Politechnika Wrocławska

FIELD OF SCIENCE: Social Sciences

DISCIPLINE OF SCIENCE: Management and Quality Studies

DOCTORAL DISSERTATION

Forecasting wholesale electricity prices to support decision-making in power companies: Use of regularization and forecast combinations

Bartosz Uniejewski

Supervisor:
prof. dr hab. Rafał Weron

Assistant supervisor:
dr Katarzyna Maciejowska

Keywords: Decision-making, Electricity price forecasting, Energy market, Forecast combinations, Probabilistic forecasting, Regularization

WROCŁAW 2023

Abstract

This doctoral thesis develops robust electricity price forecasting techniques to aid decision-making in power markets. The research focuses on five interconnected objectives, all utilizing regularization techniques. Firstly, a LASSO-type regularization approach is employed to identify the most relevant explanatory variables. Secondly, a fully automated approach is developed to average a pool of individual forecasts using PCA and LASSO. Thirdly, quantile regression and regularization are utilized to construct more accurate algorithms for probabilistic forecasting of electricity prices. Fourthly, a trading strategy is designed to evaluate the economic value of probabilistic forecasts. Finally, a critical review of forecasting in electricity markets is conducted, and an outlook for future research is provided. The proposed solutions significantly outperform existing literature benchmarks, and the thesis sets up directions for future research in the field.

Streszczenie

Niniejsza praca doktorska dotyczy opracowania odpornych technik prognozowania cen energii elektrycznej na potrzeby wspomagania podejmowania decyzji na rynkach energii. Badania koncentrują się na pięciu celach, których wspólnym mianownikiem jest technika regularyzacji: (i) wykorzystanie regularyzacji typu LASSO do identyfikacji najważniejszych zmiennych objaśniających, (ii) opracowanie w pełni zautomatyzowanego podejścia do uśredniania puli indywidualnych prognoz za pomocą PCA oraz LASSO, (iii) wykorzystanie regresji kwantylowej i regularyzacji do konstrukcji bardziej dokładnych algorytmów probabilistycznego prognozowania cen energii elektrycznej, (iv) zaprojektowanie strategii handlowej do oceny wartości ekonomicznej prognoz probabilistycznych, (v) przeprowadzenie krytycznego przeglądu metod prognozowania na rynkach energii elektrycznej oraz przedstawienie perspektyw przyszłych badań. Proponowane rozwiązania znacznie poprawiają dokładność prognoz cen w porównaniu z modelami istniejącymi w literaturze, a praca doktorska wyznacza kierunki dla przyszłych badań w tym obszarze.

Contents

Chapter 1

Introduction

1.1 Background

The electricity market is one of the most important, as well as one of the most unique among all commodity markets. Since the deregulation of the government-controlled power systems in the UK and Scandinavia in the early 1990s, electricity has been traded under competitive rules in many countries worldwide (Mayer and Trück, 2018). Research on *electricity price forecasting* (EPF) has become extremely important and valuable, as price predictions are an essential input to decision support and risk management systems in energy companies (Weron, 2014).

However, forecasting electricity prices is more challenging than those of other commodities or financial assets. The uniqueness of the power market is primarily due to the lack of efficient ways to store large volumes of energy (Weron, 2006). Consequently, the system has to be balanced at any given moment, i.e., all energy produced must be consumed. Over the past few years, many authors have reported an increase in the share of electricity generated from *renewable energy sources* (RES) (Papież et al., 2018), which is more challenging to plan and manage (Kiesel and Paraschiv, 2017; Maciejowska, 2020). This leads to non-intuitive situations when negative prices are recorded (Ziel and Steinert, 2018). If due to inadequate planning or technical limitations, generators deliver more energy than demanded, they may have to pay consumers for the collection of electricity. Conversely, in the presence of an abnormal surge in demand or a power plant or transmission grid failure, back-up units with much higher marginal costs may be called upon to match demand, leading to so-called price spikes (Karakatsani and Bunn, 2008; Sadowski et al., 2012).

An additional factor that makes it difficult to accurately predict electricity prices is the dependence of both production and consumption on weather and the intensity of human activities (Bunn, 2000). Depending on the season, day of the week, and time of the day, the demand for electricity changes. This means that, as Paraschiv (2013) and Weron (2006) argue, when analyzing the time series of electricity prices, we should distinguish three seasonal components at different levels – daily, weekly and annual.

Over the past few years, a market that was already volatile has become even more unstable. Rapid changes in electricity supply and demand caused by the COVID-19 pandemic and, even more, by the unstable political situation in Europe keep market participants awake at night. The electricity market remains a dynamic and complex environment, and forecasting prices is an ongoing challenge.

1.2 Aim and objectives

The aim of this thesis is to develop robust and efficient electricity price forecasting techniques to support decision-making in power companies. To address this aim, five objectives are set:

1. Use regularization to identify the most relevant predictors.
2. Develop a fully-automated approach to average a rich pool of individual forecasts using regularization and principal component analysis.
3. Utilize quantile regression and regularization to construct more accurate algorithms for probabilistic forecasting of electricity prices.
4. Design a trading strategy to evaluate the economic value of probabilistic forecasts.
5. Conduct a critical review of forecasting in the electricity markets and provide an outlook for future research in this area.

These objectives are interesting not only from the point of view of basic research but also from a managerial perspective. On the one hand, they develop and validate statistical learning techniques. On the other hand, they yield trading strategies to support decision-making.

Objective 1 is to identify the most important predictors of electricity prices using the *least absolute shrinkage and selection operator* (LASSO), which helps to build well-structured forecasting models without the need for expert knowledge. Objective 2 is to develop a fully automated method for combining a large pool of forecasts using *principal component analysis* PCA and LASSO. Objective 3 is to develop a novel, regularization-based approach for constructing probabilistic predictions, which outperform existing benchmarks.

The first three objectives have one more common factor. All of them use regularization, a statistical learning technique, to solve a range of problems in forecasting. In this thesis, I provide evidence that statistical learning techniques can be used to: identify the most relevant variables driving intraday prices (Objective 1), combine forecasts (Objective 2), and obtain probabilistic predictions (Objective 3).

In the vast majority of the literature measures based on absolute or squared errors are by far the most popular methods to evaluate forecasts. However, in practice, the aim is to support decision-making and maximize profits. Hence, Objective 4 is concerned with developing a new approach to evaluating forecasts, which focuses on the financial value of the decisions based on price predictions.

It should be noted that the objectives are complementary. The three trends discussed in the critical review (Objective 5) are visible in Objectives 1-4. To propose a solution for each of the objectives, joint research with at least one of the supervisors was carefully designed, performed, and reported in premier publishing outlets.

The remainder of the thesis is structured as follows. In Chapter 2 I discuss how my thesis contributes to the discipline of *Management and Quality Studies*. Next, in Chapter 3, which is an excerpt from Paper 5, I describe the marketplace. In Chapter 4, I summarize the introduced concepts, methods and models, as well as the results obtained in Papers 1-5 which constitute the core part of the thesis. The latter addresses the above five objectives.

Namely, in Section 4.1, I emphasize the importance of regularization in the development of automated variable selection models and in Section 4.2 I present the key finding in the context of the combining forecasts. Next, in Section 4.3, I discuss the development of probabilistic forecasting methods and in Section 4.4 I introduce a new approach to the economic evaluation of probabilistic forecasts. Finally, in Section 4.5 I comment on the forward-looking trends put forward in Paper 5. Next, in Chapter 5, I briefly discuss the articles I have published in the course of my undergraduate and graduate studies that do not constitute the core part of the thesis. Finally, in Chapter 6, I summarize the key findings and conclude. The five core articles, i.e., Papers 1-5, can be found in the Appendix.

Chapter 2

Contribution to the discipline of Management and Quality Studies

2.1 Forecasting as a decision support tool

To place this thesis in the discipline of *Management and Quality Studies* we have to start by noting that forecasting electricity prices is a subfield of *predictive analytics*. The latter, along with *descriptive* and *prescriptive analytics*, constitutes so-called *business analytics* (Lepenioti et al., 2020). Business analytics lies at the intersection of data science and operations research, and involves not only analysis but also synthesis and implementation (Delen and Ram, 2018; Rose, 2016).

When referring to decision-making, we usually think of the final act of choosing one among several options. In fact, the decision-making process is a challenging task that requires a number of activities (Elbanna, 2006). The initial stage of the decision-making process is the definition of the problem. Subsequently, we identify the possible decisions to be made, the uncertain future events (so-called *chance events*), and the associated impacts of each alternative decision and each outcome of the chance event (Anderson et al., 2015). According to the Noble Prize winner Herbert Alexander Simon (1960)[1], any management decision is composed of three principal phases:

- intelligence activity, which involves searching for environment and conditions (opportunities) for making a decision,
- design activity, i.e., inventing, developing, and analyzing possible courses of action,
- choice activity, i.e., selecting one alternative from several options.

Actually, Simon (1960) suggested that the decision-making process is far more demanding than the proposed sequence – each phase itself requires a complex decision-making process. He pointed out that dealing with a problem at any level of this sequence generates another sub-problem that requires intelligence, design, and choice activity. Kamiński et al. (2018) further elaborate that, in reality, managers must consider multi-stage decisions, involving several consecutive actions (decisions) and note that company profits depend not only on the actions of managers but also on exogenous events, which are random.

[1] In 1978 for 'his pioneering research into the decision-making process within organizations'.

More recently, Heizer et al. (2004) stated that decision-making follows a process of defining the problem, developing objectives, creating a model, evaluating alternatives, selecting the best solution, and implementing the decision. This process requires data and quantifiable variables, which can be difficult to obtain due to either lack or abundance of data.

Although forecasting is not explicitly mentioned in the three decision-making phases outlined by Simon (1960) or the six-step process outlined by Heizer et al. (2004), it is involved in the design activity and evaluating alternatives stages. Forecasting plays a critical role in assessing potential courses of action, rating considered options, and ultimately selecting the best solution. The decision maker has to predict the outcome of each possible option, which is a challenging task given the uncertain nature of the future. A misjudgment of the situation can lead to wrong decisions and financial losses.

In a recent encyclopedic article, Petropoulos et al. (2022) emphasize that forecasting has always been at the leading edge of decision-making. Nearly all decisions require some form of forecasting to support the process (Slack et al., 2022). However, the impact of forecasting on the decision-making process varies depending on the level of risk associated with a particular choice. Heizer et al. (2004) describe three distinct types of decision-making environments:

- decision-making under uncertainty,
- decision-making under risk,
- and decision-making under certainty.

Decision-making under uncertainty takes place when the set of alternatives is not fully known and the decision-maker cannot assign risk to each of them. In practice, the most common are decisions made under risk. Here, the potential payoffs or costs are associated with scenarios and can be described with probability distributions. Lastly, decision-making under certainty occurs when there is certain information about the outcome (Heizer et al., 2004).

In the context of power markets, decisions are typically made under risk. Market participants face many decision-making problems related to power plant operations and energy trading. For example, a company operating a conventional power plant has to decide about the level of production at least a day before the delivery. A different problem occurs when we consider the daily operation of renewable-energy power plants. In this case, the level of production depends on the weather conditions, so it remains unknown until a very short time before delivery. The power plant operator must predict the amount of energy that will be produced in a given hour and decide how much and where to sell it, such as through futures contracts or in the day-ahead (DA) or intraday (ID) market.

The decision-making process for the electricity market is even more challenging and risk exposed compared to other markets (Weron, 2014). On the one hand, the extraordinary variation of prices gives market participants the opportunity to make immense profits. It is important to note, that the forward contracts in the electricity market are not a remedy to the corresponding risk, as the actual demand remains unknown until a very short time before delivery (Wilson, 2002).

The main focus of this thesis, and at the same time a key factor in decision-making in power markets, is forecasting day-ahead electricity prices. This auction-based market plays a crucial role in ensuring the reliability and efficiency of the electricity system. The day-ahead market not only gives the opportunity to adjust the long-term position to the actual exposure (Kath and Ziel, 2018; Mayer and Trück, 2018; Maciejowska et al., 2019; Janczura and Wójcik, 2022), but also is a reference point for *over-the-counter* (OTC) trading and settlement.

Accurate day-ahead price forecasts are critical for market participants to make informed decisions about their bidding strategies (Narajewski and Ziel, 2022). If participants have a reliable price forecast, they can adjust their bids accordingly and increase their chances to execute profitable transactions. In addition, price forecasts can help participants to estimate their costs and revenues more accurately, which is important for their financial planning and risk management.

Each day, energy companies must make decisions and take actions that determine whether they will succeed and make profits or fail and suffer losses. Recently, more and more authors (Delarue et al., 2010; Zareipour et al., 2010; Maciejowska et al., 2019) have been trying to capture the economic value of forecasts. Researchers understand that forecasts, to be more useful, must better support decision-making and yield higher profits or reduce losses rather than minimize some statistical error measures. Although the economic impact of improving electricity price predictions is not easy to estimate, Hong (2015) concluded that the savings from a 1% reduction of errors (in terms of mean absolute percentage error) in short-term load and price forecasting translated into $600 thousand savings per year for a 1 GW peak load utility.

2.2 Improving forecast accuracy

In this thesis, the main focus is on the development of new statistical tools used for forecasting electricity prices. In power markets, statistical methods such as regression models are commonly used to forecast day-ahead electricity prices (Weron, 2014). These models use historical prices and a combination of forecasts of consumption and production figures, and weather variables to predict the future price. Regressors are selected from a set of explanatory variables assumed to be correlated with electricity prices, based on in-sample analysis. These models have the advantage of being able to provide an interpretation of their components, allowing us to better understand their behavior.

The main factor determining the success of a predictive model is its structure, i.e., the set of explanatory variables and the method used to estimate the weights corresponding to each variable. Moreover, the EPF literature provides many tweaks and tricks to develop more precise forecasting methods. Data preprocessing techniques, such as transformations (Schneider, 2011; Diaz and Planas, 2016; Uniejewski et al., 2018; Shi et al., 2021) or deseasonalization (Paraschiv, 2013; Nowotarski and Weron, 2016a; Lisi and Pelagatti, 2018; Jędrzejewski et al., 2021), as well as careful selection of the model calibration window (Marcjasz et al., 2018; Hubicka et al., 2019; Fezzi and Mosetti, 2020) turn out to be very useful.

The electricity price forecasting community is constantly coming up with new solutions to improve forecasting accuracy and provide managers with even better tools to support the decision-making process. Recently, approaches based on regularization have gained a lot of attention. Regularization is a very general technique that can be used to solve many research problems. The idea aims to improve the quality of parameter estimators by imposing a penalty function on the original model (Tikhonov, 1963). It can be used not only to predict prices (Ziel, 2016), but also to identify the most important regressors (Ziel and Weron, 2018; Uniejewski and Weron, 2018). With the help of regularization, it is also possible to average forecasts obtained with different models (Diebold and Shin, 2019). Finally, based on the idea of regularization, new methods can be created by formulating original variants of the penalty function or by using known penalty functions for other underlying models (Li and Zhu, 2008).

The rapid development of forecasting methods has led to the availability of countless solutions for obtaining predictions. Therefore, it is common practice to use multiple forecasting methods and combine their predictions to improve overall accuracy. The idea of forecast averaging assigns weights to a set of individual forecasts and then takes the weighted sum as the final prediction. In practice, the quality of predictions obtained with a given forecasting model is unknown until the real price is observed, thus it is nearly impossible to select the best-performing model ex-ante. The predictions obtained with the combinations of forecasts are much more reliable compared to those obtained with one selected model (Hibon and Evgeniou, 2005). With the abundance of data and the improvement of forecasting models, averaging techniques must cope with a very large set of possible inputs. One way to address this issue is to utilize PCA, which can successfully reduce the dimensionality of the problem (Huang and Lee, 2010; Maciejowska et al., 2020).

According to Sadowski (1980), decision-making involves uncertainty as the final outcome of a decision. Forecasting factors that influence the decision can provide additional information and reduce uncertainty. Moreover, it is important for decision-makers to understand the likelihood of the forecast in order to use it effectively. Recently, the field of electricity price forecasting has shifted towards the use of probabilistic predictions. This new approach offers a more powerful tool for decision-makers to manage risk. Instead of predicting a single expected value for future prices, probabilistic forecasts provide the whole distribution of the future price or at least an interval in which the price is likely to fall with a specified probability. This allows for risk analysis in decision-making and helps to prevent extreme price fluctuations and their associated costs (Morales et al., 2014).

Chapter 3

The marketplace[1]

As a result of the $(\ldots)$ liberalization and deregulation of the power sectors, two basic models for power markets have emerged: power pools – where trading, dispatch and transmission are managed by the *system operator* (SO), and power exchanges – where trading and initial dispatch are managed by an institution independent from the *transmission system operator* (TSO). Participation in power pools is limited to generators and is typically mandatory. The *market clearing price* (MCP) is established through a one-sided auction as the intersection of the supply curve constructed from aggregated supply bids of the generators and the demand predicted by the system operator. Often a separate price for each node in the network is calculated, so-called *locational marginal price* (LMP). Such a system was adopted in highly meshed North American networks. On the other hand, in Australia, where the network structure is simpler, zonal pricing was successfully implemented, where for areas without grid limitations a unique price is settled.

In contrast to power pools, participation in power exchanges is – except for some special cases – voluntary and open not only to generators, but also to wholesale consumers and speculators. The price is established either through a two-sided auction (DA, ID) as the intersection of the supply curve constructed from aggregated supply bids and the demand curve constructed from aggregated demand bids or in continuous trading (ID). Most market designs have adopted the uniform-price auction, where buyers who bid at or above the MCP pay that price and sellers who bid at or below the MCP are paid this price. Moreover, in auction markets the bids can be submitted until a certain time – called *gate closure* – which is the same for all load periods, see the left panel in Fig. 3.1. Hence, auction prices could be viewed as realizations of a multivariate random variable and therefore prices for all load periods should be predicted simultaneously (Ziel and Weron, 2018). On the other hand, some ID markets allow for continuous trading. They run 24/7 from an afternoon hour on day $d-1$ up until a few minutes before the delivery of electricity during a particular load period on day d, see the right panel in Fig. 3.1.

In some countries (e.g., Germany, Ireland, Poland) the DA and ID markets are complemented by the so-called balancing market. This technical market is used for pricing differences between the market schedule and actual system demand for very short time horizons before delivery. For instance, the TSO might instruct a generator to increase its output to meet a sudden surge in demand. The producer then receives a premium via the balancing market for the energy generated used to balance the grid.

The timeline of day-ahead and intraday trading activities in selected European coun-

[1] This Chapter is an excerpt from Section *The Marketplace* in Paper 5

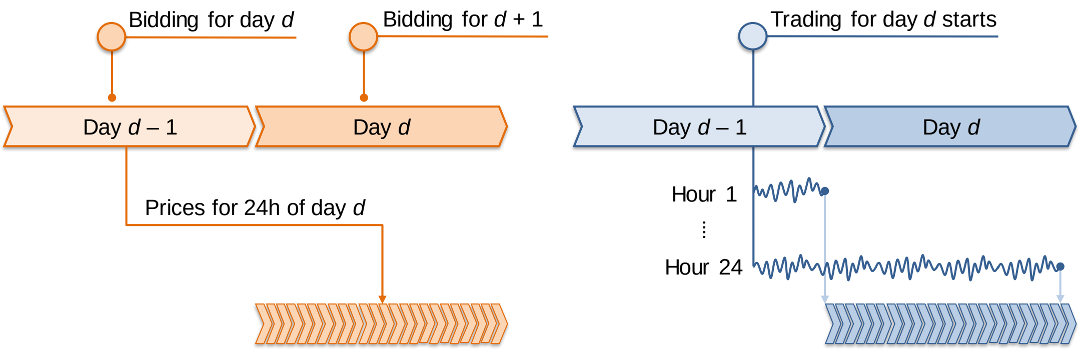


Figure 3.1: Illustration of bidding and price settlement in auction (*left*) and continuous trading (*right*) power markets. In day-ahead auctions the bids for all load periods (here: hours) of day d can be submitted until a certain hour on day $d-1$. Intraday markets which admit continuous trading run 24/7 from an afternoon hour on day $d-1$ up until a few minutes before the delivery on day d. Source: Paper 5.

tries is illustrated in Fig. 3.2. As can be seen, the DA and ID markets complement each other. Once the gate closes for day-ahead bids around noon, various intraday markets open for adjusting these bids. They are particularly important for nondispatchable, stochastic producers such as wind or solar farms, and include both auctions and continuous trading. Note that both the ID and DA contracts can concern delivery during the same load period, only the time the decision has to be made and the bid placed differs.

The presented sequence of events has important implications for study design. In the DA market the forecasting horizons typically range from 12-14 hours for the first load period of the next day to 36-38 hours for the last. However, at the time the predictions are made, i.e., the morning hours of day $d-1$, the DA prices for all load periods of this day are already known (they were settled around noon on day $d-2$). Generally, the TSO

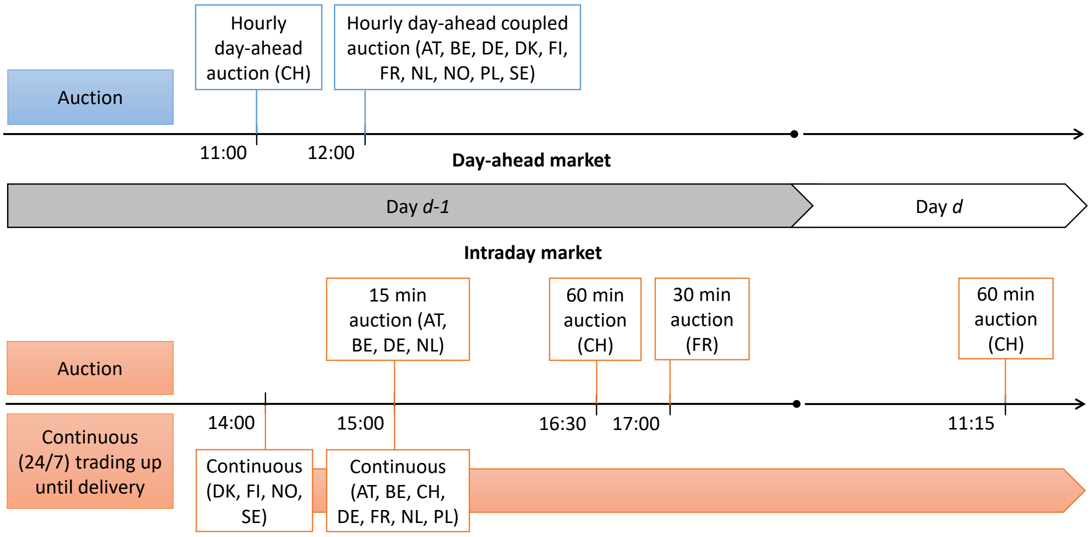

Hourly day-ahead auction (CH)
Hourly day-ahead coupled auction (AT, BE, DE, DK, FI, FR, NL, NO, PL, SE)
Auction
11:00
12:00
Day-ahead market
Day *d-1*
Day *d*
Intraday market
15 min auction (AT, BE, DE, NL)
60 min auction (CH)
30 min auction (FR)
60 min auction (CH)
Auction
14:00
15:00
16:30
17:00
11:15
Continuous (24/7) trading up until delivery
Continuous (DK, FI, NO, SE)
Continuous (AT, BE, CH, DE, FR, NL, PL)

Figure 3.2: The timeline of day-ahead (*top*) and intraday (*bottom*) trading activities for delivery of electricity on day d in selected European countries: Austria (AT), Belgium (BE), Denmark (DK), Germany (DE), Finland (FI), France (FR), the Netherlands (NL), Norway (NO), Poland (PL), Sweden (SE) and Switzerland (CH). Source: Paper 5.

day-ahead forecasts of the system load ($\approx$ demand) and the system-wide generation from *renewable energy sources* (RES) are also available to market participants at this time.

When the ID market is considered, the selection of the forecasting horizon depends on the research question. Firstly, the predictions can be made on the morning of day $d-1$, when market participants need to decide how much electricity to bid in the DA market and how much to buy/sell in the ID market or leave for the balancing market. Forecasts of the price spread between DA and ID/balancing markets can provide valuable insights for decision-making (Maciejowska et al., 2019, 2021).

Secondly, the predictions can be used for bidding in ID markets with continuous trading. Although the trading floor opens in the afternoon hours of day $d-1$, the majority of bids are placed during the last 3-4 hours before the delivery (Narajewski and Ziel, 2020a). Hence, the forecasting horizons considered typically range from a couple of minutes to 4 hours (Janke and Steinke, 2019; Uniejewski et al., 2019b; Narajewski and Ziel, 2020b). Note that different model specifications may be optimal for predicting ID prices for different horizons (Maciejowska et al., 2020). Since the bidding behavior of market participants is significantly influenced by RES generation forecasts which are available at the time of trading (Kiesel and Paraschiv, 2017; Kulakov and Ziel, 2021), ID price forecasts should not only exploit the short-term price dependencies but also updated predictions of wind and solar power generation. Interestingly, including self-exciting terms in ID models allows to better capture the empirically observed trade clustering (Kramer and Kiesel, 2021).

Chapter 4

Summary of results

4.1 Objective 1: Regularization and variable selection

4.1.1 Problem statement

In practical applications, a predictive model not only has to perform well in terms of accuracy but also has to be relatively easy to interpret. Moreover, when it comes to decision support, it is not just the accuracy of the prediction that matters, but also the time it takes to get it. To address these issues, statistical learning techniques, such as regularization, are deployed. They allow designing sparse models that are much easier to interpret and faster to estimate than dense multi-parameter models without a significant loss in forecast accuracy (James et al., 2013).

The idea of regularization is formally defined as follows:

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}\left\{f(\boldsymbol{X};\boldsymbol{\beta}) + g(\boldsymbol{\beta})\right\}, \tag{4.1}$$

where $f(X)$ is the function minimized by the 'classic' model, e.g., the *residual sum of squares* (RSS) in a regression, $g(\boldsymbol{\beta})$ is the penalty function (the so-called *regularization term*), $\boldsymbol{\beta}$ is the parameter vector and $\hat{\boldsymbol{\beta}}$ is its estimator.

The variant of regularization most frequently used defines the penalty function as a norm of order ℓ^q scaled by so-called *tuning parameter* λ. In this case, the coefficients of a regularized linear regression model are estimated by (Hastie et al., 2015):

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}\left\{\underbrace{(\boldsymbol{y}-\boldsymbol{\beta}\boldsymbol{X})^2}_{\text{RSS}} + \lambda\|\boldsymbol{\beta}\|_q^q\right\} = \operatorname{argmin}\left\{\text{RSS} + \lambda\sum_{i=1}^{n}|\beta_i|^q\right\}. \tag{4.2}$$

If we choose the penalty function appropriately, regularization becomes a tool for identifying the most significant variables in the model. A regularization procedure fits the full model with all predictors using an algorithm that shrinks coefficients of the less important explanatory variables towards zero. If we take the norm of order ℓ^1 as a regularization term, then we obtain the least absolute shrinkage and selection operator of Tibshirani (1996). The value λ in Eq. 4.2 indicates how significant the variables have to be to remain in the final model. While for $\lambda = 0$ the method reduces to *ordinary least squares* (OLS), as the parameter increases, more and more variables are considered irrelevant and eliminated from the model. An advantage of using methods of automatic variable selection is the almost unlimited size (number of explanatory variables) of the base model. Due to this property, the knowledge of experts, which is often not verified, becomes less important (Uniejewski et al., 2016).

The key point in EPF is a careful selection of explanatory variables (Dudek, 2016). Some of the first examples of formal variable selection in electricity price forecasting include Karakatsani and Bunn (2008) and Misiorek (2008), who used stepwise regression to eliminate statistically insignificant variables in parsimonious *autoregressive* (AR) and regime-switching models for individual load periods. Regularization techniques appeared in the next decade. In 2015, Ludwig et al. (2015) utilized the LASSO model as a feature selection tool to select relevant weather stations from the total of 77 available stations. The following year, Ziel (2016) employed the LASSO method to simplify very large sets of model parameters, consisting of over 100. The author used time-varying coefficients to capture the dependency structure within a day and employed a set of 24 regression models for the 24 hours of the day. Uniejewski et al. (2016) conducted an extensive study to compare automatic variable selection models and suggested that LASSO and elastic nets (Zou and Hastie, 2015) significantly outperformed their competitors.

Automatic variable selection is particularly valuable when the goal is to forecast prices in a market that has not yet been thoroughly studied. However, recently also for well-researched markets variable selection methods came in handy. As more data becomes available, models based on expert knowledge are no longer sufficient and more complex ones are required to produce accurate forecasts (Jędrzejewski et al., 2022).

4.1.2 Understanding intraday electricity markets: Variable selection and very short-term price forecasting using LASSO (Paper 1)

Knowledge of the market fundamentals and model-building experience are crucial in constructing a well-performing predictive model. That is why dealing with a new market setup is always a challenge for forecasters. In Paper 1 we study the German intraday market. This paper was the first to predict prices in this market, to be more precise, predict the so-called ID3 price index (Narajewski and Ziel, 2020a). To our best knowledge, at the time, only two articles addressed forecasting intraday electricity prices in European power markets. Both of them focused solely on the Iberian market (Andrade et al., 2017; Monteiro et al., 2016).

In this paper, we use LASSO to identify the most relevant variables among 349 to 372 (depending on the forecasted hour) potential predictors. We consider 12 models: a naive benchmark, a parsimonious structure inspired by a well-performing autoregressive EPF model, and LASSO with ten different values of the tuning parameter λ. The performed analysis yields tables with variables most frequently selected by LASSO (see Tables 3 and 4 in Paper 1). We also use regularization to build parsimonious, well-performing autoregressive models with exogenous variables (ARX). For this purpose, we utilize an additional 364-day rolling selection window and include in the final ARX model only those variables that are selected more often than a certain threshold. We test five different cut-off values ranging from 50% to 90%.

We rank the models on the basis of the *mean absolute errors* (MAE). To assess the significance of differences in prediction accuracy, we use the Diebold and Mariano (1995) test.

The key contribution of this paper is the development of the first model to forecast the German ID3 price index and the identification of the most important predictors. The most recent ID3 value and the day-ahead (DA) price with the same delivery time turn out to be the most important explanatory variables. In addition, intraday and day-ahead prices for late evening hours appear to have a significant impact on future prices. Surprisingly, unlike the forecasting models for the day-ahead, neither the price on the previous day nor the weekly dummies are found to be important predictors. Finally, we show that LASSO can be successfully used for variable selection. The parsimonious ARX model built using variables that remain significant at least 70% of the time, performs equally well as the best LASSO model. This result provides evidence that LASSO is a powerful tool for building predictive models, especially when the market is not yet fully researched.

Publication details:

- Authors: B. Uniejewski, G. Marcjasz, R. Weron
- Journal: *International Journal of Forecasting* 35, 1533-1547
- DOI: 10.1016/j.ijforecast.2019.02.001
- Publication year: 2019
- JCR Impact Factor: 7.022
- JCR classification(s): *Economics*, *Management*
- MEiN: 140 pts, assigned to the Management and Quality Studies (NZJ) discipline
- Contribution: 50%, including participation in the design of the study, development and implementation of most of the algorithms, and co-editing of the paper
- Citations according to Scopus: 59 (53 w/o autocitations)
- Award: Outstanding paper in Energy Forecasting, International Institute of Forecasters (2022)[1]

[1] `https://forecasters.org/ijf/iif-tao-hong-award-for-energy-forecasting`

4.2 Objective 2: Combining forecasts

4.2.1 Problem statement

The idea of combining or averaging forecasts appeared in the literature more than 50 years ago, with the pioneering papers of Bates and Granger (1969) and Crane and Crotty (1967). Since then many authors have suggested the superiority of forecast combinations over individual models (Timmermann, 2006; Nowotarski and Weron, 2016b; Berrisch and Ziel, 2022). Taylor (2020) remarks that when competing forecasts are available, their combination can provide a practical synthesis of the information contained in each individual prediction. Moreover, Hibon and Evgeniou (2005) argue that the advantage of combining forecasts is not that the best possible combinations perform better than the best individual forecasts (i.e., ex-post), but that it is less risky in practice to combine forecasts than to select an individual forecasting method (i.e., ex-ante).

Although many averaging schemes have been proposed in the literature, the simple average (that is, the arithmetic average of individual forecasts) stands out as the most popular and surprisingly reliable approach (Genre et al., 2013). The same applies to taking the median of forecasts, which in some cases can even outperform the average. Using OLS to derive combination weights is another easy-to-implement approach in which individual forecasts are treated as explanatory variables in linear regression. Such weights, however, may exhibit unstable behavior (so-called *bouncing betas*); even slight fluctuations in the data can cause large changes in the final forecast Weron (2014). To prevent this, Raviv et al. (2015) suggested using the constrained version of OLS, the so-called *constrained least squares* (CLS) averaging, in which the weights are only positive and add up to one.

The unwavering popularity of the simplest solutions is strong evidence of how difficult it is to choose the right tools to average forecasts. Recently, Diebold and Shin (2019) suggested that possibly we should turn the question around, and instead of wondering how to average individual forecasts, we should rather develop tools to select which forecasts we want to average in the first place. The authors introduce a new technique that combines the two approaches. The proposed regularization-based method can select and average forecasts from different models at once. Although their results were promising, there is room for improvement.

Another challenge in combining forecasts emerges with the increasing popularity of probabilistic forecasting methods. My thesis does not address this problem; however, it is briefly discussed in Section 4.3.2.

4.2.2 LASSO principal component averaging – a fully automated approach for point forecast pooling (Paper 2)

Regularization is a very general idea that can be used to solve many research problems. It can be used to increase the accuracy of future price forecasts, identify the most important explanatory variables (as in Objective 1), and also average the predictions obtained with various models. Principal component analysis, on the other hand, is a well-known tool that has been successfully applied to reduce the dimensionality of large data panels. Despite a few attempts (Chan et al., 1999; Huang and Lee, 2010) the potential of PCA has not been fully utilized in forecast averaging.

In Paper 2, we introduce *LASSO principal component averaging* (LPCA), a novel approach for averaging point forecasts. It combines LASSO estimation with the *principal component* (PC) averaging scheme proposed in our earlier study (Maciejowska et al., 2020). PCA extracts a relatively small set of orthogonal components from a large panel of forecasts, while LASSO selects the most important PCs and estimates the corresponding weights. It should be noted that compared to the method proposed in Maciejowska et al. (2020), the novel approach offers the advantage of being fully automated, meaning that the selection of the most important PCs and the corresponding weight estimation are carried out automatically.

The proposed approach is compared against nearly 20 LASSO- or PCA-based benchmarks and evaluated in four major energy markets. The underlying point predictions used for the pooling are obtained by estimating one model structure across different calibration windows ranging from 56 to 728 days, i.e., 673 different forecasts are obtained. In addition, we compare three *information criteria* (IC) to optimize the number of principal components used for PCA-based benchmarks and the value of the regularization parameter for LASSO-based methods.

The averaged forecasts are evaluated in terms of MAE and with percentage change of forecast accuracy relative to the results of a model with the longest considered calibration window. Additionally, we used the Giacomini and White (2006) test to see if the differences in the performance of the methods are statistically significant. The test is a generalization of the commonly used Diebold and Mariano (1995) test for unconditional predictive ability. The test uses information from the previous day to determine whether differences in forecast ability are affected by factors such as the business cycle phase (Giacomini and White, 2006).

Paper 2 confirms that averaging algorithms can substantially reduce forecast errors. The simple average reduces MAE by ca. 6.6% compared to the non-averaged prediction. The highest improvement (on average 10.3%) is achieved with the newly proposed method. The combination of PCA and LASSO, i.e., the LPCA model significantly outperforms both the PCA-based and LASSO-based benchmarks. Finally, of the three information criteria considered, the most robust performance is obtained for *Bayesian information criteria* (BIC).

Publication details:

- Authors: B. Uniejewski, K. Maciejowska
- Journal: *International Journal of Forecasting*, In Press / Corrected Proof
- DOI: 10.1016/j.ijforecast.2022.09.004
- Publication year: 2022 (online version)
- JCR Impact Factor: 7.022
- JCR classification(s): *Economics*, *Management*
- MEiN: 140 pts, assigned to the Management and Quality Studies (NZJ) discipline
- Contribution: 50%, the authors have contributed equally to the design of the study, implementation of the algorithms, and editing of the paper

4.3 Objective 3: Probabilistic forecasting

4.3.1 Problem statement

Compared to point predictions, probabilistic forecasts present much more information about the future price. Instead of specifying an expected value, they assign the probability that the future price will exceed a given level. Therefore, they provide a more comprehensive risk assessment and provide more information on possible unexpected or extreme price changes, allowing energy companies to avoid costs caused by unexpected fluctuations in the amount of electricity generation or consumption (Morales et al., 2014). Probabilistic EPF is a concept closely related to decision-making (Petropoulos et al., 2022). According to Taylor (2021), the availability of probabilistic predictions improves the decision-making process. In power markets, good quality probabilistic price forecasts can help producers, traders, and speculators determine optimal strategies for short-term operations (Uniejewski and Weron, 2021). Probabilistic forecasts are also used in risk management for derivative pricing, *value-at-risk* (VaR) calculations, hedging, and trading (Bunn et al., 2016).

The most common way to go from point to probabilistic forecasts is to construct a *prediction intervals* (PI). Instead of forecasting the expected value of the future price, we predict the price range in which the future price will fall with a given probability (see the left panel in Figure 4.1). If we extend this idea to multiple PIs, the final outcome will be a set of quantiles with many different levels. In the literature, it is emphasized that a dense grid of quantiles, e.g., 99 percentiles, is a sufficiently good approximation of the entire distribution (see the right panel in Figure 4.1) (Hong et al., 2016).

If we assume that the point forecast is the expected value of the future price for day d and hour h, that is, $\widehat{P}_{d,h} = \mathbb{E}(P_{d,h})$, then after observing the real value, we can calculate the forecast error: $\varepsilon_{d,h} = P_{d,h} - \widehat{P}_{d,h}$. As a result, we get:

$$F_P(x) = F_\varepsilon(x - \widehat{P}_{d,h}),$$

where F_ε is the distribution of errors corresponding to $\widehat{P}_{d,h}$, and F_P is the distribution of $P_{d,h}$. This means that both the error distribution and the price distribution have identical shapes but with different means (Gneiting and Katzfuss, 2014). Consequently, the

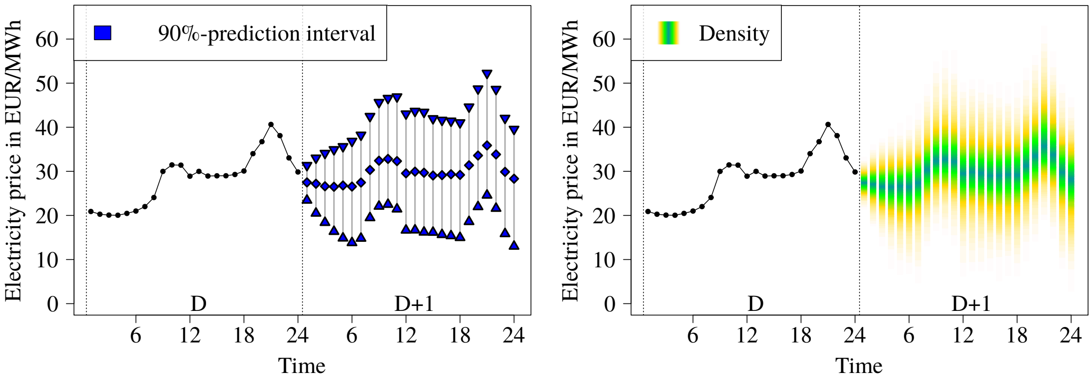


Figure 4.1: Example of probabilistic forecasts in the context of electricity price forecasting: quantiles of forecasted distributions (*left*) and density forecast (*right*). Source: Weron and Ziel (2020)

following relation applies to quantiles of distributions:

$$\widehat{q}_{\alpha,P} - \widehat{P}_{d,h} = \widehat{q}_{\alpha,\varepsilon} \text{ for } \alpha \in (0,1).$$

Using the quantile function, i.e., the inverse of the cumulative distribution function, we can formulate the relationship between the error distribution and the probabilistic forecast:

$$\widehat{F}_P^{-1}(\alpha) = \widehat{P}_{d,h} + \widehat{F}_\varepsilon^{-1}(\alpha).$$

One of the most popular probabilistic EPF methods is *quantile regression averaging* (QRA), of Nowotarski and Weron (2015). It has gained attention in academia and industry after its unprecedented success at *Global Energy Forecasting Competition* (GEFCom2014), where the method was used by the two top-winning teams in the price track (Gaillard et al., 2016; Maciejowska and Nowotarski, 2016). The idea underlying QRA is to average individual point forecasts using quantile regression.

$$\min_{\boldsymbol{\beta}_\alpha} \left[\sum (\alpha - \mathbb{1}_{\{P_{d,h} < \boldsymbol{\beta}_\alpha \boldsymbol{X}_{d,h}\}})(P_{d,h} - \boldsymbol{\beta}_\alpha \boldsymbol{X}_{d,h}) \right], \tag{4.3}$$

where $\alpha \in (0,1)$ is the order of the predicted quantile, $\boldsymbol{X}_{d,h}$ is the vector of independent variables and $\boldsymbol{\beta}_\alpha$ is the corresponding weight vector.

This approach has been reported to be successful not only in forecasting electricity prices but also in areas ranging from load (Liu et al., 2017; Zhang et al., 2018; Wang et al., 2019) to wind (Zhang et al., 2016) and solar power (Mpfumali et al., 2019).

In a recent study, (Marcjasz et al., 2020b) have revealed the vulnerability of QRA to low-quality point predictions. The method is not robust when the number of input forecasts is increased without any restrictions. To play safe Marcjasz et al. (2020a) recommend selecting as inputs in Eq. (4.3) only two or three different point forecasts. This is a serious limitation. Since the quality of the point forecasts is only known after the actual price is observed their selection is extremely difficult and at the same time crucial for the correct operation of the method. A way to tackle this was proposed by Maciejowska et al. (2016). However, their factor QRA (or FQRA) model yields only slightly more reliable forecasts and does not fully solve the problem.

4.3.2 On the importance of the long-term seasonal component in day-ahead electricity price forecasting. Part II - Probabilistic forecasting (Paper 3)

The *long-term seasonal component* (LTSC) of electricity prices is strongly influenced by external factors, such as changes in consumption and weather conditions (Paraschiv, 2013) and fossil fuel prices (Papież and Śmiech, 2015). Although the existence of the LTSC is generally accepted in the literature (Weron, 2014), most research on day-ahead EPF ignores or pays little attention to this component. For instance, Dudek (2016) argues that training models on the most recent data makes detrending of the time series or data deseasonalization unnecessary. However, according to Nowotarski and Weron (2016a), the *seasonal component autoregressive* (SCAR) modeling concept can significantly increase the point forecasting accuracy compared to models fitted to non-deseasonalized prices. The main purpose of Paper 3 is to validate whether the SCAR framework can also be used to improve prediction accuracy in the context of probabilistic forecasts.

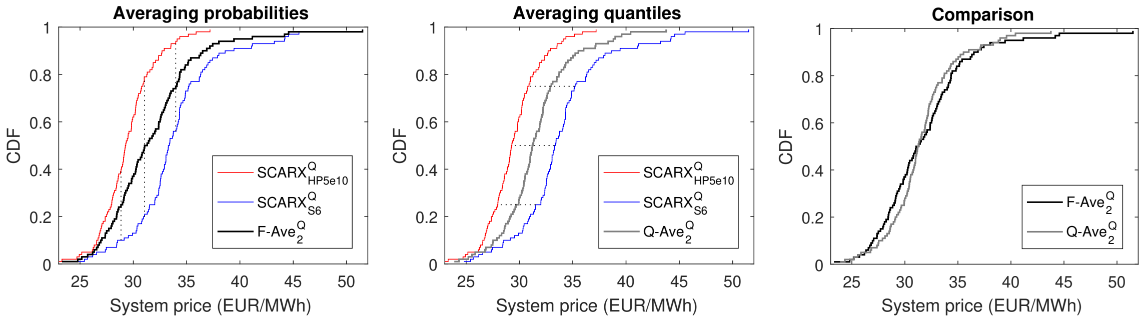


Figure 4.2: Illustration of the averaging probabilities and averaging quantiles concepts. The average probability forecast is a vertical average of the two cdfs (*left panel*), while the average quantile forecast is a horizontal average (*center panel*). Source: Figure 3 in Paper 3.

In this paper, we perform a comprehensive empirical study and consider a collection of 20 point forecasting methods including models based on the SCAR framework. To estimate the seasonal component, we use wavelet decomposition and Hodrick and Prescott (1997) filter, which, according to Rubaszek (2012), is the most commonly used filter that allows decomposing a series into a long-term trend and cycle components. Having the set of point forecasts, we apply one of three schemes for computing probabilistic forecasts (historical simulation, bootstrapping, and QRA). As a result, we obtain a pool of 59 individual probabilistic models. Furthermore, using two approaches to combine probabilistic forecasts introduced by Lichtendahl et al. (2013) we generate a set of averaged predictions and evaluate them using the pinball loss (see Section 5.1.2 in Paper 5). To draw valid conclusions, we use the Diebold and Mariano (1995) test to assess the significance of differences in the predictive accuracy.

The main contribution of Paper 3 is the introduction of a novel extension of the SCAR approach to probabilistic forecasting. One of the key findings is that SCAR-based probabilistic models nearly always significantly outperform the benchmarks. However, QRA allows obtaining, on average, more accurate predictions than historical simulation and bootstrap.

An additional contribution of Paper 3 is the development of methods to average probabilistic predictions. Despite the importance of the concept, only a few papers have considered quantile forecast combinations (Taylor, 2020). To our best knowledge, this study was the first to apply the averaging schemes proposed by Lichtendahl et al. (2013) to electricity price forecasting. The first method, called Probability (F) averaging, utilizes probabilities, which is equivalent to the vertical average of the *cumulative distribution functions* (cdf), and the latter, called Quantile (Q) averaging, considers quantiles, which is equivalent to the horizontal average of the cdfs, see Figure 4.2. In Paper 3, we observe that both averaging schemes significantly outperform both the benchmarks and the proposed (non-combined) SCAR-based models. Interestingly, in contrast to other econometric applications (Wang et al., 2022), we find that averaging over probabilities yields more accurate predictions. Note, that these two combination schemes can be generalized by considering averaging cdfs at different angles. As Taylor (2022) argued during last year's International Symposium on Forecasting, angular averaging could yield even better predictive distributions than any of the marginal cases. Checking whether this is also the case for EPF is left, however, for future research.

Publication details:

- Authors: B. Uniejewski, G. Marcjasz, R. Weron
- Journal: *Energy Economics* 79, 171-182
- DOI: 10.1016/j.eneco.2018.02.007
- Publication year: 2019
- JCR Impact Factor: 9.252
- JCR classification(s): *Economics*
- MEiN: 200 pts, assigned to the Management and Quality Studies (NZJ) discipline
- Contribution: 50%, including development and implementation of most of the algorithms and co-editing the paper
- Citations according to Scopus: 29 (22 w/o autocitations)

4.3.3 Regularized quantile regression averaging for probabilistic electricity price forecasting (Paper 4)

Regularization is a very general idea that can be used to solve many research problems. However, the vast majority of electricity price forecasting applications focus purely on point forecasting. To address this literature gap in Paper 4 we introduce a new approach to constructing probabilistic forecasts by applying the LASSO-type penalty function to regularize the QRA model. The proposed model is defined as follows:

$$\min_{\beta_\alpha} \Big\{ \textstyle\sum_{d,h} \underbrace{\left(\alpha - \mathbb{1}_{P_{d,h}<\boldsymbol{X}_{d,h}\boldsymbol{\beta}_\alpha}\right)\left(P_{d,h} - \boldsymbol{X}_{d,h}\boldsymbol{\beta}_\alpha\right)}_{\text{Quantile regression}} + \underbrace{\lambda \textstyle\sum_{n=1}^{N} \left|\beta_\alpha^{(n)}\right|}_{\text{LASSO penalty}} \Big\}, \tag{4.4}$$

where $\boldsymbol{X}_{d,h}$ is the vector of independent variables, that is, the vector of point forecasts, $\alpha \in (0, 1)$ is the order of the predicted quantile, and λ is the LASSO tuning parameter.

In this paper, we compare our approach with nearly 30 benchmarks. We consider a pool of 25 point forecasts obtained by estimating the same model structure with different amounts of data (similarly to Paper 2, see Section 4.2.2). In particular, we estimate the underlying point forecasting model using calibration windows ranging from 8 weeks to 2 years. Then we apply various variants of the standard QRA and a newly proposed regularized version of this model. In addition, we propose two automated techniques to optimize the regularization parameter. The first is based on the BIC and the second utilizes cross-validation (Stone, 1974).

In Paper 4, we test both the reliability and the sharpness of the obtained forecasts. To evaluate the reliability, we perform the Kupiec (1995) test and report the *prediction interval coverage probability* (PICP, Nowotarski and Weron (2018)). On the other hand, to compare the sharpness, we compute the aggregate pinball score and, similarly to Paper 2, we use the Giacomini and White (2006) test to assess significance. In addition, we evaluate the methods in terms of financial profits (for details, see Section 4.4).

One of the key findings is that our LASSO QRA (or LQRA) model outperforms all considered benchmarks in terms of both reliability and sharpness. What is more, we manage to overcome the issue of too narrow distributions, which can be a problem for standard QRA (Marcjasz et al., 2020b). The new approach is capable of passing the unconditional Kupiec test for most hours and significantly outperforms (by ca. 2-3% depending on the market) the QRA-based benchmarks in terms of pinball loss.

Publication details:

- Authors: B. Uniejewski, R. Weron
- Journal: *Energy Economics* 95, 105121
- DOI: 10.1016/j.eneco.2021.105121
- Publication year: 2021
- JCR Impact Factor: 9.252
- JCR classification(s): *Economics*
- MEiN: 200 pts, assigned to the Management and Quality Studies (NZJ) discipline
- Contribution: 75%, including model development, implementation of all algorithms, co-design of the study, and co-editing of the paper
- Citations according to Scopus: 20 (15 w/o autocitations)

4.4 Objective 4: Economic value of probabilistic forecasts

4.4.1 Problem statement[2]

There are only a handful of papers which examine the economic impact of EPF errors in a more systematic manner. Interestingly, most of these studies have been published in engineering, not economic or financial journals. The likely reason is that at least a basic knowledge is needed of how power markets, loads and generating units operate. Moreover, there is no standardized test ground/procedure for evaluating the economic impact. Nearly every EPF study considers a different setup.

Supply- and demand-side perspectives

In one of the earlier studies, Delarue et al. (2010) take the supply-side point of view and quantify the *profit loss* that can be expected in a price based unit commitment problem, when incorrect price forecasts are used. Simulations reveal that a combined cycle gas turbine (CCGT) is much more sensitive to EPF errors (the profit can easily lie 20% below the optimal level for a perfect price forecast) than a classic coal fired unit (profit loss rarely exceeds 10%). More interestingly, negatively biased forecasts (i.e., that predict prices lower than actual) typically yield much higher losses than positively biased predictions.

On the other hand, Zareipour et al. (2010) take the demand-side perspective and consider short-term operation scheduling of two typical loads (a process industry owning on-site generation facilities and a municipal water plant with load-shifting capabilities). They introduce the *forecast inaccuracy economic impact* index:

$$\text{FIEI} = \frac{\text{cost}(\widehat{P}) - \text{cost}(P)}{\text{cost}(\widehat{P})},$$

so that a positive value of FIEI indicates the percentage of the actual cost of buying electricity attributable to EPF errors. The authors report that a 1% improvement in the MAPE in forecasting accuracy would result in about 0.1%–0.35% cost reductions from short-term EPF, but also conclude that the MAPE is not a good measure.

An interesting concept is considered by Doostmohammadi et al. (2017), who compute the *financial loss/gain* (FLG) time series, defined as the difference between expected profit of a generator and the actual one. Then, based on the day-ahead forecasts of the FLG series, they propose a bidding strategy. However, by doing so, they do not work with the actual profits but with (another) estimate.

Maciejowska et al. (2019, 2021) take the perspective of a small RES utility (e.g., with one wind turbine) which has to decide where to sell 1 MW of electricity during each hour of the next day – in the day-ahead (DA) or the intraday (ID) market. Conditional on the decision, summarized by the *decision variable* based on price forecasts:

$$Y_{d,h} = \begin{cases} 1 & \text{if } \widehat{P}^{DA}_{d,h} > \widehat{P}^{ID}_{d,h}, \\ 0 & \text{if } \widehat{P}^{DA}_{d,h} \leq \widehat{P}^{ID}_{d,h}, \end{cases} \tag{4.5}$$

[2] This Section is an excerpt from Section *Trend #3: From Statistical to Economic Evaluation* in Paper 5

they compute the additional income over the benchmark, i.e., selling the production in the DA market, as:

$$\pi_{d,h} = Y_{d,h} P_{d,h}^{DA} + (1 - Y_{d,h}) P_{d,h}^{ID} - P_{d,h}^{DA}, \tag{4.6}$$

where $P_{d,h}^{DA}$ and $P_{d,h}^{ID}$ are the electricity prices in the DA and ID markets, respectively. While Maciejowska et al. (2019) utilize the load forecasts published by the German and Polish system operators, Maciejowska et al. (2021) additionally improve the load forecasts for Germany by applying ARX-type models. In both papers, they measure the gains from EPF as the sum of profits in the test period, $\pi = \sum_{d=1}^{D} \sum_{h=1}^{24} \pi_{d,h}$, and conclude that the statistical measures of forecast accuracy – such as the percent of correct sign classifications of the price spread between the DA and ID markets – do not necessarily coincide with economic benefits.

Trading strategies

Uniejewski et al. (2018) take a trading perspective (different from the supply- or demand-side point of views and consider a naive spot-futures trading strategy in the German market. With a perfect day-ahead forecast the buyer could always choose the lower of the two – the day-ahead price (unknown when submitting bids) or the futures price. Since this can never be achieved in reality, the authors bias (or perturb) the 'crystal-ball' forecast and show that a 0.20 EUR/MWh decrease in the MAE from using one model instead of another would result in ca. 90,000 EUR profits, for a 1 GW baseload in 2016.

Chitsaz et al. (2018) propose a trading strategy applicable in Ontario's real-time electricity market. The energy storage operator maximizes profits with optimal scheduling. The schedule is set before the trading period begins, based on the available price forecasts and then it is updated at the end of each hour with a newer price forecasts. The authors conclude that such a strategy yields higher profits when using predictions generated by the proposed ARX model with features selected via the Mutual Information technique (Amjady et al., 2011) – 62% of the potential saving for 'crystal ball' predictions, compared with a number of other EPF approaches, e.g., using the so-called Pre-Dispatch Prices (PDPs; publicly available price predictions published by the system operator IESO) – 43% of the potential saving.

Kath and Ziel (2018) propose a multivariate elastic net model for forecasting German quarter-hourly electricity prices. They demonstrate that the "sell in the high and buy in the low market" strategy performs well, leading to substantial benefits for both a net buyer and a net seller. On the other hand, the mean-variance approach does not bring economic benefits, but yields an optimal portfolio in terms of the *Sharpe ratio*:

$$SR = \frac{\bar{\pi}}{\sigma}, \tag{4.7}$$

where $\bar{\pi}$ denotes the average level of an additional revenue (i.e., $\bar{\pi} = \pi/24D$; see also Eq. (4.6)) and σ is the standard deviation of the time series of revenues. As such, the Sharpe ratio can be used to assess the trade-off between revenue and uncertainty. However, there are more performance measures (Eling and Schuhmacher, 2007; Auer, 2015), including measures based on drawdowns (e.g., Calmar ratio, Sterling ratio), based on partial moments (e.g., omega ratio, Sortino ratio) and based on the Value-at-Risk (VaR; e.g., excess return on VaR, conditional Sharpe ratio). Whether they will turn out to be useful in the EPF context remains yet to be checked.

4.4.2 Trading strategy as a way to evaluate forecasts (Paper 4 Revisited)

In Paper 4, we propose a trading strategy that can be implemented by a company that owns an energy storage system and participants in the Polish power market. In this three-step strategy, probabilistic forecasts are used to support the bidding process. First, in step I, before the day-ahead market prices are established for day $d-1$, we select two hours. Based on the quantile forecasts, we pick the hour with the lowest and the hour with the highest price on day d, respectively $h1$ and $h2$ in Figure 4.3).

In step II we submit the bid to buy 1 MWh at hour $h1$ and simultaneously the offer to sell 0.8 MWh (due to battery efficiency) at hour $h2$. A unique feature of the strategy is that the bidding levels are established based on quantile forecasts. The price for which we bid is equal to the upper bound of the PI at hour $h1$, i.e., $U^{\alpha}_{d,h1}$, and the price at which we offer to sell is equal to the lower bound of the PI at hour $h2$, i.e., $L^{\alpha}_{d,h2}$.

Finally, in step III the profits, depending on whether our bid and offer were accepted, are calculated according to Table 4.1. Note that whenever bids or offers are rejected, the transactions are executed at the same hours in the balancing market.

Table 4.1: Presentation of profit calculations from the trading strategy. The prices of electricity on day d and hour h are indicated by $P_{d,h}$ for the day-ahead market and $B_{d,h}$ for the balancing market.

	Case 1	**Case 2**	**Case 3**	**Case 4**
Bid	accepted	accepted	rejected	rejected
Offer	accepted	rejected	accepted	rejected
Profit	$0.8\ P_{d,h_2} - P_{d,h_1}$	$0.8\ B_{d,h_2} - P_{d,h_1}$	$0.8\ P_{d,h_2} - B_{d,h_1}$	0

In Paper 4 we show that using LQRA-based predictions (see Section 4.3.3) and the proposed strategy, we can gain around 20 000 PLN per year. We also provide evidence that decisions based on probabilistic forecasts can lead to a better outcome compared to those based on point forecasts. In particular, the averaged profits are 20% higher when the strategy is executed based on the LQRA approach compared to the same strategy based on point forecasts.

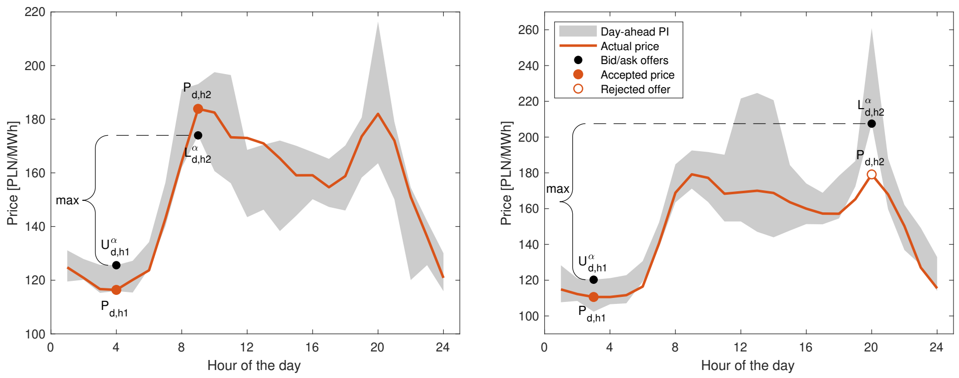


Figure 4.3: Illustration of the trading strategy using Polish TGE data for two selected days (*left panel* – 12.10.2017, *right panel* – 4.10.2017). The $\alpha\%$ prediction intervals are plotted in gray, the bid and offer prices are indicated with black dots and the actual price trajectory is in orange. Note, that on the right panel, the ask is not accepted because $P_{d,h2} < \hat{L}^{\alpha}_{d,h1}$.

4.5 Objective 5: Critical literature review

4.5.1 Forecasting electricity prices (Paper 5)

Paper 5 identified three trends in electricity price forecasting. Firstly, the tendency to consider not only point but also probabilistic and/or path (also called ensemble) forecasts. Secondly, the shift from the relatively parsimonious econometric or statistical models towards more complex, but potentially more accurate statistical learning/machine learning approaches. Finally, the recent trend to evaluate model performance not only in terms of statistical error measures (MAE, RMSE, pinball, CRPS, etc.), but also in terms of profits from scheduling or trading strategies based on price forecasts obtained from different models.

All three trends are addressed in this thesis. In particular, two methods to compute probabilistic forecasting are proposed in Papers 3 and 4 (Objective 3). Papers 1, 2, and 4 use statistical learning methods, such as LASSO, to improve forecasting performance (Objectives 1-3). Finally, in Paper 4 a trading strategy is proposed to assess the economic value of probabilistic forecasts (Objective 4).

Publication details:

- Authors: K. Maciejowska, B. Uniejewski, R. Weron
- Invited chapter in: *The Oxford Research Encyclopedia of Economics and Finance*, Accepted for publication
- DOI (arXiv): 10.48550/arXiv.2204.11735
- (Expected) publication year: 2023
- MEiN: 75 pts, assigned to the Management and Quality Studies (NZJ) discipline
- Contribution: $33\frac{1}{3}\%$, including literature review and co-editing of the paper

Chapter 5

Auxiliary results

In the course of my undergraduate and graduate studies, I have published 13 papers related to electricity price forecasting (listed below in chronological order). Five of them are an integral part of the thesis. The rest are either not directly related to the thesis or were considered less important, and thus excluded from the core part of the thesis. Nevertheless, for a complete picture of the research that I have conducted, below I briefly summarize the key findings.

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

2. Uniejewski, B., Weron, R., Ziel, F., 2018. Variance stabilizing transformations for electricity spot price forecasting. IEEE Transactions on Power Systems 33, 2219–2229.

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

4. 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 Economics 79, 171–182 → **Paper 3**.

5. Marcjasz, G., Uniejewski, B., Weron, R., 2019. On the importance of the long-term seasonal component in day-ahead electricity price forecasting with NARX neural networks. International Journal of Forecasting 35, 1520–1532.

6. Uniejewski, B., Marcjasz, G., Weron, R., 2019b. Understanding intraday electricity markets: Variable selection and very short-term price forecasting using LASSO. International Journal of Forecasting 35, 1533–1547 → **Paper 1**.

7. Serafin, T., Uniejewski, B., Weron, R., 2019. Averaging predictive distributions across calibration windows for day-ahead electricity price forecasting. Energies 12, 256.

8. Marcjasz, G., Uniejewski, B., Weron, R., 2020a. Beating the naive – combining LASSO with naive intraday electricity price forecasts. Energies 13, 1667.

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

10. Marcjasz, G., Uniejewski, B., Weron, R., 2020b. Probabilistic electricity price forecasting with NARX networks: Combine point or probabilistic forecasts? International Journal of Forecasting 35, 466–479.

11. Uniejewski, B., Weron, R., 2021. Regularized quantile regression averaging for probabilistic electricity price forecasting. Energy Economics 95, 105121 → **Paper 4**.

12. Uniejewski, B., Maciejowska, K., 2022. Lasso principal component averaging – a fully automated approach for point forecast pooling. International Journal of Forecasting, forthcoming, DOI: 10.1016/j.ijforecast.2022.09.004 → **Paper 2**.

13. Maciejowska, K., Uniejewski, B., Weron, R., 2023. Forecasting electricity prices. Oxford research encyclopedia of economics and finance, forthcoming, DOI (arXiv): 10.48550/arXiv.2204.11735 → **Paper 5**.

Uniejewski et al. (2016) was my first published research article submitted when I was still an undergraduate student. In this paper, we addressed the problem of selecting significant explanatory variables in a linear regression setting. We compared several automated variable selection methods, such as stepwise regression and regularization-based approaches. In particular, we were the first to apply the elastic net (Zou and Hastie, 2015) in the context of electricity price forecasting. We showed that regularization techniques such as LASSO or elastic net significantly outperform other competitors. As an additional contribution, we presented tables with variable selection frequencies that could help build well-performing parsimonious regression models.

In Uniejewski et al. (2018) we raised the issue of data preprocessing. We introduced new functions aiming to stabilize the variance and showed that applying appropriate *variance stabilizing transformations* (VSTs) could significantly reduce forecasting errors. In particular, we recommended the *area hyperbolic sine* (asinh) or *normal probability integral transform* (N-PIT) transformations, which in our setup improved the prediction accuracy the most. This paper has had a visible impact on the electricity forecasting community and, according to the Scopus database, is my most frequently cited article.

In Uniejewski and Weron (2018) we were searching for the optimal way to implement LASSO for electricity price forecasting. We addressed three open issues: variable selection, the choice of the tuning parameter, and the choice of VST. We concluded that selecting a fixed λ value for all days in the out-of-sample period is an acceptable option, however, if computation efficiency is not a problem, re-selecting λ on a daily basis can increase forecast accuracy. Finally, we showed that using VSTs significantly reduces the forecasting error for LASSO model.

In Marcjasz et al. (2019) we addressed a similar problem as in Paper 3 but from a different perspective. Both studies were inspired by Nowotarski and Weron (2016a) and aimed at answering the question of whether it is beneficial to separately consider the long-term seasonal component when forecasting electricity prices. The difference between this article and Paper 3 is substantial. Here, we utilized neural networks and showed that the benefits of using the seasonal component approach in a point forecasting concept are significant also for non-linear models.

In Serafin et al. (2019) we addressed the problem of averaging forecasts of the same model trained on different calibration windows. Motivated by the results of Hubicka et al. (2019) and Marcjasz et al. (2018), we proposed two quantile regression-based extensions of this approach to probabilistic forecasts. Having a large panel of point forecasts, we successfully applied Quantile Regression Averaging (QRA) of Nowotarski and Weron (2015) and Quantile Regression Machine (QRM) of Uniejewski et al. (2019a) to obtain probabilistic predictions. We showed that the more computationally efficient QRM approach outperforms QRA. In addition, we confirmed that the idea of combining forecasts obtained for short and long calibration windows improves prediction accuracy also for probabilistic forecasts.

Marcjasz et al. (2020a) is a continuation of Paper 1. The main objective was to improve the accuracy of very short-term predictions in the German intraday market. The proposed solution was a combination of the LASSO model developed in Paper 1 and the naive model of Narajewski and Ziel (2020a). Additionally, we showed that, similarly to forecasting day-ahead prices, exogenous variables such as load and wind power generation improve model performance.

Maciejowska et al. (2020) is another article on forecast combination. It has started a series of articles on the use of principal component averaging, including Paper 2 and further future research. In that paper, we used PCA to average a large panel of forecasts. The proposed method appears to be very sensitive to the number of factors considered. To overcome this problem, we use Bayesian information criteria to select the optimal number of factors. Although the results were satisfactory, in Paper 2 we achieved even better predictions using LASSO estimation.

The problem of averaging forecasts is also addressed in Marcjasz et al. (2020b). Having a pool of point forecasts, we can proceed in two different ways to obtain probabilistic forecasts via quantile regression. Firstly, we can average point forecasts to improve accuracy and use them to produce a probabilistic forecast. Secondly, we can produce a pool of probabilistic forecasts, each corresponding to a different point forecast, and then average them. The former approach turned out to yield more accurate predictions.

Chapter 6

Conclusions

The aim of this thesis was to make a significant contribution to the field of electricity price forecasting (EPF) by developing robust predictive models for decision-making in power markets. The advancement of EPF techniques is crucial as accurate price predictions can aid power market participants in energy trading and risk management decisions, thereby increasing the profits of energy companies.

To address this aim, five objectives have been set:

1. **Use regularization to identify the most relevant explanatory variables:** In this objective, the goal is to identify the most important predictors of electricity prices using statistical learning techniques. The LASSO-type regularization technique is proposed as a solution to this problem. Empirical tests, including the rapidly growing German intraday market, indicate that LASSO-based models can successfully recognize the most informative explanatory variables (Paper 1).
2. **Develop a fully-automated approach to average a rich pool of individual forecasts using regularization and Principal Component Analysis (PCA):** In this objective, a novel approach is introduced to combine forecasts obtained from different models. The rapid development of EPF methods results in countless solutions to obtain an accurate prediction, making it impossible to choose the overall best forecasting model. The proposed approach incorporates the PCA and LASSO methods to average a large pool of forecasts and provides a more accurate prediction compared to individual models (Paper 2).
3. **Utilize quantile regression and regularization to construct more accurate algorithms for probabilistic forecasting of electricity prices:** In this objective, a preprocessing technique such as series decomposition is used in the context of probabilistic forecasting (Paper 3), and a novel, LASSO-based approach is introduced to construct probabilistic predictions (Paper 4). The proposed solutions significantly outperform existing literature benchmarks in terms of accuracy and reliability.
4. **Design a trading strategy to evaluate the economic value of probabilistic forecasts:** In this objective, instead of a statistical measure we introduce a trading strategy and use it to evaluate the economic value of probabilistic forecasts, illustrating the importance of predictions in supporting decision-making processes and maximizing profits (Paper 4).
5. **Conduct a critical review of forecasting in the electricity markets and provide an outlook for future research in this area:** In this objective, a comprehensive

overview of the EPF research is provided, with a particular focus on three current trends (Paper 5). The critical review helps to integrate and organize the primary literature and sets the direction for future research in the EPF area.

The contribution of this thesis extends beyond the advancement of EPF techniques. It also highlights the importance of considering the economic value of predictions in decision-making processes and encourages future research in this area. The results of this dissertation have significant implications for power market participants and academia, providing them with valuable insights and tools to forecast electricity prices and support decision-making.

Bibliography

Amjady, N., Keynia, F., Zareipour, H., 2011. Wind power prediction by a new forecast engine composed of modified hybrid neural network and enhanced particle swarm optimization. IEEE Transactions on Sustainable Energy 2, 265–276.

Anderson, D.R., Sweeney, D.J., Williams, T.A., Camm, J.D., Cochran, J.J., 2015. An Introduction to Management Science: Quantitative Approaches to Decision Making. Cengage Learning.

Andrade, J., Filipe, J., Reis, M., Bessa, R., 2017. Probabilistic price forecasting for day-ahead and intraday markets: Beyond the statistical model. Sustainability 9, 1990.

Auer, B., 2015. Does the choice of performance measure influence the evaluation of commodity investments? International Review of Financial Analysis 38, 142–150.

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

Berrisch, J., Ziel, F., 2022. CRPS learning. Journal of Econometrics, forthcoming, DOI: 10.1016/j.jeconom.2021.11.008.

Bunn, D., 2000. Forecasting loads and prices in competitive power markets. Proceedings of the IEEE 88, 163–169.

Bunn, D., Andresen, A., Chen, D., Westgaard, S., 2016. Analysis and forecasting of electricity price risks with quantile factor models. Energy Journal 37, 101–122.

Chan, Y.L., Stock, J.H., Watson, M.W., 1999. A dynamic factor model framework for forecast combination. International Journal of Forecasting 22, 283–300.

Chitsaz, H., Zamani-Dehkordi, P., Zareipour, H., Parikh, P.P., 2018. Electricity price forecasting for operational scheduling of behind-the-meter storage systems. IEEE Transactions on Smart Grid 9, 6612–6622.

Crane, D., Crotty, J., 1967. A two-stage forecasting model: exponential smoothing and multiple regression. Management Science 13, B501–B507.

Delarue, E., Van Den Bosch, P., D'haeseleer, W., 2010. Effect of the accuracy of price forecasting on profit in a price based unit commitment. Electric Power Systems Research 80, 1306–1313.

Delen, D., Ram, S., 2018. Research challenges and opportunities in business analytics. Journal of Business Analytics 1, 2–12.

Diaz, G., Planas, E., 2016. A note on the normalization of Spanish electricity spot prices. IEEE Transactions on Power Systems 31, 2499–2500.

Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. Journal of Business and Economic Statistics 13, 253–263.

Diebold, F.X., Shin, M., 2019. Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives. International Journal of Forecasting 35, 1679–1691.

Doostmohammadi, A., Amjady, N., Zareipour, H., 2017. Day-ahead financial loss/gain modeling and prediction for a generation company. IEEE Transactions on Power Systems 32, 3360–3372.

Dudek, G., 2016. Multilayer perceptron for GEFCom2014 probabilistic electricity price forecasting. International Journal of Forecasting 32, 1057–1060.

Elbanna, S., 2006. Strategic decision-making: Process perspectives. International Journal of Management Reviews 8, 1–20.

Eling, M., Schuhmacher, F., 2007. Does the choice of performance measure influence the evaluation of hedge funds? Journal of Banking and Finance 31, 2632–2647.

Fezzi, C., Mosetti, L., 2020. Size matters: Estimation sample length and electricity price forecasting accuracy. The Energy Journal 41, 231–254.

Gaillard, P., Goude, Y., Nedellec, R., 2016. Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. International Journal of Forecasting 32, 1038–1050.

Genre, V., Kenny, G., Meyler, A., Timmermann, A., 2013. Combining expert forecasts: Can anything beat the simple average? International Journal of Forecasting 29, 108–121.

Giacomini, R., White, H., 2006. Tests of conditional predictive ability. Econometrica 74, 1545–1578.

Gneiting, T., Katzfuss, M., 2014. Probabilistic forecasting. Annual Review of Statistics and Its Application 1, 125–151.

Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press.

Heizer, J.H., Render, B., Weiss, H.J., 2004. Operations Management. Pearson Prentice Hall.

Hibon, M., Evgeniou, T., 2005. To combine or not to combine: Selecting among forecasts and their combinations. International Journal of Forecasting 21, 15–24.

Hodrick, R.J., Prescott, E.C., 1997. Postwar U.S. business cycles: An empirical investigation. Journal of Money, Credit and Banking 29, 1–16.

Hong, T., 2015. Crystal ball lessons in predictive analytics. EnergyBiz, Spring , 35–37.

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

Huang, H., Lee, T.H., 2010. To combine forecasts or to combine information? Econometric Reviews 29, 534–570.

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.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning with Applications in R. Springer, New York.

Janczura, J., Wójcik, E., 2022. 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.

Janke, T., Steinke, F., 2019. Forecasting the price distribution of continuous intraday electricity trading. Energies 12, 4262.

Jędrzejewski, A., Lago, J., Marcjasz, G., Weron, R., 2022. Electricity price forecasting: The dawn of machine learning. IEEE Power & Energy Magazine 20, 24–31.

Jędrzejewski, A., Marcjasz, G., Weron, R., 2021. Importance of the long-term seasonal component in day-ahead electricity price forecasting revisited: Parameter-rich models estimated via the lasso. Energies 14, 3249.

Kamiński, B., Jakubczyk, M., Szufel, P., 2018. A framework for sensitivity analysis of decision trees. Central European Journal of Operations Research 26, 135–159.

Karakatsani, N., Bunn, D., 2008. Forecasting electricity prices: The impact of fundamentals and time-varying coefficients. International Journal of Forecasting 24, 764–785.

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

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

Kramer, A., Kiesel, R., 2021. Exogenous factors for order arrivals on the intraday electricity market. Energy Economics 97, 105186.

Kulakov, S., Ziel, F., 2021. The impact of renewable energy forecasts on intraday electricity prices. Economics of Energy and Environmental Policy 10, 79–104.

Kupiec, P.H., 1995. Techniques for verifying the accuracy of risk measurement models. The Journal of Derivatives 3, 73–84.

Lepenioti, K., Bousdekis, A., Apostolou, D., Mentzas, G., 2020. Prescriptive analytics: Literature review and research challenges. International Journal of Information Management 50, 57–70.

Li, Y., Zhu, J., 2008. L1-norm quantile regression. Journal of Computational and Graphical Statistics 17, 163–185.

Lichtendahl, K.C., Grushka-Cockayne, Y., Winkler, R.L., 2013. Is it better to average probabilities or quantiles? Management Science 59, 1594–1611.

Lisi, F., Pelagatti, M., 2018. Component estimation for electricity market data: Deterministic or stochastic? Energy Economics 74, 13–37.

Liu, B., Nowotarski, J., Hong, T., Weron, R., 2017. Probabilistic load forecasting via Quantile Regression Averaging on sister forecasts. IEEE Transactions on Smart Grid 8, 730–737.

Ludwig, N., Feuerriegel, S., Neumann, D., 2015. Putting big data analytics to work: Feature selection for forecasting electricity prices using the LASSO and random forests. Journal of Decision Systems 24, 19–36.

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

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

Maciejowska, K., Nowotarski, J., 2016. A hybrid model for GEFCom2014 probabilistic electricity price forecasting. International Journal of Forecasting 32, 1051–1056.

Maciejowska, K., Nowotarski, J., Weron, R., 2016. Probabilistic forecasting of electricity spot prices using Factor Quantile Regression Averaging. International Journal of Forecasting 32, 957–965.

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, forthcoming, DOI (arXiv): 10.48550/arXiv.2204.11735.

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

Marcjasz, G., Uniejewski, B., Weron, R., 2019. On the importance of the long-term seasonal component in day-ahead electricity price forecasting with NARX neural networks. International Journal of Forecasting 35, 1520–1532.

Marcjasz, G., Uniejewski, B., Weron, R., 2020a. Beating the naive – Combining lasso with naive intraday electricity price forecasts. Energies 13, 1667.

Marcjasz, G., Uniejewski, B., Weron, R., 2020b. Probabilistic electricity price forecasting with NARX networks: Combine point or probabilistic forecasts? International Journal of Forecasting 35, 466–479.

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

Misiorek, A., 2008. Short-term forecasting of electricity prices: Do we need a different model for each hour? Medium Econometrisch Toepassingen 16, 8–13.

Monteiro, C., Ramirez-Rosado, I., Fernandez-Jimenez, L., Conde, P., 2016. Short-term price forecasting models based on artificial neural networks for intraday sessions in the Iberian electricity market. Energies 9, 721.

Morales, J.M., Conejo, A.J., Madsen, H., Pinson, P., Zugno, M., 2014. Integrating Renewables in Electricity Markets: Operational Problems. Springer.

Mpfumali, P., Sigauke, C., Bere, A., Mulaudzi, S., 2019. Day ahead hourly global horizontal irradiance forecasting – Application to South African data. Energies 12, 3569.

Narajewski, M., Ziel, F., 2020a. Econometric modelling and forecasting of intraday electricity prices. Journal of Commodity Markets 19, 100107.

Narajewski, M., Ziel, F., 2020b. Ensemble forecasting for intraday electricity prices: Simulating trajectories. Applied Energy 279, 115801.

Narajewski, M., Ziel, F., 2022. Optimal bidding in hourly and quarter-hourly electricity price auctions: Trading large volumes of power with market impact and transaction costs. Energy Economics 110, 105974.

Nowotarski, J., Weron, R., 2015. Computing electricity spot price prediction intervals using quantile regression and forecast averaging. Computational Statistics 30, 791–803.

Nowotarski, J., Weron, R., 2016a. On the importance of the long-term seasonal component in day-ahead electricity price forecasting. Energy Economics 57, 228–235.

Nowotarski, J., Weron, R., 2016b. To combine or not to combine? Recent trends in electricity price forecasting. ARGO 9, 7–14.

Nowotarski, J., Weron, R., 2018. Recent advances in electricity price forecasting: A review of probabilistic forecasting. Renewable and Sustainable Energy Reviews 81, 1548–1568.

Papież, M., Śmiech, S., 2015. Modelowanie i prognozowanie cen surowców energetycznych. C.H. Beck.

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.

Paraschiv, F., 2013. Price dynamics in electricity markets, in: Handbook of Risk Management in Energy Production and Trading. Springer, pp. 47–69.

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.

Raviv, E., Bouwman, K.E., van Dijk, D., 2015. Forecasting day-ahead electricity prices: Utilizing hourly prices. Energy Economics 50, 227–239.

Rose, B., 2016. Defining analytics: A conceptual framework. ORMS Today 43, 36–40.

Rubaszek, M., 2012. Modelowanie polskiej gospodarki z pakietem R. SGH.

Sadowski, J., Zawisza, M., Kamiński, B., 2012. Major drivers for price spikes' occurrence on balancing market (in Polish). Rynek Energii 99, 73–78.

Sadowski, W., 1980. Forecasting and decision-making, in: Quantitative Wirtschafts-und Unternehmensforschung. Springer, pp. 92–102.

Schneider, S., 2011. Power spot price models with negative prices. Journal of Energy Markets 4, 77–102.

Serafin, T., Uniejewski, B., Weron, R., 2019. Averaging predictive distributions across calibration windows for day-ahead electricity price forecasting. Energies 12, 256.

Shi, W., Wang, Y., Chen, Y., Ma, J., 2021. An effective two-stage electricity price forecasting scheme. Electric Power Systems Research 199, 107416.

Simon, H.A., 1960. The New Science of Management Decision. Harper and Brothers.

Slack, N., Brandon-Jones, A., Burgess, N., 2022. Operations Management 10th edition. Pearson.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society: Series B (Methodological) 36, 111–133.

Taylor, J.W., 2020. Forecast combinations for value at risk and expected shortfall. International Journal of Forecasting 36, 428–441.

Taylor, J.W., 2021. Evaluating quantile-bounded and expectile-bounded interval forecasts. International Journal of Forecasting 37, 800–811.

Taylor, J.W., 2022. Angular combining of forecasts of probability distributions. Presentation during the International Symposium on Forecasting (ISF2022).

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society B 58, 267–288.

Tikhonov, A.N., 1963. Solution of incorrectly formulated problems and the regularization method. Soviet Mathematics Doklady 4, 1035–1038.

Timmermann, A.G., 2006. Forecast combinations, in: Handbook of economic forecasting. Elsevier, pp. 135–196.

Uniejewski, B., K.Maciejowska, 2022. Lasso Principal Component Averaging – A fully automated approach for point forecast pooling. International Journal of Forecasting, forthcoming, DOI: 10.1016/j.ijforecast.2022.09.004.

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 Economics 79, 171–182.

Uniejewski, B., Marcjasz, G., Weron, R., 2019b. Understanding intraday electricity markets: Variable selection and very short-term price forecasting using lasso. International Journal of Forecasting 35, 1533–1547.

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

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

Uniejewski, B., Weron, R., 2021. Regularized quantile regression averaging for probabilistic electricity price forecasting. Energy Economics 95, 105121.

Uniejewski, B., Weron, R., Ziel, F., 2018. Variance stabilizing transformations for electricity spot price forecasting. IEEE Transactions on Power Systems 33, 2219–2229.

Wang, X., Hyndman, R.J., Li, F., Kang, Y., 2022. Forecast combinations: An over 50-year review. International Journal of Forecasting, forthcoming, DOI: 10.1016/j.ijforecast.2022.11.005.

Wang, Y., Zhang, N., Tan, Y., Hong, T., Kirschen, D., Kang, C., 2019. Combining probabilistic load forecasts. IEEE Transactions on Smart Grid 10, 3664–3674.

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

Weron, R., 2014. Electricity price forecasting: A review of the state-of-the-art with a look into the future. International Journal of Forecasting 30, 1030–1081.

Weron, R., Ziel, F., 2020. Electricity price forecasting, in: Handbook of Energy Economics. Routledge, pp. 506–521.

Wilson, R., 2002. Architecture of power markets. Econometrica 70, 1299–1340.

Zareipour, H., Canizares, C.A., Bhattacharya, K., 2010. Economic impact of electricity market price forecasting errors: A demand-side analysis. IEEE Transactions on Power Systems 25, 254–262.

Zhang, W., Quan, H., Srinivasan, D., 2018. Parallel and reliable probabilistic load forecasting via quantile regression forest and quantile determination. Energy 160, 810–819.

Zhang, Y., Liu, K., Qin, L., An, X., 2016. Deterministic and probabilistic interval prediction for short-term wind power generation based on variational mode decomposition and machine learning methods. Energy Conversion and Management 112, 208–219.

Ziel, F., 2016. Forecasting electricity spot prices using lasso: On capturing the autoregressive intraday structure. IEEE Transactions on Power Systems 31, 4977–4987.

Ziel, F., Steinert, R., 2018. Probabilistic mid- and long-term electricity price forecasting. Renewable and Sustainable Energy Reviews 94, 251–266.

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

Zou, H., Hastie, T., 2015. Regularization and variable selection via the elastic nets. Journal of the Royal Statistical Society B 67, 301–320.

Paper 1

Understanding intraday electricity markets: Variable selection and very short-term price forecasting using LASSO

Bartosz Uniejewski, Grzegorz Marcjasz, Rafał Weron

International Journal of Forecasting 35 (2019) 1533–1547

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Understanding intraday electricity markets: Variable selection and very short-term price forecasting using LASSO

Bartosz Uniejewski [a,b], Grzegorz Marcjasz [a,b], Rafał Weron [a,*]

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

ARTICLE INFO

Keywords:
Intraday electricity market
Variable selection
Price forecasting
LASSO
ARX model
Diebold-Mariano test
Trading strategy

ABSTRACT

We use a unique set of prices from the German EPEX market and take a closer look at the fine structure of intraday markets for electricity, with their continuous trading for individual load periods up to 30 min before delivery. We apply the *least absolute shrinkage and selection operator* (LASSO) in order to gain statistically sound insights on variable selection and provide recommendations for very short-term electricity price forecasting.

1. Introduction

Since the deregulation of government-controlled power sectors in the 1990s and 2000s and the introduction of competitive markets in many countries worldwide, electricity has been being traded under market rules like any other commodity (Mayer & Trück, 2018). The workhorse of power trading in Europe has been the uniform price auction conducted a day before delivery, and the vast majority of research studies and applications have concerned day-ahead (DA) electricity prices (Weron, 2014). However, the expansion of renewable generation (mostly wind and solar), the modernization of power grids (including an increase in interconnector capacity) and active demand-side management (smart meters, smart appliances) have made the electricity demand/supply and prices more volatile and less predictable than ever before (Hong & Fan, 2016; Kiesel & Kusterman, 2016). This has amplified the importance of intraday markets, which can be used to balance the deviations resulting from differences between positions in day-ahead contracts and the actual demand (Gianfreda, Parisio, & Pelagatti, 2016; Märkle-Huß, Feuerriegel, & Neumann, 2018; Zaleski & Klimczak, 2015). As a result, the last few years have observed a shifting of volume from the DA to intraday markets across Europe (EPEX, 2018).

This article uses a unique set of prices from the German EPEX market and takes a closer look at the fine structure of intraday markets, with their continuous trading for individual load periods up to 30 min before delivery. We apply the *least absolute shrinkage and selection operator* (LASSO) of Tibshirani (1996) in order to gain statistically sound insights on variable selection and provide recommendations for very short-term *electricity price forecasting* (EPF).[1] Given that the literature on the forecasting of intraday prices in European power markets is very scarce — being limited, to the best of our knowledge, to only two papers dealing with Spanish data (Andrade, Filipe, Reis, & Bessa, 2017; Monteiro, Ramirez-Rosado, Fernandez-Jimenez, & Conde, 2016) — our study is a major step towards understanding the intraday price dynamics and developing predictive models that perform

* Corresponding author.
E-mail addresses: uniejewskibartosz@gmail.com (B. Uniejewski), gelusz@hotmail.co.uk (G. Marcjasz), rafal.weron@pwr.edu.pl (R. Weron).

[1] We use EPF as the abbreviation for both *electricity price forecasting* and *electricity price forecast*. The plural form, i.e., forecasts, is abbreviated EPFs.

https://doi.org/10.1016/j.ijforecast.2019.02.001
0169-2070/© 2019 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

well for a market that many participants see as the future of electricity trading. The importance of our study is emphasized further by the fact that electricity price forecasts are now fundamental inputs to energy companies' decision-making mechanisms, alongside weather and demand predictions (Nowotarski & Weron, 2018).

The remainder of the paper is structured as follows. Section 2 reviews the literature on intraday electricity markets and on variable selection for EPF. Section 3 begins by introducing the EPEX dataset and the rolling window scheme, then discusses variance stabilization, and finally describes the model structures considered. Section 4 first compares the predictive performance in terms of two commonly-used error measures and the Diebold and Mariano (1995) test, then takes a closer look at the best LASSO-estimated model in order to identify the most important explanatory variables, and thus provide guidelines to the structuring of better-performing models for intraday electricity markets. Finally, Section 5 wraps up the results and concludes.

2. Literature review

2.1. Intraday markets for electricity

There is a small but growing body of literature on intraday electricity markets. Most publications take a fundamental, market behavior or energy policy perspective. For instance, Pape, Hagemann, and Weber (2016) investigate the explanatory power of a fundamental modeling approach in the German power market, explicitly accounting for must-run operations of combined heat and power plants (CHP) and intraday peculiarities such as a shortened intraday supply stack. González-Aparicio and Zucker (2015) analyze the influence of wind power forecasting errors in the time period between the closure of the day-ahead market and the opening of the first intraday session in the Spanish market. Similarly, Ziel (2017) studies the impact of wind and solar generation forecast errors on price formation in the German intraday market. Kiesel and Paraschiv (2017) investigate bidding behaviors in the intraday market by looking at both last prices and continuous bidding, in the context of a reduced-form econometric analysis. They find that intraday prices adjust asymmetrically to both forecasting errors in renewables and the volume of trades dependent on a threshold variable that reflects the expected non-renewable generation in the DA market. Aïd, Gruet, and Pham (2016) analyze the trading in intraday electricity markets and develop an optimal bidding strategy. They consider as their modeling framework a continuous-time stochastic process for the net position of sales and purchases of electricity. Märkle-Huß et al. (2018) investigate the introduction of 15-min contracts to the German EPEX market and argue that these products are used to balance the intra-hour volatility of renewable energy sources. Finally, Maciejowska, Nitka, and Weron (2019) take the perspective of a small renewable energy generator that trades via a larger company and only has to decide how much electricity it will sell in the DA market; the rest will be sold in either the intraday market (Germany) or the balancing market (Poland). They forecast the price spread between the DA and intraday/balancing markets using autoregressive and probit models and show that statistical measures of the forecast accuracy, such as the percentage of correct sign classifications, do not necessarily coincide with economic benefits.

However, when it comes to the forecasting of intraday electricity prices in European power markets, the literature is very scarce. To the best of our knowledge, only two studies have addressed this important problem, and these only in the context of the Iberian electricity market (MIBEL), which features a very specific design with six intraday sessions of between 9 and 27 delivery hours. Monteiro et al. (2016) utilize neural networks (multilayer perceptron, one hidden layer) with up to 21 input variables selected on an ad-hoc basis: dummies (hour-of-the-day, day-of-the-week), hourly prices on previous days (lags 1 and 7), price values of the daily session, price values of previous intraday sessions, and weather, demand and wind power generation forecasts. They find that the best models for intraday sessions #1 to #5 use only the hourly prices of the daily session, the hourly prices of previous intraday sessions and the seasonal dummies, while the best model for intraday session #6 uses only the hourly prices of previous intraday sessions #3–#5 and the seasonal dummies. Andrade et al. (2017) reach similar conclusions by utilizing a linear quantile regression (LQR) model, namely that high quality point and probabilistic forecasts of intraday prices can be obtained by just exploring the prices from previous sessions (plus deterministic variables for modeling the daily, weekly and annual seasonalities), despite the fact that they consider a large set of fundamental variables.

2.2. Variable selection for electricity price forecasting

The conclusions from these last two studies suggest clearly that variable selection is a very important issue in EPF, and that it may be even more critical for intraday markets than for DA markets because of the vast amounts of data available. In this context, high-dimensional statistical modeling techniques that deal with large amounts of data may come in handy.

The earliest known examples of statistically-sound variable selection in day-ahead EPF include the studies by Karakatsani and Bunn (2008) and Misiorek (2008), who used *stepwise regression* to eliminate statistically insignificant variables from parsimonious regression-type models, and Amjady and Keynia (2009), who introduced a feature selection algorithm based on *mutual information*. We believe that Barnes and Balda (2013) were the first to apply regularization in day-ahead EPF. In a study concerning the profitability of battery storage, they utilized *ridge regression* to compute EPFs for a model with more than 50 regressors. Ludwig, Feuerriegel, and Neumann (2015) used *random forests* and the LASSO to choose which of the 77 available weather stations were relevant, while Keles, Scelle, Paraschiv, and Fichtner (2016) combined the *k-nearest-neighbor* algorithm with *backward elimination* to select the most appropriate inputs out of more than 50 fundamental parameters or lagged versions of these parameters.

A qualitative change came with the papers by Ziel (2016) and Ziel, Steinert, and Husmann (2015), who used LASSO to sparsify very large (100+) sets of model parameters, utilizing B-splines either in a univariate setting or, more efficiently, within a multivariate framework. In the first thorough comparative study, Uniejewski, Nowotarski, and Weron (2016) evaluated six automated selection and shrinkage procedures (single-step elimination, forward and backward stepwise regression, ridge regression, LASSO, and *elastic nets*) applied to a baseline model with 100+ regressors. They concluded that the use of LASSO and elastic nets can achieve significant accuracy gains relative to commonly-used EPF models. In a study on the optimal model structure for day-ahead EPF, Ziel and Weron (2018) considered autoregressive models with 200+ potential explanatory (but not exogenous) variables, and concluded that both uni- and multivariate LASSO-implied structures outperform autoregressive benchmarks significantly, and that combining their forecasts can achieve further improvements in predictive accuracy. Finally, Uniejewski and Weron (2018) show that using a complex regression model with nearly 400 explanatory variables, a well-chosen variance-stabilizing transformation (asinh or N-PIT), and a procedure that recalibrates the LASSO regularization parameter once or twice a day leads to significant accuracy gains compared to the EPF models that are considered typically.

This study follows the approach set forth in the last three articles and considers LASSO models based on hundreds of potential regressors and calibrated to asinh-transformed prices. However, we do not consider recalibrating the LASSO regularization parameter, as this slows down the forecasting procedure considerably.

3. Methodology

3.1. The dataset

The main German 'spot' market is operated by EPEX SPOT SE and allows the trading of power supply contracts with hourly and quarter-hourly delivery. The participants have the option of bidding on hourly products in the day-ahead (DA) auction that is conducted at noon on the day before delivery (i.e., $d-1$), or trading hourly and quarter hourly contracts in the continuous intraday market that opens at 16:00 on day $d-1$ and closes 30 min before the delivery starts (since March 2017, five minutes for transactions within the delivery zone; see EPEX, 2018).

The leading reference price for the intraday market is the recently-introduced ID3 index for contracts with an hourly delivery, which is also an underlying instrument of exchange-traded derivative products (see https://www.eex.com). The index is based exclusively on hourly and 15-min products traded in the German intraday continuous market (i.e., intraday auction data are excluded), and is computed as the volume-weighted average price of all trades performed over the last three hours before the delivery starts. Moreover, cross-trades (i.e., trades with the same entity selling on one side and buying on the other side) are excluded, while cross-border trades with one leg (buy/sell) in Germany are taken into account (EPEX, 2015).

The exchange publishes the index, but the period covered is too short for a proper evaluation of our models. Thus, we have reconstructed an ID3-like time series from the individual transactions. It differs slightly from the actual index: (i) we have not excluded cross-trades (since the data that we have access to are anonymized), and (ii) we have not considered the trades conducted between 30 and 5 min before the delivery starts, because such trades have been allowed only since March 2017. For each hourly product, only the transactions with timestamps between 180 and 30 min were chosen. For products with no transactions in this period, the window was extended to contain transactions conducted from the start of trading to 30 min before the delivery starts. There was no product without transactions in the expanded window.

In addition to the ID3 index (actually its approximation), we are also using the DA prices as external regressors. Both time series are of an hourly resolution and span the 1216 days from 1.01.2015 to 30.04.2018, see Fig. 1. Like many EPF studies, we consider a rolling window scheme and use a 364-day window in order to estimate our models on a sample which is a multiple of the weekly seasonality and covers a full year; for a discussion of calibration window lengths, see Hubicka, Marcjasz, and Weron (2019). Initially, we fit our models to data from 1.01.2015 h 1 to 30.12.2015 h 24, and compute the price forecasts for the first hour of 31.12.2015. Next, the window is rolled forward by one hour, the models are re-estimated, and the predictions for the second hour of 31.12.2015 are generated. This procedure is repeated until forecasts for the last hour in the 852-day out-of-sample test period (i.e., 30.04.2018 h 24) have been made.

3.2. The forecasting framework

We denote the intraday and day-ahead electricity prices at time (hour) $t = 24d+h$ by P_t and S_t respectively, where $d = 0, 1, \ldots, 1215$ is the day in our sample and $h = 1, 2, \ldots, 24$ the hour of the day. For each hour t in our out-of-sample test period we make a prediction at time $t-4$ of the closing value of the ID3 index for that hour, i.e., P_t. This is illustrated in Fig. 2 using actual transaction data for the period from 12.09.2016 16:00 to 13.09.2016 24:00. Observe the four-hour time lag between the moment when the forecast is made and the time when the delivery starts. For instance, at 12:00 on 13.09.2016 we are forecasting the price for 16:00 (denoted by $\longrightarrow$). The most recent intraday price is for 12:00 (i.e., the hourly contract with delivery between 12:00 and 13:00 on 13.09.2016, denoted by $*$), while the most 'forward-looking' (i.e., beyond the target hour) DA price is for hour 24 on 13.09.2016. One hour later, at 13:00 on 13.09.2016, we are forecasting the price for 17:00 and the most recent intraday price is for 13:00. However, since the day-ahead auction results are known a few minutes after 12:00, the most 'forward looking' DA price is for hour 24 on 14.09.2016.

Following the recommendations set forth by Uniejewski, Weron, and Ziel (2018), we calibrate our models (except for the **Naive** benchmark; see Section 3.3) not to raw prices but to transformed; i.e., $X_t = f(P_t)$, where $f(\cdot)$ is

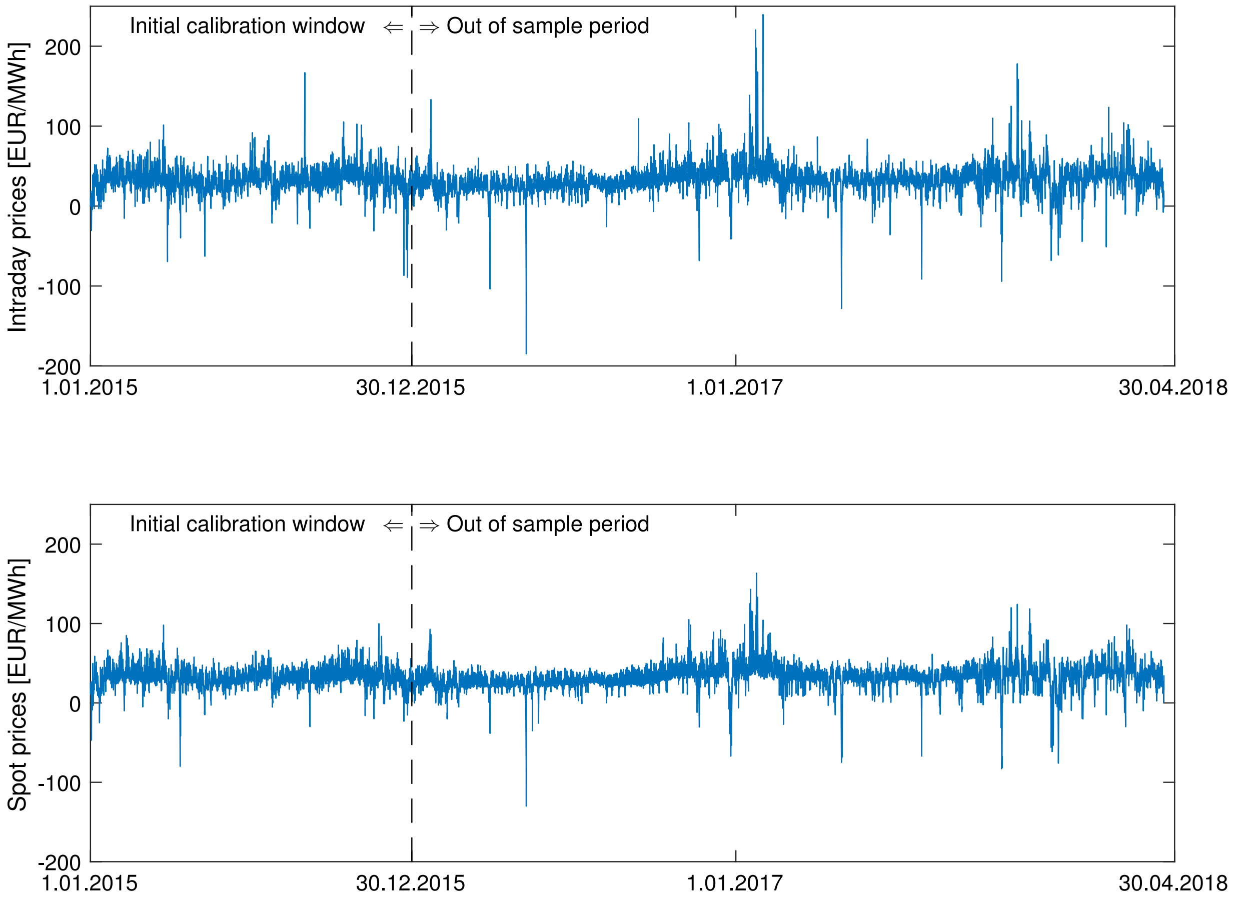


Fig. 1. EPEX hourly intraday (top) and day-ahead (bottom) prices for the period 1.01.2015 to 30.04.2018. The vertical dashed lines mark the beginning of the 852-day out-of-sample test period.

an appropriately chosen *variance stabilizing transformation* (VST). The idea underlying a VST is that of reducing the price variation. As was argued by Janczura, Trück, Weron, and Wolff (2013), a lower variation and/or less spiky behavior of the input data usually allows the forecasting model to yield more accurate predictions.

For electricity markets with only positive prices, the logarithm is the most popular choice for a VST. However, the log-transform is not feasible in our case, since EPEX prices exhibit negative values.[2] Instead, we utilize the *area hyperbolic sine* transformation:

$$X_t = \textbf{asinh}(p_t) \equiv \log\left(p_t + \sqrt{p_t^2 + 1}\right), \tag{1}$$

where $p_t = \frac{1}{b}(P_t - a)$ are 'normalized' prices, a is the median of P_t in the calibration window, and b is the sample *median absolute deviation* (MAD) around the median; the latter two parameters are recomputed every day, separately for the ID3 index and the DA prices. Note that the median of all prices after applying the **asinh** transformation is zero, and the variance is close to one. Note also that this transformation can be used for negative data, and its implementation is straightforward. Moreover, it has been found to perform well in a number of EPF studies (Schneider, 2011; Uniejewski & Weron, 2018; Ziel & Weron, 2018). The inverse transformation is the *hyperbolic sine*, i.e., $p_t = \sinh(X_t)$. After computing the forecasts, we apply it to obtain the price predictions:

$$\widehat{P}_t = b \sinh(\widehat{X}_t) + a. \tag{2}$$

We transform the exogenous series analogously: $Y_t = \textbf{asinh}(s_t)$, where s_t is the normalized day-ahead price S_t.

3.3. Benchmark models

The first benchmark, denoted by **Naive**, is based on the assumption that the day-ahead and intraday markets are driven by similar data generating processes. It is defined by $\widehat{P}_t = S_t$, where S_t is the DA price for the same day and hour (recall that it is set at noon on day $d-1$). The second benchmark is a parsimonious autoregressive structure inspired by the well-performing $expert_{DoW,nl}$ model of Ziel and Weron (2018). In this model, denoted by **ARX**, the VST-transformed price at time t is given by:

$$X_t = \beta_1 X_{t-4} + \beta_2 X_{t-24} + \beta_3 X_{t-48} + \beta_4 X_{t-168} + \beta_5 Y_t + \sum_{i=1}^{7} \beta_{5+i} D_i + \varepsilon_t, \tag{3}$$

where X_{t-24}, X_{t-48} and X_{t-168} account for the autoregressive effects of the previous days (the same hour yesterday,

[2] Note that negative prices are natural in electricity trading: since plant flexibility is limited (especially for coal-fired power plants) and costly, incurring a negative price for a few hours can actually be economically optimal (Gianfreda, Parisio, & Pelagatti, 2018; Schneider, 2011; Weron, 2006).

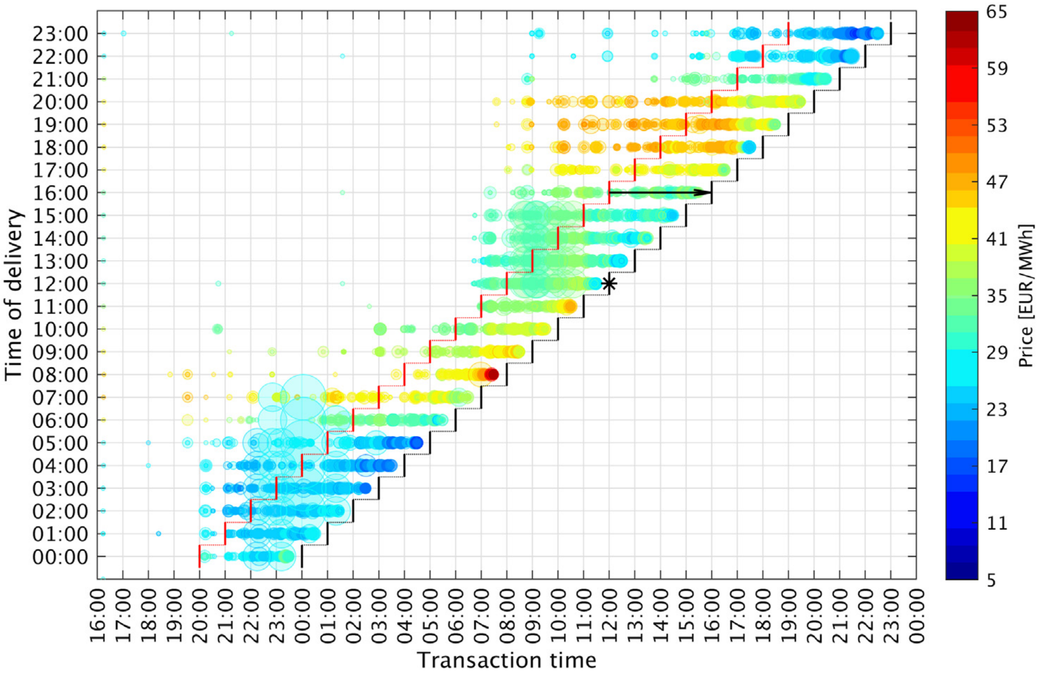


Fig. 2. Illustration of the forecasting framework using actual transaction data for the period from 12.09.2016 16:00 to 13.09.2016 24:00. The black step function indicates the time when the delivery starts (every hour of 13.09.2016), the circles refer to actual trades (the circle size represents the traded volume, from 0.1 to 200 MWh, while the color represents the price, see the colorbar on the right), and the red step function represents the time when the forecasts are made. For instance, when forecasting the price for 16:00 (⟶) at 12:00 on 13.09.2016, the most recent intraday price is for 12:00 (∗).

two days ago, and one week ago), X_{t-4} is the last observed intraday price, and Y_t is the VST-transformed DA price for the same day and hour. The seven dummy variables $D_1, \ldots, D_7$ account for the weekly seasonality, and are defined as $D_i = 1$ for day of the week i and zero otherwise. Finally, the ε_ts are assumed to be independent and identically distributed normal variables. The **ARX** model is estimated via ordinary least squares (OLS).

3.4. LASSO-estimated models

One advantage of using automatic variable selection is the ability to start out by considering an almost unlimited number of explanatory variables. This study utilizes a baseline model with between 349 (for hour 16) and 372 (for hour 17) potential regressors: seven dummy variables (to account for the weekly seasonality, as in the **ARX** benchmark), 165 last known prices from the intraday market (i.e., nearly the whole week), 169 prices from the DA market (i.e., one week of past prices) and between 8 (for hour 16) and 31 (for hour 17) 'forward-looking' prices from the DA market (i.e., $Y_{t+1}, \ldots, Y_{t+31}$; see also Table 4):

$$X_t = \underbrace{\sum_{k=1}^{7} \beta_k D_k}_{\text{weekday dummies}} + \underbrace{\sum_{i=4}^{168} \beta_{i+4} X_{t-i}}_{\text{past intraday prices}} + \underbrace{\sum_{j=0}^{168} \beta_{173+j} Y_{t-j}}_{\text{current and past DA prices}} + \underbrace{\sum_{j=-31}^{-1} \beta_{373+j} Y_{t-j}}_{\text{'forward-looking' DA prices}} + \varepsilon_t. \quad (4)$$

Note that the price for each hour is predicted with a four-hour time lag (which is why the second sum in the formula above starts with $i = 4$) and using the most recent information. Note also that Eq. (4) does not consider any fundamental regressors. However, the results of Andrade et al. (2017) and Monteiro et al. (2016) suggest that fundamentals (historical and predicted demand, generation and weather) do not have much explanatory power when forecasting intraday electricity prices. Perhaps the DA price for the same day and hour already includes this information.

In order to explain the LASSO scheme, let us rewrite Eq. (4) in a more compact form:

$$X_t = \sum_{i=1}^{n} \beta_i V_t^i, \quad (5)$$

where V_t^i are the regressors and β_i are the corresponding coefficients.

The LASSO of Tibshirani (1996) can be treated as a generalization of a linear regression, where instead of minimizing only the *residual sum of squares* (RSS), we minimize the sum of RSS and a linear penalty function of the β_is:

$$\hat{\boldsymbol{\beta}}^L = \min_{\boldsymbol{\beta}} \{\text{RSS} + \lambda \|\boldsymbol{\beta}\|_1\} = \min_{\boldsymbol{\beta}} \left\{ \text{RSS} + \lambda \sum_{i=1}^{n} |\beta_i| \right\}, \quad (6)$$

where $\lambda \geq 0$ is a *tuning* (or *regularization*) parameter. Note that for $\lambda = 0$ we get the standard OLS estimator; for large λs all β_is become zero; and for intermediate values of λ there is a balance between minimizing the RSS and

Table 1
Mean absolute errors (MAE) and root mean squared errors (RMSE) for the two benchmarks (**Naive**, **ARX**) and the ten **LASSO**($\boldsymbol{\lambda}_i$) models, with $\lambda_i = 10^{-\frac{19-i}{6}}$, $i = 1, \ldots, 10$, over the 852-day out-of-sample test period, see Fig. 1. A heat map is used to indicate better (→ green) and worse (→ red) performing models.

	Naive	ARX	LASSO									
			λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
MAE	5.0323	4.7234	5.2198	4.9454	4.7323	4.5728	4.4672	4.4135	4.4128	4.4680	4.5704	4.6928
RMSE	8.1098	7.6513	8.0331	7.6670	7.4196	7.2370	7.1122	7.0721	7.1095	7.2356	7.4416	7.6834

shrinking the coefficients towards zero (and each other), and hence performing variable selection.

Selecting a ‘good’ value for λ is critical. Based on the results of Uniejewski and Weron (2018), we have limited our computations to a log-spaced grid of ten values: $\lambda_i = 10^{-\frac{19-i}{6}}$ for $i = 1, \ldots, 10$. Since our dataset is relatively short compared to that used by Uniejewski and Weron (2018), we have not opted to select one optimal λ based on the model’s performance in a validation period. Also, we have not chosen the value of λ that maximizes an in-sample information criterion (as per Ziel & Weron, 2018), as this could lead to underperforming models. Since the focus of this study is on variable selection, not on the implementation of the LASSO, we have decided to show the results for all ten λs instead. The models are denoted later in the text by **LASSO**($\boldsymbol{\lambda}_i$), or simply by $\boldsymbol{\lambda}_i$.

4. Empirical results

4.1. Forecast evaluation

We use the *mean absolute error* (MAE) and the *root mean squared error* (RMSE) for the full out-of-sample test period of $D = 852$ days (i.e., 31.12.2015 to 30.04.2018; see Fig. 1) as the main evaluation criteria:

$$\text{MAE} = \frac{1}{24D}\sum_{t=1}^{24D}|\varepsilon_t| \quad \text{and}$$

$$\text{RMSE} = \sqrt{\frac{1}{24D}\sum_{t=1}^{24D}\varepsilon_t^2}, \tag{7}$$

where $\varepsilon_t = P_t - \widehat{P}_t$ is the prediction error at time (hour) t. Recall that the RMSE is the optimal measure for least square problems, but is more sensitive to outliers than the MAE. The MAE and RMSE values obtained can be used to provide a ranking of models, but do not allow us to draw statistically significant conclusions as to the outperformance of the forecasts of one model by those of another. Therefore, we use the (Diebold & Mariano, 1995) test, which is simply an asymptotic z-test of the hypothesis that the mean of the loss differential series:

$$\Delta_{A,B,t} = |\varepsilon_{A,t}|^i - |\varepsilon_{B,t}|^i \tag{8}$$

is zero, where $\varepsilon_{Z,t}$ is the prediction error of model Z at time t and $i = 1, 2$ correspond to the absolute and squared loss, respectively; here, ‘model Z’ is used to refer to one of the benchmarks or the LASSO-estimated models. For each model pair and each dataset we compute the p-values of two one-sided tests: (i) a test with the null hypothesis $H_0 : E(\Delta_{A,B,t}) \leq 0$, i.e., the outperformance of the forecasts of B by those of A, and (ii) the complementary test with the reverse null $H_0^R : E(\Delta_{A,B,t}) \geq 0$, i.e., the outperformance of the forecasts of A by those of B. The loss differential series thus obtained are covariance stationary.

Table 1 reports the MAE and RMSE errors for the benchmarks and the LASSO models, while Fig. 3 depicts the results of the DM tests. We use a heat map to indicate the range of the p-values: the closer they are to zero (→ dark green), the more significant the difference between the forecasts of a set on the X-axis (better) and the forecasts of a set on the Y-axis (worse). For instance, the first row in both panels of Fig. 3 is green except for one black square, indicating that — irrespective of whether we are considering absolute or squared losses — the forecasts of the **Naive** benchmark are outperformed significantly by those of all other models except **LASSO**($\boldsymbol{\lambda}_1$). On the other hand, the column that corresponds to the **LASSO**($\boldsymbol{\lambda}_6$) model is green in the right panel, meaning that this model leads to significantly better forecasts than all others when considering squared losses.

Table 1 clearly shows that the **Naive** benchmark and the **LASSO**($\boldsymbol{\lambda}_1$) model are the worst predictors. Somewhat surprisingly, the **Naive** benchmark even significantly outperforms the worst LASSO model for absolute losses; see the green square in the left panel of Fig. 3 in the **Naive** column. Obviously, the DA price S_t is a good predictor of the intraday price P_t for the same day and hour. However, we can do better than that.

Indeed, even the reasonably parsimonious **ARX** model with only 12 regressors, including seven weekday dummies, significantly outperforms the naive benchmark and one or two of the worst LASSO models, namely $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ (for absolute losses only); see the green squares in the columns labeled **ARX** in Fig. 3. Moreover, it is not outperformed significantly by the naive benchmark or the three worst LASSO models, i.e., $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$ and $\boldsymbol{\lambda}_3$ (for absolute losses) or $\boldsymbol{\lambda}_{10}$ (for squared losses); see the black squares in the rows labeled **ARX**. Now, comparing the LASSO-estimated models among themselves, we can see that the best predictions are obtained for $\boldsymbol{\lambda}_7$ (according to MAE) and $\boldsymbol{\lambda}_6$ (according to RMSE). However, while $\boldsymbol{\lambda}_6$ is a clear winner for squared losses, there are no significant

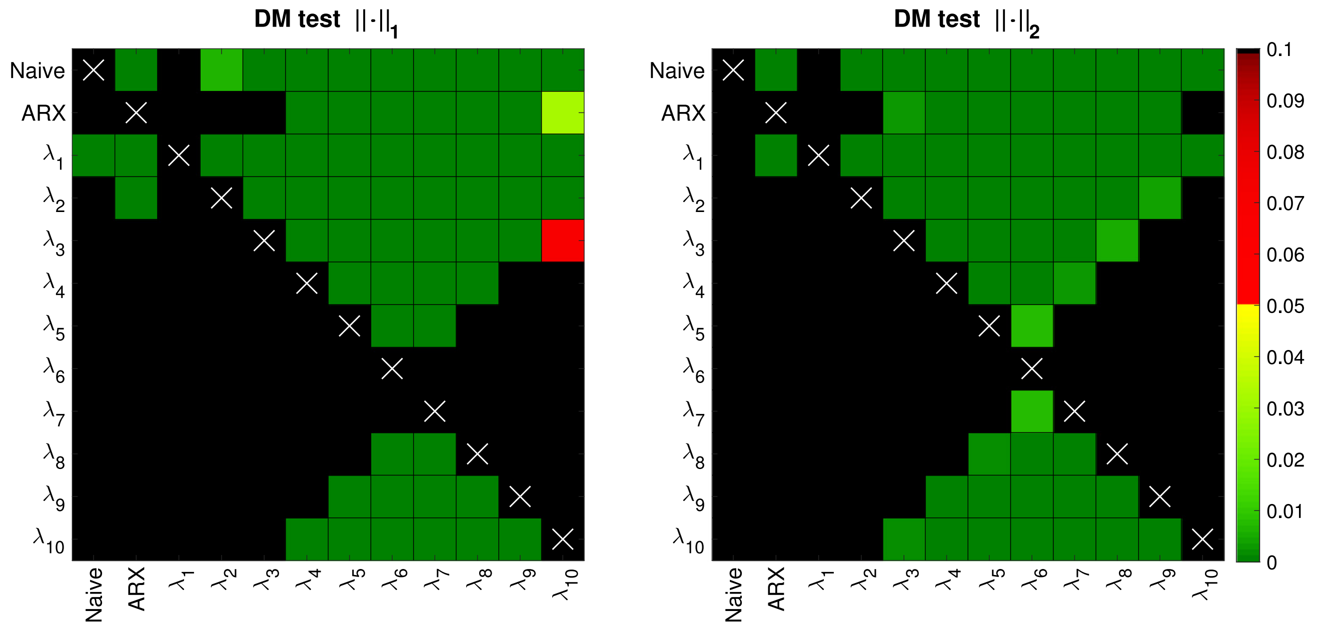


Fig. 3. Results of the Diebold–Mariano (DM) test for the absolute (left) and squared (right) prediction errors, for the same models as in Table 1. A heat map is used to indicate the range of p-values: the closer they are to zero (→ dark green), the more significant the difference is between the forecasts of a model on the X-axis (better) and those of a model on the Y-axis (worse). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

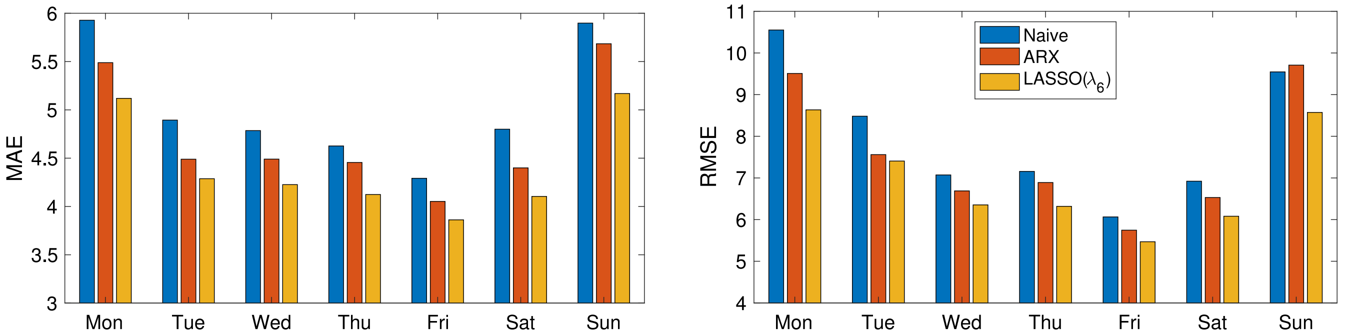


Fig. 4. MAE (left panel) and RMSE (right panel) errors for the two benchmarks (**Naive**, **ARX**) and the **LASSO**($\boldsymbol{\lambda}_6$) model, separately for each day of the week in the 852-day out-of-sample test period, see Fig. 1.

differences in predictive accuracy between $\boldsymbol{\lambda}_6$ and $\boldsymbol{\lambda}_7$ for absolute losses, see Fig. 3. Overall, the picture is consistent with other recent studies that have used the LASSO for day-ahead EPF (Uniejewski et al., 2016; Uniejewski & Weron, 2018; Ziel, 2016; Ziel & Weron, 2018).

4.2. Economic benefits from a simple trading strategy

We now give the error measures above a financial interpretation by considering a simple trading strategy that a participant in the German intraday market can execute. Assume that we want to trade a unit of electricity, say 1 MWh, for each hourly delivery period throughout the 852-day out-of-sample test period; see Fig. 1. We take a long position if the intraday price three hours before the delivery starts (to be precise: the first transaction price after the three hours to delivery time stamp) is lower than our **LASSO**($\boldsymbol{\lambda}_6$) forecast $\widehat{P}_t$ of the ID3 index, and a short position otherwise. We close the position at the first price that exceeds $\widehat{P}_t$; i.e., the first price that is higher than $\widehat{P}_t$ in the case of a long position and the first price that is lower than $\widehat{P}_t$ in the case of a short position. If our ID3 forecast is not breached, then we close the position at the last traded price (to be precise: the last transaction price before the 30 min to delivery time stamp), possibly with a loss. Assuming that there are no transaction costs, this simple strategy leads to a profit of 0.29 EUR/MWh on average across the whole out-of-sample test period. This result clearly shows the usefulness of our approach relative to a profit of 0.03 EUR/MWh from using the **ARX** model and a loss of 0.18 EUR/MWh from using the **Naive** benchmark.

4.3. Performances across days of the week and price regimes

Let us now have a closer look at model performances across days of the week. Fig. 4 plots the MAE (left panel)

Table 2
MAE (top panel) and RMSE (bottom panel) values for the benchmarks (**Naive**, **ARX**) and the ten **LASSO**($\boldsymbol{\lambda}_i$) models with $\lambda_i = 10^{-\frac{19-i}{6}}$, $i = 1, \ldots, 10$, in the 852-day out-of-sample test period, see Fig. 1. As in Table 1, a heat map is used to indicate better (→ green) and worse (→ red) performing models.

MAE	Naive	ARX	LASSO									
			λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
positive spike	29.7105	28.3006	25.8022	26.2783	26.5414	26.7997	27.2811	27.8916	28.2491	28.9377	29.7935	31.1033
normal range	4.6745	4.3522	4.9090	4.6296	4.4127	4.2455	4.1309	4.0652	4.0535	4.0947	4.1805	4.2814
negative spike	30.1848	33.6440	28.2448	27.7597	27.5162	28.0629	28.6342	29.6337	30.9765	32.4805	34.2896	36.3411

RMSE	Naive	ARX	LASSO									
			λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
positive spike	40.0163	36.9870	34.1606	34.4266	34.7280	35.1616	35.6512	36.1430	36.6081	37.3888	38.5202	40.1076
normal range	6.7899	6.2017	6.9715	6.5417	6.2307	5.9752	5.7921	5.6875	5.6587	5.7086	5.8241	5.9659
negative spike	34.6055	39.6889	34.3290	33.8993	33.9294	34.2932	34.4089	35.1922	36.5425	38.2854	40.4355	42.4274

and RMSE (right panel) values for the two benchmarks (**Naive**, **ARX**) and the best LASSO model, namely $\boldsymbol{\lambda}_6$, separately for each day of the week in the 852-day out-of-sample test period; see Fig. 1. Clearly, the **LASSO**($\boldsymbol{\lambda}_6$) model is better than the benchmarks across all days of the week and for both error measures. As expected, **ARX** outperforms the much simpler **Naive** benchmark, except under the RMSE on Sundays (which is rather surprising).

Following Uniejewski et al. (2018), we obtain a better understanding of model performances across price regimes by considering an evaluation conducted separately for three subsamples defined using the 3σ-rule: (i) a *positive spike* regime: $\mu + 3\sigma < P_t$, (ii) the *normal range*: $\mu - 3\sigma < P_t < \mu + 3\sigma$, and (iii) a *negative spike* regime: $P_t < \mu - 3\sigma$, where μ is the sample mean and σ is the sample standard deviation of P_t in the 852-day out-of-sample test period, see Fig. 1. Overall, there are only 169 (0.83%) positive and 121 (0.59%) negative spikes, and as many as 20,158 (98.58%) 'normal' prices in the test period.

Table 2 reports the MAE and RMSE values for the benchmarks and the LASSO models across the three price regimes. According to both measures, the normal range requires larger λs than the spike regimes, which corresponds to a smaller number of regressors. The best predictions are obtained for $\boldsymbol{\lambda}_7$ (according to both error measures), with $\boldsymbol{\lambda}_6$ following closely behind. On the other hand, positive spikes are captured best by the model with the smallest smoothing parameter, namely $\boldsymbol{\lambda}_1$, while negative spikes are captured best by either $\boldsymbol{\lambda}_2$ or $\boldsymbol{\lambda}_3$. Apparently, extreme prices require much more complex models, utilizing dependencies among many regressors. Finally, as expected, the MAE and RMSE values for the normal range in Table 2 are smaller than those for the full sample in Table 1.

4.4. Variable selection

Let us now comment on variable selection by looking at the results of the model that performed best overall, namely **LASSO**($\boldsymbol{\lambda}_6$). Tables 3 and 4 report the mean occurrences (in %) of model parameters across the 852-day out-of-sample test period. A heat map is used to indicate more (→ green) and less (→ red) commonly selected variables. Several interesting conclusions can be drawn:

- The most important variables are the most recent intraday price (i.e., X_{t-4}; see the bottom row in Table 3 with '100' in all columns) and the DA price (i.e., Y_t; see the row labeled '0' in Table 4 with '100' in nearly all columns) that correspond to the predicted hour. Interestingly, the DA prices for the nearby hours (lags -2, -1 and 1) also tend to be selected by the LASSO, which may be an indication that Y_t is not a perfect estimate of X_t and that the prices for the neighboring hours include valuable information.
- Surprisingly, the impact of the previous day's intraday price for the same hour is hardly visible. Hence, there is no reason to have X_{t-24} as an explanatory variable in parsimonious expert models for the intraday market. This is in stark contrast to day-ahead EPF models, where the previous day's price for the same hour is typically one of the most important regressors (Amjady & Keynia, 2009; Karakatsani & Bunn, 2008; Keles et al., 2016; Uniejewski et al., 2016; Ziel & Weron, 2018).
- As was observed by Maciejowska and Nowotarski (2016) and Ziel (2016), the prices for not only hour 24, but also the nearby evening hours, are important predictors. This can be seen by the yellow-green diagonals in Table 3. Note that the first full diagonal (from the bottom, i.e., lags 4 to 27) corresponds to hour 21 of day $d-1$, the second (lags 5 to 28) to hour 20 of day $d-1$, etc. A similar effect can be observed for DA prices; note the yellow-green diagonals starting at the rows labeled '1' (corresponding to hour 24) and '2' (corresponding to hour 23) in Table 4.
- Somewhat surprisingly, the 'forward-looking' DA prices are rarely selected. A notable exception is the price for hour 1 on day $d+1$; see the mostly green

Table 3
Mean occurrence (in %) of model parameters across the 852-day out-of-sample test period. The columns represent the hours ($h = 1, \ldots, 24$) for which the price predictions are made and the rows represent the parameters of the **LASSO**(λ_6) model, see Eq. (4), that correspond to the seven daily dummies and to the past intraday prices X_{t-i}, $i = 168, 167, \ldots, 4$. A heat map is used to indicate more (→ green) and less (→ red) commonly-selected variables.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Sun	0	12	12	13	14	18	39	57	63	41	7	5	0	0	17	17	58	31	0	4	1	8	8	3
Mon	1	0	10	19	12	0	0	2	22	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0
Tue	0	0	0	1	2	4	25	32	19	6	27	4	1	5	27	18	13	12	1	0	4	0	0	0
Wed	10	8	19	1	0	0	2	0	0	0	0	0	0	0	0	0	1	3	9	15	0	0	0	0
Thu	11	2	2	2	1	0	0	0	0	0	3	9	11	0	0	0	8	2	3	0	0	0	18	0
Fri	0	0	0	0	0	4	2	0	0	0	0	0	4	0	0	0	0	0	1	0	7	0	0	0
Sat	0	1	0	0	0	0	14	18	0	15	28	16	3	7	4	0	2	34	29	4	10	0	0	0
168	7	0	19	12	8	1	21	47	29	43	48	14	22	38	44	9	48	8	35	20	3	13	9	34
167	1	41	12	15	1	16	4	16	1	14	33	2	19	49	16	45	0	17	2	7	2	6	5	0
166	45	12	2	0	1	3	17	0	0	65	2	4	55	0	34	12	5	0	15	0	66	20	0	0
165	19	2	14	3	12	0	0	0	69	1	0	49	6	8	31	13	12	4	1	60	36	0	28	21
164	0	0	1	1	6	8	4	42	5	6	47	39	5	17	26	35	0	34	61	54	0	0	2	15
163	0	0	32	0	10	2	2	4	0	19	32	9	14	37	37	0	29	50	61	0	19	0	1	0
162	7	22	0	0	7	0	10	9	9	6	2	1	0	15	10	28	34	84	7	30	3	2	10	0
161	25	1	0	2	0	9	1	20	9	13	0	2	9	10	23	37	85	0	34	4	24	25	5	35
160	0	6	1	13	0	16	6	16	23	11	0	22	9	10	13	32	0	0	2	13	21	0	14	63
159	0	6	12	0	0	7	6	35	4	1	29	36	30	1	12	3	0	10	14	13	0	0	42	6
158	9	3	0	1	0	32	19	7	8	2	9	35	3	8	1	1	0	0	21	0	2	47	3	9
157	8	13	1	14	31	5	0	10	2	0	40	9	23	0	19	2	2	0	0	0	4	0	28	0
156	2	0	30	9	61	2	43	10	32	49	9	12	3	6	1	12	12	11	0	24	2	35	28	4
155	7	40	0	38	3	1	1	39	41	52	10	6	7	12	6	10	3	1	6	11	66	19	14	10
154	38	0	16	25	0	0	27	39	25	2	0	17	3	4	11	45	5	21	14	18	5	11	12	1
153	9	2	13	0	0	4	41	8	5	2	0	0	2	0	67	16	18	13	30	2	0	1	12	46
152	0	0	3	0	1	34	25	4	3	1	0	1	1	42	38	29	6	7	0	2	9	5	22	4
151	7	17	3	0	14	51	0	12	10	0	2	1	33	41	9	5	0	8	2	14	3	18	1	24
150	3	9	3	11	32	1	18	0	0	0	12	15	44	0	4	0	26	28	23	0	27	1	40	1
149	38	0	8	34	0	0	2	0	0	12	13	29	0	12	8	3	1	18	0	8	22	54	14	0
148	2	4	32	4	7	0	1	8	0	1	24	3	11	11	12	26	7	0	4	20	48	13	0	18
147	5	6	3	3	1	0	4	0	12	37	0	8	16	25	27	3	9	3	8	5	8	0	22	65
146	0	4	17	4	0	11	0	2	11	18	15	4	6	13	11	21	2	25	1	1	0	13	38	1
145	17	17	8	27	0	0	33	58	36	13	8	1	0	14	10	3	37	26	6	3	25	15	0	2
144	18	5	7	14	9	20	29	29	2	34	7	15	36	2	0	35	29	14	38	41	14	1	6	25
143	1	7	11	0	10	1	30	3	45	0	50	39	2	9	38	9	16	1	5	3	5	24	10	15
142	4	0	5	38	0	4	12	18	0	29	10	0	7	35	5	14	0	0	0	15	56	0	51	9
141	0	19	26	0	47	63	26	27	19	34	0	8	26	0	34	0	0	7	38	36	15	46	21	0
140	10	11	0	57	69	0	7	19	0	15	24	21	0	37	9	10	0	10	22	20	23	0	7	22
139	2	14	54	82	0	3	13	0	19	6	26	1	40	22	13	22	1	40	9	0	4	13	28	3
138	12	72	71	0	0	0	9	4	32	34	0	30	19	16	40	1	42	5	0	0	14	1	2	13
137	63	52	0	0	0	32	7	0	22	0	13	7	8	44	4	12	16	7	0	0	1	3	11	13
136	32	0	0	0	31	7	4	0	0	11	1	3	26	15	13	12	12	0	7	0	2	4	23	28
135	4	13	3	34	13	9	2	2	11	6	0	23	20	3	3	6	5	17	6	27	17	14	15	0
134	25	13	0	14	17	21	41	13	0	16	12	42	3	17	0	8	26	3	36	0	13	8	5	5
133	10	3	11	0	10	36	1	11	31	10	25	6	3	0	0	10	0	30	3	2	8	9	11	1
132	0	4	5	6	27	0	0	15	10	27	40	16	0	2	0	11	12	0	9	18	10	5	15	16
131	12	1	10	0	0	0	26	5	24	25	20	0	0	8	20	0	0	20	21	8	2	3	0	2
130	0	37	0	0	2	3	6	28	29	21	0	1	8	14	6	1	28	18	22	1	0	3	11	26
129	27	0	1	52	24	0	5	14	21	13	0	9	20	5	12	25	14	5	7	10	0	0	3	2
128	7	0	40	2	1	4	23	32	11	4	8	8	14	8	28	30	12	1	15	7	3	3	5	0
127	3	42	2	36	0	29	42	16	9	4	9	27	2	1	38	22	22	25	6	13	26	0	2	12
126	19	0	25	0	19	27	1	0	0	6	21	8	0	25	27	26	27	9	10	6	4	0	19	2
125	0	38	7	34	0	0	0	0	9	25	3	1	28	13	0	13	4	14	19	3	2	29	4	0
124	42	1	49	6	0	52	0	12	11	16	1	48	12	12	17	0	2	13	0	1	24	23	0	9
123	0	36	3	0	36	0	17	0	25	7	16	20	3	0	1	0	15	0	0	46	15	2	0	43
122	0	0	3	1	0	0	0	12	13	12	58	0	0	0	11	11	1	4	50	1	1	9	57	7
121	1	3	0	0	0	3	18	9	8	51	3	1	8	2	0	2	20	15	7	2	42	11	20	9
120	5	2	0	0	0	12	17	40	38	18	9	4	0	3	9	16	4	30	8	77	8	13	0	13
119	0	1	0	1	0	7	37	15	12	19	14	6	22	12	24	13	32	9	4	7	0	2	13	0
118	0	1	0	19	0	35	56	1	29	3	3	15	25	9	1	16	10	1	8	1	4	14	8	2
117	0	0	6	6	19	71	22	15	16	7	9	20	29	0	3	12	0	20	3	10	23	0	0	10
116	3	4	33	43	26	4	11	34	0	15	28	21	7	18	15	8	27	1	0	13	4	0	1	26
115	8	43	45	11	23	0	23	3	2	12	14	10	15	9	21	17	14	1	57	23	0	2	18	13
114	21	56	0	1	0	26	12	10	2	14	25	20	5	12	26	30	5	30	17	0	0	11	15	1
113	22	10	0	11	6	2	24	0	5	4	17	28	11	8	8	11	11	0	0	2	3	1	9	0
112	12	0	13	0	3	1	4	0	9	8	5	5	4	14	9	2	6	0	17	0	1	3	1	5
111	1	0	0	2	8	0	0	20	1	7	10	0	13	10	3	4	0	16	0	0	5	19	6	62
110	5	4	33	3	3	8	0	5	2	25	34	5	1	25	4	0	5	10	0	3	25	5	24	28
109	11	9	2	2	0	17	8	0	33	49	5	0	25	0	0	2	50	2	19	29	54	0	48	17
108	7	19	18	1	10	8	7	44	57	30	0	15	4	0	9	53	26	0	4	47	20	13	2	3
107	4	19	1	29	2	8	46	39	20	0	13	24	3	0	33	0	10	0	51	13	22	5	18	0
106	7	37	3	7	0	2	35	27	0	7	4	7	0	5	0	12	18	5	27	35	5	4	10	27
105	42	9	0	1	5	53	16	0	3	1	0	19	9	12	0	39	2	1	21	1	3	18	12	11
104	11	1	0	34	18	14	7	0	20	0	18	6	8	19	17	27	0	0	1	0	9	0	0	9
103	12	0	19	14	7	1	1	0	0	4	31	7	23	25	40	0	2	5	1	1	0	18	10	0
102	3	1	21	0	0	0	1	0	0	7	12	32	21	43	0	41	43	1	0	0	12	3	20	0
101	1	9	0	11	8	0	2	2	1	17	37	39	62	0	2	25	0	1	4	17	3	7	20	1
100	5	0	2	2	6	5	16	31	5	43	43	13	0	1	19	0	0	1	18	3	39	3	4	4
99	13	4	0	6	5	3	41	0	50	34	19	0	13	3	20	2	18	7	2	0	0	0	7	16
98	25	0	9	1	2	15	5	18	14	1	23	20	17	16	0	3	1	4	1	37	6	0	9	0
97	2	14	0	0	13	32	0	4	1	4	34	13	4	3	2	0	0	4	35	16	9	14	19	54
96	10	1	4	8	15	25	26	20	21	38	4	0	17	5	12	0	0	51	22	69	68	41	20	0
95	18	3	6	3	13	7	22	4	32	4	0	16	3	11	0	0	40	9	46	17	8	9	2	0
94	12	13	8	5	26	28	10	21	14	1	8	0	16	8	3	15	17	24	7	0	0	0	2	24
93	0	0	4	5	32	12	25	20	2	4	3	23	4	15	21	19	25	19	0	3	33	1	43	1
92	0	2	16	19	7	5	9	0	14	5	15	3	21	4	10	18	1	3	6	34	11	6	1	0
91	0	12	15	11	3	31	6	1	27	21	1	22	0	2	0	33	24	10	2	17	12	1	22	7
90	11	1	12	3	17	0	0	29	22	19	2	1	3	1	6	6	31	7	15	9	6	25	24	4

(continued on next page)

Table 3 (*continued*).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
89	2	14	12	18	1	12	7	3	1	4	3	4	5	24	7	37	19	1	12	2	30	42	42	0
88	20	7	5	16	1	2	15	2	4	1	1	2	10	1	4	8	2	17	0	28	46	33	8	41
87	17	0	7	12	1	15	3	33	21	0	35	4	29	1	29	23	7	0	17	29	17	12	38	5
86	0	4	11	0	12	4	17	21	1	31	0	0	17	23	3	19	3	7	15	15	4	5	8	0
85	4	25	18	6	0	0	21	0	24	4	12	22	18	28	0	25	0	6	3	1	38	7	1	1
84	6	13	3	2	0	2	8	33	1	6	30	2	2	5	3	0	8	0	0	23	6	21	7	53
83	13	6	6	0	7	4	9	0	25	31	9	2	39	4	7	34	8	7	15	22	44	30	34	10
82	3	8	0	34	3	1	0	2	39	20	0	47	0	4	46	5	2	23	25	25	41	7	18	9
81	13	0	21	3	2	5	12	29	22	1	5	0	7	46	4	11	16	18	25	24	0	19	2	12
80	1	24	33	1	34	10	6	17	2	19	3	3	41	4	4	15	13	0	10	0	2	8	22	0
79	11	35	0	14	24	23	22	0	2	4	0	37	29	29	23	22	8	16	18	3	6	12	0	0
78	0	9	16	26	36	0	0	0	6	20	14	31	28	19	2	6	9	45	0	10	21	0	1	22
77	35	34	27	1	1	0	23	37	13	10	23	35	12	12	43	21	7	0	18	20	26	8	12	35
76	35	11	8	25	0	39	25	14	4	6	42	0	16	46	30	1	7	3	3	19	28	30	33	86
75	0	2	60	8	35	27	25	10	21	22	6	0	18	23	2	8	1	3	0	8	22	37	91	34
74	0	78	33	54	31	36	9	17	2	31	19	2	9	20	1	10	0	0	4	15	0	6	16	9
73	69	11	4	24	24	11	1	23	35	11	1	29	16	0	2	9	0	0	12	1	6	7	30	20
72	22	7	38	27	34	3	26	27	7	6	13	0	3	2	3	8	2	25	12	0	21	23	35	25
71	6	8	20	14	27	40	3	8	2	2	8	0	15	2	11	13	22	26	0	6	17	33	12	1
70	4	4	35	41	19	1	8	39	1	4	11	25	1	9	32	38	20	19	12	1	11	21	5	0
69	2	15	13	5	0	56	45	12	26	10	32	17	0	26	38	24	11	4	6	10	17	2	34	0
68	4	16	3	9	24	7	0	10	2	29	13	0	37	20	38	2	21	3	0	12	2	28	5	2
67	22	7	3	35	23	0	13	7	21	13	6	44	2	29	3	23	0	19	12	44	24	3	9	0
66	22	3	23	0	23	11	7	7	13	9	7	1	41	10	21	0	26	32	45	0	25	21	27	26
65	10	20	0	29	0	0	9	0	16	1	3	19	15	7	10	3	11	33	0	1	23	3	19	30
64	4	0	3	2	2	5	1	6	2	14	12	41	1	25	4	13	1	6	0	0	3	1	43	2
63	5	0	0	32	6	1	3	28	39	2	41	3	20	15	24	20	0	0	0	6	0	33	37	0
62	6	0	0	1	0	2	29	36	8	39	27	16	27	17	0	5	5	0	26	0	37	16	2	7
61	4	0	0	0	12	38	14	13	17	16	12	39	25	0	2	0	0	29	8	53	12	13	15	6
60	4	2	1	6	40	0	27	7	13	5	27	61	5	7	2	0	7	22	37	31	10	36	32	1
59	3	8	4	28	4	6	5	6	0	24	70	5	8	10	2	2	42	48	46	8	39	6	31	3
58	5	0	28	0	9	10	22	0	44	71	4	22	4	0	10	41	34	23	1	33	1	40	41	13
57	2	20	0	24	1	8	13	10	40	5	29	0	1	20	21	19	13	14	5	0	17	42	5	7
56	0	0	10	0	2	0	44	7	13	29	0	0	23	2	37	2	9	0	12	0	41	1	0	3
55	0	13	19	4	0	20	3	1	6	4	0	1	22	21	9	23	0	22	8	45	11	0	19	9
54	0	1	31	15	36	2	6	3	2	1	2	6	10	12	14	1	23	17	38	28	7	0	6	3
53	23	4	47	36	0	3	0	14	0	0	3	6	16	15	0	22	28	55	50	6	30	19	0	13
52	39	38	32	2	3	66	0	8	1	9	20	0	5	0	30	27	35	19	38	44	28	46	28	2
51	5	21	0	5	47	0	14	2	38	33	15	9	6	10	7	14	9	0	14	13	19	49	0	0
50	38	0	22	1	4	10	0	35	25	17	13	6	10	10	0	0	2	3	7	0	42	0	12	7
49	0	28	0	7	28	0	18	0	23	12	5	22	8	3	0	11	6	1	13	11	11	16	23	15
48	42	0	14	0	12	9	0	23	8	4	11	16	10	2	2	13	29	71	11	28	48	51	23	5
47	0	8	1	30	30	15	21	46	0	23	18	7	0	7	23	28	84	24	5	11	12	21	0	22
46	2	0	15	28	22	0	36	2	3	12	3	4	0	24	39	79	6	7	10	0	38	4	3	0
45	0	26	26	13	3	24	0	0	9	2	9	6	17	37	31	0	5	13	0	46	1	5	2	22
44	0	7	23	9	15	0	2	13	0	18	5	6	21	0	0	0	14	1	39	10	0	25	4	21
43	0	24	20	19	29	33	4	5	7	10	0	0	11	21	12	16	24	22	7	5	0	2	43	8
42	55	32	12	26	22	0	25	6	33	5	2	2	9	64	16	20	0	12	5	28	7	11	7	7
41	15	34	37	13	0	21	15	34	9	12	21	4	69	18	0	14	5	7	36	1	7	1	0	27
40	38	31	24	0	0	14	12	4	5	17	15	38	16	1	23	27	0	41	2	23	4	12	59	28
39	0	15	0	7	4	11	2	0	7	14	19	38	1	6	36	0	52	7	4	9	12	30	21	7
38	37	4	12	12	0	0	0	1	10	2	13	2	20	39	1	18	2	2	7	3	22	10	25	0
37	0	0	2	26	0	5	8	29	33	22	2	14	23	2	0	12	6	4	22	25	36	30	0	10
36	0	18	19	18	19	1	17	15	24	2	18	2	1	0	9	0	1	12	11	15	14	41	16	1
35	5	21	5	10	0	44	31	33	6	21	6	0	7	5	9	10	7	6	21	25	26	7	11	0
34	27	3	0	14	2	2	0	13	38	14	0	13	1	9	9	27	21	12	38	3	7	1	0	2
33	0	6	9	0	0	3	2	48	12	2	18	0	3	0	7	7	1	47	2	11	1	0	1	0
32	0	17	2	1	18	1	17	11	1	0	8	0	0	5	33	17	36	3	1	7	0	0	0	9
31	35	8	5	20	7	16	22	19	1	14	5	0	0	41	40	24	41	13	3	16	12	0	5	27
30	0	31	28	12	65	1	7	4	3	2	3	1	17	47	7	37	20	0	2	8	0	18	59	27
29	31	34	3	43	28	14	7	1	13	17	4	7	28	3	1	4	0	15	14	5	7	87	4	0
28	15	3	2	41	16	9	1	0	24	12	5	40	15	3	27	0	33	25	36	5	90	55	2	38
27	0	1	17	3	0	0	0	14	0	7	21	13	19	24	8	20	34	43	7	100	30	0	10	28
26	40	11	17	1	0	2	1	4	13	22	10	5	4	6	17	47	4	4	47	22	0	0	26	18
25	6	21	14	2	1	8	6	28	21	7	0	5	3	0	6	1	2	3	0	0	27	0	22	57
24	0	0	36	25	28	32	23	30	20	0	10	18	0	14	2	1	11	38	2	100	52	71	73	45
23	0	33	28	34	32	36	47	9	0	27	22	0	21	37	21	13	35	9	100	18	91	40	55	33
22	36	4	30	39	19	22	2	10	3	17	1	28	42	27	6	5	15	91	4	48	22	39	22	6
21	28	63	15	10	23	0	11	1	6	0	26	56	15	18	13	7	74	0	1	27	28	62	4	22
20	46	59	3	50	0	14	1	13	13	0	39	7	31	11	25	37	0	4	29	22	60	35	13	0
19	35	17	16	0	3	4	0	11	2	7	9	19	8	26	34	3	13	59	15	51	20	1	3	48
18	6	0	0	8	0	0	1	5	9	11	25	22	25	11	5	8	10	2	47	2	0	0	9	3
17	6	13	21	0	0	20	3	7	12	2	20	21	5	15	2	11	34	48	2	0	2	0	30	10
16	1	15	9	1	43	6	0	11	26	6	30	3	21	16	38	41	17	2	2	7	0	13	27	22
15	17	1	3	70	11	9	32	27	8	10	14	49	28	66	18	17	2	3	8	4	0	42	19	21
14	13	0	48	40	2	13	30	1	0	7	72	31	59	19	1	2	5	0	0	1	58	63	0	12
13	4	8	43	7	0	24	4	0	27	56	37	72	8	10	20	0	3	1	7	34	63	0	8	0
12	14	16	7	0	14	1	6	0	39	16	45	5	6	12	0	9	0	14	35	77	16	3	4	26
11	17	11	0	0	4	0	6	57	21	32	1	15	9	17	21	0	10	37	63	26	44	4	43	0
10	6	0	0	19	1	18	27	23	51	5	8	2	15	5	10	37	44	13	76	68	16	20	10	5
9	5	2	4	0	35	38	36	36	4	5	0	3	23	18	28	47	0	87	74	7	21	0	32	8
8	0	13	40	76	25	2	24	5	10	0	3	5	26	26	45	0	73	33	42	7	4	4	4	25
7	17	20	88	42	0	2	25	15	5	4	0	26	57	32	23	71	12	77	17	0	11	7	31	7
6	19	84	61	0	7	10	82	12	56	16	16	38	59	11	18	8	46	65	0	2	7	4	1	27
5	65	9	29	10	3	10	28	36	39	2	33	75	0	0	0	3	51	0	0	1	3	4	33	37
4	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

Table 4
Mean occurrence (in %) of model parameters across the 852-day out-of-sample test period. The columns represent the hours ($h = 1, \ldots, 24$) for which the price predictions are made and the rows represent the parameters of the **LASSO**(λ_6) model, see Eq. (4), that correspond to the past day-ahead (DA) prices Y_{t-i}, $i = 168, 167, \ldots, 1$, the DA price for the target hour Y_t, and the 'forward-looking' DA prices Y_{t-i}, $i = -1, -2, \ldots, -31$. A heat map is used to indicate more (→ green) and less (→ red) commonly-selected variables.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
168	4	5	2	3	4	9	5	1	34	29	22	11	43	21	12	11	38	54	81	57	43	10	18	14
167	0	4	0	4	8	27	0	72	17	17	12	33	3	15	2	18	30	27	11	34	3	7	0	7
166	0	0	1	10	7	5	15	8	9	0	23	19	2	0	7	1	35	0	5	0	0	0	4	0
165	0	0	10	0	20	8	1	7	0	0	27	1	0	7	1	13	0	7	0	1	28	1	5	12
164	1	6	0	30	29	13	3	0	19	30	1	0	14	9	0	0	0	0	1	2	0	13	0	0
163	1	2	43	21	0	1	0	3	19	4	0	5	9	0	3	0	0	27	0	0	3	0	0	0
162	0	38	14	2	0	0	1	6	13	0	0	18	7	0	0	3	18	1	0	0	6	0	12	18
161	64	29	8	0	39	0	3	47	0	0	15	8	9	1	9	14	12	5	0	0	0	6	12	5
160	18	16	1	53	6	0	1	5	0	23	20	0	0	9	6	36	4	0	0	4	11	7	8	24
159	20	7	0	1	0	6	0	7	21	3	0	0	0	11	26	24	0	0	2	34	4	7	1	6
158	7	4	0	0	24	0	17	7	21	16	0	0	0	36	11	0	0	1	23	0	17	0	0	21
157	2	0	0	14	0	9	5	14	18	0	4	0	37	14	0	22	30	40	0	1	5	0	0	0
156	0	0	35	0	8	2	6	14	2	5	0	0	7	0	35	16	8	0	3	3	0	0	0	8
155	0	32	0	0	2	16	24	10	7	10	0	0	9	11	0	4	0	2	26	1	0	5	33	34
154	11	0	6	0	3	24	3	0	15	0	0	1	4	0	3	1	0	31	3	5	30	21	10	6
153	0	0	1	10	15	0	0	7	12	4	1	3	1	0	0	0	31	1	1	36	16	14	3	2
152	0	0	0	0	6	1	0	2	0	0	4	6	0	0	17	25	7	2	20	24	5	3	10	6
151	0	0	13	4	6	2	18	8	0	41	21	0	1	8	6	0	9	0	22	0	2	9	1	0
150	1	17	8	10	32	0	0	0	49	23	0	0	0	15	0	0	1	4	1	0	3	0	1	2
149	37	11	0	7	0	0	0	23	11	1	0	6	32	8	0	18	2	5	5	23	0	2	3	3
148	1	0	13	0	0	0	0	0	0	0	6	33	9	0	19	15	21	5	9	0	48	8	2	27
147	1	8	0	1	0	0	0	0	0	2	36	4	7	15	13	47	0	0	0	34	14	0	0	0
146	0	0	22	8	0	16	0	0	1	11	4	7	0	13	17	0	0	0	2	0	0	2	4	3
145	2	0	7	4	3	2	21	4	3	0	0	0	3	2	3	4	0	0	0	10	14	32	29	3
144	8	5	6	4	9	6	5	0	0	0	1	10	1	6	17	1	23	0	29	77	75	41	29	7
143	1	2	3	22	0	1	0	1	0	0	5	10	1	11	1	27	6	0	2	31	7	0	0	7
142	0	0	1	0	8	4	19	23	6	1	5	4	0	0	0	13	4	1	0	6	5	0	0	0
141	0	0	2	0	15	35	31	3	21	3	0	0	0	1	7	41	0	0	9	8	2	5	0	4
140	0	1	0	7	30	0	17	0	0	5	0	13	0	4	28	6	1	0	3	10	8	0	17	0
139	0	13	36	29	0	37	0	1	8	1	11	4	13	8	2	9	0	11	4	0	0	0	4	6
138	2	7	7	3	36	0	3	19	1	24	1	17	1	0	11	0	9	10	21	2	11	0	31	0
137	9	18	11	21	2	22	9	0	38	8	0	3	0	30	0	6	1	62	14	1	5	21	1	3
136	19	13	6	22	3	19	0	31	1	0	0	0	6	1	0	1	50	10	7	7	7	0	17	42
135	14	0	0	9	23	0	29	0	0	0	13	0	12	7	0	35	0	1	7	35	0	0	9	0
134	0	0	0	21	0	3	11	0	2	13	0	0	15	10	12	0	12	0	0	5	0	3	4	0
133	2	0	23	3	0	5	13	2	0	0	2	13	9	2	0	0	0	0	0	31	0	0	4	0
132	0	5	0	3	0	2	14	0	2	4	8	3	0	0	3	0	0	0	8	10	6	0	0	52
131	0	0	0	0	9	0	0	20	11	1	1	1	2	5	0	0	1	6	7	2	0	11	12	2
130	1	2	0	0	11	0	38	21	0	13	1	0	2	0	0	6	5	0	16	2	8	13	0	36
129	0	0	24	4	2	7	53	2	5	0	2	0	6	0	0	2	10	0	0	13	2	0	0	0
128	0	7	8	12	5	0	0	2	0	0	0	0	29	0	0	25	1	2	19	1	7	0	0	0
127	19	15	0	2	0	6	1	0	2	10	5	43	0	0	2	5	0	23	7	16	1	0	6	0
126	4	0	0	3	0	8	0	7	6	2	41	0	2	8	0	0	0	12	14	10	0	14	5	1
125	13	0	5	2	17	0	0	0	0	45	0	9	23	0	3	0	8	11	10	1	9	63	4	37
124	0	6	17	41	0	0	0	1	16	3	17	35	2	0	0	0	6	8	4	0	27	1	50	23
123	2	19	18	0	0	0	0	0	32	4	30	0	7	0	0	1	3	0	12	8	5	25	35	0
122	30	19	0	0	0	17	0	23	1	12	4	10	12	0	11	7	0	12	1	0	19	13	0	27
121	0	0	0	0	22	8	0	7	7	5	7	3	4	7	0	0	0	9	4	5	1	2	37	38
120	0	0	2	12	13	0	7	21	9	9	3	0	7	6	3	7	8	2	11	18	7	16	45	41
119	0	2	10	8	0	2	27	3	13	1	3	0	17	24	2	11	0	5	5	2	5	3	33	0
118	0	3	4	0	0	21	21	25	12	0	0	0	17	0	18	14	6	11	22	0	8	16	0	8
117	0	9	0	1	15	5	19	0	1	20	0	28	0	10	12	33	0	10	0	25	2	0	0	0
116	11	0	1	6	21	5	0	7	25	23	10	0	7	18	49	3	6	0	6	6	0	0	0	10
115	0	15	19	27	4	0	19	0	25	2	0	1	14	28	0	1	5	0	1	8	9	1	0	1
114	4	0	1	0	0	1	8	0	7	4	1	11	22	8	1	0	7	0	0	0	1	0	0	12
113	5	22	5	0	8	2	0	1	0	1	10	0	0	0	0	1	7	0	0	1	0	0	0	14
112	1	7	0	1	4	10	0	5	4	3	0	0	0	1	0	36	8	0	2	1	3	0	1	4
111	0	0	12	3	2	0	0	8	0	0	2	0	0	0	19	12	0	0	2	3	1	17	6	0
110	1	0	3	18	2	1	0	2	3	11	5	13	14	16	0	3	0	0	0	0	43	7	11	7
109	0	5	4	4	0	4	4	1	12	8	12	14	12	15	0	2	0	0	0	32	0	4	0	4
108	6	5	4	1	3	0	4	12	15	8	18	5	21	2	0	0	0	2	0	12	2	15	0	3
107	0	3	0	0	6	10	19	10	8	47	0	14	1	3	0	0	0	12	11	6	10	1	3	9
106	19	0	0	6	50	23	16	0	37	0	8	22	7	0	0	4	2	52	5	0	13	5	5	1
105	0	0	16	29	25	0	0	12	0	15	6	6	1	0	5	1	38	3	13	12	22	0	2	0
104	0	0	7	17	0	1	3	6	17	0	1	1	1	9	6	57	0	1	3	27	0	0	0	0
103	0	1	5	0	2	10	1	0	0	0	0	7	21	4	29	0	37	0	10	2	0	0	0	2
102	1	2	2	7	16	2	0	0	0	0	5	27	4	17	5	40	1	0	4	0	4	0	6	0

(*continued on next page*)

'half-diagonal' in Table 4 starting at the row labeled '−8' for the column (i.e., target hour) labeled '17'. A plausible explanation is that the DA price for the first hour of the next day includes important information about the price evolution for the late night intraday prices; it is known a few minutes after noon, and hence is available for forecasting intraday prices for deliveries starting at hour 17.

- Finally, which also comes as a surprise, the dummies are usually removed by the LASSO (except for two morning hours and one afternoon hour on Sunday), see the top seven rows in Table 3. Again, this is in contrast to day-ahead EPF models, where the weekday dummies are typically selected (Uniejewski et al., 2016; Ziel & Weron, 2018).

Table 4 (*continued*).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
101	34	11	16	37	0	0	3	0	0	8	19	10	5	9	10	12	0	2	2	0	6	37	3	4
100	17	12	10	0	0	2	0	0	7	19	14	0	3	16	31	8	0	0	12	0	23	30	4	25
99	0	17	1	0	1	1	0	0	16	5	0	0	12	11	39	0	0	0	12	0	1	1	14	0
98	0	2	0	0	2	0	0	1	2	36	0	4	9	29	3	0	0	0	1	0	5	2	4	1
97	1	0	0	1	2	4	2	13	26	0	1	1	0	1	0	1	15	20	48	10	4	1	5	25
96	22	1	9	0	6	0	7	12	3	19	0	23	9	2	0	17	40	75	31	42	25	25	38	19
95	3	13	0	0	0	6	7	8	23	7	35	4	12	0	6	30	46	6	1	18	16	22	0	31
94	7	0	0	0	7	0	1	6	1	20	30	31	6	8	28	2	0	0	17	0	0	1	0	0
93	0	0	0	25	14	0	0	2	9	19	25	6	43	17	2	4	0	4	0	0	15	0	0	0
92	0	0	9	27	1	0	7	0	4	5	3	52	20	9	17	16	0	1	2	33	0	3	0	0
91	0	0	33	10	0	1	0	0	0	9	15	13	17	20	3	0	0	1	42	0	2	0	5	0
90	0	8	18	0	8	0	0	0	0	15	1	16	21	0	0	1	2	16	3	16	0	1	21	0
89	0	49	1	0	0	0	14	2	0	0	0	8	0	11	15	0	4	7	21	0	0	14	2	0
88	21	3	0	1	2	4	56	15	15	2	0	14	8	10	22	8	0	0	2	0	12	0	1	0
87	0	0	0	1	2	7	9	14	0	0	17	15	10	26	2	5	0	0	2	16	4	1	0	9
86	0	5	0	13	19	17	23	0	0	16	4	12	19	4	0	0	8	2	15	2	11	1	7	8
85	1	0	10	0	0	13	0	0	14	0	15	1	2	0	0	6	0	3	2	14	8	6	0	5
84	3	15	0	0	0	3	0	14	31	11	2	19	0	1	7	0	0	2	15	3	2	18	0	0
83	23	0	0	0	10	20	13	31	7	1	5	0	0	6	0	0	0	8	1	0	12	0	0	0
82	2	0	0	15	31	10	5	17	7	11	2	0	2	0	0	0	5	2	4	5	0	0	0	24
81	4	0	1	35	8	52	12	7	4	0	0	0	0	7	2	27	12	2	23	0	8	1	1	5
80	3	2	33	6	32	15	0	0	0	0	0	1	1	0	36	0	0	10	1	0	1	2	10	1
79	12	24	6	39	18	23	0	9	0	0	0	0	0	36	13	5	17	3	6	0	0	10	2	0
78	8	1	46	25	21	0	0	0	0	4	0	0	36	7	2	18	26	14	0	10	27	9	24	0
77	1	7	1	16	2	0	0	5	0	0	0	28	8	8	0	32	6	1	42	38	12	44	0	17
76	18	0	0	7	0	0	15	1	0	0	2	30	7	3	5	24	2	51	20	0	28	16	19	0
75	0	0	0	0	0	25	0	0	0	2	18	26	0	14	20	0	0	5	0	30	5	15	0	1
74	22	2	17	5	34	36	6	6	8	4	33	0	13	3	4	0	0	0	1	3	0	0	1	17
73	0	1	0	29	18	13	33	0	6	0	8	1	0	0	1	0	0	0	1	0	40	3	22	47
72	0	7	24	39	0	32	0	15	0	21	5	0	0	0	15	0	1	0	34	39	4	38	48	27
71	4	2	21	15	0	0	18	7	0	0	0	0	0	22	8	0	0	16	11	21	16	12	5	3
70	0	12	8	0	0	4	29	0	1	1	0	13	19	7	0	0	0	13	33	0	0	0	0	0
69	0	18	0	0	14	13	8	4	17	0	4	26	13	0	4	3	0	27	14	3	3	3	0	2
68	0	10	0	13	39	0	0	1	2	3	26	1	0	0	4	1	12	6	10	4	0	0	12	0
67	4	0	3	9	0	0	3	0	2	22	0	0	0	0	0	39	5	15	1	0	0	0	1	1
66	8	0	12	0	8	3	0	0	3	16	2	0	0	8	32	1	32	0	9	16	0	1	2	0
65	1	0	0	9	1	0	0	0	14	4	0	0	12	37	16	12	4	14	5	21	0	0	7	20
64	1	0	12	9	0	0	0	19	0	2	0	0	36	3	5	13	7	1	48	5	1	0	20	4
63	2	9	2	0	0	20	18	0	16	0	0	21	14	0	17	2	6	31	8	2	0	13	5	9
62	0	0	1	1	0	19	6	20	6	1	4	13	4	7	5	0	24	0	5	0	19	8	0	2
61	0	0	0	0	0	4	1	11	0	6	0	11	4	6	0	3	0	0	0	17	3	2	2	0
60	9	0	0	0	0	5	5	0	11	0	0	1	2	0	0	1	0	0	7	20	0	2	0	13
59	0	1	0	1	0	17	0	8	6	0	0	0	6	4	16	16	0	9	21	10	11	3	16	12
58	6	0	2	0	0	7	34	1	0	1	6	4	1	16	39	0	41	0	3	1	8	8	6	2
57	2	23	0	0	0	14	11	0	2	2	7	0	6	42	5	20	6	6	0	1	11	6	0	3
56	15	0	0	3	0	0	0	8	13	7	13	10	3	0	11	20	42	0	1	4	1	0	0	0
55	32	0	1	1	1	0	4	0	0	0	0	1	3	26	18	46	4	36	7	7	14	1	5	37
54	0	0	4	10	17	0	2	0	2	0	0	0	34	13	31	3	21	6	7	0	0	0	11	0
53	0	0	11	5	0	0	0	0	6	0	0	37	0	15	12	17	25	3	1	0	0	4	4	7
52	8	0	0	0	15	1	3	0	2	0	34	0	0	9	43	46	1	1	20	0	0	19	6	1
51	5	0	0	8	3	6	0	0	0	36	13	0	2	20	44	0	0	0	1	0	1	0	2	0
50	0	4	2	0	12	18	0	0	39	10	0	1	3	22	2	0	0	0	1	9	0	17	17	0
49	17	12	0	13	24	3	0	49	8	0	1	3	0	16	12	0	0	3	31	29	17	10	15	7
48	0	0	0	9	8	24	23	14	0	1	7	0	14	10	0	0	8	45	30	80	41	10	36	15
47	0	6	11	4	36	21	25	2	3	0	10	6	9	4	3	1	1	19	45	12	4	7	20	5
46	0	6	2	47	7	16	0	0	8	16	5	0	7	2	0	0	1	0	0	0	3	2	16	2
45	0	7	53	3	1	0	16	2	3	2	0	3	2	4	0	0	5	0	14	2	1	0	1	5
44	0	48	2	4	10	0	13	0	0	3	4	3	8	0	4	12	1	0	7	0	2	0	2	4
43	21	13	4	8	3	5	9	4	0	8	4	29	4	7	19	1	0	0	5	6	0	7	19	26
42	15	2	9	9	17	0	3	6	11	6	19	18	0	2	12	0	2	0	8	0	27	14	0	0
41	20	0	16	0	2	0	10	19	0	4	16	7	0	26	0	0	2	3	2	15	27	2	3	1
40	0	9	1	2	0	4	1	1	0	1	11	4	27	1	0	5	5	6	29	1	0	6	30	16
39	10	1	0	1	1	8	2	11	1	4	2	24	0	0	6	0	10	2	4	0	6	17	5	10
38	5	0	3	0	4	0	5	4	5	0	10	0	0	11	0	5	0	25	0	0	14	0	0	4
37	2	3	0	0	11	0	18	2	0	5	0	8	8	4	0	4	1	0	0	9	10	0	0	0
36	0	3	7	25	1	3	0	7	6	1	10	8	7	0	27	2	0	0	12	17	2	0	3	40
35	16	4	45	27	0	0	5	9	0	64	26	22	0	42	0	3	0	37	1	4	0	0	28	0

(*continued on next page*)

4.5. Using the LASSO to build parsimonious ARX-type models

Now that we know which variables are selected by the LASSO, we can ask ourselves whether this information can be used to build well-performing, parsimonious ARX-type models. We address this question by computing the frequency of selecting each of the 349 (for hour 16) to 372 (for hour 17) regressors of the **LASSO**(λ_6) model, see Eq. (4), in a 364-day rolling 'selection' window that follows the 364-day calibration window directly. Then, we forecast prices in the remaining 488-day out-of-sample test period by building ARX-type models that include as regressors only those variables that have been selected at least $x\% = 50\%, 60\%, \ldots, 90\%$ of the time. The resulting models are denoted by $\mathbf{ARX}_{x\%}$.

The MAE and RMSE values are reported in Table 5. Note that they differ from those reported for the benchmarks and the **LASSO**(λ_6) model in Table 1 because of the much shorter out-of-sample test period (488 vs. 852 days). Surprisingly, some of the $\mathbf{ARX}_{x\%}$ models perform

Table 4 (*continued*).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
34	6	35	16	16	0	14	11	0	68	13	16	0	44	0	4	0	14	1	9	0	0	0	0	5
33	50	17	12	0	19	9	3	4	5	13	0	41	17	3	0	17	25	4	0	0	0	3	0	19
32	1	26	0	10	13	8	3	0	0	0	0	0	5	0	0	28	21	0	0	29	0	12	0	1
31	20	4	14	1	0	25	0	0	0	0	0	2	3	4	3	17	0	0	0	0	36	1	0	0
30	0	2	0	4	14	21	20	0	15	0	17	0	0	0	1	0	0	8	0	21	3	4	3	0
29	7	0	0	5	2	2	0	0	0	7	0	0	0	0	11	0	1	1	19	15	3	3	4	0
28	2	5	7	18	0	0	0	0	0	1	4	0	0	8	6	1	0	24	26	17	11	35	0	0
27	0	29	28	0	0	0	0	4	11	18	5	0	9	4	22	0	14	2	0	0	34	0	0	6
26	34	4	1	0	0	0	0	6	10	4	4	16	5	20	2	20	2	0	0	2	0	0	17	0
25	8	0	5	0	2	0	1	11	9	1	33	5	4	5	1	2	0	0	12	0	29	24	12	35
24	4	0	0	4	13	6	2	34	14	55	6	7	12	0	0	0	0	10	1	19	29	46	42	0
23	0	0	0	40	14	7	43	6	12	3	0	12	0	0	0	0	7	0	44	1	1	2	0	2
22	0	0	13	9	0	62	19	6	32	4	10	0	0	7	0	7	0	15	0	0	0	9	6	2
21	0	0	5	4	96	0	0	29	10	11	0	0	30	0	10	20	5	0	0	16	8	0	0	0
20	0	1	1	53	0	0	17	12	15	0	4	23	0	14	16	4	1	5	14	26	0	0	0	0
19	0	32	5	13	0	0	0	8	0	2	5	3	15	11	8	22	12	46	38	8	12	0	0	13
18	41	0	0	0	0	40	16	0	0	23	14	8	40	0	38	17	37	44	31	31	0	0	3	17
17	0	14	0	4	21	30	0	0	8	32	32	39	0	23	16	33	80	33	13	27	3	4	1	51
16	16	16	0	5	51	8	0	25	29	36	39	0	0	27	30	59	19	9	42	2	4	1	35	2
15	18	7	4	20	16	0	36	5	13	49	0	1	30	40	78	8	0	26	3	6	5	45	4	36
14	2	11	25	17	0	0	17	11	71	32	0	33	36	76	1	1	14	0	45	6	40	22	27	15
13	1	18	10	2	1	14	10	74	36	8	40	20	85	6	0	7	0	52	9	40	3	7	19	3
12	13	28	17	0	0	2	19	32	15	25	48	90	4	0	2	0	52	5	44	3	21	2	0	8
11	1	7	0	0	0	25	36	0	13	42	93	5	0	0	0	27	16	30	4	23	16	5	15	0
10	24	0	9	0	0	0	3	2	35	56	19	0	0	1	12	35	38	5	28	23	0	17	14	10
9	0	3	0	4	0	0	0	38	79	51	7	2	6	13	9	23	37	30	33	33	0	54	19	34
8	6	1	3	29	1	0	97	86	38	24	21	7	7	6	23	2	22	21	51	16	31	3	14	3
7	7	12	45	22	38	100	62	38	24	48	27	6	0	24	4	32	36	64	26	53	9	54	43	17
6	18	67	55	12	96	0	53	18	57	51	2	0	28	26	37	71	19	85	90	37	81	51	4	42
5	51	31	0	66	0	28	12	87	82	51	29	25	55	90	71	38	61	66	48	76	68	7	33	34
4	15	0	6	0	20	0	96	59	60	34	83	100	100	54	50	69	88	42	18	37	32	73	93	97
3	0	3	0	0	0	0	7	30	13	3	11	51	0	4	0	0	0	0	0	8	3	9	0	35
2	38	14	0	0	0	0	1	0	0	80	10	0	0	0	0	0	0	0	3	17	0	2	0	0
1	42	0	1	27	12	9	0	1	100	78	0	8	57	36	20	43	61	44	54	15	69	4	14	89
0	100	100	65	100	100	100	99	100	100	92	100	100	100	100	100	100	100	100	100	100	100	100	100	100
-1	96	54	70	74	36	43	94	100	50	64	92	66	83	86	60	68	55	27	24	37	53	30	2	80
-2	56	69	27	58	9	93	64	2	35	84	23	44	68	44	34	67	0	6	16	0	0	1	100	29
-3	47	11	36	2	26	51	2	10	45	0	30	72	4	62	42	10	6	0	2	4	17	100	21	35
-4	6	27	10	0	5	1	16	48	0	4	33	35	51	38	3	4	7	6	5	10	100	27	0	6
-5	34	39	1	0	38	0	0	12	9	53	14	51	15	16	22	9	8	3	9	100	25	2	0	0
-6	39	0	0	2	13	5	4	10	12	31	41	11	33	13	1	21	11	10	80	21	8	0	0	16
-7	5	0	0	10	0	0	1	0	44	17	19	38	17	0	28	20	3	76	23	45	3	0	0	7
-8	9	0	17	4	0	1	0	8	16	9	41	9	10	37	29	79	57	5	38	9	1	0	23	28
-9	0	7	1	0	0	0	34	21	15	64	8	35	37	19	65		14	56	0	0	0	47	26	0
-10	9	0	1	0	3	11	23	25	55	31	80	11	45	77			62	0	0	0	63	31	10	54
-11	1	11	0	1	25	42	6	14	25	93	1	42	80				5	0	0	39	61	19	56	7
-12	10	1	9	22	44	9	36	55	57	8	22	16					0	0	5	48	25	64	4	24
-13	0	30	15	69	21	12	2	20	19	18	11						11	29	0	11	40	8	5	5
-14	0	5	80	9	6	16	28	12	33	9							32	2	12	9	6	0	29	7
-15	0	77	12	15	0	8	14	18	13								14	5	18	47	4	0	13	16
-16	79	7	11	14	23	19	0	1									2	3	88	10	0	31	12	0
-17	3	0	19	6	11	0	16										0	19	5	11	18	22	2	10
-18	0	31	6	43	27	1											48	56	23	1	41	4	9	0
-19	13	0	32	0	37												50	23	0	24	2	0	18	7
-20	0	24	12	36													19	3	3	0	3	8	5	12
-21	7	8	20														9	2	1	5	4	4	1	1
-22	92	39															6	0	28	0	0	6	0	5
-23	8																4	47	7	25	50	9	0	68
-24																	41	8	30	20	36	34	62	7
-25																	2	21	0	13	12	40	5	
-26																	23	9	2	0	0	6		
-27																	25	3	2	8	21			
-28																	1	8	16	25				
-29																	2	7	27					
-30																	3	40						
-31																	8							

exceptionally well. In particular, **ARX**$_{70\%}$ is slightly better in terms of RMSE than **LASSO**(λ_6), despite utilizing only 13.4 explanatory variables on average, i.e., about 27 times fewer than the baseline LASSO model in Eq. (4) and nearly 3.5 times fewer than **LASSO**(λ_6), which selects an average of 46 regressors. However, the difference is not statistically significant, see the DM plots in Fig. 5. On the other hand, the **LASSO**(λ_6) model is the best in terms of MAE, being significantly better than all of its competitors except **ARX**$_{70\%}$.

5. Conclusions

We have used a unique set of electricity prices from the German EPEX market to address the problem of the optimal choice of explanatory variables for forecasting intraday prices. Given that the literature on this topic is very scarce, our study is a major step towards understanding the intraday price dynamics and developing well-performing predictive models for a market that many participants see as the future of electricity trading.

Table 5
MAE and RMSE values for the two benchmarks (**Naive**, **ARX**), the **LASSO**($\boldsymbol{\lambda}_6$) model and five $\mathbf{ARX}_{x\%}$ models built on the latter with cutoffs of $x\% = 50\%, 60\%, \ldots, 90\%$, over the 488-day out-of-sample test period (see Section 4.5 for details).

	Naive	ARX	LASSO	ARX				
			λ_6	50%	60%	70%	80%	90%
MAE	5.7051	5.3095	4.9330	5.0119	4.9711	4.9539	5.0021	5.0936
RMSE	9.2912	8.5738	7.8343	7.8883	7.8552	7.8325	7.9419	8.1578

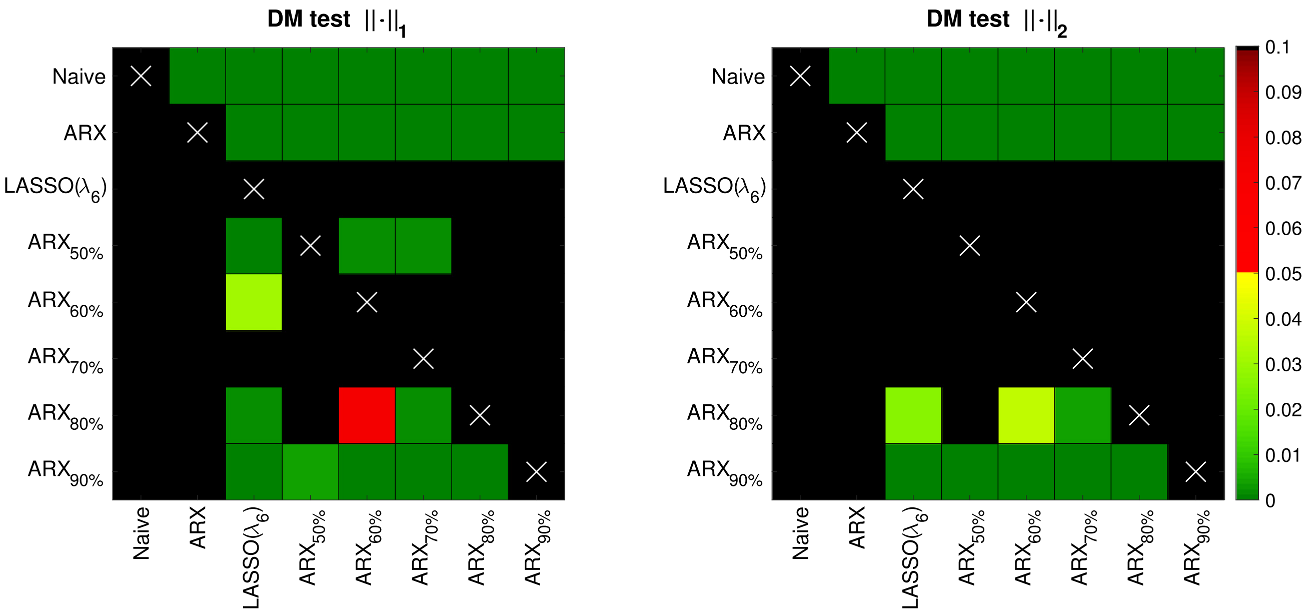


Fig. 5. Results of the Diebold–Mariano (DM) test for the absolute (left) and squared (right) prediction errors, using the same models as in Table 5. A heat map is used to indicate the range of the p-values: the closer they are to zero (→ dark green), the more significant the difference is between the forecasts of a model on the X-axis (better) and the forecasts of a model on the Y-axis (worse). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

To this end, we have considered 12 models, namely a naive benchmark, a parsimonious autoregressive structure inspired by the well-performing $expert_{DoW,nl}$ model of Ziel and Weron (2018), and a LASSO-estimated model with 349 (for hour 16) to 372 (for hour 17) potential regressors and ten different values of the *tuning* parameter. We have found that, for an appropriately chosen value of λ, the LASSO model significantly outperforms its competitors, as measured by the Diebold–Mariano test.

The most important explanatory variables turned out to be the most recent intraday price and the day-ahead (DA) price that corresponds to the same hour. The intraday and — to a lesser extent — DA prices for late evening hours could also be considered as regressors. On the other hand, in contrast to day-ahead EPF models, neither the previous day's price for the same hour nor weekday dummies were found to be important predictors.

Finally, we have shown that the LASSO can be used to build well-performing, parsimonious ARX-type models. In particular, the performance of an OLS-estimated model with regressors that have been selected by the LASSO at least 70% of the time in a 364-day rolling 'selection' window is comparable to that of the best LASSO model. At the same time, it utilizes an average of only 13.4 explanatory variables, i.e., about 27 times fewer than the baseline model and nearly 3.5 times fewer than the best LASSO model, **LASSO**($\boldsymbol{\lambda}_6$).

Acknowledgments

This work was partially supported by the German Research Foundation (DFG, Germany) and the National Science Center (NCN, Poland) through BEETHOVEN grant no. 2016/23/G/HS4/01005. The authors would like to thank the participants of the IMMORTAL kick-off meeting (9-11.05.2018, Wrocław, Poland), the 7th International Ruhr Energy Conference INREC2018 (24-25.09.2018, Essen, Germany) and the Energy Finance Christmas Workshop EFC18 (12-13.12.2018, Bolzano, Italy) for fruitful discussions and suggestions.

References

Aïd, R., Gruet, P., & Pham, H. (2016). An optimal trading problem in intraday electricity markets. *Mathematics and Financial Economics*, *10*(1), 49–85.

Amjady, N., & Keynia, F. (2009). Day-ahead price forecasting of electricity markets by mutual information technique and cascaded neuro-evolutionary algorithm. *IEEE Transactions on Power Systems*, *24*(1), 306–318.

Andrade, J., Filipe, J., Reis, M., & Bessa, R. (2017). Probabilistic price forecasting for day-ahead and intraday markets: beyond the statistical model. *Sustainability*, *9*(11), 1990.

Barnes, A. K., & Balda, J. C. (2013). Sizing and economic assessment of energy storage with real-time pricing and ancillary services. In *PEDG 2013 Conference Proceedings*. http://dx.doi.org/10.1109/PEDG.2013.6785651.

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

EPEX (2015). New ID3-price index on German intraday continuous market. Report no. 2015-21, ver. July 2015, http://www.epexspot.com/en/extras/download-center.

EPEX (2018). Annual Report 2017, http://www.epexspot.com/en/extras/download-center.

Gianfreda, A., Parisio, L., & Pelagatti, M. (2016). The impact of RES in the Italian day-ahead and balancing markets. *Energy Journal*, *37*, 161–184.

Gianfreda, A., Parisio, L., & Pelagatti, M. (2018). A review of balancing costs in Italy before and after RES introduction. *Renewable & Sustainable Energy Reviews*, *91*, 549–563.

González-Aparicio, I., & Zucker, A. (2015). Impact of wind power uncertainty forecasting on the market integration of wind energy in Spain. *Applied Energy*, *159*, 334–349.

Hong, T., & Fan, S. (2016). Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, *32*, 914–938.

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*(1), 321–323.

Janczura, J., Trück, S., Weron, R., & Wolff, R. (2013). Identifying spikes and seasonal components in electricity spot price data: a guide to robust modeling. *Energy Economics*, *38*, 96–110.

Karakatsani, N., & Bunn, D. (2008). Forecasting electricity prices: the impact of fundamentals and time-varying coefficients. *International Journal of Forecasting*, *24*, 764–785.

Keles, D., Scelle, J., Paraschiv, F., & Fichtner, W. (2016). Extended forecast methods for day-ahead electricity spot prices applying artificial neural networks. *Applied Energy*, *162*, 218–230.

Kiesel, R., & Kusterman, M. (2016). Structural models for coupled electricity markets. *Journal of Commodity Markets*, *3*(1), 16–38.

Kiesel, R., & Paraschiv, F. (2017). Econometric analysis of 15-minute intraday electricity prices. *Energy Economics*, *64*, 77–90.

Ludwig, N., Feuerriegel, S., & Neumann, D. (2015). Putting big data analytics to work: Feature selection for forecasting electricity prices using the LASSO and random forests. *Journal of Decision Systems*, *24*(1), 19–36.

Maciejowska, K., Nitka, W., & Weron, T. (2019). Day-ahead vs. intraday – forecasting the price spread to maximize economic benefits. *Energies*, *12*(4), 631.

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

Märkle-Huß, J., Feuerriegel, S., & Neumann, D. (2018). Contract durations in the electricity market: causal impact of 15 min trading on the EPEX SPOT market. *Energy Economics*, *69*, 367–378.

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

Misiorek, A. (2008). Short-term forecasting of electricity prices: Do we need a different model for each hour? *Medium Econometrisch Toepassingen*, *16*(2), 8–13.

Monteiro, C., Ramirez-Rosado, I., Fernandez-Jimenez, L., & Conde, P. (2016). Short-term price forecasting models based on artificial neural networks for intraday sessions in the Iberian electricity market. *Energies*, *9*(9), 721.

Nowotarski, J., & Weron, R. (2018). Recent advances in electricity price forecasting: a review of probabilistic forecasting. *Renewable & Sustainable Energy Reviews*, *81*, 1548–1568.

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.

Schneider, S. (2011). Power spot price models with negative prices. *Journal of Energy Markets*, *4*(4), 77–102.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B.*, *58*, 267–288.

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

Uniejewski, B., & Weron, R. (2018). Efficient forecasting of electricity spot prices with expert and lasso models. *Energies*, *11*, 2039.

Uniejewski, B., Weron, R., & Ziel, F. (2018). Variance stabilizing transformations for electricity spot price forecasting. *IEEE Transactions on Power Systems*, *33*, 2219–2229.

Weron, R. (2006). *Modeling and forecasting electricity loads and prices: A statistical approach*. John Wiley & Sons, Chichester.

Weron, R. (2014). 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.

Zaleski, P., & Klimczak, D. (2015). Prospects for the rise of renewable sources of energy in Poland. balancing renewables on the intraday market. In K. Zamasz (Ed.), *Capacity market in contemporary economic policy* (pp. 124–138). Difin.

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

Ziel, F. (2017). Modeling the impact of wind and solar power forecasting errors on intraday electricity prices. In *IEEE conference proceedings — EEM17*. http://dx.doi.org/10.1109/EEM.2017.7981900.

Ziel, F., Steinert, R., & Husmann, S. (2015). Efficient modeling and forecasting of electricity spot prices. *Energy Economics*, *47*, 89–111.

Ziel, F., & Weron, R. (2018). Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate modeling frameworks. *Energy Economics*, *70*, 396–420.

Bartosz Uniejewski is a Master student of Applied Mathematics at the Wrocław University of Science and Technology, Poland. Despite being a (very) early stage researcher, he has given talks on electricity price forecasting at conferences and workshops in Essen (Germany), Kraków (Poland), Milan (Italy), Oxford (UK), Padova (Italy) and Trondheim (Norway). He is the recipient of the prestigious scholarship for outstanding students of the Ministry of Higher Education in Poland (2016, 2017, 2018).

Grzegorz Marcjasz is a Master student of Applied Mathematics at the Wrocław University of Science and Technology, Poland. He is a proficient programmer in Matlab and Python. His research interests include energy forecasting and machine learning. He is the recipient of the prestigious scholarship for outstanding students of the Ministry of Higher Education in Poland (2017, 2018) and the Best Student Paper prize (runner-up) at the Computational Management Science conference in Trondheim (2018).

Rafał Weron is Professor of Economics in the Department of Operations Research, Wrocław University of Science and Technology, Poland. His research focuses on developing risk management and forecasting tools for the energy industry and computational statistics as applied to finance and insurance. He is the author of the widely acclaimed Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach (Wiley, 2006) and co-author of over 80 peer-reviewed book chapters and journal articles. His review paper on electricity price forecasting was selected for the 2017 edition of the Emerald Citations of Excellence and received the 2017 IIF Tao Hong Award for the best International Journal of Forecasting paper on energy forecasting. With a Ph.D. (1999) in Financial Mathematics and a habilitation (2009) and professor title (2015) in Economics, he is periodically engaged as a consultant to financial, energy and software engineering companies.

Paper 2

LASSO Principal Component Averaging – a fully automated approach for point forecast pooling

Bartosz Uniejewski, Katarzyna Maciejowska

Contents lists available at ScienceDirect

International Journal of Forecasting

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

LASSO principal component averaging: A fully automated approach for point forecast pooling

Bartosz Uniejewski [*], Katarzyna Maciejowska

Department of Operations Research and Business Intelligence, Wrocław University of Science and Technology, 50-370 Wrocław, Poland

ARTICLE INFO

Keywords:
Electricity price forecasting
EPF
Averaging
PCA
Principal component analysis
Regularization
LASSO
Day-ahead market
Risk management

ABSTRACT

This paper develops a novel, fully automated forecast averaging scheme which combines LASSO estimation with principal component averaging (PCA). LASSO-PCA (LPCA) explores a pool of predictions based on a single model but calibrated to windows of different sizes. It uses information criteria to select tuning parameters and hence reduces the impact of researchers' ad hoc decisions. The method is applied to average predictions of hourly day-ahead electricity prices over 650 point forecasts obtained with various lengths of calibration windows. It is evaluated on four European and American markets with an out-of-sample period of almost two and a half years and compared to other semi- and fully automated methods, such as the simple mean, AW/WAW, LASSO, and PCA. The results indicate that LASSO averaging is very efficient in terms of forecast error reduction, whereas PCA is robust to the selection of the specification parameter. LPCA inherits the advantages of both methods and outperforms other approaches in terms of the mean absolute error, remaining insensitive to the choice of a tuning parameter.

1. Introduction

Electricity price forecasting (EPF) is nowadays perceived as fundamental for decision making in energy markets. As short-term transactions provide a tool for adjusting long-term positions and a benchmark in over-the-counter trading, the day-ahead, intraday, and balancing prices play a key role in day-to-day operations (Kath & Ziel, 2018; Maciejowska, Nitka, & Weron, 2019; Mayer & Trück, 2018; Weron, 2014). In the last decades, the market share of renewable energy sources has rapidly increased. As a result, intermittent changes in the generation level and structure have become more likely. This leads to an increase in market imbalances and electricity price volatility (Gianfreda, Parisio, & Pelagatti, 2016; Kowalska-Pyzalska, 2018; Maciejowska, 2020). Hence, reliable methods dedicated to EPF are essential for managing energy companies.

One way to increase the prediction accuracy is to combine forecasts obtained with different models. The idea of forecast averaging started about half a century ago. The pioneering papers of Bates and Granger (1969) and Crane and Crotty (1967) inspired many authors to develop new methods and contribute to the area. Since the late 1960s, hundreds of papers have suggested the superiority of forecast combinations over individual models (Nowotarski & Weron, 2016; Timmermann, 2006; Wallis, 2011). Hibon and Evgeniou (2005) state that the main advantage of combining forecasts is the fact that, in practice, it is less risky to combine forecasts than to select an individual forecasting method.

Recently, experts have paid more attention to the selection of the calibration window used for model estimation (see Pesaran & Timmermann, 2007). Marcos, Bunn, Bello, and Reneses (2020) claim that in rapidly developing markets, such as an energy market, researchers

[*] Corresponding author.
E-mail addresses: bartosz.uniejewski@pwr.edu.pl (B. Uniejewski), katarzyna.maciejowska@pwr.edu.pl (K. Maciejowska).

https://doi.org/10.1016/j.ijforecast.2022.09.004
0169-2070/© 2022 The Author(s). Published by Elsevier B.V. on behalf of International Institute of Forecasters. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

should take into account structural breaks and adjust model parameters to market changes. The simplest solution to the issue is to work with short data, which describe only the most recent events. This approach has some severe drawbacks, as it decreases the estimation accuracy and limits the complexity of applied models. On the other hand, one may try to estimate the time of a structural break and include it directly in a forecasting model. The assumption of a discrete shift in model parameters is, however, unsuitable for more complex evolution patterns (Marcos et al., 2020). In the literature, there is no agreement on which solution is the best, so the majority of research on EPF applies an arbitrarily chosen calibration window length. In recent articles, Hubicka, Marcjasz, and Weron (2019), Marcjasz, Serafin, and Weron (2018), Serafin, Uniejewski, and Weron (2019) suggest using a pool of different in-sample data sizes and averaging the resulting forecasts. The outcomes presented in these papers suggest that the choice of three 'short' and three 'long' calibration windows provides robust results that outperform all individual predictions. This conclusion is questioned by Maciejowska, Uniejewski, and Serafin (2020), who show that the suggested solution is not valid for all the electricity markets and has to be adjusted to a market specification.

The estimation of a single model with various calibration windows enables us to obtain a large number of predictions. For example, in Maciejowska et al. (2020) a panel of 673 forecasts is built. Moreover,it could be observed that predictions in such a pool are very similar to each other,because a slight change in the estimation window does not alter the model parameters much. Thus, it is natural to search for methods that would help to reduce the dimension of the problem, without losing useful information. In this context, two approaches are natural candidates: the principal component (PC) method, which summarizes the panel with a small number of components (see Bai & Ng, 2002; Stock & Watson, 2002); and the least absolute shrinkage and selection operator (LASSO, Tibshirani (1996)), which reduces the dimension of a model by assigning a penalty to non-zero parameters. Here, we propose a novel approach which combines these two methods for forecast averaging. Principal components is a well-known tool that has been successfully applied for analyzing big panels of data. It has been used to directly predict the variables of interest (Boivin & Ng, 2005; Stock & Watson, 2012) or to augment a small-scale econometric model (Banerjee, Marcellino, & Masten, 2014). Factor models have been extended to account for dynamic relationships (see Forni, Hallin, Lippi, & Reichlin, 2000; Forni & Lippi, 2001) and used to create economic indicators (Stock & Watson, 1998). Although the potential of principal component averaging (PCA) for forecast averaging was recognized by Chan, Stock, and Watson (1999) and Huang and Lee (2010), there are only a few papers that illustrate its performance. Stock and Watson (2004) and Poncela, Rodriguez, Sanchez-Mangas, and Senra (2011) used PCA to predict macroeconomic variables. They estimated components from a panel of forecasts coming from either different models or different experts. In both cases, the panels were relatively small and diversified. Maciejowska et al. (2020) proposed an algorithm that extracts PCs from a large standardized panel of predictions coming from a single model (as in Hubicka et al. (2019), Marcjasz et al. (2018), and Serafin et al. (2019)) and uses them to calculate the final forecasts via linear regression. In their study, 1–4 components were used. The results indicated that PCA is a robust method for forecast pooling. The major issue with Maciejowska et al. (2020) is that the number of PCs is either chosen a priori or selected from a small number of alternatives. Moreover, it is not clear how the approach performs if a larger number of components are considered.

The literature proposes many methods of dealing with a large set of potential explanatory variables. Two major approaches could be distinguished: selecting an optimal model (Gaillard, Goude, & Nedellec, 2016; Ludwig, Feuerriegel, & Neumann, 2015; Uniejewski, Nowotarski, & Weron, 2016; Ziel, Steinert, & Husmann, 2015) or averaging across models. Here, we adopt the first approach and apply LASSO, which was introduced by Tibshirani (1996) and is one of the most popular and important regularization methods. Because of its linear penalty function, the LASSO estimator shrinks the coefficients of the less important explanatory variables to zero. It is a tool for automated variable selection, as it identifies significant variables and excludes redundant ones (Uniejewski et al., 2016; Uniejewski & Weron, 2018). In the context of prediction pooling, the LASSO technique has been successfully used in both point (Diebold & Shin, 2019) and probabilistic (Bayer, 2018; Bracale, Carpinelli, & Falco, 2019; Uniejewski & Weron, 2021) forecasting. To our knowledge, LASSO averaging has not been applied to point forecasting of electricity prices, and therefore there is a need to evaluate its performance in this field.

The main novelty of this paper is a fully automated forecast averaging scheme, called LPCA, that utilizes both PCA and LASSO regularization techniques. We present an algorithm that extends the approach described in Maciejowska et al. (2020) and allows for the use of an arbitrarily large number of components. Thanks to the LASSO estimation method, irrelevant PCs are excluded and hence the corresponding noise is reduced. Since LASSO depends on a tuning parameter, information criteria are applied to select its optimal value. Unlike in typical LASSO averaging, the inputs in LPCA are orthogonal to each other. Moreover, although one could use all PCs, a smaller number of components than individual forecasts should be sufficient. Hence, LPCA should be much easier and faster to compute than the full-panel LASSO. As a result, the proposed methodology does not require any expert knowledge or intuition to obtain predictions of future prices and should be less computationally burdensome than existing methods.

The paper is structured as follows. First, we present the datasets that consist of day-ahead price series, as well as the exogenous variables. At the end of Section 2, we describe a data transformation. Next, in Section 3, we present the methodology to obtain point forecasts and to average them. In the same section, we introduce a new algorithm for a fully automated approach designed to combine forecasts. In Section 4, we present the results of our study, and in Section 5, we conclude the research.

Table 1
Exogenous variables.

Description	Notation	Availability
Load	$L_{d,h}$	EPEX, NP, OMIE, PJM
Zonal load	$Z_{d,h}$	PJM
Wind power generation (WPG)	$W_{d,h}$	EPEX, NP, OMIE
Photovoltaic generation (PVG)	$S_{d,h}$	EPEX, OMIE

2. Datasets

The datasets used in this study cover five years and describe four different markets: German (EPEX), Scandinavian (Nord Pool, NP), Spanish (OMIE), and American (PJM). All time series have an hourly resolution and span 1826 days from 1.01.2015 to 31.12.2019 (the data are not extended to 2020, as the COVID-19 pandemic changed the market dynamics). Missing or 'doubled' values (corresponding to the time change) are replaced by the average of the closest observations, for the missing hours, and the arithmetic mean of the two values, for 'doubled' hours. Note that the data are double-indexed, with d denoting the day and h the hour of an observation.

2.1. Day-ahead electricity prices

This research focuses on electricity prices from day-ahead markets, which are established simultaneously around noon on the day preceding the delivery. A more detailed description of the day-ahead market design can be found in Weron (2014). As a result, market participants can utilize only the information available at the time of bidding. This also impacts the forecasters, who should include in their models only the data published before the noon (see Huisman, Huurman, & Mahieu, 2007).

The following day-ahead prices, $DA_{d,h}$, are considered:

- The German market EPEX spot (top panel in Fig. 1(a)); the data are taken from the transparency platform (https://transparency.entsoe.eu).
- The Scandinavian market Nord Pool (top panel in Fig. 1(b)); the data are taken from the Nord Pool website (https://www.nordpoolgroup.com).
- The Spanish market OMIE (top panel in Fig. 2(a)); the data are taken from the OMIE website (https://www.omie.es).
- The American market PJM COMED (top panel in Fig. 2(b)); the data are taken from the PJM data miner (https://dataminer2.pjm.com).

2.2. Exogenous variables

The literature indicates that various exogenous factors, such as the generation structure and fuel prices, have an important impact on electricity prices and can be used to forecast these prices Following Maciejowska et al. (2020), in this study, we consider day-ahead predictions of fundamental variables describing the demand and supply of electricity provided by transmission system operators. A description of the data can be found in Table 1. Notice that the set of exogenous variables changes between markets and depends on the data availability.

The day-ahead forecasts for all exogenous variables are plotted in Figs. 1 and 2. The variables, in particular load and solar generation, exhibit strong yearly seasonality, with the load also following a weekly pattern.

2.3. Variance stabilizing transformation

As can be seen in Figs. 1 and 2, electricity prices exhibit spiky behavior. Uniejewski, Weron, and Ziel (2018) argue that it is possible to reduce the influence of such extreme values on forecasts by using a variance stabilizing transformation (VST). These findings are confirmed by the literature (Marcjasz et al., 2018; Uniejewski & Weron, 2018). Here, we follow the recommendation of Uniejewski et al. (2018) and apply the N-PIT transformation (to all variables in the dataset). Let us recall that the N-PIT transformation is based on the so-called probability integral transform. Let us consider a time series $Y_{d,h}$. Its transformation, $\tilde{Y}_{d,h}$, is given by:

$$\tilde{Y}_{d,h} = N^{-1}\left(\hat{F}_Y(Y_{d,h})\right), \tag{1}$$

where $\hat{F}_Y(\cdot)$ is the empirical cumulative distribution function of the in-sample Y, and N^{-1} is the quantile function of the normal distribution. After the models are estimated on the transformed time series, we apply the inverse transformation to obtain the final forecast of electricity prices:

$$Y_{d,h} = \hat{F}_Y\left(N(\tilde{Y}_{d,h})\right), \tag{2}$$

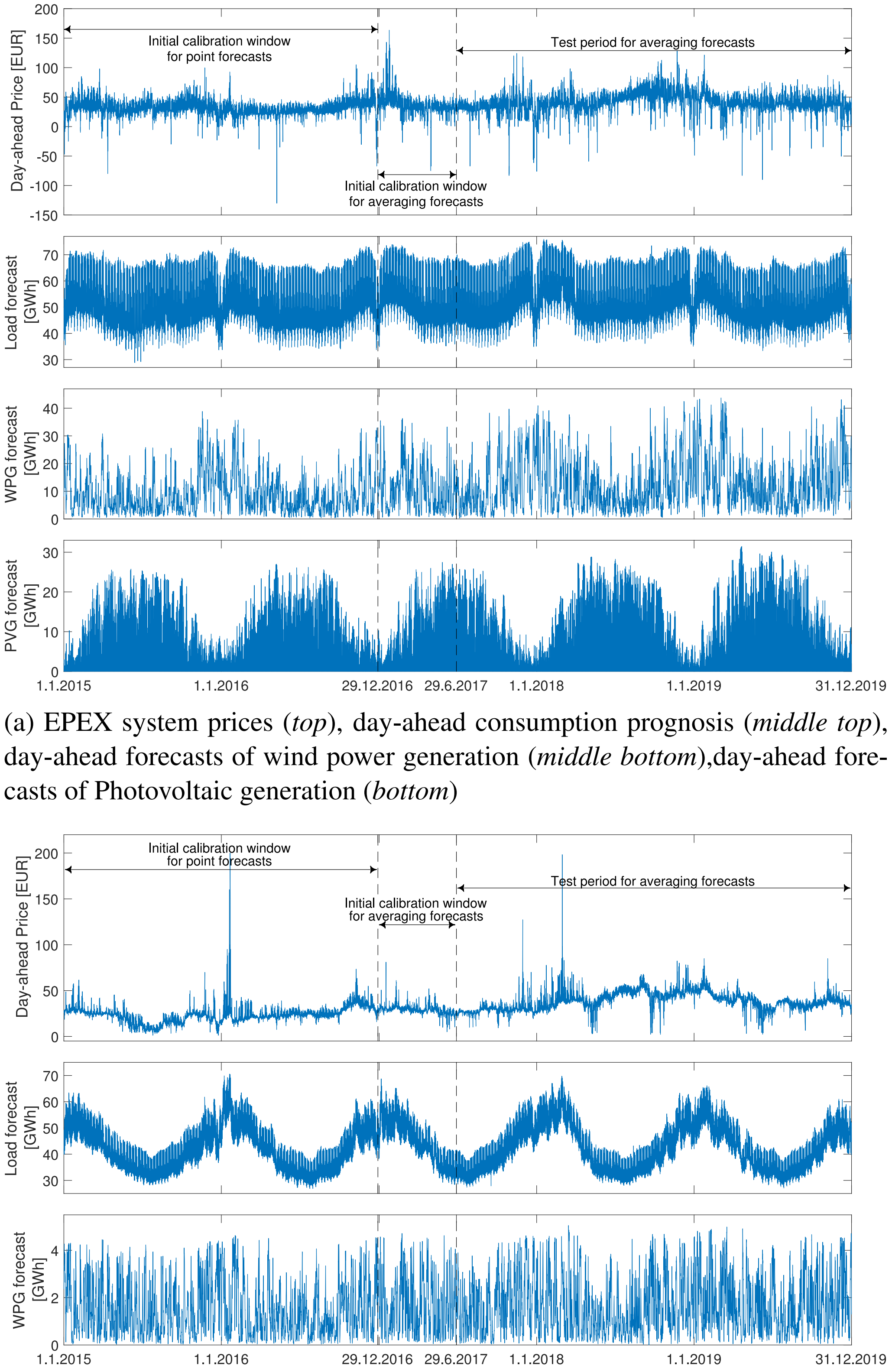
where the time series Y corresponds to the price series DA.

3. Methodology

3.1. Experiment design

The majority of research in EPF arbitrarily chooses the length of a calibration window. In last years, various research Hubicka et al. (see 2019), Maciejowska et al. (see 2020), Marcjasz et al. (see 2018), Serafin et al. (see 2019) has shown that averaging predictions based on different in-sample data lead to an improvement in forecast accuracy. Here, we follow this idea and use a pool of 673 calibration window lengths, ranging from 56 days (around two months) to 728 days (around two years). Unlike in previous papers, this research focuses on automating the averaging process in order to make it independent of the ad hoc decisions of forecasters.

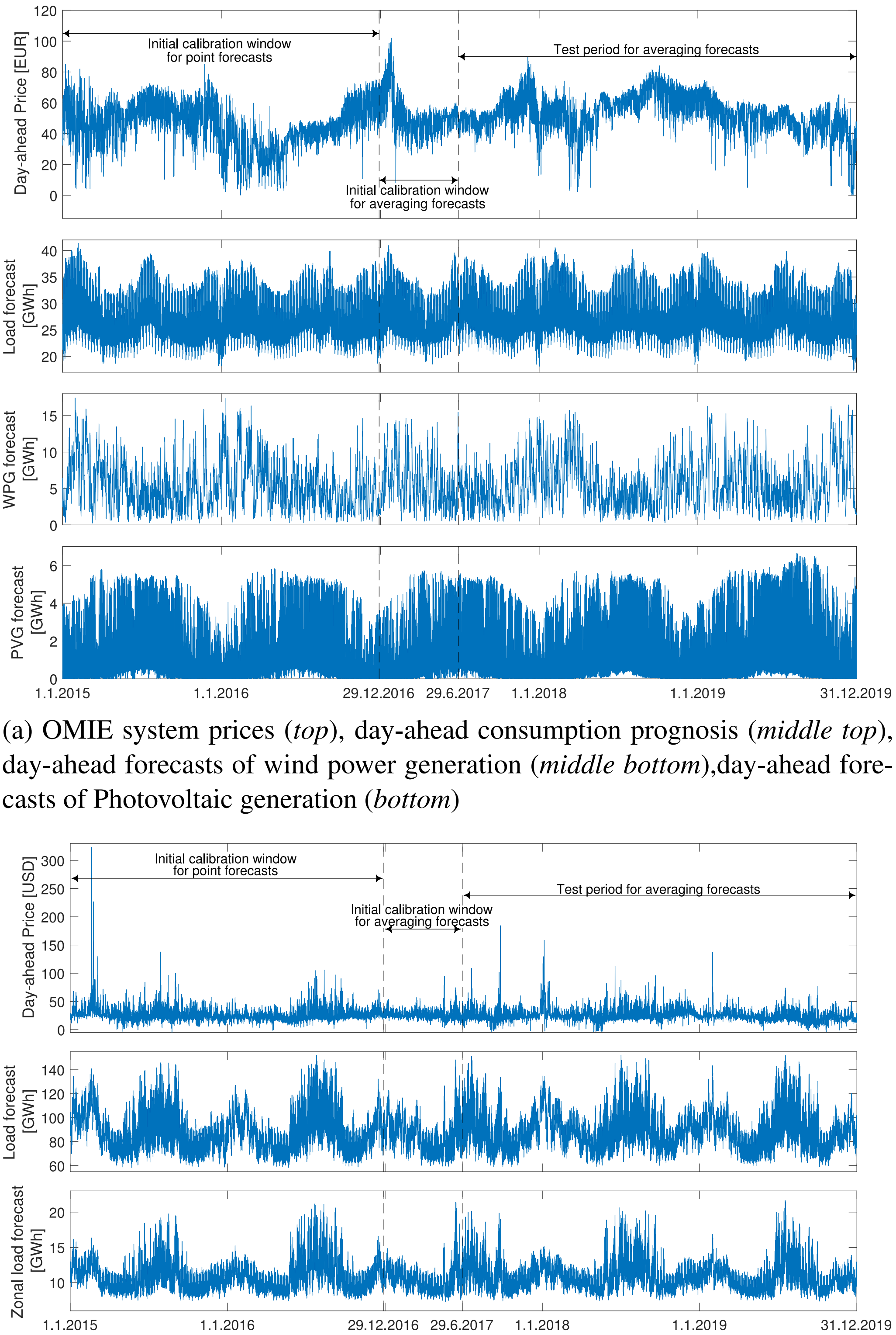
The pool of forecasts is obtained with a rolling window procedure, a standard procedure in EPF literature (Weron, 2014). To be more specific, the first 728 days are used for model estimation (for shorter windows, the calibration



(a) EPEX system prices (*top*), day-ahead consumption prognosis (*middle top*), day-ahead forecasts of wind power generation (*middle bottom*),day-ahead forecasts of Photovoltaic generation (*bottom*)

(b) Nord Pool system prices (*top*), day-ahead consumption prognosis (*middle*), day-ahead forecasts of wind power generation (*bottom*)

Fig. 1. Day-ahead prices and exogenous time series from 1 January 2015 to 31 December 2019. The vertical dashed lines respectively mark the beginning of the out-of-sample test period for point forecasts (29 December 2016; also the beginning of the initial 182-day calibration window for averaging forecasts) and the beginning of the out-of-sample test period for averaging forecasts (29 June 2017). The first 728 days constitute the initial calibration window for point forecasts.



(a) OMIE system prices (*top*), day-ahead consumption prognosis (*middle top*), day-ahead forecasts of wind power generation (*middle bottom*),day-ahead forecasts of Photovoltaic generation (*bottom*)

(b) PJM system prices (*top*), day-ahead system load prognosis (*middle*), day-ahead zonal (COMED) load prognosis (*bottom*)

Fig. 2. Day-ahead prices and exogenous time series from 1 January 2015 to 31 December 2019. The vertical dashed lines respectively mark the beginning of the out-of-sample test period for point forecasts (29 December 2016; also the beginning of the initial 182-day calibration window for averaging forecasts) and the beginning of the out-of-sample test period for averaging forecasts (29 June 2017). The first 728 days constitute the initial calibration window for point forecasts.

sample is left truncated, so it ends on the same day). Next, 24 point forecasts are computed, one for each hour of the day, and finally the window is moved one day forward. The procedure is repeated until the last out-of-sample day is reached. Once the pool of predictions is created, a rolling window of 182 days (around half a year) is used to calibrate the averaging methods (see Section 3.3). The final predictions are evaluated using the last 916 days of the sample. The divisions of the point forecast, averaging, and out-of-sample periods are marked by dashed lines in Figs. 1 and 2. The first line marks the end of the initial 728-day calibration window for point forecasts (i.e., 1 January 2015 to 28 December 2016). The second indicates the end of the initial 182-day calibration window for averaging forecasts (i.e., 28 June 2017), which is also the beginning of the evaluation period.

3.2. Forecasting models

In this research, forecasts for all 24 h of the next day are computed simultaneously a day in advance. Similarly to Maciejowska et al. (2020), we consider a parsimonious autoregressive structure used in a number of EPF studies (Uniejewski et al., 2016; Uniejewski & Weron, 2018; Uniejewski et al., 2018; Ziel & Weron, 2018). The originally proposed setup is expanded to include the exogenous variables presented in Section 2.2. The final model is denoted by **DA**. The price $DA_{d,h}$ for day d and hour h is described by the following formula:

$$\begin{aligned} DA_{d,h} = {} & \underbrace{\beta_{h,1}DA_{d-1,h} + \beta_{h,2}DA_{d-2,h} + \beta_{h,3}DA_{d-7,h}}_{\text{autoregressive effects}} + \\ & + \underbrace{\beta_{h,4}DA_{d-1,min} + \beta_{h,5}DA_{d-1,max}}_{\text{non-linear effects}} + \underbrace{\beta_{h,6}DA_{d-1,24}}_{\text{midnight price}} + \\ & + \underbrace{\sum_{i=1}^{7}\beta_{h,6+i}D_i}_{\text{weekday dummies}} + \underbrace{\theta_h X_{d,h}}_{\text{exogenous variables}} + \varepsilon_{d,h}, \qquad (3) \end{aligned}$$

where $DA_{d-1,h}$, $DA_{d-2,h}$, and $DA_{d-7,h}$ are the lagged day-ahead prices from one, two, and seven days before. $DA_{d-1,min}$ and $DA_{d-1,max}$ respectively refer to the minimum and the maximum price from day $d-1$. $DA_{d-1,24}$ is the last already-known price, corresponding to the previous day at midnight. $D_1, \ldots, D_7$ denote dummies that capture weekly seasonality. Finally, vector $X_{d,h}$ describes the exogenous variables. As stated in Section 2.2, $X_{d,h}$ differs across markets. The day-ahead forecasts of the load ($L_{d,h}$) are included in $X_{d,h}$ for all the countries, whereas the presence of other variables is restricted by their availability. For example, $W_{d,h}$ is used for all European countries but is not included in $X_{d,h}$ for the PJM market. Therefore, in the case of the USA, the zonal load forecasts ($Z_{d,h}$) are added instead. Additionally, for Germany and Spain, the photovoltaic generation $S_{d,h}$ is included. Note that, as in Maciejowska et al. (2019) and Maciejowska et al. (2020), $S_{d,h}$ is admitted in the model (3) only for hours 9–17, because during the night and early morning hours, solar generation is too weak to impact the electricity price.

3.3. Averaging methods

According to recent literature, the forecasting performance of statistical models is sensitive to the choice of the calibration window (Hubicka et al., 2019). Hence, it may be beneficial to average forecasts based on windows of different lengths (Hubicka et al., 2019; Pesaran & Timmermann, 2007), as this allows us to explore both the local and long-run behavior. Although estimating the same model with different datasets seems straightforward, forecast averaging remains a demanding task. First, a large number of predictions based on long windows are almost identical. Extending the sample by one observation from, for example, 727 to 728 days, does not alter the parameter estimates much. This feature impedes the usage of typical regressions for choosing averaging weights, as a large number of forecasts are almost co-linear. On the other hand, there are relatively few predictions based on short windows, which are distinct. Unfortunately, these forecasts are also more variable and typically burdened with a larger forecast error. Finally, it is not clear how to balance the impact of the short and long windows on the final prediction.

In this paper, we consider three types of forecast combination methods. First, predictions are computed either as a simple or weighted mean of individual forecasts. Next, the weights are selected with the LASSO method, which is a regression-based approach. LASSO allows us to include a large number of input variables and shrinks the parameters toward zero. Hence, it can help to select the optimal window lengths. Finally, the information included in the panel of forecasts is summarized by a set of common factors (computed as principal components, PCs), which are next used to compute the predictions of interest.

3.3.1. Linear average (simple average, AW, and WAW)

We consider three methods based on a linear average. The literature indicates that the arithmetic mean is a simple but very efficient approach (Genre, Kenny, Meyler, & Timmermann, 2004). Here, we compute the mean of all considered window sizes ranging from 56 days to 728 days, called the simple average. Second, following Hubicka et al. (2019), a subset of six calibration window lengths is selected, which consists of three short (56-, 84-, and 112-day) and three long (714-, 721-, and 728-day) in-sample sizes. Forecasts based on these chosen window sizes are then averaged. This approach is denoted AW(56, 84, 112, 714, 721, 728) or simply AW. Unfortunately, both the simple average and AW assume that the weights are equal and constant over time. Therefore, they cannot adapt to changing market conditions, for example, a rising share of renewable energy sources in the generation mix.

In order to overcome this problem, Marcjasz et al. (2018) proposed to extend AW to allow for data-driven weights. Similar to Hubicka et al. (2019), a small subset of available forecasts is first selected. Then, instead of taking a simple average, Marcjasz et al. (2018) use the forecast errors from the previous day to assign weights to each individual prediction. The forecasts are evaluated with the mean absolute error (MAE), and those that are more accurate are assigned higher weights (for more details, see Eq. (5) in Marcjasz et al. (2018)). Here, following Maciejowska et al. (2020), we use the whole averaging window (182 days) to compute the weights. Similar to AW, the weighted AW is denoted as WAW(56, 84, 112, 714, 721, 728) or simply WAW.

An application of linear averages is associated with some issues. First, when computing the simple average, the majority of inputs come from long calibration windows which provide very similar forecasts. Hence, the long windows dominate and reduce the impact of local behavior. This drawback is reduced in the AW and WAW approaches, as they include the same number of short and long windows and balance the impact of different window sizes. Unfortunately, AW/WAW, unlike the simple average, requires pre-selecting the number and lengths of calibration windows used for averaging. Hence, it cannot be considered a robust approach, because a subset that works well for one market may not be plausible for the other.

3.3.2. *LASSO averaging*

The idea of the regularization of an estimation process can be viewed as an optimization problem:

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}\left\{f(\boldsymbol{X};\boldsymbol{\beta}) + g(\boldsymbol{\beta})\right\}, \tag{4}$$

where $\boldsymbol{\beta}$ is a parameter vector, and $\boldsymbol{X}$ is a dataset. In Eq. (4), $f(\boldsymbol{X};\boldsymbol{\beta})$ denotes a loss function—e.g., the residual sum of squares (RSS), as in the least squares estimation method—while $g(\boldsymbol{\beta})$ is the penalty function (Tikhonov, 1963).

In the literature, it is common to use a scaled ℓ^q norm as $g(\boldsymbol{\beta})$. The most popular variant of the regularization, called LASSO, was introduced by Tibshirani (1996). It sets $q = 1$ and $f(\boldsymbol{X};\boldsymbol{\beta}) = \text{RSS}$ (see (5)). Due to its properties, it is a tool for automated variable selection and can successfully identify the most important variables (Uniejewski et al., 2016; Uniejewski & Weron, 2018).

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \operatorname{argmin}\left\{\text{RSS} + \lambda\sum_{i=1}^{n}|\beta_{h,i}|\right\}\\ &\equiv \operatorname{argmin}\left\{\sum_{d,h}\left(p_{d,h} - \sum_{i=1}^{n}\beta_{h,i}X_{d,h,i}\right)^2 + \lambda\sum_{i=1}^{n}|\beta_{h,i}|\right\},\end{aligned} \tag{5}$$

LASSO is also one of the most popular solutions to combine point forecasts. It has become a gold standard in the literature, especially for high-dimensional problems (i.e., when the number of individual predictions exceeds the number of in-sample observations). It has the property of selecting only a few individual point forecasts, even in the case of rich pools, which improves the accuracy. In a recent paper, Uniejewski and Weron (2021) showed that linear penalty regularization also works in probabilistic forecasting.

Here, LASSO regression is used to average all (673) point forecasts from the pool (see Section 3.1). We consider a log-scaled grid of 20 λ parameters (LASSO(λ)) and choose its optimal value via information criteria: the Akaike information criterion (AIC), Bayesian information criterion (BIC), and the Hannan–Quinn information criterion (HQC). The procedure to select the tuning parameter is taken from Ziel and Weron (2018) and its results are denoted by LASSO(BIC), LASSO(AIC), and LASSO(HQC).

3.3.3. *Principal component averaging (PCA)*

Many forecast averaging methods strongly depend on expert knowledge. For example, AW and WAW require pre-selecting the window lengths used in the forecast pooling. In order to overcome this issue, Maciejowska et al. (2020) proposed principal component averaging (PCA) to automate the procedure of averaging over a rich pool of predictions. The authors applied the principal component method to a panel of over 650 point forecasts obtained with models calibrated with different in-sample sizes. Next, they used the estimated components in a linear regression to form the final predictions. In such a way, they overcame the problem of the co-linearity of forecasts stemming from the same model calibrated on similar windows. Their results indicated that the PCA forecast averaging leads to more accurate predictions of electricity prices in terms of the MAE than the simple average, AW, or WAW.

The step-by-step algorithm of PCA is described below. In the algorithm, d_f denotes the forecasted day and $\tau = 56, 57, \ldots, 728$ stands for the length (in days) of a calibration window used to calculate the predictions. Moreover, during the averaging, all the hourly predictions are treated as time series and indexed with t. The averaging window includes the predicted day d_f and 182 proceeding days: $t \in \{24d + h : d_f - 182 \leq d \leq d_f, 1 \leq h \leq 24\}$. In the following parts of the paper, $\hat{P}_{t,\tau}$ denotes the predicted electricity prices for period t obtained with a τ-day calibration window, whereas P_t stands for their actual level.

1. For each time period, t, in an averaging window, estimate the mean ($\hat{\mu}_t$) and standard deviation ($\hat{\sigma}_t$) of forecasts ($\hat{P}_{t,\tau}$) across $\tau = 56, 57, \ldots, 728$.
2. Standardize the forecasts and the real price with the previously estimated $\hat{\mu}_t$ and $\hat{\sigma}_t$:
$$\hat{Z}_{t,\tau} = \frac{\hat{P}_{t,\tau} - \hat{\mu}_t}{\hat{\sigma}_t}, \quad Z_t = \frac{P_t - \hat{\mu}_t}{\hat{\sigma}_t}. \tag{6}$$
Notice that at the time of forecasting, the last 24 elements of Z_t, corresponding to the predicted day d_f, are not known.
3. Estimate the first K principal components, $(PC_{t,1}, PC_{t,2}, \ldots, PC_{t,K})$, of a panel $\{\hat{Z}_{t,\tau}\}$ using the method described by Bai and Ng (2002), Stock and Watson (2004). Notice that PCs include the information of the price forecasts for all hours in a 182-day averaging calibration window, as well as for the forecasted day.
4. Estimate linear regression parameters with least squares using observations from the averaging window (without day d_f):
$$Z_t = \alpha + \sum_{k=1}^{K}\beta_k PC_{t,k} + \varepsilon_t. \tag{7}$$

5. Using estimated parameters, compute the prediction of the normalized price Z_t for $t \in \big(24d_f + 1, 24d_f + 24\big)$ corresponding to all hours in forecasted day d_f:

$$\hat{Z}_t = \hat{\alpha} + \sum_{k=1}^{K} \hat{\beta}_k PC_{t,k} \tag{8}$$

and transform it back into its original level

$$\hat{P}_t = \hat{Z}_t \cdot \hat{\sigma}_t + \hat{\mu}_t. \tag{9}$$

Although PCA allows us to explore the information included in the whole panel of forecasts, it still requires selecting the number of components used in a regression, K. Therefore, similar to Maciejowska et al. (2020), we consider the method based on k-first PCs and denote them by PCA(k). For illustrative purposes, we also choose the ex post optimal (fixed) number of PCs taken for averaging, denoted PCA(best).

Next, three variants of PCA are applied, which are based on information criteria. This allows a data-driven adjustment of the number of PCs used in the regression (7). We consider the same information criteria, which are used to select λ in the LASSO procedure. The results are denoted consecutively by PCA(BIC), PCA(AIC), and PCA(HQC).

3.3.4. LASSO principal component averaging (LPCA)

In this paper, we propose a novel approach which combines a PCA-based procedure with LASSO estimation. First, similar to Maciejowska et al. (2020), K components are extracted from the standardized panel of point predictions (see Section 3.3.3 for a detailed description of the algorithm). Unlike in previous work, the number of PCs is substantial (here, 20 components) and can be arbitrarily big. Next, the PCs are used as input variables in the regression (7). In order to estimate the model's parameters, the LASSO method is applied. This approach enables the calibration of the model even when the number of PCs is larger than the size of the averaging calibration window. Moreover, it shrinks the parameters toward zero and hence reduces the noise induced by redundant components. Finally, the predictions of all hours of day d_f are calculated (8) and transformed back into the original units (9).

The LASSO optimization algorithm depends on a parameter λ which specifies the impact of the penalty function. Similar to LASSO averaging, we consider a log-scaled grid of 20 λ and select the optimal value via information criteria. The outcomes are denoted either by LPCA(λ) or by LPCA(BIC), LPCA(AIC), and LPCA(HQC).

The LPCA does not require any prior decision on the size of the calibration windows used for averaging (as in AW/WAW), and it is not restrictive in terms of the number of PCs (as in PCA). As such, it can be perceived as a fully automated method. Moreover, thanks to the orthogonality of the PCs, the estimation algorithm is faster than LASSO averaging.

4. Results

We use the mean absolute error (MAE) for the full out-of-sample test period of $D = 916$ days (i.e., 29.06.2017 to 31.12.2019; see Fig. 1 or 2) as the main evaluation criterion. The MAE is one of the most commonly used measures for evaluating forecast accuracy. In the case of electricity markets, it reflects the average deviation of the revenue from selling 1 MWh from its expected level. We consider two MAE-based measures:

$$MAE_d^{(i)} = \frac{1}{24} \sum_{h=1}^{24} |\varepsilon_{d,h}^{(i)}| \tag{10}$$

$$MAE^{(i)} = \frac{1}{D} \sum_{d=1}^{D} MAE_d^{(i)} \tag{11}$$

where $\varepsilon_{d,h}^{(i)} = P_{d,h} - \hat{P}_{d,h}^{(i)}$ is the forecast error for hour h in day d, obtained with either different calibration window lengths, τ, or averaging methods. The first measure, $MAE_d^{(i)}$, describes the forecast accuracy for a single day d and is used for a statistical comparison between individual approaches. Finally, $MAE^{(i)}$ describes the overall performance in the whole out-of-sample period.

As an auxiliary measure, we define a percentage change of forecast accuracy relative to the results of a model with the longest considered calibration window, the 728-day window ($MAE^{(728)}$):

$$\%\text{chng}_i = \frac{MAE^{(i)} - MAE^{(728)}}{MAE^{(728)}} \times 100\%. \tag{12}$$

The relative change in the accuracy of a given model shows how different the model is from the usual approach of taking calibration windows that are as long as possible. Note that a positive sign of the measure indicates that a given model is worse than the benchmark, while a negative value appears when a given model outperforms the longest-window approach.

Given a number of datasets, it is hard to rank the models' accuracy. To solve this issue, we use a mean of the $\%\text{chng}_i$ over four datasets to obtain the final ranking:

$$\text{m.p.d.b.}_i = \frac{1}{4} \sum_{m=1}^{4} \%\text{chng}_i^m, \tag{13}$$

where m indicates one of four datasets (EPEX, NP, OMIE, or PJM).

The obtained MAE values can be used to provide a ranking of forecasts. Unfortunately, they do not allow us to draw statistically significant conclusions on the outperformance of one prediction over another. Therefore, the conditional predictive ability (CPA) test of Giacomini and White (2006) is used to compare competitive outcomes. The test statistic is computed using the vector of average daily MAE_d:

$$\Delta_{i,j,d} = MAE_d^{(i)} - MAE_d^{(j)}. \tag{14}$$

For each pair (i, j), the p-value of the CPA test is computed.

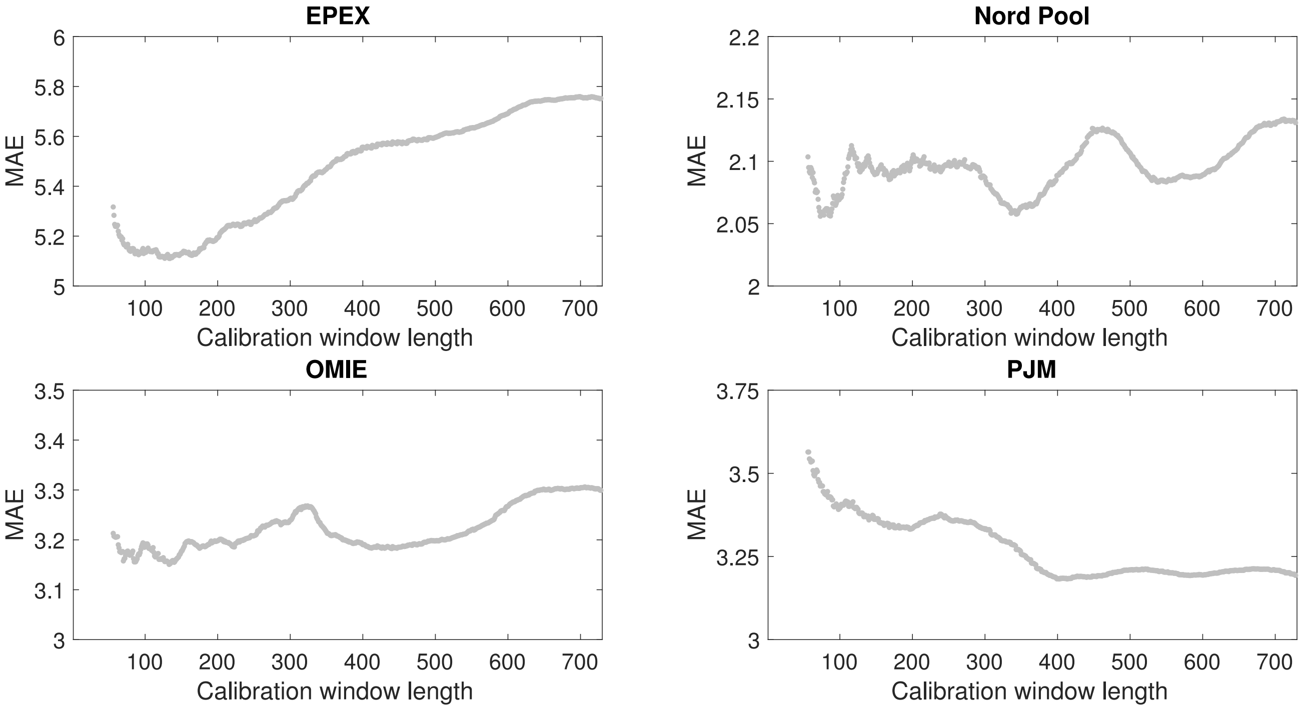


Fig. 3. Mean absolute errors (MAEs) for the EPEX, Nord Pool, OMIE, and PJM datasets for the period from 29.06.2017 to 31.12.2019 as a function of the calibration window length ranging from 56 to 728 days.

4.1. Individual forecasts

The performance of individual forecasts is presented in Fig. 3, which shows the MAE for different calibration window lengths in the four analyzed markets. It can be observed that the strategy for selecting the optimal size of the calibration window differs between the datasets. For some markets, such as EPEX, the longer the calibration window we take, the worse the autoregressive model performs. For others, such as PJM, it is beneficial to use long samples to estimate the model parameters. Finally, for Nord Pool and OMIE, the MAE plots are not monotonic and it is difficult to make an optimal decision. Hence, the results confirm the previous findings of Hubicka et al. (2019) and Marcjasz et al. (2018) and prove that it is impossible to ex ante choose the length of the optimal calibration window size.

Table 2 presents the detailed results for three selected window sizes: 56 days (8 weeks), 364 days (1 year), and 728 days (2 years). They are next compared with the benchmark, which is the longest available calibration window. The outcomes are augmented with the results for the optimal window size, which is selected ex post and hence is not available for real-time usage. The results indicate that the selection of the calibration window length may have a great impact on the forecast accuracy. The gains from its proper choice reach up to 12.527% (EPEX market).

4.2. Averaging results

Tables 3 and 4 present the MAE and %chng results for the forecasts obtained with different averaging techniques. Here, two approaches are evaluated separately: semi-automated and fully automated. In the first group of methods, arbitrary decisions of researchers about the number of components to be averaged are allowed. Moreover, the penalty parameter λ in the LASSO method is pre-defined for the whole sample. In the second group, the methods are fully automated, which means that the forecaster is not involved in the averaging process.

4.2.1. Semi-automated averaging methods

Let us first analyze the outcomes of semi-automated approaches, in which the researcher decides a priori on the selection of forecasts used for averaging. In all considered methods, the inputs are chosen once for the whole evaluation period and do not adjust as the calibration and averaging windows move. The results are reported in Table 3. First, the outcomes of the AW and WAW methods are presented based only on a small subset of individual point forecasts (three short and three long windows). It can be observed that both approaches yield results which are far better than the benchmark. By averaging forecasts stemming from just six different calibration windows, the MAE is reduced by more than 10% for EPEX, NP, and OMIE, and at least 3% for PJM. When both methods are compared, it can be observed that the weighted approach is better than AW, which assigns equal weights for all predictions.

Next, the error measures for LASSO, PCA, and LPCA with parameters selected ad hoc, based on existing literature and experience, are presented. For each method, the first three rows show outcomes for exemplary specifications described either by the number of components, k, in PCA(k) or by λ in LASSO(λ) and LPCA(λ). The forth row reports results for the best ex post value of these parameters. The outcomes confirm that using forecast averaging techniques is beneficial. Similar to AW/WAW,

Table 2
Mean absolute errors (MAEs) and the percentage change (%chng) compared to the simple average benchmark of the price forecasts for the whole 916-day out-of-sample period from 29.06.2017 to 31.12.2019. The results are presented for selected calibration window lengths of 56, 364, and 728 days.

Calib. window length	EPEX		NP		OMIE		PJM	
	MAE	%chng	MAE	%chng	MAE	%chng	MAE	%chng
56	5.339	−9.126%	2.210	−1.365%	3.181	−2.545%	3.674	11.075%
364	5.599	−4.695%	2.163	−3.450%	3.141	−3.767%	3.352	1.317%
728	5.875	0%	2.241	0%	3.264	0%	3.308	0%
best	5.139	−12.527%	2.159	−3.651%	3.100	−5.040%	3.299	−0.280%

Table 3
Mean absolute errors (MAEs) and the percentage change (%chng) compared to the simple average benchmark of the price forecasts for the whole 916-day out-of-sample period from 29.06.2017 to 31.12.2019. In this panel, we report the results obtained with averaging setups that depend on the forecaster's knowledge/intuition. Note that in each column, the best result is shown in bold.

Averaging	EPEX		NP		OMIE		PJM		
	MAE	%chng	MAE	%chng	MAE	%chng	MAE	%chng	m.p.d.b.
AW	5.059	−13.895%	1.970	−12.101%	2.917	−10.629%	3.206	−3.099%	−9.931%
WAW	5.014	−14.650%	1.966	−12.264%	2.913	−10.755%	3.204	−3.148%	−10.204%
LASSO(10^{-2})	5.416	−7.822%	2.408	7.464%	3.029	−7.216%	3.657	10.534%	0.740%
LASSO(10^{-1})	4.954	−15.671%	2.018	−9.954%	2.886	−11.575%	3.255	−1.595%	−9,699%
LASSO(10^{0})	**4.962**	**−15.536**%	1.984	−11.458%	**2.893**	**−11.356**%	3.230	−2.372%	−10.180%
LASSO(best)	4.924	−16.182%	1.963	−12.37%	2.872	−12.023%	3.200	−3.268%	−10.961%
PCA(1)	5.030	−14.380%	2.025	−9.612%	2.963	−9.210%	3.269	−1.195%	−8.599%
PCA(5)	5.007	−14.771%	1.980	−11.647%	2.913	−10.766%	3.220	−2.672%	−9.964%
PCA(20)	5.080	−13.524%	2.069	−7.663%	2.944	−9.803%	3.278	−0.915%	−7.976%
PCA(best)	4.965	−15.495%	1.969	−12.103%	2.913	−10.766%	3.210	−2.972%	−10.334%
LPCA(10^{-3})	4.998	−14.930%	2.022	−9.745%	2.914	−10.728%	3.244	−1.942%	−9.336%
LPCA(10^{-2})	4.979	−15.253%	**1.970**	**−12.064**%	2.904	−11.045%	**3.202**	**−3.209**%	**−10.393**%
LPCA(10^{-1})	5.107	−13.066%	2.058	−8.151%	3.014	−7.663%	3.270	−1.149%	−7.507%
LPCA(best)	4.970	−15.406%	1.961	−12.473%	2.893	−11.361%	3.197	−3.369%	−10.652%

Table 4
Mean absolute errors (MAEs) and the percentage change (%chng) compared to the simple average benchmark of the price forecasts for the whole 916-day out-of-sample period from 29.06.2017 to 31.12.2019. The results correspond to the fully automated approaches to averaging. Note that in each column, the best result is shown in bold.

Fully automated	EPEX		NP		OMIE		PJM		
	MAE	%chng	MAE	%chng	MAE	%chng	MAE	%chng	m.p.d.b.
simple average	5.275	−10.221%	2.068	−7.706%	3.021	−7.448%	3.270	−1.149%	−6.631%
LASSO(AIC)	5.853	−0.379%	2.304	2.810%	3.285	0.632%	3.998	20.866%	5.982%
LASSO(BIC)	5.005	−14.811%	1.989	−11.235%	**2.898**	**−11.206**%	3.259	−1.480%	−9.683%
LASSO(HQC)	5.221	−11.131%	2.084	−6.971%	2.968	−9.071%	3.542	7.069%	−5.026%
PCA(AIC)	5.012	−14.694%	2.014	−10.105%	2.949	−9.664%	3.249	−1.791%	−9.064%
PCA(BIC)	5.018	−14.590%	1.987	−11.342%	2.953	−9.540%	3.252	−1.691%	−9.291%
PCA(HQC)	5.005	−14.806%	2.004	−10.578%	2.945	−9.792%	3.251	−1.737%	−9.228%
LPCA(AIC)	4.947	−15.796%	1.988	−11.260%	2.931	−10.200%	3.221	−2.626%	−9.971%
LPCA(BIC)	**4.923**	**−16.212**%	**1.979**	**−11.689**%	2.924	−10.426%	**3.217**	**−2.759**%	**−10.271**%
LPCA(HQC)	4.927	−16.137%	1.988	−11.292%	2.932	−10.182%	3.221	−2.624%	−10.059%

all three methods enable a substantial reduction of the MAE, with the following specifications being the best: LASSO(10^0), PCA(5), and LPCA(10^{-2}).

When the LASSO averaging scheme is considered, it can be observed that the results depend strongly on the parameter λ. There are substantial differences between LASSO(10^{-2}) and LASSO(10^0), which reach 18.922% of the benchmark MAE for NP and 12.906% for PJM. Moreover, LASSO(10^{-2}) is the worst of the averaging schemes and provides predictions that are less accurate than the two-year calibration window for the NP and PJM markets.

The performance of PCA is more robust to the selection of the specification parameter, k. The relation between the MAE of PCA and the number of components, k, is non-monotonic. First, as the number of PCs increases, the forecasts become more accurate. As it reaches the optimal level of k, additional components introduce noise and lead to a higher MAE. Hence, increasing the number of components does not improve the overall performance of the method.

When the results of LPCA are analyzed, it can be observed that LPCA inherits the positive features of both PCA and LASSO and reduces their weaknesses. Similar to PCA, LPCA is robust to the choice of the tuning parameter, λ. On the other hand, it allows us to use a large number of components without a loss of efficiency, because LASSO allows for a reduced parameter space.

Finally, when LASSO(best), PCA(best), and LPCA(best) are compared, LPCA and LASSO are both the best in two out of four markets, with the PCA scheme never reaching the top of the podium. The aggregated results, summarized by m.p.d.b., confirm that LPCA yields the most accurate predictions among the alternatives.

The results for the non-automated averaging approaches can be summarized as follows:

- Almost all averaging approaches (except LASSO(10^{-2}) outperform the 'longest-window' model by a large margin, often by more than 10%.
- The most accurate forecast can be obtained with LASSO and LPCA; both are the best for two out of four datasets.
- The performance of LASSO depends strongly on λ, whereas PCA and LPCA are more robust to the choice of the specification parameters.
- The idea of AW and WAW, introduced by Hubicka et al. (2019) and Marcjasz et al. (2018), performs very well. However, it can be outperformed by more sophisticated approaches.

4.2.2. Fully automated averaging methods

We considered four fully automated forecast averaging methods. These are approaches that do not require any expert knowledge to select the inputs used for forecast averaging or to specify parameters such as the number of components, k, in PCA and a value of the LASSO tuning parameter, λ. The results are presented in Table 4, which (similar to Table 3) shows the MAE forecast accuracy measure and %chng.

The first method is a simple average. It is an automated approach because it does not require any pre-selection of the predictions used for pooling. This method provides forecasts that are far better than the benchmark. It reduces the MAE by 1.149%–10.221%, which is slightly less than in the case of AW/WAW.

Next, three methods are analyzed: LASSO, PCA, and LPCA. Unlike in the previous section, here the tuning parameters k and λ are selected with information criteria (AIC, BIC, and HQ). This modification has two major advantages. First, it does not require a priori knowledge of the specification of these methods in a particular application. Hence, it can be easily used to predict the prices of other commodities or for any other forecasting exercise. Second, the parameters can evolve as new data arrive and adjust to the market situation.

First, it can be noticed that the LASSO method is sensitive to the choice of information criterion. For the AIC, it provides forecasts which are less accurate than a benchmark in three out of four analyzed markets. For PJM, the loss of accuracy exceeds 20%. Even for the EPEX market, for which the gains are the highest, LASSO(AIC) is only slightly better than the predictions obtained with the longest calibration window. Moreover, LASSO(HQC), although better than LASSO(AIC), does not provide satisfactory results. It improves the predictions for EPEX, NP, and OMIE but worsens them for PJM by more than 7%. Only LASSO(BIC) gives results that are consistently better than the benchmark.

Similar to LASSO, the performance of the LPCA approach depends on the choice of information criterion. In this case, the differences between information criteria are less pronounced, with LPCA(BIC) providing the most accurate predictions. Hence, for the LPCA (as with the standard LASSO), the BIC should be used to select the parameter λ. It is worth noting that all three LPCA methods produce the best forecasts in terms of the MAE for the EPEX, NP, and PJM markets. They are outperformed only by LASSO(BIC) in the case of OMIE.

In the case of the PCA method, it is hard to choose a clear winner between different information criteria. For each dataset, a different approach provides the most accurate results. The differences, however, are not substantial, so the optimal number of PCs can be successfully selected via any of the considered information criteria. Although it is the most robust, this approach is never the best choice in terms of MAE accuracy, as it is outperformed by either LPCA or LASSO.

The last column of Table 4 presents m.p.d.b., the aggregated measure of forecast accuracy. The outcomes show that well-designed averaging models can outperform the most popular approach of an arithmetic mean. Moreover, they confirm previous findings obtained using semi-automated methods and indicate that LPCA reduces the MAE more than other averaging approaches.

To formally investigate the advantages of using our newly proposed averaging method, we apply the conditional predictive ability (CPA; see Giacomini and White (2006)) test for significant differences in the forecasting performance. The outcomes are presented in Fig. 4, where a non-black square indicates that the forecasts of the model on the x-axis are statistically more accurate than the forecasts of a model on the y-axis. The results confirm the previous findings and show the LPCA extension of the standard PCA approach significantly outperforms the other methods, in particular the simple mean and PCA, for each considered dataset. What is more, two out of four times, it is significantly better than the LASSO, and never worse.

Finally, it can be noticed that the simple average is outperformed by other averaging approaches almost every time. This result shows that the arithmetic mean is useful as a benchmark for the newly introduced methodology, but it should not be treated as a gold standard.

To sum up:

- Almost all averaging approaches (except LASSO(AIC) and LASSO(HQC)) can easily beat the 'longest-window' model by a large margin.
- Among the forecasts based on LASSO or LPCA methods, the most accurate results are obtained with the BIC.
- The PCA method is the most robust to the choice of information criterion. None of the information criteria dominates and all of them provide similar results.
- Overall, the best result can be obtained with LPCA(BIC).

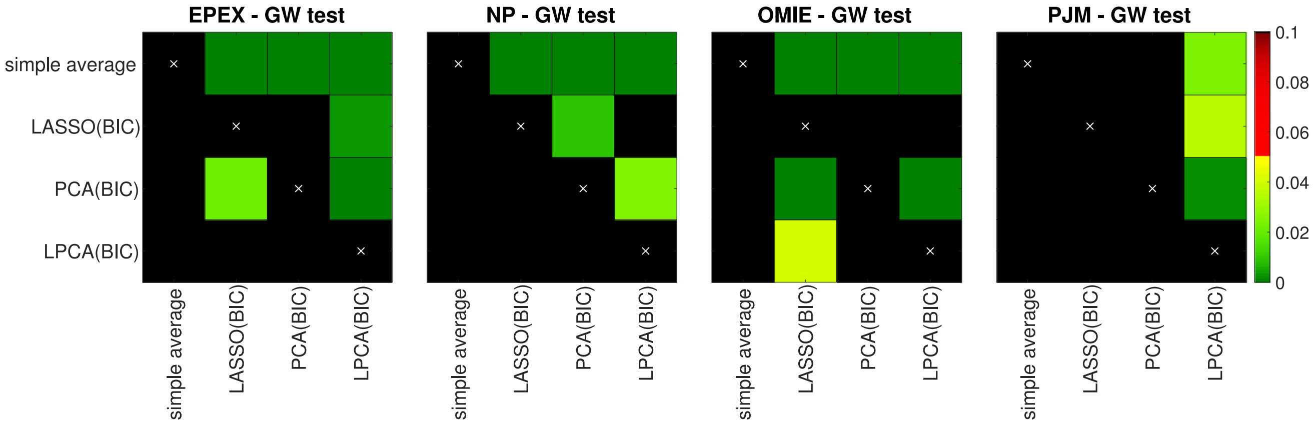


Fig. 4. Results of the conditional predictive ability (CPA) test of Giacomini and White (2006) for forecasts of selected models for the EPEX (left), Nord Pool (left center), OMIE (right center), and PJM (right) datasets. A heatmap indicates the range of the *p*-values: the closer they are to zero (→ dark green), the more significant the difference between the forecasts of a model on the *x*-axis (better) and the forecasts of a model on the *y*-axis (worse).

Table 5
Comparison of averaging methods based on m.p.d.b. across different specifications: choice of tuning parameter, k or λ, for semi-automated or information criteria for fully automated approaches. MAD, mean absolute deviation.

	Top	Mean	MAD
Semi-automated methods			
LASSO	−10.961%	−5.981%	5.165%
PCA	−10.334%	−9.232%	0.614%
LPCA	−10.652%	−9.582%	0.802%
Fully automated methods			
LASSO	−9.683%	−2.909%	5.927%
PCA	−9.291%	−9.194%	0.087%
LPCA	−10.271%	−10.100%	0.114%

4.3. Discussion

We analyzed the performance of different averaging schemes based on forecasts obtained with different calibration windows. We found that it is beneficial to pool predictions, even when they come from a single model. A large number of individual forecasts available for averaging becomes both an advantage and the main issue with this idea, which makes it difficult to fully automate the computations. Here, two approaches were explored based on information and parameter-space reduction. The PCA method can summarize the data described by a panel of forecasts with a relatively small set of orthogonal components, whereas LASSO shrinks the model's parameters toward zero and hence increases the estimation efficiency. The study demonstrates that the application of both approaches can result in a substantial increase in forecast accuracy. Unfortunately, the methods are burdened with the uncertainty associated with the choice of tuning parameters. The dependence of the results on this selection is illustrated in Table 5, which shows the best outcomes in terms of m.p.d.b. together with the mean and the mean absolute deviation (MAD) of m.p.d.b. across different specifications. The results indicate that although LASSO(10^0) and LASSO(BIC) are among the best forecast averaging approaches, the LASSO method is sensitive to the selection of the tuning parameter and the information criterion. Its average m.p.d.b. is slightly less than 6% and 3% for semi- and fully automated approaches, respectively. At the same time, LPCA improves forecasts by 9.582% and 10.1%, respectively. Moreover, the PCA and LPCA methods are characterized by low values of MAD, which are far smaller than in the case of LASSO.

The difference in performance of the LASSO, PCA, and LPAC forecast averaging methods results from their construction. When the PCA approach is considered, it should be emphasized that the components used for averaging are orthogonal to each other, which enables an efficient estimation of (8) parameters. However, unlike LPCA, this approach includes all PCs from 1 to k in the regression. The application of LASSO to (8) can reduce the parameter space. The method not only eliminates insignificant components but also shrinks the weights corresponding to less important variables. Uniejewski and Weron (2018) compared LASSO with a two-step procedure including variable selection via LASSO and estimating weights (of selected variables) via ordinary least squares. It turned out that LASSO significantly outperformed the two-step procedure. A similar situation was observed in our research. The limited study presented in Table 6 shows that applying the two-step procedure does not improve (on average) the forecast accuracy compared to PCA(BIC). This indicates that shrinkage is even more important than selection in our task. The regularization improves the averaging accuracy not because it allows for a better selection of the number of PCs, but because the LASSO shrunken weights are better to use in this setup.

Finally, when the LASSO and LPCA methods are compared, it can be noticed that LASSO has many more inputs than LPCA. Extracting information from the panel of forecasts via PCs reduces the dimension of the regression. Moreover, unlike with PCs, the individual forecasts are highly correlated and almost co-linear. Due to these features, LASSO is more sensitive to the specification of the tuning parameter. Moreover, the CPU needed to compute the forecast with LASSO is 900 times higher compared to the time needed to perform LPCA.

Table 6
Mean absolute errors (MAEs) and the percentage change (%chng) compared to the simple average benchmark of the price forecasts for the whole 916-day out-of-sample period from 29.06.2017 to 31.12.2019. The results are presented to compare LASSO with the two-step procedure proposed by Uniejewski and Weron (2018).

Averaging	EPEX		NP		OMIE		PJM		
	MAE	%chng	MAE	%chng	MAE	%chng	MAE	%chng	m.p.d.b
PCA(BIC)	5.018	−14.590%	1.987	−11.342%	2.953	−9.540%	3.252	−1.691%	−9.291
LPCA(BIC)	4.923	−16.212%	1.979	−11.689%	2.924	−10.426%	3.217	−2.759%	−10.272
2-step(BIC)	5.057	−13.919%	2.019	−9.901%	2.959	−9.348%	3.238	−2.125%	−8.823

5. Conclusions

In this paper, a novel approach to point forecast pooling was presented that combines LASSO estimation and the PCA scheme introduced by Maciejowska et al. (2020). PCA can summarize the information included in a panel of forecasts with a relatively small set of orthogonal components, whereas LASSO shrinks the model's parameters toward zero and hence increases the estimation efficiency. The performance of the approach was evaluated on datasets from four major energy markets. Following Marcjasz et al. (2018) and Hubicka et al. (2019), the point predictions used for pooling stemmed from a single ARX-type model calibrated to windows of different sizes. The forecasts were evaluated with the MAE, and the results were presented relative to the outcomes obtained with the longest available calibration window, which included two years of observations.

The results confirmed previous findings of Marcjasz et al. (2018) and Maciejowska et al. (2020) that the longest estimation window does not necessarily lead to the most accurate predictions. Hence, it is not possible to select a prior optimal length of the sample used for calibration. At the same time, averaging algorithms can substantially reduce the MAE and improve the forecast accuracy relative to the benchmark, by −6.631% for a simple average and by −10.271% for the LPCA(BIC) approach.

When forecast averaging methods were considered, the outcomes indicated that fully automated approaches, which use information criteria to select an optimal specification, yielded significantly better results than the benchmark or the simple average. The performance of the presented pooling methods depended on the applied information criterion, however. The results showed that the BIC was the most robust choice, leading to the lowest relative MAE for all approaches. Based on a comparison of LASSO, PCA, and LPCA, we draw the following conclusions:

- The PCA method is the most robust to the choice of information criterion. However, it reduces the MAE less than the methods using LASSO.
- LASSO is extremely sensitive to the choice of the tuning parameter and information criterion.
- Overall, LPCA outperforms other approaches: it improves the forecast accuracy the most and is relatively robust to the selection of the tuning parameter.

The LPCA approach, which combines LASSO with PCA, was shown to be successful at forecasting day-ahead electricity prices. This research could be viewed as a first step in mixing PCA with automated variable selection methods. Future analysis may include more complex models, such as the elastic net, adaptive LASSO, or neural network-based models. Moreover, the research may be extended to interval and probabilistic forecasting and to other commodity markets.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by the Ministry of Science and Higher Education (MNiSW, Poland) through Diamond Grant No. 0199/DIA/2019/48 (to BU), and by the National Science Centre (NCN, Poland) through SONATA BIS Grant No. 2019/34/E/HS4/00060 (to KM).

References

Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, *70*(1), 191–221.

Banerjee, A., Marcellino, M., & Masten, I. (2014). Forecasting with factor-augmented error correction models. *International Journal of Forecasting*, *30*(3), 589–612.

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly*, *20*(4), 451–468.

Bayer, S. (2018). Combining value-at-risk forecasts using penalized quantile regressions. *Econometrics and Statistics*, *8*, 56–77.

Boivin, J., & Ng, S. (2005). Understanding and comparing factor-based forecasts. *International Journal of Central Banking*, *1*(3).

Bracale, A., Carpinelli, G., & Falco, P. D. (2019). Developing and comparing different strategies for combining probabilistic photovoltaic power forecasts in an ensemble method. *Energies*, *12*(6), 1–16.

Chan, Y. L., Stock, J. H., & Watson, M. W. (1999). A dynamic factor model framework for forecast combination. *International Journal of Forecasting*, *22*, 283–300.

Crane, D., & Crotty, J. (1967). A two-stage forecasting model: Exponential smoothing and multiple regression. *Management Science*, *13*(8), B501âeB507.

Diebold, F. X., & Shin, M. (2019). Machine learning for regularized survey forecast combination: Partially-egalitarian LASSO and its derivatives. *International Journal of Forecasting*, *35*, 1679–1691.

Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2000). The generalized dynamic factor model: Identification and estimation. *The Review of Economics and Statistics*, *82*(4), 540–554.

Forni, M., & Lippi, M. (2001). The generalized dynamic factor model: Representation theory. *Economic Theory*, *17*, 1113–1141.

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

Genre, V., Kenny, G., Meyler, A., & Timmermann, A. (2004). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, *29*(1), 108–121.

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica*, *74*(6), 1545–1578.

Gianfreda, A., Parisio, L., & Pelagatti, M. (2016). The impact of RES in the Italian day-ahead and balancing markets. *Energy Journal*, *37*, 161–184.

Hibon, M., & Evgeniou, T. (2005). To combine or not to combine: Selecting among forecasts and their combinations.. *International Journal of Forecasting*, *21*, 15–24.

Huang, H., & Lee, T.-H. (2010). To combine forecasts or to combine information? *Econometric Reviews*, *29*(5–6), 534–570.

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*(1), 321–323.

Huisman, R., Huurman, C., & Mahieu, R. (2007). Hourly electricity prices in day-ahead markets. *Energy Economics*, *29*, 240–248.

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

Kowalska-Pyzalska, A. (2018). What makes consumers adopt to innovative energy services in the energy market? A review of incentives and barriers. *Renewable and Sustainable Energy Reviews*, *82*, 3570–3581.

Ludwig, N., Feuerriegel, S., & Neumann, D. (2015). Putting big data analytics to work: Feature selection for forecasting electricity prices using the LASSO and random forests. *Journal of Decision Systems*, *24*(1), 19–36.

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

Maciejowska, K., Nitka, W., & Weron, T. (2019). Day-ahead vs. Intraday – Forecasting the price spread to maximize economic benefits. *Energies*, *12*(4), 631.

Maciejowska, K., Uniejewski, B., & Serafin, T. (2020). PCA forecast averaging – Predicting day-ahead and intraday electricity prices. *Energies*, *13*, 3530.

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

Marcos, R. A. d., Bunn, D. W., Bello, A., & Reneses, J. (2020). Short-term electricity price forecasting with recurrent regimes and structural breaks. *Energies*, *13*(20).

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

Nowotarski, J., & Weron, R. (2016). To combine or not to combine? Recent trends in electricity price forecasting. *ARGO*, *9*, 7–14.

Pesaran, M., & Timmermann, A. (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics*, *137*(1), 134–161.

Poncela, P., Rodriguez, J., Sanchez-Mangas, R., & Senra, E. (2011). Forecast combination through dimension reduction techniques. *International Journal of Forecasting*, *27*, 224–237.

Serafin, T., Uniejewski, B., & Weron, R. (2019). Averaging predictive distributions across calibration windows for day-ahead electricity price forecasting. *Energies*, *12*(13), 256.

Stock, J. H., & Watson, M. W. (1998). Testing for common trends. *Journal of the American Statistical Association*, *83*, 1097–1107.

Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, *97*(460), 1167–1179.

Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, *23*, 405–430.

Stock, J. H., & Watson, M. W. (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics*, *30*(4), 481–493.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, *58*, 267–288.

Tikhonov, A. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics—Doklady*, *4*, 1035âe1038.

Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. W. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting* (pp. 135–196). Elsevier.

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

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

Uniejewski, B., & Weron, R. (2021). Regularized quantile regression averaging for probabilistic electricity price forecasting. *Energy Economics*, *95*, Article 105121.

Uniejewski, B., Weron, R., & Ziel, F. (2018). Variance stabilizing transformations for electricity spot price forecasting. *IEEE Transactions on Power Systems*, *33*, 2219–2229.

Wallis, K. (2011). Combining forecasts âe Forty years later. *Applied Financial Economics*, *21*, 33–41.

Weron, R. (2014). 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.

Ziel, F., Steinert, R., & Husmann, S. (2015). Efficient modeling and forecasting of electricity spot prices. *Energy Economics*, *47*, 89–111.

Ziel, F., & Weron, R. (2018). Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate modeling frameworks. *Energy Economics*, *70*, 396–420.

Paper 3

On the importance of the long-term seasonal component in day-ahead electricity price forecasting Part II — Probabilistic forecasting

Bartosz Uniejewski, Grzegorz Marcjasz, Rafał Weron

Energy Economics 79 (2019) 171–182

Contents lists available at ScienceDirect

Energy Economics

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

On the importance of the long-term seasonal component in day-ahead electricity price forecasting Part II — Probabilistic forecasting

Bartosz Uniejewski[a, b], Grzegorz Marcjasz[a, b], Rafał Weron[a, *]

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

ARTICLE INFO

Article history:
Received 6 May 2017
Received in revised form 9 January 2018
Accepted 16 February 2018
Available online 22 February 2018

JEL classification:
C14
C22
C51
C53
Q47

Keywords:
Electricity spot price
Long-term seasonal component
Seasonal Component AutoRegressive (SCAR) model
Probabilistic forecasting
Risk management
Quantile Regression Averaging (QRA)
Pinball score

ABSTRACT

A recent electricity price forecasting study has shown that the *Seasonal Component AutoRegressive* (SCAR) modeling framework, which consists of decomposing a series of spot prices into a trend-seasonal and a stochastic component, modeling them independently and then combining their forecasts, can yield more accurate point predictions than an approach in which the same autoregressive model is calibrated to the prices themselves. Here, we show that further accuracy gains can be achieved when the explanatory variables (load forecasts) are deseasonalized as well. More importantly, considering a novel extension of the SCAR concept to probabilistic forecasting and applying two methods of combining predictive distributions, we find that (i) SCAR-type models nearly always significantly outperform the autoregressive benchmark but are in turn outperformed by combined SCAR forecasts, (ii) predictive distributions computed using Quantile Regression Averaging (QRA) outperform those obtained from historical simulation and bootstrap methods, and (iii) averaging over predictive distributions generally yields better probabilistic forecasts of electricity spot prices than averaging over quantiles. Given that probabilistic forecasting is a concept closely related to risk management, our study has important implications for risk officers and portfolio managers in the power sector.

1. Introduction

Most day-ahead *electricity price forecasting* (EPF) studies treat the daily and weekly seasonalities as an inherent feature of any EPF model, but ignore the *long-term seasonal component* (LTSC; also called the *trend-seasonal component*, see Weron, 2014, for a recent review). However, as Nowotarski and Weron (2016) have recently shown, decomposing a series of spot prices into a LTSC and a stochastic component, modeling them independently and then combining their forecasts can yield more accurate point predictions than an approach in which the same autoregressive model is calibrated to the prices themselves. The authors have dubbed their approach/model the *Seasonal Component AutoRegressive* (SCAR) model. They have also conjectured that by considering SCAR models with different LTSCs a forecaster should be able to generate a pool of accurate, yet to a large extent independent forecasts that could be combined to yield even better predictions (an idea which is similar in spirit to combining so-called *sister forecasts* in load forecasting, see Liu et al., 2017; Nowotarski et al., 2016).

The main aim of this paper is to validate the latter conjecture in the context of probabilistic forecasts. To this end we perform an extensive empirical study which involves:

- two 1.5-year long, hourly resolution test periods from two distinct power markets (GEFCom2014 and Nord Pool),
- a relatively well performing, parsimonious autoregressive structure[1] (denoted in the text by **ARX**), originally proposed

[*] Corresponding author.
E-mail addresses: rafal.weron@pwr.edu.pl, rafal.weron@pwr.wroc.pl (R. Weron).

[1] Since such models are built on some prior knowledge of experts, following Uniejewski et al. (2016) and Ziel (2016), we refer to them as expert models.

https://doi.org/10.1016/j.eneco.2018.02.007
0140-9883/© 2018 Elsevier B.V. All rights reserved.

by Misiorek et al. (2006) and later used in a number of EPF studies (Gaillard et al., 2016; Kristiansen, 2012; Maciejowska et al., 2016; Nowotarski and Weron, 2016; Nowotarski et al., 2014; Serinaldi, 2011; Uniejewski et al., 2016; Weron, 2006; Weron and Misiorek, 2008, Ziel, 2016),

- a range of *Seasonal Component AutoRegressive* (SCAR) type models introduced by Nowotarski and Weron (2016) and built on the **ARX** structure (denoted in the text by **SCARX** with a subscript representing the LTSC used),
- two well-performing LTSC model classes — the Hodrick-Prescott and wavelet filters, as advocated for EPF by Janczura et al. (2013), Lisi and Nan (2014), Nowotarski et al. (2013) and Weron and Zator (2015),
- three methods of constructing probabilistic forecasts — historical simulation, bootstrapping and Quantile Regression Averaging (see Nowotarski and Weron, 2018, for a review),
- two approaches to combining probabilistic forecasts — averaging quantiles and averaging predictive distributions (see Lichtendahl et al., 2013, for a discussion),
- forecast evaluation in terms of the robust *weekly-weighted mean absolute error* (WMAE; see Weron, 2014) for point forecasts, the *pinball loss* function for probabilistic forecasts (Gneiting, 2011; Hong et al., 2016), and the Diebold and Mariano (1995) test for significant differences in forecasting performance,

and draw statistically significant conclusions with far reaching consequences for probabilistic EPF. Moreover, given that probabilistic forecasting is a concept closely related to risk management, our study provides producers, retailers and speculators with efficient quantitative tools that can be used as an aid in determining optimal strategies for short-term operations, Value-at-Risk calculations, hedging and trading (Bunn et al., 2016).

We should also note that our study draws heavily on Nowotarski and Weron (2016), but there are some important differences as well. In particular, we use the same datasets, the same **ARX** structure and the same collection of SCAR-type models built on the latter using a Hodrick-Prescott or wavelet filter. On the other hand, the differences include (i) using only one autoregressive structure (the mARX model yielded similar conclusions at the qualitative level and its performance is not reported here), consequently only one class of SCAR-type models (i.e., **SCARX**; mSCARX are not considered), (ii) utilizing system – not zonal – loads for the GEFCom2014 dataset as they have turned out to yield slightly better forecasts, (iii) using deseasonalized – instead of original – exogenous variables in **SCARX** models, and – most importantly – (iv) considering a whole plethora of algorithms to construct, combine and evaluate probabilistic forecasts.

The remainder of the paper is structured as follows. In Section 2, we present the datasets. Then in Section 3, we describe the techniques considered for price forecasting: the baseline autoregressive model structure, two LTSC model classes, the set of SCAR models, and the methods of constructing and combining probabilistic forecasts. In Section 4, we summarize the empirical findings and in Section 5 wrap up the results and conclude.

2. Datasets

The datasets considered in this empirical study include the same two day-ahead time series as studied by Nowotarski and Weron (2016). The first one comes from the Global Energy Forecasting Competition 2014 (GEFCom2014) and includes three time series at hourly resolution: locational marginal prices (LMPs), day-ahead predictions of zonal loads and day-ahead predictions of system loads (available as supplementary material in Hong et al., 2016). It covers the period from January 1, 2011 to December 16, 2013.[2] In this paper, we only use two subseries — LMPs and day-ahead predictions of system loads, see Fig. 1. In contrast to Nowotarski and Weron (2016), we use system, not zonal loads, as they have turned out to yield slightly better forecasts. Note, that the origin of the data has never been revealed by the organizers, but given its features it quite likely comes from the U.S.

The second dataset comprises Nord Pool (NP) system prices and *consumption prognosis* for four Nordic countries (Denmark, Finland, Norway and Sweden) for every hour in the period January 1, 2013–December 24, 2015, see Fig. 2. The time series were constructed using data published by the Nordic power exchange Nord Pool (www.nordpoolspot.com) and preprocessed to account for missing values and changes to/from the daylight saving time (like in Weron, 2006, Section 4.3.7). The missing data values were substituted by the arithmetic average of the neighboring values. The 'doubled' values (corresponding to the changes from the daylight saving/summer time) were substituted by the arithmetic average of the two values for the 'doubled' hour.

Like in Nowotarski and Weron (2016), the day-ahead point forecasts of the hourly electricity price are determined within a rolling window scheme, using a 360-day calibration window. First, all considered models (their short-term and long-term components) are calibrated to data from the initial calibration period, i.e. January 1 to December 26 (year 2011 for GEFCom2014 and 2013 for Nord Pool) and forecasts for all 24 h of the next day (December 27) are determined. Then the window is rolled forward by one day and forecasts for all 24 h of December 28 are computed. This procedure is repeated until the predictions for the last day in the 103-week (for GEFCom2014) or 104-week (for Nord Pool) test sample are made.

Once the point predictions are made, they are used to provide probabilistic forecasts. Due to the nature of the methods used, the procedure is different for historical/QRA and bootstrapped forecasts, see Section 3.3. The former two require a subsample of one-day ahead prediction errors. Hereby, a 182-day (or 26-week) rolling calibration window is used for computing quantiles of the error distribution (historical *prediction intervals*, PIs) or weights of the QRA approach. On the other hand, the bootstrapped PIs are computed directly from point forecasts in a 360-day rolling window. Formally, the latter are available from December 27 (2011 for GEFCom2014 and 2013 for Nord Pool). However, to allow direct comparisons they are computed only for the same test period as historical/QRA forecasts: June 26, 2012–December 16, 2013 for GEFCom (77 full weeks) and June 27, 2014–December 24, 2015 for Nord Pool (78 full weeks).

3. Methodology

The modeling is implemented separately across the hours, leading to 24 sets of parameters for each day the forecasting exercise is performed. This 'multivariate' approach is inspired by the fact that each hour displays a rather distinct price profile (reflecting the daily variation of demand, costs and operational constraints), by the extensive research on demand forecasting (which has generally favored the multi-model specification for short-term predictions) and the results of a recent comparative study of Ziel and Weron (2018) which concludes that on average the 'multivariate' approach has a minor edge over 'univariate' models in predictive performance (although it does not consistently outperform them across all datasets, seasons of the year or hours of the day).

[2] The last day in the GEFCom2014 dataset is actually December 17, 2013. However, we end the analysis on December 16, 2013, so that the test period length is a multiple of 7 days.

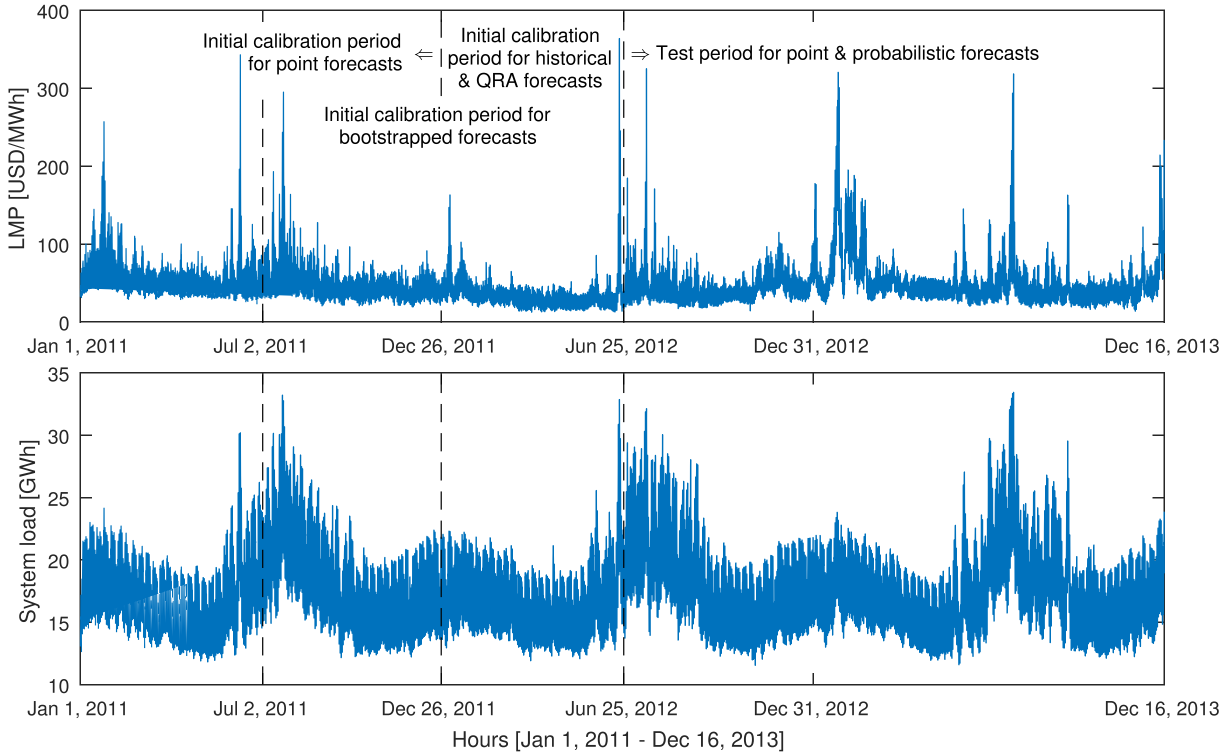


Fig. 1. GEFCom2014 hourly locational marginal prices (LMP; *top*) and hourly day-ahead predictions of system load (*bottom*) for the period January 1, 2011–December 16, 2013. The vertical dashed lines mark (i) the beginning of the initial calibration window for bootstrapped forecasts, (ii) the beginning of the initial calibration window for historical and QRA forecasts (at the same time — the end of the initial calibration period for point forecasts and the beginning of the test period for point forecasts) and (iii) the beginning of the test period for point and probabilistic forecasts.

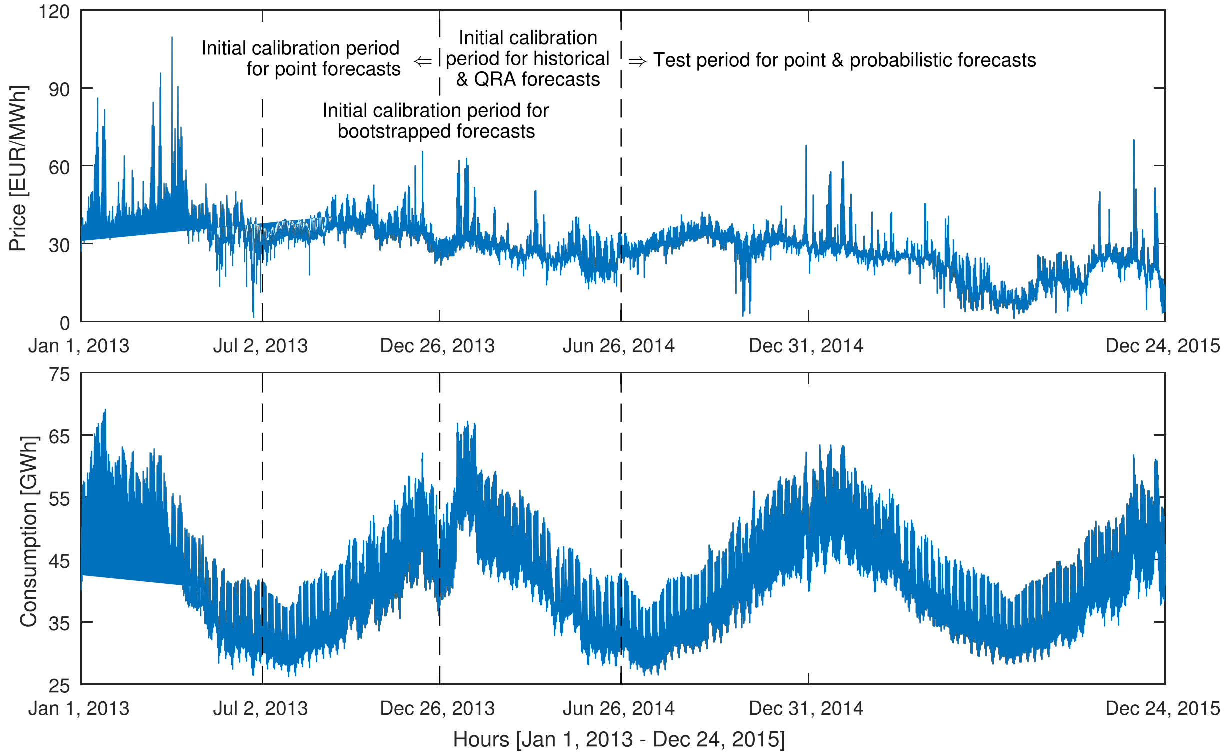


Fig. 2. Nord Pool hourly system prices (*top*) and hourly consumption prognosis (*bottom*) for the period January 1, 2013–December 24, 2015. The vertical dashed lines mark (i) the beginning of the initial calibration window for bootstrapped forecasts, (ii) the beginning of the initial calibration window for historical and QRA forecasts (at the same time — the end of the initial calibration period for point forecasts and the beginning of the test period for point forecasts) and (iii) the beginning of the test period for point and probabilistic forecasts.

3.1. The benchmarks

Similarly to Nowotarski and Weron (2016), we consider benchmark models. The first one belongs to the class of similar-day techniques. Most likely, it was introduced to the EPF literature by Nogales et al. (2002) and dubbed the *naïve method*. It proceeds as follows: hour h on Monday is similar to the same hour on Monday of the previous week, and the same rule applies for Saturdays and Sundays; hour h on Tuesday is similar to the same hour on Monday, and the same rule applies for Wednesdays, Thursdays and Fridays. As was argued by Conejo et al. (2005) and Nogales et al. (2002), forecasting procedures that are not calibrated carefully fail to pass this 'naïve test' surprisingly often. We denote this benchmark by **Naïve**.

The second benchmark is a parsimonious autoregressive structure originally proposed by Misiorek et al. (2006) and later used in a number of EPF studies (Gaillard et al., 2016; Kristiansen, 2012; Maciejowska et al., 2016; Nowotarski and Weron, 2016; Nowotarski et al., 2014; Serinaldi, 2011; Uniejewski et al., 2016; Weron, 2006; Weron and Misiorek, 2008; Ziel, 2016). Within this model, the demeaned natural logarithm of the electricity spot price on day d and hour h, i.e., $p_{d,h} = \log(P_{d,h})$, is given by the following formula:

$$p_{d,h} = \underbrace{\beta_{h,1}p_{d-1,h} + \beta_{h,2}p_{d-2,h} + \beta_{h,3}p_{d-7,h}}_{\text{autoregressive effects}} + \underbrace{\beta_{h,4}p^{\min}_{d-1}}_{\text{non-linear effect}} + \underbrace{\beta_{h,5}z_t}_{\text{load forecast}} + \underbrace{\sum\nolimits_{i=1}^{3}\beta_{h,5+i}D_i}_{\text{Mon, Sat, Sun dummies}} + \varepsilon_{d,h}, \qquad (1)$$

where the lagged log-prices $p_{d-1,h}$, $p_{d-2,h}$ and $p_{d-7,h}$ account for the autoregressive effects of the previous days (the same hour yesterday, two days ago and one week ago), while $p^{\min}_{d-1}$ is the minimum of the previous day's 24 hourly log-prices, which creates a link between bidding and price signals from the entire previous day. The variable z_t refers to the logarithm of hourly system load of a U.S. utility or Nordic consumption (actually to forecasts made a day before, see Section 2). The three dummy variables – D_1, D_2 and D_3 (for Monday, Saturday and Sunday, respectively) – account for the weekly seasonality. Finally, the ε_t's refer to error terms for log-prices, i.e., $\varepsilon_{d,h} = p_{d,h} - \hat{p}_{d,h} = \log(P_{d,h}) - \log(\hat{P}_{d,h})$, and are assumed to be independent and identically distributed (i.i.d.) normal variables. We denote this autoregressive benchmark by **ARX** to reflect the fact that the load (or consumption) forecast is used as the eXogenous variable in Eq. (1). Note, that compared to Nowotarski and Weron (2016) we use here an explicit day-hour 'multivariate' notation. Naturally, it is linked to the 'univariate' notation via $P_t = P_{24d+h} = P_{d,h}$.

3.2. Seasonal Component AutoRegressive (SCAR) models

Recall from Nowotarski and Weron (2016) that a *Seasonal Component AutoRegressive* (SCAR) model consists of two elements — a LTSC and an autoregressive structure[3] that forms the backbone of a family of SCAR models. The approach consists of (i) decomposing a series of electricity log-prices into a LTSC and a stochastic component (or residual), (ii) modeling them independently and (iii) combining their forecasts. We decompose the electricity spot log-price series $p_{d,h}$ into a sum of two independent parts:

- $q_{d,h} = X_{d,h} + s_{d,h}$, i.e. the stochastic component $X_{d,h}$ with weekly periodicities $s_{d,h}$,
- and $T_{d,h}$, i.e. the long-term seasonal component.

Motivated by a series of recent articles on modeling and forecasting the LTSC of electricity spot prices (see Janczura et al., 2013; Lisi and Nan, 2014; Nowotarski et al., 2013; Weron and Zator, 2015, among others), we consider two well-performing model classes — the Hodrick-Prescott (HP) and wavelet filters.

3.2.1. The Hodrick-Prescott (HP) filter

The Hodrick and Prescott (1997) filter was originally proposed in macroeconomics for decomposing the series of GDP values into a long-term growth component and the business cycle. However, the mechanics of the HP filter are universal and it has been found to perform well in EPF (Lisi and Nan, 2014; Weron and Zator, 2015). When applied to electricity spot prices it splits the series into a smooth part — the LTSC, and a volatile part — the stochastic component with weekly (short-term) periodicities. For a 'noisy' input series of electricity log-prices p_t, the HP filter returns a smoothed series T_t which minimizes (here the 'univariate' notation, i.e., $p_t = p_{24d+h} = p_{d,h}$ and $T_t = T_{24d+h} = T_{d,h}$, is more convenient):

$$\min_{T_t}\left\{\sum_{t=1}^{\tau}(p_t - T_t)^2 + \lambda\sum_{t=2}^{\tau-1}[(T_{t+1} - T_t) - (T_t - T_{t-1})]^2\right\}, \qquad (2)$$

where τ is the number of observations (in this study: $360 \times 24 = 8640$ h of the calibration window) and λ is a smoothing parameter. To find the optimal value of λ, we use a similar grid as in Weron and Zator (2015), the only difference is that in this study the price series are in hourly (not daily) resolution and the values of λ have to be larger. We use eight different λs: 10^8, $5\cdot10^8$, 10^9, $5\cdot10^9$, 10^{10}, $5\cdot10^{10}$, 10^{11} and $5\cdot10^{11}$.

3.2.2. Wavelet smoothing

Recall, that any function or signal (here: the electricity log-price series, $p_{d,h}$) can be built up as a sequence of projections onto one father wavelet (the smooth component or approximation) and a sequence of mother wavelets (or details): $p_{d,h} = S_J + D_J + D_{J-1} + \ldots + D_1$, where 2^J is the maximum scale sustainable by the number of observations (Percival and Walden, 2000; Weron, 2006). At the coarsest scale, the signal can be estimated by S_J. At a higher level of refinement, the signal can be approximated by $S_{J-1} = S_J + D_J$. At each step, by adding a mother wavelet D_j of a lower scale $j = J-1, J-2, \ldots$, we obtain a better estimate of the original signal. This procedure, known as *wavelet smoothing* or *lowpass filtering*, yields a traditional linear smoother. Basing on the results of Janczura et al. (2013) and Nowotarski et al. (2013), we use the Daubechies family of order 24 as they make a reasonable trade-off between how compactly they are localized in time and their smoothness. To provide a comprehensive analysis, we consider ten smoothing levels: $J = 5, \ldots, 14$, respectively $S_5, \ldots, S_{14}$ approximations. This corresponds to a range of smoothers, roughly from daily ($2^5 = 32$ h) to nearly biannual (2^{14} h or ca. 683 days).

3.2.3. The original SCAR algorithm and an innovation

The original SCAR modeling framework, as proposed by Nowotarski and Weron (2016), consists of the following four steps:

1. Decomposing the series of electricity log-prices $p_{d,h}$ from the calibration window into a trend-seasonal component $T_{d,h}$ and a stochastic component with short-term periodicities $q_{d,h}$, using

[3] Note, however, that the SCAR approach is not restricted to autoregressive structures. Neural network models can be used as well. In fact, as Marcjasz et al. (2018) have shown recently in a point forecasting context, the gains from using the Seasonal Component approach are even higher for NARX-type neural networks than for AR-type models, though achieved at a much higher computational cost.

one of the ten wavelet smoothers ($S_5, \ldots, S_{14}$) or one of the eight HP filters ($\lambda = 10^8, \ldots, 5 \cdot 10^{11}$). Then computing persistent forecasts of the LTSC independently for each of the 24 h of the next day, i.e., $\hat{T}_{d^*+1,h} \equiv T_{d^*,h}$, where d^* is the last day in the calibration window and $h = 1, \ldots, 24$.

2. Calibrating the **ARX** model defined by Eq. (1) to $q_{d,h}$ and computing forecasts for the 24 h of the next day, i.e., $\hat{q}_{d^*+1,h}$. Note, that unlike the seasonal decomposition in Step 1, which is made for the whole 360 × 24 = 8640 hour long calibration sample, here we split the data into 24 hourly series.
3. Adding forecasts of the **ARX** model computed in Step 2 to the persistent forecasts of the LTSC to yield log-price forecasts, i.e., $\hat{p}_{d^*+1,h}$.
4. Taking the exponent of the log-price forecasts computed in Step 3 to convert them into price forecasts of the **SCARX** model: $\hat{P}_{d^*+1,h} = \exp(\hat{p}_{d^*+1,h})$.

In this study, we add an additional step, say 1(b), in which the exogenous variable (the logarithm of the system load or consumption forecast) is deseasonalized using the same LTSC as prices prior to using it in Eq. (1). As it turns out, this innovation significantly improves the forecasting performance of the above algorithm (see Table 1 and compare with Table 1 in Nowotarski and Weron, 2016). Interestingly, it improves the **SCARX** models, but not the underlying **ARX** model. For instance, for the GEFCom2014 dataset and **ARX** models with system load deseasonalized using S_8, S_9 and S_{10} wavelet smoothers, the WMAE errors are respectively 11.254, 11.277 and 11.282 (vs. 11.232 for the **ARX** benchmark, see Table 1), while for the Nord Pool dataset and consumption prognosis deseasonalized using the same three wavelet smoothers the errors are respectively 8.528, 8.525 and 8.514 (vs. 8.500 for the **ARX** benchmark, see Table 1).

One may wonder if using the same LTSC for loads (or consumption) and prices is optimal or maybe a different LTSC should be used for loads and a different for prices. To address this issue, we have conducted an empirical study where each of the 19 (= 18 + 1; the '+1′ stands for 'no LTSC') price LTSCs was combined with each of the 19 possible load LTSCs to yield 361 $\text{LTSC}_{\text{price}}$–$\text{LTSC}_{\text{load}}$ pairs. For the GEFCom2014 dataset the best combination turns out to be S_{10}–HP_{1e8} with WMAE of 10.192. It is ca. 2.4% better than the best 'identical LTSC' pair, i.e., HP_{1e9}–HP_{1e9} with WMAE of 10.437, see Table 1 and the discussion in Section 4.1. Generally, better combinations are achieved for more fluctuating (or less stable) load than price LTSCs, e.g., HP_{1e9}–HP_{5e8} is better than HP_{1e9}–HP_{1e9}, while S_{10}–HP_{1e8} is better than S_{10}–HP_{5e8} and S_{10}–HP_{1e9}. For the Nord Pool dataset, the difference between the best combination, i.e., S_9–HP_{5e9} with WMAE of 8.101, and the best 'identical LTSC' pair, i.e., S_9–S_9 with WMAE of 8.147, is even smaller — only 0.6%. Interestingly, this time the best combinations are achieved for similarly fluctuating load and price LTSCs, but possibly from different classes, e.g., a wavelet price LTSC and a HP-filter load LTSC as in the S_9–HP_{5e9} pair. Although the results for the Nood Pool dataset justify our approach of using 'identical' LTSCs for loads and prices, the results are not that clear cut for the GEFCom2014 dataset. Nevertheless, the efficiency loss is not that substantial and for the sake of parsimony we recommend to use the same LTSC for prices and loads.

3.3. Probabilistic forecasting

With the introduction of smart grids and renewable integration requirements, probabilistic load and price forecasting has become more important to energy systems planning and operations (Hong and Fan, 2016). And probabilistic forecasting has a lot to offer, in particular, improved assessment of future uncertainty, ability to plan different strategies for the range of possible outcomes, increased effectiveness of submitted bids and possibility of more thorough forecast comparisons (Chatfield, 2000; Amjady and Hemmati, 2006; Nowotarski and Weron, 2018). Last but not least, probabilistic forecasting is a concept closely related to risk management. Namely, the most commonly used risk measure – the Value-at-Risk (VaR) – is a quantile risk metric, i.e., the $\alpha\%$ VaR is the α quantile of the profit and loss (P&L) distribution of a portfolio. Hence, to estimate the h-day ahead VaR, we need to find the α quantile of the h-day ahead P&L distribution (Alexander, 2008). In other words, VaR is nothing else but a quantile forecast, which lies at the heart of probabilistic forecasting. The more accurate are the probabilistic forecasts – especially for the extreme quantiles – the better is the VaR estimate.

3.3.1. Constructing prediction intervals

The most common extension from point to probabilistic forecasts is to construct *prediction intervals* (PIs). A number of methods can be used for this purpose, the most popular take into account both the point forecast and the corresponding error (Weron and Misiorek, 2008; Maciejowska et al., 2016): the center of the PI at the $(1-\alpha)$ confidence level is set equal to $\hat{P}_{d,h}$ and its bounds are defined by the $\frac{\alpha}{2}$th and $(1-\frac{\alpha}{2})$th quantiles of the distribution of $\varepsilon_{d,h}$. For

Table 1
Average WMAE (in percent) for all 103 weeks of the GEFCom2014 (*upper half*) or all 104 weeks of the Nord Pool (*lower half*) test period for point forecasts. WMAE errors for **SCARX** models smaller (better) than those for the **ARX** benchmark are underlined. Emphasized in bold are the results for the best performing model in each of the two parts of the table. Compare with Table 1 in Nowotarski and Weron (2016).

GEFCom2014									
Benchmarks & SCARX with HP filter (λ)									
Naive	ARX	10^8	$5 \cdot 10^8$	10^9	$5 \cdot 10^9$	10^{10}	$5 \cdot 10^{10}$	10^{11}	$5 \cdot 10^{11}$
14.716	11.232	10.519	10.447	**10.437**	10.495	10.559	10.798	10.897	11.060
SCARX with wavelet approximation									
S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}
12.917	12.226	11.106	10.849	10.732	10.776	10.843	10.824	11.096	11.072
Nord Pool									
Benchmarks & SCARX with HP filter (λ)									
Naive	ARX	10^8	$5 \cdot 10^8$	10^9	$5 \cdot 10^9$	10^{10}	$5 \cdot 10^{10}$	10^{11}	$5 \cdot 10^{11}$
9.661	8.500	8.475	8.512	8.536	8.601	8.621	8.655	8.663	8.670
SCARX with wavelet approximation									
S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}
9.834	9.761	8.411	8.205	**8.147**	8.169	8.319	8.351	8.484	8.389

instance, the 5% and 95% quantiles of the error term yield the 90% PI. We later denote such a PI of the spot price on day d and hour h by $\left[\hat{L}_{d,h},\hat{U}_{d,h}\right]$, where $\hat{L}_{d,h}$ and $\hat{U}_{d,h}$ are the lower and upper bounds, respectively. We skip the nominal rate $(1-\alpha)$ for simplicity. A forecaster may further extend her study and construct multiple PIs — the final outcome may be a set of quantiles on many levels. Here, like in the GEFCom2014 competition, we consider all 99 percentiles ($q = 1\%, 2\%, \ldots, 99\%$). This is a reasonably accurate approximation of the predictive distribution.

Moreover, we use three methods of constructing PIs (for more detailed descriptions, see Nowotarski and Weron, 2018) :

1. *Historical simulation*, which is a model-independent approach that consists of computing sample quantiles of the empirical distribution of $\varepsilon_{d,h}$ (Weron, 2014). Later in the text, we use superscript **H** to denote probabilistic forecasts obtained using historical simulation. EPF studies where PIs are obtained using this approach include Misiorek et al. (2006), Nowotarski and Weron (2015), Weron (2006) and Weron and Misiorek (2008) among others.
2. *Bootstrapping*, which first generates pseudo-prices recursively using sampled normalized residuals, then computes desired quantiles of the bootstrapped prices. The advantage of the bootstrap over historical simulation is that it takes into account not only historical forecast errors but also parameter uncertainty. The disadvantage is the significantly increased computational burden. Later in the text, we use superscript **B** to denote probabilistic forecasts obtained via the bootstrap. EPF studies where this approach is used to compute PIs include Alonso et al. (2011), Chen et al. (2012), Khosravi et al. (2013), Wan et al. (2014) and Ziel and Steinert (2016) among others.
3. *Quantile Regression Averaging* (QRA), proposed by Nowotarski and Weron (2015), which involves applying quantile regression to a pool of point forecasts of individual (i.e., not combined) forecasting models. As such, it directly works with the distribution of the electricity spot price, $\hat{F}_{P_{d,h}}$, without the need to split the probabilistic forecast into a point forecast and the distribution of the error term. Later in the text, we use superscript **Q** to denote probabilistic forecasts obtained with QRA. The very good forecasting performance of QRA has been verified by a number of authors (Gaillard et al., 2016; Maciejowska and Nowotarski, 2016; Maciejowska et al., 2016), not only in the area of EPF (Liu et al., 2017; Zhang et al., 2016). However, its most spectacular success came during the GEFCom2014 competition — the top two winning teams in the price track used variants of QRA (Gaillard et al., 2016; Maciejowska and Nowotarski, 2016).

Note, that all three approaches require that one-day ahead prediction errors, hence point predictions, are available in the calibration window for probabilistic forecasts. Hereby, a 182-day (or 26-week) rolling calibration window that directly follows the 360-day calibration window for point forecasts is used for computing quantiles of the error distribution (historical *prediction intervals*, PIs) or weights of the QRA approach. Consequently, the test periods start on the 543rd (= 360 + 182 + 1) day in each sample, i.e., June 26, 2012 (for GEFCom2014) and June 27, 2014 (for Nord Pool). On the other hand, the bootstrapped PIs are computed directly from point forecasts in a 360-day rolling window, see Figs. 1 and 2.

3.3.2. Averaging probabilistic forecasts

Given a set of n probabilistic forecasts we can combine them in one of two ways: by averaging either probabilities or quantiles (see Lichtendahl et al., 2013, for a discussion). If we denote by $\hat{F}_i(x)$ the ith distributional forecast and by $\hat{Q}_i(x) = \hat{F}_i^{-1}(x)$ the ith quantile forecast, then the average probability forecast $\textbf{F-Ave}_n^* \equiv \frac{1}{n}\sum_{i=1}^n \hat{F}_i(x)$ can be regarded as a vertical average of the corresponding predictive distributions while the average quantile forecast $\textbf{Q-Ave}_n^* \equiv \hat{Q}^{-1}(x)$ with $\hat{Q}(x) = \frac{1}{n}\sum_{i=1}^n \hat{Q}_i(x)$ as a horizontal average; the asterisk (*) denotes here one of three methods of constructing PIs (**H**, **B** or **Q**; see Section 3.3.1). This is illustrated in Fig. 3 for a sample day and hour and the Nord Pool dataset. Note that the average quantile forecast is always sharper, i.e., $\textbf{Q-Ave}_n^*$ has lower variance than $\textbf{F-Ave}_n^*$. While this feature is an advantage in many forecasting problems (Lichtendahl et al., 2013), in EPF it may not necessarily be so. To check this, in the empirical section we consider both averaging schemes.

4. Empirical results

We are now ready to present day-ahead forecasting results for the two considered datasets: GEFCom2014 and Nord Pool. For point forecasts, we use 2-year out-of-sample test periods, for probabilistic — slightly shorter, 1.5-year test periods (for the reasons discussed in Section 3.3.1). Recall from Section 2, that models are re-estimated on a daily basis. Price forecasts $\hat{P}_{d^*+1,1}, \ldots, \hat{P}_{d^*+1,24}$ for all 24 h of the next day are determined at the same point in time and the 360-day calibration window is rolled forward by one day: $d^* \rightarrow d^* + 1$.

4.1. Evaluation of point forecasts in terms of WMAE

Following Conejo et al. (2005), Weron and Misiorek (2008) and Nowotarski and Weron (2016), we compare the models in terms of the *Weekly-weighted Mean Absolute Error* (WMAE) loss function.

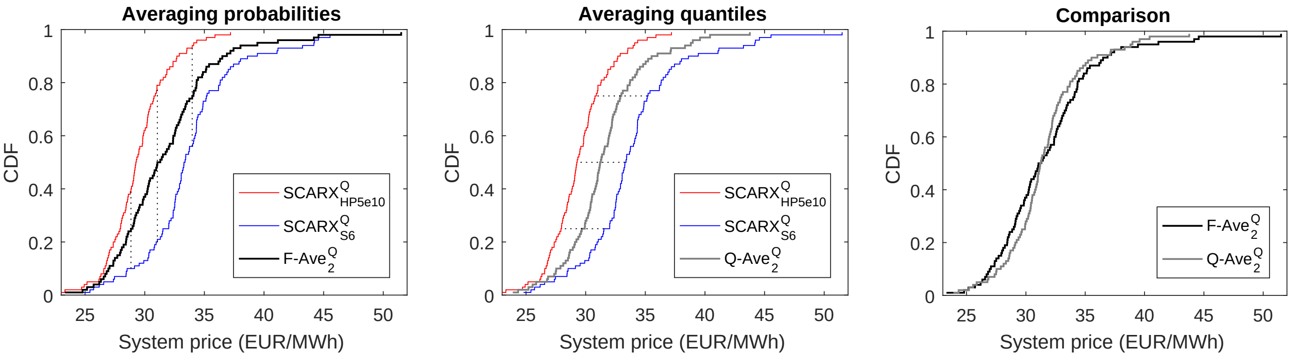


Fig. 3. Illustration of the averaging probabilities and averaging quantiles concepts for the $\textbf{SCARX}^{\textbf{Q}}_{\textbf{HP5e10}}$ and $\textbf{SCARX}^{\textbf{Q}}_{\textbf{S6}}$ predictive distributions for Nord Pool data on June 27th, 2014, hour 17–18. The average probability forecast $\textbf{F-Ave}^{\textbf{Q}}_{2}$ is a vertical average of the two CDFs (*left panel*), while the average quantile forecast $\textbf{Q-Ave}^{\textbf{Q}}_{2}$ is a horizontal average (*center panel*). The latter is always sharper, i.e., $\textbf{Q-Ave}^{\textbf{Q}}_{2}$ has lower variance than $\textbf{F-Ave}^{\textbf{Q}}_{2}$ (*right panel*).

WMAE is a robust measure similar to the Mean Absolute Percentage Error (MAPE) but with the absolute error normalized by the mean weekly price to avoid the adverse effect of negative and close to zero electricity spot prices. We evaluate the forecasting performance using weekly time intervals, each with 24 × 7 = 168 hourly observations. For each week, we calculate the WMAE for model *i* as:

$$\mathrm{WMAE}_i = \frac{1}{168 \cdot \bar{P}_{168}} \sum_{t=1}^{168} \left| P_t - \hat{P}_t^i \right|, \tag{3}$$

where $\bar{P}_{168}$ is the mean price for a given week, P_t $(= P_{24d+h} = P_{d,h})$ is the actual price at time t (i.e, for day d and hour h), not the log-price p_t, and $\hat{P}_t^i$ is the predicted price obtained from model i. Note, that WMAE requires the test period to be a multiple of a week (or 168 h). Hence, when computing WMAE we consider 103 full weeks (December 27, 2011–December 16, 2013) for the GEFCom2014 dataset and 104 full weeks (December 27, 2013–December 24, 2015) for the Nord Pool dataset.

In Table 1, we report the average WMAE (in percent) in the forecasting period for two benchmarks (**Naïve**, **ARX**) and two SCARX model classes — $\mathbf{SCARX_{HP^*}}$ with HP-filter-based LTSCs and $\mathbf{SCARX_{S^*}}$ with wavelet-based LTSCs. The asterisk (*) denotes here either λ (for the eight HP-filter LTSCs) or decomposition level (for the ten wavelet LTSCs). Note, that the **Naïve** and **ARX** benchmarks are identical to those in Nowotarski and Weron (2016), however, the WMAE errors for the former model were mistakenly given in the cited article as 20.475 for GEFCom2014 and 12.663 for Nord Pool.

Comparing Table 1 with Table 1 in Nowotarski and Weron (2016), we can clearly observe the improvement in forecasting accuracy from using deseasonalized log-loads in the **SCARX** models in this paper (see the discussion at the end of Section 3.2.3). Except for one case ($\mathbf{SCARX_{S13}}$ for Nord Pool), all WMAE errors in this paper are lower (better) than in Nowotarski and Weron (2016), sometimes by as much as 11%. The improvement is mostly seen for the more fluctuating LTSCs (HP1e8, 5e8, 1e9 and S_6, S_7, S_8), i.e., 6–11% for GEFCom2014 and 2.1–2.2% for Nord Pool, while for the more stable LTSCs the differences are much lower — only 0.7–2%. In particular, the best performing wavelet-based LTSC is now S_9 for both datasets. It is better than the best SCARX model in Nowotarski and Weron (2016) by ca. 3.4% for GEFCom2014 and ca. 2% for Nord Pool.

Interestingly, the performance of the HP-filter-based LTSCs has improved more than of the wavelet-based. To the extent that the best performing LTSC for GEFCom2014 data is now HP1e9 with an improvement of ca. 6% over S_{12} in Nowotarski and Weron (2016). The conjecture made in the cited study, that the variability at the edges of the sample is the reason for the disappointing performance of HP-filter-based LTSCs is not supported by the current results. When the exogenous variable (load forecast) is deseasonalized, the $\mathbf{SCARX}\ _{\mathbf{HP^*}}$ models excel in forecasting, despite their susceptibility to a volatile behavior at the edges.

4.2. Evaluation of probabilistic forecasts

As the primary focus of this paper is not on point predictions, let us now consider probabilistic forecasts. Like in the GEFCom2014 competition (Hong et al., 2016), we use the *pinball loss* function to measure the sharpness (i.e., concentration) of predictive distributions. Note, that although we formally do not compute density (or distributional) forecasts, we approximate them pretty well by a set of 99 quantile forecasts spanned on a grid of 99 percentiles of the predictive distribution.

We consider a battery of probabilistic models built on 20 point forecasting structures (see Table 1). Except for the **Naïve** benchmark, all point forecasting models lead to three types of probabilistic predictions — with PIs obtained via historical simulation ($\mathbf{model^H}$), bootstrapping ($\mathbf{model^B}$) and QRA ($\mathbf{model^Q}$). In total, we consider 59 probabilistic models:

- Two based on the **Naïve** benchmark → $\mathbf{Na\ddot{i}ve^H}$ and $\mathbf{Na\ddot{i}ve^Q}$ (note, that the bootstrap procedure relies on randomness, while the **Naïve** method is purely deterministic),
- Three based on the **ARX** benchmark → $\mathbf{ARX^H}$, $\mathbf{ARX^B}$ and $\mathbf{ARX^Q}$,
- 24 based on the $\mathbf{SCARX_{HP}}$ family with a HP-filter LTSC → $\mathbf{SCARX^*_{H\,P?}}$, where the question mark denotes $\lambda = 10^8$,$5 \cdot 10^8$,...,$5 \cdot 10^{11}$(i.e., ? = 1e8, 5e8, ..., 5e11) and the asterisk one of three methods of constructing PIs (i.e., $* =$ **H**, **B**, **Q**),
- 30 based on the $\mathbf{SCARX_{S?}}$ family with a wavelet LTSC → $\mathbf{SCARX}\,^*_{\mathbf{S?}}$, where the question mark denotes the decomposition level S_5, S_6, ..., S_{14} (i.e., ? $= 5, 6, \ldots, 14$) and the asterisk one of three methods of constructing PIs (i.e., $* =$ **H**, **B**, **Q**).

Furthermore, following Lichtendahl et al. (2013), we consider two ways of combining probabilistic forecasts: by averaging either probabilities (denoted by $\mathbf{F\text{-}Ave}^*_n$) or quantiles (denoted by $\mathbf{Q\text{-}Ave}^*_n$), where n is the number of averaged forecasts and $* =$ **H**, **B**, **Q**. Since the number of all possible subsets of the listed above 59 probabilistic models is too large, we use only one selection scheme. Namely, we rank all point forecasting models in terms of WMAE in the 'probabilistic' calibration window (to be precise: the 26-week window preceding day $d^* + 1$; naturally, we repeat this procedure for every day in the test period), then combine the probabilistic forecasts of the best $n = 1, 2, \ldots, 19$ models, but only within each **H**, **B** or **Q** family. For instance, $\mathbf{F\text{-}Ave^Q_5}$ is the average probability forecast over the best five models with PIs obtained via QRA, while $\mathbf{Q\text{-}Ave^B_1}$ is simply the best performing model in terms of WMAE with PIs obtained via bootstrapping. Note, that we use the WMAE in the calibration window and not the pinball loss to rank the forecasts, because the latter is not known at the time the forecast is made.

4.2.1. Sharpness and the pinball loss

Sharpness is a measure of concentration of the predictive distribution — the more concentrated the distribution the better. Sharpness can be evaluated using so-called *proper* scoring rules, for instance, the pinball loss (Gneiting, 2011; Hong et al., 2016):

$$\mathrm{Pinball}\left(\hat{Q}_{P_t}(q), P_t, q\right) = \begin{cases} (1-q)\left(\hat{Q}_{P_t}(q) - P_t\right), & \text{for } P_t < \hat{Q}_{P_t}(q), \\ q\left(P_t - \hat{Q}_{P_t}(q)\right), & \text{for } P_t \geq \hat{Q}_{P_t}(q), \end{cases} \tag{4}$$

where $\hat{Q}_{P_t}(q)$ is the price forecast at the q-th quantile and P_t is the actually observed price. Naturally, a lower score indicates a better probabilistic forecast. The pinball loss defined by Eq. (4) is a measure of fit for one quantile only. However, to provide an aggregate score it can be averaged (i) across a certain time period, e.g., all hours in the test period (as we do here), and (ii) across different quantiles, e.g., all 99 percentiles (as in the GEFCom2014 competition and here) or only the upper ten and lower ten percentiles (as we do here).

In Table 2, we summarize the sharpness of all 59 probabilistic models across all 99 percentiles. We can observe that:

- The S_9 clearly stands out as the best performing LTSC. For both datasets, the $\mathbf{SCARX^*_{S9}}$ models outperform all other within each group (**H**, **B** and **Q**).
- The QRA implied predictions (**Q**) nearly always outperform the two other techniques (**H** and **B**). The only two exceptions are $\mathbf{SCARX^Q_{S13}}$ and $\mathbf{SCARX^Q_{S14}}$ for the Nord Pool dataset, which are just barely outperformed by the respective bootstrapped forecasts.
- Despite the much more sophisticated and computationally intensive algorithm, the bootstrapped forecasts are not much

Table 2
The pinball loss defined by Eq. (4) averaged across all 99 percentiles and all hours in the 1.5-year 'probabilistic' test period: 77 weeks for GEFCom2014 (*upper half*) or 78 weeks for Nord Pool (*lower half*). Except for the **Naïve** benchmark, all models are in three variants — with probabilistic forecasts obtained via historical simulation (**model$^{\mathbf{H}}$**), bootstrapping (**model$^{\mathbf{B}}$**) and QRA (**model$^{\mathbf{Q}}$**). Scores smaller (better) than those for the **ARX$^{\mathbf{H}}$** benchmark are underlined. Emphasized in bold are the results for the best performing model in each of the two parts of the table.

GEFCom2014										
Benchmarks & SCARX with HP filter (λ)										
	Naive	ARX	10^8	$5\cdot10^8$	10^9	$5\cdot10^9$	10^{10}	$5\cdot10^{10}$	10^{11}	$5\cdot10^{11}$
Historical (H)	3.269	2.472	2.421	2.389	2.380	2.386	2.400	2.445	2.459	2.469
Bootstrap (B)	–	2.468	2.427	2.390	2.380	2.385	2.402	2.447	2.462	2.475
QRA (Q)	3.189	2.431	2.340	2.322	2.321	2.340	2.360	2.425	2.450	2.467
SCARX with wavelet approximation										
	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}
Historical (H)	2.772	2.694	2.487	2.410	2.371	2.381	2.390	2.388	2.456	2.440
Bootstrap (B)	2.998	2.810	2.526	2.425	2.368	2.380	2.386	2.372	2.496	2.482
QRA (Q)	2.702	2.615	2.420	2.349	**2.313**	2.333	2.366	2.343	2.450	2.418
Nord Pool										
Benchmarks & SCARX with HP filter (λ)										
	Naive	ARX	10^8	$5\cdot10^8$	10^9	$5\cdot10^9$	10^{10}	$5\cdot10^{10}$	10^{11}	$5\cdot10^{11}$
Historical (H)	0.884	0.744	0.739	0.745	0.748	0.758	0.762	0.768	0.771	0.778
Bootstrap (B)	–	0.742	0.738	0.743	0.747	0.756	0.759	0.764	0.766	0.766
QRA (Q)	0.859	0.739	0.721	0.721	0.723	0.732	0.735	0.735	0.735	0.731
SCARX with wavelet approximation										
	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}
Historical (H)	0.842	0.835	0.731	0.713	0.712	0.718	0.735	0.741	0.744	0.734
Bootstrap (B)	0.874	0.852	0.731	0.709	0.708	0.714	0.734	0.733	0.736	0.727
QRA (Q)	0.832	0.826	0.723	0.700	**0.696**	0.703	0.714	0.720	0.740	0.731

better (if better at all) than the much simpler historical simulation predictions. For the GEFCom2014 dataset they are actually more often outperformed by the latter than not. However, for both datasets the best bootstrapped model (**SCARX$_{\mathbf{S9}}^{\mathbf{B}}$**) outperforms the best historical simulation model (**SCARX$_{\mathbf{S9}}^{\mathbf{H}}$**).

- The point and probabilistic forecasts are not always consistent. In particular, for the GEFCom2014 dataset the wavelet-based LTSCs lead to generally better probabilistic but worse point predictions (for the Nord Pool dataset the wavelet-based LTSCs are better in both cases).
- Finally, the **SCARX** models – especially QRA-based – nearly always significantly (as we will see in Section 4.2.2) outperform the benchmarks (**Naïve** and **ARX**), which justifies the SCAR modeling framework in EPF.

Table 3
The pinball loss defined by Eq. (4) averaged across 20 extreme percentiles (1, …, 10 and 90, …, 99, i.e., corresponding to confidence levels typically considered in risk management) and all hours in the 1.5-year 'probabilistic' test period: 77 weeks for GEFCom2014 (*upper half*) or 78 weeks for Nord Pool (*lower half*). Except for the **Naïve** benchmark, all models are in three variants — with probabilistic forecasts obtained via historical simulation (**model$^{\mathbf{H}}$**), bootstrapping (**model$^{\mathbf{B}}$**) and QRA (**model$^{\mathbf{Q}}$**). Scores smaller (better) than those for the **ARX$^{\mathbf{H}}$** benchmark are underlined. Emphasized in bold are the results for the best performing model in each of the two parts of the table.

GEFCom2014										
Benchmarks & SCARX with HP filter (λ)										
	Naive	ARX	10^8	$5\cdot10^8$	10^9	$5\cdot10^9$	10^{10}	$5\cdot10^{10}$	10^{11}	$5\cdot10^{11}$
Historical (H)	1.690	1.211	1.202	1.178	1.173	1.178	1.183	1.188	1.185	1.173
Bootstrap (B)	–	1.218	1.278	1.236	1.221	1.221	1.225	1.223	1.228	1.218
QRA (Q)	1.610	1.089	1.026	**1.018**	1.018	1.027	1.035	1.052	1.058	1.068
SCARX with wavelet approximation										
	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}
Historical (H)	1.310	1.299	1.241	1.165	1.152	1.152	1.139	1.153	1.183	1.187
Bootstrap (B)	1.872	1.618	1.379	1.243	1.185	1.181	1.152	1.155	1.225	1.215
QRA (Q)	1.173	1.139	1.073	1.020	1.026	1.025	1.033	1.046	1.078	1.075
Nord Pool										
Benchmarks & SCARX with HP filter (λ)										
	Naive	ARX	10^8	$5\cdot10^8$	10^9	$5\cdot10^9$	10^{10}	$5\cdot10^{10}$	10^{11}	$5\cdot10^{11}$
Historical (H)	0.483	0.380	0.379	0.375	0.374	0.373	0.373	0.375	0.377	0.383
Bootstrap (B)	–	0.379	0.387	0.382	0.382	0.379	0.379	0.379	0.380	0.383
QRA (Q)	0.418	0.356	0.355	0.350	0.350	**0.349**	0.349	0.350	0.352	0.357
SCARX with wavelet approximation										
	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}
Historical (H)	0.445	0.433	0.381	0.376	0.372	0.371	0.375	0.376	0.374	0.374
Bootstrap (B)	0.521	0.480	0.393	0.376	0.369	0.369	0.375	0.371	0.370	0.368
QRA (Q)	0.409	0.402	0.365	0.355	0.351	0.350	0.353	0.352	0.352	0.349

In a risk management context, which concentrates on the tail behavior of the P&L distribution, we would probably wish to consider only those quantiles that corresponded to confidence levels used in the financial sector, i.e., definitely above 90% and typically in the range of 95–99 % (Alexander, 2008). Hence, let us now look at Table 3 and focus on the extreme quantiles only, i.e., on the lower ten and the upper ten percentiles. We note that some of the observed above features are even more pronounced than in Table 2. For instance, the QRA-based predictions (**Q**) now significantly outperform the other two techniques (**H** and **B**) across nearly all LTSCs. In particular, if we exclude the two most volatile wavelet-based LTSCs (i.e., S_5 and S_6), then all $\mathbf{model^Q}$ pinball scores are smaller (better) than all $\mathbf{model^H}$ and $\mathbf{model^B}$ scores. On the other hand, the historical simulation and bootstrapped distributional forecasts perform very much alike, with the former being now consistently better for GEFCom2014 — the best $\mathbf{model^H}$ outperforms the best $\mathbf{model^B}$ (both are based on the S_{11} LTSC).

In Fig. 4, we plot the pinball loss defined by Eq. (4) for selected models and forecast combinations, averaged either across all 99 percentiles or 20 extreme percentiles. Three types of average quantile forecasts ($\textbf{Q-Ave}^*_n$, with $* = \mathbf{H}$, $\mathbf{B}$ or $\mathbf{Q}$) and three types of average probability forecasts ($\textbf{F-Ave}^*_n$, with $* = \mathbf{H}$, $\mathbf{B}$ or $\mathbf{Q}$) for $n = 1$ to 19 combined forecasts are compared against three **ARX** benchmarks and the best performing (*ex-post*) in terms of the pinball loss **SCARX** model. Note, that the latter is in general different than the 'average' quantile or probability forecast for $n = 1$, since it is selected *ex-post* based the average pinball loss in the test period, while $\textbf{F-Ave}^*_1 = \textbf{Q-Ave}^*_1$ is selected *ex-ante* as the best performing model in terms of WMAE in the calibration period.

While for $n = 1$ the best performing **SCARX** models outperform the average quantile or probability forecasts (which is clear), for $n > 2, 3$ the situation changes. The average probability forecasts – and in most cases also the average quantile forecasts – for QRA-based predictive distributions significantly outperform the *ex-post* selected best performing individual, i.e., not combined, model. But there is more to be noted:

- While for individual models the historical simulation and bootstrap-based distributional forecasts perform very much alike, for combined forecasts $\textbf{F/Q-Ave}^{\mathbf{B}}_n$ tend to outperform $\textbf{F/Q-Ave}^{\mathbf{H}}_n$ (except for GEFCom2014 and the extreme percentiles). This could justify undertaking the substantially higher computational burden related to the bootstrap procedure.
- In contrast to Lichtendahl et al. (2013), we observe that average probability forecasts yield generally better predictions (in our study: smaller pinball scores) than average quantile forecasts. Recall from Section 3.3.2, that the latter are always sharper, i.e., $\textbf{Q-Ave}^*_n$ has lower variance than $\textbf{F-Ave}^*_n$ (see Fig. 3 for an illustration). It seems that in EPF this feature is a disadvantage, possibly due to the extremely spiky price distributions.
- It is hard to decide *ex-ante* how many forecasts should be combined. However, we may suggest $n = 7$ as a 'rule of thumb'.
- We do not see qualitative differences when comparing pinball scores across 20 extreme percentiles with scores for all 99 percentiles. Perhaps, only the differences between QRA-based and historical simulation and bootstrap-based predictive distributions are more pronounced. But to formally check this (and other observations made earlier), we resort to the Diebold and Mariano (1995) test in Section 4.2.2.

4.2.2. Diebold-Mariano (DM) tests

The WMAE values analyzed in Section 4.1 or pinball scores studied in Section 4.2.1 can be used to provide a ranking of models,

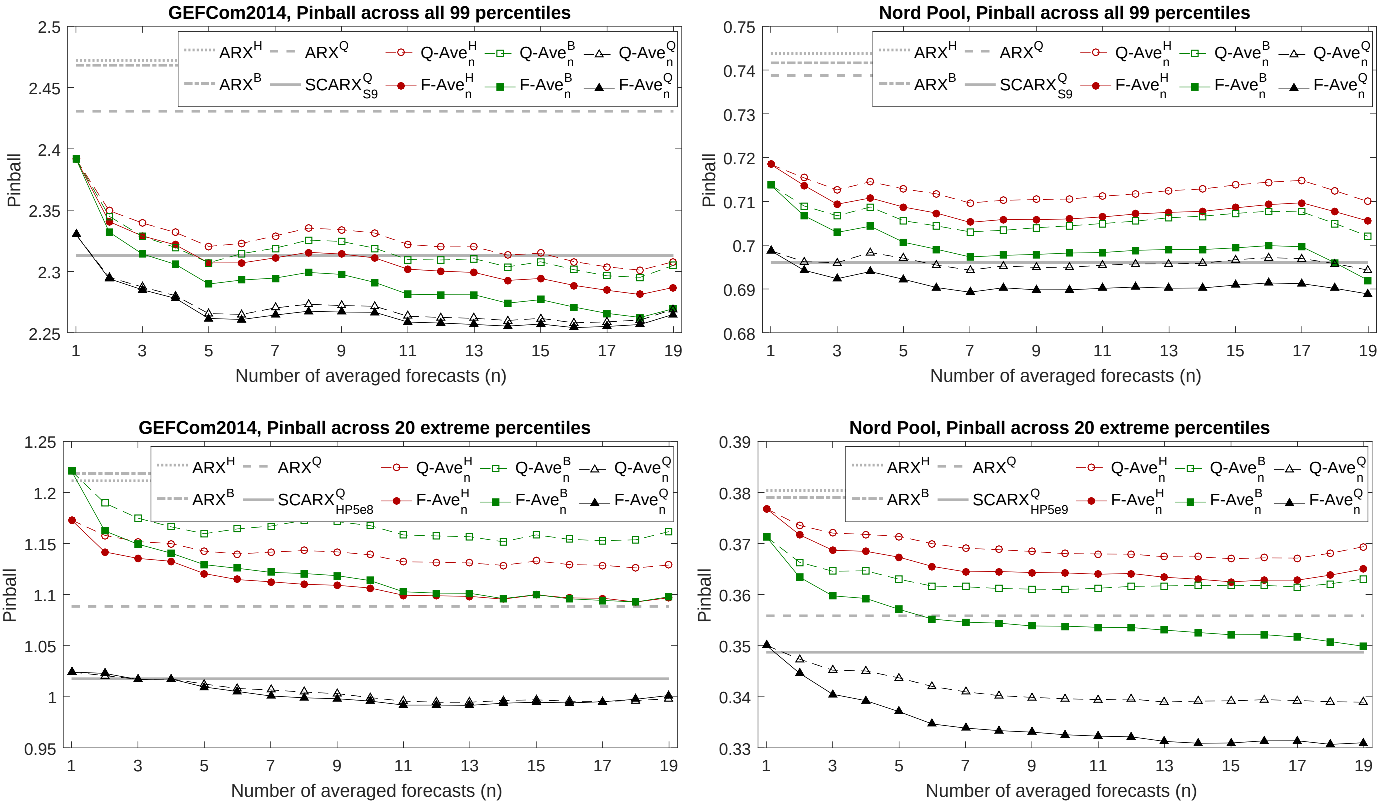


Fig. 4. The pinball loss defined by Eq. (4) for selected models and combinations, averaged across all 99 percentiles (*top panels*) or 20 extreme percentiles (1, …, 10 and 90, …, 99; *bottom panels*) and all hours in the 1.5-year 'probabilistic' test period — 77 weeks for GEFCom2014 (*left panels*) or 78 weeks for Nord Pool (*right panels*). Three types of average quantile forecasts ($\textbf{Q-Ave}^*_n$) and three types of average probability forecasts ($\textbf{F-Ave}^*_n$) for $n = 1, 2, \ldots, 19$ combined forecasts are compared against three **ARX** benchmarks and the best performing in terms of the pinball loss **SCARX** model.

but not statistically significant conclusions on the outperformance of the forecasts of one model by those of another. In this section, we compute the Diebold and Mariano (1995) test (abbreviated 'DM test'), which takes into account the correlation structure of prediction errors and performs a pairwise comparison.

In the EPF literature, the DM test is usually conducted separately for each of the load periods of the day (Bordignon et al., 2013; Nowotarski and Weron, 2016; Weron, 2014). However, Ziel and Weron (2018) recently introduced a different approach, where only one statistic for each pair of models is computed based on the 24-dimensional vector of errors (or scores) for each day, and called it the *multivariate* or *vectorized* DM test. If we denote by $\pi_{X,d} = (\pi_{X,d,1}, \ldots, \pi_{X,d,24})'$ and $\pi_{Y,d} = (\pi_{Y,d,1}, \ldots, \pi_{Y,d,24})'$ the vectors of pinball scores for day d of models X and Y, respectively, then the multivariate loss differential series in the $\|\cdot\|_1$-norm is given by:

$$\Delta_{X,Y,d} = \|\pi_{X,d}\|_1 - \|\pi_{Y,d}\|_1, \qquad (5)$$

where $\|\pi_{X,d}\|_1 = \sum_{h=1}^{24} |\pi_{X,d,h}|$. For each model pair and each dataset, we compute the p-value of two one-sided DM tests: (i) a test with the null hypothesis $H_0 : E(\Delta_{X,Y,d}) \le 0$, i.e., the outperformance of the probabilistic forecasts of Y by those of X, and (ii) the complementary test with the reverse null $H_0^R : E(\Delta_{X,Y,d}) \ge 0$, i.e., the outperformance of the probabilistic forecasts of X by those of Y. As in the standard DM test, we assume that the loss differential series is covariance stationary.

In Fig. 5, we plot the results for the multivariate DM-test for 14 selected models, both datasets and separately across all 99 and across 20 extreme percentiles. The models include both **Naïve** benchmarks, all three **ARX** benchmarks, the best *ex-post* **SCARX$^{\mathbf{H}}_{*}$**, **SCARX$^{\mathbf{B}}_{*}$** and **SCARX$^{\mathbf{Q}}_{*}$** models, the best *ex-post* **Q-Ave$^{\mathbf{H}}_{n}$**, **Q-Ave$^{\mathbf{B}}_{n}$** and **Q-Ave$^{\mathbf{Q}}_{n}$** average quantile forecasts, and the best *ex-post* **F-Ave$^{\mathbf{H}}_{n}$**, **F-Ave$^{\mathbf{B}}_{n}$** and **F-Ave$^{\mathbf{Q}}_{n}$** average probability forecasts.

In all four panels, we see the corresponding p-values of the conducted pairwise comparisons: green and yellow squares indicate statistical significance at the 5% level (with the darkest green corresponding to close to zero p-values), red squares indicate weak significance with a p-value between 5% and 10%, while black denote no significance (i.e., a p-value of 10% or more). For instance, we see in the bottom right panel that the first row is dark green, so that

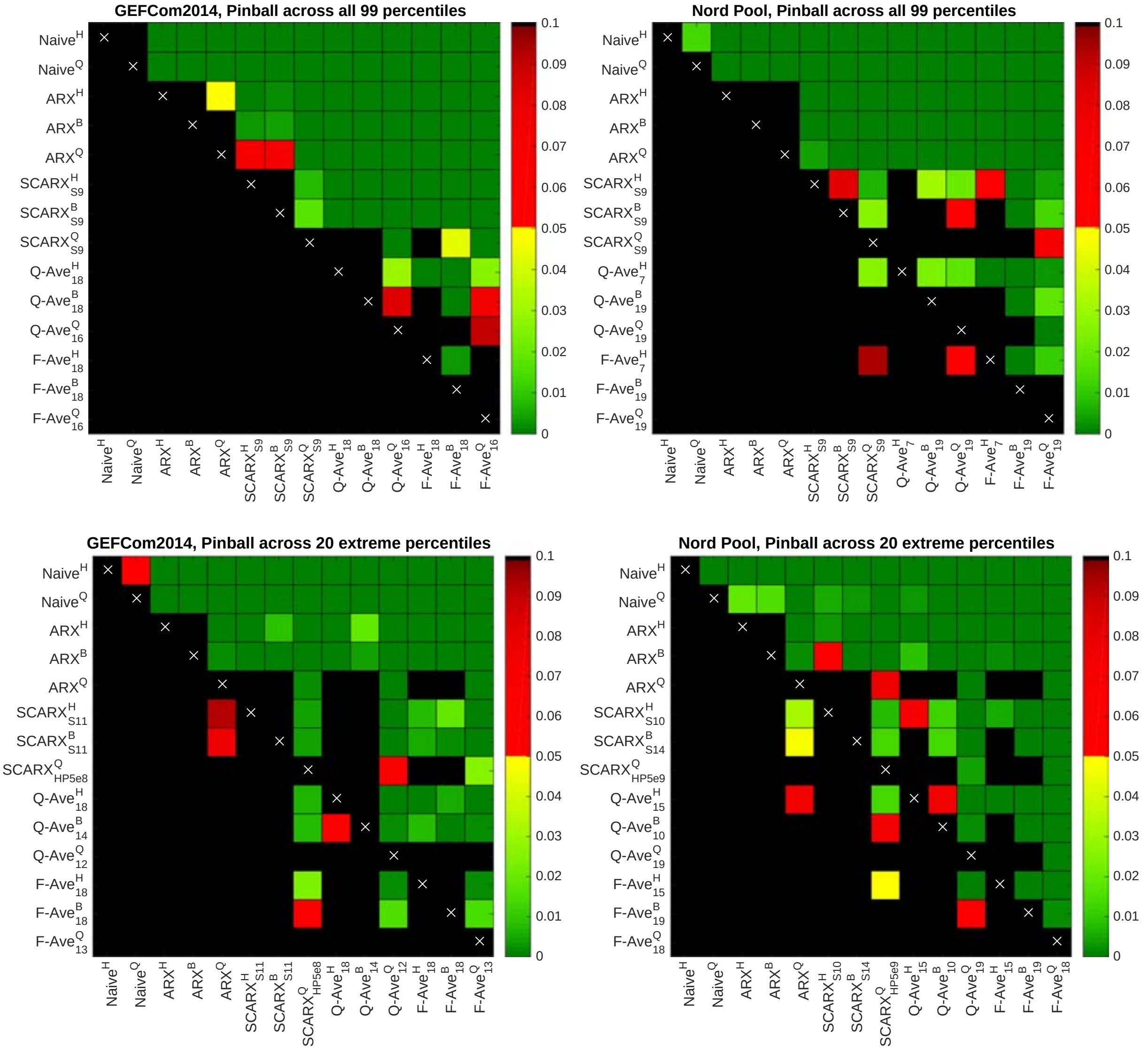


Fig. 5. Results of the multivariate DM test defined by the multivariate loss differential series in Eq. (5) for 14 selected models across all 99 percentiles (*top panels*) and only 20 extreme percentiles (*bottom panels*), for the GEFCom2014 (*left panels*) and Nord Pool (*right panels*) datasets. We use a heat map to indicate the range of the p-values — the closer they are to zero (→ dark green), 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). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the forecasts of every model significantly outperform those of the **Naïve$^{\mathbf{H}}$** benchmark. In the same panel we see that the column which corresponds to **F-Ave$^{\mathbf{Q}}_{18}$** is dark green, meaning that this combination leads to significantly better forecasts than all other models. As can be seen in all four panels, the model classes are ordered from the worst to the best performing (on average). Within each class, models with QRA-based distributional forecasts typically significantly outperform the **H** and **B** type forecasts. Overall, the best models are **F-Ave$^{\mathbf{Q}}_{n}$** and **Q-Ave$^{\mathbf{Q}}_{n}$** as they significantly outperform nearly all competitors. On the other hand, all benchmarks except **ARX$^{\mathbf{Q}}$** are always significantly outperformed by the **SCARX** and combined models. The latter clearly demonstrates the usefulness of the SCAR concept, and forecast averaging in particular.

5. Conclusions

Most day-ahead electricity price forecasting (EPF) studies ignore the long-term seasonal component (LTSC). However, as Nowotarski and Weron (2016) have shown for point forecasts, the *Seasonal Component AutoRegressive* (SCAR) modeling concept can bring significant accuracy gains compared to models fitted to non-deseasonalized prices. The main aim of this paper was to validate – in the context of probabilistic forecasts – the conjecture made by Nowotarski and Weron (2016) that combining forecasts of SCAR-type models with different LTSCs should further improve the predictive accuracy.

To this end, we have considered a collection of 20 point forecasting models: the **Naïve** benchmark, an autoregressive expert model dubbed **ARX** and 18 **SCARX** models (i.e., SCAR-type models built on Hodrick-Prescott filter or wavelet-based LTSCs with an exogenous variable). Then by applying one of three schemes for computing probabilistic forecasts (historical simulation, bootstrapping and QRA), we have obtained a battery of 59 individual (i.e., non-combined) ‘probabilistic’ models and using one of two approaches to combining probabilistic forecasts (averaging probabilities or averaging quantiles) — a pool of averaged predictions. The point forecasts have been then compared using the robust Weekly-weighted Mean Absolute Error (WMAE) and the probabilistic forecasts using the pinball loss function. The significance of differences in predictive accuracy (in terms of the pinball loss) has been tested using the Diebold and Mariano (1995) test. We should note at this point, however, that the study is based on two datasets which may not be representative of all electricity markets in the world. Hence, there is a possibility that some of our conclusions will not hold for some markets. Testing this is left for future research.

The results of our comprehensive EPF study have been discussed in detail in Section 4. Here, let us only briefly recap the most important contributions and findings:

- In the original SCAR modeling framework, as proposed by Nowotarski and Weron (2016), only the prices were deseasonalized. We find that deseasonalizing the exogenous variable (here: the system load or consumption forecast) using the same LTSC as used to deseasonalize prices significantly enhances the forecasting performance of the **SCARX** models. Interestingly, this innovation does not improve the efficiency of the underlying **ARX** model.
- We introduce a novel extension of the SCAR approach to probabilistic forecasts by applying one of three schemes – historical simulation, bootstrapping or QRA – to the residuals (i.e., day-ahead prediction errors) of the **SCARX** models, and thus obtain ‘probabilistic’ **SCARX** models.
- We find that ‘probabilistic’ **SCARX** models – especially QRA-based – nearly always significantly outperform the benchmarks (**Naïve** and **ARX**), which further justifies the SCAR approach in EPF.
- We observe that QRA-based ‘probabilistic’ **SCARX** models nearly always significantly outperform historical simulation and bootstrap-based models.
- To our best knowledge, we are the first to apply two alternative averaging schemes – the average probability forecast (which is a vertical average of the predictive distribution functions) and the average quantile forecast (which is a horizontal average) – to probabilistic EPFs.
- We observe that both averaging schemes generally significantly outperform the benchmarks and the individual (i.e., non-combined) **SCARX** models.
- Finally, we find that averaging over probabilities generally yields better probabilistic EPFs than averaging over quantiles. This is in contrast to typically encountered economic forecasting problems (Lichtendahl et al., 2013).

Acknowledgments

This work was supported in part by the National Science Center (NCN, Poland) through grant 2015/17/B/HS4/00334 and in part by the Ministry of Science and Higher Education (MNiSW), Poland, Core Funding for Statutory Research and Development Activities.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eneco.2018.02.007.

References

Alexander, C., 2008. Market Risk Analysis IV: Value at Risk Models. Wiley.

Alonso, A., Garcia-Martos, C., Rodriguez, J., Sanchez, M., 2011. Seasonal dynamic factor analysis and bootstrap inference: application to electricity market forecasting. Technometrics 53, 137–151.

Amjady, N., Hemmati, M., 2006. Energy price forecasting. IEEE Power Energ. Mag. 20–29. March/April.

Bordignon, S., Bunn, D.W., Lisi, F., Nan, F., 2013. Combining day-ahead forecasts for British electricity prices. Energy Econ. 35, 88–103.

Bunn, D., Andresen, A., Chen, D., Westgaard, S., 2016. Analysis and forecasting of electricity price risks with quantile factor models. Energy J. 37 (1), 101–122.

Chatfield, C., 2000. Time-series forecasting. Chapman & Hall/CRC, Boca Raton, Florida.

Chen, X., Dong, Z., Meng, K., Xu, Y., Wong, K., Ngan, H., 2012. Electricity price forecasting with extreme learning machine and bootstrapping. IEEE Trans. Power Syst. 27 (4), 2055–2062.

Conejo, A.J., Contreras, J., Espínola, R., Plazas, M.A., 2005. Forecasting electricity prices for a day-ahead pool-based electric energy market. Int. J. Forecast. 21, 435–462.

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

Gaillard, P., Goude, Y., Nedellec, R., 2016. Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. Int. J. Forecast. 32 (3), 1038–1050.

Gneiting, T., 2011. Quantiles as optimal point forecasts. Int. J. Forecast. 27 (2), 197–207.

Hodrick, R.J., Prescott, E.C., 1997. Postwar U.S. business cycles: an empirical investigation. J. Money Credit Bank. 29 (1), 1–16.

Hong, T., Fan, S., 2016. Probabilistic electric load forecasting: a tutorial review. Int. J. Forecast. 32, 914–938.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., Hyndman, R.J., 2016. Probabilistic energy forecasting: global energy forecasting competition 2014 and beyond. Int. J. Forecast. 32 (3), 896–913.

Janczura, J., Trück, S., Weron, R., Wolff, R., 2013. Identifying spikes and seasonal components in electricity spot price data: a guide to robust modeling. Energy Econ. 38, 96–110.

Khosravi, A., Nahavandi, S., Creighton, D., 2013. Quantifying uncertainties of neural network-based electricity price forecasts. Appl. Energy 112, 120–129.

Kristiansen, T., 2012. Forecasting Nord Pool day-ahead prices with an autoregressive model. Energy Policy 49, 328–332.

Lichtendahl, K.C., Grushka-Cockayne, Y., Winkler, R.L., 2013. Is it better to average probabilities or quantiles? Manag. Sci. 59 (7), 1594–1611.

Lisi, F., Nan, F., 2014. Component estimation for electricity prices: procedures and comparisons. Energy Econ. 44, 143–159.

Liu, B., Nowotarski, J., Hong, T., Weron, R., 2017. Probabilistic load forecasting via Quantile Regression Averaging on sister forecasts. IEEE Trans. Smart Grid 8, 730–737.

Maciejowska, K., Nowotarski, J., 2016. A hybrid model for GEFCom2014 probabilistic electricity price forecasting. Int. J. Forecast. 32 (3), 1051–1056.

Maciejowska, K., Nowotarski, J., Weron, R., 2016. Probabilistic forecasting of electricity spot prices using Factor Quantile Regression Averaging. Int. J. Forecast. 32 (3), 957–965.

Marcjasz, G., Uniejewski, B., Weron, R., 2018. On the importance of the long-term seasonal component in day-ahead electricity price forecasting with NARX neural networks. Int. J. Forecast. https://doi.org/10.1016/j.ijforecast.2017.11.009.

Misiorek, A., Trück, S., Weron, R., 2006. Point and interval forecasting of spot electricity prices: linear vs. non-linear time series models. Stud. Nonlinear Dyn. Econ. 10 (3).2.

Nogales, F.J., Contreras, J., Conejo, A.J., Espinola, R., 2002. Forecasting next-day electricity prices by time series models. IEEE Trans. Power Syst. 17, 342–348.

Nowotarski, J., Liu, B., Weron, R., Hong, T., 2016. Improving short term load forecast accuracy via combining sister forecasts. Energy 98, 40–49.

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.

Nowotarski, J., Tomczyk, J., Weron, R., 2013. Robust estimation and forecasting of the long-term seasonal component of electricity spot prices. Energy Econ. 39, 13–27.

Nowotarski, J., Weron, R., 2015. Computing electricity spot price prediction intervals using quantile regression and forecast averaging. Comput. Stat. 30 (3), 791–803.

Nowotarski, J., Weron, R., 2016. On the importance of the long-term seasonal component in day-ahead electricity price forecasting. Energy Econ. 57, 228–235.

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

Percival, D.B., Walden, A.T., 2000. Wavelet Methods for Time Series Analysis. Cambridge University Press.

Serinaldi, F., 2011. Distributional modeling and short-term forecasting of electricity prices by Generalized Additive Models for location, scale and shape. Energy Econ. 33, 1216–1226.

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

Wan, C., Xu, Z., Wang, Y., Dong, Z., Wong, K., 2014. A hybrid approach for probabilistic forecasting of electricity price. IEEE Trans. Smart Grid 5 (1), 463–470.

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

Weron, R., 2014. Electricity price forecasting: a review of the state-of-the-art with a look into the future. Int. J. Forecast. 30 (4), 1030–1081.

Weron, R., Misiorek, A., 2008. Forecasting spot electricity prices: a comparison of parametric and semiparametric time series models. Int. J. Forecast. 24, 744–763.

Weron, R., Zator, M., 2015. A note on using the Hodrick-Prescott filter in electricity markets. Energy Econ. 48, 1–6.

Zhang, Y., Liu, K., Qin, L., An, X., 2016. Deterministic and probabilistic interval prediction for short-term wind power generation based on variational mode decomposition and machine learning methods. Energy Convers. Manag. 112, 208–219.

Ziel, F., 2016. Forecasting electricity spot prices using LASSO: on capturing the autoregressive intraday structure. IEEE Trans. Power Syst. 31 (6), 4977–4987.

Ziel, F., Steinert, R., 2016. Electricity price forecasting using sale and purchase curves: the X-model. Energy Econ. 59, 435–454.

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 4

Regularized quantile regression averaging for probabilistic electricity price forecasting

Bartosz Uniejewski, Rafał Weron

Energy Economics 95 (2021) 105121

Contents lists available at ScienceDirect

Energy Economics

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

Regularized quantile regression averaging for probabilistic electricity price forecasting

Bartosz Uniejewski, Rafał Weron *

Department of Operations Research and Business Intelligence, Wrocław University of Science and Technology, 50-370 Wrocław, Poland

ARTICLE INFO

Article history:
Received 18 November 2019
Received in revised form 1 September 2020
Accepted 9 January 2021
Available online 16 January 2021

Keywords:
Electricity price forecasting
Probabilistic forecasting
Risk management
Quantile Regression Averaging (QRA)
LASSO
Bayesian Information Criterion (BIC)
Cross-validation
Kupiec test
Pinball score
Conditional predictive accuracy
Trading strategy
Financial profits

ABSTRACT

Quantile Regression Averaging (QRA) has sparked interest in the electricity price forecasting community after its unprecedented success in the Global Energy Forecasting Competition 2014, where the top two winning teams in the price track used variants of QRA. However, recent studies have reported the method's vulnerability to low quality predictors when the set of regressors is larger than just a few. To address this issue, we consider a regularized variant of QRA, which utilizes the Least Absolute Shrinkage and Selection Operator (LASSO) to automatically select the relevant regressors. We evaluate the introduced technique – dubbed LASSO QRA or LQRA for short – using datasets from the Polish and Nordic power markets. By comparing against a number of benchmarks, we provide evidence for its superior predictive performance in terms of the Kupiec test, the pinball score and the test for conditional predictive accuracy, as well as financial profits for a range of trading strategies, especially when the regularization parameter is selected *ex-ante* using the Bayesian Information Criterion (BIC). As such, we offer an efficient tool that can be used to boost the profitability of energy trading activities, help with bidding in day-ahead markets and improve risk management practices in the power sector.

1. Introduction

The Global Energy Forecasting Competition 2014 (GEFCom2014) has changed the landscape of energy forecasting. Instead of focusing on point forecasts, with all their limitations, the organizers have advocated computing *probabilistic* – more precisely – *quantile* forecasts and using the pinball score to evaluate them. The *electricity price forecasting* (EPF) literature has reacted lively, with the number of probabilistic EPF papers growing steadily in the following years (Hong et al., 2020). This has had the welcome effect of making the energy forecasting literature more interesting for economists and the financial industry. After all, the most commonly used risk measure – the Value-at-Risk (VaR) – is nothing else but a quantile forecast. The better the probabilistic forecast, particularly for the extreme quantiles, the more accurate the VaR estimate (Bunn et al., 2016; Uniejewski et al., 2019). And, obviously, this applies also to energy portfolios. Yet, risk management is not the only application of probabilistic energy forecasting. For instance, using predictive distributions as a basis for trading imbalances in an electricity market can be significantly more profitable and financially less risky than relying upon mean value-based point forecasts (Bunn et al., 2018). Hence the need for a development of efficient probabilistic EPF techniques.

Formally introduced by Nowotarski and Weron (2015), *Quantile Regression Averaging* (QRA) has sparked interest in the EPF community after its unprecedented success in GEFCom2014, where the top two winning teams in the price track used variants of QRA (Gaillard et al., 2016; Maciejowska and Nowotarski, 2016). The idea underlying QRA is to apply quantile regression (QR; see Koenker, 2005) to point – not probabilistic – forecasts of a pool of models, i.e., use the individual point predictions as regressors and the observed spot price as the predicted variable. This is of substantial practical value. QRA is able to leverage developments in point forecasting, which still is much more popular than its probabilistic counterpart, while offering an accurate approximation of the predictive distribution. Since it is a general forecasting technique not limited to electricity prices, a number of authors have reported its successful application in areas ranging from load (Liu et al., 2017; Zhang et al., 2018; Wang et al., 2019) to wind power (Zhang et al., 2016) and irradiance forecasting (Mpfumali et al., 2019).

However, a recent study of Marcjasz et al. (2020) has revealed the method's vulnerability to low quality predictors when the set of regressors is larger than just a few. The authors emphasize the importance of

* Corresponding author.
E-mail addresses: bartosz.uniejewski@pwr.edu.pl (B. Uniejewski), rafal.weron@pwr.edu.pl (R. Weron).

https://doi.org/10.1016/j.eneco.2021.105121
0140-9883/© 2021 Elsevier B.V. All rights reserved.

correctly selecting the individual point forecasts for QRA, whose performance is known only *ex post*, and even recommend using only two or three point forecasts as inputs. A way to tackle this problem via principal component analysis (PCA) has been proposed by Maciejowska et al. (2016). However, their Factor QRA (or FQRA) model yields suboptimal and nearly identical forecasts to the so-called Quantile Regression Machine (QRM) of Uniejewski et al. (2019), which first aggregates point predictions of the individual methods, then runs quantile regression on the combined forecasts.

Here, we introduce a different approach that significantly outperforms not only these two, but also a number of other benchmarks. It can be regarded as a regularized variant of QRA, which utilizes the *Least Absolute Shrinkage and Selection Operator* (LASSO) of Tibshirani (1996) to select the relevant regressors. We evaluate the introduced technique – dubbed LASSO QRA or LQRA for short – using datasets from the Polish and Nordic power markets. By comparing against nearly 30 benchmarks, we provide evidence for its superior predictive performance in terms of the Kupiec (1995) test, the pinball score (see, e.g., Gneiting, 2011; Nowotarski and Weron, 2018) and the test for conditional predictive accuracy of Giacomini and White (2006), especially when the regularization parameter is selected *ex-ante* using the Bayesian Information Criterion (BIC). As such, our study provides energy generators, retailers and brokers with a tool that can be used to boost the profitability of their trading activities, help with bidding in day-ahead markets and improve their risk management practices.

The remainder of the paper is structured as follows. In Section 2 we briefly describe the datasets and the forecasting scheme used in this study. Then, in Section 3 we discuss data preprocessing, describe the expert model used for computing the point forecasts and introduce LQRA. In Section 4, we first evaluate the predictive performance of the point forecasts, then discuss coverage and sharpness of the obtained probabilistic forecasts. In particular, we report the results for the 90% prediction intervals, which are based on quantiles that correspond to confidence levels often used in risk management, i.e., 5% and 95%. Given that electricity price forecasts, no matter how accurate, are of limited value if they cannot be used to devise a trading strategy that yields profits, in Section 4 we also evaluate a range of strategies that involve buying electricity when it is cheap, storing it and selling when it is expensive. Finally, in Section 5 we wrap up the results and conclude.

2. Datasets

2.1. Two distinct markets

We consider datasets from two neighboring, but distinct European power markets – the Polish Power Exchange (POLEX) and Nord Pool (NP). The first one was downloaded from the websites of POLEX (www.tge.pl) and the Polish Transmission System Operator (TSO; www.pse.pl). It comprises three time series at hourly resolution (see Fig. 1):

- day-ahead electricity prices for the main auction at POLEX (so-called 'Fixing #1'),
- day-ahead predictions of the system-wide load in Poland,
- day-ahead predictions of the generation of centrally dispatched generating units in Poland (denoted by JWCD),

and covers the period from 2 October 2014 to 31 December 2019. The Polish market is fossil fuel dominated, with a slowly decreasing share of hard coal and lignite (from ca. 86% in 2014 to ca. 75% in 2019).

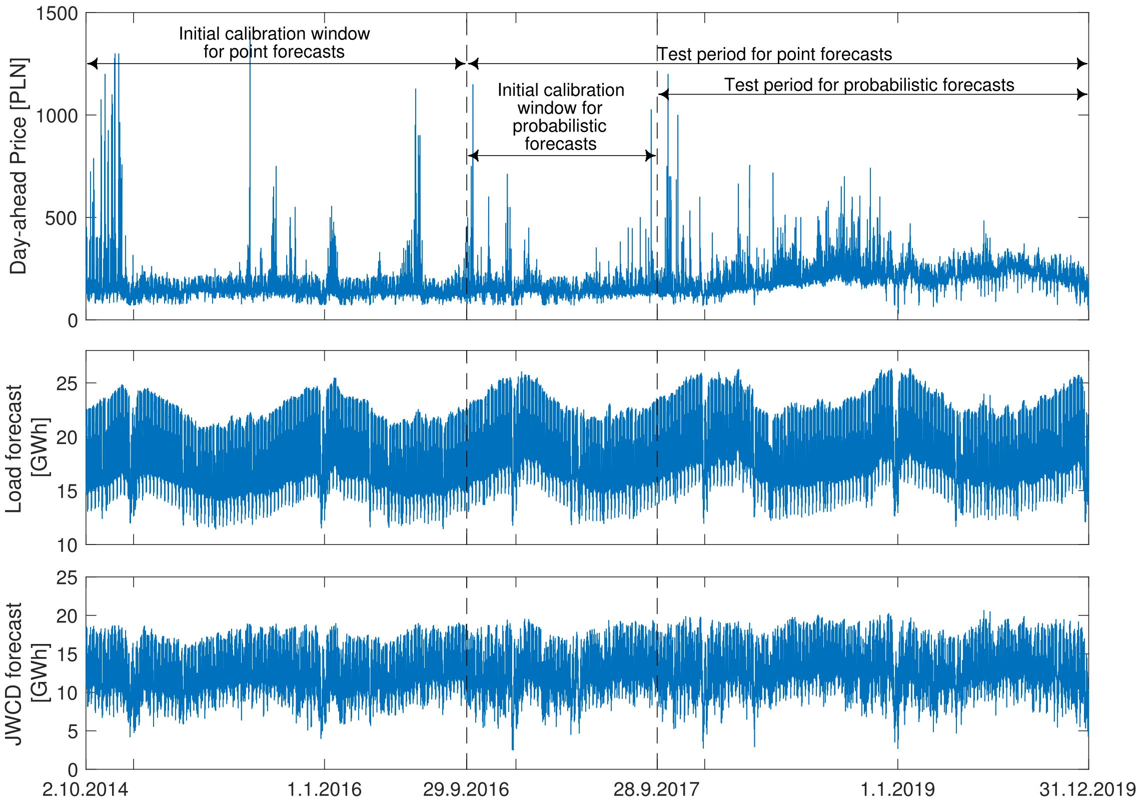


Fig. 1. Polish Power Exchange (POLEX) day-ahead prices (*top*), day-ahead system load forecasts (*middle*) and day-ahead forecasts of the generation of centrally dispatched generating units (*bottom*) from 2 October 2014 to 31 December 2019. The vertical dashed lines mark respectively the beginning of the 1189-day out-of-sample test period for point forecasts (29 September 2016; also the beginning of the initial 364-day calibration window for probabilistic forecasts) and the beginning of the 825-day out-of-sample test period for probabilistic forecasts (28 September 2017). The first 728 days constitute the initial calibration window for point forecasts.

The second dataset was downloaded from the Nord Pool website (www.nordpoolgroup.com) and also includes three time series at hourly resolution (see Fig. 2):

- day-ahead system prices for the whole Nordic region,
- day-ahead load forecasts (so-called *consumption prognosis*) aggregated for Denmark, Finland, Norway and Sweden,
- day-ahead forecasts of wind power generation in Denmark,

and covers the period from 1 January 2013 to 31 December 2019. In contrast to the Polish market, the Nordic one is hydro dominated (particularly Norway) with a large share of nuclear (Sweden, Finland) and wind power generation (Denmark).

All time series were preprocessed to account for changes to/from the daylight saving time. The missing values (corresponding to changes to the summertime) were substituted by the arithmetic average of the observations from neighboring hours. The doubled values (corresponding to the changes from the summertime) were replaced by their arithmetic mean.

2.2. The forecasting scheme

The idea underlying QRA is to apply quantile regression to a pool of point forecasts. Hence, our forecasting scheme is composed of two stages. First, point forecasts of the so-called *expert*, regression-type model defined in Section 3.1.2 are obtained via ordinary least squares (OLS). Then they are used to construct probabilistic forecasts by running quantile regression; the resulting models are described in Section 3.2.5. Since both stages require calibration windows, the first probabilistic predictions are obtained for the 1093rd day in each dataset, see the rightmost vertical dashed lines in Figs. 1 and 2.

Like the majority of EPF studies, we use a rolling window scheme. However, following Hubicka et al. (2019), Marcjasz et al. (2018) and Maciejowska et al. (2020), we do not consider an arbitrarily chosen calibration window length for computing point forecasts, but rather a pool of 701 window lengths – ranging from 28 (ca. one month) to 728 days (ca. two years) – and combine the obtained predictions. Initially, the first 728 (=2 × 364) days are used for calibration of the expert model. For shorter windows, the calibration sample is left-truncated, so that it ends on the same day as the 728-day window. Then for each day between 29 September 2016 and 31 December 2018 for POLEX and between 30 December 2014 and 31 December 2019 for Nord Pool, we simultaneously compute 24 point forecasts, one for each hour of the day. These are the 1189-day (POLEX) and 1828-day (NP) out-of-sample test periods for point forecasts, spanning from the leftmost vertical dashed lines in Figs. 1–2 to the last day in the datasets. To obtain probabilistic forecasts from point predictions, for each hour of the days between 28 September 2017 and 31 December 2018 for POLEX and between 29 December 2015 and 31 December 2019 for Nord Pool, we use a rolling 364-day calibration window and one of the models described in Section 3.2.5. The 825-day (POLEX) and 1464-day (NP) out-of-sample test periods for probabilistic forecasts span from the rightmost vertical dashed lines in Figs. 1-2 to the last day in the datasets; the probabilistic forecasts themselves are evaluated in Section 4.2.

3. Methodology

Since our forecasting scheme is composed of two stages, we first describe the point forecasting concepts in Section 3.1, then the probabilistic ones in Section 3.2. Each of these Sections ends with a summary which wraps up the discussed ideas and presents the models.

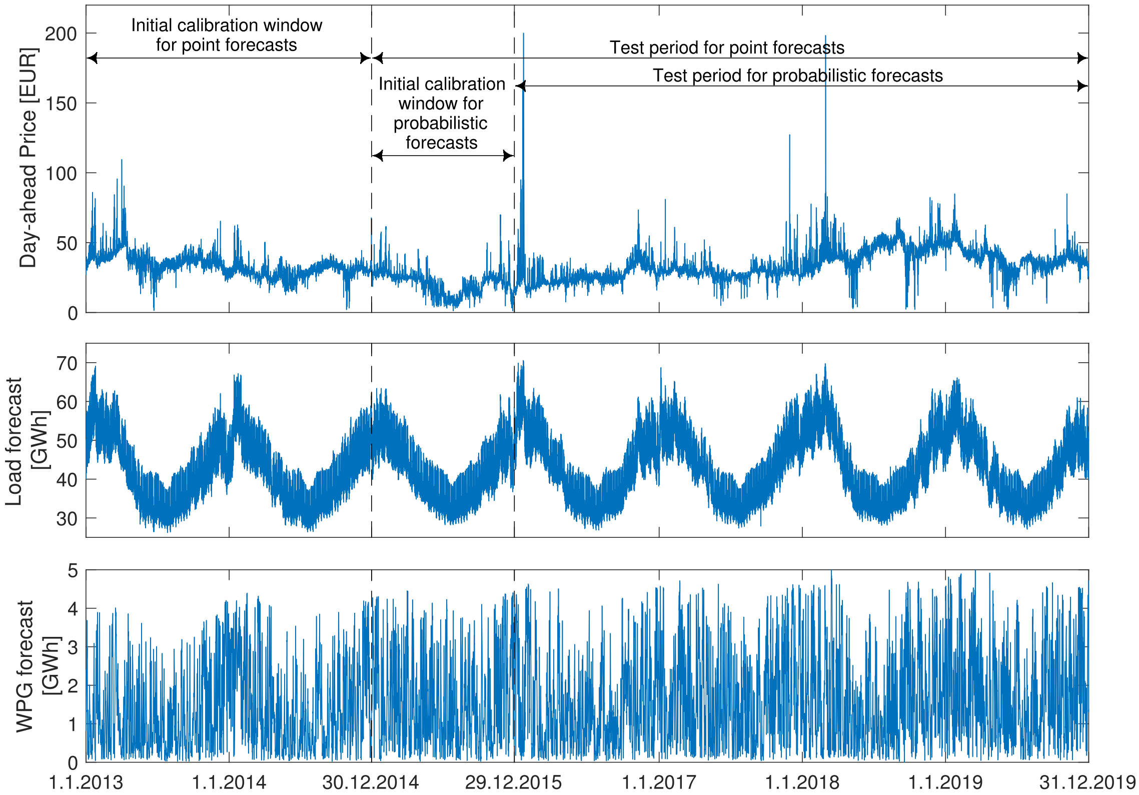


Fig. 2. Nord Pool (NP) system prices (*top*), day-ahead consumption prognosis (*middle*) and day-ahead forecasts of wind power generation (*bottom*) from 1 January 2013 to 31 December 2019. The vertical dashed lines mark respectively the beginning of the 1828-day out-of-sample test period for point forecasts (30 December 2014; also the beginning of the initial 364-day calibration window for probabilistic forecasts) and the beginning of the 1464-day out-of-sample test period for probabilistic forecasts (29 December 2015). The first 728 days constitute the initial calibration window for point forecasts.

3.1. Point forecasts

3.1.1. Variance stabilizing transformations

The estimation of electricity price models via maximum likelihood (ML) or ordinary least squares (OLS) is typically hampered by spikes. These 'outliers' tend to pull model estimates towards values that yield a better fit for the extreme observations, but at the same time increase the in-sample errors for the non-spiky prices. A statistically sound modeling framework would either require robust estimation algorithms, as in Grossi and Nan (2019), or models with explicit spike components, for a review see Weron (2014). Neither of these are popular in the EPF literature. Instead the authors have resorted either to treating electricity price series with a 'reasonable' filter (Janczura et al., 2013; Lisi and Pelagatti, 2018; Afanasyev and Fedorova, 2019) and calibrating the model to spike-filtered data or to transforming the original series, fitting the model to transformed prices and applying the inverse transformation to obtain the forecasts (Diaz and Planas, 2016; Uniejewski et al., 2018; Narajewski and Ziel, 2020); the latter approach has the advantage of using spiky prices, though their impact on parameter estimates is reduced. For markets with only positive and distinctly different from zero prices, the logarithmic transform can be used, which leads to popular in finance models for log-prices. However, spikes come in all shapes and sizes – historically they have been mostly positive, but with the increasing penetration of renewables significant price drops have started appearing (Maciejowska, 2020). In some markets, like in Germany, negative spikes are more common nowadays than positive ones (Hagfors et al., 2016b). This has triggered the development of alternative transformations that can handle negative and close to zero values. In an extensive empirical study involving two model classes (regression models, neural networks) and datasets from 12 diverse power markets, Uniejewski et al. (2018) have evaluated 16 *variance stabilizing transformations* (VSTs), ranging from simple threshold-type cutoffs, through generalized Box-Cox type transforms, to the *probability integral transform* (PIT) based approaches. Although there was no clear-cut winner, the N-PIT, the *mirror-log* and the *area hyperbolic sine* (asinh) transforms performed reasonably well, with the N-PIT having an edge over the competitors for the considered regression model and asinh performing very well for the neural network. Given that asinh is much simpler to compute, like Schneider (2011), Ziel and Weron (2018) and Marcjasz (2020), we use it in our study.

Each time series (prices and exogenous variables) is first standardized by subtracting the sample median and dividing by the sample *Median Absolute Deviation* (MAD) corrected by the 75% quantile of the standard normal distribution $z_{0.75}$; note, that according to Uniejewski et al. (2018), the (median, MAD) standardization yields better results than the more commonly used in statistics (mean, standard deviation) pair. For instance, for prices $P_{d,h}$ we have:

$$p_{d,h} = \frac{P_{d,h} - a}{b} \equiv \frac{P_{d,h} - \mathrm{med}(P_{d,h})}{\frac{\mathrm{MAD}(P_{d,h})}{z_{0.75}}}. \tag{1}$$

Then we asinh-transform the standardized series. For instance, for prices we obtain:

$$Y_{d,h} = \mathrm{asinh}(p_{d,h}) = \log\left(p_{d,h} + \sqrt{p_{d,h}^2 + 1}\right), \tag{2}$$

for each d and h. Once the transformed price forecast for day d and hour h, i.e., $\widehat{Y}_{d,h}$, is computed, we apply the inverse transformation to obtain the price forecast itself. Here, we follow Narajewski and Ziel (2020) and define the back-transformation as:

$$\widehat{P}_{d,h} = a + \frac{b}{D}\sum_{i=1}^{D} \sinh\left(\widehat{Y}_{d,h} + \varepsilon_i\right), \tag{3}$$

where D is the number of days in the calibration window, ε_i are the in-sample residuals of the OLS procedure, and a and b are defined in Eq. (1). Note, that the back-transformation proposed by Uniejewski et al. (2018) is simpler, but may be less accurate when $\widehat{Y}_{d,h}$ is far from zero.

3.1.2. The expert model

For computing point forecasts we use a parsimonious autoregressive structure, inspired by the ARX model of Misiorek et al. (2006) and later used in a number of EPF studies (see Ziel and Weron, 2018, for a review). Following Uniejewski et al. (2016) and Ziel (2016), we refer to it as an *expert* model, since it is built on some prior knowledge of experts. The model for the asinh-transformed price on day d and hour h is given by:

$$\begin{aligned} Y_{d,h} = \ & \underbrace{\beta_1 Y_{d-1,h} + \beta_2 Y_{d-2,h} + \beta_3 Y_{d-7,h}}_{autoregressive\ \ effects} + \underbrace{\beta_4 Y_{d-1,24}}_{end-of-day} + \underbrace{\beta_5 Y_{d-1}^{max} + \beta_6 Y_{d-1}^{min}}_{non-linear\ \ effects} \\ & + \underbrace{\beta_7 Z_{d,h}^{(1)} + \beta_8 Z_{d,h}^{(2)}}_{exogenous\ \ variables} + \underbrace{\beta_9 D_{Sat} + \beta_{10} D_{Sun} + \beta_{11} D_{Mon}}_{weekday\ \ dummies} + \varepsilon_{d,h}, \end{aligned} \tag{4}$$

where $Y_{d-1,h}$, $Y_{d-2,h}$ and $Y_{d-7,h}$ account for the autoregressive effects and correspond to prices from the same hour of the previous day, two days before and a week before, $Y_{d-1,24}$ is the last known price at the time the prediction is made and provides information about the end-of-day price level, $Y_{d-1}^{max} \equiv \max_{h=1,\ \ldots,\ 24}\{Y_{d-1,h}\}$ and $Y_{d-1}^{min} \equiv \min_{h=1,\ \ldots,\ 24}\{Y_{d-1,h}\}$ represent previous day's variation in prices, $Z_{d,\ h}^{(1)}$ is the asinh-transformed day-ahead system load forecast for day d and hour h, $Z_{d,h}^{(2)}$ is the asinh-transformed day-ahead forecast for day d and hour h of either generation of centrally dispatched generating units (for POLEX) or of wind power generation (for Nord Pool), D_{Sat}, D_{Sun} and D_{Mon} are respectively Saturday, Sunday and Monday dummies, and the $\varepsilon_{d,h}$'s are assumed to be independent and identically distributed normal variables. The model weights are estimated via OLS, independently for each day and hour.

3.1.3. Point forecasting models

Summing up, for each day in the out-of-sample test period for point forecasts, i.e., $d = 1, 2, \ldots, 1189$ (for POLEX) and $d = 1, 2, \ldots, 1828$ (for NP), each hour of the day, i.e., $h = 1, 2, \ldots, 24$, and each calibration window length, i.e., $T = 28, 29, \ldots, 728$ days (see Section 2 for details), we follow the same routine:

$$P_{d,h} \overset{\text{standardize}}{\rightarrow} p_{d,h} \overset{\text{VST}}{\rightarrow} Y_{d,h} \overset{\text{predict}}{\rightarrow} \widehat{Y}_{d,h} \overset{\text{inverseVST}}{\rightarrow} \widehat{P}_{d,h}. \tag{5}$$

Namely, we (*i*) standardize the prices in the calibration window using Eq. (1), (*ii*) VST-transform the standardized prices in the calibration window using the area hyperbolic sine defined in Eq. (2), (*iii*) estimate parameters of the expert model (4) via OLS and compute the one-step ahead point predictions, and finally (*iv*) apply the inverse of the asinh transformation given by Eq. (3) to obtain the price forecasts. Note, that we also standardize and asinh-transform the exogenous variable time series prior to fitting the expert model in step (*iii*). As a result of this routine, for each day d and hour h we obtain 701 different point forecasts of the same expert model estimated on data from 701 calibration windows. We can treat these forecasts as coming from 701 different 'models'.

3.2. Probabilistic forecasts

3.2.1. Quantile Regression Averaging (QRA)

QRA, as introduced by Nowotarski and Weron (2015), is based on averaging $n = 1, \ldots, N$ point predictions of electricity prices $\widehat{P}_{d,h}^{(n)}$ using quantile regression (QR; see Koenker, 2005):

$$q(\alpha|\mathbf{X}_{d,h}) = \mathbf{X}_{d,h}\boldsymbol{\beta}_{\alpha}, \tag{6}$$

where $q(\alpha|\cdot)$ is the conditional αth quantile of the electricity price distribution, $\mathbf{X}_{d,h}$ is the vector of independent variables, i.e., $\widehat{P}_{d,h}^{(n)}$ for

$n = 1, \ldots, N$, while $\boldsymbol{\beta}_\alpha = [\beta_\alpha^{(1)}, \ldots, \beta_\alpha^{(N)}]'$ is the corresponding vector of weights for given α. To obtain parameter estimates we minimize the sum of the so-called *check functions*:

$$\hat{\boldsymbol{\beta}}_\alpha = \underset{\boldsymbol{\beta}_\alpha}{\operatorname{argmin}} \sum_{d,h} \underbrace{\left(\alpha - 1_{P_{d,h} < \mathbf{X}_{d,h}\boldsymbol{\beta}_\alpha}\right)\left(P_{d,h} - \mathbf{X}_{d,h}\boldsymbol{\beta}_\alpha\right)}_{check\ \ function}, \tag{7}$$

where 1_X is the characteristic function of set X. Following the unprecedented success in the Global Energy Forecasting Competition 2014, QRA became a popular technique for probabilistic energy forecasting (see, e.g., Maciejowska et al., 2016; Zhang et al., 2016; Liu et al., 2017; Zhang et al., 2018; Kostrzewski and Kostrzewska, 2019; Mpfumali et al., 2019; Serafin et al., 2019; Uniejewski et al., 2019; Wang et al., 2019; Kath and Ziel, 2020). However, a recent study of Marcjasz et al. (2020) has revealed the method's vulnerability to low quality predictors when the set of regressors is larger than just a few. To tackle this problem, in what follows we introduce a *regularized* (or *penalized*) variant of QRA.

QRA is a special case of QR where the regressors are point (mean value) forecasts of the predicted variable, i.e., $\widehat{P}_{d,h}^{(n)}$ for $n = 1, \ldots, N$. Note, however, that QR is a general technique that can yield electricity price forecasts also for more diverse sets of explanatory variables, including past prices and/or fundamental variables, as in Bunn et al. (2016) or Hagfors et al. (2016a), potentially also higher moments of the predictive distribution or quantile forecasts. To our best knowledge, the latter idea has not been utilized in the EPF context thus far.

3.2.2. Regularization

In simple terms the idea of regularization can be formulated as an optimization problem:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \{ f(\mathbf{X}; \boldsymbol{\beta}) + g(\boldsymbol{\beta}) \}, \tag{8}$$

where $f(\mathbf{X}; \boldsymbol{\beta})$ is the minimized function, $g(\boldsymbol{\beta})$ is the penalty function and $\boldsymbol{\beta}$ is the vector of parameters (Tikhonov, 1963). The most commonly used variant of regularization defines the penalty function as an ℓ^q norm scaled with an additional *tuning* (or *regularization*) parameter λ. When $q = 1$ and $f(\mathbf{X}; \boldsymbol{\beta})$ is the Residual Sum of Squares (RSS) we obtain the *Least Absolute Shrinkage and Selection Operator* (LASSO) of Tibshirani (1996):

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \underbrace{\sum\nolimits_{d,h} \left(P_{d,h} - \mathbf{X}_{d,h}\boldsymbol{\beta}\right)^2}_{\text{RSS}} + \lambda \sum_{i=1}^{n} |\beta_i| \right\}, \tag{9}$$

which not only shrinks the β's towards zero, but also eliminates some of them. For $\lambda = 0$ the method is equivalent to the original formulation of the problem, i.e., OLS. As λ increases, more and more variables are eliminated. A clear advantage of LASSO is that it can handle an almost unlimited number of explanatory variables. Hence, it can be used as an automated variable selection tool (Gaillard et al., 2016; Uniejewski et al., 2016; Ziel, 2016; Uniejewski and Weron, 2018; Ziel and Weron, 2018).

3.2.3. LASSO QRA

Regularization applied to quantile regression is not a completely new idea (for a mathematical treatment of this topic see Li and Zhu, 2008). There are even a few forecasting papers where this approach has been considered. On one hand, there are publications where LASSO is combined with QR to construct probabilistic forecasts of economic variables (Manzan, 2015), global radiation (Ben Bouallègue, 2017), electric demand (Lebotsa et al., 2018; He et al., 2019) or photovoltaic power (Agoua et al., 2019), for a pool of explanatory variables (but not point forecasts as in QRA). On the other, there are studies where penalized (via LASSO) QR is used to obtain a quantile of the predictive distribution (e.g., Value at Risk at the 95% level) from a pool of forecasts of this particular quantile (Bayer, 2018; Bracale et al., 2019). However, to our best knowledge, there is no paper that combines regularization with QRA, i.e., with mean value (not quantile) forecasts taken as independent variables in Eq. (6).

The LASSO QRA or LQRA for short is constructed by applying the check function defined in Eq. (7) to the ℓ^1 regularization in Eq. (9):

$$\hat{\boldsymbol{\beta}}_\alpha = \underset{\boldsymbol{\beta}_\alpha}{\operatorname{argmin}} \left\{ \sum\nolimits_{d,h} \underbrace{\left(\alpha - 1_{P_{d,h} < \mathbf{X}_{d,h}\boldsymbol{\beta}_\alpha}\right)\left(P_{d,h} - \mathbf{X}_{d,h}\boldsymbol{\beta}_\alpha\right)}_{check\ \ function,\ \ \text{see}\ \ \text{Eqn.}\ \ (7)} + \underbrace{\lambda \sum_{n=1}^{N} \left|\beta_\alpha^{(n)}\right|}_{\text{LASSO}\ \ \text{penalty}} \right\}, \tag{10}$$

where, as in Eq. (6), $\mathbf{X}_{d,h}$ is the vector of independent variables, i.e., $\widehat{P}_{d,h}^{(n)}$ for $n = 1, \ldots, N$, for day d and hour h, and $\mathbf{X}_{d,h}\boldsymbol{\beta}_\alpha = \sum_{n=1}^{N} \beta_\alpha^{(n)} \widehat{P}_{d,h}^{(n)}$. Note, that Eq. (10) can be rewritten to fit into the definition of a linear problem, thus to estimate $\widehat{\boldsymbol{\beta}}_\alpha$ we use a basic linear programming tool, i.e., the Simplex algorithm (Dantzig et al., 1955).

Due to the linear penalty factor, LQRA selects the most informative point forecasts, so there is no need for experts to select them a priori. What is more, because of regularization, the final forecasts of different quantiles may be constructed using different sets of point forecasts. Potentially, this may improve the predictive performance. Actually, we expect the model to be at least as accurate as QRA and at the same time much more robust to a large number of point forecast (inputs). The higher the λ, the fewer point forecasts will finally be used to construct the quantile forecasts. Note, that λ may or may not depend on α. In Section 4.2 we present results for models with one fixed λ for all 99 percentiles as well as a model which selects the optimal value of λ for each percentile and each hour.

3.2.4. Selection of λ

To check the performance of LQRA we conduct empirical tests on a logarithmic grid of 19 λ's ranging from 10^{-1} to 10^3. We additionally include $\lambda = 0$, which corresponds to the original QRA method; in total we report results for 20 different values of λ. We also consider two procedures for selecting the optimal value of λ. The first, denoted by **LQRA (BIC)**, utilizes the *Bayesian Information Criterion* (BIC) to select the tuning parameter based on the in-sample fit and uses the formulation proposed by Lee et al. (2014). Namely, for quantile α, day d and hour h we select the λ that corresponds to the lowest BIC value:

$$\begin{aligned} \text{BIC}_{d,h}^{\alpha}(\lambda) = \ & \log\left(\sum_{d^*=d-364}^{d} \left(\alpha - 1_{P_{d^*,h} < \hat{P}_{d^*,h}^{\alpha}}\right)\left(P_{d^*,h} - \hat{P}_{d^*,h}^{\alpha}\right)\right) \\ & + m \frac{\log(n)}{2n} \log(p), \end{aligned} \tag{11}$$

where $P_{d,\ h}$ is the observed price, $\widehat{P}_{d,h}^{\alpha}$ is the corresponding model-predicted α-quantile of the price distribution (in-sample fit), m is the number of non-zero parameters in the regularized model, $n = 364$ is the number of observations in the calibration window and p is the number of potential predictors (here: 25).

The second procedure, denoted by **LQRA(CV)**, utilizes cross-validation to select the optimal value of λ, see, e.g., Hastie et al. (2015) for a discussion and examples. We divide the in-sample data into $k = 7$ disjunctive subsets of $\frac{364}{7} = 52$ days; note, that the length of the calibration window for probabilistic forecasts is 364 days. Then we calibrate the model k times, each time we leave out one of the k subsets for testing. Finally, we compare the performance for each λ by looking at the Pinball Score, defined in Eq. (13) below, averaged across all days in the calibration window. The λ with the lowest score is chosen, independently for each percentile.

3.2.5. Probabilistic forecasting models

The probabilistic forecasts are obtained for QR-type models calibrated on a rolling window of 364 days; this is in contrast to point forecasts, which are fitted on 701 rolling windows of lengths ranging from 28 to 728 days. For example, in case of the Polish data, the initial calibration window for probabilistic forecasts covers the period from 29 September 2016 to 27 September 2017 and yields forecasts for all 24 h of the next day (28 September). Afterwards, the window is rolled forward by one day and the models are re-estimated to obtain forecasts for 29 September. This procedure is repeated until the predictions for 31 December 2019 are made.

To obtain probabilistic forecasts of the day-ahead price, we use only 25 out of 701 individual point forecasts available (see Section 4.1). Marcjasz et al. (2018) claim that it is not beneficial to take into consideration models that are calibrated on windows of fewer then 56 days. What is more, forecasts obtained for windows of similar lengths provide almost identical results. Thus, as the potential variables for LQRA we have selected point forecasts obtained for models calibrated on windows of 56,84,112, …, 728 days; see the black dots in Fig. 3. The resulting models are denoted by **LQRA**(λ_i), where $\lambda_i = 10^{\frac{2(i-1)}{9}-1}$ for $i = 1, \ldots, 19$ is a logarithmic grid of λ's ranging from 10^{-1} to 10^3.

Note, that $\lambda = 0$ in Eq. (10) corresponds to 'pure' or non-regularized QRA. The latter is treated as the first benchmark and denoted later in the text by **QRA**. The next 25 benchmarks use QR applied to individual point forecasts; we denote them by **QR**(T), where $T = 56,84,112, \ldots, 728$. The last three benchmarks are motivated by the results of Marcjasz et al. (2020). The authors report that averaging probabilistic forecasts can further improve the accuracy. Thus, we additionally average probabilistic forecasts **QR**(T), both 'horizontally' and 'vertically', see Section 3.3.2 in the cited article. We denote them by **Q-Ave** and **F-Ave**, respectively. Finally, since it may be beneficial to first average point forecasts and then apply QR (instead of averaging predictions on the probabilistic level), we also consider the Quantile Regression Machine (QRM) approach of Uniejewski et al. (2019). In this model, denoted by **QRM**, all 25 individual point forecasts are first averaged and the result is then used as an input to the QR algorithm. Interestingly, QRM yields nearly identical forecasts to the Factor QRA (or FQRA) model of Maciejowska et al. (2016), because the latter applies QR to the first factor (which approximates the mean) extracted from a pool of forecasts.

Finally, note that in this study we forecast 99 percentiles, which can be treated as a relatively good approximation of the entire distribution. Due to a possible lack of monotonicity of the quantile forecasts, after obtaining predictions of the 99 percentiles we sort the results, independently for each day and hour.

4. Results

4.1. Point forecasts

The point forecasts are evaluated using the *Mean Absolute Error*:

$$\mathrm{MAE} = \frac{1}{24D}\sum_{d=1}^{D}\sum_{h=1}^{24}|\widehat{P}_{d,h} - P_{d,h}|,$$

where D is the number of days in the out-of-sample test period. The results are presented in Fig. 2. They are very similar to those reported in Hubicka et al. (2019) and Marcjasz et al. (2018). As can be seen for both datasets, the optimal (in terms of MAE) calibration window length is around 100-days; this can hardly be selected *ex-ante*. For both datasets the shortest windows are not recommended. What was also reported by Marcjasz et al. (2018) for some of the analyzed markets, the longer calibration window we take, the worse the autoregressive model performs. We observe this behavior both for the POLEX and NP datasets.

4.2. Probabilistic forecasts

4.2.1. Error measures and testing for predictive ability

When evaluating probabilistic forecasts, we follow the paradigm of 'maximizing sharpness subject to reliability' (Gneiting and Katzfuss, 2014; Nowotarski and Weron, 2018). To check the *reliability* we compute the empirical coverage:

$$I^{\alpha}_{d.h} = \begin{cases} 1, & \text{for } P_{d,h} \in \left[\hat{L}^{\alpha}_{d,h}\hat{U}^{\alpha}_{d,h}\right], \\ 0, & \text{for } P_{d,h} \notin \left[\hat{L}^{\alpha}_{d,h}\hat{U}^{\alpha}_{d,h}\right], \end{cases} \tag{12}$$

where $\hat{L}^{\alpha}_{d,h}$ is the lower and $\hat{U}^{\alpha}_{d,h}$ the upper bound of the prediction interval (PI), and compare it to the nominal coverage. To this end, we perform the Kupiec (1995) test for the 50% and 90% PIs.

Then, we use the Pinball Score (or Pinball Loss) to evaluate the *sharpness*. It is a proper scoring rule and a special case of an asymmetric piecewise linear loss function (Gneiting, 2011):

$$\mathrm{PS}\left(\hat{q}_{\alpha,P}, P_{d,h}, \alpha\right) = \begin{cases} (1-\alpha)\left(\hat{q}_{\alpha,P} - P_{d,h}\right) & \text{for } P_{d,h} < \hat{q}_{\alpha,P}, \\ \alpha\left(P_{d,h} - \hat{q}_{\alpha,P}\right) & \text{for } P_{d,h} \geq \hat{q}_{\alpha,P}, \end{cases} \tag{13}$$

where $\hat{q}_{\alpha,P}$ is the price quantile of order $\alpha \in (0,1)$ and $P_{d,h}$ is the observed price for day d and hour h. The lower PS is, the more accurate are the probabilistic forecasts for a given quantile. The Pinball Score can be averaged across all percentiles, i.e., $\alpha = 0.01, 0.02, \ldots, 0.99$ as in the GEFCom2014 competition, and all hours in the whole out-of-sample test period. This yields the *Aggregate Pinball Score* (APS), which is

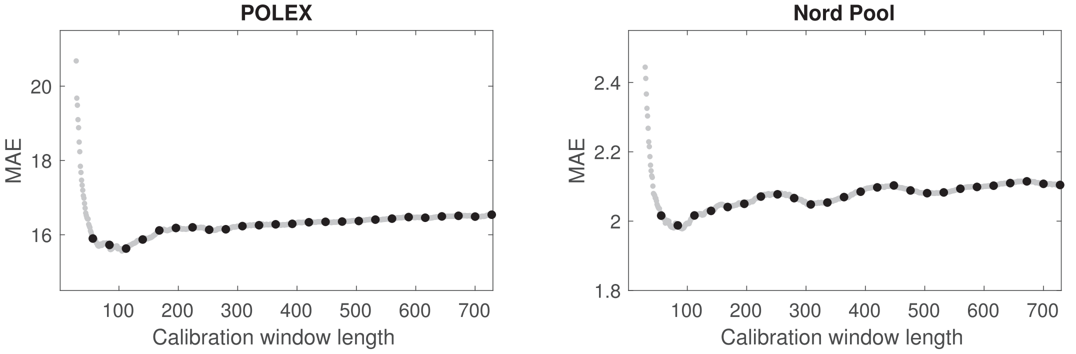


Fig. 3. The *Mean Absolute Errors* (MAE) for the POLEX dataset from the period 29 September 2016 to 31 December 2019 (*left panel*) and the Nord Pool dataset from the period 30 December 2014 to 31 December 2019 (*right panel*), as a function of the calibration window length ranging from 28 to 728 days. Black dots represent forecasts used as inputs for QR-based procedures, gray dots – predictions for all considered window lengths.

equivalent to computing the quantile representation of the so-called Continuous Ranked Probability Score (CRPS), see Laio and Tamea (2007).

Finally, to draw statistically significant conclusions on the outperformance of forecasts of one model by those of another, we use the Giacomini and White (2006) test for *conditional predictive ability* (CPA). It can be regarded as a generalization of the commonly used Diebold-Mariano test for *unconditional* predictive ability. Here, one statistic for each pair of models is computed based on the 24-dimensional vector of Pinball Scores for each day:

$$\Delta_{X,Y,d} = \|APS_{X,d}\| - \|APS_{Y,d}\|,$$

where $\|APS_{X,d}\| = \sum_{h=1}^{24}\sum_{\alpha=0.01}^{0.99} PS\left(\widehat{q}_{\alpha,P}, P_{d,h}, \alpha\right)$ for model X. For each pair of models and both datasets we compute the p-values of the CPA test with null $H_0 : \boldsymbol{\phi} = 0$ in the regression: $\Delta_{X,Y,d} = \boldsymbol{\phi}'\mathbb{X}_{d-1} + \varepsilon_d$, where $\mathbb{X}_{d-1}$ contains information of day $d-1$, i.e., a constant and lags of $\Delta_{X,Y,d}$.

4.2.2. Empirical coverage

In Fig. 4 we depict the so-called *Prediction Interval Coverage Probability* (PICP; see, e.g., Nowotarski and Weron, 2018) or empirical coverage of the 90% PIs constructed from the 5th and 95th percentiles, i.e., for $\alpha = 0.05$ and 0.95. Plotted are the results of empirical coverage aggregated for all hours in the out-of-sample test period, i.e., $825 \times 24 = 19\,800$ observations for POLEX and $1464 \times 24 = 35\,136$ for Nord Pool. Several conclusions can be drawn:

- Firstly, we can observe that the dashed lines indicating the expected level of 90% are rarely reached. This means that most of the PIs are too narrow; the exceptions include **F-Ave** (for both markets) and **Q-Ave** (for NP).
- As far as benchmarks are concerned, it is hard to formulate clear-cut conclusions. For the Polish dataset, **F-Ave** and **Q-Ave** outperform **QR** (*T*) for all *T*. However, for Nord Pool, **Q-Ave** and **QR**(*T*) perform similarly and **F-Ave** yields too wide PIs; **QRM** yields similar coverage to **QR**(*T*) for both datasets.
- As emphasized by Lichtendahl et al. (2013), vertical averaging of probabilities (→ **F-Ave**) yields wider PIs than horizontal averaging of quantiles (→ **Q-Ave**), leading to less sharp predictive distributions. While this often helps in EPF (Uniejewski et al., 2019; Marcjasz et al., 2020), for the considered Nord Pool dataset the **F-Ave**-implied PIs are clearly too wide. Possibly this is a consequence of the much less spiky prices in 2019 compared to the earlier years, see Fig. 2.
- **LQRA** strongly depends on λ. Rather larger values of the tuning parameter are preferable.
- Both methods for automated selection of λ, i.e., **LQRA(BIC)** and **LQRA (CV)**, are slightly outperformed by the best (*ex-post*) fixed λ values. At the same time, for both markets they yield similar coverage to **QRM** and **QR**(*T*).
- The most reliable results are obtained with **F-Ave** for POLEX and **QR (168)** for Nord Pool. They yield a coverage of 90.45% and 89.67%, respectively.

Similar conclusions can be drawn for the 50% PIs (not plotted here for the sake of clarity).

4.2.3. The Kupiec test

The test is performed to check whether the obtained PIs are close enough to the nominal values of PINC = 50% or 90%, or equivalently whether the Average Coverage Errors, defined as ACE = PICP − PINC, are close enough to zero. Recall, that the Kupiec (1995) test checks whether the probability of an actual price falling into the PI (→ ‘hit’) is equal to PINC, under the assumption that the ‘hits’ and ‘misses’ are independent. The test rejects the null hypothesis of ACE equal to zero if the actual fraction of ‘hits’ is statistically different from PINC (Nowotarski and Weron, 2018). In what follows, we report the results for three benchmarks (**Q-Ave**, **F-Ave**, **QRM**) and three LQRA models – one with an *ex-post* selected, relatively well performing value of $\lambda_{14} = 77.43 \approx 77$, denoted by **LQRA(77)**, and two that utilize automated selection of the tuning parameter, i.e., **LQRA(BIC)** and **LQRA(CV)**.

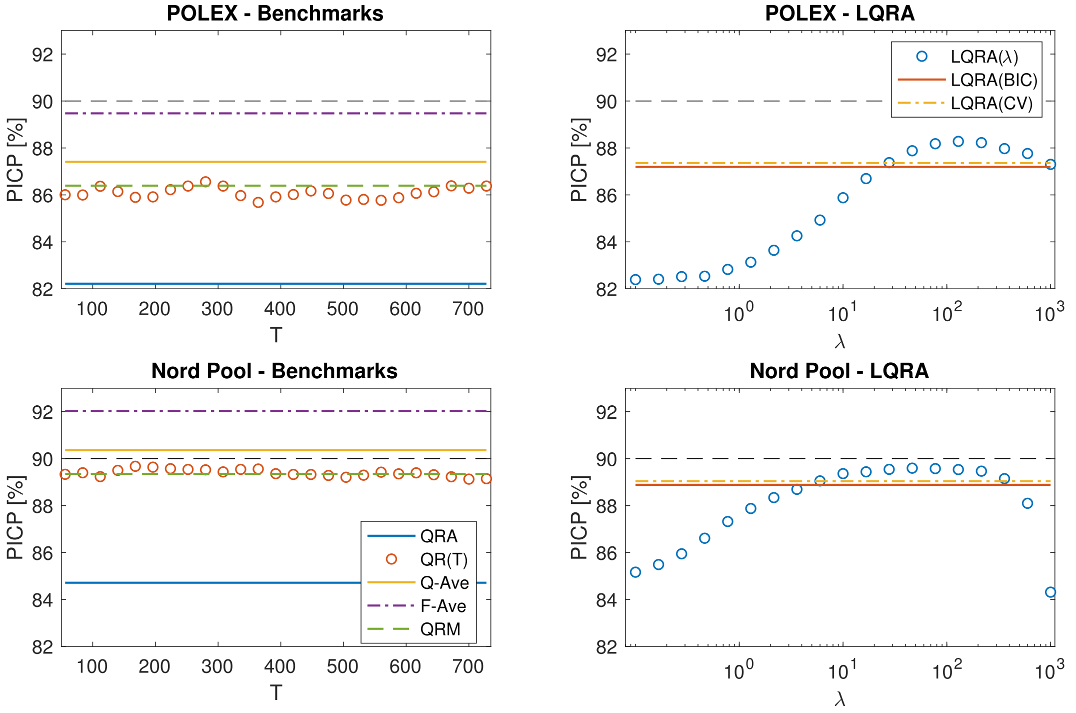


Fig. 4. The PI Coverage Probability (PICP) of the probabilistic forecasts for the POLEX dataset from 29 September 2016 to 31 December 2019 (*top panels*) and the Nord Pool dataset from 30 December 2014 to 31 December 2019 (*bottom panels*). The black dashed line corresponds to the PI Nominal Coverage (PINC) of 90%. Results are divided into two plots, the coverage for benchmark models (*left panels*; see Section 3.2.5) and **LQRA** as a function of λ and the two automated procedures for λ selection (*right panels*; see Section 3.2.4).

For both datasets, the Kupiec test is conducted for the whole out-of-sample test period, but separately for each of the 24 hourly series. In Fig. 4 we depict the ACE (in percent) for the 50% and 90% PIs generated by the six selected probabilistic forecasting methods. The filled symbols indicate significance of the Kupiec test at the 1% level. The latter results, also at the 5% level, are summarized in Table 1. We can draw the following conclusions:

- Nearly in all cases the PIs for off-peak prices (i.e., night hours, early mornings and late evenings) are too narrow (ACE <0), while for on-peak prices they are usually too wide (ACE >0).
- As far as benchmarks are concerned, the best performing is **QRM**, particularly for on-peak hours. On the other hand, **F-Ave** seems to perform slightly better for off-peak hours.
- All three **LQRA** models outperform the benchmarks. They pass the Kupiec test at the 1% significance level for nearly all on-peak hours, and for as many as 17–22 h for the NP dataset. At the 5% significance level the results are slightly less convincing, but still much better than for the benchmarks.
- No significant differences between the performance of **LQRA(BIC)** and **LQRA(CV)** can be observed; interestingly, they are at par with **LQRA**

Table 1
The number of hours of the day (i.e., out of 24) for which the null of the Kupiec test is not rejected at the 1% and 5% significance levels, for both datasets. Compare with Fig. 5.

	POLEX				**Nord Pool**			
PI	*90%*		*50%*		*90%*		*50%*	
Sig. level	*1%*	*5%*	*1%*	*5%*	*1%*	*5%*	*1%*	*5%*
Q-Ave	6	3	6	6	5	3	8	8
F-Ave	2	2	5	4	7	6	5	2
QRM	11	8	10	10	15	8	12	8
LQRA(77)	14	11	18	16	20	17	20	18
LQRA(BIC)	12	10	17	14	17	17	22	18
LQRA(CV)	12	10	19	16	17	17	21	18

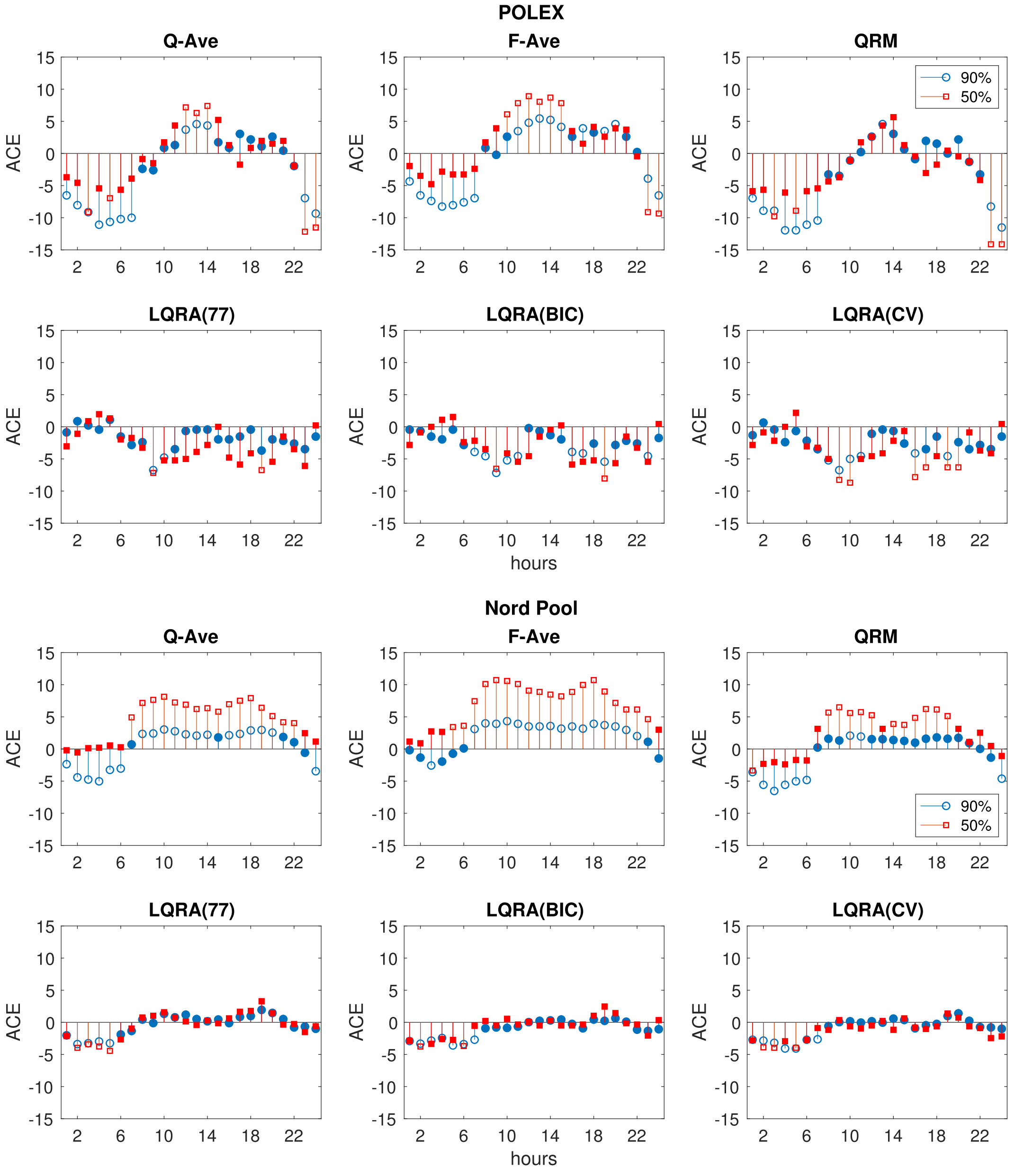


Fig. 5. The Average Coverage Errors (ACE = PICP − PINC; in percent) for the 50% and 90% PIs generated by the six selected probabilistic forecasting methods, separately for each hour of the day and the POLEX (*top six panels*) and NP datasets (*bottom six panels*). Filled symbols indicate significance of the Kupiec test at the 1% level, compare with Table 1.

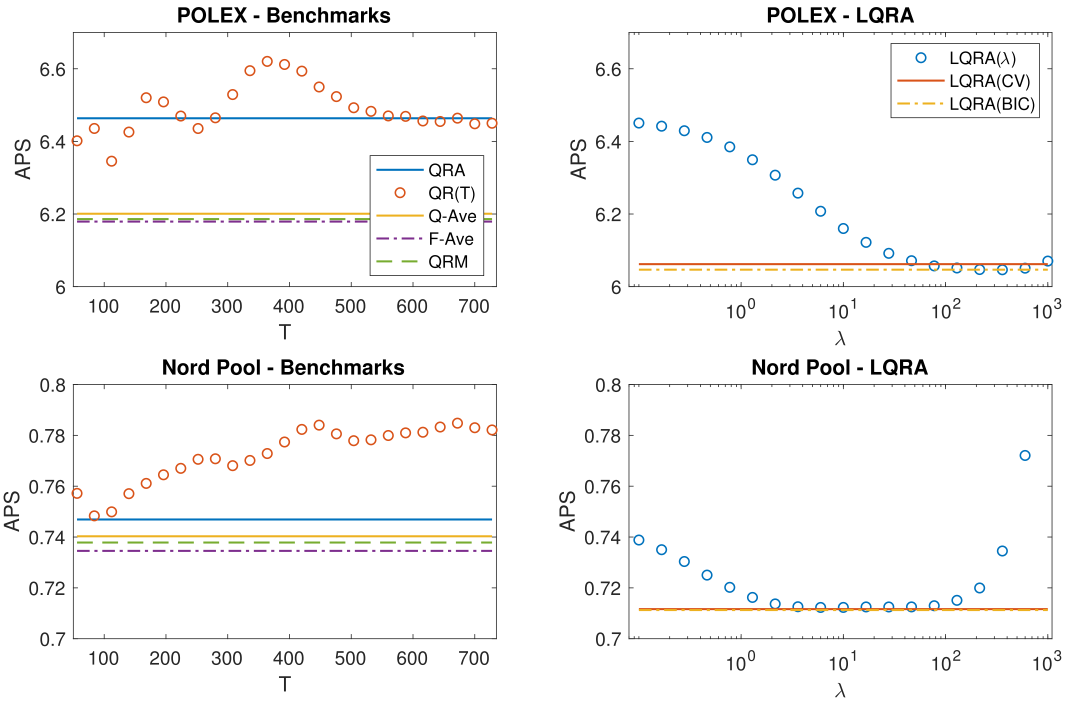


Fig. 6. The Aggregate Pinball Score (APS) for the POLEX dataset from 29 September 2016 to 31 December 2019 (*top panels*) and the Nord Pool dataset from 30 December 2014 to 31 December 2019 (*bottom panels*), plotted separately for the benchmark (*left panels*) and **LQRA** models (*right panels*).

(77), for which λ was selected *ex-ante*. Given that **LQRA(BIC)** involves a significantly lower computational cost than **LQRA(CV)**, it can be recommended based on the results of the Kupiec test.

4.2.4. Sharpness

In Fig. 6 we depict the Aggregate Pinball Score (APS) defined in Section 4.2.1, i.e., the PS averaged across all 99 percentiles and all days and hours in the out-of-sample test period for probabilistic forecasts. Like in Section 4.2.2, we present the results for all benchmarks and all considered values of λ for **LQRA**. We can observe that:

- Both for the Polish and Nordic markets the best performing benchmark is **F-Ave** with an APS of 6.18 (for POLEX) and 0.73 (for NP).
- For the POLEX dataset, all results obtained for **LQRA** with $\lambda > 10$ outperform the benchmarks by a large margin. The APS as function of λ has a 'minimum' at 6.05, ca. 2.1% lower compared to the APS of the best benchmark.
- For the Nordic market we can also conclude that most of the forecasts obtained for **LQRA** outperform the benchmarks by a large margin. The APS as a function of λ takes its 'minimum' at 0.71, around 3% lower compared to the best benchmark.
- Both for the Polish and Nordic markets **LQRA(BIC)** and **LQRA(CV)** perform comparably well, with the former having a slight edge for the POLEX dataset. Both are significantly better than all benchmarks and essentially at par with **LQRA(77)**, i.e., with the *ex-post* selected λ.

4.2.5. Testing conditional predictive ability

As in Section 4.2.3, we perform the CPA test of Giacomini and White (2006) for three benchmarks (**Q-Ave**, **F-Ave**, **QRM**) and three LQRA models – **LQRA(77)** with an *ex-post* selected λ and two that utilize automated selection of the tuning parameter, i.e., **LQRA(BIC)** and **LQRA(CV)**. In Fig. 7 we illustrate the obtained *p*-values using 'chessboards'

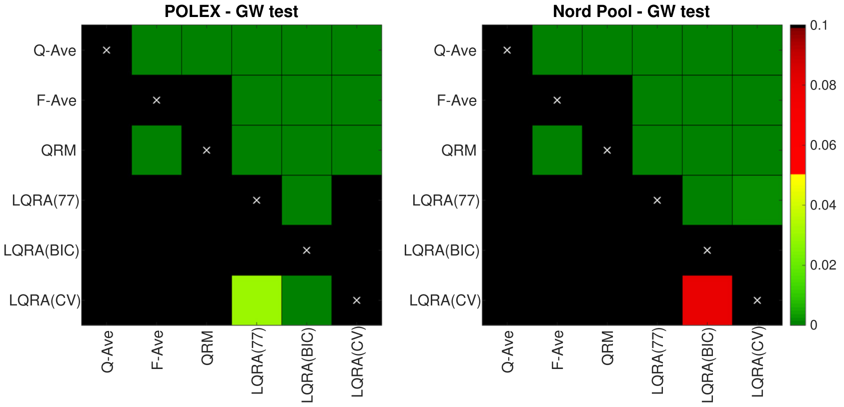


Fig. 7. Results of the conditional predictive ability (CPA) test of Giacomini and White (2006) for forecasts of selected models for the POLEX (*left*) and Nord Pool (*right*) datasets. We use a heat map to indicate the range of the *p*-values – the closer they are to zero (→ dark green) the more significant 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).

(analogously as in Ziel and Weron, 2018; Serafin et al., 2019; Uniejewski et al., 2019), i.e., we use a heat map to indicate the range of the p-values – the closer they are to zero (→ dark green) 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). Clearly, the CPA test results confirm and emphasize the observations made in Section 4.2.4. In particular, **F-Ave** significantly outperforms all other benchmarks and for both datasets the best performing model overall is **LQRA(BIC)**, that is never significantly outperformed by any other method. Given that **LQRA(BIC)** is computationally less demanding than **LQRA(CV)**, it is the recommended option.

4.3. Measuring financial profits

Electricity price forecasts, no matter how accurate, are of limited value if they cannot be used to devise a trading strategy that yields profits. Hence, to put the above results in a different perspective, we consider strategies that can be implemented by a company that possesses an energy-storage system, like the Virtual Power Plant analyzed in Sikorski et al. (2019). Without loss of generality, let us assume that the company owns a 1.25 MWh battery, for both economic and technical reasons it cannot be discharged below 0.25 MWh (or 20% of the nominal capacity) and its efficiency is 80% for a single charge-discharge cycle. We compare a range of quantile-based, intra-day trading strategies which aim at buying 1 MWh of electricity to charge the battery when it is the cheapest (typically very early morning hours), storing it and discharging the unit to sell 0.8 MWh, i.e., 80% of 1 MWh, of electricity when it is the most expensive (typically in the afternoon). We use data from the Polish day-ahead (see Section 2) and balancing (source: www.pse.pl) markets to illustrate the strategies and evaluate the financial value of different electricity price forecasts.

4.3.1. Quantile-based trading strategies

All considered quantile-based trading strategies consist of the same three steps. In step (i), before bidding in the day-ahead market on day $d-1$, we select two hours – one with a low and one with a high expected price of electricity on day d. More precisely, the ‘low-price hour’ h_1 and the ‘high-price hour’ h_2 are chosen such that $h_1 < h_2$ and the price spread between the α-quantile of the price distribution for hour h_2 and the $(1-\alpha)$-quantile for hour h_1 is maximized, see Fig. 8 for an illustration.

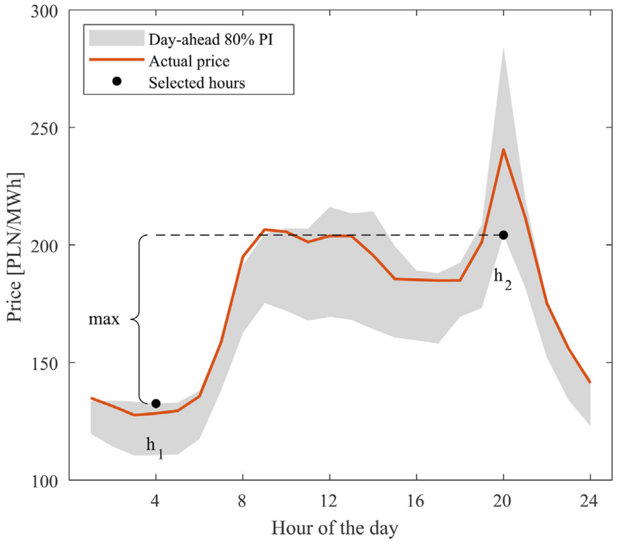


Fig. 8. Illustration of the quantile-based trading strategies discussed in Section 4.3.1. When bidding for 29 September 2017, based on the 80% PIs available a day earlier, the price spread is maximized for $\widehat{P}_{d,4}^{90\%} = 132.51$ (left •) and $\widehat{P}_{d,20}^{10\%} = 204.23$ PLN/MWh (right •).

We consider six levels of α (1%, 5%, 10%, 15%, 20% and 25%) and formulate the above as a linear integer programming problem with:

- decision variables x_{sell}^{h} and x_{buy}^{h} equal to one if we respectively sell or buy at hour h and zero otherwise, satisfying $x_{buy}^{h*} + \sum_{h=1}^{h*} x_{sell}^{h} \le 1$ for $h^* = 1, 2, ..., 24$;
- and objective function:

$$\max \sum_{h=1}^{24} \left(0.8\, \widehat{P}_{d,h}^{\alpha} \cdot x_{sell}^{h} - \widehat{P}_{d,h}^{1-\alpha} \cdot x_{buy}^{h} \right), \tag{14}$$

where $\widehat{P}_{d,h}^{\alpha}$ is the model-predicted α-quantile of the price distribution on day d and hour h. We use the *intlinprog* function from the Matlab Optimization Toolbox to solve it.

In step (ii) we submit to POLEX the bid to buy 1 MWh for $\widehat{P}_{d,h_1}^{1-\alpha}$ at hour h_1 and simultaneously the offer to sell 0.8 MWh at $\widehat{P}_{d,h_2}^{\alpha}$ at hour h_2. For instance, when bidding for 29 September 2017 with $\alpha = 10\%$ the price spread is maximized for $\widehat{P}_{d,4}^{90\%} = 132.51$ and $\widehat{P}_{d,20}^{10\%} = 204.23$ PLN/MWh, see Fig. 8. Note, that the probability that each of our bids will be accepted is $1-\alpha$ (independently); in this case it is 90%, both for the bid to buy at $h_1 = 4$ and the offer to sell at $h_2 = 20$. Finally, in step (iii):

- If both the bid and the offer are accepted, on day d and hour h_1 we buy 1 MWh in the day-ahead market for the auction settled price $P_{d,h1}$ and charge the battery. Then at hour $h_2 > h_1$ we discharge the battery and sell 0.8 MWh in the day-ahead market at $P_{d,h2}$. The daily profit is $0.8\, P_{d,h2} - P_{d,h1}$ PLN.
- If both the bid and the offer are rejected, we do nothing; the daily profit is 0 PLN.
- If only the offer is accepted, on day d and hour h_1 we buy 1 MWh in the balancing market for $B_{d,h1}$ and charge the battery. Then at hour $h_2 > h_1$ we discharge the battery and sell 0.8 MWh in the day-ahead market. The daily profit is $0.8\, P_{d,h2} - B_{d,h1}$ PLN.
- Analogously, if only the bid is accepted, on day d and hour h_1 we buy 1 MWh in the day-ahead market and charge the battery. Then at hour $h_2 > h_1$ we discharge the battery and sell 0.8 MWh in balancing market at $B_{d,h2}$. The daily profit is $0.8\, B_{d,h2} - P_{d,h1}$ PLN.

4.3.2. Benchmark strategies

To show that the probabilistic forecasts are worth considering we compare the quantile-based strategies defined in Section 4.3.1 against two benchmarks. The first one is a **naive** strategy that always buys 1 MWh at hour 4 am and sells 0.8 MWh at 12 pm; those are the hours with the – on average – lowest and highest price in the Polish day-ahead market. The second, **point forecasts-based** benchmark is constructed analogously to the quantile-based strategies, but instead of using quantile forecasts $\widehat{P}_{d,h}^{\alpha}$ it utilizes point forecasts $\widehat{P}_{d,h}$.

4.3.3. Results

In Table 2 we report the profits (in PLN) obtained when using one of the benchmark or quantile-based strategies defined in Sections 4.3.1–4.3.2 in the POLEX market for the whole out-of-sample, 853-day period, see Fig. 1. We can draw the following conclusions:

- For any considered quantile-based strategy the obtained profit is significantly higher than for the naive or point forecasts-based benchmarks.
- For all considered α’s the highest profits are obtained for one of the **LQRA**-based probabilistic forecasts; see the underlined values in each column.
- **LQRA(BIC)** with $\alpha = 20\%$ yields the highest profit overall. The value of 45,396.21 PLN for 853 days is equivalent to an average daily profit of 53.22 PLN (or ca. 12 EUR).

Table 2
Profits (in PLN) obtained when using one of the benchmark or quantile-based strategies defined in Sections 4.3.1–4.3.2 in the POLEX market for the whole out-of-sample, 853-day period, see Fig. 1. The highest profits in each column are underlined, the best overall result is emphasized in bold.

Strategy	Profit					
Naive (4 am–12 pm)	33,065.29					
Point forecasts-based	37,722.39					
Quantile-based	*1–99%*	*5–95%*	*10–90%*	*15–85%*	*20–80%*	*25–75%*
Q-Ave	41,317.92	43,328.89	43,432.31	43,793.54	43,289.09	43,033.88
F-Ave	39,848.26	43,369.44	44,052.04	44,071.39	44,088.11	43,130.34
QRM	41,163.29	43,054.28	43,124.12	43,197.42	43,731.54	42,240.25
LQRA(77)	42,360.05	44,135.49	44,320.03	44,713.52	44,684.40	43,624.65
LQRA(BIC)	42,886.10	43,993.23	44,502.81	43,836.79	**45,396.21**	42,741.57
LQRA(CV)	41,693.80	43,971.88	44,238.45	44,213.62	45,073.19	43,103.88

5. Conclusions

In this paper we have introduced a new technique for computing probabilistic forecasts – LASSO QRA or LQRA for short. It is a regularized variant of Quantile Regression Averaging, which utilizes the Least Absolute Shrinkage and Selection Operator (LASSO) to select the valuable and eliminate the redundant point forecasts used as inputs to quantile regression. Although illustrated on datasets comprising day-ahead electricity prices, we believe that LQRA can be a useful approach also for other forecasting applications, in particular, in risk management.

We have evaluated the new technique using datasets from the Polish and Nordic power markets, and compared against nearly 30 benchmarks. Furthermore, we have introduced two automated techniques for *ex-ante* selection of the regularization parameter – **LQRA(BIC)** that uses the Bayesian Information Criterion to select λ based on the in-sample fit, and a more computationally demanding procedure which utilizes cross-validation. We provide evidence for the superior predictive performance of **LQRA(BIC)** in terms of the Kupiec test for (unconditional) coverage, the pinball score and the test for conditional predictive accuracy (CPA), as well as financial profits for a range of trading strategies. Overall, we recommend **LQRA(BIC)** as an accurate and computationally efficient method for computing probabilistic forecasts that can be used to boost the profitability of energy trading activities, help with bidding in day-ahead markets and improve risk management practices in the power sector.

Acknowledgments

This work was partially supported by the Ministry of Science and Higher Education (MNiSW, Poland) through Diamond Grant No. 0199/DIA/2019/48 (to BU) and the National Science Center (NCN, Poland) through MAESTRO grant No. 2018/30/A/HS4/00444 (to RW).

Appendix A. Supplementary data

Supplementary data and Matlab codes can be found online at https://doi.org/10.1016/j.eneco.2021.105121.

References

Afanasyev, D., Fedorova, E., 2019. On the impact of outlier filtering on the electricity price forecasting accuracy. Appl. Energy 236, 196–210.

Agoua, X., Girard, R., Kariniotakis, G., 2019. Probabilistic models for spatio-temporal photovoltaic power forecasting. IEEE Transactions on Sustainable Energy 10, 780–789.

Bayer, S., 2018. Combining value-at-risk forecasts using penalized quantile regressions. Econometrics and Statistics 8, 56–77.

Ben Bouallègue, Z., 2017. Statistical postprocessing of ensemble global radiation forecasts with penalized quantile regression. Meteorol. Z. 26, 253–264.

Bracale, A., Carpinelli, G., Falco, P.D., 2019. Developing and comparing different strategies for combining probabilistic photovoltaic power forecasts in an ensemble method. Energies 12, 1–16.

Bunn, D., Andresen, A., Chen, D., Westgaard, S., 2016. Analysis and forecasting of electricity price risks with quantile factor models. Energy Journal 37, 101–122.

Bunn, D., Gianfreda, A., Kermer, S., 2018. A trading-based evaluation of density forecasts in a real-time electricity market. Energies 11, 2658.

Dantzig, G.B., Orden, A., Wolfe, P., et al., 1955. The generalized simplex method for minimizing a linear form under linear inequality restraints. Pac. J. Math. 5, 183–195.

Diaz, G., Planas, E., 2016. A note on the normalization of Spanish electricity spot prices. IEEE Trans. Power Syst. 31, 2499–2500.

Gaillard, P., Goude, Y., Nedellec, R., 2016. Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. Int. J. Forecast. 32, 1038–1050.

Giacomini, R., White, H., 2006. Tests of conditional predictive ability. Econometrica 74, 1545–1578.

Gneiting, T., 2011. Quantiles as optimal point forecasts. Int. J. Forecast. 27, 197–207.

Gneiting, T., Katzfuss, M., 2014. Probabilistic forecasting. Annual Review of Statistics and Its Application 1, 125–151.

Grossi, L., Nan, F., 2019. Robust forecasting of electricity prices: simulations, models and the impact of renewable sources. Technological Forecasting & Social Change 141, 305–318.

Hagfors, L., Bunn, D., Kristoffersen, E., Staver, T., Westgaard, S., 2016a. Modeling the UK electricity price distributions using quantile regression. Energy 102, 231–243.

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

Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press.

He, Y., Qin, Y., Lei, X., Feng, N., 2019. A study on short-term power load probability density forecasting considering wind power effects. International Journal of Electrical Power and Energy Systems 113, 502–514.

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. https://doi.org/10.1109/OAJPE.2020.3029979.

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.

Janczura, J., Trück, S., Weron, R., Wolff, R., 2013. Identifying spikes and seasonal components in electricity spot price data: a guide to robust modeling. Energy Econ. 38, 96–110.

Kath, C., Ziel, F., 2020. Conformal prediction interval estimation and applications to day-ahead and intraday power markets. Int. J. Forecast. https://doi.org/10.1016/j.ijforecast.2020.09.006 In press.

Koenker, R.W., 2005. Quantile Regression. Cambridge University Press.

Kostrzewski, M., Kostrzewska, J., 2019. Probabilistic electricity price forecasting with Bayesian stochastic volatility models. Energy Econ. 80, 610–620.

Kupiec, P.H., 1995. Techniques for verifying the accuracy of risk measurement models. The Journal of Derivatives 3, 73–84.

Laio, F., Tamea, S., 2007. Verification tools for probabilistic forecasts of continuous hydrological variables. Hydrol. Earth Syst. Sci. Discuss. 11, 1267–1277.

Lebotsa, M.E., Sigauke, C., Bere, A., Fildes, R., Boylan, J.E., 2018. Short term electricity demand forecasting using partially linear additive quantile regression with an application to the unit commitment problem. Appl. Energy 222, 104–118.

Lee, E.R., Noh, H., Park, B.U., 2014. Model selection via Bayesian information criterion for quantile regression models. J. Am. Stat. Assoc. 109, 216–229.

Li, Y., Zhu, J., 2008. L1-norm quantile regression. J. Comput. Graph. Stat. 17, 163–185.

Lichtendahl, K.C., Grushka-Cockayne, Y., Winkler, R.L., 2013. Is it better to average probabilities or quantiles? Manag. Sci. 59, 1594–1611.

Lisi, F., Pelagatti, M., 2018. Component estimation for electricity market data: deterministic or stochastic? Energy Econ. 74, 13–37.

Liu, B., Nowotarski, J., Hong, T., Weron, R., 2017. Probabilistic load forecasting via Quantile regression averaging on sister forecasts. IEEE Transactions on Smart Grid 8, 730–737.

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

Maciejowska, K., Nowotarski, J., 2016. A hybrid model for GEFCom2014 probabilistic electricity price forecasting. Int. J. Forecast. 32, 1051–1056.

Maciejowska, K., Nowotarski, J., Weron, R., 2016. Probabilistic forecasting of electricity spot prices using factor quantile regression averaging. Int. J. Forecast. 32, 957–965.

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

Manzan, S., 2015. Forecasting the distribution of economic variables in a data-rich environment. J. Bus. Econ. Stat. 33, 144–164.

Marcjasz, G., 2020. Forecasting electricity prices using deep neural networks: a robust hyper-parameter selection scheme. Energies 13 (18), 4605.

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

Marcjasz, G., Uniejewski, B., Weron, R., 2020. Probabilistic electricity price forecasting with NARX networks: combine point or probabilistic forecasts? Int. J. Forecast. 36, 466–479.

Misiorek, A., Trück, S., Weron, R., 2006. Point and interval forecasting of spot electricity prices: linear vs. non-linear time series models. Studies in Nonlinear Dynamics & Econometrics 10 (Article 2).

Mpfumali, P., Sigauke, C., Bere, A., Mulaudzi, S., 2019. Day ahead hourly global horizontal irradiance forecasting – application to south African data. Energies 12, 3569.

Narajewski, M., Ziel, F., 2020. Econometric modelling and forecasting of intraday electricity prices. J. Commod. Mark. 19, 100107.

Nowotarski, J., Weron, R., 2015. Computing electricity spot price prediction intervals using quantile regression and forecast averaging. Comput. Stat. 30, 791–803.

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

Schneider, S., 2011. Power spot price models with negative prices. Journal of Energy Markets 4, 77–102.

Serafin, T., Uniejewski, B., Weron, R., 2019. Averaging predictive distributions across calibration windows for day-ahead electricity price forecasting. Energies 12, 256.

Sikorski, T., Jasinski, M., Ropuszynska-Surma, E., Weglarz, M., Kaczorowska, D., Kostyla, P., Leonowicz, Z., Lis, R., Rezmer, J., Rojewski, W., Sobierajski, M., Szymanda, J., Bejmert, D., Janik, P., 2019. A case study on distributed energy resources and energy-storage systems in a virtual power plant concept: economic aspects. Energies 12, 4447.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society B 58, 267–288.

Tikhonov, A.N., 1963. Solution of incorrectly formulated problems and the regularization method. Soviet Mathematics Doklady 4, 1035–1038.

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., Weron, R., Ziel, F., 2018. Variance stabilizing transformations for electricity spot price forecasting. IEEE Trans. Power Syst. 33, 2219–2229.

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 Econ. 79, 171–182.

Wang, Y., Zhang, N., Tan, Y., Hong, T., Kirschen, D., Kang, C., 2019. Combining probabilistic load forecasts. IEEE Transactions on Smart Grid 10, 3664–3674.

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.

Zhang, Y., Liu, K., Qin, L., An, X., 2016. Deterministic and probabilistic interval prediction for short-term wind power generation based on variational mode decomposition and machine learning methods. Energy Convers. Manag. 112, 208–219.

Zhang, W., Quan, H., Srinivasan, D., 2018. Parallel and reliable probabilistic load forecasting via quantile regression forest and quantile determination. Energy 160, 810–819.

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

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

Forecasting Electricity Prices

Katarzyna Maciejowska, Bartosz Uniejewski, Rafał Weron

Forecasting Electricity Prices

Katarzyna Maciejowska, Bartosz Uniejewski, Rafał Weron

Department of Operations Research and Business Intelligence,
Wrocław University of Science and Technology, 50-370 Wrocław, Poland

Summary

Forecasting electricity prices is a challenging task and an active area of research since the 1990s and the deregulation of the traditionally monopolistic and government-controlled power sectors. It is interdisciplinary by nature. It requires expertise in econometrics, statistics or machine learning for developing well-performing predictive models, in finance for understanding market mechanics, and in electrical engineering for comprehension of the fundamentals driving electricity prices.

Although electricity price forecasting aims at predicting both spot and forward prices, the vast majority of research is focused on short-term horizons which exhibit dynamics unlike in any other market. The reason is that power system stability calls for a constant balance between production and consumption, while being weather (both demand and supply) and business activity (demand only) dependent. The recent market innovations do not help in this respect. The rapid expansion of intermittent renewable energy sources is not offset by the costly increase of electricity storage capacities and modernization of the grid infrastructure.

On the methodological side, this leads to three visible trends in electricity price forecasting research. Firstly, there is a slow, but more noticeable with every year, tendency to consider not only point but also probabilistic (interval, density) or even path (also called ensemble) forecasts. Secondly, there is a clear shift from the relatively parsimonious econometric (or statistical) models towards more complex and harder to comprehend, but more versatile and eventually more accurate statistical/machine learning approaches. Thirdly, statistical error measures are regarded as only the first evaluation step. Since they may not necessarily reflect the economic value of reducing prediction errors, more and more often, they are complemented by case studies comparing profits from scheduling or trading strategies based on price forecasts obtained from different models.

Keywords: price forecasting, electricity market, quantile regression, probabilistic forecasting, statistical learning, deep learning, forecast evaluation, economic value, trading strategy

Introduction

Electricity price forecasting (EPF)[1] as a research area of its own appeared in the early 1990s with the liberalization and deregulation of the power sectors in the UK and Scandinavia. The late 1990s and 2000s were marked by the widespread conversion from the traditionally monopolistic and government-controlled power sectors to competitive power markets in Europe, North America, Australia and eventually in Asia (Mayer and Trück, 2018). Over the years, EPFs

[1] Here, EPF refers to both electricity price forecasting and electricity price forecast(s). The plural form, i.e., forecasts, is abbreviated EPFs.

have become a fundamental input to companies' decision-making mechanisms (Weron, 2014). As Hong (2015) estimates, for a medium-sized utility with a 5-gigawatt annual peak load[2], improving the day-ahead demand forecasts by 1% leads to annual savings of ca. 1.5 million USD. With the additional price forecasts, the savings double. Clearly, the time invested in developing EPF models can pay off.

For newcomers to this research area it is important to realize that the literature has generally focused on horizons of up to 48 hours, since short-term price dynamics is what makes electricity special. In the longer term, prices are averaged across weekly, monthly or annual delivery periods and lose much of their uniqueness. In the short-term, on the other hand, electricity prices exhibit significant seasonality at different levels (daily, weekly and in many markets also annual), short-lived and generally unanticipated price spikes (ranging up to two orders of magnitude), and in some markets even negative values.

However, the "short-term" is not a particular horizon, but a whole spectrum of horizons ranging from a few minutes ahead (real-time, *intraday*, ID; also called "spot" in North America) to a *day-ahead* (DA; called "spot" in Europe). Note that from a financial perspective, both the ID and DA contracts can be regarded as very short-term forwards, with delivery during a particular hourly (half- or quarter-hourly) load period on the same or the next day. Since each day can be divided into a finite number of load periods $h = 1, 2, ..., H$ with $H = 24, 48$ or 96, it is common to use double indexing when referring to the electricity price. Here, the price for day d and load period h is denoted by $P_{d,h}$, its prediction by $\widehat{P}_{d,h}$, and its predictive distribution by $\widehat{F}_P$ or $\widehat{F}_{P_{d,h}}$.

To address the structural changes initiated by market liberalization and deregulation on one hand and the rapid expansion of intermittent renewable energy sources on the other, the EPF literature has evolved over the years. The relatively parsimonious linear regression and neural network models of the 2000s and 2010s have been gradually replaced by techniques able to cope with the increasing complexity of the data and inflated expectations of the market participants. Three trends are clearly visible in the 2020s:

#1: increasing popularity of *probabilistic* (interval, density) and *path* (also called *ensemble*) forecasts,

#2: a visible shift towards *statistical/machine learning* (SL/ML), and

#3: evaluating the economic value of price predictions.

A detailed discussion of these trends follows a brief description of the marketplace and the typical forecasting tasks considered.

The Marketplace

As a result of the aforementioned liberalization and deregulation of the power sectors, two basic models for power markets have emerged: power pools – where trading, dispatch and transmission are managed by the *system operator* (SO), and power exchanges – where trading and initial dispatch are managed by an institution independent from the *transmission system*

[2] The annual peak load is the highest electrical power demand in a (calendar) year. The power consumed or generated is measured in multiples of the watt (W). Smaller power plants can generate tens of megawatts (1 MW = 10^6 W), the largest tens of gigawatts (1 GW = 10^9 W). The amount of electricity consumed or generated over a specific period of time is typically measured in megawatt-hours (MWh); it is also the basic unit used in trading electricity.

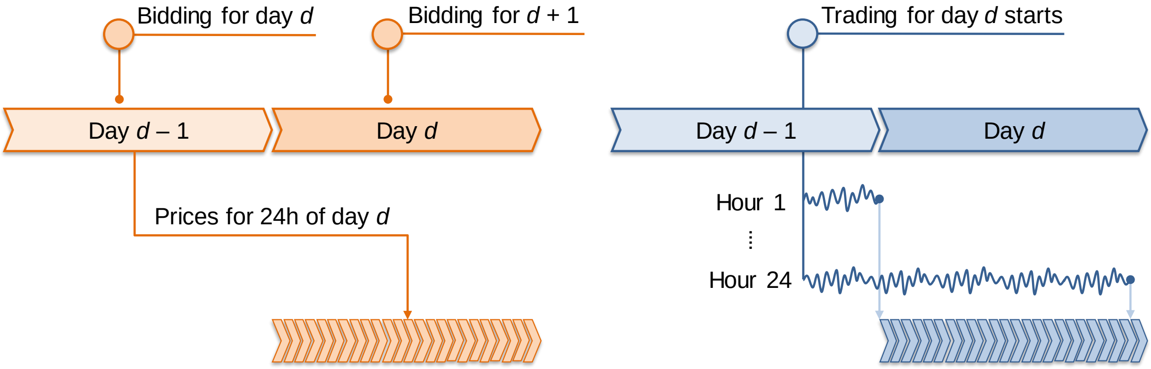


Figure 1: Illustration of bidding and price settlement in auction (*left*) and continuous trading (*right*) power markets. In day-ahead auctions the bids for all load periods (here: hours) of day d can be submitted until a certain hour on day $d-1$. Intraday markets which admit continuous trading run 24/7 from an afternoon hour on day $d-1$ up until a few minutes before the delivery on day d.

operator (TSO). Participation in power pools is limited to generators and is typically mandatory. The *market clearing price* (MCP) is established through a one-sided auction as the intersection of the supply curve constructed from aggregated supply bids of the generators and the demand predicted by the system operator. Often a separate price for each node in the network is calculated, so-called *locational marginal price* (LMP). Such a system was adopted in highly meshed North American networks. On the other hand, in Australia, where the network structure is simpler, zonal pricing was successfully implemented, where for areas without grid limitations a unique price is settled.

In contrast to power pools, participation in power exchanges is – except for some special cases – voluntary and open not only to generators, but also to wholesale consumers and speculators. The price is established either through a two-sided auction (DA, ID) as the intersection of the supply curve constructed from aggregated supply bids and the demand curve constructed from aggregated demand bids or in continuous trading (ID). Most market designs have adopted the uniform-price auction, where buyers who bid at or above the MCP pay that price and sellers who bid at or below the MCP are paid this price. Moreover, in auction markets the bids can be submitted until a certain time – called *gate closure* – which is the same for all load periods, see the left panel in Fig. 1. Hence, auction prices could be viewed as realizations of a multivariate random variable and therefore prices for all load periods should be predicted simultaneously (Ziel and Weron, 2018). On the other hand, some ID markets allow for continuous trading. They run 24/7 from an afternoon hour on day $d-1$ up until a few minutes before the delivery of electricity during a particular load period on day d, see the right panel in Fig. 1.

In some countries (e.g., Germany, Ireland, Poland) the DA and ID markets are complemented by the so-called balancing market. This technical market is used for pricing differences between the market schedule and actual system demand for very short time horizons before delivery. For instance, the TSO might instruct a generator to increase its output to meet a sudden surge in demand. The producer then receives a premium via the balancing market for the energy generated used to balance the grid.

The timeline of day-ahead and intraday trading activities in selected European countries is

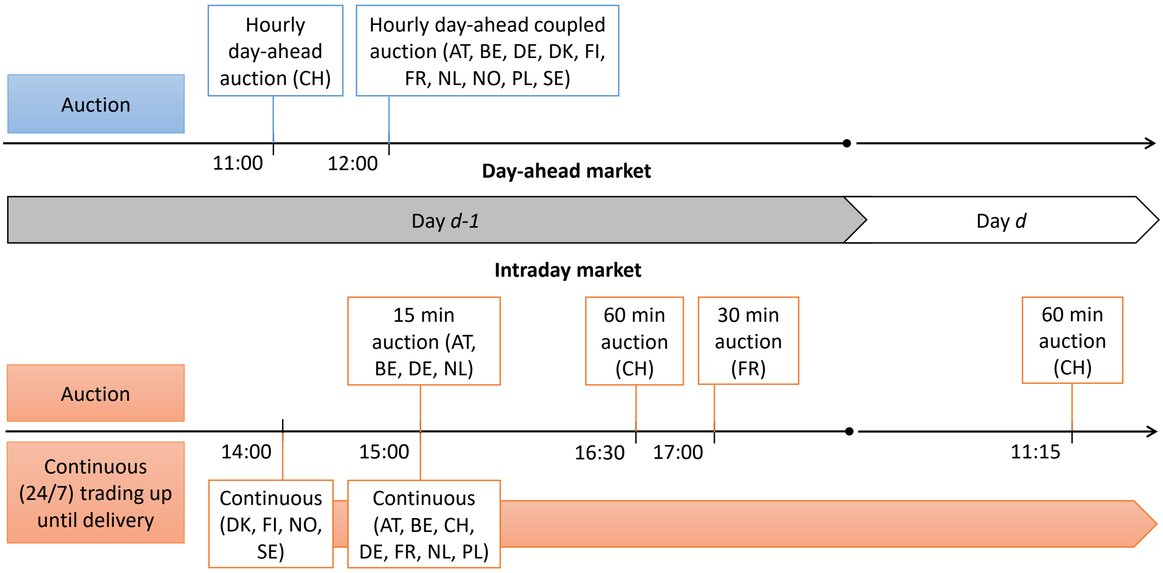

Hourly day-ahead auction (CH)
Hourly day-ahead coupled auction (AT, BE, DE, DK, FI, FR, NL, NO, PL, SE)
Auction
11:00
12:00
Day-ahead market
Day d-1
Day d
Intraday market
15 min auction (AT, BE, DE, NL)
60 min auction (CH)
30 min auction (FR)
60 min auction (CH)
Auction
14:00
15:00
16:30
17:00
11:15
Continuous (24/7) trading up until delivery
Continuous (DK, FI, NO, SE)
Continuous (AT, BE, CH, DE, FR, NL, PL)

Figure 2: The timeline of day-ahead (*top*) and intraday (*bottom*) trading activities for delivery of electricity on day d in selected European countries: Austria (AT), Belgium (BE), Denmark (DK), Germany (DE), Finland (FI), France (FR), the Netherlands (NL), Norway (NO), Poland (PL), Sweden (SE) and Switzerland (CH).

illustrated in Fig. 2. As can be seen, the DA and ID markets complement each other. Once the gate closes for day-ahead bids around noon, various intraday markets open for adjusting these bids. They are particularly important for nondispatchable, stochastic producers such as wind or solar farms, and include both auctions and continuous trading. Note that both the ID and DA contracts can concern delivery during the same load period, only the time the decision has to be made and the bid placed differs.

The presented sequence of events has important implications for study design. In the DA market the forecasting horizons typically range from 12-14 hours for the first load period of the next day to 36-38 hours for the last. However, at the time the predictions are made, i.e., the morning hours of day $d-1$, the DA prices for all load periods of this day are already known (they were settled around noon on day $d-2$). Generally, the TSO day-ahead forecasts of the system load ($\approx$ demand) and the system-wide generation from *renewable energy sources* (RES) are also available to market participants at this time.

When the ID market is considered, the selection of the forecasting horizon depends on the research question. Firstly, the predictions can be made on the morning of day $d-1$, when market participants need to decide how much electricity to bid in the DA market and how much to buy/sell in the ID market or leave for the balancing market. Forecasts of the price spread between DA and ID/balancing markets can provide valuable insights for decision-making (Maciejowska et al., 2019, 2021).

Secondly, the predictions can be used for bidding in ID markets with continuous trading. Although the trading floor opens in the afternoon hours of day $d-1$, the majority of bids are placed during the last 3-4 hours before the delivery (Narajewski and Ziel, 2020a). Hence, the forecasting horizons considered typically range from a couple of minutes to 4 hours (Janke and Steinke, 2019; Uniejewski et al., 2019b; Narajewski and Ziel, 2020b). Note that different model specifications may be optimal for predicting ID prices for different horizons (Maciejowska et al., 2020). Since the bidding behavior of market participants is significantly influenced by RES generation forecasts which are available at the time of trading (Kiesel and Paraschiv, 2017; Kulakov and Ziel, 2021), ID price forecasts should not only exploit the short-term price dependencies but also updated predictions of wind and solar power generation. Interestingly, including self-exciting terms in ID models allows to better capture the empirically observed trade clustering (Kramer and Kiesel, 2021).

Trend #1: From Point to Probabilistic and Ensemble Forecasts

By far point forecasts are the most popular. Despite a few early attempts, often inspired by developments in wind forecasting (Hong et al., 2020), probabilistic forecasting was not part of the mainstream EPF literature until 2014 and the Global Energy Forecasting Competition (GEFCom2014; Hong et al., 2016). Probabilistic EPF quickly gained momentum and energy analysts have become aware of its importance in energy systems planning and operations. A variety of approaches have been considered, including bootstrapping (Chen et al., 2012), Quantile Regression Averaging (QRA; Nowotarski and Weron, 2015), Bayesian statistics (Kostrzewski and Kostrzewska, 2019) and deep learning (Mashlakov et al., 2021; Jędrzejewski et al., 2022). Nevertheless, no more than 15% of Scopus-indexed articles concern interval or distributional EPF. Path (also called *ensemble*) forecasts, which focus on the multidimensional temporal distribution, are even less popular. Yet, path-dependency is crucial for many optimization problems arising in power plant scheduling, energy storage and trading, and this has been recognized in the recent EPF literature (Janke and Steinke, 2020; Narajewski and Ziel, 2020b).

Error and Price Distributions

There are two main approaches to probabilistic forecasting: the more elegant one directly considers the distribution of the electricity price, while the more popular one builds on the point forecast and the distribution of errors associated with it. In both cases, the focus can be on prediction intervals, selected quantiles or the whole predictive distribution. For reviews on short- and medium-term probabilistic EPF see Nowotarski and Weron (2018) and Ziel and Steinert (2018), respectively, while for a general treatment – the seminal review of Gneiting and Katzfuss (2014).

Given the current information set and assuming that the point forecast is the expected price[3] at a future time point, i.e, $\widehat{P}_{d,h} = \mathbb{E}(P_{d,h})$, the price can be written as[4]:

$$P_{d,h} = \widehat{P}_{d,h} + \varepsilon_{d,h}, \tag{1}$$

and the distribution F_ε of errors associated with $\widehat{P}_{d,h}$ is identical to the distribution F_P of prices, except for a horizontal shift by $\widehat{P}_{d,h}$:

$$F_P(x) \equiv \mathbb{P}(P_{d,h} \le x) = \mathbb{P}(\widehat{P}_{d,h} + \varepsilon_{d,h} \le x) = \mathbb{P}(\varepsilon_{d,h} \le x - \widehat{P}_{d,h}) \equiv F_\varepsilon(x - \widehat{P}_{d,h}).$$

[3] Although this is the most common assumption, the point forecast does not have to be the expected value. For instance, it can be the median or any quantile of the predictive distribution.

[4] For notational brevity conditioning is not used here. More formally, in a typical day-ahead setup, forecasts for all load periods of day d are made at hour h_0 in the morning of day $d-1$, using information available at that time, see Fig 1. In this case, $\widehat{P}_{d,h} \equiv \widehat{P}_{d,h|d-1,h_0}$.

This, however, implies that the inverse empirical *cumulative distribution functions* (also called *quantile functions*) satisfy:

$$\widehat{F}_P^{-1}(\alpha) = \widehat{P}_{d,h} + \widehat{F}_\varepsilon^{-1}(\alpha), \tag{2}$$

i.e., they are identical except for a shift by $\widehat{P}_{d,h}$, but this time on the vertical axis. Equation (2) provides the basic framework for constructing probabilistic forecasts from prediction errors. If a dense grid of quantiles is considered, e.g., 99 percentiles, then $\widehat{F}_P$ can be approximated pretty well (Hong et al., 2016; Nowotarski and Weron, 2018; Uniejewski and Weron, 2021).

If F_P has a density f_P, then a density forecast $\widehat{f_P}$ can be provided as well. However, Ziel and Steinert (2016) argue against using such an approach. Analyzing the fine structure of aggregated supply and demand curves in the German market they found that F_P was multimodal with significant jumps (corresponding to point masses) at certain 'round' prices.

Quantile Regression Averaging

Quantile regression (QR; see Koenker, 2005) is one of the most popular methods for directly modeling the distribution of a random variable. QR approximates the target quantile with a linear function of a set of explanatory variables. In the EPF context, these variables typically contain publicly available market information (load forecasts, generation structure, historical electricity prices, etc.; Bunn et al., 2016; Maciejowska, 2020) and/or point predictions of electricity prices (Weron, 2014). The later case leads to the so-called *Quantile Regression Averaging* (QRA) introduced by Nowotarski and Weron (2015) and originally developed for Team Poland's participation in the GEFCom2014 competition (Maciejowska and Nowotarski, 2016). It is a forecast combination approach to the computation of quantile forecasts, which bridges the gap between point and probabilistic forecasts. QRA involves applying QR to the point forecasts of a small number of individual forecasting models or experts:

$$q(\alpha|\boldsymbol{X}_{d,h}) = \boldsymbol{X}_{d,h}\boldsymbol{\beta}_\alpha, \tag{3}$$

where $q(\alpha|\cdot)$ is the conditional αth quantile of F_P, $\boldsymbol{X}_{d,h}$ is the vector of point forecasts and $\boldsymbol{\beta}_\alpha$ is the corresponding vector of weights. The latter is estimated by minimizing the following sum of *check functions*:

$$\widehat{\boldsymbol{\beta}}_\alpha = \underset{\boldsymbol{\beta}_\alpha}{\operatorname{argmin}}\Big\{ \textstyle\sum_{d,h} \underbrace{\left(\alpha - \mathbb{1}_{P_{d,h}<\boldsymbol{X}_{d,h}\boldsymbol{\beta}_\alpha}\right)\left(P_{d,h} - \boldsymbol{X}_{d,h}\boldsymbol{\beta}_\alpha\right)}_{\text{check function}} \Big\}. \tag{4}$$

The very good forecasting performance of QRA has been verified by a number of authors, not only in the area of EPF (Liu et al., 2017; Kostrzewski and Kostrzewska, 2019; Kath and Ziel, 2021; Uniejewski and Weron, 2021). However, its most spectacular success came during the GEFCom2014 competition, when teams using variants of QRA (Gaillard et al., 2016; Maciejowska and Nowotarski, 2016) were ranked in the top two places in the price track.

Paths and Ensembles

Although the concept of probabilistic EPF is much more general than of point forecasting, it is not sufficient to support operational decisions that depend on future trajectories of electricity prices. For instance, in Germany renewable energy producers can receive less subsidies if the electricity price is negative for 6 hours in a row. Hence, instead of looking at the 24 hourly univariate price distributions $F_{P_{d,1}}, ..., F_{P_{d,24}}$, the focus should be on the multidimensional distribution $\boldsymbol{F_P}$ of the 24-dimensional price vector $\boldsymbol{P_d} = (P_{d,1}, \ldots, P_{d,24})'$. However, many models considered in the literature cannot output such a multidimensional forecast.

Ensemble forecasts provide a practical remedy. An ensemble is a collection of simulated price *paths*, also called *trajectories* or *scenarios*. For a large number of paths the ensemble approximates the underlying distribution $\boldsymbol{F_P}$ arbitrarily well (Weron and Ziel, 2020). In practice 'large' means thousands or millions of paths, which may be a computational challenge (Narajewski and Ziel, 2020b). It should be noted that, on one hand, the same or similar concepts have been used in different disciplines under different names, e.g., *simultaneous prediction intervals*, *prediction bands*, *spatio-temporal trajectories*, *numerical weather prediction ensembles*. On the other, the term ensemble is also used to refer to any averaging of – point or probabilistic – forecasts (Hong et al., 2020).

Trend #2: From Regression to Statistical and Machine Learning

Until the mid 2010s, the EPF literature was dominated by relatively parsimonious linear regression and neural network models. They were characterized by a small number – a dozen or two – of *inputs* (also called *features*, *input features*, *explanatory variables*, *regressors*, or *predictors*) and complex data pre-/post-processing:

- replacing outliers, i.e., price spikes, by more 'normal' values before estimating the model (Contreras et al., 2003; Bierbrauer et al., 2007; Janczura et al., 2013) or utilizing robust estimation methods (Grossi and Nan, 2019);
- averaging forecasts, both across models (Nan, 2009; Bordignon et al., 2013; Nowotarski et al., 2014) and across calibration windows for the same model (Marcjasz et al., 2018; Hubicka et al., 2019), also in a probabilistic EPF setting (Serafin et al., 2019);
- using so-called variance stabilizing transformations (VSTs; Schneider, 2011; Diaz and Planas, 2016; Uniejewski et al., 2018; Narajewski and Ziel, 2020a; Shi et al., 2021) to make the marginal distributions less heavy-tailed (Box-Cox family, area hyperbolic sine) or Gaussian (Probability Integral Transform);
- deseazonalizing the data with respect to the long-term seasonal component (LTSC) before estimating the model (Janczura et al., 2013; Nowotarski and Weron, 2016; Lisi and Pelagatti, 2018; Afanasyev and Fedorova, 2019; Uniejewski et al., 2019a; Marcjasz et al., 2020).

However, as more data and computational power became available, the models became more complex to the extent that expert knowledge was no longer enough to handle them (Jędrzejewski et al., 2022). This paved the way for statistical/machine learning in EPF. Arguably, *statistical learning* (SL) and *machine learning* (ML) are synonyms.[5] They have just originated in different communities – computer science/artificial intelligence (Mitchell, 1997) or computational statistics (James et al., 2021). Both SL and ML refer to a vast set of (computational, statistical) tools for understanding data, both can improve "automatically" through training. In either case, learning can be supervised or unsupervised. In EPF, the authors are typically interested

[5] Januschowski et al. (2020) even argue that the distinction between *statistical* and *machine learning* forecasting is dubious, as this distinction does not stem from fundamental differences in the methods assigned to either class, but rather is of a "tribal" nature.

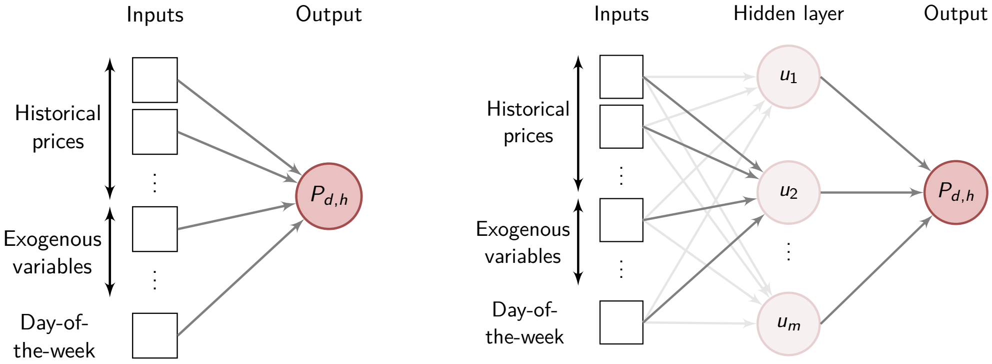


Figure 3: Visualization of a linear regression model (*left*) and a shallow neural network (*right*) with identical inputs and output, i.e., the electricity price for day d and hour h. White squares represent the inputs and $u_1, u_2, ..., u_m$ the hidden nodes (or neurons). Arrows indicate the flow of information.

in supervised learning, which involves building a model for predicting a known output or outputs based on a set of inputs.

The Expert Benchmark

A class of commonly used EPF benchmarks is based on a parsimonious *autoregressive* (AR) structure with exogenous variables and calendar effects, originally proposed by Misiorek et al. (2006). Since expert knowledge is used to select the regressors, such benchmarks are often called *expert* models (Ziel and Weron, 2018). One of the most popular structures represents the electricity price for day d and hour h by:

$$\begin{aligned} P_{d,h} \;=\; & \underbrace{\beta_1 P_{d-1,h} + \beta_2 P_{d-2,h} + \beta_3 P_{d-7,h}}_{\text{autoregressive effects}} + \underbrace{\beta_4 P_{d-1,24}}_{\text{end-of-day}} + \underbrace{\beta_5 P_{d-1}^{max} + \beta_6 P_{d-1}^{min}}_{\text{non-linear effects}} \\ & + \underbrace{\beta_7 X_{d,h}^1 + \beta_8 X_{d,h}^2}_{\text{exogenous variables}} + \underbrace{\sum\nolimits_{j=1}^{7} \beta_{h,j+8} D_j}_{\text{weekday dummies}} + \varepsilon_{d,h}, \end{aligned} \tag{5}$$

where $P_{d-1,h}$, $P_{d-2,h}$ and $P_{d-7,h}$ account for the autoregressive effects and correspond to prices from the same hour h of the previous day, two days before and a week before, $P_{d-1,24}$ is the last known price at the time the prediction is made and provides information about the end-of-day price level, P_{d-1}^{max} and P_{d-1}^{min} represent previous day's maximum and minimum prices, $X_{d,h}^1$ and $X_{d,h}^2$ are exogenous variables, $D_1, ..., D_7$ are weekday dummies and $\varepsilon_{d,h}$ is the noise term (i.i.d. variables with finite variance). The β_i's are estimated using *ordinary least squares* (OLS).

Autoregression or more generally linear regression is one of the two most commonly used classes of EPF models (Weron, 2014). The other is the *multi-layer perceptron* (MLP). The simplest neural network, a single-layer perceptron, contains no hidden layers (only inputs and the output) and is equivalent to a linear regression – both represent $P_{d,h}$ by a linear combination of input features, see the left panel in Fig. 3. On the other hand, the MLP includes at least one hidden layer and utilizes a feed-forward architecture – the outputs of the nodes (or neurons) in one layer are inputs to the next one, see the right panel in Fig. 3. Since the output of a node is a weighted sum of all of the inputs transformed by a typically nonlinear activation function, unlike in linear regression, NNs can tackle complex dependence structures encountered in power market data (Keles et al., 2016).

Exogenous variables typically include day-ahead predictions of the system load and RES generation (Lago et al., 2021). Days with high demand and low RES generation are characterized by relatively high prices. On the other hand, high RES generation pulls prices down; in periods of low demand – holidays and/or at night – even below zero (Zhou et al., 2016). Other fundamental variables may possess explanatory power as well. For instance, fuel and CO_2 allowance prices, especially for medium-term EPF (Maciejowska and Weron, 2016; Ziel and Steinert, 2018). Due to the merit order effect, i.e., dispatching units characterized by the lowest marginal cost of production, the fuel–electricity price relationship changes throughout the day. Natural gas prices impact mainly the peak hours, whereas coal prices influence the off-peak hours. Finally, the day-of-the-week input feature visible in Fig. 3 can be a set of weekday dummies, as in Eq. (5), or a single multi-valued variable, which is more common in NN models.

Regularization and the LEAR Model

Selecting regressors is a cumbersome task and expert knowledge does not always identify the relevant ones. In a series of papers in the mid 2010s, Ludwig et al. (2015), Ziel et al. (2015), Gaillard et al. (2016), Uniejewski et al. (2016) and Ziel (2016) introduced the concept of regularization to EPF. In simple terms, the idea behind this approach is to add a penalty term to the *residual sum of squares* (RSS) in OLS regression:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \mathrm{RSS} + \lambda \sum_{i=1}^{n} |\beta_i|^q \right\}, \tag{6}$$

where λ is the *tuning* or *regularization* hyperparameter. Note, that *hyperparameters* are model parameters that cannot be optimized during the training (estimation) phase, but have to be set or calibrated beforehand, e.g., using cross-validation (James et al., 2021). For $q = 2$ Equation (6) yields *ridge regression* (Hoerl and Kennard, 1970) and for $q = 1$ the *least absolute shrinkage and selection operator* (LASSO; Tibshirani, 1996). The latter can shrink β_i's not only towards zero but actually to zero itself, thus effectively eliminating some regressors from the model. If both terms, i.e., $\lambda_1 \sum |\beta_i| + \lambda_2 \sum \beta_i^2$, are admitted, then Eq. (6) yields the so-called *elastic net* (Zou and Hastie, 2015).

In the EPF setting, all three variants were compared in Uniejewski et al. (2016). Ridge regression easily outperformed expert models and stepwise regression techniques, but was significantly worse than the LASSO and the elastic net. At the cost of an additional parameter, the elastic net generally yields more accurate predictions than the LASSO. Nevertheless, the latter has become the golden standard in EPF (Uniejewski and Weron, 2018; Ziel and Weron, 2018; Janke and Steinke, 2019; Narajewski and Ziel, 2020a; Marcjasz, 2020; Zhang et al., 2020; Özen and Yıldırım, 2021). It was even utilized by Lago et al. (2021) to construct a well-performing EPF benchmark – the *LASSO-Estimated AutoRegressive* (LEAR) model. The starting point for

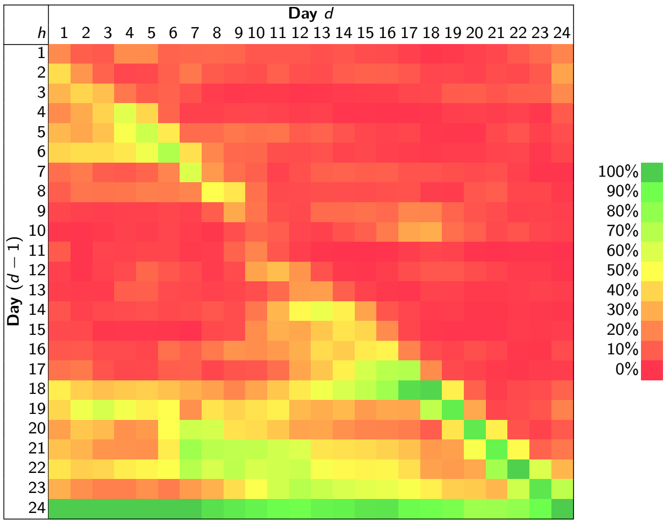


Figure 4: Mean occurrence of non-zero $\beta_{h,i}$'s across datasets from 12 power markets. Columns represent the predicted hours on day d and rows the first 24 variables, i.e., $\sum_{i=1}^{24} \beta_{h,i} P_{d-1,i}$ in Eq. (7), of a LEAR-type model considered by Ziel and Weron (2018). A heat map is used to indicate more (→ green) and less (→ red) commonly-selected variables.

the LEAR model is a parameter-rich regression:

$$\begin{aligned} P_{d,h} = \sum_{i=1}^{24} & (\beta_{h,i} P_{d-1,i} + \beta_{h,i+24} P_{d-2,i} + \beta_{h,i+48} P_{d-3,i} + \beta_{h,i+72} P_{d-7,i}) \\ & + \sum_{i=1}^{24} \left(\beta_{h,i+96} X^1_{d,i} + \beta_{h,i+120} X^1_{d-1,i} + \beta_{h,i+144} X^1_{d-7,i} \right) \\ & + \sum_{i=1}^{24} \left(\beta_{h,i+168} X^2_{d,i} + \beta_{h,i+192} X^2_{d-1,i} + \beta_{h,i+216} X^2_{d-7,i} \right) \\ & + \sum_{k=1}^{7} \beta_{h,240+k} D_k + \varepsilon_{d,h}, \end{aligned} \tag{7}$$

which differs from the expert model in Eq. (5) mainly by allowing for cross-hourly dependencies. In general, the price for hour h may depend on the prices for all 24 hours yesterday, the day before, etc. In practice, only a dozen or two of the potential 247 regressors turn out to be relevant. However, they need not be the ones included in the expert model.

This is visualized in Figure 4 for the first 24 variables, i.e., $\sum_{i=1}^{24} \beta_{h,i} P_{d-1,i}$, of a LEAR-type model considered by Ziel and Weron (2018) and across datasets from 12 power markets.[6] The yellow-green diagonal indicates that the price for hour h on day $d-1$ is a good predictor of the price for the same hour on day d. The yellow-green bottom rows were a surprising finding

[6] BELPEX price for Belgium, EPEX prices for Switzerland, Germany–Austria and France, EXAA price for Germany–Austria, GEFCom2014 competition data, Nord Pool prices for West Denmark, East Denmark and the system price, OMIE prices for Spain and Portugal, and OTE price for the Czech Republic. The GEFCom2014 dataset covers a 3-year period (2011–2013; see Hong et al., 2016), the remaining datasets a 6-year period (July 2010 – July 2016; see Ziel and Weron, 2018).

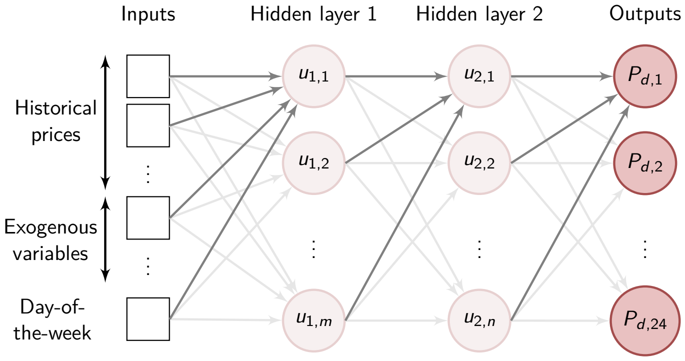


Figure 5: Visualization of the DNN model of Lago et al. (2018), i.e., a feed-forward neural network with two hidden layers and 24 outputs. Like in Figure 3, white squares represent the inputs, $u_{1,1}, ..., u_{1,m}, u_{2,1}, ..., u_{2,n}$ the hidden nodes, and arrows indicate the flow of information.

at the time Ziel and Weron (2018) published their paper. They simply mean that late evening prices for day $d - 1$ and particularly the last known price, i.e., $P_{d-1,24}$, are good predictors for all hours of the next day. Since then, terms like $\beta_4 P_{d-1,24}$ in Eq. (5) have been added to expert models. Interestingly, the performance of LEAR-type models can be further improved by deseazonalizing the data with respect to the long-term seasonal component (LTSC) before estimation (Jędrzejewski et al., 2021), just like in the case of parsimonious regression (Nowotarski and Weron, 2016) and neural network models (Marcjasz et al., 2020).

Deep Learning and the DNN Model

Starting in the mid 2010s, the EPF research shifted towards models with a larger number of inputs and automatic feature engineering, like the LEAR, and architectures that employ *deep learning* (DL) to obtain better hidden data representations. Both families of models are examples of a recent trend called *data-centric* ML, where emphasis is not put on the model, but on input data quality and consistency. Both families use SL/ML methods as means to increase the number of (potential) input features and to reduce the need for human interaction during feature engineering and data processing. The difference is that the second family uses deep architectures, e.g., neural networks with more than one hidden layer (see Goodfellow et al., 2016, for an excellent introduction to DL).

Deep learning EPF models can be traced back to Wang et al. (2017), who proposed an architecture built on stacked *denoising autoencoders* that take a partially corrupted (or noisy) input and are trained to recover the original undistorted input. The DL models that followed were primarily based on the MPL with features modeled as hyperparameters. The most prominent example is probably the DNN model of Lago et al. (2018) that has been shown to improve upon parameter-rich linear regression models estimated via the LASSO.

The DNN is a feed-forward network with two hidden layers of m and n nodes, and 24 outputs, i.e., it jointly predicts 24 hourly prices $P_{d,1}, ..., P_{d,24}$, see Fig. 5. Its hyperparameters and input features are optimized using the tree-structured Parzen estimator (Bergstra et al., 2011). This is achieved by modeling the features as hyperparameters, with each hyperparameter representing a binary variable that selects whether or not a specific feature is included in the model. Other hyperparameters include the number of neurons per layer, the activation function, the dropout rate, the learning rate, etc. In practice, the model structure can be quite large, Lago et al. (2018)

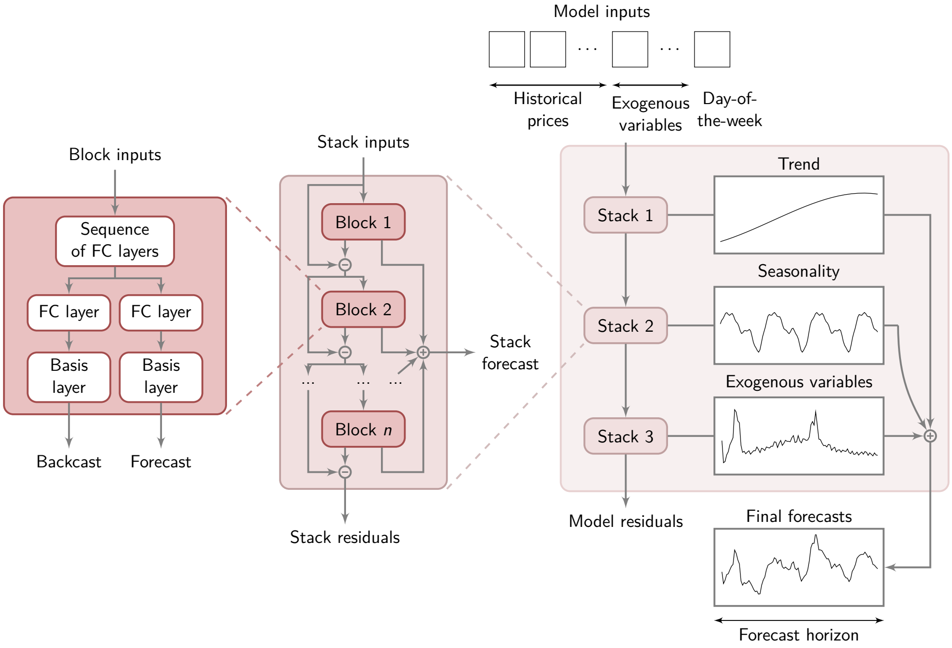


Figure 6: Visualization of the NBEATSx model of Olivares et al. (2022).

report the optimal values in their study of the Belgian market to be $m = 239$ and $n = 162$.

Given the optimal hyperparameters and features, the DNN is recalibrated on a daily basis to provide next day's electricity price forecasts. Although not strictly mandatory, periodic (e.g., monthly) recalibration of features and hyperparameters can be beneficial. The starting set of input features is the same as for the LEAR model in Eq. (7), with the only difference that, for the sake of simplicity, the day-of-the-week is modeled with a multi-valued variable, not a set of 7 dummies (Lago et al., 2021). The open-source Python codes for the DNN (and the LEAR) model are available from GitHub (`https://epftoolbox.readthedocs.io`).

Interpretability and the NBEATSx Model

The architecture of the *neural basis expansion analysis for time series* (NBEATS) model introduced by Oreshkin et al. (2020) has the ability to structurally decompose signals making the outputs easily interpretable. A feature whose absence has made it difficult to apply neural networks in many contexts. Moreover, it has demonstrated state-of-the-art performance on multiple large-scale datasets, including those used in the M4 competition (Makridakis et al., 2020), and it is computationally efficient exhibiting a linear cost with respect to the input size. Recently, it has been successfully applied in mid-term electricity load forecasting (Oreshkin et al., 2021).

In general, the decomposition in the NBEATS model is performed by projecting the objective time series onto basis functions in the fundamental blocks of the network structure. Each fundamental block (the dark red rectangles labeled "block 1", ..., "block n" in Fig. 6) consists of two parts: (i) a sequence of fully-connected layers (FC) ended with a fork that returns estimated backward and forward expansion coefficients, and (ii) the backward and forward basis layers that map these coefficients via the basis functions onto two block outputs called the backcast and the forecast. The former is the best estimate of the block inputs given the functional space used in the considered block, whereas the latter is the partial prediction that contributes to the final forecast.

The blocks are lined up so that the backcast of each is removed from its inputs, and the residuals are passed to the following block as new inputs. Such a residual recursion is performed consecutively over all blocks in the network. The block forecasts, on the other hand, are summed up to produce the final prediction. The NBEATS architecture groups blocks into stacks that specialize in different types of basis functions. Separate stacks can account for the trend and seasonality by modelling these functions as polynomials and harmonic functions, respectively. Consequently, the final forecasts can be decomposed into interpretable components returned by individual stacks.

The NBEATSx model introduced by Olivares et al. (2022) adds to this structure a stack (light red rectangle labeled "stack 3" in Fig. 6) that performs the projection onto exogenous variables. Such an exogenous stack helps to predict the effects induced by holidays and fundamentals (like electric load or RES generation forecasts), and is crucial for EPF. While Olivares et al. report no significant differences between the NBEATS model and the *exponential smoothing recurrent neural network* (ESRNN) of Smyl (2020) that has excelled in the M4 competition, the NBEATSx architecture improves over NBEATS by nearly 20% and up to 5% over the LEAR and DNN models.

The hyperparameters and the input features are optimized in the same way as for the DNN model. However, compared to the DNN, the hyperparameter list also includes: the type and the number of stacks, the number of blocks per stack, the degree of trend polynomials, and the number of Fourier bases. The optimization algorithm also selects the best-performing order of stacks. Open-source Python codes are available from PythonRepo (`https://pythonrepo.com/repo/cchallu-nbeatsx-python-deep-learning`).

Trend #3: From Statistical to Economic Evaluation

Over the years a number of authors have criticized the exclusive use of statistical error measures to evaluate and compare forecasts. However, a standardized test ground/procedure for evaluating the economic impact of predictions has not been developed, not only in EPF (Hong et al., 2020), but in forecasting in general (Petropoulos et al., 2022). And this, despite the fact that already three decades ago Murphy (1993) postulated that the "goodness" of a forecast can be assessed in terms of consistency, quality, and value.

While *quality* can be readily quantified by commonly used error metrics, the other two characteristics require an explanation. As Murphy (1993) defines it, *consistency* refers to the correspondence between forecasters' internal, i.e., recorded only in the forecaster's mind, judgments and their forecasts. Since such judgments are, by definition, unavailable to others, consistency cannot be assessed directly. Yet, some authors explicitly mention using expert knowledge to ex-post correct the results from a statistical or a ML model. For instance, Maciejowska and Nowotarski (2016) 'manually' expanded or tightened the PIs in their top performing GEFCom2014 competition approach.

The third characteristic, i.e., *value*, refers to the (incremental) economic and/or other benefits to decision makers from using the predictions. For instance, it may reflect additional revenue resulting from improved forecasts or reduced uncertainty as measured by revenue volatility. As Yardley and Petropoulos (2021) argue, it is a construct that not only incorporates considerations of the utility to the forecaster, which is discussed in subsection *Economic Measures*, but also the computational and opportunity costs. While numerous papers report them, the computational costs are rarely used to compare different methods. One of a few exceptions is an article by

Nikolopoulos and Petropoulos (2018), who study the trade-off between optimal versus suboptimal (but less costly) solutions and find that choosing the latter does not necessarily reduce forecast accuracy. Finally, the opportunity costs reflect the resources wasted on implementing a complex method that eventually is not used, because the decision-makers do not have confidence in a model they do not understand (Green and Armstrong, 2015). Yet, both cited papers do not concern EPF.

Statistical Error Measures

Point Forecasts

The most commonly used error metrics for point forecasts include the *mean absolute error* (MAE) and the *root mean squared error* (RMSE), typically across all $H = 24$ hours (48 half-hours or 96 quarter-hours) in the test period:

$$\text{MAE} = \frac{1}{DH}\sum_{d=1}^{D}\sum_{h=1}^{H}|\widehat{\varepsilon}_{d,h}|, \qquad \text{RMSE} = \sqrt{\frac{1}{DH}\sum_{d=1}^{D}\sum_{h=1}^{H}\widehat{\varepsilon}_{d,h}^{\,2}}, \tag{8}$$

where $\widehat{\varepsilon}_{d,h} = P_{d,h} - \widehat{P}_{d,h}$ and D is the number of days in the test period. It is advised to report both absolute and squared errors, especially if regression and neural network models are compared. The reason is that regression-type models are typically estimated using OLS or its variants, as in Eq. (6), while NNs are often trained by minimizing absolute errors (Lago et al., 2018; Smyl, 2020; Olivares et al., 2022).

Both MAE and RMSE are scale-dependent and hence hard to compare across different datasets. The often used in other forecasting contexts *mean absolute percentage error* (MAPE) and its “symmetric” variant (sMAPE; see, e.g., Makridakis et al., 2020) are sensitive to values close to zero and may lead to absurd results in EPF. Hyndman and Koehler (2006) advocate using the *mean absolute scaled error* (MASE) which is simply the MAE in Eq. (8) scaled by the in-sample MAE of a naive[7] forecast. However, the MASE is not recommended for comparisons of models using different calibration windows, since for each model it will be based on a different scaling factor. Instead, Lago et al. (2021) recommend using relative measures. For instance, the *relative* MAE (rMAE) which normalizes the MAE by the out-of-sample (not in-sample) MAE of a naive forecast.

The significance of differences in EPF accuracy is usually evaluated using the Diebold and Mariano (1995) test for (*unconditional*) *predictive ability* or its generalization – the Giacomini and White (2006) test for *conditional predictive ability.* Both tests can be used for nested and non-nested models, as long as the calibration window does not grow with the sample size, but only the latter accounts for parameter estimation uncertainty. However, energy forecasters are not restricted to these two tests, there is a plethora of available approaches (for a review see, e.g., Section 2.12.6 in Petropoulos et al., 2022).

The Diebold-Mariano (DM) test is an asymptotic z-test of the hypothesis that the mean of the loss differential series is zero. It is based upon the observation that the DM statistic:

$$\text{DM} = \sqrt{DH}\frac{\hat{\mu}}{\hat{\sigma}}, \tag{9}$$

[7] E.g., a random walk forecast. Note that for seasonal time series of period τ, the time lag should be equal to τ. For instance, in EPF it is common to take $\widehat{P}_{d,h}^{naive} = P_{d-7,h}$.

is asymptotically standard normal under the assumption of covariance stationarity of the *loss differential* series:

$$\Delta_{d,h} = L_{1,d,h} - L_{2,d,h}, \tag{10}$$

where $L_{i,d,h}$ is the *score* or *loss function* of model i for day d and load period (e.g., hour) h, while $\hat{\mu}$ and $\hat{\sigma}$ are respectively the sample mean and standard deviation of $\Delta_{d,h}$. Covariance stationarity may not be satisfied by forecasts in day-ahead electricity markets, since the H predictions for the next day are made at the same time, using the same information set. Hence, either H independent tests (one for each load period of the day; Bordignon et al., 2013; Nowotarski et al., 2014; Uniejewski et al., 2016; Lago et al., 2018; Gianfreda et al., 2020) or a multivariate variant proposed by Ziel and Weron (2018) are performed (Uniejewski et al., 2018; Hubicka et al., 2019; Marcjasz et al., 2019; Maciejowska et al., 2021; Özen and Yıldırım, 2021). The latter jointly tests forecasting accuracy across all H load periods using the 'daily' or 'multivariate' loss differential series:

$$\Delta_d = ||\varepsilon_{1,d}||_p - ||\varepsilon_{2,d}||_p, \tag{11}$$

where $\varepsilon_{i,d}$ is the H-dimensional vector of prediction errors of model i for day d, $||\varepsilon_{i,d}||_p = (\sum_{h=1}^{H} |\varepsilon_{i,d,h}|^p)^{1/p}$ is the p-th norm of that vector, with $p = 1$ for absolute or 2 for squared losses.

Like in the DM test, also in the Giacomini-White (GW) test the object of interest is the loss differential series – univariate or multivariate. Tested is the null $H_0 : \boldsymbol{\phi} = 0$ in the following regression (here in the multivariate variant):

$$\Delta_d = \boldsymbol{\phi}' X_{d-1} + \epsilon_d, \tag{12}$$

where X_{d-1} contains elements from the information set on day $d-1$, i.e., a constant and lags of Δ_d, and ϵ_d is an error term. Notice that $\epsilon_d \in R$ is not the 24-dimensional vector $\varepsilon_{i,d}$ of prediction errors from Eq. (11). Sample applications of the GW test in the context of EPF include Marcjasz et al. (2018), Lago et al. (2021) and Olivares et al. (2022).

Probabilistic Forecasts

While defining error measures for point predictions is relatively straightforward, for probabilistic ones this becomes tricky. The problem is that the true price distribution F_P cannot be observed, only a single draw from it can, i.e., the observed price $P_{d,h}$. Therefore, evaluation of probabilistic forecasts relies on so-called scoring rules and the notions of reliability, sharpness and resolution. A *scoring rule* – also, as in Eq. (10), called score or loss function – assigns a numerical score $S(\widehat{F}_P, P_{d,h})$ based on the predictive distribution $\widehat{F}_P$ and the observed price. A scoring rule is (*strictly*) *proper* if it is (uniquely) optimized in expectation by the true distribution (Gneiting and Raftery, 2007). *Reliability* (also called *calibration* or *unbiasedness*) refers to the statistical consistency between $\widehat{F}_P$ and $P_{d,h}$. For instance, a 95% prediction interval (PI) is reliable if it covers exactly 95% of the observed prices. *Sharpness* refers to how concentrated is $\widehat{F}_P$. Finally, *resolution* refers to how much the predicted density varies over time. Since sharpness and resolution are equivalent when probabilistic forecasts have perfect reliability, evaluating probabilistic predictions boils down to "maximizing sharpness subject to reliability" (Gneiting and Katzfuss, 2014; Nowotarski and Weron, 2018).

The most intuitive approach to formally check the reliability of a prediction interval to compute the empirical coverage based on the indicator series of 'hits and misses' defined as: $I_{d,h} = 1$ if $P_{d,h} \in$ PI and zero otherwise. EPF studies typically report the empirical coverage itself (*PI coverage probability*, PICP) or the *average coverage error*: ACE = PICP − PINC, where PINC = α is the *PI nominal coverage*. To formally check whether $\mathbb{P}(I_{d,h} = 1) = \alpha$, i.e., the so-called *unconditional coverage* (UC), the Kupiec (1995) test can be used; it verifies whether $I_{d,h}$ is i.i.d. Bernoulli with mean α. Since the latter cannot distinguish between randomly distributed and clustered PI exceedances, Christoffersen (1998) introduced the *independence* and *conditional coverage* (CC) tests. The former is tested against a first-order Markov alternative and the latter is a joint test for independence and UC; note, that both can be run for lags larger than one (Berkowitz et al., 2011). In a continuous setting, i.e., when testing $\widehat{F}_P$, not just selected PIs, the most common approach is to use the *Probability Integral Transform*:

$$\mathrm{PIT}_{d,h} = \widehat{F}_P(P_{d,h}), \tag{13}$$

which is independent and uniformly distributed if the distributional forecast is perfect. The PIT can be assessed visually (Nowotarski and Weron, 2018) or formally evaluated using the approach of Berkowitz (2001), which jointly tests for independence and normality, i.e., for conditional coverage.

Unlike reliability, sharpness is a property of the forecasts only – the narrower the PI or the more concentrated the predictive distribution the better. Consequently, the PI width itself is a good measure of sharpness. A more elaborate approach relies on proper scoring rules, which actually assess reliability and sharpness simultaneously (Gneiting and Katzfuss, 2014). Among them, arguably the most popular is the *pinball* loss, also known as the *linlin*, *bilinear* or *newsboy* loss (Elliott and Timmermann, 2016) and has become popular in EPF after the Global Energy Forecasting (GEFCom2014) competition (Dudek, 2016; Hong et al., 2016; Maciejowska and Nowotarski, 2016). It is defined by:

$$\mathrm{pinball}^{\alpha} = \begin{cases} (1-\alpha)\left(\widehat{P}^{\alpha}_{d,h} - P_{d,h}\right), & \text{for } P_{d,h} < \widehat{P}^{\alpha}_{d,h}, \\ \alpha\left(P_{d,h} - \widehat{P}^{\alpha}_{d,h}\right), & \text{for } P_{d,h} \geq \widehat{P}^{\alpha}_{d,h}, \end{cases} \tag{14}$$

where $\widehat{P}^{\alpha}_{d,h}$ is the αth quantile of the predictive distribution for day d and load period (e.g., hour) h; note, that the pinball score is the function minimized in quantile regression, see Eq. (4). The pinball can be averaged across different quantiles, e.g., 99 percentiles, and across load periods of the target day, e.g., 24 hours, to provide the *aggregate pinball score* (APS). If the grid of quantiles is arbitrarily dense, then the average converges to the *Continuous Ranked Probability Score* (Gneiting and Raftery, 2007):

$$\mathrm{CRPS}(\widehat{F}_P, P_{d,h}) = \underbrace{\mathbb{E}|\widehat{P}_{d,h} - P_{d,h}|}_{\text{reliability}} - \underbrace{\frac{1}{2}\mathbb{E}|\widehat{P}_{d,h} - \widehat{P}^{*}_{d,h}|}_{\text{lack of sharpness}}, \tag{15}$$

where random variables $\widehat{P}_{d,h}$ and $\widehat{P}^{*}_{d,h}$ are two independent $\widehat{F}_P$-distributed copies. Probabilistic forecasts can be tested for equal predictive performance using the DM and GW tests, just like point forecasts. In this case $L_{i,d,h}$ is replaced by $S_i(\widehat{F}_P, P_{d,h})$ in Eq. (10). For sample EPF applications see, e.g., Serafin et al. (2019), Abramova and Bunn (2020), Marcjasz et al. (2020), Muniain and Ziel (2020) and Uniejewski and Weron (2021).

Path Forecasts

Compared to evaluating point or probabilistic predictions, evaluating *path* (also called *ensemble*) forecasts constitutes a challenge – it requires utilizing scoring rules for multivariate distributions

(Scheuerer and Hamill, 2015). The commonly used *Dawid-Sebastiani* and *variogram scores* are not strictly proper in the multivariate setting, while the *log-score* requires forecasts of a multivariate density, which may be not available. Hence, the recommended option is the *energy score* proposed by Gneiting and Raftery (2007), which is a generalization of the pinball and CRPS scores:

$$ES_{d,h} = \underbrace{\frac{1}{M}\sum_{i=1}^{M}||P^i_{d,h} - P_{d,h}||_2}_{\text{distance from the prices}} - \underbrace{\frac{1}{2}\frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}||P^i_{d,h} - P^j_{d,h}||_2}_{\text{distance between paths}}, \tag{16}$$

where $P^i_{d,h}$ for $i = 1, \dots, M$ is the i-th price path forecast and $||\cdot||_2$ is the Euclidean norm. When minimizing the energy score, the average distance between the simulated paths and the actual price trajectory is minimized and at the same time the average distance between the paths is maximized. Its use in EPF is limited, though, probably due to the much higher complexity of the problem (Muniain and Ziel, 2020; Narajewski and Ziel, 2020b).

Economic Measures

There are only a handful of papers which examine the economic impact of EPF errors in a more systematic manner. Interestingly, most of these studies have been published in engineering, not economic or financial journals. The likely reason is that at least a basic knowledge is needed of how power markets, loads and generating units operate. Moreover, there is no standardized test ground/procedure for evaluating the economic impact. Nearly every EPF study considers a different setup.

Supply- and Demand-Side Perspectives

In one of the earlier studies, Delarue et al. (2010) take the supply-side point of view and quantify the *profit loss* that can be expected in a price based unit commitment problem, when incorrect price forecasts are used. Simulations reveal that a combined cycle gas turbine (CCGT) is much more sensitive to EPF errors (the profit can easily lie 20% below the optimal level for a perfect price forecast) than a classic coal fired unit (profit loss rarely exceeds 10%). More interestingly, negatively biased forecasts (i.e., that predict prices lower than actual) typically yield much higher losses than positively biased predictions.

On the other hand, Zareipour et al. (2010) take the demand-side perspective and consider short-term operation scheduling of two typical loads (a process industry owning on-site generation facilities and a municipal water plant with load-shifting capabilities). They introduce the *forecast inaccuracy economic impact* index: FIEI $= [\text{cost}(\widehat{P}) - \text{cost}(P)]/\text{cost}(\widehat{P})$, so that a positive value of FIEI indicates the percentage of the actual cost of buying electricity attributable to EPF errors. The authors report that a 1% improvement in the MAPE in forecasting accuracy would result in about 0.1%–0.35% cost reductions from short-term EPF, but also conclude that the MAPE is not a good measure.

An interesting concept is considered by Doostmohammadi et al. (2017), who compute the *financial loss/gain* (FLG) time series, defined as the difference between expected profit of a generator and the actual one. Then, based on the day-ahead forecasts of the FLG series, they propose a bidding strategy. However, by doing so, they do not work with the actual profits but with (another) estimate.

Maciejowska et al. (2019, 2021) take the perspective of a small RES utility (e.g., with one wind turbine) which has to decide where to sell 1 MW of electricity during each hour of the next day – in the day-ahead (DA) or the intraday (ID) market. Conditional on the decision, summarized by the *decision variable* based on price forecasts:

$$Y_{d,h} = \begin{cases} 1 & \text{if } \widehat{P}^{DA}_{d,h} > \widehat{P}^{ID}_{d,h}, \\ 0 & \text{if } \widehat{P}^{DA}_{d,h} \le \widehat{P}^{ID}_{d,h}, \end{cases} \tag{17}$$

they compute the additional income over the benchmark, i.e., selling the production in the DA market, as:

$$\pi_{d,h} = Y_{d,h} P^{DA}_{d,h} + (1 - Y_{d,h}) P^{ID}_{d,h} - P^{DA}_{d,h}, \tag{18}$$

where $P^{DA}_{d,h}$ and $P^{ID}_{d,h}$ are the electricity prices in the DA and ID markets, respectively. While Maciejowska et al. (2019) utilize the load forecasts published by the German and Polish system operators, Maciejowska et al. (2021) additionally improve the load forecasts for Germany by applying ARX-type models. In both papers, they measure the gains from EPF as the sum of profits in the test period, $\pi = \sum_{d=1}^{D} \sum_{h=1}^{24} \pi_{d,h}$, and conclude that the statistical measures of forecast accuracy – such as the percent of correct sign classifications of the price spread between the DA and ID markets – do not necessarily coincide with economic benefits.

Trading Strategies

Uniejewski et al. (2018) take a trading perspective (different from the supply- or demand-side point of views and consider a naive spot-futures trading strategy in the German market. With a perfect day-ahead forecast the buyer could always choose the lower of the two – the day-ahead price (unknown when submitting bids) or the futures price. Since this can never be achieved in reality, the authors bias (or perturb) the 'crystal-ball' forecast and show that a 0.20 EUR/MWh decrease in the MAE from using one model instead of another would result in ca. 90,000 EUR profits, for a 1 GW baseload in 2016.

Chitsaz et al. (2018) propose a trading strategy applicable in Ontario's real-time electricity market. The energy storage operator maximizes profits with optimal scheduling. The schedule is set before the trading period begins, based on the available price forecasts and then it is updated at the end of each hour with a newer price forecasts. The authors conclude that such a strategy yields higher profits when using predictions generated by the proposed ARX model with features selected via the Mutual Information technique (Amjady et al., 2011) – 62% of the potential saving for 'crystal ball' predictions, compared with a number of other EPF approaches, e.g., using the so-called Pre-Dispatch Prices (PDPs; publicly available price predictions published by the system operator IESO) – 43% of the potential saving.

Kath and Ziel (2018) propose a multivariate elastic net model for forecasting German quarter-hourly electricity prices. They demonstrate that the "sell in the high and buy in the low market" strategy performs well, leading to substantial benefits for both a net buyer and a net seller. On the other hand, the mean-variance approach does not bring economic benefits, but yields an optimal portfolio in terms of the *Sharpe ratio*:

$$SR = \frac{\bar{\pi}}{\sigma}, \tag{19}$$

where $\bar{\pi}$ denotes the average level of an additional revenue (i.e., $\bar{\pi} = \pi/24D$; see also Eq. (18)) and σ is the standard deviation of the time series of revenues. As such, the Sharpe ratio

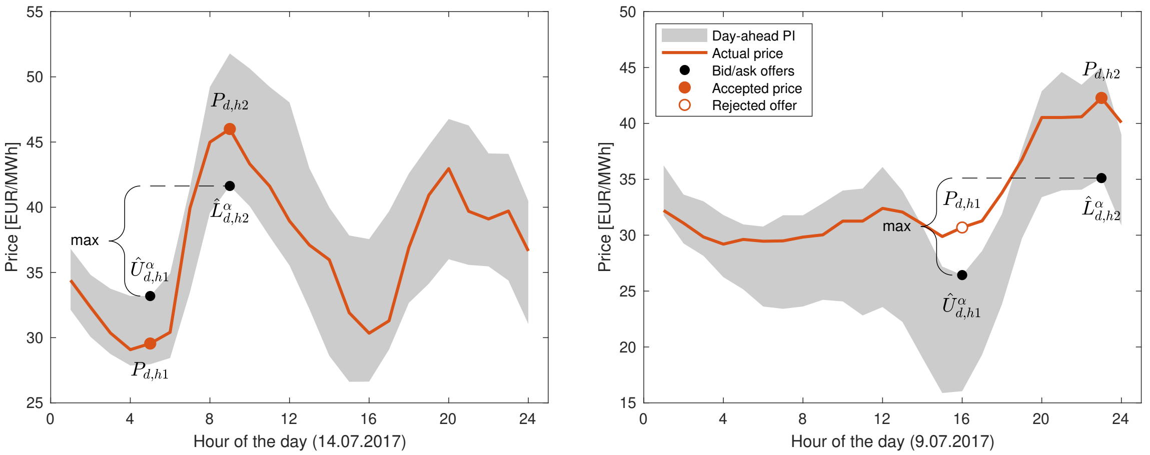


Figure 7: Illustration of the trading strategy considered by Uniejewski and Weron (2021) using German EPEX data for Friday, July 14th (*left panel*) and Sunday, July 9th (*right panel*) 2017. The day-ahead forecast of the PI is plotted in gray, the bids for the selected hours are indicated with black dots and the actual price trajectory is in orange. Note, that on July 9th the buy order is not accepted because $P_{d,h1} > \hat{U}^{\alpha}_{d,h1}$.

can be used to assess the trade-off between revenue and uncertainty. However, there are more performance measures (Eling and Schuhmacher, 2007; Auer, 2015), including measures based on drawdowns (e.g., Calmar ratio, Sterling ratio), based on partial moments (e.g., omega ratio, Sortino ratio) and based on the Value-at-Risk (VaR; e.g., excess return on VaR, conditional Sharpe ratio). Whether they will turn out to be useful in the EPF context remains yet to be checked.

Uniejewski and Weron (2021) propose a strategy for market participants having access to storage capacity. They consider a realistic setup, inspired by the Virtual Power Plant analyzed in Sikorski et al. (2019), in which the company owns a 1.25 MW battery with an efficiency of 80% per charge and discharge cycle, that cannot be discharged below 20% of the nominal capacity (i.e., 0.25 MW) due to technical limitations. The strategy is straightforward: each day buy energy and charge the battery when the price is low (generally in the early morning hours) and discharge and sell when the price is high (generally in the afternoon hours). Using probabilistic forecasts of the DA prices in the Polish market, the authors determine both the time (buy $h1$ and sell $h2$ hours) and the prices of limit orders submitted to the power exchange. They formulate and solve the following maximization problem:

$$\max_{\{h1,h2\}} \left(0.8\ \hat{L}^{\alpha}_{d,h2} - \hat{U}^{\alpha}_{d,h1}\right) \qquad \text{subject to} \qquad h1 < h2. \tag{20}$$

The optimizer selects the lowest price of a given day based on the upper quantile forecast $\hat{U}^{\alpha}_{d,h1}$ and the highest price based on the lower quantile forecast $\hat{L}^{\alpha}_{d,h2}$. The company then submits the bid to buy 1 MW for $\hat{U}^{\alpha}_{d,h1}$ at hour $h1$ and simultaneously the offer to sell 0.8 MW at $\hat{L}^{\alpha}_{d,h2}$ at hour $h2$; two sample solutions for the German EPEX market are depicted in Fig. 7. If both offers are accepted in the day-ahead market, as in the left panel of Fig. 7, the profit for a given day equals $0.8P_{d,h2} - P_{d,h1}$. However, the probability of each offer to be accepted in the market is equal to $\frac{1-\alpha}{2}$. If one of them is rejected, as in the right panel of Fig. 7 for hour $h1$, the energy has to be bought or sold in the balancing market.

This strategy is further modified by Uniejewski (2022), who restricts all trading to the day-ahead market. In such a setting, a twice larger energy storage capacity (2.5 MW) is required to trade the same volume (0.8-1 MW). The idea is to always remain in an intermediate state of the battery, for which both charging and discharging 1 MW is possible. When the bid or the ask is rejected, Uniejewski proposes to close the position the next day by submitting a market order (i.e., with no price limit). He concludes that the proposed Smoothing Quantile Regression Averaging (SQRA) approach outperforms the benchmarks in terms of statistical error metrics (Kupiec test, GW test for the pinball score) in all four considered markets (German EPEX, Scandinavian Nord Pool, Iberian OMIE, North American PJM). However, when the trading strategy is executed, SQRA forecasts lead to higher profits only in two markets (EPEX, PJM). The author hypothesizes that the poor performance for NP and OMIE is due to a twice lower average intraday price spread, i.e., the gap between the maximum and the minimum hourly price for a given day.

Further Reading

The review literature on EPF is not very rich. A couple of publications touch upon this topic, however, they usually concentrate on modeling the price dynamics for derivatives valuation and risk management purposes. In the context of day-ahead price forecasting the following are worth recommending.

In one of the first reviews, Bunn (2000) writes that "the forecasting of loads and prices are mutually intertwined activities" and that game theory and the economic perspective cannot be "an accurate basis for daily forecasts". He recommends using methods which involve variable segmentation (separate models for each load period), neural networks (for modeling the nonlinear behavior) and averaging forecasts.

The first comprehensive review and a standard reference for EPF is Weron (2014). The article not only explains the strengths and weaknesses of the available techniques, but also postulates the need for objective comparative studies and speculates on the future research directions.

The first thorough treatments of probabilistic EPF – Nowotarski and Weron (2018) and Ziel and Steinert (2018) – present much needed guidelines for the rigorous use of methods, measures and tests, in line with the paradigm of maximizing sharpness subject to reliability (Gneiting and Katzfuss, 2014). The former concentrates on short-term horizons, while the latter on mid- and long-term.

Hong et al. (2020) review energy (load, price, wind and solar generation) forecasting and discuss two challenging problems that deserve rigorous investigation – close-loop forecasting and (economic) valuation of forecasts.

Lago et al. (2021) is the first thorough review of deep learning in EPF. It also provides a set of guidelines/best practices and introduces the `epftoolbox`[8] with Python codes for two highly competitive benchmark models (LEAR, DNN).

Finally, Jędrzejewski et al. (2022) is a popular science article on the evolution of machine learning models in EPF. It is recommended for less research-oriented readers who want a light introduction to the fascinating world of electricity price forecasting.

[8] Freely available for download from: `https://epftoolbox.readthedocs.io/en/latest`.

References

Abramova, E., Bunn, D., 2020. Forecasting the intra-day spread densities of electricity prices. Energies 13, 687.

Afanasyev, D., Fedorova, E., 2019. On the impact of outlier filtering on the electricity price forecasting accuracy. Applied Energy 236, 196–210.

Amjady, N., Keynia, F., Zareipour, H., 2011. Wind power prediction by a new forecast engine composed of modified hybrid neural network and enhanced particle swarm optimization. IEEE Transactions on Sustainable Energy 2, 265–276.

Auer, B., 2015. Does the choice of performance measure influence the evaluation of commodity investments? International Review of Financial Analysis 38, 142–150.

Bergstra, J., Bardenet, R., Bengio, Y., Kégl, B., 2011. Algorithms for hyper-parameter optimization, in: Advances in Neural Information Processing Systems, pp. 2546–2554.

Berkowitz, J., 2001. Testing density forecasts, with applications to risk management. Journal of Business & Economic Statistics 19, 465–474.

Berkowitz, J., Christoffersen, P., Pelletier, D., 2011. Evaluating value-at-risk models with desk-level data. Management Science 57, 2213–2227.

Bierbrauer, M., Menn, C., Rachev, S.T., Trück, S., 2007. Spot and derivative pricing in the EEX power market. Journal of Banking & Finance 31, 3462–3485.

Bordignon, S., Bunn, D.W., Lisi, F., Nan, F., 2013. Combining day-ahead forecasts for British electricity prices. Energy Economics 35, 88–103.

Bunn, D., Andresen, A., Chen, D., Westgaard, S., 2016. Analysis and forecasting of electricity price risks with quantile factor models. Energy Journal 37, 101–122.

Bunn, D.W., 2000. Forecasting loads and prices in competitive power markets. Proceedingsof the IEEE 88, 163–169.

Chen, X., Dong, Z., Meng, K., Xu, Y., Wong, K., Ngan, H., 2012. Electricity price forecasting with extreme learning machine and bootstrapping. IEEE Transactions on Power Systems 27, 2055–2062.

Chitsaz, H., Zamani-Dehkordi, P., Zareipour, H., Parikh, P.P., 2018. Electricity price forecasting for operational scheduling of behind-the-meter storage systems. IEEE Transactions on Smart Grid 9, 6612–6622.

Christoffersen, P., 1998. Evaluating interval forecasts. International Economic Review 39, 841–862.

Contreras, J., Espínola, R., Nogales, F., Conejo, A., 2003. ARIMA models to predict next-day electricity prices. IEEE Transactions on Power Systems 18, 1014–1020.

Delarue, E., Van Den Bosch, P., D'haeseleer, W., 2010. Effect of the accuracy of price forecasting on profit in a price based unit commitment. Electric Power Systems Research 80, 1306–1313.

Diaz, G., Planas, E., 2016. A note on the normalization of Spanish electricity spot prices. IEEE Transactions on Power Systems 31, 2499–2500.

Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. Journal of Business and Economic Statistics 13, 253–263.

Doostmohammadi, A., Amjady, N., Zareipour, H., 2017. Day-ahead financial loss/gain modeling and prediction for a generation company. IEEE Transactions on Power Systems 32, 3360–3372.

Dudek, G., 2016. Multilayer perceptron for GEFCom2014 probabilistic electricity price forecasting. International Journal of Forecasting 32, 1057–1060.

Eling, M., Schuhmacher, F., 2007. Does the choice of performance measure influence the evaluation of hedge funds? Journal of Banking and Finance 31, 2632–2647.

Elliott, G., Timmermann, A., 2016. Economic Forecasting. Princeton University Press.

Gaillard, P., Goude, Y., Nedellec, R., 2016. Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. International Journal of Forecasting 32, 1038–1050.

Giacomini, R., White, H., 2006. Tests of conditional predictive ability. Econometrica 74, 1545–1578.

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., Katzfuss, M., 2014. Probabilistic forecasting. Annual Review of Statistics and Its Application 1, 125–151.

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

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press. Freely available from `http://www.deeplearningbook.org/`.

Green, K., Armstrong, J., 2015. Simple versus complex forecasting: The evidence. Journal of Business Research 68, 1678–1685.

Grossi, L., Nan, F., 2019. Robust forecasting of electricity prices: Simulations, models and the impact of renewable sources. Technological Forecasting and Social Change 141, 305–318.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67.

Hong, T., 2015. Crystal ball lessons in predictive analytics. EnergyBiz, Spring , 35–37.

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

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.

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.

Hyndman, R., Koehler, A., 2006. Another look at measures of forecast accuracy. International Journal of Forecasting 22, 679–688.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2021. An Introduction to Statistical Learning with Applications in R (2nd ed.). Springer, New York.

Janczura, J., Trück, S., Weron, R., Wolff, R., 2013. Identifying spikes and seasonal components in electricity spot price data: A guide to robust modeling. Energy Economics 38, 96–110.

Janke, T., Steinke, F., 2019. Forecasting the price distribution of continuous intraday electricity trading. Energies 12, 4262.

Janke, T., Steinke, F., 2020. Probabilistic multivariate electricity price forecasting using implicit generative ensemble post-processing, in: 2020 International Conference on Probabilistic Methods Applied to Power Systems, PMAPS 2020 - Proceedings, p. 9183687.

Januschowski, T., Gasthaus, J., Wang, Y., Salinas, D., Flunkert, V., Bohlke-Schneider, M., Callot, L., 2020. Criteria for classifying forecasting methods. International Journal of Forecasting 36, 167 – 177.

Jędrzejewski, A., Lago, J., Marcjasz, G., Weron, R., 2022. Electricity price forecasting: The dawn of machine learning. IEEE Power & Energy Magazine 20, 24–31.

Jędrzejewski, A., Marcjasz, G., Weron, R., 2021. Importance of the long-term seasonal component in day-ahead electricity price forecasting revisited: Parameter-rich models estimated via the lasso. Energies 14, 3249.

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

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.

Keles, D., Scelle, J., Paraschiv, F., Fichtner, W., 2016. Extended forecast methods for day-ahead electricity spot prices applying artificial neural networks. Applied Energy 162, 218–230.

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

Koenker, R.W., 2005. Quantile Regression. Cambridge University Press.

Kostrzewski, M., Kostrzewska, J., 2019. Probabilistic electricity price forecasting with Bayesian stochastic volatility models. Energy Economics 80, 610–620.

Kramer, A., Kiesel, R., 2021. Exogenous factors for order arrivals on the intraday electricity market. Energy Economics 97, 105186.

Kulakov, S., Ziel, F., 2021. The impact of renewable energy forecasts on intraday electricity prices. Economics of Energy and Environmental Policy 10, 79–104.

Kupiec, P.H., 1995. Techniques for verifying the accuracy of risk measurement models. The Journal of Derivatives 3, 73–84.

Lago, J., De Ridder, F., De Schutter, B., 2018. Forecasting spot electricity prices: deep learning approaches and empirical comparison of traditional algorithms. Applied Energy 221, 386–405.

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.

Lisi, F., Pelagatti, M., 2018. Component estimation for electricity market data: Deterministic or stochastic? Energy Economics 74, 13–37.

Liu, B., Nowotarski, J., Hong, T., Weron, R., 2017. Probabilistic load forecasting via Quantile Regression Averaging on sister forecasts. IEEE Transactions on Smart Grid 8, 730–737.

Ludwig, N., Feuerriegel, S., Neumann, D., 2015. Putting big data analytics to work: Feature selection for forecasting electricity prices using the LASSO and random forests. Journal of Decision Systems 24, 19–36.

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

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

Maciejowska, K., Nowotarski, J., 2016. A hybrid model for GEFCom2014 probabilistic electricity price forecasting. International Journal of Forecasting 32, 1051–1056.

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

Maciejowska, K., Weron, R., 2016. Short- and mid-term forecasting of baseload electricity prices in the U.K.: The impact of intra-day price relationships and market fundamentals. IEEE Transactions on Power Systems 31, 994–1005.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2020. The M4 competition: 100,000 time series and 61 forecasting methods. International Journal of Forecasting 36, 54–74.

Marcjasz, G., 2020. Forecasting electricity prices using deep neural networks: A robust hyper-parameter selection scheme. Energies 13, 13184605.

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

Marcjasz, G., Uniejewski, B., Weron, R., 2019. On the importance of the long-term seasonal component in day-ahead electricity price forecasting with NARX neural networks. International Journal of Forecasting 35, 1520–1532.

Marcjasz, G., Uniejewski, B., Weron, R., 2020. Probabilistic electricity price forecasting with NARX networks: Combine point or probabilistic forecasts? International Journal of Forecasting 36, 466–479.

Mashlakov, A., Kuronen, T., Lensu, L., Kaarna, A., Honkapuro, S., 2021. Assessing the performance of deep learning models for multivariate probabilistic energy forecasting. Applied Energy 285, 116405.

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

Misiorek, A., Trück, S., Weron, R., 2006. Point and interval forecasting of spot electricity prices: Linear vs. non-linear time series models. Studies in Nonlinear Dynamics & Econometrics 10, Article 2.

Mitchell, T., 1997. Machine Learning. McGraw Hill, New York.

Muniain, P., Ziel, F., 2020. Probabilistic forecasting in day-ahead electricity markets: Simulating peak and off-peak prices. International Journal of Forecasting 36, 1193–1210.

Murphy, A., 1993. What is a good forecast? an essay on the nature of goodness in weather forecasting. Weather & Forecasting 8, 281–293.

Nan, F., 2009. Forecasting next-day electricity prices: From different models to combination. URL: `http://paduaresearch.cab.unipd.it/2147`. PhD Thesis, Universita degli Studi di Padova, Italy.

Narajewski, M., Ziel, F., 2020a. Econometric modelling and forecasting of intraday electricity prices. Journal of Commodity Markets 19, 100107.

Narajewski, M., Ziel, F., 2020b. Ensemble forecasting for intraday electricity prices: Simulating trajectories. Applied Energy 279, 115801.

Nikolopoulos, K., Petropoulos, F., 2018. Forecasting for big data: Does suboptimality matter? Computers and Operations Research 98, 322–329.

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

Nowotarski, J., Weron, R., 2015. Computing electricity spot price prediction intervals using quantile regression and forecast averaging. Computational Statistics 30, 791–803.

Nowotarski, J., Weron, R., 2016. On the importance of the long-term seasonal component in day-ahead electricity price forecasting. Energy Economics 57, 228–235.

Nowotarski, J., Weron, R., 2018. Recent advances in electricity price forecasting: A review of probabilistic forecasting. Renewable and Sustainable Energy Reviews 81, 1548–1568.

Olivares, K.G., Challu, C., Marcjasz, G., Weron, R., Dubrawski, A., 2022. Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with NBEATSx. International Journal of Forecasting , forthcoming.

Oreshkin, B., Carpov, D., Chapados, N., Bengio, Y., 2020. N-BEATS: neural basis expansion analysis for interpretable time series forecasting, in: 8th International Conference on Learning Representations, ICLR 2020.

Oreshkin, B., Dudek, G., Pełka, P., Turkina, E., 2021. N-BEATS neural network for mid-term electricity load forecasting. Applied Energy 293, 116918.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., *et al.*, 2022. Forecasting: Theory and practice. International Journal of Forecasting (doi: 10.1016/j.ijforecast.2021.11.001).

Scheuerer, M., Hamill, T.M., 2015. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. Monthly Weather Review 143, 1321–1334.

Schneider, S., 2011. Power spot price models with negative prices. Journal of Energy Markets 4, 77–102.

Serafin, T., Uniejewski, B., Weron, R., 2019. Averaging predictive distributions across calibration windows for day-ahead electricity price forecasting. Energies 12, 256.

Shi, W., Wang, Y., Chen, Y., Ma, J., 2021. An effective two-stage electricity price forecasting scheme. Electric Power Systems Research 199, 107416.

Sikorski, T., Jasinski, M., Ropuszynska-Surma, E., Weglarz, M., Kaczorowska, D., Kostyla, P., Leonowicz, Z., Lis, R., Rezmer, J., Rojewski, W., Sobierajski, M., Szymanda, J., Bejmert, D., Janik, P., 2019. A case study on distributed energy resources and energy-storage systems in a virtual power plant concept: Economic aspects. Energies 12, 4447.

Smyl, S., 2020. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. International Journal of Forecasting 36, 75–85.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society B 58, 267–288.

Uniejewski, B., 2022. Smoothing quantile regression averaging: A new approach to probabilistic forecasting of electricity prices. Working paper.

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 Economics 79, 171–182.

Uniejewski, B., Marcjasz, G., Weron, R., 2019b. Understanding intraday electricity markets: Variable selection and very short-term price forecasting using lasso. International Journal of Forecasting 35, 1533–1547.

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

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

Uniejewski, B., Weron, R., 2021. Regularized quantile regression averaging for probabilistic electricity price forecasting. Energy Economics 95, 105121.

Uniejewski, B., Weron, R., Ziel, F., 2018. Variance stabilizing transformations for electricity spot price forecasting. IEEE Transactions on Power Systems 33, 2219–2229.

Wang, L., Zhang, Z., Chen, J., 2017. Short-term electricity price forecasting with stacked denoising autoencoders. IEEE Transactions on Power Systems 32, 2673–2681.

Weron, R., 2014. Electricity price forecasting: A review of the state-of-the-art with a look into the future. International Journal of Forecasting 30, 1030–1081.

Weron, R., Ziel, F., 2020. Electricity price forecasting, in: Soytas, U., Sari, R. (Eds.), Handbook of Energy Economics. Routledge, pp. 506–521.

Yardley, E., Petropoulos, F., 2021. Beyond error measures to the utility and cost of the forecasts. Foresight Q4, 36–45.

Zareipour, H., Canizares, C., Bhattacharya, K., 2010. Economic impact of electricity market price forecasting errors: A demand-side analysis. IEEE Transactions on Power Systems 25, 254–262.

Zhang, R., Li, G., Ma, Z., 2020. A deep learning based hybrid framework for day-ahead electricity price forecasting. IEEE Access 8, 143423–143436.

Zhou, Y., Scheller-Wolf, A., Secomandi, N., Smith, S., 2016. Electricity trading and negative prices: Storage vs. disposal. Management Science 62, 880–898.

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

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

Ziel, F., Steinert, R., 2018. Probabilistic mid- and long-term electricity price forecasting. Renewable and Sustainable Energy Reviews 94, 251–266.

Ziel, F., Steinert, R., Husmann, S., 2015. Efficient modeling and forecasting of electricity spot prices. Energy Economics 47, 89–111.

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

Zou, H., Hastie, T., 2015. Regularization and variable selection via the Elastic Nets. Journal of the Royal Statistical Society B 67, 301–320.

Özen, K., Yıldırım, D., 2021. Application of bagging in day-ahead electricity price forecasting and factor augmentation. Energy Economics 103, 105573.