On the importance of the long-term seasonal component in day-ahead electricity price forecasting

Jakub Nowotarski, Rafał Weron*

Department of Operations Research, Wrocław University of Technology, Wrocław, Poland

ARTICLE INFO

ABSTRACT

In day-ahead *electricity price forecasting* (EPF) the daily and weekly seasonalities are always taken into account, but the long-term seasonal component (LTSC) is believed to add unnecessary complexity to the already parameter-rich models and is generally ignored. Conducting an extensive empirical study involving state-of-the-art time series models we show that (i) decomposing a series of electricity prices into a LTSC and a stochastic component, (ii) modeling them independently and (iii) combining their forecasts can bring – contrary to a common belief – an accuracy gain compared to an approach in which a given time series model is calibrated to the prices themselves.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Without doubt *electricity price forecasting* (EPF) is of prime importance to the functioning of today's energy business. Alongside load forecasting, *short-term* (also called *spot* or *day-ahead*; for a discussion see Weron, 2014) EPF has become the core process of an energy company's planning activities at the operational level. Although it is very hard to quantify the benefits of improving load and/or price forecasts, Hong (2015) provides interesting back-of-the-envelope calculations. Based on U.S. data from the last decade, he concludes that for a typical medium-size utility with a 5 GW peak load, savings from a 1% reduction in the Mean Absolute Percentage Error (MAPE) are as much as $1.5 million per year from short-term load forecasting and $3 million per year from short-term load and price forecasting! Hong's simplified calculations coincide quite well with more technical studies of Hobbs et al. (1999), who conclude that on average a reduction of 1% in MAPE for short-term load forecasts

decreases variable generation costs by 0.1%–0.3% when MAPE is in the typical range of 3%–5%, and of Zareipour et al. (2010), who find that when MAPE is the 5% to 15% range commonly observed for short-term price forecasts, a 1% improvement in forecast accuracy would result in about 0.1%–0.35% cost reductions. In both studies, the level of actual savings depends to a large extent on generator characteristics.

As has been noted in a number of studies, a key point in electricity spot price modeling and forecasting is the appropriate treatment of seasonality (Janczura et al., 2013; Keles et al., 2016; Lisi and Nan, 2014; Maciejowska, 2014; Nowotarski et al., 2013). For mid-term horizons – ranging from a few days to a few months ahead and typically considered in derivatives pricing and risk management applications – the daily profile is usually regarded as irrelevant (Ignatieva and Trück, 2016; Janczura, 2014). In fact, most mid-term EPF models work with average daily prices and focus on the *annual* or *long-term seasonal component* (LTSC; also called the *trend-seasonal component*). However, in short-term EPF the daily and weekly seasonalities are always taken into account, but the LTSC is believed to add unnecessary complexity to the already parameter-rich models and is generally ignored (for recent reviews see Garcia-Martos and Conejo, 2013; Weron, 2014). But is this the right approach? Should

---

* Corresponding author.
  E-mail addresses: jakub.nowotarski@pwr.edu.pl (J. Nowotarski),
rafal.weron@pwr.edu.pl (R. Weron).

the LTSC be included in day-ahead EPF models, contrary to a common belief that it is redundant in the short-term?

It is exactly the aim of this paper to address these two important questions, that have not been investigated in the EPF literature to date. We perform an extensive empirical study which involves:

- two 2-year long, hourly resolution test periods from two distinct power markets (GEFCom2014 and Nord Pool),
- two autoregressive model structures — one originally proposed by Misiorek et al. (2006) and later used in a number of EPF studies (Gaillard et al., 2016; Kristiansen, 2012; Maciejowska et al., 2016; Nowotarski et al., 2014; Nowotarski and Weron, 2016; Serinaldi, 2011; Weron, 2006; Weron and Misiorek, 2008; Ziel, 2016) and one which evolved from it during the successful participation of TEAM POLAND in the Global Energy Forecasting Competition 2014 (GEFCom2014; see Hong et al., 2016; Maciejowska and Nowotarski, 2016),
- two novel *Seasonal Component AutoRegressive* (SCAR) models that combine a 24 hour-ahead extrapolation of an estimated LTSC with the forecasts of autoregressive models,
- two well-performing LTSC model classes — wavelet smoothing and the Hodrick and Prescott (1997) filter, as advocated by Janczura et al. (2013), Lisi and Nan (2014), Nowotarski et al. (2013) and Weron and Zator (2015),
- model validation in terms of the robust *weekly-weighted mean absolute error* (WMAE; see Weron, 2014) and the Diebold and Mariano (1995) test,

and draw statistically significant conclusions with far reaching consequences for day-ahead EPF.

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

2. Datasets

The datasets used in this empirical study include two day-ahead time series. The first one comes from the Global Energy Forecasting Competition 2014 (GEFCom2014) — the largest energy forecasting competition to date, both in terms of the diversity of competition topics and wide geographic coverage of the participants (for details see Hong et al., 2016). The dataset includes three time series at hourly resolution: locational marginal prices, day-ahead predictions of zonal loads and day-ahead predictions of system loads and covers the period from January 1, 2011 to December 17, 2013. During the competition the information set was being extended on a weekly basis to prevent 'peeking' into the future. However, now it is available in whole as supplementary material (Appendix A) accompanying Hong et al. (2016). In this paper we only use two subseries — locational marginal prices and day-ahead predictions of zonal loads, see Fig. 1. The origin of the data has never been revealed by the organizers.

The second dataset comes from one of the major European power markets — Nord Pool (NP). It comprises hourly system prices and hourly *consumption prognosis* for four Nordic countries (Denmark, Finland, Norway and Sweden) for the period January 1, 2013– December 26, 2015, see Fig. 2. The time series were constructed using data published by the Nordic power exchange Nord Pool (www.nordpoolspot.com) and preprocessed to account for missing values and changes to/from the daylight saving time (like in Weron (2006), Section 4.3.7). The missing data values were substituted by the arithmetic average of the neighboring values. The 'doubled' values (corresponding to the changes from the daylight saving/summer

time) were substituted by the arithmetic average of the two values for the 'doubled' hour.

For both markets, the day-ahead forecasts of the hourly electricity price are determined within a rolling window scheme, using a 360-day calibration window. First, all considered models (their short-term and long-term components) are calibrated to data from the initial calibration period, i.e. January 1 to December 26 (year 2011 for GEFCom2014 and 2013 for Nord Pool) and forecasts for all 24 h of the next day (December 27) are determined. Then the window is rolled forward by one day and forecasts for all 24 h of December 28 are computed. This procedure is repeated until the predictions for the last day in the sample – December 17, 2013 (for GEFCom2014) or December 26, 2015 (for Nord Pool) – are made.

3. Methodology

3.1. The benchmarks

Our choice of the benchmark models is guided by previous literature on electricity price forecasting and experience gained during the successful participation of TEAM POLAND in the GEFCom2014 competition. The modeling is implemented separately across the hours, leading to 24 sets of parameters for each day the forecasting exercise is performed. This approach is inspired by the fact that each hour displays a rather distinct price profile, reflecting the daily variation of demand, costs and operational constraints, and by the extensive research on demand forecasting, which has generally favored the multi-model specification for short-term predictions (see Weron, 2014 for a review).

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

The second model is a parsimonious autoregressive structure originally proposed by Misiorek et al. (2006) and later used in a number of EPF studies (Gaillard et al., 2016; Kristiansen, 2012; Maciejowska et al., 2016; Nowotarski et al., 2014; Weron, 2006; Weron and Misiorek, 2008; Ziel, 2016). Within this model the natural logarithm of the electricity spot price, $p_t = log(P_t)$, is given by the following formula:

$$p_t = \phi_1 p_{t-24} + \phi_2 p_{t-48} + \phi_7 p_{t-168} + \phi_8 mp_t + \psi_1 z_t + \sum_{i=1}^{3} d_i D_i + \varepsilon_t, \quad (1)$$

where the lagged log-prices p_{t-24}, p_{t-48} and p_{t-168} account for the autoregressive effects of the previous days (the same hour yesterday, two days ago and one week ago), while mp_t creates the link between bidding and price signals from the entire previous day (it is the minimum of the previous day's 24 hourly log-prices). The variable z_t refers to the hourly zonal load of a US utility or Nordic consumption (actually to forecasts made a day before, see Section 2). The three dummy variables – D_1, D_2 and D_3 (for Monday, Saturday and Sunday, respectively) – account for the weekly seasonality. Finally, the ε_ts are assumed to be independent and identically distributed (i.i.d.) normal variables. We will denote this autoregressive benchmark by ARX to reflect the fact that the load (or consumption) forecast is used as the eXogenous variable in Eq. (1).

The third benchmark is an extension of the ARX model, which takes into account the experience gained during the GEFCom2014
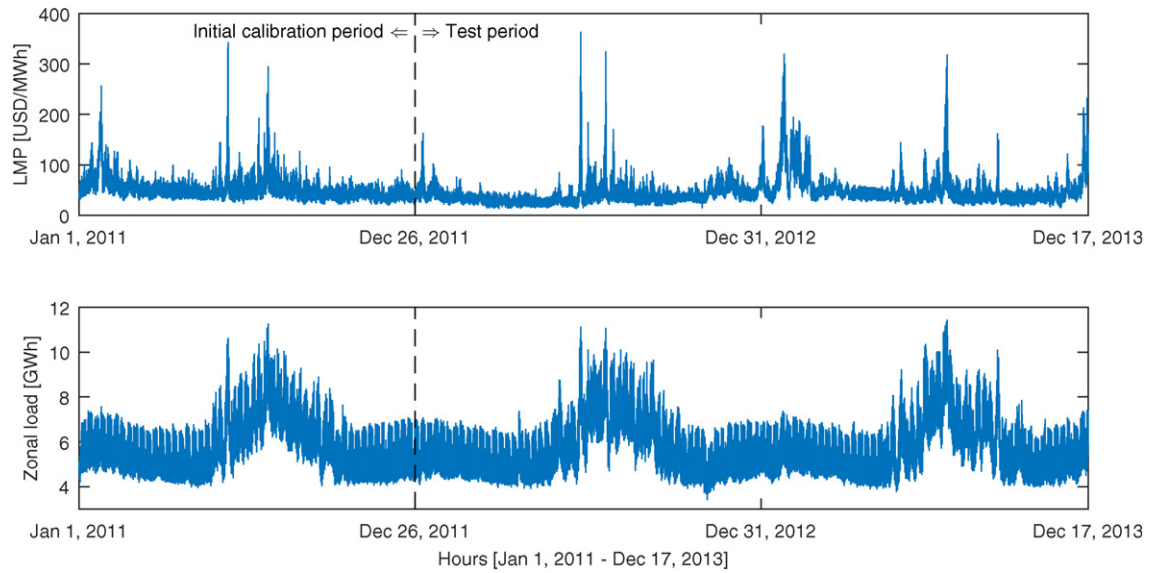
Fig. 1. GEFCom2014 hourly locational marginal prices (LMP; *top*) and hourly day-ahead predictions of zonal load (*bottom*) for the period January 1, 2011–December 17, 2013. The vertical dashed lines mark the end of the initial calibration window. Each day the 360-day long window is rolled forward by 24 h and price forecasts for the 24 h of the next day are computed.

competition (for a detailed discussion see Maciejowska and Nowotarski, 2016). It turns out that it may be beneficial to use different model structures for different days of the week, not only different parameter sets. Hence, the multi-day ARX model (denoted later in the text by mARX) for the natural logarithm of the electricity price is given by the following formula:

$$p_t = \left(\sum_{i=0}^{3} \phi_{1,i} D_i \right) p_{t-24} + \phi_2 p_{t-48} + \phi_3 D_1 p_{t-72} + \phi_7 p_{t-168}$$

$$+ \phi_8 mp_t + \psi_1 z_t + \sum_{i=1}^{3} d_i D_i + \varepsilon_t, \tag{2}$$

where $D_0 \equiv 1$ and the term $D_1 p_{t-72}$ accounts for the autoregressive effect of Friday's prices on the prices for the same hour on Monday.

Finally, note that both autoregressive models – ARX and mARX – are estimated in this study in Matlab using Least Squares.

3.2. Seasonal Component AutoRegressive (SCAR) models

Let us now return to the question whether the long-term seasonal component should be included in day-ahead EPF models or not. To provide an answer, we will investigate whether (i) decomposing a series of electricity log-prices into a LTSC and a stochastic component (or residual), (ii) modeling them independently and (iii) combining their forecasts brings an accuracy gain compared to an approach in which a given autoregressive model structure is calibrated to raw log-prices. Hence, we will compare ARX and mARX models (calibrated to raw log-prices) with the corresponding *Seasonal Component AutoRegressive* (SCAR) models — SCARX and mSCARX
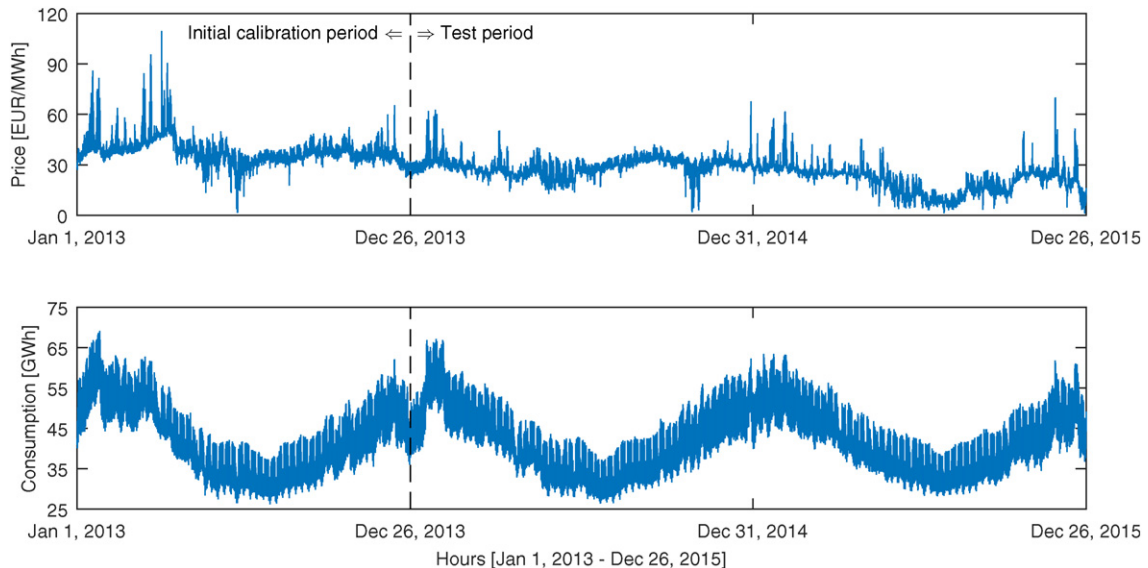

Fig. 2. Nord Pool hourly system prices (*top*) and hourly consumption prognosis (*bottom*) for the period January 1, 2013–December 26, 2015. The vertical dashed lines mark the end of the initial calibration window. Each day the 360-day long window is rolled forward by 24 h and price forecasts for the 24 h of the next day are computed.

(the same autoregressive structures, but calibrated to seasonally decomposed log-prices, then combined with LTSC forecasts).

The standard approach to seasonal decomposition splits the series under investigation into the trend-cycle or long-term seasonal component (LTSC; T_t), the periodic short-term seasonal component (STSC; s_t) and remaining variability (i.e. the stochastic component; X_t), either in an additive or a multiplicative fashion (Hyndman and Athanasopoulos, 2013; Janczura et al., 2013). In the short-term EPF literature it is quite common to consider s_t and X_t jointly. We also do it here. Namely, we decompose the electricity spot log-price series p_t into a sum of two independent parts:

- $q_t = X_t + s_t$, i.e. the stochastic component with short-term periodicities,
- and T_t, i.e. the long-term seasonal component.

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

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

The Hodrick and Prescott (1997) filter was originally proposed in macroeconomics for decomposing the series of GDP values into a long-term growth component and the business cycle. The mechanics of the HP filter are, however, universal. When applied to electricity spot prices it splits the series into a smooth part – the LTSC, and a volatile part – the stochastic component with short-term periodici-ties. Weron and Zator (2015) advocated the use of the HP filter as a simpler, yet equally flexible alternative to wavelet smoothing. Also Lisi and Nan (2014) found it to be a very well performing smoother. For a noisy (or volatile) input series of electricity spot log-prices p_t, the HP filter returns a smoothed series T_t which minimizes:

$$\min_{T_t} \left\{ \sum_{t=1}^{\tau} (p_t - T_t)^2 + \lambda \sum_{t=2}^{\tau-1} [(T_{t+1} - T_t) - (T_t - T_{t-1})]^2 \right\}, \quad (3)$$

where τ is the number of observations (in this study: $360 \times 24 = 8640$ h of the calibration window) and λ is a smoothing parameter. To find the optimal value of λ we use a similar grid as in Weron and Zator (2015), the only difference is that in this study the price series are in hourly (not daily) resolution and the values of λ have to be larger. We use eight different λs: 1×10^8, 5×10^8, 1×10^9, 5×10^9, 1×10^{10}, 5×10^{10}, 1×10^{11} and 5×10^{11}.

A potential problem with the SCAR approach is that the LTSC has to be predicted (which is not a trivial task, see Nowotarski et al., 2013) before combining it with the day-ahead forecasts of the

autoregressive model for q_t. What makes the problem even more challenging is that the predictions of the trend-seasonal component cannot be properly evaluated, since the LTSC itself is an outcome of model assumptions and is not observed in reality. However, in the short-term we may assume that a simple *persistent* (or *random walk*) forecast is good enough for our purposes. Note that Lisi and Nan (2014) used a similar approach but for daily, not hourly data.

Summarizing, the SCAR modeling framework consists of the following four steps:

1. Decomposing the series of electricity log-prices p_t from the calibration window into a trend-seasonal component T_t and a stochastic component with short-term periodicities q_t, using one of the ten wavelet smoothers ($S_5, \ldots, S_{14}$) or one of the eight HP filters ($\lambda = 1 \times 10^8, \ldots, 5 \times 10^{11}$). Then computing persistent forecasts of the LTSC independently for each of the 24 h of the next day, i.e. $\hat{T}_{t^*+1} \equiv T_{t^*-23}, \ldots, \hat{T}_{t^*+24} \equiv T_{t^*}$, where t^* is the time index of the last observation in the calibration window.
2. Calibrating autoregressive models defined by Eqs. (1) and (3.1) to q_t and computing forecasts for the 24 h of the next day, i.e. $\hat{q}_{t^*+1}, \ldots, \hat{q}_{t^*+24}$. Note, that unlike the seasonal decomposition in Step 1, which is made for the whole $360 \times 24 = 8640$ h long calibration sample, here we split the data into 24 hourly series (like for the ARX and mARX benchmarks).
3. Adding forecasts of the autoregressive models computed in Step 2 to the persistent forecasts of the LTSC to yield log-price forecasts: $\hat{p}_{t^*+1}, \ldots, \hat{p}_{t^*+24}$.
4. Taking the exponent of the log-price forecasts computed in Step 3 to convert them into price forecasts of the SCARX and mSCARX models: $\hat{P}_t = \exp(\hat{p}_t)$.

4. Empirical results

We now present day-ahead forecasting results for the two considered datasets: GEFCom2014 hourly locational marginal prices and Nord Pool hourly system prices. We use long, 2-year out-of-sample test periods to make sure the obtained results are reliable. Recall from Section 2, that models are re-estimated on a daily basis. Price forecasts $\hat{p}_{t^*+1}, \ldots, \hat{p}_{t^*+24}$ for all 24 h of the next day are deter-mined at the same point in time and the 360-day calibration window is rolled forward by one day: $t^* \to t^* + 24$ h.

4.1. Performance evaluation in terms of WMAE

Following Conejo et al. (2005), Weron and Misiorek (2008) and Nowotarski et al. (2014), we compare the models in terms of the Weekly-weighted Mean Absolute Error (WMAE) loss function. WMAE is a robust measure similar to MAPE but with the absolute error normalized by the mean weekly price to avoid the adverse effect of negative and close to zero electricity spot prices. We evalu-ate the forecast performance using weekly time intervals, each with $24 \times 7 = 168$ hourly observations. Note that we also analyzed the forecasts using squared error losses, however, results were qualita-tively similar and are omitted here due to space limitations. For each week we calculate the WMAE for model i as:

$$\text{WMAE}_i = \frac{1}{\bar{P}_{168}} \text{MAE}_i = \frac{1}{168 \cdot \bar{P}_{168}} \sum_{h=1}^{168} \left| P_h - \hat{P}_h^i \right|, \quad (4)$$

where P_h is the actual price for hour h (not the log-price p_h), $\hat{P}_h^i$ is the predicted price for that hour obtained from model i and $\bar{P}_{168} = \frac{1}{168} \sum_{h=1}^{168} P_h$ is the mean price for a given week. Note, that WMAE requires the test period to be a multiple of a week

Table 1

Average WMAE for all 103 weeks of the GEFCom2014 out-of-sample test period (*upper half*) or all 104 weeks of the Nord Pool out-of-sample test period (*lower half*). WMAE errors for the SCARX and mSCARX models smaller (better) than those of the ARX and mARX benchmarks, respectively, are indicated in bold. Underlined are the results for the best performing model in each of the four parts of the table. Note, that the Naïve benchmark is much worse than any of the other models. Note also, that the mARX and mSCARX models are almost always better than the corresponding ARX and SCARX models.

GEFCom2014

Naïve	ARX	SCARX									
		Wavelet approximation									
		S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}
		13.530	13.686	12.466	11.558	11.378	11.264	11.263	**_11.112_**	**11.221**	11.245
20.475	11.232	*HP filter* λ									
		1×10^8	5×10^8	1×10^9	5×10^9	1×10^{10}	5×10^{10}	1×10^{11}	5×10^{11}		
		11.775	11.586	11.527	11.425	11.396	11.376	11.362	11.287		

	mARX	mSCARX									
		Wavelet approximation									
		S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}
		13.482	13.647	12.233	11.379	**11.216**	**11.213**	11.312	**_10.901_**	**10.976**	**11.130**
	11.252	*HP filter* λ									
		1×10^8	5×10^8	1×10^9	5×10^9	1×10^{10}	5×10^{10}	1×10^{11}	5×10^{11}		
		11.580	11.414	11.369	11.347	11.381	11.548	11.612	11.598		

Nord Pool

Naïve	ARX	SCARX									
		Wavelet approximation									
		S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}
		9.949	9.988	8.598	**8.389**	**_8.309_**	**8.332**	**8.417**	**8.453**	**8.463**	**8.475**
12.663	8.500	*HP filter* λ									
		1×10^8	5×10^8	1×10^9	5×10^9	1×10^{10}	5×10^{10}	1×10^{11}	5×10^{11}		
		8.665	8.697	8.718	8.760	8.766	8.766	8.757	8.729		

	mARX	mSCARX									
		Wavelet approximation									
		S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}
		9.954	10.049	8.558	8.286	**8.157**	**_8.154_**	8.331	8.471	8.428	8.361
	8.341	*HP filter* λ									
		1×10^8	5×10^8	1×10^9	5×10^9	1×10^{10}	5×10^{10}	1×10^{11}	5×10^{11}		
		8.516	8.504	8.513	8.526	8.530	8.561	8.578	8.644		

(or 168 h). Hence, when computing WMAE we consider the first 103 full weeks (December 27, 2011–December 16, 2013) for the GEFCom2014 dataset and the first 104 full weeks (December 27, 2013–December 24, 2015) for the Nord Pool dataset.

In Table 1 we report the average WMAE in the forecasting period for five different model classes: three benchmarks – Naïve, ARX and mARX, and two SCAR-type models that combine a 24 hour-ahead extrapolation of an estimated LTSC with the forecasts of autoregressive models – SCARX and mSCARX. Each of the SCAR-type models is computed for ten wavelet-based and eight HP filter-based LTSC. We use suffixes -S_J and -HP$_\lambda$ to denote SCAR-type models with an S_J wavelet-based LTSC and a HP filter-based LTSC with smoothing parameter λ, respectively.

Several important conclusions can be drawn. Firstly, all autoregressive models beat the Naïve benchmark by a large margin. This indicates that they all are highly efficient forecasting tools. Secondly, a SCAR-type model with a properly selected wavelet-based seasonal component can improve the accuracy of day-ahead forecasts compared to the corresponding autoregressive benchmark. Thirdly, somewhat surprisingly, this is not the case for the HP filter-based LTSC. In fact, no matter what is the value of the smoothing parameter λ, SCAR-type models with HP filter-based LTSC never outperform the benchmark autoregressive models. Finally, the multi-day autoregressive models – mARX and mSCARX – are almost always better than the corresponding ARX and SCARX models.

For the GEFCom2014 dataset, the best performing wavelet approximation is S_{12}, which roughly corresponds to half-year ($2^{12} =$

4096 h or 170 days) smoothing, the second best is S_{13}, equivalent to annual smoothing. Generally, all approximations between S_9 and S_{14} work pretty well, however, not all can beat the corresponding benchmark autoregressive models, especially when non-multi-day specifications are considered. For the Nord Pool dataset, the best performing wavelet approximations are S_9 and S_{10}, which roughly correspond to 3- and 6-week smoothing. Like for the GEFCom2014 dataset, more approximations work pretty well (this time from S_8 to S_{14}) but not all can beat the corresponding benchmark autoregressive models, especially when multi-day specifications are considered. Generally, lower wavelet decomposition levels do not perform well, probably because the S_5, S_6 and S_7 smoothers are too sensitive to fluctuations in spot prices at the far end of the calibration window.

Quite likely this variability at the edges of the sample is also the reason for the disappointing performance of the HP filter-based LTSC. Indeed, this can be seen to some extent in Fig. 2 of Weron and Zator (2015), where a wavelet-based and a HP filter-based LTSC are fitted to a simulated trajectory of the PJM spot prices. While in the middle of the sample they roughly match or the wavelet-based LTSC fluctuates even more, at the edges the HP filter-based LTSC is more volatile. To validate this conjecture we have computed standard deviations, denoted by σ(LTSC), of the differences between the fitted LTSC and the actual prices across all 24 h of the last day in the calibration window. Recall that the last 24 hourly values of the LTSC are used as the day-ahead persistent forecasts of the LTSC in our SCAR-type models (see Section 3.2). It turns out that for wavelet
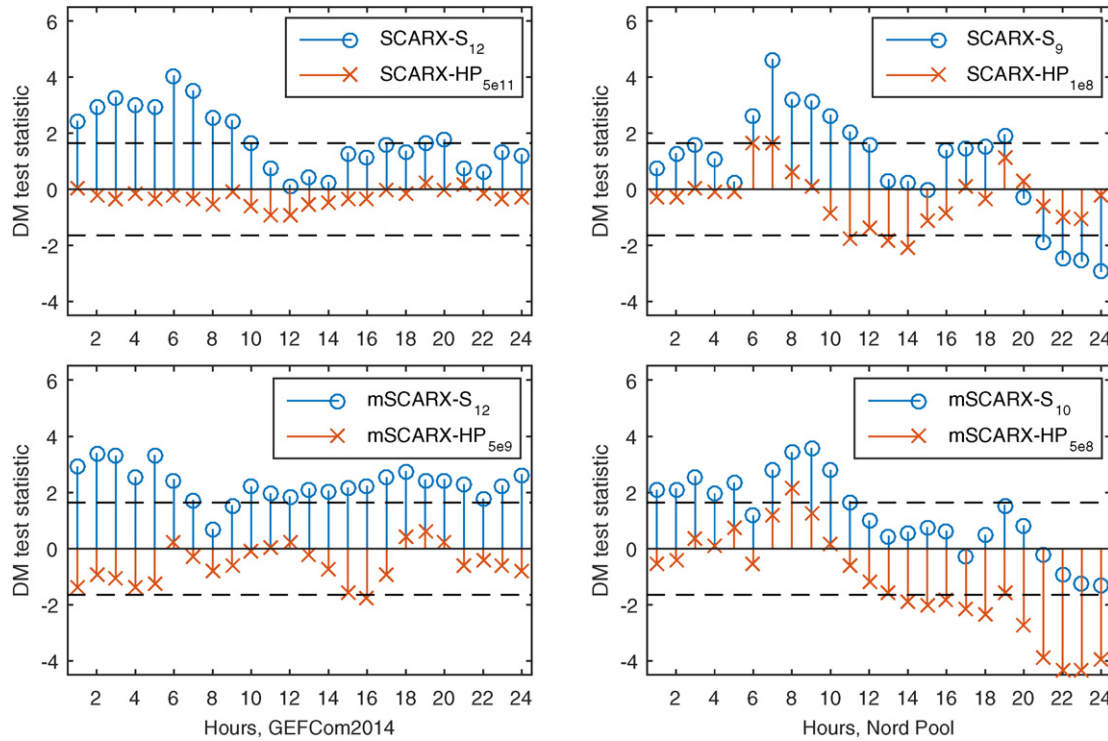
Fig. 3. Results for conducted one-sided Diebold–Mariano tests for both datasets: GEFCom2014 (*left panels*) and Nord Pool (*right panels*). The SCARX models are tested against the ARX benchmark (*upper panels*) and the mSCARX models against the mARX benchmark (*lower panels*). The tests were conducted separately for each of the 24 h. Only points lying above (i.e. better than the benchmark) or below (i.e. worse than the benchmark) the dashed lines are significant at the 5% level. See also Table 2.

and HP-filter based LTSC of similar variability (e.g. S_{10} and $HP_{5\times10^{10}}$, S_{12} and $HP_{5\times10^{11}}$) the former always yield lower values of $\sigma(LTSC)$. For instance, for the GEFCom2014 dataset, $\sigma(S_{10})$ is on average 25% lower than $\sigma(HP_{5\times10^{10}})$ and $\sigma(S_{12})$ is on average 27% lower than $\sigma(HP_{5\times10^{11}})$.

4.2. Diebold–Mariano (DM) tests

In order to formally investigate the advantages from using SCAR-type models over the autoregressive benchmark models, we apply the Diebold and Mariano (1995) test. Since predictions for all 24 h of the next day are made at the same time using the same information set, forecast errors for a particular day will typically exhibit high serial correlation. Therefore, like Bordignon et al. (2013) and Nowotarski et al. (2014), we conduct the DM tests for each of the 24 load periods separately, using absolute error losses of the model

forecast: $L(\varepsilon_t) = |\varepsilon_t| = |P_t - \hat{P}_t|$. For each best performing SCAR-type model in its class:

- SCARX-S_{12}, -$HP_{5\times10^{11}}$, mSCARX-S_{12}, -$HP_{5\times10^{9}}$ for GEFCom2014,
- SCARX-S_9, -$HP_{1\times10^{8}}$, mSCARX-S_{10}, -$HP_{5\times10^{8}}$ for Nord Pool,

and for each hour independently, we calculate the loss differential series:

$$d_t = L(\varepsilon_t^{model}) - L(\varepsilon_t^{benchmark}). \tag{5}$$

We then conduct the DM tests for significant differences with respect to the performance of the corresponding benchmark model (ARX or mARX). We perform two one-sided DM tests at the 5% significance level: (i) a standard test with the null hypothesis H_0 : $E(d_t) \leq 0$, i.e. the outperformance of the benchmark by a given SCAR-type model, and (ii) the complementary test with the reverse null

Table 2
The percentage and the number (in parentheses) of hours for which the best wavelet-based and the best HP filter-based SCAR-type model is able to outperform the corresponding autoregressive benchmark model (*#better*) and for which the benchmark model is able to outperform SCAR-type model (*#worse*), at the 5% significance level for the conducted Diebold–Mariano test. Compare with Fig. 3.

	GEFCom2014		Nord Pool		Both datasets	
	SCARX-S_{12}	SCARX-$HP_{5\times10^{11}}$	SCARX-S_9	SCARX-$HP_{1\times10^{8}}$	Wavelets	HP filter
#better	46% (11)	0% (0)	29% (7)	0% (0)	38% (18)	0% (0)
#worse	0% (0)	0% (0)	17% (4)	13% (3)	8% (4)	6% (3)
	mSCARX-S_{12}	mSCARX-$HP_{5\times10^{9}}$	mSCARX-S_{10}	mSCARX-$HP_{5\times10^{8}}$	Wavelets	HP filter
#better	92% (22)	0% (0)	42% (10)	4% (1)	67% (32)	2% (1)
#worse	0% (0)	4% (1)	0% (0)	40% (10)	0% (0)	23% (11)

$H_0 : E(d_t) \geq 0$, i.e. the outperformance of a given SCAR-type model by the benchmark. Fig. 3 provides a graphical representation of the DM test statistic for each hour and model for both considered markets. The results are also summarized in Table 2.

The obtained DM-test results support our observations from Section 4.1 on WMAE errors. Again, we can conclude that SCAR-type models with properly selected wavelet-based LTSC improve forecasting accuracy, especially when the generally more accurate multi-day models are considered. For 67% of the hours the best mSCARX-S models significantly outperform the mARX benchmark; at the same time they are never significantly outperformed by the benchmark. For the GEFCom2014 dataset all test statistics are positive and for 22 h they are significant at the 5% level. For the Nord Pool dataset the results are not that optimistic, yet still only for 5 h the test statistics are negative and all are insignificant at the 5% level.

For the on average less accurate specifications based on Eq. (1), the improvement in predictive performance is not that clear-cut. For 38% of the hours the best SCARX-S models significantly outperform the ARX benchmark. However, for the Nord Pool dataset, SCARX-S$_9$ is significantly outperformed by the benchmark for 4 h. Overall, the SCAR-type models are still generally better than the ARX benchmark.

Finally, SCAR-type models with the HP filter-based LTSC again do not perform well. Regardless of the selection of λ, this approach is not able to significantly outperform the benchmark models (except for 1 h for the Nord Pool dataset and the multi-day specifications). On the other hand, these models are significantly worse than the ARX benchmark for 6% of the hours and the mARX benchmark for as many as 23% of the hours.

5. Conclusions

In day-ahead electricity price forecasting (EPF) the daily and weekly seasonalities are always taken into account, but the long-term seasonal component (LTSC) is believed to add unnecessary complexity to the already parameter-rich models and is generally ignored (Weron, 2014). But is this the right approach? Should the LTSC be included in day-ahead EPF models, contrary to a common belief that it is redundant in the short-term?

Conducting an extensive empirical study involving state-of-the-art time series models and two 2-year long, hourly resolution test periods from two distinct power markets (GEFCom2014 and Nord Pool), we have addressed these important questions. To this end, we have introduced a new class of models – *Seasonal Component AutoRegressive* (SCAR) models – that combine a 24 hour-ahead extrapolation of an estimated LTSC with the forecasts of autoregressive models. We have compared the predictive performance of these models to that of three benchmarks in terms of the robust *weekly-weighted mean absolute error* (WMAE) and the Diebold and Mariano (1995) test.

We have shown that, all considered autoregressive models – both SCAR-type and the ARX-type benchmarks – beat the naïve benchmark of Nogales et al. (2002) by a large margin, which indicates that they all are highly efficient forecasting tools. Furthermore, we have provided evidence that a SCAR-type model with a properly selected wavelet-based seasonal component can significantly (in terms of the DM test) improve the accuracy of day-ahead forecasts compared to the ARX-type benchmarks. We expect that this result will have far reaching consequences for day-ahead EPF in the near future. Somewhat surprisingly, we have also found that it does make a huge difference what model is used for the LTSC. Despite evidence provided by Weron and Zator (2015) that the Hodrick and Prescott (1997) filter is a very flexible alternative to wavelet smoothing, the predictive performance of SCAR-type models with HP filter-based LTSC was disappointing and generally inferior to the ARX-type benchmarks.

Finally, an added value of our SCAR methodology is that it creates a natural opportunity for considering forecast combinations, an approach which has been shown to provide additional accuracy gains in day-ahead forecasting of electricity spot prices (Aneiros et al., 2016; Bordignon et al., 2013; Keles et al., 2016; Nowotarski et al., 2014; Nowotarski and Weron, 2016). Although the optimal forecast combination design assumes aggregation of independent predictions, it may not be feasible in practice. Individual forecasts need to be produced by different experts and hence the costs of implementing such practices are often unaffordable to utilities. By considering SCAR models with different LTSC a forecaster should be able to generate a pool of accurate, yet to a large extent independent forecasts that may be combined to yield even better predictions. This idea is similar is spirit to combining so-called *sister forecasts* in load forecasting (Liu et al., 2016; Nowotarski et al., 2016), i.e. predictions generated from a single model structure calibrated to different datasets (two vs. three year calibration windows, individual vs. all 24 hourly loads, etc.; see Wang et al., 2016). The validation of this approach is left, however, for future research.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.eneco.2016.05.009.

References

Aneiros, G., Vilar, J., Raña, P., 2016. Short-term forecast of daily curves of electricity demand and price. Electr. Power Energy Syst. 80, 96–108.

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

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

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

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

Garcia-Martos, C., Conejo, A.J., 2013. Price forecasting techniques in power systems. Wiley Encyclopedia of Electrical and Electronics Engineering. Wiley. pp. 1–23

Hobbs, Benjamin.F., Jitprapaikulsarn, Suradet, Maratukulam, Dominic.J., 1999. Analysis of the value for unit commitment of improved load forecasts. IEEE Trans. Power Syst. 14 (4), 1342–1348.

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

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

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., Hyndman, R.J., 2016. Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. Int. J. Forecast. 32 (3), 896–913. July-September.

Hyndman, R., Athanasopoulos, G., 2013. Forecasting: Principles and Practice. http://otexts.org/fpp/.

Ignatieva, K., Trück, S., 2016. Modeling spot price dependence in Australian electricity markets with applications to risk management. Comput. Oper. Res. 66, 415–433.

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

Janczura, J., 2014. Pricing electricity derivatives within a Markov regime-switching model: a risk premium approach. Math. Meth. Oper. Res. 79 (1), 1–30.

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

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

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

Liu, B., Nowotarski, J., Hong, T., Weron, R., 2016. Probabilistic load forecasting via Quantile Regression Averaging on sister forecasts. IEEE Trans. Smart Grid http://dx.doi.org/10.1109/TSG.2015.2437877.

Maciejowska, K., 2014. Fundamental and speculative shocks, what drives electricity prices?. IEEE Conference Proceedings - EEM'14 http://dx.doi.org/10.1109/EEM.2014.6861289.

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

Maciejowska, K., Nowotarski, J., Weron, R., 2016. Probabilistic forecasting of electricity spot prices using factor quantile regression averaging. Int. J. Forecast. 32 (3), 957–965. July-September.

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

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

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

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

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

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

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

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

Wang, P., Liu, B., Hong, T., 2016. Electric load forecasting with recency effect: A big data approach. Int. J. Forecast. 32 (3), 585–597.

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

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

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

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

Zareipour, H., Canizares, C.A., Bhattacharya, K., 2010. Economic impact of electricity market price forecasting errors: A demand-side analysis. IEEE Trans. Power Syst. 25 (1), 254–262.

Ziel, F., 2016. Forecasting electricity spot prices using LASSO: On capturing the autoregressive intraday structure. IEEE Trans. Power Syst. http://dx.doi.org/10.1109/TPWRS.2016.2521545.