DEM Working Papers

N. 2018/10

**Size matters: Estimation sample length and
electricity price forecasting accuracy**

Carlo Fezzi, Luca Mosetti

UNIVERSITÀ DEGLI STUDI DI TRENTO
Dipartimento di Economia e Management

Università degli Studi di Trento

Department of Economics and Management, University of Trento, Italy.

Editors

Luciano ANDREOZZI   luciano.andreozzi@unitn.it

Roberto GABRIELE   roberto.gabriele@unitn.it

Technical officer

Marco TECILLA   marco.tecilla@unitn.it

Guidelines for authors

Papers may be written in Italian or in English. Faculty members of the Department must submit to one of the editors in pdf format. Management papers should be submitted to R. Gabriele. Economics Papers should be submitted to L. Andreozzi. External members should indicate an internal faculty member that acts as a referee of the paper.

Typesetting rules:

1. papers must contain a first page with title, authors with emails and affiliations, abstract, keywords and codes. Page numbering starts from the first page;
2. a template is available upon request from the managing editors.

Size matters:
Estimation sample length and electricity price forecasting accuracy¹

by Carlo Fezzi² and Luca Mosetti³

December 2018

Abstract

Electricity price forecasting models are typically estimated via rolling windows, i.e. by using only the most recent observations. Nonetheless, the current literature does not provide much guidance on how to select the 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 with a selection of ten different forecasting models. Incidentally, our empirical application reveals that simple models, such as the linear regression, can perform surprisingly well if estimated on extremely short samples.

Keywords: electricity price forecasting, day-ahead market, parameter instability, bandwidth selection, artificial neural networks.

JEL classification: C22, C45, C51, C53, Q47.

---

¹ Many thanks to Fany Nan and Rafal Weron for sharing with us the IPEX and Nord Pool data. An older version of this paper was presented at the fourth EEM conference in Cracow. We are thankful to all participants for their helpful comments. We both have no declaration of interests.

² University of Exeter Business School (UEBS), University of Exeter and Department of Economics and Management (DEM), University of Trento.

³ Department of Computer Science, ETH Zürich.

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. For example, recent back-of-the-envelope calculations (Hong, 2015) show that a 1% increase in forecasting accuracy would generate about \$3 million per years in savings for a typical medium-size production company.

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, 2016). The accurate short-term predictions that typically characterize demand (which, in this literature, is often referred as “load”) are still a long way to go for prices and, ultimately, might turn out to be unfeasible (Bunn, 2000).

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.

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 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. Conejo et al., 2005; Dudek, 2016; Nowotarski and Weron, 2016), 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.^{4,5}

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 articles studying this issue in EPF: Hubicka et al. (2018) and Marcjasz et al. (2018a). Both papers develop weighting schemes to average predictions from linear regression models estimated on different windows sizes. They show that using appropriate weights improves upon selecting a single window size when averaging prediction errors across all 24 trading hours of the day.

In this paper, we develop 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, encompassing time series, regressions and computational intelligence methods. Importantly, rather than evaluating the best performing model using averages across all 24 hourly predictions in a day, we evaluate predictions for each hourly trading period separately. This comparison reveals that both optimal window size and best performing models change greatly

---

⁴ 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. (2018b) use one year windows.

⁵ 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.

across hours, with the stable baseload hours favoring long windows and complex models, and the more volatile peak hours selecting short samples and relatively simple specifications. Therefore, we recommend using different models and window sizes for different hours.

Furthermore, our empirical analysis, based on datasets from two large power markets (the Nord Pool and the Italian Power Exchange, IPEX), exposes what is perhaps an unexpected result. Very straightforward models, such as the simple linear regression, can perform surprisingly well if estimated on extremely small samples of just six or seven observations. The simplicity of such models becomes particularly handy when market conditions are the most volatile, i.e. during peak-demand hours. In the Nord Pool, for example, after allowing the rolling window size to vary, a simple linear regression with only forecasted load as explanatory variable becomes the overall best performing model in 50% of the 24 price time series.

The rest of the paper is organized as follows. Section 2 introduces the two markets and datasets. Section 3 presents our selection of forecasting models and the two-steps approach. Section 4 illustrates our findings, first focusing on the Nord Pool, then on the IPEX, and finally providing an overall analysis of the benefits provided by our rolling window size selection method. Section 5 concludes, indicating avenues for further research.

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 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.⁶ In both the NP and the IPEX (as in most electricity markets) the clearing prices and quantiles for

---

⁶ 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.

each hour of the day are generated via 24 simultaneous auctions taking place the day before the delivery.

The two markets are of roughly 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 about here]

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.

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 import (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. It 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) and Gianfreda et al. (2016).

[Figure 2 about here]

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 approach 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 how we 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) and *b*) computational intelligence algorithms, in particular neural networks (e.g. Dudek, 2016; Marcjasz et al., 2018b). Therefore, rather than trying to replicate the extensive and continuously growing collection of EPF methods (a virtually

impossible task), we illustrate our two-step methodology by considering 10 different models belonging to these two large and established classes of approaches.

Our selection, summarized in Table 1, includes models with different levels of complexity, going from a simple linear regression with only two coefficients to an artificial neural network of 64 parameters. 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.⁷ 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”.⁸

[Table 1 about here]

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 surpass this apparently easy test (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):

$$(1) p_t = \beta_0 + \beta_1 q_t + \varepsilon_t ,$$

---

⁷ 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 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.

⁸ 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.

Where t represents time, ε_t is the residual component and β_0, β_l are the parameters that we estimate via Ordinary Least Squares (OLS). Assuming an inelastic short-term demand, β_l can be interpreted as the price-elasticity of electricity supply. While this simple approach has not found many applications in the literature, our analysis will show that this neglected model can provide impressive results by using particularly short estimation windows.

We then consider a number of time series models, 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:

$$(2) \Phi_i(B)p_t = \Theta_j(B)\varepsilon_t ,$$

where $\Phi_i(B) = 1 - \phi_1 B - \dots - \phi_i B^i$ is the autoregressive polynomial; $\Theta_j(B) = \theta_0 + \theta_1 B + \dots + \theta_j B^j$ 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 $\Phi_i(B)$ and $\Theta_j(B)$ 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) models.

ARMA models can be augmented by including explanatory (or exogenous) variables. This class of specifications, refereed 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:

$$(3) \Phi_i(B)[p_t - \beta_0 - \boldsymbol{\beta}'\mathbf{X}_t] = \Theta_j(B)\varepsilon_t ,$$

where $\mathbf{X}_t$ 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 forecasted logarithm of load as explanatory variable, and an ARMAXd(1,1) with forecasted logarithm of load and three dummy variables for Saturday, Sunday and Monday (Misiorek et al., 2006).

The last and most complex model belonging to the class of statistical approaches is the *multi-day ARX*, or mARX, introduced by Maciejowska and Nowotarski (2016) and further developed by Nowotarski and Weron (2016). This is slightly richer version of the model used by Hubicka et al. (2018) to study the effect of averaging forecast across different estimation windows. It is essentially an ARX with a different set of explanatory variables depending on the day of the week it is predicting. In compact notation:

$$(3) p_t = \sum_{i=0}^3 d_i D_i + (\sum_{i=0}^3 \phi_{1,i} D_i) p_t + \phi_2 p_{t-2} + \phi_3 D_1 p_{t-3} + \phi_7 p_{t-7} + \phi_8 p_{m,t-1} + \beta_1 q_t + \varepsilon_t,$$

where $p_{m,t-1}$ is the minimum of the 24 hourly prices in the previous day and $D_1, \dots, D_3$ are respectively dummy variables for Monday, Saturday and Sunday (D_0 is a column of ones so d_i is the model intercept). The parameter ϕ_3 represents the effect of price in the previous Friday on Monday prices, while ϕ_8 create the link with the overall supply function of the previous day.

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., 2018b). 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 refer to this latter parameter as the “bias”. 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 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 mARX. 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. We estimate both ANNs 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

Our proposed approach is extremely simple and consists in two steps designed to compare forecasting performance both across models and across window sizes, and select the best performing specification for each hour.

In the first step, we identify, for each model and each hour, which window size (λ) provides the best performance. We indicate this “optimal” estimation sample length with λ^* . In order to determine λ^* , we compare rolling windows of size $k, k+1, k+2, \dots, 100, 150, 200, \dots, 350$ with $k =$ number of model parameters, which corresponds to the minimum number of observations necessary for estimation. This means that our rolling windows vary in size from 2 days to almost one year. We also test the performance of using a recursive window, i.e. including all available previous observations, effectively increasing the sample size by one with each passing day. 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 windows (a) and (b) in Figure 1 and 2.

In the second step, we estimate each model using its specific λ^* and evaluate 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 window (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):

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

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 mean squared error, and we do not report them in order to preserve space.

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: mARX, SLR and ARMA(1,1). Performances vary considerably, particularly for the first two specifications. Interestingly, they exhibit contrasting behaviors. On one hand, the accuracy of the mARX steadily improves with sample size, reaching a MAE of €3.23 with the highest value of λ . On the other hand, the SLR achieves impressive results with small λ s (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 about here]

These reversed dynamics can be explained by the different structures of the two models. The mARX is relatively complex: it includes 13 parameters in order to capture all the main features of

electricity price dynamics, such as different types of seasonality, relationship with the overall daily supply function, etc. Because of its complexity, it also requires a large window 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 λ s. 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 *a priori*, 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 the same size, i.e. 300 observations. Perhaps surprisingly, but in line with previous findings (Conejo et al., 2005), most models do not outperform the naïve benchmark. Only four models pass this apparently simple test: the ARMAX(1,1), ARMAXd(1,1), mARX and ANN(7). The first one is the best model, with a MAE of €3.17. 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 about here]

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 ARMAX(1,1), its MAE improving to €3.05. This is a relatively simple model and, in fact, its optimal window size is only 24 observations. 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 mARX. 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 undoubtedly improves forecasting performance and significantly changes model ranking.

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. 2016. In Table 3, we focus on the three models estimated via OLS (SLR, ARX(1) and mARX) and compare them with the naïve benchmark. We report estimates for all trading hours using λ^* and $\lambda = 300$. Several findings emerge.

[Table 3 about here]

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 mARX. This feature supports our previous findings showing how simple models perform better with short estimation windows, while complex models require larger samples. *Fourth:* simpler model forecast better during peak hours, while more complex models are better suited for the baseload. More specifically, the SLR estimated on extremely small samples (six or seven observations) is the best performing model for all hours between 8am to 9pm, while mARX and ARX(1) are the preferred models for the evening and early morning hours. *Fifth:* comparing within models but across hours, we notice how the baseload 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 baseload hours select a much larger λ^* and even recursive samples.

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 baseload hours favor more complex specifications. The optimal windows size λ^* changes accordingly, both across-models and within the same model.

[Table 4 about here]

Finally, 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. As for the Nordic market, Table 5 focuses on the SLR, ARX(1), mARX 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 peak hours, and the more complex approaches providing better forecasts during baseload. Perhaps surprisingly, in a few trading periods (e.g. hour 20 and 21) the mARX performs better with $\lambda = 300$ than 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 5 about here]

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 forecast than NP ones, with the average MAEs being much higher than the ones of the Nordic market for all models. 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. The best models are the ARMAX(1,1) and ARMAXd(1,1). 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 more than half of the 24 hours.

[Table 6 about here]

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 43 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 hours. The best models are SLRs with very small samples for the peak hours and more complex models (mainly ARMAX and ANNs) with large, typically recursive samples for the baseload. Interestingly, this means that a fixed estimation window of 300 observations is certainly too long for the peak, but also too short for the baseload, when the best samples are typically recursive. This difference of optimal window size (and models) between peak and baseload can explain why Marcjasz 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 about here]

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 mARX models. However, the two-step method selects again smaller windows during the peak and larger samples for the baseload. This apparently small change allows our approach to outperform the standard practice in 19 out of 24 hours (in two additional hours the approaches are indistinguishable, since the selected window size is equal to the size fixed *a priori*, i.e. 300 observations).

[Figure 4 about here]

Is this improvement in forecasting accuracy statistically significant? To answer this question, 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. The improvement in forecasting performance delivered by the two-step approach is significant at the 5% level for 26 hours and at the 10% level for an additional 5 hours, for a total of 65% of all trading periods. On the other hand, in the only three hours in which the fixed-window approach performs better, MAEs are not significantly different from the ones produced by the two-step method. Overall, selecting the appropriate rolling window size for each trading hour appears to be a very promising and relatively simple strategy to improve forecasting models for electricity prices.

5. Concluding remarks

This analysis investigated the performance of 10 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 hour of the day prior to estimation.

Our results leave little doubt on the advantages provided by selecting suitable rolling window sizes for EPF. Considering both markets, the two-steps method outperforms the standard approach in 45 hours out of 48. In 31 hours, this difference is significant according to Diebold-Mariano (1995) tests. Perhaps surprisingly, using a fixed window provides better forecasting accuracy in 3 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 the development of algorithms for 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 that allow the window size to vary with each new observations and more advanced selection algorithms (e.g. 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. (2018).

Comparing our results across trading periods, it is clear how peak hours typically select simple models and small window sizes, while baseload 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, therefore 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 window sizes.

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.

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.

6. References

Bigerna S., Bollino C.A. (2015) A system of hourly demand in the Italian electricity market, *Energy Journal*, vol. 36, pp. 129-147.

Bello A., Reneses J., Muñoz A., Delgadillo A. (2016) Probabilistic forecasting of hourly electricity prices in the medium-term using spatial interpolation techniques, *International Journal of Forecasting*, vol. 32, pp. 966-980

Bordignon S., Bunn D.W., Lisi F., Nan F. (2013) Combining day-ahead forecasts for British electricity prices, *Energy Economics*, vol. 35, pp. 88–103.

Bunn D. (2000) Forecasting load and prices in competitive power markets, *Proceedings of the IEEE*, vol. 88, pp. 163-169,

Conejo A.J., Contreras J., Espinola R., Plazas M.A. (2005) Forecasting electricity prices for a day-ahead pool-based electric energy market, *International Journal of Forecasting*, vol. 21, pp. 435-462.

Cuaresma J.C., Hlouskova J., Kossmeier S., Obersteiner M. (2004) Forecasting electricity spot prices using linear univariate time-series models, *Applied Energy*, vol. 77, pp. 87-106.

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

Dudek G. (2016) Multilayer perceptron for GEFCom2014 probabilistic electricity price forecasting, *International Journal of Forecasting*, vol. 32, pp. 1057-1060.

Fabra N., Toro J. (2005) Price wars and collusion in the Spanish electricity market, *International Journal of Industrial Organization*, vol. 23, pp. 155-181.

Fezzi C., Bunn D. (2010) Structural analysis of electricity demand and supply interactions, *Oxford Bulletin of Economics and Statistics*, vol. 72, pp. 827-856

Haldrup N., Nielsen M.Ø. (2006) A regime switching long memory model for electricity prices, *Journal of Econometrics*, vol. 135, pp. 349–376.

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

Grossi L., Nan F. (2015) Robust estimation of regime switching models, in Morlini I., Minerva T., Vichi M., (Eds.) *Advances in Statistical Models for Data Analysis*, Springer, pp 125–135.

Hong, T. (2015) Crystal ball lessons in predictive analytics, *EnergyBiz*, vol. 12, pp. 35-37.

Hortaçsu A., Puller S.L. (2008) Understanding strategic bidding in multi-unit auctions: a case study of the Texas electricity spot market, *The RAND Journal of Economics*, vol. 39, pp. 86-114.

Hubicka K., Marcjasz G., Weron R. (2018) A note on averaging day-ahead electricity price forecasts across calibration windows, *IEEE Transactions on Sustainable Energy*, 2018 (forthcoming)

Huisman R., Hurman C., Mahieu R. (2007) Hourly electricity prices in day-ahead markets, *Energy Economics*, vol. 29, pp. 240-248.

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

Ito K., Reguant M. (2016) Sequential markets, market power and arbitrage, *American Economic Review*, vol. 106, pp. 1921-1957.

Inoue A., Jin L., Rossi B. (2017) Rolling window selection of out-of-sample forecasting with time-varying parameters, *Journal of Econometrics*, vol. 196, pp. 55-67.

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

Koopman S.J., Ooms M., Carnero M.A. (2007) Periodic seasonal reg-ARFIMA-GARCH models for daily electricity spot prices, *Journal of the American Statistical Association*, vol. 102, pp. 16-27.

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

Maciejowska K., Nowotarski J., Weron R. (2016) Probabilistic forecasting of electricity spot prices using Factor Quantile Regression Averaging, *International Journal of Forecasting*, vol. 32, pp. 957-965.

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

Marcjasz M., Uniejewski B., Weron R. (2018b) On the importance of the long-term seasonal component in day-ahead electricity price forecasting with NARX neural networks, *International Journal of Forecasting* (in press).

Mirakyan A., Meyer-Renschhausen M., Koch A. (2017) Composite forecasting approach, application for next-day electricity price forecasting, *Energy Economics*, vol. 66, pp. 228-237.

Misiorek A., Trueck S., Weron R. (2006) Point and interval forecasting of spot electricity prices: linear vs. non-linear time series models, *Studies in Nonlinear Dynamics and Econometrics*, vol. 10, article 2.

Nogales F.J., Contreras J., Conejo A.J., Espinola R. (2002). Forecasting next-day electricity prices by time series models, *IEEE Transactions on Power Systems*, vol. 17, pp. 342–348.

Nowotarski J., Raviv E., Trück S., Weron R. (2014) An empirical comparison of alternative schemes for combining electricity spot price forecasts, *Energy Economics*, vol. 46, pp. 395-412.

Nowotarski J., Weron R. (2016) On the importance of the long-term seasonal component in day-ahead electricity price forecasting, *Energy Economics*, vol. 57, pp. 228-235.

Riedmiller M., Braun H. (1993) A direct adaptive method for faster backpropagation learning: The RPROP algorithm. *Proceedings of the IEEE International Conference on Neural Networks (ICNN)*, pages 586-591. San Francisco.

Singhal D., Swarup K.S. (2011) Electricity price forecasting using artificial neural networks, *Electrical Power and Energy Systems*, vol. 33, pp. 550-555.

Uniejewski B., Weron R. Ziel F. (2018) Variance stabilizing transformations for electricity spot price forecasting, *IEEE Transactions on Power Systems*, vol. 33, pp. 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*, vol. 30, pp. 1030–1081.

Weron R., Misiorek A. (2008) Forecasting spot electricity prices: a comparison of parametric and semiparametric time series models, *International Journal of Forecasting*, vol. 24, pp. 744-763.

Zhang G. Patuwo B.E., Hu M.Y. (1998) Forecasting with artificial neural networks: the state of the art, *International Journal of Forecasting*, vol. 14, pp. 35-62.

Tables and Figures

Table 1: Models summary

Model	Class	Parameters	Estimation	Explanatory variables
Naïve	benchmark	N/A	N/A	N/A
SLR	statistical	2	OLS	Forecasted load
ARMA(1,1)	statistical	3	ML	Lagged prices and errors
ARMA(2,2)	statistical	5	ML	Lagged prices and errors
ARX(1)	statistical	3	OLS	Forecasted load and lagged prices
ARMAX(1,1)	statistical	4	ML	Forecasted load, lagged prices and random component
ARMAXd(1,1)	statistical	7	ML	Forecasted load, daily dummies, lagged prices and random component
mARX	statistical	13	OLS	Forecasted load, daily dummies, lagged prices and random component
ANN(4)	computational intelligence	25	RBA	Forecasted load, daily dummies, lagged prices
ANN(7)	computational intelligence	64	RBA	Forecasted load, daily dummies, lagged prices squared of the forecasted load

Notes: OLS = Ordinary Least Squares, ML = Maximum Likelihood, RBA = Resilient Backpropagation Algorithm.

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

Model	Fixed $\lambda=300$		$\lambda=\lambda^*$	
	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
mARX	3.30	300	3.23	100
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.

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

hour	Naïve	SLR		ARX(1)		mARX	
		$\lambda = 300$	$\lambda = \lambda^*$	$\lambda = 300$	$\lambda = \lambda^*$	$\lambda = 300$	$\lambda = \lambda^*$
1	1.69	5.55	1.45 (6)	1.53	1.45 (33)	1.54	1.24 (rec)
2	1.74	5.69	1.50 (6)	1.63	1.36 (rec)	1.56	1.74 (350)
3	1.89	5.89	1.68 (11)	1.89	1.51 (rec)	1.80	2.02 (350)
4	2.03	6.07	1.71 (6)	2.07	1.61 (rec)	1.84	1.54 (rec)
5	2.05	6.22	1.66 (6)	2.16	1.68 (rec)	2.00	1.59 (rec)
6	1.84	6.47	1.53 (6)	2.20	1.71 (rec)	1.79	1.53 (rec)
7	1.84	6.62	1.49 (6)	2.56	1.69 (19)	1.74	1.47 (rec)
8	2.31	6.74	1.81 (6)	3.32	1.96 (14)	2.16	2.27 (350)
9	3.84	7.28	2.87 (6)	5.11	3.22 (14)	3.57	3.39 (100)
10	4.06	7.62	3.19 (6)	5.41	3.67 (23)	3.94	3.83 (100)
11	3.74	7.25	2.93 (6)	4.75	3.32 (23)	3.64	3.40 (rec)
12	3.03	6.50	2.31 (6)	3.66	2.63 (23)	2.91	2.70 (rec)
13	2.56	6.29	1.95 (6)	3.03	2.26 (23)	2.47	2.21 (rec)
14	2.35	6.30	1.80 (6)	2.78	2.09 (14)	2.23	2.02 (rec)
15	2.36	6.22	1.79 (6)	2.81	2.10 (14)	2.19	1.96 (rec)
16	2.35	6.25	1.76 (6)	2.68	2.09 (23)	2.15	1.91 (rec)
17	2.62	6.62	1.93 (6)	2.89	2.28 (23)	2.52	2.16 (rec)
18	3.46	7.69	2.49 (6)	3.55	2.97 (23)	3.33	2.87 (rec)
19	4.32	8.14	3.07 (7)	4.29	3.60 (23)	4.15	3.74 (rec)
20	3.70	7.13	2.77 (7)	3.76	3.16 (25)	3.54	3.14 (rec)
21	2.26	5.96	1.75 (6)	2.09	1.88 (28)	2.04	1.84 (rec)
22	1.71	5.60	1.39 (6)	1.47	1.43 (31)	1.46	1.33 (rec)
23	1.49	5.33	1.23 (6)	1.15	1.19 (100)	1.17	1.11 (rec)
24	1.47	5.26	1.28 (5)	1.10	1.01 (rec)	1.13	1.07 (rec)

Notes: MAE for the naïve model and the three 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.

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

Model	Number of trading hours in which the model is ranked:				$\overline{MAE}$
	1 st	2 nd	3 rd	4 th or more	
SLR	13		1	10	1.97
ARMA(1,1)		1		23	2.58
ARMA(2,2)				24	2.41
ARX(1)			5	19	2.16
ARMAX(1,1)	3	16	1	4	2.01
ARMAXd(1,1)	6	6	7	5	2.03
mARX			5	19	2.17
ANN(4)	2		2	20	2.47
ANN(7)		1	3	20	2.33

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

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

hour	Naïve	SLR		ARX(1)		mARX	
		$\lambda = 300$	$\lambda = \lambda^*$	$\lambda = 300$	$\lambda = \lambda^*$	$\lambda = 300$	$\lambda = \lambda^*$
1	4.13	8.13	3.77 (7)	3.89	3.91 (rec)	3.55	3.53 (350)
2	4.07	7.75	3.56 (7)	3.90	3.85 (350)	3.54	3.49 (350)
3	3.91	7.56	3.33 (7)	3.65	3.63 (350)	3.37	3.28 (350)
4	3.95	7.21	3.33 (7)	3.64	3.66 (350)	3.42	3.35 (350)
5	3.94	6.99	3.42 (13)	3.75	3.27 (32)	3.50	3.42 (350)
6	3.65	7.35	3.31 (8)	3.76	3.24 (32)	3.31	3.27 (350)
7	3.71	8.81	3.39 (7)	4.99	3.81 (35)	3.45	3.45 (300)
8	4.47	10.62	3.82 (7)	6.67	4.38 (21)	4.09	4.06 (350)
9	5.69	11.74	5.00 (7)	8.01	5.42 (10)	5.30	5.30 (300)
10	5.67	11.47	4.08 (7)	7.29	5.20 (11)	5.08	5.12 (250)
11	5.28	11.23	4.40 (7)	6.40	4.57 (25)	4.72	4.75 (350)
12	5.07	10.80	4.42 (9)	6.17	4.55 (42)	4.53	4.53 (350)
13	4.46	10.12	3.98 (11)	5.39	5.32 (350)	4.02	4.07 (350)
14	4.51	10.16	4.03 (7)	6.25	4.22 (29)	4.21	4.13 (350)
15	5.21	10.71	4.61 (7)	7.13	4.79 (29)	4.90	4.88 (350)
16	4.93	11.16	4.35 (7)	6.95	4.56 (19)	4.72	4.79 (350)
17	5.16	11.88	4.51 (7)	6.75	4.89 (11)	4.85	5.33 (rec)
18	5.09	13.84	4.85 (7)	6.05	4.91 (10)	4.80	4.90 (350)
19	6.00	13.41	5.65 (8)	6.38	5.65 (22)	5.55	5.55 (350)
20	6.44	12.98	5.92 (8)	6.41	6.49 (250)	5.64	5.68 (rec)
21	6.48	11.18	5.83 (7)	6.30	6.16 (350)	5.71	5.66 (rec)
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	3.21 (350)
24	3.16	8.44	2.86 (8)	3.01	2.98 (350)	2.64	2.60 (350)

Notes: MAE for the naïve model and the three 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.

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

Model	Number of trading hours in which the model is ranked:				$\overline{\text{MAE}}$
	1 st	2 nd	3 rd	4 th or more	
SLR	2	1	10	11	4.23
ARMA(1,1)				24	5.33
ARMA(2,2)	2	4	3	15	4.94
ARX(1)	1		1	22	4.48
ARMAX(1,1)	9	8	1	6	4.08
ARMAXd(1,1)	7	10	6	1	4.09
mARX	3	1	3	17	4.28
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.

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

hour	Nord Pool				IPEX			
	$\lambda=\lambda^*$		$\lambda=300$		$\lambda=\lambda^*$		$\lambda=300$	
	Model (λ^*)	MAE	Model	MAE	Model (λ^*)	MAE	Model	MAE
1	ARMAX11 (rec)	1.23	ARMAX11	1.35	ARMAXd11 (rec)	3.34	ARMAXd11	3.46
2	ARMAXd11 (rec)	1.18	ARMAX11	1.33	ARMAXd11 (350)	3.39	ARMAXd11	3.44
3	ARMAXd11 (rec)	1.33	ANN4	1.48	ARMA22 (350)	3.26	ARMA22	3.27
4	ANN4 (rec)	1.45	ANN4	1.57	ARMA22 (350)	3.29	ARMA22	3.28
5	ANN4 (rec)	1.44	ARMAX11	1.67	ARX (32)	3.27	ARMA22	3.29
6	ARMAXd11 (rec)	1.39	ARMAX11	1.58	ARMAX11 (63)	3.14	ARMA22	3.22
7	ARMAXd11 (rec)	1.43	ARMAX11	1.59	ARMAXd11 (350)	3.28	ARMAXd11	3.41
8	SLR (6)	1.81	ARMAX11	1.94	SLR (7)	3.82	ARMAXd11	4.01
9	SLR (6)	2.87	ARMAX11	3.32	ARMAX11 (250)	4.86	ARMAX11	4.85
10	SLR (6)	3.19	ARMAX11	3.87	SLR (7)	4.80	ARMAX11	4.85
11	SLR (6)	2.93	ARMAX11	3.50	ARMAX11 (86)	4.36	ARMAX11	4.54
12	SLR (6)	2.31	ARMAX11	2.71	ARMAX11 (86)	4.14	ARMAX11	4.38
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	ARMAXd11	4.52
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	ARMAX11 (200)	5.61	mARX	5.71
22	ARMAXd11 (rec)	1.20	ARMAX11	1.26	mARX (300)	4.42	mARX	4.42
23	ARMAXd11 (rec)	1.03	ARMAX11	1.06	mARX (350)	3.21	ARMA22	3.23
24	ARMAX11 (rec)	0.97	ARMAX11	1.02	mARX (350)	2.60	mARX	2.64

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.

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

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

Figure 3: 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. SLR = simple linear regression (SLR), ARMA(1,1) = autoregressive moving average, mARX = mARX regression (Maciejowska and Nowotarski, 2016). The dot identifies the best MAE for each model.

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.

Appendix: Additional Tables and Figures

In this appendix, we report some of 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.

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

hours	ARMAX(1,1)		ARMAX(1,1)d		ARMA(1,1)		ARMA(2,2)	
	MAE	λ^*	MAE	λ^*	MAE	λ^*	MAE	λ^*
1	1.233	rec	1.233	rec	1.247	rec	1.379	97
2	1.210	rec	1.175	rec	1.295	rec	1.406	97
3	1.365	rec	1.334	rec	1.426	rec	1.420	rec
4	1.494	rec	1.460	rec	1.538	rec	1.528	rec
5	1.517	rec	1.489	rec	1.604	rec	1.563	rec
6	1.409	rec	1.384	rec	1.725	rec	1.638	rec
7	1.435	rec	1.426	rec	2.249	rec	2.050	rec
8	1.850	23	1.887	rec	2.971	24	2.794	rec
9	3.083	25	3.273	rec	4.788	24	4.284	rec
10	3.510	24	3.703	rec	5.035	22	4.582	rec
11	3.080	27	3.346	rec	4.551	350	4.003	rec
12	2.472	24	2.617	rec	3.459	26	3.106	rec
13	2.153	200	2.225	200	2.807	25	2.598	rec
14	1.976	200	2.060	200	2.687	24	2.433	rec
15	1.905	rec	1.933	rec	2.621	24	2.428	rec
16	1.818	rec	1.846	rec	2.480	24	2.321	rec
17	2.077	rec	2.114	rec	2.805	24	2.539	rec
18	2.794	rec	2.820	rec	3.474	24	3.201	rec
19	3.744	250	3.488	rec	4.012	350	3.870	rec
20	3.224	300	3.041	rec	3.857	25	3.267	rec
21	1.658	rec	1.672	rec	1.884	rec	1.957	350
22	1.255	300	1.201	rec	1.353	rec	1.432	200
23	1.064	300	1.026	rec	1.062	rec	1.079	rec
24	0.966	rec	0.970	rec	1.017	rec	1.060	100

- continues in the next page -

- Table A1 continued -

hours	SLR		ARX		mARX		ANN(4)		ANN(7)	
	MAE	λ^*	MAE	λ^*	MAE	λ^*	MAE	λ^*	MAE	λ^*
1	1.447	6	1.448	33	1.240	rec	1.482	350	1.279	rec
2	1.497	6	1.365	rec	1.744	350	1.276	rec	1.395	350
3	1.680	11	1.512	rec	2.020	350	1.404	350	1.409	rec
4	1.715	6	1.609	rec	1.536	rec	1.451	rec	1.456	rec
5	1.660	6	1.679	rec	1.594	rec	1.444	rec	1.507	rec
6	1.529	6	1.710	rec	1.532	rec	1.435	rec	1.426	rec
7	1.493	6	1.688	19	1.470	rec	1.740	350	1.529	rec
8	1.807	6	1.964	14	2.272	350	2.399	250	2.225	350
9	2.867	6	3.224	14	3.390	100	3.587	150	3.981	350
10	3.188	6	3.674	23	3.831	100	4.201	75	3.826	rec
11	2.934	6	3.325	23	3.401	rec	4.242	350	4.132	350
12	2.307	6	2.631	23	2.701	rec	3.034	100	2.700	rec
13	1.951	6	2.261	23	2.211	rec	2.933	350	2.661	350
14	1.801	6	2.089	14	2.017	rec	2.543	350	2.344	350
15	1.791	6	2.101	14	1.956	rec	2.662	350	2.113	rec
16	1.759	6	2.093	23	1.906	rec	2.632	350	1.978	rec
17	1.932	6	2.280	23	2.161	rec	3.022	350	2.275	rec
18	2.492	6	2.965	23	2.873	rec	3.766	350	4.042	350
19	3.072	7	3.600	23	3.736	rec	4.610	350	4.809	350
20	2.769	7	3.161	25	3.144	rec	3.824	350	3.084	rec
21	1.752	6	1.877	28	1.842	rec	2.072	rec	1.920	rec
22	1.387	6	1.428	31	1.325	rec	1.407	rec	1.396	rec
23	1.226	6	1.195	100	1.107	rec	1.104	rec	1.135	rec
24	1.281	5	1.011	rec	1.071	rec	1.121	350	1.246	300

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

hours	ARMAX(1,1)		ARMAX(1,1)d		ARMA(1,1)		ARMA(2,2)	
	MAE	λ^*	MAE	λ^*	MAE	λ^*	MAE	λ^*
1	3.467	rec	3.338	rec	3.606	62	3.439	rec
2	3.422	350	3.391	350	3.669	350	3.463	350
3	3.298	350	3.260	350	3.464	350	3.256	350
4	3.357	350	3.317	350	3.441	37	3.286	350
5	3.388	33	3.344	350	3.412	44	3.287	350
6	3.138	63	3.187	65	3.291	63	3.221	300
7	3.339	350	3.280	350	4.748	350	4.210	350
8	3.828	50	4.008	250	6.655	150	6.161	300
9	4.859	250	4.860	250	8.359	150	8.018	67
10	4.917	200	4.961	250	6.872	25	7.113	300
11	4.355	86	4.611	250	6.630	350	5.933	350
12	4.139	86	4.401	250	6.146	250	5.371	300
13	3.824	250	3.829	250	5.125	150	4.603	300
14	3.947	81	3.692	79	5.848	150	5.158	300
15	4.468	84	4.366	79	6.936	150	6.198	300
16	4.238	96	4.227	96	7.217	300	6.136	300
17	4.261	30	4.410	97	6.559	250	5.963	350
18	4.593	22	4.505	200	6.180	200	5.612	350
19	5.318	300	5.376	300	6.274	25	6.061	250
20	5.434	200	5.537	300	6.249	200	5.888	300
21	5.610	200	5.746	250	6.219	250	5.785	250
22	4.557	83	4.533	350	4.817	25	4.458	300
23	3.312	300	3.226	350	3.397	300	3.235	350
24	2.748	350	2.743	350	2.812	61	2.682	350

- continues in the next page -

- Table A2 continued -

hours	SLR		ARX		mARX		ANN(4)		ANN(7)	
	MAE	λ^*	MAE	λ^*	MAE	λ^*	MAE	λ^*	MAE	λ^*
1	3.770	7	3.909	999	3.531	350	4.144	999	4.001	999
2	3.556	7	3.854	350	3.488	350	3.864	300	3.779	999
3	3.325	7	3.632	350	3.277	350	3.632	300	3.632	999
4	3.333	7	3.656	350	3.348	350	3.769	350	3.567	350
5	3.415	13	3.273	32	3.416	350	3.673	350	3.712	999
6	3.307	8	3.236	32	3.269	350	3.783	350	3.782	999
7	3.389	7	3.813	35	3.451	300	4.648	350	3.994	350
8	3.820	7	4.378	21	4.06	350	6.125	350	4.522	350
9	5.004	7	5.422	10	5.296	300	7.364	300	5.760	350
10	4.795	7	5.194	11	5.123	250	8.126	999	5.825	999
11	4.403	7	4.568	25	4.749	350	6.984	999	5.318	999
12	4.423	9	4.551	42	4.53	350	5.863	350	5.473	300
13	3.984	11	5.323	350	4.069	350	5.117	350	4.367	999
14	4.027	7	4.217	29	4.129	350	5.390	350	4.450	350
15	4.605	7	4.787	29	4.877	350	6.598	999	5.078	999
16	4.354	7	4.557	19	4.792	350	6.279	350	5.133	350
17	4.509	7	4.891	11	5.325	999	7.375	999	5.683	300
18	4.847	7	4.909	10	4.897	350	6.796	999	5.114	999
19	5.646	8	5.648	22	5.545	350	6.467	350	5.656	999
20	5.915	8	6.487	250	5.68	999	7.119	350	7.334	350
21	5.827	7	6.161	350	5.66	999	6.459	350	7.074	350
22	4.767	7	4.674	21	4.418	300	5.187	250	5.070	300
23	3.523	8	3.458	22	3.21	350	3.562	350	3.659	300
24	2.861	8	2.978	350	2.599	350	2.950	350	2.895	350