection of experts

The forecaster's second decision is the selection of experts that are aggregated in the ensemble. Following [20], each ensemble we consider in this study contains a set of four DDNN forecasts, either DDNN_N_{1-4} or DDNN_JSU_{1-4}. Furthermore, to diversify the pool of experts we additionally include benchmark quantile regression-based forecasts – either LEAR_QRA and LEAR_QRM or DNN_QRA and DNN_QRM. While they have been demonstrated to perform significantly worse on their own, using them can lead to a higher prediction accuracy of the ensembles by avoiding overfitting [29]. Thus, the resulting naming convention for ensembles is:

```
DDNN_{distribution}_{averaging}_{experts}
```

with distribution={N,JSU} denoting whether normal or JSU forecasts were used, averaging={qEns, CRPS} indicating the use of equal or CRPS learning-derived weights, and experts={LEAR,DNN} added when additional experts – respectively LEAR_QRA and LEAR_QRM or DNN_QRA and DNN_QRM – were included in the ensemble. A graphical illustration of the steps performed in order to construct the ensemble forecasts is shown in Figure 2.

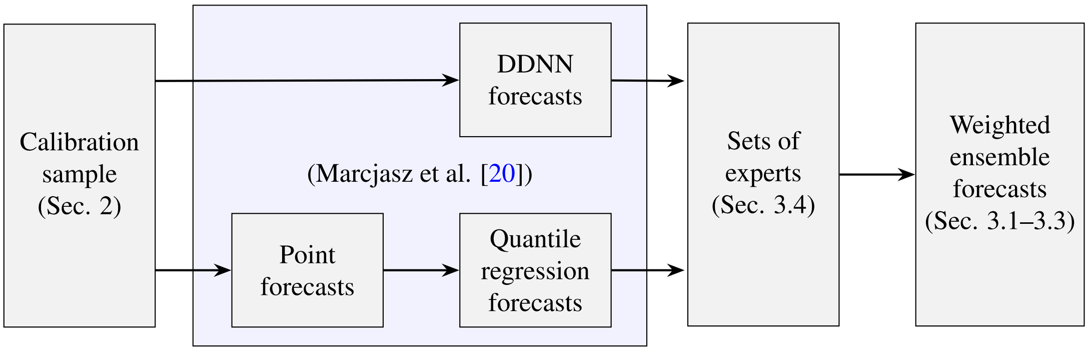

Calibration sample (Sec. 2)
DDNN forecasts
(Marcjasz et al. [20])
Point forecasts
Quantile regression forecasts
Sets of experts (Sec. 3.4)
Weighted ensemble forecasts (Sec. 3.1–3.3)

Figure 2. Schematic illustration of the process of generating ensemble forecasts

It should be noted that additional forecast combinations were explored during the course of this research. The complete list included smaller ensembles (four DDNN forecasts and a single QR-based forecast; best 5 performing models), other combinations of quantile regression forecasts (e.g., four DDNN forecasts, LEAR_QRA, DNN_QRA) and larger ensembles (four DDNN forecasts and four QR-based forecasts; eight DDNN forecasts as in [2]; all available forecasts). They offered comparable or, in the case of the largest ensembles, significantly inferior performance to the combinations listed above, and have been omitted from the presentation of results for the sake of clarity.

In an online learning setting, the forecaster may decide to discard the initial part of the test sample as a burn-in period, which is beneficial for the stability of weights and hyperparameters, see, e.g., [2], which uses a burn-in period of 182 days for combinations of DDNN forecasts. However, the quantile regression forecasts are only available within the 554-day out-of-sample test period, the entirety of which has been used in Berrisch et al. [2] and Marcjasz et al. [20] for evaluation. Since the majority of ensembles we consider in this study include these quantile regression forecasts, for the sake of consistency no burn-in period has been applied.

4. Forecast evaluation

In this section, the generated ensemble forecasts are compared to each other and individual expert models. The evaluation is divided into two parts. First, in Section 4.1, we measure the predictive accuracy in terms of statistical error metrics:

- the mean absolute error (MAE) and the root mean squared error (RMSE) for median and mean forecasts, respectively [6],
- the continuous ranked probability score (CRPS) for probabilistic forecasts [8, 23].

Then, in Sections 4.2–4.3, we measure the predictive accuracy in terms of profits – total and per trade – from a day-ahead bidding strategy that utilizes probabilistic forecasts [20, 25]. Note that the CRPS is approximated by a sum of pinball scores on a grid of 99 percentiles, see eq. (2). The statistical significance of differences in CRPS scores is assessed using the Diebold–Mariano test [5].

4.1. Evaluation in terms of statistical error measures

As Gneiting et al. [7, 8] argue, the goal of probabilistic forecasting is to *maximize the sharpness of the predictive distributions subject to calibration*. Here, calibration (also called reliability or unbiasedness) refers to the statistical consistency between the probabilistic forecasts and the observations, e.g., whether the 50% prediction interval (PI) covers 50% of the actual observations. Sharpness, on the other hand, refers to the concentration of the predictive distributions. For instance, given two reliable 50% PIs, the sharper or more narrow one is better. The CRPS introduced in Section 3 assesses calibration and sharpness simultaneously [7]. Moreover, for a point forecast, it is equal to the MAE [23].

The CRPS values for all models ordered from the lowest/best to the highest/worst are shown in the left panel of Figure 3; the corresponding MAE and RMSE errors in the right panel. Clearly, the two `LEAR` forecasts perform the worst, while the `DDNN_JSU_4` and two `DNN` forecasts the best out of the individual models. Moreover, the individual experts are outclassed by all `DDNN_N` ensembles, which are further outperformed by the `DDNN_JSU` combinations. The `DDNN_JSU_CRPS_LEAR` ensemble achieves the lowest CRPS score. For all ensembles, both weighting schemes typically result in very similar forecasts and thus accuracy. Nevertheless, CRPS learning yields slightly better predictions on average. A similar, though not identical, ordering can be observed for the point forecasting error metrics. The most significant differences are obtained for the `DDNN_JSU_{1,3,4}` experts in terms of the RMSE. As in [20], we calculate the MAE for the median (i.e., the 50th percentile) and the RMSE for the expected value of each distribution; the errors presented for the individual forecasts as well as the `DDNN_N_qEns` and `DDNN_JSU_qEns` ensembles are consistent with the results reported in [20].

To assess the statistical significance of differences in CRPS scores, we perform the Diebold–Mariano (DM) test [5]. In order to correct for daily seasonality in CRPS values, following [14] and [20], we consider a multivariate loss differential series defined for a pair of models A and B as:

$$\Delta_d^{A,B} = \|L_d^A\|_1 - \|L_d^B\|_1 \tag{3}$$

where $L_d^X = \{L_{d,1}^X, \ldots, L_{d,24}^X\}$ is the 24-dimensional vector of hourly CRPS values for model X on day d and $\|L_d^X\|_1$ is its L_1 norm. For each pair of models we apply two one-sided DM tests.

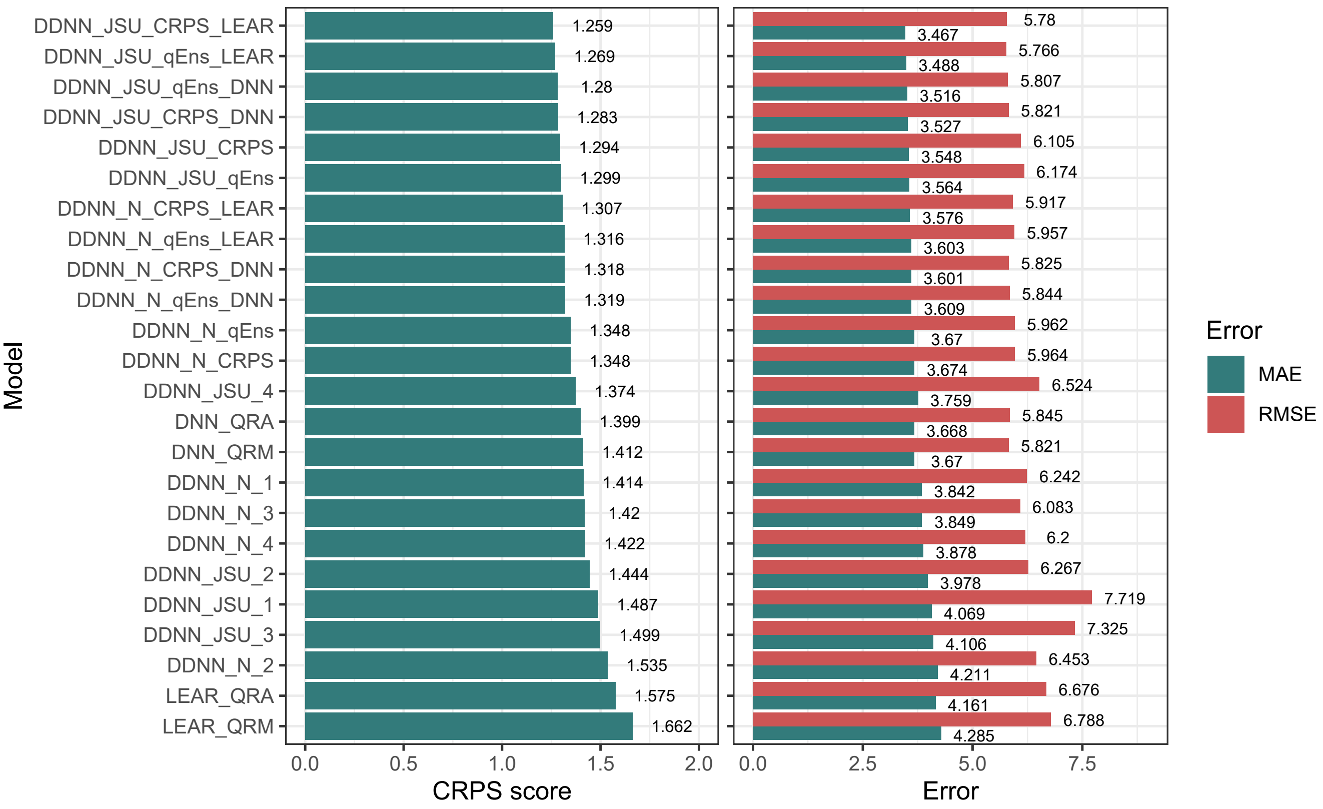


Figure 3. CRPS scores (left panel) and MAE and RMSE errors (right panel) for all models and ensembles ordered from the lowest to the highest CRPS. Compare with a CRPS of 1.284 of the best performing model of Berrisch and Ziel [2]

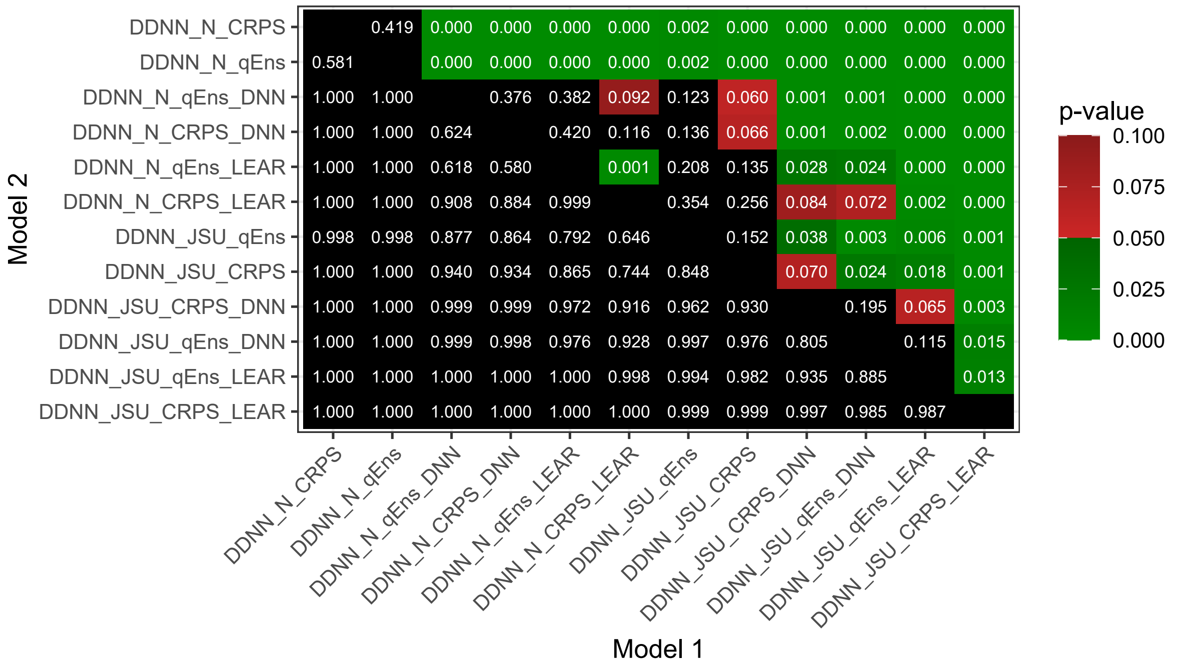


Figure 4. Results (p-values) of the Diebold–Mariano test for the CRPS loss; the lower it is the more significant is the difference between the forecasts of a model on the X-axis (better) and the forecasts of a model on the Y-axis (worse). We use a coloring scheme to highlight the differences

A heatmap of the respective p-values is presented in Figure 4. The results indicate that the predictions of the DDNN_JSU_CRPS_LEAR ensemble are significantly better than those of all competing models. The predictions of the remaining ensembles within the top 4 do not significantly differ from each other. Another ensemble whose forecasts are significantly better than those ranked lower in terms of the CRPS is DDNN_JSU_CRPS_DNN, while most other ensembles do not yield significantly better predictions than ensembles similarly ranked in terms of the CRPS.

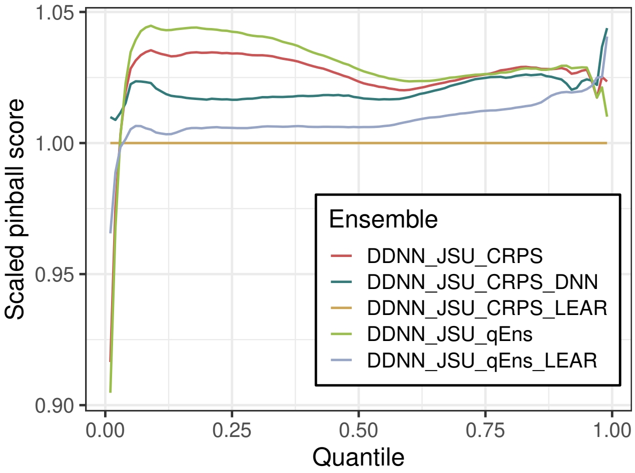


Figure 5. Pinball scores of selected best performing ensembles across quantiles, relative to the DDNN_JSU_CRPS_LEAR ensemble. A lower score corresponds to better performance

The CRPS score provides a single number for all quantiles (and each time point in the test period). To see how the pinball scores for individual percentiles contribute to the CRPS, in Figure 5 we plot them for selected best-performing ensembles. To enhance readability, all values are plotted with respect to the pinball scores of the best-performing ensemble, i.e., DDNN_JSU_CRPS_LEAR. Clearly, the relative performance of the ensembles is not uniform across the entire distributions. The largest disparity can be seen below the median, with the relative ranking of the ensembles changing for the three lowest percentiles. Above the median, the ensembles perform similarly.

4.2. Day-ahead bidding

Following Uniejewski [25], we consider a realistic trading strategy that utilizes battery storage and day-ahead bidding based on probabilistic price forecasts. The goal is to buy electricity cheaply at hour $h1$ and charge the battery, then discharge it and sell the electricity expensively at hour $h2 > h1$. To minimize the risk of losses, limit orders are submitted to the power exchange with the limits determined by selected – based on the trader's risk appetite – quantiles of the predictive distributions.

We assume that the efficiency of charging as well as discharging the battery is 90%. Hence, $1/0.9 \approx 1.1$ MWh is needed to charge the battery by 1 MWh. Similarly, discharging 1 MWh generates only 0.9 MWh. Further, we assume that the total usable capacity of the battery is $B = 2$ MWh and that at the beginning of the simulation period, the battery starts halfway charged ($B = 1$). If both orders are executed on the next day, this state persists. If $B = 0$ at the beginning of a day, an unlimited bid to buy 1 MWh is placed at hour $h* < h2$, and if $B = 2$, an unlimited offer to sell 1 MWh is placed at hour $h* < h1$.

For each day in the out-of-sample test period, the following two steps are performed. First, based on median price forecasts $Y^{0.5}_{d,h}$ for day d and hours $h = 1, 2, ..., 24$ computed on day $d-1$, hours $h1$, $h2$ and $h*$ are selected to maximize the profit:

$$\Pi_d = -\frac{1}{0.9}\hat{Y}^{0.5}_{d,\,h1} + 0.9\hat{Y}^{0.5}_{d,\,h2} - \mathbb{1}_{\{B=0\}}\frac{1}{0.9}\hat{Y}^{0.5}_{d,\,h*} + \mathbb{1}_{\{B=2\}}0.9\hat{Y}^{0.5}_{d,\,h*} \tag{4}$$

When the battery is halfway charged ($B = 1$), this optimization problem reduces to selecting hours with the lowest and the highest predicted median price. In other cases, linear programming is used to optimize the selection of $h1$, $h2$, and $h*$.

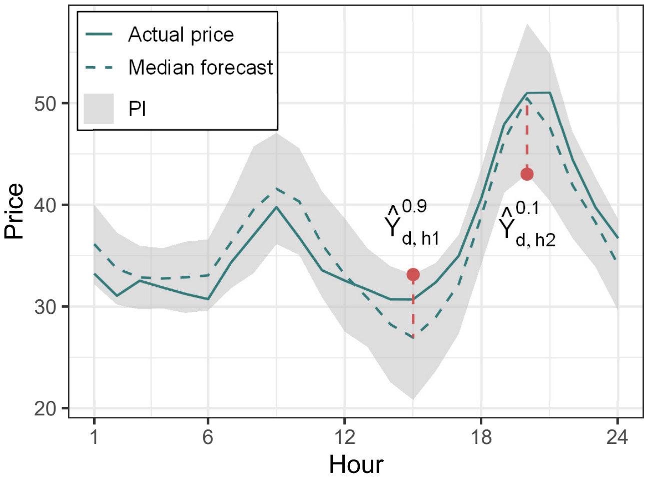


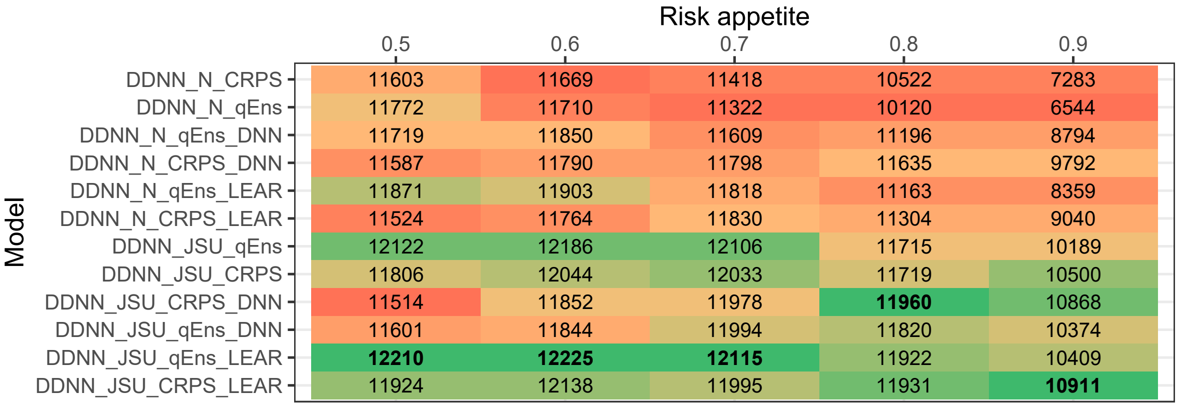
Figure 6. Illustration of the trading strategy with limit orders defined by the 80% PIs, corresponding to a risk appetite of 0.8. Red dots indicate the price limits for the selected hours

Next, following Marcjasz et al. [20], a profitability condition is checked. If the transaction is expected to be profitable, i.e., the sum of the first two terms in eq. (4) is greater than zero, a buy order with price limit $\hat{Y}^{1-q}_{d,\,h1}$ and a sell order with price limit $\hat{Y}^{q}_{d,\,h2}$ are placed. Here $q = (1-\alpha)/2$ and α is trader's risk appetite, i.e., the PI level, set only once for the whole test period. This is illustrated in Figure 6 for a sample day and forecasts generated by the `DDNN_N_qEns` ensemble. In this example, both orders would be accepted since the actual price falls within the 80% PI, corresponding to a risk appetite of $\alpha = 0.8$. However, hour h_2 is predicted suboptimally, a slightly higher price was observed for hour 21.

4.3. Evaluation in terms of trading profits

The total profits are presented in Table 1 for five values of risk appetite α ranging from 0.5 to 0.9; the minor differences between the reported values and those in [20] for the `DDNN_N_qEns` and `DDNN_JSU_qEns` ensembles are a result of correcting a bug in the original software. The profitability results somewhat correspond to the CRPS results, although with a few notable exceptions. On average, the `DDNN_JSU` ensembles achieve higher total profits than the `DDNN_N` ensembles, mirroring their better performance in terms of the CRPS. The detailed ranking of those ensembles is where the outcomes start to vary. While the CRPS weighting scheme outperforms equal weights in terms of forecast accuracy, this trend is mostly reversed in the financial results. This is especially true for lower values of the risk appetite.

Table 1. Total profits from the quantile-based trading strategy in the whole test period for risk appetite ranging from 0.5 to 0.9. The highest values in each column are in bold. Cells are colored independently in each column from the best (→ green) to the worst (→ red)



Model	0.5	0.6	0.7	0.8	0.9
DDNN_N_CRPS	11603	11669	11418	10522	7283
DDNN_N_qEns	11772	11710	11322	10120	6544
DDNN_N_qEns_DNN	11719	11850	11609	11196	8794
DDNN_N_CRPS_DNN	11587	11790	11798	11635	9792
DDNN_N_qEns_LEAR	11871	11903	11818	11163	8359
DDNN_N_CRPS_LEAR	11524	11764	11830	11304	9040
DDNN_JSU_qEns	12122	12186	12106	11715	10189
DDNN_JSU_CRPS	11806	12044	12033	11719	10500
DDNN_JSU_CRPS_DNN	11514	11852	11978	**11960**	10868
DDNN_JSU_qEns_DNN	11601	11844	11994	11820	10374
DDNN_JSU_qEns_LEAR	**12210**	**12225**	**12115**	11922	10409
DDNN_JSU_CRPS_LEAR	11924	12138	11995	11931	**10911**

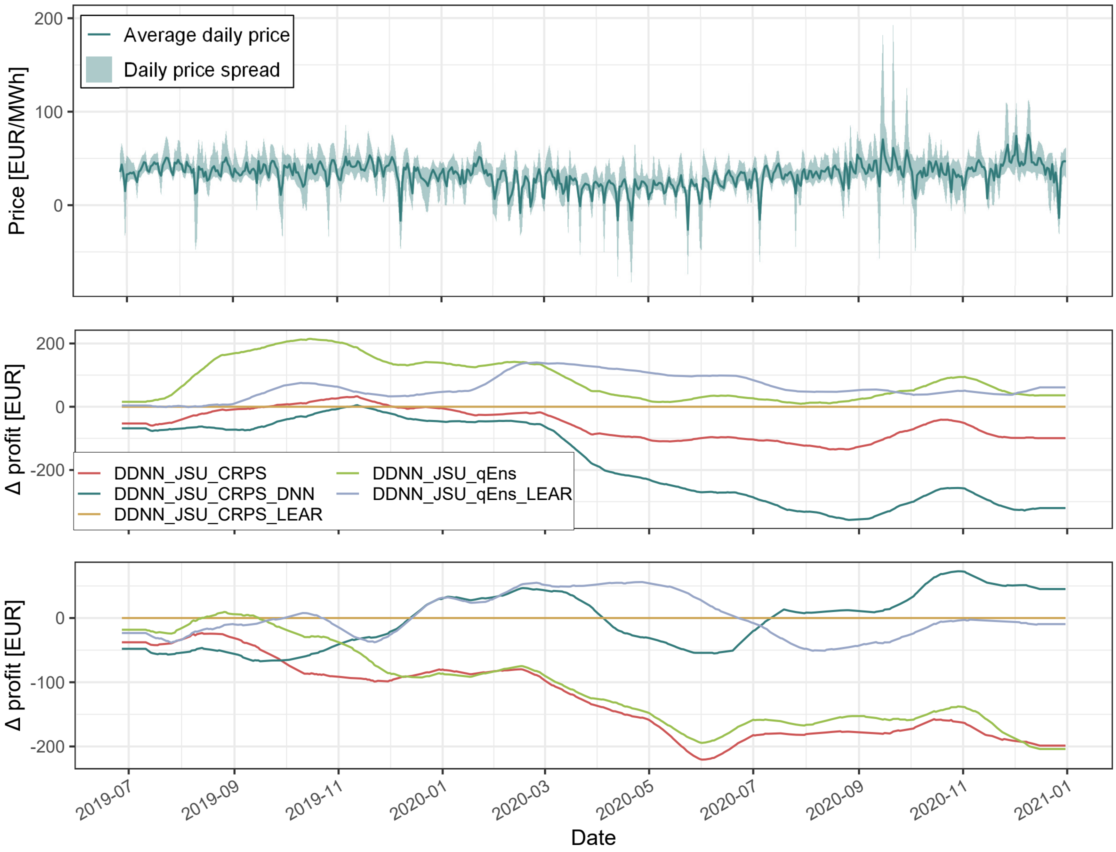


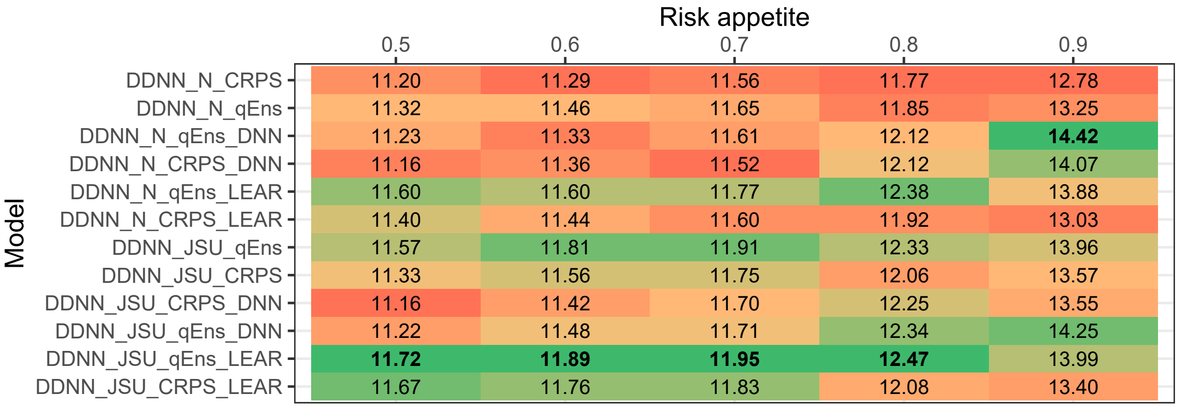
Figure 7. Average daily (dark green) and the minimum and maximum hourly (light green) prices in Germany from 27 June 2019 to 31 December 2020 (top panel). 30-day moving average of cumulative profit for the best-performing strategies, shown as a difference between the cumulative profit of each ensemble and `DDNN_JSU_CRPS_LEAR` for risk appetite $\alpha = 0.6$ (center panel) and $\alpha = 0.8$ (bottom panel)

For instance, the most accurate in terms of the CRPS ensemble, i.e., `DDNN_JSU_CRPS_LEAR`, yields lower profits than its `qEns` counterpart. This is likely due to the fact that while the CRPS weighting is more accurate on average, it is significantly outperformed by the naive weighting for the few lowest percentiles, see Figure 5, giving the latter an advantage during the initial stage of the COVID-19 pandemic (middle part of the test period), see Figure 7. A similar behavior can be observed for the

DDNN_JSU_CRPS_DNN ensemble, which performs worse than its competitors for the extreme quantiles, being at a significant disadvantage in the second half of the evaluation period, when the daily price spread is higher than in the beginning.

It can be expected that an optimal trading strategy would result in executing exactly two trades per day, buying on the low and selling on the high. With such a "crystal ball" strategy, the trader would earn 13,587 € throughout the whole evaluation period. Conversely, taking the worst possible decisions would lead to a total loss of $-21,425$ €. On this scale of possible profits, all evaluated ensembles rank relatively well. The lowest profit presented in Table 1 reaches 80% of the maximum, while the best of all forecasts as much as 96%. For comparison, a naive strategy of placing bids at fixed hours selected *ex-post* as having the highest price spread on average – buying at hour 3 and selling at hour 19 – would lead to total profits of 8048 €, or 84% of the maximum, see [20].

Table 2. Profits per trade from the quantile-based trading strategy in the whole test period for risk appetite ranging from 0.5 to 0.9. The highest values in each column are in bold. Cells are colored independently in each column from the best (→ green) to the worst (→ red).



	Risk appetite				
Model	0.5	0.6	0.7	0.8	0.9
DDNN_N_CRPS	11.20	11.29	11.56	11.77	12.78
DDNN_N_qEns	11.32	11.46	11.65	11.85	13.25
DDNN_N_qEns_DNN	11.23	11.33	11.61	12.12	**14.42**
DDNN_N_CRPS_DNN	11.16	11.36	11.52	12.12	14.07
DDNN_N_qEns_LEAR	11.60	11.60	11.77	12.38	13.88
DDNN_N_CRPS_LEAR	11.40	11.44	11.60	11.92	13.03
DDNN_JSU_qEns	11.57	11.81	11.91	12.33	13.96
DDNN_JSU_CRPS	11.33	11.56	11.75	12.06	13.57
DDNN_JSU_CRPS_DNN	11.16	11.42	11.70	12.25	13.55
DDNN_JSU_qEns_DNN	11.22	11.48	11.71	12.34	14.25
DDNN_JSU_qEns_LEAR	**11.72**	**11.89**	**11.95**	**12.47**	13.99
DDNN_JSU_CRPS_LEAR	11.67	11.76	11.83	12.08	13.40

The profit per trade results reported in Table 2 are less clear-cut, with the DDNN_N ensembles no longer being completely outclassed by the DDNN_JSU ensembles, especially when there are fewer total trades. As profits per trade can be seen as an indicator of the trader's risk, this disparity is consistent with literature findings [12]. It is worth noticing that, perhaps unintuitively, higher values of the risk appetite correspond to higher risk aversion. This is an effect of the final step of the strategy, which checks the income of the worst case scenario. With higher values of the risk appetite, this predicted profit tends to be lower, leading the trader to act more cautiously. This seemingly leads to a dominance of models which are more accurate across the entire distribution rather than only in the extreme quantiles, compare the center (risk appetite $\alpha = 0.6$) and bottom (risk appetite $\alpha = 0.8$) panels of Figure 7 with Figure 5. However, this is only a conjecture, as the relationship between daily price levels, price spreads, and PIs is not direct or linear.

5. Conclusions and discussion

In this article, we address the question of whether minimizing the continuous ranked probability score (CRPS) – the standard error metric for probabilistic forecasts – leads to optimal decisions in day-ahead bidding. Conducting an extensive empirical study, we find that introducing diversity to a pool of forecasts is highly beneficial, both in terms of forecast accuracy measured by the CRPS and profits from a trading strategy implemented in the German day-ahead power market. Also optimizing combination weights with CRPS learning positively impacts forecast accuracy. This is likely caused by the uneven performance of experts across time and quantiles, which is an outcome consistent with the literature.

While trading profits generally follow forecast accuracy, the benefits of using CRPS learning are not as pronounced in the trading scenario, especially considering the ca. 500 times higher computational burden. The precise cause-and-effect relationships between the predictive accuracy and profits are difficult to disentangle. However, the performance for the extreme quantiles of the distribution seems to be related to some of the observed patterns. In general, using any of the considered ensembles leads to achieving satisfactory profits, especially when compared to the best-case and worst-case scenarios.

For the sake of clarity, only selected forecast sets were considered. Extending the pool of experts and ensembles could lead to a more comprehensive evaluation. Other possible extensions of this study include comparisons with other weighting schemes, as well as automated methods for expert selection.

Acknowledgement

This research was partially supported by the Ministry of Science and Higher Education (MNiSW, Poland) through Diamond Grant No. 0027/DIA/2020/49 (to W.N.) and the National Science Center (NCN, Poland) through grant No. 2018/30/A /HS4/00444 (to R.W.).

References

[1] Berrisch, J., and Ziel, F. CRPS learning. *Journal of Econometrics* (2021), 105221. DOI: 10.1016/j.jeconom.2021.11.008.

[2] Berrisch, J., and Ziel, F. Multivariate probabilistic CRPS learning with an application to day-ahead electricity prices, 2023. DOI: 10.48550/arXiv.2303.10019. Working paper version available from arXiv: https://arxiv.org/abs/2303.10019.

[3] Berrisch, J., and Ziel, F. *The profoc Package: An R package for probabilistic forecast combination using CRPS Learning*, 2023. R package version 1.2.0.

[4] Blanc, S. M., and Setzer, T. Bias–variance trade-off and shrinkage of weights in forecast combination. *Management Science 66*, 12 (2020), 5720–5737.

[5] Diebold, F. X., and Mariano, R. S. Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*, 3 (1995), 253–263.

[6] Gneiting, T. Making and evaluating point forecasts. *Journal of the American Statistical Association 106*, 494 (2011), 746–762.

[7] Gneiting, T., and Katzfuss, M. Probabilistic forecasting. *The Annual Review of Statistics and Its Application 1* (2014), 125–151.

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

[9] Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., and Hyndman, R. J. Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. *International Journal of Forecasting 32*, 3 (2016), 896–913.

[10] Hong, T., Pinson, P., Wang, Y., Weron, R., Yang, D., and Zareipour, H. Energy forecasting: A review and outlook. *IEEE Open Access Journal of Power and Energy 7* (2020), 376–388.

[11] Hubicka, K., Marcjasz, G., and Weron, R. A note on averaging day-ahead electricity price forecasts across calibration windows. *IEEE Transactions on Sustainable Energy 10*, 1 (2019), 321–323.

[12] Janczura, J., and Puć, A. ARX-GARCH probabilistic price forecasts for diversification of trade in electricity markets—variance stabilizing transformation and financial risk-minimizing portfolio allocation. *Energies 16*, 2 (2023).

[13] Janczura, J., and Wójcik, E. Dynamic short-term risk management strategies for the choice of electricity market based on probabilistic forecasts of profit and risk measures. The German and the Polish market case study. *Energy Economics 110* (2022), 106015.

[14] Lago, J., Marcjasz, G., De Schutter, B., and Weron, R. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy 293* (2021), 116983.

[15] Lichtendahl, K. C., Grushka-Cockayne, Y., and Winkler, R. L. Is it better to average probabilities or quantiles? *Management Science 59*, 7 (2013), 1594–1611.

[16] Maciejowska, K. Portfolio management of a small RES utility with a structural vector autoregressive model of electricity markets in Germany. *Operations Research and Decisions 32*, 4 (2022), 75-90.

[17] Maciejowska, K., Nitka, W., and Weron, T. Enhancing load, wind and solar generation for day-ahead forecasting of electricity prices. *Energy Economics 99* (2021), 105273.

[18] Maciejowska, K., Uniejewski, B., and Weron, R. Forecasting electricity prices. In *Oxford Research Encyclopedia of Economics and Finance*. Oxford University Press, 2023. DOI: 10.1093/acrefore/9780190625979.013.667. Working paper version available from arXiv: https://doi.org/10.48550/arXiv.2204.11735.

[19] Makridakis, S., Spiliotis, E., and Assimakopoulos, V. The M4 Competition: Results, findings, conclusion and way forward. *International Journal of Forecasting 34*, 4 (2018), 802–808.

[20] Marcjasz, G., Narajewski, M., Weron, R., and Ziel, F. Distributional neural networks for electricity price forecasting. *Energy Economics 125* (2023), 106843.

[21] Marcjasz, G., Uniejewski, B., and Weron, R. Probabilistic electricity price forecasting with NARX networks: Combine point or probabilistic forecasts? *International Journal of Forecasting 36*, 2 (2020), 466–479.

[22] Nowotarski, J., and Weron, R. Computing electricity spot price prediction intervals using quantile regression and forecast averaging. *Computational Statistics 30*, 3 (2015), 791–803.

[23] Nowotarski, J., and Weron, R. Recent advances in electricity price forecasting: A review of probabilistic forecasting. *Renewable and Sustainable Energy Reviews 81* (2018), 1548–1568.

[24] Timmermann, A. Forecast Combinations. In *Handbook of Economic Forecasting*, G. Elliott, C. W. J. Granger, and A. Timmermann, Eds., vol. 1. Elsevier, 2006, pp. 135–196.

[25] Uniejewski, B. Smoothing Quantile Regression Averaging: A new approach to probabilistic forecasting of electricity prices, 2023. DOI: 10.48550/arXiv.2302.00411. Working paper version available from arXiv: https://arxiv.org/abs/2302.00411.

[26] Uniejewski, B., and Maciejowska, K. Lasso principal component averaging: A fully automated approach for point forecast pooling. *International Journal of Forecasting* (2022). DOI: 10.1016/j.ijforecast.2022.09.004, (in press).

[27] Uniejewski, B., and Weron, R. Regularized quantile regression averaging for probabilistic electricity price forecasting. *Energy Economics 95* (2021), 105121.

[28] Vogler, A., and Ziel, F. Event-based evaluation of electricity price ensemble forecasts. *Forecasting 4*, 1 (2022), 51–71.

[29] Wang, X., Hyndman, R. J., Li, F., and Kang, Y. Forecast combinations: An over 50-year review. *International Journal of Forecasting* (2022). DOI: 10.1016/j.ijforecast.2022.11.005, (in press).

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

[31] Yardley, E., and Petropoulos, F. Beyond error measures to the utility and cost of the forecasts. *Foresight: The International Journal of Applied Forecasting 63* (2021), 36–45.