Size Matters: Estimation Sample Length and Electricity Price Forecasting Accuracy

Carlo Fezzi[a] *and Luca Mosetti*[b]

ABSTRACT

Short-term electricity price forecasting models are typically estimated via rolling windows, i.e. by using only the most recent observations. Nonetheless, the literature does not provide guidelines on how to select the optimal size of such windows. This paper shows that determining the appropriate window prior to estimation dramatically improves forecasting performances. In addition, it proposes a simple two-step approach to choose the best performing models and window sizes. The value of this methodology is illustrated by analyzing hourly datasets from two large power markets (Nord Pool and IPEX) with a selection of eleven different forecasting models. Incidentally, our empirical application reveals that simple models, such as a simple linear regression (SLR) with only two parameters, can perform unexpectedly well if estimated on extremely short samples. Surprisingly, in the Nord Pool, such SLR is the best performing model in 13 out 24 trading periods.

Keywords: Electricity price forecasting, Day-ahead market, Parameter instability, Bandwidth selection, Statistical models, Artificial neural networks.

https://doi.org/10.5547/01956574.41.4.cfez

1. INTRODUCTION

During the past 30 years, the widespread liberalization of the energy sector has entrusted the electricity price formation process to the law of supply and demand. In most developed economies, electricity is now traded in high-frequency (hourly or half-hourly) wholesale markets, where power companies sell directly to retailers and large consumers. In this relatively new environment, developing effective short-term, day-ahead, Electricity Price Forecasting (EPF) tools can be of tremendous value (e.g. Hong, 2015).

EPF has proven to be particularly challenging. Electricity prices are characterized by a level of variability that is unobserved in any other commodity or financial asset, and peculiar dynamics such as abrupt and short-lived spikes, heteroscedasticity, and pronounced daily, weekly and yearly seasonality (e.g. Weron, 2014), which follows the dynamics of demand, often referred as "load" in this literature. Given this background, it is not surprising the recent proliferation of EPF techniques, which include statistical models (e.g. linear and non-linear regressions, time series models), machine-learning algorithms (e.g. neural networks) and various hybrid methods. Despite this significant effort, the comprehensive review by Weron (2014) concludes that a leading, best-performing methodology is yet to emerge.

a Corresponding author. Department of Economics and Management (DEM), University of Trento, and Land Use and Economics Policy (LEEP) Institute, University of Exeter Business School (UEBS). E-mail: carlo.fezzi@unitn.it.

b Department of Mathematics, ETH Zürich.

 Copyright © 2020 by the IAEE. All rights reserved.

Why is EPF so difficult? Arguably, the main cause is the substantial instability that characterizes the process of price formation. The physical laws governing the electric grid always require production and consumption to be perfectly balanced, making economically-sound storage virtually impossible. For this reason, minor changes in demand, which frequently go unnoticed, can sometimes have tremendous repercussions on prices, particularly when margin (i.e. the additional generation capacity available for production) is low. In such cases, even relatively small utilities can exercise a significant amount of market power, and influence prices substantially (e.g. Fabra and Toro, 2005; Hortaçsu and Puller, 2008; Ito and Reguant, 2016). Although margin can sometimes be predicted, unobservable determinants, such strategic behavior and asymmetric information, create an unstable price formation process that continuously evolves through time.

In forecasting, a common approach to handle time-instability is to estimate parameters using moving windows that include only the most recent observations: the so-called "rolling estimation" method (e.g. Inoue et al., 2017). While rolling estimation is the standard approach also in EPF (e.g. Dudek, 2016; Nowotarski and Weron, 2016; Steinert and Ziel, 2019), perhaps surprisingly, there is no established strategy to guide the window-size selection process. The typical approach is to simply set a relatively large window (usually between 180 and 365 observations, i.e. from 6 months to 1 year of data) *a priori* for all models and markets.[1,2]

This paper demonstrates that such stylized approach produces subpar results. Window size dramatically affect EPF models' performance, and selecting the optimal rolling sample prior to estimation significantly reduces forecasting errors. To the best of our knowledge, there are only two contributions studying this issue in EPF: Hubicka et al. (2018) and Marcjasz et al. (2018a). Both articles explore the performance of weighting schemes constructed by averaging predictions across models estimated on different window sizes. The first work analyzes the performance of a regression model and an artificial neural network under different weighting schemes, while the second article focuses on regression models and explores a larger selection of weighting techniques. Both papers conclude that using appropriate weights improves upon selecting a single window size when averaging prediction errors across all 24 trading hours of the day.

This paper develops a different and highly complementary approach. We propose a simple, two-step strategy to select both best performing models and window sizes. Our analysis includes a wide selection of models including time series, regressions and computational intelligence methods for a total of eleven different approaches. In addition, rather than evaluating the best performing model using average measures across all 24 hourly predictions within a day, we evaluate predictions for each hourly trading period separately. This finer analysis reveals that both optimal window size and best performing model change greatly across hours, with the stable off-peak hours favoring long windows and complex (i.e. with a large number of parameters) models, and the more volatile peak hours selecting short samples and relatively simple (i.e. with only a few parameters) specifications. We conclude that different models and window sizes should be used for different hours. Furthermore, our simple two-step strategy significantly outperforms the standard fixed rolling window approach in the majority of the trading periods we investigate.

1. For example, Misiorek et al. (2006), Weron and Misiorek (2008), Bordignon et al. (2013) use 9 months rolling windows, Nowotarski et al. (2014) employs 10 months windows, Maciejowska et al. (2016), Nowotarski and Weron (2016) and Marcjasz et al. (2019), Steinert and Ziel (2019) use one-year windows.

2. Another strategy for dealing with unstable environments is combining predictions from multiple models. This approach has recently found a few applications in EPF (e.g. Bordignon et al., 2013, Nowotarski et al., 2014, Mirakyan et al., 2017). However, even when combining predictions from different models, the issue of selecting the optimal estimation window for each model still stands.

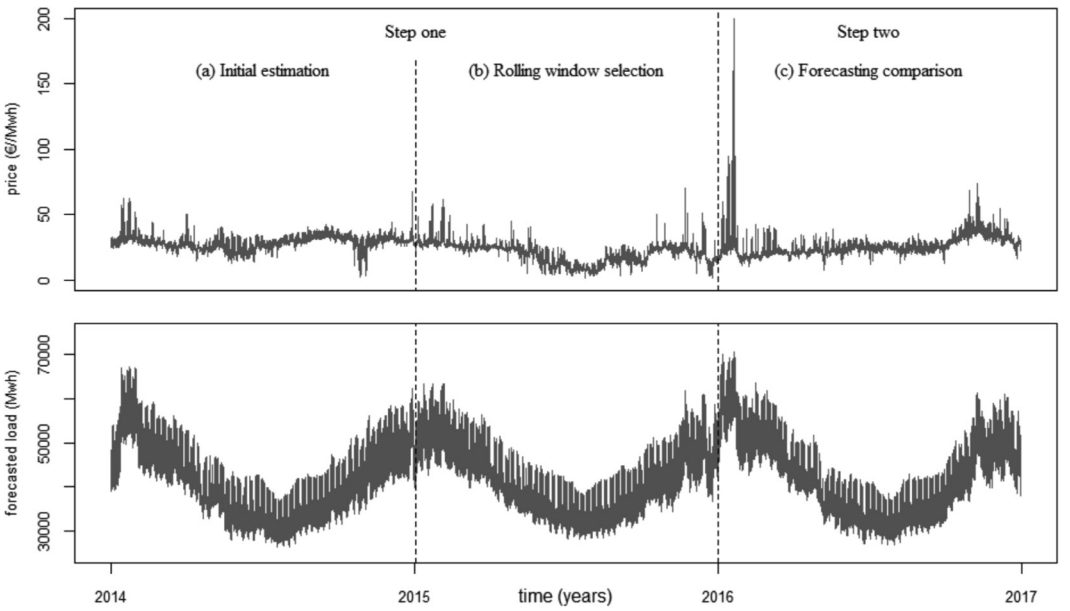
All rights reserved. Copyright © 2020 by the IAEE.

2. DATA

We analyze the day-ahead time series of two large but very different wholesale electricity markets: the Nord Pool (NP) and the Italian Power Exchange (IPEX). For both markets, we consider hourly prices and forecasted load (published on the day-ahead by the system operator and, therefore, available to all market participants for price forecasting) for three years covering the period from January 1, 2014 to December 31, 2016. These data are accessible from the system operator websites (www.nordpoolspot.com and www.mercatoelettrico.org) and from the webpage of the corresponding Author of this paper.[3] In both the NP and the IPEX (as in most electricity markets) the clearing prices and quantiles for each hour of the day are generated via 24 simultaneous auctions taking place on the day before the delivery.

The two markets are of similar size (a peak of about 70GWh for the NP and one of 55GWh for the IPEX) but have very different histories, generation mixes, demand patterns and resulting price dynamics. The NP was inaugurated in 1991 and it is now one of the oldest liberalized power markets in the world, encompassing Denmark, Finland, Norway and Sweden. Its price and load dynamics have been subject to extensive research (e.g. Haldrup and Nielsen, 2006; Weron and Misiorek, 2008; Nowotarski and Weron, 2016; Marcjasz et al., 2018a). Figure 1 presents the NP hourly day-ahead price and forecasted load time series.

Figure 1: Nord Pool hourly system prices (top) and forecasted load (bottom) for the period January 1, 2014–December 31, 2016.



Notes: In the first step, we estimate the initial parameters (sample a) and select the most appropriate rolling window for each model (sample b). In sample (c) we compare the one-step ahead forecasting performance of the different models, each estimated with the rolling window length selected in sample (b). Each hour of the day is modelled separately leading to 24 different sets of predictions in each day.

3. As standard practice (e.g. Weron 2014, Nowotarski and Weron, 2016), we pre-process the time series by substituting daylight savings hours' missing data with the arithmetic average of the neighboring values and replace the "doubled" hours with the arithmetic mean of the two values.

Copyright © 2020 by the IAEE. All rights reserved.

Load displays a strong yearly seasonality with peaks in the winter months that, however, are not always mirrored by high prices. In fact, the NP is characterized by a large share of hydropower (Norway, for example, is almost entirely reliant on this type of generation) which generates prices that are typically lower than in other European markets. However, significant spikes are present when demand is high and water storage is low (e.g. during early 2016). Figure A1 in the Appendix illustrates the daily seasonality. Both load and prices are characterized by two daily peaks (around hour 10 and hour 20), which also present the highest volatility and, therefore, are the most challenging for forecasting. This daily seasonality is essentially the same in both weekdays and weekends, while, of course, weekdays peaks are generally much higher for both load and price.

The IPEX opened in year 2004, which makes it one of the youngest power markets in Europe. Consumption is met with a mix of fossil fuels (about 65%), renewables (21%) and direct imports (14%). This mixture of relatively expensive generation and large share of imports makes the IPEX price rather high and subject to frequent spikes, as highlighted in Figure 2. Yearly seasonality is not very strong, since in Italy the main source of heating is natural gas and, therefore, there is not a winter demand peak. On the other hand, as show in Figure A2 in the Appendix, daily seasonality is quite pronounced. Such seasonality is characterized by two distinctive peaks, with the highest being the one in the evening, around hour 20. Volatility is highest when power plants are increasing production in order to reach this peak. Recent analyses of IPEX prices are Bigerna and Bollino (2015), Grossi and Nan (2015), Gianfreda et al. (2016, 2019), Lisi and Edoli (2018).

3. METHODOLOGY

As mentioned in the previous section, the NP and IPEX generate clearing prices and quantities via 24 simultaneous day-ahead auctions. This mechanism breaks down the temporal structure of the time series, since the information available to traders is updated every day, and not every hour (Huisman et al., 2007). Acknowledging this feature, it has become standard practice to model the prices of the 24 hours of the day as separate series. This approach is also superior from a purely forecasting perspective, since it recognizes that electricity generators faces very different constraints throughout the daily cycle (e.g. Weron, 2014). Our analysis follows this convention, generating 24 different sets of estimates and predictions for each one of the models we analyze. We focus on day-ahead predictions, i.e. one-step ahead, which is the most common short-term EPF exercise. Section 3.1 presents the different forecasting models in detail, while Section 3.2 illustrates how we determine the optimal sample length for each model and evaluate forecasting performance.

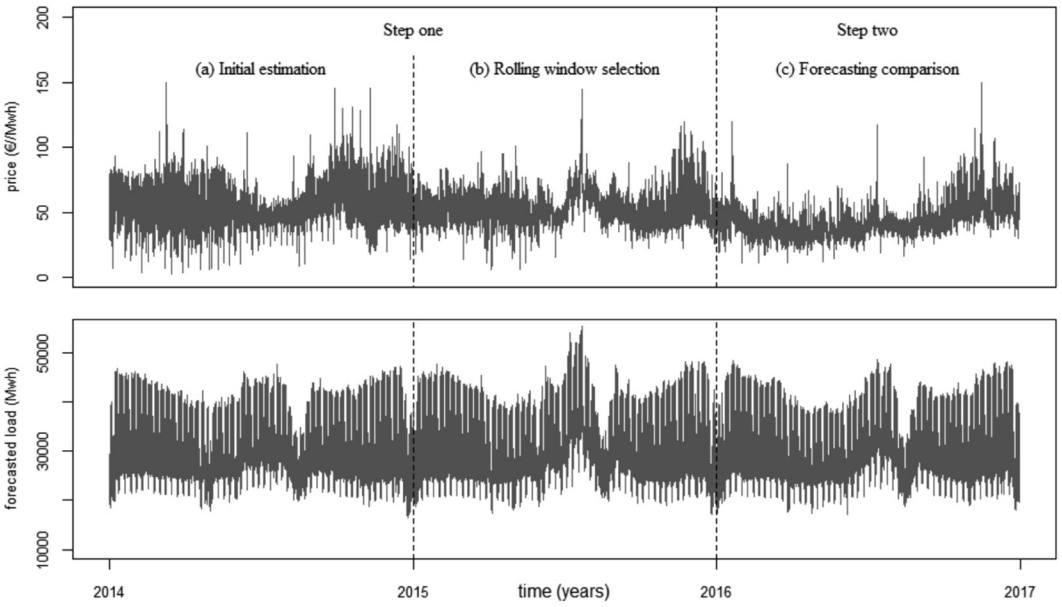
3.1 Forecasting models

The literature proposes a truly extensive collection of short-term EPF techniques. According to the review by Weron (2014), the two most popular classes of methods are: a) *statistical models*, which include both time series (e.g. Conejo et al., 2005; Weron and Misiorek, 2008; Koopman et al., 2009) and multiple regression models (e.g. Maciejowska and Nowotarski, 2016; Nowotarski and Weron, 2016; Marcjasz et al., 2018a; Steinert and Ziel, 2019) and b) *computational intelligence algorithms*, in particular neural networks (e.g. Dudek, 2016; Hubicka et al., 2018; Marcjasz et al., 2019). Rather than trying to replicate the extensive and continuously growing collection of EPF methods (a virtually impossible task), we consider eleven different models proposed in the literature and belonging to these two established classes of methods.

Our selection, summarized in Table 1, includes models with different levels of complexity, going from a simple linear regression with only two parameters to an artificial neural network of

All rights reserved. Copyright © 2020 by the IAEE.

Figure 2: IPEX hourly system prices (top) and forecasted load (bottom) for the period January 1, 2014–December 31, 2016.



Notes: In the first step, we estimate the initial parameters (sample a) and select the most appropriate rolling window for each model (sample b). In sample (c) we compare the one-step ahead forecasting performance of the different models, each estimated with the rolling window length selected in sample (b). Each hour of the day is modelled separately leading to 24 different sets of predictions in each day.

64 coefficients. We consider the explanatory variables included in the majority of EPF studies, i.e. lagged price, forecasted load and dummy variables for different days of the week.[4] To reduce the volatility and improve predictions, as recommended by Uniejewski et al. (2018), before estimating our models we apply the natural logarithm to the variables as a "variance stabilizing transformation".[5]

As benchmark, we employ the *naïve model*, or the similar-day approach, introduced by Conejo et al. (2005). In this simple model, if the day is a Tuesday, Wednesday, Thursday or Friday, the price forecast is equal to the price in the previous day, while if the day is a Saturday, Sunday or Monday the forecast is equal to the price of the same day of the previous week. Perhaps surprisingly, models that are not well designed often fail to perform better than this apparently ingenuous benchmark (Conejo et al., 2005).

Our first and simplest approach is a *simple linear regression* (SLR) of the logarithm of price (p_t) as function of the logarithm of the forecasted load (q_t):

$$p_t = \beta_0 + \beta_1 q_t + \varepsilon_t, \tag{1}$$

4. We do not consider other potential electricity price predictors such as the price of fossil fuels or the availability of renewable energy sources. While these variables are certainly important determinants of the electricity supply function (e.g. Fezzi and Bunn, 2010), their use in short-term EPF is, so far, the exception (e.g. Gianfreda et al., 2016; 2019) rather the rule (Weron, 2014). The reason is that lagged electricity prices already convey the information on the state of the supply curve relevant for short term forecasting, making additional descriptors meaningful only in some particular situations (e.g. Carmona et al., 2013). In any case, our methodological approach and the guidelines derived from our analysis should be general enough to be applicable to models including virtually any potential electricity price predictor.

5. We also tested our models estimated on the un-transformed variables and, in fact, their performance were (slightly) less satisfactory. Nevertheless, our findings remained consistent.

Copyright © 2020 by the IAEE. All rights reserved.

Table 1: Models summary

Model	Class	Parameters	Estimation	Explanatory variables	References
Naïve	benchmark	NA	NA	NA	Conejo et al. (2005), Weron and Misiorek (2008)
SLR	statistical	2	OLS	Forecasted load	NA
ARMA(1,1)	statistical	3	ML	Lagged prices and errors	Cuaresma et al. (2004), Conejo et al. (2005), Weron and Misiorek (2008)
ARMA(2,2)	statistical	5	ML	Lagged prices and errors	Cuaresma et al. (2004), Conejo et al. (2005), Weron and Misiorek (2008)
ARX(1)	statistical	3	OLS	Forecasted load and lagged prices	Misiorek et al. (2006); Weron and Misiorek (2008);
ARMAX(1,1)	statistical	4	ML	Forecasted load, lagged prices and random component	Weron and Misiorek (2008); Kristiansen (2012); Bordignon et al. (2013)
ARMAXd(1,1)	statistical	7	ML	Forecasted load, daily dummies, lagged prices and random component	Misiorek et al. (2006); Kristiansen (2012); Bordignon et al. (2013)
mARX1	statistical	9	OLS	Forecasted load, daily dummies, lagged prices, min lagged price	Nowotarski and Weron (2016), Gaillard et al. (2016), Hubicka et al. (2018)
mARX2	statistical	14	OLS	Forecasted load, daily dummies, lagged prices, min and max lagged price	Ziel and Weron (2018), Marcjasz et al. (2018a)
ANN(4)	computational intelligence	25	RBA	Forecasted load, daily dummies, lagged prices	Conejo et al. (2005); Singhal and Swarup (2011); Dudek (2016); Marcjasz et al. (2019)
ANN(7)	computational intelligence	64	RBA	Forecasted load, daily dummies, lagged prices squared of the forecasted load	Conejo et al. (2005); Singhal and Swarup (2011); Dudek (2016); Marcjasz et al. (2019)

Notes: OLS = Ordinary Least Squares, ML = Maximum Likelihood, RBA = Resilient Backpropagation Algorithm. NA = Not Available.

where t represents time, ε_t is the error component and β_0, β_1 are the parameters that we estimate via Ordinary Least Squares (OLS). Assuming an inelastic short-term demand, β_1 can be interpreted as the price-elasticity of electricity supply. This is the only model we use that, to the best of our knowledge, has not found previous applications in the EPF literature. Our analysis will show that this neglected specification can, instead, provide impressive results by using particularly short estimation windows.

We then consider a number of time series approaches, including different autoregressive and moving average (*ARMA*) specifications. The general form of this class of models is the ARMA(i,j), which can be written as:

$$\Phi_{\mathrm{i}}(B)\, p_t = \Theta_{\mathrm{j}}(B)\, \varepsilon_t, \tag{2}$$

where $\Phi_{\mathrm{i}}(B) = 1 - \phi_1 B_1 - \ldots - \phi_{\mathrm{i}} B_i$ is the autoregressive polynomial; $\Theta_{\mathrm{j}}(B) = \theta_0 + \theta_1 B_1 + \ldots + \theta_{\mathrm{j}} B_j$ is the moving average polynomial, B is the backward shift operator (i.e. $B_k p_t \equiv p_{t-k}$) and the other variables are defined as before. The parameters in $\Phi_{\mathrm{i}}(B)$ and $\Theta_{\mathrm{j}}(B)$ can be estimated via Maximum Likelihood (ML). ARMA models have been implemented in this context by Nogales et al. (2002), Cuaresma et al. (2004), Conejo et al. (2005), Weron and Misiorek (2008) and many others. We consider ARMA(1,1) and ARMA(2,2) specifications.

All rights reserved. Copyright © 2020 by the IAEE.

ARMA models can be augmented by including explanatory (or exogenous) variables. This class of specifications, referred as *ARMAX*, is sometimes re-written as a transfer function (e.g. Weron, 2014). In this study, we opt for the "regression with ARMA errors" representation, which is equally effective for forecasting while allowing a simpler interpretation of the parameters (Hyndman and Athanasopoulos, 2014). This specification can be written as:

$$\Phi_{\mathrm{i}}(B)\left[p_t-\beta_0-\boldsymbol{\beta}'\boldsymbol{X}_t\right]=\Theta_{\mathrm{j}}(B)\varepsilon_t, \tag{3}$$

where $\boldsymbol{X}_t$ indicates the vector of explanatory variables including, for example, forecasted load. Most EPF studies (e.g. Conejo et al., 2005; Misiorek et al., 2006; Weron and Misiorek, 2008; Kristiansen, 2012; Bordignon et al., 2013) find that these models typically outperform their ARMA counterparts. Here we consider three possible specifications: an ARX(1) and an ARMAX(1,1), both with logarithm of forecasted load as explanatory variable, and an ARMAXd(1,1) with logarithm of forecasted load and three dummy variables for Saturday, Sunday and Monday (Misiorek et al., 2006).

We then consider the two regression models evaluated by Marcjasz et al. (2018a), that we indicate with the term mARX1 and mARX2, which stands for *multi-day ARX* (Nowotarski and Weron, 2016). The mARX1 was introduced by Misiorek et al. (2006) and further developed by Maciejowska and Nowotarski (2016), Nowotarski and Weron (2016), Gaillard et al. (2016), Hubicka et al. (2018), among others. It is essentially an ARX with 9 parameters and a rich set of explanatory variables:

$$p_t=\phi_0+\phi_1 p_{t-1}+\phi_2 p_{t-2}+\phi_7 p_{t-7}+\phi_3 p_{min,t-1}+\beta_1 q_t+\sum\nolimits_{i=1}^{3} d_i D_i+\varepsilon_t, \tag{4}$$

where $p_{min,t-1}$ is the minimum of the 24 hourly prices in the previous day and $D_1,\ldots,D_3$ are respectively dummy variables for Monday, Saturday and Sunday. These three variables capture the fact that Saturdays and Sundays are characterized by lower demand and that Monday is the day after the weekend and, therefore, the one in which demand and price raise the quickest. The parameters ϕ_1, ϕ_2 and ϕ_7 create the link with the historical prices and ϕ_3 with the overall supply function of the previous day. The second regression model (mARX2), was introduced by Ziel and Weron (2018), and it is our most complex statistical approach. It expands upon the mARX1 to include up to 14 parameters:

$$p_t=\phi_1 p_{t-1}+\phi_2 p_{t-2}+\phi_7 p_{t-7}+\phi_3 p_{min,t-1}+\phi_4 p_{max,t-1}+\phi_5 p_{h24,t-1}+\beta_1 q_t+\sum\nolimits_{i=1}^{7} d_i D_i+\varepsilon_t, \tag{5}$$

where $p_{max,t-1}$ is maximum of the 24 hourly prices in the previous day, $p_{h24,t-1}$ is the price for the last hour of the previous day (important to predict the early morning hours) and the daily dummies are now expanded to one for each day of the week.

Regarding computational intelligence models, we consider two different *Artificial Neural Networks* (ANNs). ANNs are becoming a very popular forecasting tool. They are data-driven, in the sense that they require little assumptions and yet are able to capture functional relationships that are unknown *a priori* or hard to describe. In addition, they can approximate any continuous non-linear function to any desired accuracy (Zhang et al., 1998). This flexibility comes at a cost: ANNs include many parameters and, therefore, require large and stable samples for their estimation. Here we consider two different feed-forward ANNs with a single hidden layer. This established ANN structure has found several applications in EPF (e.g. Conejo et al, 2005; Singhal and Swarup, 2011; Dudek, 2016; Marcjasz et al., 2019). ANNs can be viewed as high-dimensional nonlinear regression models. In particular, a feedforward ANN with k explanatory variables (inputs) $\boldsymbol{X}_t$, and with one hidden layer of q neurons can be represented as:

$$ANN\left(\boldsymbol{X}_t\right)=f_0\left(\alpha+\sum_{h=1}^{q} w_h\phi\left(\tilde{\alpha}_h+\sum_{j=1}^{p}\tilde{w}_{jh}x_j\right)\right) \tag{6}$$

Copyright © 2020 by the IAEE. All rights reserved.

Where $\phi(.)$ is the activation function, $f_0(.)$ is typically chosen as the identity function (and we do so in our specification) and $w_{hk}, \alpha, \tilde{w}_h, \tilde{\alpha}_h$ are the unknown parameters to be estimated. For both networks, we impose the common assumption that the number of neurons in the hidden layer is equal to number of input variables plus one (e.g. Dudek, 2016). The ANNs literature often refers to this latter parameter as the "bias" and in (5) is represented by α and $\tilde{\alpha}_h$. Its role is analogous to that of the intercept in linear models. As activation function, we use the hyperbolic tangent (Zhang et al., 1998). In the first and less complex neural network, which we indicate with ANN(4), we include in $\boldsymbol{X}_t$ four input variables: the electricity price of the previous day, the forecasted load and two dummy variables for Saturday and Sunday. This ANN has 25 parameters, which is almost two times the parameters of the most complex statistical model, the mARX2. In the second and more complex neural network, ANN(7), we also include a Monday dummy variable (to mirror the ARX and mARX specifications), the square product of forecasted load (as in Dudek, 2016) and the price at lag two. This generates 64 parameters. In both ANNs, we standardize all the variables before estimation. We carry out estimation via the resilient backpropagation algorithm, which is characterized by a faster learning rate than the standard backpropagation (Riedmiller and Braun, 1993).

3.2 Window-size selection approach and forecasting evaluation

In order to clearly illustrate how rolling-window size selection affects EPF performance, we follow a deliberately simple approach. It consists in two steps designed to compare forecasting performance both across models and across window sizes, and thereby select the best performing specification for each hour. We compare this strategy with the standard approach of using a fixed sample size selected *a priori*.

In the first step of our approach, we identify, for each model and each hour, which window size (λ) provides the best performance. We estimate the initial values of the parameters using the data in year 2014 (or part of, depending on the value of λ) and we compare one-step ahead forecasts during year 2015. Therefore, in this initial step, we use data in samples (a) and (b) in Figure 1 and 2. In order to determine λ^* (the optimal window size for each model and hour) we compare rolling windows of size k, $k+1$, $k+2$, ..., 100, 150, 200, ..., 350 with k = number of model parameters, which corresponds to the minimum number of observations necessary for estimation. This means that our rolling windows vary from 2 days to almost one year. We also test the performance of using a recursive window, i.e. a window that includes all available previous observations. While the size of each rolling window is constant, the size of the recursive window increases by one with each passing day. For example, the recursive window for $t=1$ (i.e. for 1/1/2015) includes 365 observations, for $t=2$ includes 366 observations, and so forth, up to 730 observations for $t=366$ (i.e. 31/12/2015), the last day in sample (b).

In the second step, we estimate each model using its specific λ^* and evaluate their forecasting accuracy. We compare results across models and against the standard approach of estimating all models using the same sample length, which in this literature is typically picked between 6 months and 1 year of data (here we use $\lambda=300$). To make sure this is a fair comparison, we evaluate forecasting performances on a data range that is different to the one used to identify λ^*, i.e. we use the data in year 2016, corresponding to sample (c) in Figure 1 and 2.

In both steps, as a measure of accuracy during each hour we use the commonly applied Mean Absolute Error (MAE):

$$\mathrm{MAE}_h = \frac{1}{T}\sum_{t=1}^{T}\left|\hat{p}_{t,h} - p_{t,h}\right|,$$

All rights reserved. Copyright © 2020 by the IAEE.

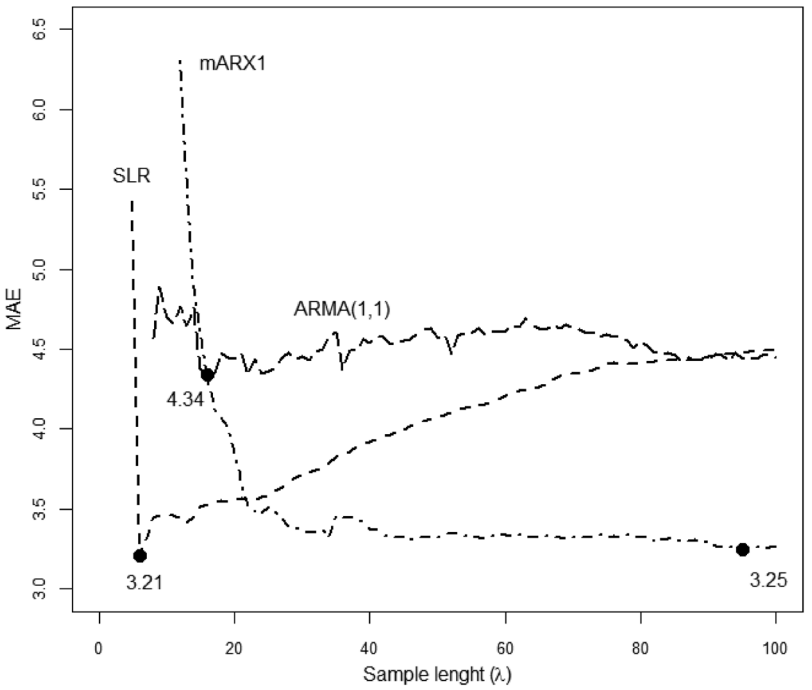
where the "hat" sign indicates the one-step ahead forecast, T is the total number of forecasting steps and h indicates the hourly trading period. Results remain consistent using alternative measures of forecasting power, such as the root mean squared error, and we do not report them in order to preserve space (the summary of results obtained using the root mean squared error is reported in Table A3 in the Appendix).

4. RESULTS

4.1 Nord Pool

We start by exploring in detail how models perform in the highest peak period (hour 10) and, in particular, how forecasting accuracy varies with λ. Figure 3 compares MAEs for values ranging from k to 100 for three models: SLR, ARMA(1,1) and mARX1 (Figure A3 in the Appendix displays the same comparison for all the remaining models). Performances vary considerably, particularly for the two regression models. Interestingly, these two models exhibit contrasting behaviors. On one hand, the accuracy of the mARX1 steadily improves with sample size, reaching a MAE of €3.25 with one of the highest values of λ. On the other hand, the SLR achieves impressive results with small λs (the best MAE is €3.21 for $\lambda = 6$) but, after that, steadily deteriorates, reaching a MAE above €4.5 for the largest sample sizes.

Figure 3: Forecasting performance at different estimation sample lengths (Nord Pool, hour 10)



Notes: Forecasting performance measured via the Mean Absolute Error (MAE). All values refer to hour 10 (peak), Nord Pool price for year 2015. SLR = simple linear regression (SLR), ARMA(1,1) = autoregressive moving average, mARX1 = mARX1 regression (Marcjasz et al., 2018a). The dot identifies the best MAE for each model.

These reversed dynamics can be explained by the different structures of the two models. The mARX1 is relatively complex: it includes 9 parameters in order to capture all the main features

Copyright © 2020 by the IAEE. All rights reserved.

of electricity price dynamics, such as different types of seasonality and the relationship with the overall daily supply function. Because of its complexity, it also requires large windows for precise estimation. On the other hand, the SLR is extremely simple and can quickly adapt to changes in market conditions without the need of modelling them explicitly. For this reason, it performs well with small λs. Employing large samples and, therefore, implicitly assuming a slower evolution of the parameters over time, nullifies such advantages and deteriorates its forecasting ability. Interestingly, the ARMA(1,1) is somewhat in a middle ground: it is not complex enough to capture the main features of price dynamics, but also not simple enough to be estimated precisely on very small samples. Therefore, its performance varies less significantly with λ, and remains inferior to the other two models for most window sizes.

In order to explore further the importance of sample length and provide a first comparison with the standard approach of using a fixed sample size, Table 2 reports MAEs for all models, again focusing on peak hour 10. The first column follows the established approach and estimates all models using rolling windows of 300 observations. Perhaps surprisingly, but in line with previous findings (Conejo et al., 2005), most models do not outperform the naïve benchmark. Only the five most complex models pass this apparently simple test: the two ARMAX, the two mARX, and the ANN(7). The mARX2 is the best model, with a MAE of €2.87. For comparison, the SLR has a MAE of €6.03, which is almost two times the one of the naïve benchmark, making it the worst overall model.

Table 2: Forecasting performance for hour 10, Nord Pool price in year 2015 (window b)

	Fixed λ=300		$\lambda=\lambda^*$	
Model	**MAE**	**λ**	**MAE**	**λ^***
Naïve	3.64		3.64	
SLR	6.03	300	3.21	6
ARMA(1,1)	4.41	300	4.34	22
ARMA(2,2)	4.38	300	4.23	rec
ARX(1)	4.21	300	3.32	23
ARMAX(1,1)	3.17	300	3.05	24
ARMAXd(1,1)	3.21	300	3.10	rec
mARX1	3.38	300	3.25	95
mARX2	**2.87**	300	**2.79**	rec
ANN(4)	3.94	300	3.84	75
ANN(7)	3.60	300	3.43	rec

Notes: "rec" stands for recursive estimation, i.e. augmenting the estimation sample with one additional observation each day. MAE of best performing models highlighted and in bold. Models' descriptions reported in Table 1.

In the third column, we allow λ to vary and compare models using their optimal window size. Not surprisingly, all models improve significantly. The best performing one is still the mARX2, its MAE improving to €2.79. This result is achieved using a continuously expanding, recursive window. The most improved model is the SLR which, as already mentioned, with $\lambda^*=6$ reaches a MAE of €3.21. This is quite impressive for a linear regression with just two parameters: the SLR is now the third-best model, performing better than much more complex approaches such as the two neural networks and the mARX1. Indeed, the two ANNs do not seem to forecast particularly well, with the best of the two doing only marginally better than the naïve model. Finally, the worst specifications are the two ARMAs, both of them failing to outperform the benchmark for any λ. Altogether, allowing sample size to vary has two main impacts: a) improve forecasting performance and b) significantly change model ranking.

All rights reserved. Copyright © 2020 by the IAEE.

We now move to the second step of our comparison. For each hour and each model, we use the λ^* selected in the previous step to evaluate forecasting performance on the last year of our data, i.e. year 2016. In Table 3, we focus on three of the models estimated via OLS (SLR, ARX(1) and mARX2) and compare them with the naïve benchmark. We report estimates for all trading hours using λ^* and $\lambda = 300$. Several findings emerge.

Table 3: Forecasting performance for all hours, Nord Pool price in year 2016 (window c)

		SLR		ARX(1)		mARX2	
hour	**Naïve**	**λ=300**	**λ=λ***	**λ=300**	**λ=λ***	**λ=300**	**λ=λ***
1	1.69	5.55	1.45 (6)	1.53	1.45 (33)	1.29	**1.17** (rec)
2	1.74	5.69	1.50 (6)	1.63	1.36 (rec)	**1.01**	1.08 (56)
3	1.89	5.89	1.68 (11)	1.89	1.51 (rec)	**1.23**	1.34 (58)
4	2.03	6.07	1.71 (6)	2.07	1.61 (rec)	**1.42**	1.53 (57)
5	2.05	6.22	1.66 (6)	2.16	1.68 (rec)	1.59	**1.55** (59)
6	1.84	6.47	1.53 (6)	2.20	1.71 (rec)	1.63	**1.42** (57)
7	1.84	6.62	**1.49** (6)	2.56	1.69 (19)	1.64	1.64 (300)
8	2.31	6.74	**1.81** (6)	3.32	1.96 (14)	1.88	2.37 (rec)
9	3.84	7.28	**2.87** (6)	5.11	3.22 (14)	3.29	3.20 (rec)
10	4.06	7.62	**3.19** (6)	5.41	3.67 (23)	3.83	3.41 (rec)
11	3.74	7.25	**2.93** (6)	4.75	3.32 (23)	3.43	2.99 (rec)
12	3.03	6.50	**2.31** (6)	3.66	2.63 (23)	2.69	2.40 (rec)
13	2.56	6.29	**1.95** (6)	3.03	2.26 (23)	2.26	2.07 (rec)
14	2.35	6.30	**1.80** (6)	2.78	2.09 (14)	2.04	1.96 (rec)
15	2.36	6.22	**1.79** (6)	2.81	2.10 (14)	2.01	1.95 (rec)
16	2.35	6.25	**1.76** (6)	2.68	2.09 (23)	1.96	1.92 (rec)
17	2.62	6.62	**1.93** (6)	2.89	2.28 (23)	2.40	2.24 (rec)
18	3.46	7.69	**2.49** (6)	3.55	2.97 (23)	3.17	2.91 (rec)
19	4.32	8.14	**3.07** (7)	4.29	3.60 (23)	4.01	3.61 (rec)
20	3.70	7.13	**2.77** (7)	3.76	3.16 (25)	3.38	3.10 (rec)
21	2.26	5.96	**1.75** (6)	2.09	1.88 (28)	1.82	1.79 (rec)
22	1.71	5.60	1.39 (6)	1.47	1.43 (31)	1.36	**1.31** (rec)
23	1.49	5.33	1.23 (6)	1.15	1.19 (100)	1.16	**1.10** (rec)
24	1.47	5.26	1.28 (5)	1.10	**1.01** (rec)	1.15	1.07 (rec)

Notes: MAE for the naïve model and the four models estimated via OLS. Best sample length (λ^*) selected prior to estimation using 2015 data (window b) reported in parenthesis next to each MAE. Highlighted and bold is the best MAE for each hour.

First, in line with the literature, peak forecasting is significantly harder than baseload forecasting, with the MAEs of all models significantly increasing during the peak. *Second*, the models estimated using λ^* (estimated in the previous step, i.e. on year 2015 data) consistently outperform the models estimated using the standard fixed rolling-window approach. This is true for all models and all hours, leaving little doubt on the advantages provided by an appropriate window size selection in EPF. *Third*, comparing across models, there is a clear and consistent increase in λ^* when moving from the SLR, to the ARX(1) and finally to the mARX2. This feature supports our previous findings showing how simple models perform better with short estimation windows, while complex models require larger samples. *Fourth*, simpler models forecast better during peak hours, while more complex models are better suited for the off-peak. More specifically, the SLR estimated on extremely small samples (six or seven observations) is the best performing model for all hours between 7am to 9pm, while the mARX2 is the preferred model for the evening and early morning hours. *Fifth*, comparing within models but across hours, we notice how the off-peak seem to favor larger samples, while peak hours select smaller ones. This is particularly evident in the ARX(1) column, which reports peak hours' λ^* between 14 and 25 observations, while most off-peak hours select a much larger λ^* and even recursive samples.

Copyright © 2020 by the IAEE. All rights reserved.

The last two points can be explained by the higher volatility and instability of peak hours. In such unstable circumstances, simple models have an edge over complex ones, since they can be estimated precisely on extremely small samples and, therefore, can quickly adapt to changing conditions. On the other hand, the stable dynamics of the off-peak favor more complex specifications. The optimal windows size λ^* changes accordingly, both across-models and within the same model.

Finally, perhaps surprisingly, in a few trading periods (hours 2, 3 and 4) the mARX2 performs better with $\lambda = 300$ then with $\lambda = \lambda^*$. These differences are small, but indicate that the optimal window size selected in one year is not necessarily also the optimal one for the following year. This result does not undermine the importance of selecting the appropriate sample length but, on the other hand, should stimulate the development of more advanced window size selection approaches.

Table 4: Model forecasting ranking, Nord Pool price in year 2016 (window c)

	Number of trading hours in which the model is ranked:				
Model	1st	2nd	3rd	4th or more	$\overline{\text{MAE}}$
SLR	13		2	9	1.97
ARMA(1,1)		1		23	2.58
ARMA(2,2)				24	2.41
ARX(1)			2	22	2.16
ARMAX(1,1)	2	10	7	5	2.01
ARMAXd(1,1)	5	6	6	7	2.03
mARX1			1	23	2.18
mARX2	2	6	2	14	2.09
ANN(4)	2			22	2.47
ANN(7)		1	2	21	2.33

Notes: All models estimated using the best sample length (λ^*) selected using 2015 data (window b). $\overline{\text{MAE}}$ indicates the average MAE across all hours.

Table 4 summarizes the performance of all models estimated with λ* across all hours (the MAEs for all models and hours are reported in Table A1 in the Appendix). Surprisingly, the best overall model is the SLR, which has not only the lowest average MAE but it is also the best forecasting model for more than half of the hours of the day. The ARMAX(1,1) and ARMAXd(1,1) are second and third, achieving comparable results to the SLR but with considerably larger optimal samples. On the other side of the spectrum, the ARMAs present the largest MAEs, confirming the importance of including forecasted load as explanatory variable when modelling price.

4.2 IPEX

We replicate the NP analysis on the IPEX market. To preserve space, we present directly the results of the second step, i.e. comparing models using λ^* (again, selected using year 2015 data) and $\lambda = 300$ using one-step ahead forecasts for year 2016 (the performance of all models for different λs in hour 10 are reported in Figure A4 in the Appendix). As for the Nordic market, Table 5 focuses on the SLR, ARX(1), mARX2 and the naïve benchmark. Results are in line with those observed on the NP, with the SLR estimated on short windows outperforming the other two specifications during most peak hours, and the mARX2 providing the best forecasts during the off-peak. As before, the most complex specification selects larger estimation samples, including recursive windows. Again, for the mARX2 the optimal λ^* is sometimes outperformed by the fixed $\lambda = 300$.

As with the NP, we summarize the performance of all models estimated with λ^* across all hours. Results are presented in Table 6 (the full set of results with the MAEs for all models and hours is in Table A2 in the Appendix). The last column shows how IPEX prices are considerably harder to

All rights reserved. Copyright © 2020 by the IAEE.

Table 5: Forecasting performance for all hours, IPEX price in year 2016 (window c)

		SLR		ARX(1)		mARX2	
hour	**Naïve**	**λ=300**	**λ=λ***	**λ=300**	**λ=λ***	**λ=300**	**λ=λ***
1	4.13	8.13	3.77 (7)	3.89	3.91 (rec)	2.61	**2.65** (350)
2	4.07	7.75	3.56 (7)	3.90	3.85 (350)	2.63	**2.62** (350)
3	3.91	7.56	3.33 (7)	3.65	3.63 (350)	**2.52**	2.53 (350)
4	3.95	7.21	3.33 (7)	3.64	3.66 (350)	**2.65**	2.69 (350)
5	3.94	6.99	3.42 (13)	3.75	3.27 (32)	**2.80**	2.82 (350)
6	3.65	7.35	3.31 (8)	3.76	3.24 (32)	2.76	**2.74** (350)
7	3.71	8.81	3.39 (7)	4.99	3.81 (35)	3.11	**3.03** (350)
8	4.47	10.62	**3.82** (7)	6.67	4.38 (21)	4.07	4.17 (250)
9	5.69	11.74	5.00 (7)	8.01	5.42 (10)	**4.79**	4.86 (250)
10	5.67	11.47	**4.08** (7)	7.29	5.20 (11)	4.77	4.86 (250)
11	5.28	11.23	**4.40** (7)	6.40	4.57 (25)	4.48	4.54 (250)
12	5.07	10.80	4.42 (9)	6.17	4.55 (42)	**4.28**	4.31 (250)
13	4.46	10.12	3.98 (11)	5.39	5.32 (350)	**3.80**	3.85 (250)
14	4.51	10.16	4.03 (7)	6.25	4.22 (29)	**4.02**	4.12 (250)
15	5.21	10.71	4.61 (7)	7.13	4.79 (29)	4.54	**4.37** (350)
16	4.93	11.16	**4.35** (7)	6.95	4.56 (19)	4.42	4.37 (350)
17	5.16	11.88	**4.51** (7)	6.75	4.89 (11)	4.68	4.58 (250)
18	5.09	13.84	**4.85** (7)	6.05	4.91 (10)	4.85	4.88 (350)
19	6.00	13.41	5.65 (8)	6.38	5.65 (22)	5.48	**5.34** (rec)
20	6.44	12.98	5.92 (8)	6.41	6.49 (250)	5.67	**5.49** (rec)
21	6.48	11.18	5.83 (7)	6.30	6.16 (350)	5.66	**5.61** (350)
22	5.17	9.79	4.77 (7)	4.89	4.67 (21)	4.42	**4.42** (300)
23	3.90	8.92	3.53 (8)	3.66	3.46 (22)	3.24	**3.20** (350)
24	3.16	8.44	2.86 (8)	3.01	2.98 (350)	2.73	**2.68** (350)

Notes: MAE for the naïve model and the four models estimated via OLS. Best sample length (λ^*) selected prior to estimation using 2015 data (window b) reported in parenthesis next to each MAE. Highlighted and bold is the best MAE for each hour.

forecast then NP ones, with the average MAEs being much higher than the ones of the Nordic market. This feature has to be expected, given the higher volatility of the Italian price, which is evident even from a quick comparison of Figure 1 and 2. In this market the best model is the mARX2, which provides the best results for 10 out of 24 trading periods. The ARMAX(1,1) and ARMAXd(1,1) are, again, second and third. The SLR does not repeat the impressive performance accomplished on the Nord Pool, but still provides reasonable forecasts, being among the first three models for seven of the 24 hours.

Table 6: Model forecasting ranking, IPEX price in year 2016 (window c)

	Number of trading hours in which the model is ranked:				
Model	1st	2nd	3rd	4th or more	$\overline{\text{MAE}}$
SLR	2	1	4	17	4.23
ARMA(1,1)				24	5.33
ARMA(2,2)		2	3	19	4.94
ARX(1)		1		23	4.48
ARMAX(1,1)	7	6	3	8	4.08
ARMAXd(1,1)	4	6	7	7	4.09
mARX1	1	2	4	17	4.28
mARX2	10	6	3	5	4.06
ANN(4)				24	5.47
ANN(7)				24	4.79

Notes: All models estimated using the best sample length (λ^*) selected using 2015 data (window b). $\overline{\text{MAE}}$ indicates the average MAE across all hours.

Copyright © 2020 by the IAEE. All rights reserved.

4.3 The benefits of selecting the appropriate window size

Table 7 provides an overall summary of the improvements delivered by the two-steps window size selection method over the standard fixed-window approach. It presents the best models and MAEs for each hour of the day in the two markets according to the two strategies. In 39 hours out of 48, the two-steps method provides lower MAEs. Focusing on the Nord Pool, columns two to five show that our approach outperforms the benchmark in all but three off-peak hours. The best models are SLRs with very small samples for the peak hours and more complex models (mainly ARMAX and mARX2) with large, typically recursive samples for the off-peak. Interestingly, this means that a fixed estimation window of 300 observations is certainly too long for the peak, but also too short for the off-peak, when the best samples are typically recursive. This difference of optimal window size (and models) between on-peak and off-peak can explain why Marcjasz et al. (2018a) and Hubicka et al. (2018) find that using a weighted average of different windows performs better than both long and short windows in terms of overall daily average MAE. Our results suggest that varying λ (or weights) across hours should produce even better results.

Table 7: Forecasting performance for the best models in each hour, Nord Pool and IPEX price in year 2016 (window c)

	Nord Pool				IPEX			
	$\lambda=\lambda^*$		$\lambda=300$		$\lambda=\lambda^*$		$\lambda=300$	
hour	**Model (λ^*)**	**MAE**	**Model**	**MAE**	**Model (λ^*)**	**MAE**	**Model**	**MAE**
1	mARX2 (rec)	**1.17**	mARX2	1.29	mARX2 (350)	2.65	mARX2	**2.61**
2	mARX2 (56)	1.08	mARX2	**1.01**	mARX2 (350)	**2.62**	mARX2	2.63
3	ARMAXd11 (rec)	1.33	mARX2	**1.23**	mARX2 (350)	2.53	mARX2	**2.52**
4	ANN4 (rec)	1.45	mARX2	**1.42**	mARX2 (350)	2.69	mARX2	**2.65**
5	ANN4 (rec)	**1.44**	mARX2	1.59	mARX2 (350)	2.82	mARX2	**2.80**
6	ARMAXd11 (rec)	**1.39**	ARMAX11	1.58	mARX2 (350)	**2.74**	mARX2	2.76
7	ARMAXd11 (rec)	**1.43**	ARMAX11	1.59	mARX2 (350)	**3.03**	mARX2	3.11
8	SLR (6)	**1.81**	mARX2	1.88	SLR (7)	**3.82**	ARMAXd11	4.01
9	SLR (6)	**2.87**	mARX2	3.29	ARMAX11 (250)	4.86	mARX2	**4.79**
10	SLR (6)	**3.19**	mARX2	3.83	SLR (7)	4.80	mARX2	**4.77**
11	SLR (6)	**2.93**	mARX2	3.43	ARMAX11 (86)	**4.36**	mARX2	4.48
12	SLR (6)	**2.31**	mARX2	2.69	ARMAX11 (86)	**4.14**	mARX2	4.28
13	SLR (6)	**1.95**	ARMAX11	2.21	ARMAX11 (250)	**3.82**	ARMAX11	3.78
14	SLR (6)	**1.80**	ARMAX11	2.04	ARMAXd11 (79)	**3.69**	ARMAXd11	3.86
15	SLR (6)	**1.79**	ARMAX11	1.98	ARMAXd11 (79)	**4.37**	ARMAX11	4.45
16	SLR (6)	**1.76**	ARMAX11	1.89	ARMAXd11 (96)	**4.23**	mARX2	4.42
17	SLR (6)	**1.93**	ARMAX11	2.20	ARMAX11 (30)	**4.26**	ARMAX11	4.66
18	SLR (6)	**2.49**	ARMAX11	2.99	ARMAXd11 (200)	**4.51**	ARMAX11	4.62
19	SLR (7)	**3.07**	ARMAX11	3.71	ARMAX11 (300)	**5.32**	ARMAX11	**5.32**
20	SLR (7)	**2.77**	ARMAX11	3.22	ARMAX11 (200)	**5.43**	ARMAX11	5.48
21	ARMAX11 (rec)	**1.66**	ARMAX11	1.74	mARX2 (350)	**5.61**	mARX2	5.66
22	ARMAXd11 (rec)	**1.20**	ARMAX11	1.26	mARX2 (300)	**4.42**	mARX2	**4.42**
23	ARMAXd11 (rec)	**1.03**	ARMAX11	1.06	mARX2 (350)	**3.20**	ARMA22	3.23
24	ARMAX11 (rec)	**0.97**	ARMAX11	1.02	mARX1 (350)	**2.64**	ARMA22	2.68

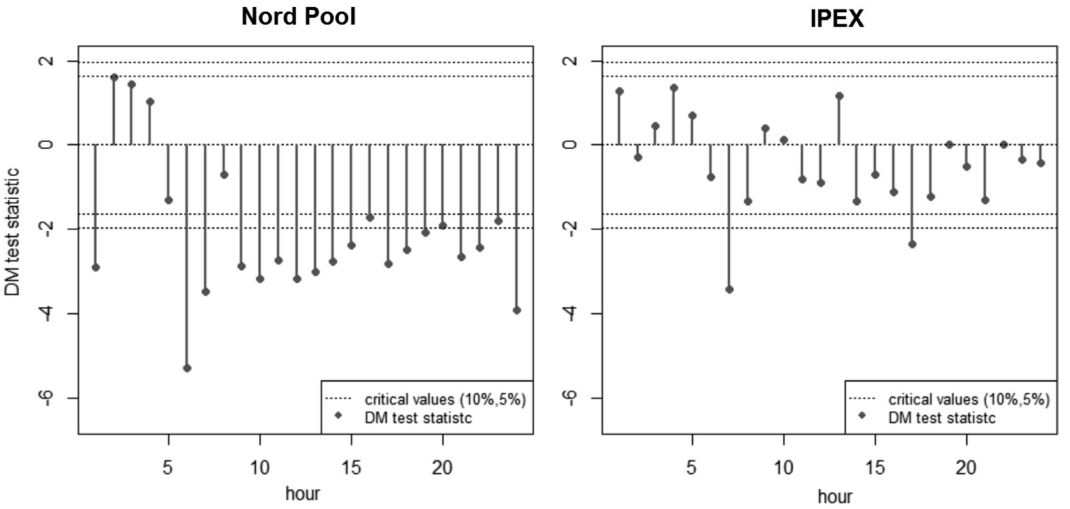
Notes: best models and sample lengths (in parenthesis next to each model name) for each hour chosen by running one-step head forecasts in year 2015 (window b). MAEs calculated on the data in year 2016 (window c). Highlighted and bold is the best MAE for each hour and market.

The situation in the IPEX, reported in columns six to nine, appears to be less diverse across approaches, with both methods selecting mostly ARMAX and mARX2 models. However, the two-step method selects again smaller windows during the peak and larger samples for the off-peak. This apparently small change allows our approach to outperform the standard practice in 18 out of 24

All rights reserved. Copyright © 2020 by the IAEE.

hours. However, for 6 hours the fixed λ works better. We do not interpret this as a sign that selecting the optimal window length is unimportant but, rather, that selection methods can be significantly improved, for example by allowing the size of the window to vary with each passing day and not simply employing the optimal window of the previous year.

Figure 4: Diebold-Mariano tests: best model with $\lambda=\lambda^*$ vs. best model with $\lambda=300$



Notes: Best models (and sample λ^*) for each hour chosen by running one-step head forecasts in year 2015 (window b). Diebold- Mariano tests calculated on the data in year 2016 (window c). The best models for each hour are reported in Table 7.

Nevertheless, even our simple two-step approach provides a significant improvement in forecasting accuracy. Figure 4 reports Diebold-Mariano (1995) tests comparing, in each hour and market, the best specifications according to the two-step and the standard methods.[6] The improvement delivered by the two-step approach is significant at the 5% level for 18 hours and at the 10% level for an additional 3 hours, for a total of about 45% of all trading periods. Most of these significant values are in the Nord Pool dataset. On the other hand, in the hours in which the fixed-window approach performs better, MAEs are never significantly different from the ones produced by the two-step method. Overall, these results reveal how selecting the appropriate rolling window size for each trading hour is a very promising and relatively simple strategy to improve forecasting models for electricity prices.

5. CONCLUDING REMARKS

This analysis investigated the performance of eleven different EPF models (including both statistical methods and computational intelligence techniques) on two large European power markets, the NP and the IPEX. We compared the common approach of implementing all models on a rolling window whose size is fixed *a priori* against a novel and simple two-step method that selects the appropriate window size for each model and trading period prior to estimation.

6. We do not compare models but forecasts, i.e. we ask whether the difference in forecasting performance we observe in year 2016 is significant or not. In such context, the Diebold-Mariano (1995) test is the appropriate approach, as illustrated by Diebold (2014). We implement the modified test by Harvey et al. (1997). For model comparison see, for example, Hansen and Timmerman (2015).

Copyright © 2020 by the IAEE. All rights reserved.

Our results leave little doubt on the advantages provided by selecting suitable rolling window sizes. Considering both markets, the two-steps method outperforms the standard approach in 39 trading periods out of 48. In 21 of such periods, the difference is significant according to Diebold-Mariano (1995) tests. Perhaps surprisingly, using a fixed window provides better forecasts in 9 hours, but none of those differences is significant at any standard level. Nevertheless, this result implies that there is likely to be significant room for improvement in algorithms of window size selection, and that our two-steps approach is just a "first step" in the right direction. Further research should investigate the design of methods allowing window size to vary with each new observation and more advanced selection algorithms (e.g. Pesaran and Timmerman, 2007; Inoue et al., 2017). Another promising approach is combining our two-step strategy with the weighted average of models estimated with different rolling windows recently developed by Hubicka et al. (2018) and Marcjasz et al. (2018a).

Comparing our results across trading periods, it is clear how peak hours typically select simple models and small window sizes, while off-peak hours favor more complex specifications and longer (often recursive) samples. A reasonable justification for this difference is the higher instability characterizing the price formation process during the peak. Such instability is likely to be a byproduct of the characteristics of electricity supply and demand and, in particular, of the non-storability of this peculiar commodity. Since the electric grid needs always to be balanced, strategic and time-varying bidding behavior (e.g. Fabra and Toro, 2005, Ito and Reguant, 2016) are likely to play a more significant role during the peak, i.e. when margin is lower, thereby generating evolutionary strategic dynamics. The implication for EPF is that simple models that can quickly "adapt" to varying conditions can perform extremely well in peak hours if estimated on very small rolling windows. In our analysis, if we consider peak hours, SLRs estimated using samples of only six or seven observation tend to perform better than much more complex specifications, such as ANNs and advanced ARMAX models. This result was so far overlooked because EPF comparisons typically focuses on relatively large rolling samples, which implicitly favor more complex models. Another important result is that different trading hours require very different EPF windows, therefore, window-size selection mechanisms are best studied on hourly basis rather than on daily averages.

Taken together, our results clearly show that the standard practice of testing EPF models on fixed estimation windows produces subpar results. A straightforward and yet powerful approach to improve forecasting accuracy is to determine optimal sample lengths (for each trading hour) prior to estimation or, at least, to evaluate models using different rolling-window sizes. Given its simplicity, we are hopeful that this practice will establish itself as standard in electricity price forecasting. As a general guideline, the size of the optimal window appears to depend on two main features: a) the complexity of the model and b) the instability of the data generating process. These relations are valid in both our markets and we think there are no reasons to believe that they would not hold in virtually all power markets across the world.

Finally, our analysis focuses on point predictions and completely bypasses the issue of probabilistic (or interval) forecasting. However, probabilistic forecasting in electricity markets has been the subject of significant recent work (e.g. Bello et al., 2016; Dudek, 2016) and it is becoming a fundamental tool to understand price volatility and risk. While it is reasonable to expect that selecting the appropriate rolling window size will produce significant improvements also to probabilistic forecasts and, therefore, that our approach will benefit also this area of inquiry, we leave a formal investigation of this issue to further work.

All rights reserved. Copyright © 2020 by the IAEE.

ACKNOWLEDGMENTS

Many thanks to Fany Nan and Rafal Weron for sharing with us the IPEX and Nord Pool data. We are indebted to the Editor, Adonis Yatchew, and four anonymous referees for their helpful comments. A preliminary version of this article was presented at the fourth EEM conference in Cracow and at the ECF 2018 in Bolzano. We are thankful to all participants for their suggestions. We both have no declaration of interests.

REFERENCES

Bigerna, S. and C.A. Bollino (2015). "A system of hourly demand in the Italian electricity market." *The Energy Journal* 36: 129–147. https://doi.org/10.5547/01956574.36.4.sbig.

Bello A., J. Reneses, A. Muñoz, and A. Delgadillo (2016). "Probabilistic forecasting of hourly electricity prices in the medium-term using spatial interpolation techniques." *International Journal of Forecasting* 32: 966–980. https://doi.org/10.1016/j.ijforecast.2015.06.002.

Bordignon, S., D.W. Bunn, F. Lisi, and F. Nan (2013). "Combining day-ahead forecasts for British electricity prices." *Energy Economics* 35: 88–103. https://doi.org/10.1016/j.eneco.2011.12.001.

Conejo, A.J., Contreras J., Espinola R., and M.A. Plazas (2005). "Forecasting electricity prices for a day-ahead pool-based electric energy market." *International Journal of Forecasting* 21: 435–462. https://doi.org/10.1016/j.ijforecast.2004.12.005.

Cuaresma, J.C., J. Hlouskova, S. Kossmeier, and M. Obersteiner (2004). "Forecasting electricity spot prices using linear univariate time-series models." *Applied Energy* 77: 87–106. https://doi.org/10.1016/S0306-2619(03)00096-5.

Diebold, F.X. (2014) "Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold-Mariano tests." *Journal of Business & Economic Statistics* 33: 1–9. https://doi.org/10.1080/07350015.2014.983236.

Diebold, F.X. and R.S. Mariano (1995). "Comparing predictive accuracy." *Journal of Business and Economic Statistics* 13: 253–263. https://doi.org/10.1080/07350015.1995.10524599.

Dudek, G. (2016). "Multilayer perceptron for GEFCom2014 probabilistic electricity price forecasting." *International Journal of Forecasting* 32: 1057–1060. https://doi.org/10.1016/j.ijforecast.2015.11.009.

Fabra, N. and J. Toro (2005). "Price wars and collusion in the Spanish electricity market." *International Journal of Industrial Organization* 23: 155–181. https://doi.org/10.1016/j.ijindorg.2005.01.004.

Fezzi, C. and D. Bunn (2010). "Structural analysis of electricity demand and supply interactions." *Oxford Bulletin of Economics and Statistics* 72: 827–856. https://doi.org/10.1111/j.1468-0084.2010.00596.x.

Gaillard, P., Y. Goude, and R. Nedellec (2016). "Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting." *International Journal of Forecasting* 32: 1038–1050. https://doi.org/10.1016/j.ijforecast.2015.12.001.

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

Gianfreda, A., I. Parisio, and M. Pelagatti (2019). "The RES induced switching effect across fossil fuels: an analysis of day-ahead and balancing prices." *The Energy Journal* 40: 365–384. https://doi.org/10.5547/01956574.40.1.agia.

Grossi, L. and F. Nan (2015). "Robust estimation of regime switching models." In *Advances in Statistical Models for Data Analysis*, I. Morlini, T. Minerva T. and M. Vichi, eds. New York: Springer, pp. 125–135. https://doi.org/10.1007/978-3-319-17377-1_14.

Haldrup, N. and M.Ø. Nielsen (2006). "A regime switching long memory model for electricity prices." *Journal of Econometrics* 135: 349–376. https://doi.org/10.1016/j.jeconom.2005.07.021.

Hansen, P.R and A. Timmermann (2015). "Equivalence between out-of-sample forecast comparisons and Wald statistics." *Econometrica* 83: 2485–2505. https://doi.org/10.3982/ECTA10581.

Harvey, D., S. Leybourne, and P. Newbold (1997). "Testing the equality of prediction mean squared errors." *International Journal of Forecasting* 13: 281–291. https://doi.org/10.1016/S0169-2070(96)00719-4.

Hong, T. (2015). "Crystal ball lessons in predictive analytics." *EnergyBiz* 12: 35–37.

Hortaçsu, A. and S.L. Puller (2008). "Understanding strategic bidding in multi-unit auctions: a case study of the Texas electricity spot market." *The RAND Journal of Economics* 39: 86–114. https://doi.org/10.1111/j.0741-6261.2008.00005.x.

Hubicka, K., G. Marcjasz, and R. Weron (2019). "A note on averaging day-ahead electricity price forecasts across calibration windows." *IEEE Transactions on Sustainable Energy* 10: 321–323. https://doi.org/10.1109/TSTE.2018.2869557.

Huisman, R., C. Huurman, and R. Mahieu (2007). "Hourly electricity prices in day-ahead markets." *Energy Economics* 29: 240–248. https://doi.org/10.1016/j.eneco.2006.08.005.

Hyndman, R.J and G. Athanasopoulos (2014). *Forecasting: principles and practice*. Heathmont: OTexts Publishing.

Copyright © 2020 by the IAEE. All rights reserved.

Ito, K. and M. Reguant (2016). "Sequential markets, market power and arbitrage." *American Economic Review* 106: 1921–1957. https://doi.org/10.1257/aer.20141529.

Inoue, A., I. Jin, and B. Rossi (2017). "Rolling window selection of out-of-sample forecasting with time-varying parameters." *Journal of Econometrics* 196: 55–67. https://doi.org/10.1016/j.jeconom.2016.03.006.

Kristiansen, T. (2012). "Forecasting Nord Pool day-ahead prices with an autoregressive model." *Energy Policy* 49: 328–332. https://doi.org/10.1016/j.enpol.2012.06.028.

Koopman, S.J., M. Ooms, and M.A. Carnero (2007). "Periodic seasonal reg-ARFIMA-GARCH models for daily electricity spot prices." *Journal of the American Statistical Association* 102: 16–27. https://doi.org/10.1198/016214506000001022.

Lisi, F. and F. Edoli (2018) "Analyzing and forecasting zonal imbalance signs in the Italian electricity market." *The Energy Journal* 39: 1–19. https://doi.org/10.5547/01956574.39.5.flis.

Maciejowska, K. and J. Nowotarski (2016). "A hybrid model for GEFCom2014 probabilistic electricity price forecasting." *International Journal of Forecasting* 32: 1051–1056. https://doi.org/10.1016/j.ijforecast.2015.11.008.

Maciejowska, K., J. Nowotarski, and R. Weron (2016) "Probabilistic forecasting of electricity spot prices using Factor Quantile Regression Averaging." *International Journal of Forecasting* 32: 957–965. https://doi.org/10.1016/j.ijforecast.2014.12.004.

Marcjasz, M., T. Serafin, and R. Weron (2018a). "Selection of calibration windows for day-ahead electricity price forecasting." *Energies* 11: 2364. https://doi.org/10.3390/en11092364.

Marcjasz, M., B. Uniejewski, and R. Weron (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. https://doi.org/10.1016/j.ijforecast.2017.11.009.

Mirakyan, A., M. Meyer-Renschhausen, and A. Koch (2017). "Composite forecasting approach, application for next-day electricity price forecasting." *Energy Economics* 66: 228–237. https://doi.org/10.1016/j.eneco.2017.06.020.

Misiorek, A., S. Trueck, and R. Weron (2006). "Point and interval forecasting of spot electricity prices: linear vs. non-linear time series models." *Studies in Nonlinear Dynamics and Econometrics* 10: 2. https://doi.org/10.2202/1558-3708.1362.

Nogales, F.J., J. Contreras, A.J. Conejo, and R. Espinola (2002). "Forecasting next-day electricity prices by time series models." *IEEE Transactions on Power Systems* 17: 342–348. https://doi.org/10.1109/TPWRS.2002.1007902.

Nowotarski, J., E. Raviv, S. Trück, and R. Weron (2014). "An empirical comparison of alternative schemes for combining electricity spot price forecasts." *Energy Economics* 46: 395–412. https://doi.org/10.1016/j.eneco.2014.07.014.

Nowotarski, J. and R. Weron (2016). "On the importance of the long-term seasonal component in day-ahead electricity price forecasting." *Energy Economics* 57: 228–235. https://doi.org/10.1016/j.eneco.2016.05.009.

Pesaran, M.H. and A. Timmermann (2007). "Selection of estimation window in the presence of breaks." *Journal of Econometrics* 137: 134–161. https://doi.org/10.1016/j.jeconom.2006.03.010.

Riedmiller, M. and H. Braun (1993). "A direct adaptive method for faster backpropagation learning: The RPROP algorithm." *Proceedings of the IEEE International Conference on Neural Networks (ICNN)* 586–591. https://doi.org/10.1109/ICNN.1993.298623.

Singhal, D. and K.S. Swarup (2011). "Electricity price forecasting using artificial neural networks." *Electrical Power and Energy Systems* 33: 550–555. https://doi.org/10.1016/j.ijepes.2010.12.009.

Steinert, R. and F. Ziel (2019). "Short- to mid-term day-ahead electricity price forecasting using futures." *The Energy Journal* 40: 105–127. https://doi.org/10.5547/01956574.40.1.rste.

Uniejewski, B., R. Weron, and F. Ziel (2018). "Variance stabilizing transformations for electricity spot price forecasting." *IEEE Transactions on Power Systems* 33: 2219–2229. https://doi.org/10.1109/TPWRS.2017.2734563.

Weron, R. (2006). *Modeling and forecasting electricity loads and prices: a statistical approach*. Chichester: John Wiley & Sons. https://doi.org/10.1002/9781118673362.

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. https://doi.org/10.1016/j.ijforecast.2014.08.008.

Weron, R. and A. Misiorek (2008). "Forecasting spot electricity prices: a comparison of parametric and semiparametric time series models." *International Journal of Forecasting* 24: 744–763. https://doi.org/10.1016/j.ijforecast.2008.08.004.

Zhang, G., B.E. Patuwo, and M.Y. Hu (1998). "Forecasting with artificial neural networks: the state of the art." *International Journal of Forecasting* 14: 35–62. https://doi.org/10.1016/S0169-2070(97)00044-7.

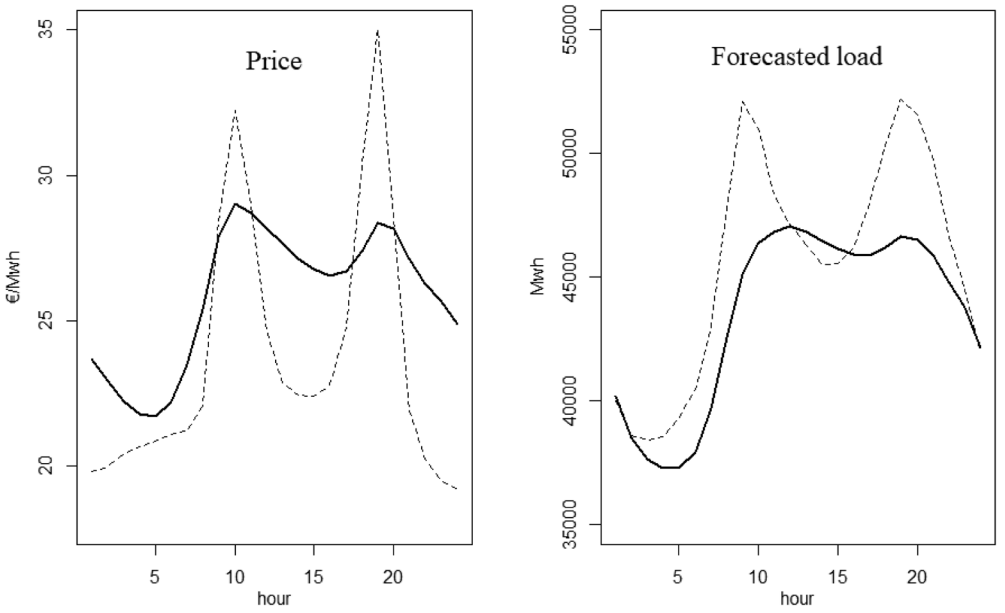
Ziel, F. and R. Weron (2018). "Day-ahead electricity price forecasting with high-dimensional structures: univariate vs. multivariate modeling frameworks." *Energy Economics* 70: 396–420. https://doi.org/10.1016/j.eneco.2017.12.016.

All rights reserved. Copyright © 2020 by the IAEE.

APPENDIX: ADDITIONAL TABLES AND FIGURES

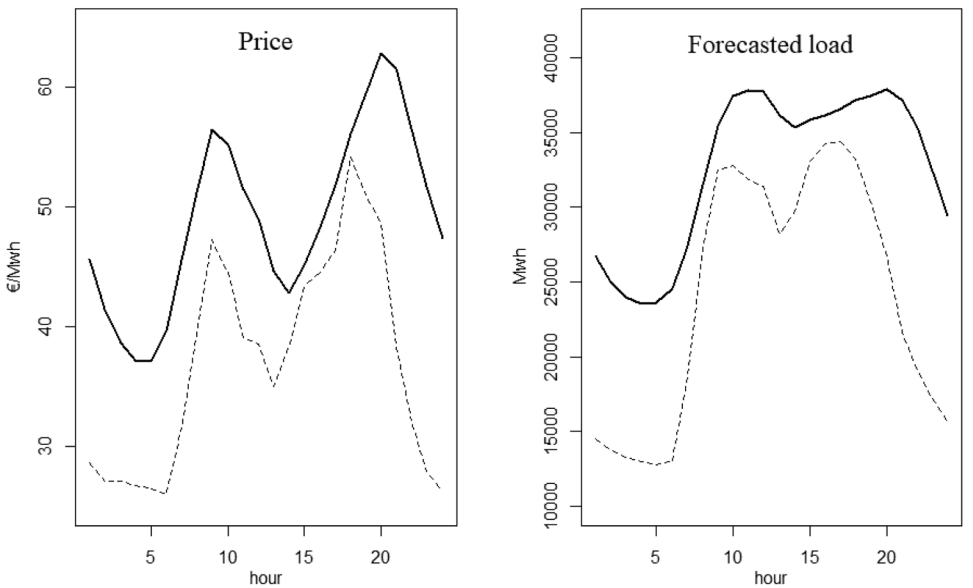
In this appendix, we report some Tables and Figures that may provide additional valuable information to the interested reader, but that we removed from the main manuscript in order to preserve space.

Figure A1: Nord Pool hourly averages (bold) and standard deviations (dashed)



Notes: standard deviation of price multiplied by three and standard deviation of load multiplied by five to preserve scale.

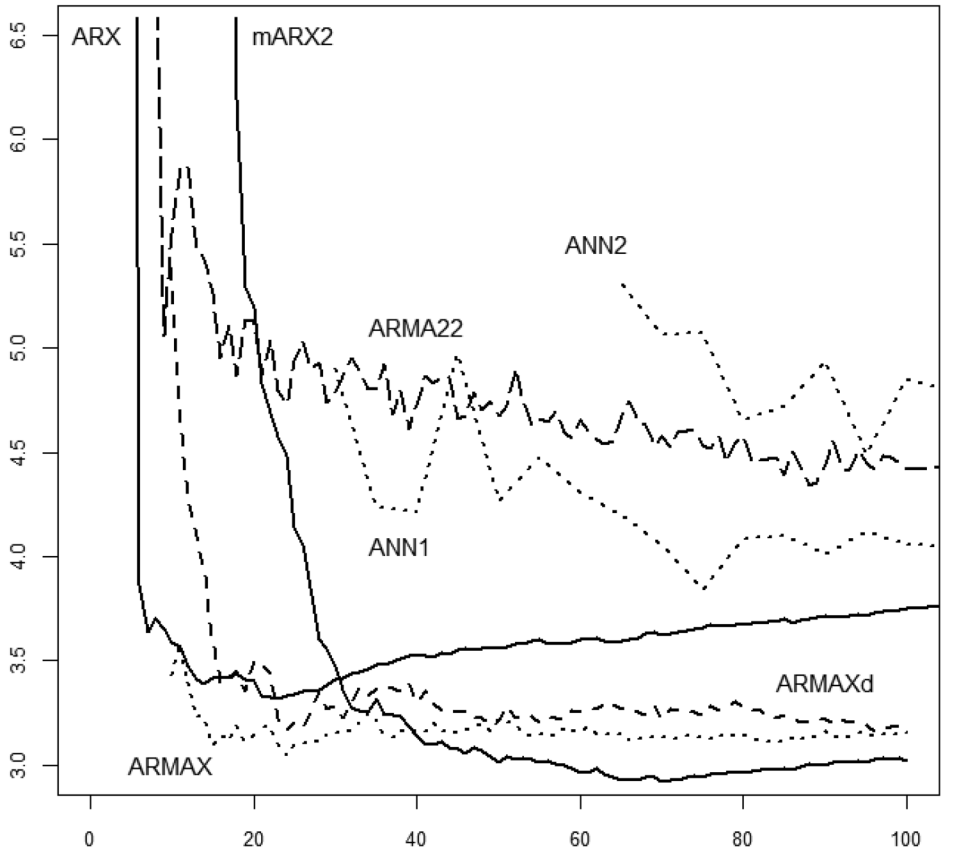
Figure A2: IPEX hourly averages (bold) and standard deviations (dashed)



Notes: standard deviation of price multiplied by three and standard deviation of load multiplied by five to preserve scale.

Copyright © 2020 by the IAEE. All rights reserved.

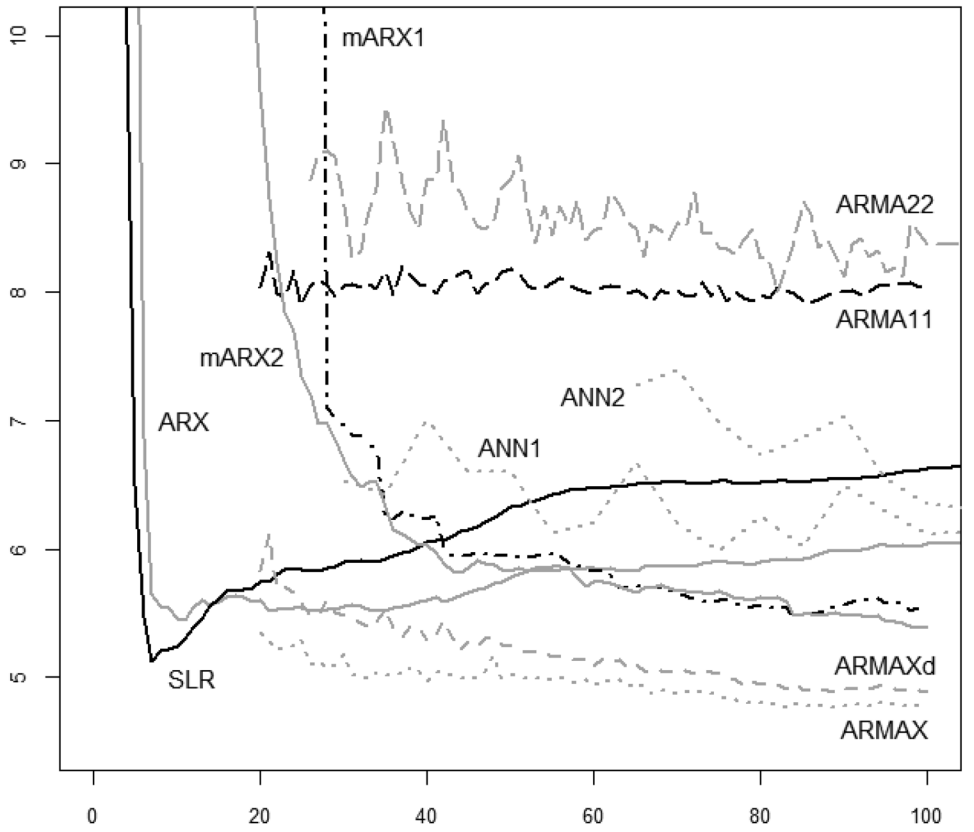
Figure A3: Forecasting performance at different estimation sample lengths



Notes: Forecasting performance measured via the Mean Absolute Error (MAE). All values refer to hour 10 (peak), Nord Pool price for year 2015. This picture represents all the models non reported in Figure 3 of the main paper.

All rights reserved. Copyright © 2020 by the IAEE.

Figure A4: Forecasting performance at different estimation sample lengths (IPEX)



Notes: Forecasting performance measured via the Mean Absolute Error (MAE). All values refer to hour 10 (peak), IPEX price for year 2015. This picture includes all the models in the main paper, with the black lines representing the three models in Figure 3 and the gray lines all the additional models in Figure A3.

Copyright © 2020 by the IAEE. All rights reserved.

Table A1: MAEs for all models estimated with λ^*, window (c), Nord Pool

	ARMAX(1,1)		ARMAX(1,1)d		ARMA(1,1)		ARMA(2,2)		mARX2		SLR		ARX		mARX1		ANN(4)		ANN(7)	
hours	**MAE**	λ^*	**MAE**	λ^*	**MAE**	λ^*	**MAE**	λ^*	**MAE**	λ^*	**MAE**	λ^*	**MAE**	λ^*	**MAE**	λ^*	**MAE**	λ^*	**MAE**	λ^*
1	1.233	*rec*	1.233	*rec*	1.247	*rec*	1.379	*97*	1.168	*rec*	1.447	*6*	1.448	*33*	1.315	*rec*	1.482	*350*	1.279	*rec*
2	1.210	*rec*	1.175	*rec*	1.295	*rec*	1.406	*97*	1.082	*56*	1.497	*6*	1.365	*rec*	1.766	*350*	1.276	*rec*	1.395	*350*
3	1.365	*rec*	1.334	*rec*	1.426	*rec*	1.420	*rec*	1.336	*58*	1.680	*11*	1.512	*rec*	1.407	*rec*	1.404	*350*	1.409	*rec*
4	1.494	*rec*	1.460	*rec*	1.538	*rec*	1.528	*rec*	1.530	*57*	1.715	*6*	1.609	*rec*	1.487	*rec*	1.451	*rec*	1.456	*rec*
5	1.517	*rec*	1.489	*rec*	1.604	*rec*	1.563	*rec*	1.550	*59*	1.660	*6*	1.679	*rec*	1.568	*rec*	1.444	*rec*	1.507	*rec*
6	1.409	*rec*	1.384	*rec*	1.725	*rec*	1.638	*rec*	1.417	*57*	1.529	*6*	1.710	*rec*	1.569	*rec*	1.435	*rec*	1.426	*rec*
7	1.435	*rec*	1.426	*rec*	2.249	*rec*	2.050	*rec*	1.637	*rec*	1.493	*6*	1.688	*19*	1.611	*rec*	1.740	*350*	1.529	*rec*
8	1.850	*23*	1.887	*rec*	2.971	*24*	2.794	*rec*	2.368	*rec*	1.807	*6*	1.964	*14*	2.019	*rec*	2.399	*250*	2.225	*350*
9	3.083	*25*	3.273	*rec*	4.788	*24*	4.284	*rec*	3.197	*rec*	2.867	*6*	3.224	*14*	3.364	*95*	3.587	*150*	3.981	*350*
10	3.510	*24*	3.703	*rec*	5.035	*22*	4.582	*rec*	3.412	*rec*	3.188	*6*	3.674	*23*	3.799	*95*	4.201	*75*	3.826	*rec*
11	3.080	*27*	3.346	*rec*	4.551	*350*	4.003	*rec*	2.989	*rec*	2.934	*6*	3.325	*23*	3.554	*95*	4.242	*350*	4.132	*350*
12	2.472	*24*	2.617	*rec*	3.459	*26*	3.106	*rec*	2.396	*rec*	2.307	*6*	2.631	*23*	3.065	*350*	3.034	*100*	2.700	*rec*
13	2.153	*200*	2.225	*200*	2.807	*25*	2.598	*rec*	2.069	*rec*	1.951	*6*	2.261	*23*	2.585	*350*	2.933	*350*	2.661	*350*
14	1.976	*200*	2.060	*200*	2.687	*24*	2.433	*rec*	1.964	*rec*	1.801	*6*	2.089	*14*	2.061	*rec*	2.543	*350*	2.344	*350*
15	1.905	*rec*	1.933	*rec*	2.621	*24*	2.428	*rec*	1.953	*rec*	1.791	*6*	2.101	*14*	2.014	*rec*	2.662	*350*	2.113	*rec*
16	1.818	*rec*	1.846	*rec*	2.480	*24*	2.321	*rec*	1.917	*rec*	1.759	*6*	2.093	*23*	1.960	*rec*	2.632	*350*	1.978	*rec*
17	2.077	*rec*	2.114	*rec*	2.805	*24*	2.539	*rec*	2.238	*rec*	1.932	*6*	2.280	*23*	2.239	*rec*	3.022	*350*	2.275	*rec*
18	2.794	*rec*	2.820	*rec*	3.474	*24*	3.201	*rec*	2.909	*rec*	2.492	*6*	2.965	*23*	3.264	*350*	3.766	*350*	4.042	*350*
19	3.744	*250*	3.488	*rec*	4.012	*350*	3.870	*rec*	3.614	*rec*	3.072	*7*	3.600	*23*	4.075	*350*	4.610	*350*	4.809	*350*
20	3.224	*300*	3.041	*rec*	3.857	*25*	3.267	*rec*	3.100	*rec*	2.769	*7*	3.161	*25*	3.460	*350*	3.824	*350*	3.084	*rec*
21	1.658	*rec*	1.672	*rec*	1.884	*rec*	1.957	*350*	1.787	*rec*	1.752	*6*	1.877	*28*	1.822	*rec*	2.072	*rec*	1.920	*rec*
22	1.255	*300*	1.201	*rec*	1.353	*rec*	1.432	*200*	1.309	*rec*	1.387	*6*	1.428	*31*	1.305	*rec*	1.407	*rec*	1.396	*rec*
23	1.064	*300*	1.026	*rec*	1.062	*rec*	1.079	*rec*	1.103	*rec*	1.226	*6*	1.195	*100*	1.101	*rec*	1.104	*rec*	1.135	*rec*
24	0.966	*rec*	0.970	*rec*	1.017	*rec*	1.060	*100*	1.070	*rec*	1.281	*5*	1.011	*rec*	1.065	*rec*	1.121	*350*	1.246	*300*

All rights reserved. Copyright © 2020 by the IAEE.

Table A2: MAEs for all models estimated with λ^*, window (c), IPEX

	ARMAX(1,1)		ARMAX(1,1)d		ARMA(1,1)		ARMA(2,2)		mARX2		SLR		ARX		mARX1		ANN(4)		ANN(7)	
hours	**MAE**	λ^*	**MAE**	λ^*	**MAE**	λ^*	**MAE**	λ^*	**MAE**	λ^*	**MAE**	λ^*	**MAE**	λ^*	**MAE**	λ^*	**MAE**	λ^*	**MAE**	λ^*
1	3.467	*rec*	3.338	*rec*	3.606	*62*	3.439	*rec*	2.647	*350*	3.770	*7*	3.909	*rec*	3.411	*350*	4.144	*rec*	4.001	*rec*
2	3.422	*350*	3.391	*350*	3.669	*350*	3.463	*350*	2.623	*350*	3.556	*7*	3.854	*350*	3.415	*350*	3.864	*300*	3.779	*rec*
3	3.298	*350*	3.260	*350*	3.464	*350*	3.256	*350*	2.530	*350*	3.325	*7*	3.632	*350*	3.234	*350*	3.632	*300*	3.632	*rec*
4	3.357	*350*	3.317	*350*	3.441	*37*	3.286	*350*	2.691	*350*	3.333	*7*	3.656	*350*	3.273	*350*	3.769	*350*	3.567	*350*
5	3.388	*33*	3.344	*350*	3.412	*44*	3.287	*350*	2.820	*350*	3.415	*13*	3.273	*32*	3.361	*350*	3.673	*350*	3.712	*rec*
6	3.138	*63*	3.187	*65*	3.291	*63*	3.221	*300*	2.735	*350*	3.307	*8*	3.236	*32*	3.275	*350*	3.783	*350*	3.782	*rec*
7	3.339	*350*	3.280	*350*	4.748	*350*	4.210	*350*	3.029	*350*	3.389	*7*	3.813	*35*	3.450	*350*	4.648	*350*	3.994	*350*
8	3.828	*50*	4.008	*250*	6.655	*150*	6.161	*300*	4.167	*250*	3.820	*7*	4.378	*21*	4.300	*350*	6.125	*350*	4.522	*350*
9	4.859	*250*	4.860	*250*	8.359	*150*	8.018	*67*	4.859	*250*	5.004	*7*	5.422	*10*	5.523	*300*	7.364	*300*	5.760	*350*
10	4.917	*200*	4.961	*250*	6.872	*25*	7.113	*300*	4.863	*250*	4.795	*7*	5.194	*11*	5.305	*250*	8.126	*rec*	5.825	*rec*
11	4.355	*86*	4.611	*250*	6.630	*350*	5.933	*350*	4.541	*250*	4.403	*7*	4.568	*25*	4.809	*250*	6.984	*rec*	5.318	*rec*
12	4.139	*86*	4.401	*250*	6.146	*250*	5.371	*300*	4.305	*250*	4.423	*9*	4.551	*42*	4.703	*250*	5.863	*350*	5.473	*300*
13	3.824	*250*	3.829	*250*	5.125	*150*	4.603	*300*	3.847	*250*	3.984	*11*	5.323	*350*	4.204	*250*	5.117	*350*	4.367	*rec*
14	3.947	*81*	3.692	*79*	5.848	*150*	5.158	*300*	4.117	*250*	4.027	*7*	4.217	*29*	4.054	*84*	5.390	*350*	4.450	*350*
15	4.468	*84*	4.366	*79*	6.936	*150*	6.198	*300*	4.373	*350*	4.605	*7*	4.787	*29*	4.938	*150*	6.598	*rec*	5.078	*rec*
16	4.238	*96*	4.227	*96*	7.217	*300*	6.136	*300*	4.371	*350*	4.354	*7*	4.557	*19*	4.918	*350*	6.279	*350*	5.133	*350*
17	4.261	*30*	4.410	*97*	6.559	*250*	5.963	*350*	4.581	*250*	4.509	*7*	4.891	*11*	4.970	*250*	7.375	*rec*	5.683	*300*
18	4.593	*22*	4.505	*200*	6.180	*200*	5.612	*350*	4.884	*350*	4.847	*7*	4.909	*10*	5.161	*350*	6.796	*rec*	5.114	*rec*
19	5.318	*300*	5.376	*300*	6.274	*25*	6.061	*250*	5.342	*rec*	5.646	*8*	5.648	*22*	5.750	*250*	6.467	*350*	5.656	*rec*
20	5.434	*200*	5.537	*300*	6.249	*200*	5.888	*300*	5.488	*rec*	5.915	*8*	6.487	*250*	5.774	*300*	7.119	*350*	7.334	*350*
21	5.610	*200*	5.746	*250*	6.219	*250*	5.785	*250*	5.605	*350*	5.827	*7*	6.161	*350*	5.737	*350*	6.459	*350*	7.074	*350*
22	4.557	*83*	4.533	*350*	4.817	*25*	4.458	*300*	4.418	*300*	4.767	*7*	4.674	*21*	4.567	*300*	5.187	*250*	5.070	*300*
23	3.312	*300*	3.226	*350*	3.397	*300*	3.235	*350*	3.195	*350*	3.523	*8*	3.458	*22*	3.234	*350*	3.562	*350*	3.659	*300*
24	2.748	*350*	2.743	*350*	2.812	*61*	2.682	*350*	2.682	*350*	2.861	*8*	2.978	*350*	2.642	*350*	2.950	*350*	2.895	*350*

Copyright © 2020 by the IAEE. All rights reserved.

Table A3: Forecasting performance for the best models in each hour, Nord Pool and IPEX price in year 2016 (window c)

	Nord Pool				IPEX			
	$\lambda=\lambda^*$		$\lambda=300$		$\lambda=\lambda^*$		$\lambda=300$	
hour	**Model (λ^*)**	**RMSE**	**Model**	**RMSE**	**Model (λ^*)**	**RMSE**	**Model**	**RMSE**
1	mARX2 (rec)	**2.39**	mARX2	2.48	mARX2 (350)	3.57	mARX2	**3.54**
2	mARX2 (56)	1.49	mARX2	**1.28**	mARX2 (350)	**3.49**	mARX2	3.49
3	ARMAXd11 (rec)	1.76	mARX2	**1.53**	mARX2 (350)	3.29	mARX2	**3.28**
4	ANN4 (rec)	1.87	mARX2	**1.78**	mARX2 (350)	**3.38**	mARX2	3.36
5	ANN4 (rec)	**1.89**	mARX2	2.06	mARX2 (350)	3.58	mARX2	**3.57**
6	ARMAXd11 (rec)	**1.87**	ARMAX11	2.06	mARX2 (350)	**3.55**	mARX2	3.58
7	ARMAXd11 (rec)	**1.87**	ARMAX11	2.04	mARX2 (350)	**4.05**	mARX2	4.15
8	SLR (6)	2.66	mARX2	**2.63**	SLR (7)	**5.51**	ARMAXd11	5.66
9	SLR (6)	**6.40**	mARX2	7.10	ARMAX11 (250)	6.85	mARX2	**6.71**
10	SLR (6)	**7.32**	mARX2	9.08	SLR (7)	7.02	mARX2	**6.86**
11	SLR (6)	**6.42**	mARX2	7.67	ARMAX11 (86)	**6.05**	mARX2	6.14
12	SLR (6)	**4.50**	mARX2	5.05	ARMAX11 (86)	**5.67**	mARX2	5.74
13	SLR (6)	**3.07**	ARMAX11	3.46	ARMAX11 (250)	5.07	ARMAX11	**5.02**
14	SLR (6)	**2.80**	ARMAX11	3.1	ARMAXd11 (79)	**4.97**	ARMAXd11	5.05
15	SLR (6)	**2.81**	ARMAX11	3.06	ARMAXd11 (79)	6.23	ARMAX11	**6.19**
16	SLR (6)	**2.73**	ARMAX11	2.91	ARMAXd11 (96)	**5.95**	mARX2	6.07
17	SLR (6)	**3.49**	ARMAX11	4.26	ARMAX11 (30)	**6.78**	ARMAX11	7.05
18	SLR (6)	**5.48**	ARMAX11	7.62	ARMAXd11 (200)	**6.98**	ARMAX11	7.27
19	SLR (7)	**7.06**	ARMAX11	10.35	ARMAX11 (300)	**8.32**	ARMAX11	**8.32**
20	SLR (7)	**5.62**	ARMAX11	7.54	ARMAX11 (200)	**7.91**	ARMAX11	7.95
21	ARMAX11 (rec)	3.10	ARMAX11	**3.03**	mARX2 (350)	**7.78**	mARX2	7.91
22	ARMAXd11 (rec)	2.08	ARMAX11	**2.05**	mARX2 (300)	**6.02**	mARX2	**6.02**
23	ARMAXd11 (rec)	**1.46**	ARMAX11	1.46	mARX2 (350)	**4.44**	ARMA22	4.60
24	ARMAX11 (rec)	**1.32**	ARMAX11	1.36	mARX1 (350)	**3.53**	ARMA22	3.58

Notes: This table replicates Table 7 in the main paper using, as a measure of forecasting performance, Root Mean Squared Error (RMSE) instead of Mean Absolute Error (MAE). Like in Table 7, RMSE is calculated on the data in year 2016. Highlighted and bold is the best RMSE for each hour and market.

All rights reserved. Copyright © 2020 by the IAEE.