Energy Systems manuscript No.
(will be inserted by the editor)

Benchmarking time series based forecasting models for electricity balancing market prices

Gro Klæboe · Anders Lund Eriksrud · Stein-Erik Fleten

Received: date / Accepted: date

Abstract In the trade-off between bidding in the day-ahead electricity market and the real time balancing market, producers need good forecasts for balancing market prices to make informed decisions. A range of earlier published models for forecasting of balancing market prices, including a few extensions, is benchmarked. The models are benchmarked both for one hour-ahead and day-ahead forecast, and both point and interval forecasts are compared. None of the benchmarked models produce informative day-ahead point forecasts, suggesting that information available before the closing of the day-ahead market is efficiently reflected in the day-ahead market price rather than the balancing market price. Evaluation of the interval forecasts reveals that models without balancing state information overestimate variance, making them unsuitable for scenario generation.

Keywords Balancing market price · Forecasting · Electricity markets · Ancillary service market · Regulating market · Electricity price scenarios

1 Introduction

According to the European Wind Energy Association [8], solar PV and wind power accounted for the two largest share of new installed capacity in the EU in 2012. The increased penetration of intermittent renewable sources in the power system will increase the need for and cost of balancing reserves in the

G. Klæboe
Department of Electric Power Engineering,
Norwegian University of Science and Technology, 7491 Trondheim, Norway
E-mail: gro.klaeboe@elkraft.ntnu.no

A.L. Eriksrud · S-E Fleten
Department of Industrial Economics and Technology Management,
Norwegian University of Science and Technology, 7491 Trondheim, Norway
E-mail: ae2459@columbia.edu, E-mail: stein-erik.fleten@iot.ntnu.no

power system [11]. For flexible generators, higher prices and volumes in the balancing market offer new opportunities for profit. This raises the issue of how the producers should allocate their capacity between the day-ahead market and the balancing market. The question is analyzed by Glachant and Saguan [10] who study equilibrium relationships between the day-ahead and the balancing market, and by Boomsma et al [1] who constructed a stochastic programming model, with scenarios for day-ahead and balancing market prices.

In order to formulate good bidding strategies that take all the subsequent physical markets into account, the market participants need good price forecast so that the trade-off between sales in the day-ahead market and sales in the shorter-term markets can be properly evaluated. Fleten and Pettersen [9] and Boomsma et al [1] formulate such bidding models from the retailer's and producer's perspective, respectively. A challenge in this respect is to build price forecasts for the balancing market that deliver useful information before the closure of the day-ahead market, i.e. 12–36 hours ahead. The balancing market is designed to take care of unforseen events or variations in the power system, and if such events are known before the closing of the day-ahead market, they are no longer unforseen and therefore reflected in the day-ahead market price rather than in the balancing market price. The major question is if it is possible at all to create informative forecasts for such a market?

Whereas the papers on modelling day-ahead electricity markets are numerous, the modelling of balancing market prices has received less attention. Weron and Misiorek [21] offer a good survey of day-ahead forecast methods and benchmark a range of the time series methods against each other. Other articles that benchmarks day-ahead forecasting methods include the work of Conejo et al [4], who compare three different time series models, neural network and wavelet models for the PJM Interconnection day-ahead price, and that of Nogales et al [16], who compare two different time series models for the Spanish and the Californian day-ahead market prices. When it comes to balancing market prices, no survey exists to our knowledge. However, case studies exist, including Skytte [19], Fleten and Pettersen [9],Olsson and Söder [17], Jaehnert et al [12], Brolin and Söder [2] and Boomsma et al [1] - all of them from the Nordic market. The contribution of this article is a systematic review and benchmarking of time series based methods for balancing market price forecasting. Emphasis is laid on the day-ahead horizon, but both one hour ahead and one day ahead forecasts are benchmarked.

Since European balancing markets are operated by the national transmission system operators (TSOs), the balancing markets have more country specific rules than the day-ahead wholesale electricity exchanges. This study will focus on only one balancing market - the Nord Pool price zone NO2 in Norway. Price models for this area have previously been built by Skytte [19] and Jaehnert et al [12].

2 The role of the balancing market

In most deregulated markets, electricity is traded numerous times before actual production and consumption takes place. There exist non-physical forward- and future markets used by producers and retailers for hedging. The day-ahead market (also referred to as the spot market) is normally the largest market for physical trade, where producers and retailers commit to a physical injection or withdrawal in the grid the following day. However, as Glachant and Saguan [10] correctly point out, the day-ahead market is in fact also a forward market, since the price the producers or retailers in the last instance is exposed to, are the real-time balancing market price.

The electricity market is very special in the sense that demand and supply must be in equilibrium in at every moment. Unlike other markets, failure of supply or unusual high demand will not only result in a share of customers not being served, but may, if not well perceived and proper measures taken, lead to deterioration of the whole power system, and in the worst case, to a black out. With a black out, no customer will be served, and it may have large adverse effects on production and transmission equipment. To ensure safe operation of the electricity grid, a third party is given monopoly on trading in the last minutes (usually 60 or more) before real time operation with the mandate of ensuring instantaneous balance between demand and supply. In Europe, this third party is usually a national body named Transmission System Operator (TSO), whereas in the various US markets the role is given to an Independent System Operator (ISO).

The power system has three levels of protection, on various timescales. ENTSO-E [6] defines these as frequency containment, frequency restoration and replacement reserves, commonly referred to as primary, secondary and tertiary reserves, respectively. Due to the timescale, only replacement reserves can be traded ex post an event. Thus, the balancing market in this context is understood as replacement reserve trade between operator (SO) and the producers.

2.1 Properties of the balancing market

Glachant and Saguan [10] point out that the pricing policy in the balancing market can follow one of two main philosophies: Either the balancing market is seen as a real time market, within a special institutional framework, or the balancing market is perceived as a measure of last resort for producers or retailers who fail to fulfill their commitments, with price mechanisms built in to discourage trading. Depending on technology mix and the general regulatory environment, the balancing market implementations usually lie somewhere in between these two extremes. Based on the work of Rivero et al [18] and van der Veen et al [20], we find six properties useful to describe a specific implementation of a balancing market. These are:

- Settlement granularity: What is the program time unit of the balancing market? Is it equal to or different from the program time unit of the day-ahead market.
- Remuneration: Is the producer paid for reservation of capacity or energy delivered or both?
- Pricing scheme: Is the balancing market price set by the marginal cost of the last activated unit, or by a cost covering scheme?
- Activation: Are replace reserves activated manually or automatically?
- Single vs dual pricing: Is the balancing market price paid to all producers who deviate from their day-ahead market scheduling in a beneficial direction, or only to those who have actively entered the balancing market as suppliers?
- Price caps/floors: Is the balancing market price floored by the spot market price for upward regulation, and capped by it for downward regulation, or can the balancing market price take any value?

In the Nordic countries, balancing market prices are hourly; they have the same program time unit as the day-ahead market. This is in contrast to the German system, where the balancing market is settled and prices defined for every 15 minutes [15]. Remuneration and pricing schemes vary quite a lot throughout Europe; the reader is referred to [18] for details. In the Norwegian market, producers are remunerated for the utilization of balancing power, and not for the reservation of capacity. [1] Norwegian producers are paid for their balancing power based on marginal price for the most costly activated bid that hour - in contrast to, for instance, France, Germany and Italy where the producer is paid-as-bid [18].When it comes to activation, the Nordic system is somewhat special in the sense that tertiary reserves are activated manually, implying a certain inertia from imbalance occurs until the activation of reserves takes place. Also, the Nordic system has a dual imbalance pricing scheme, which means that producers have to state their intention of participating in the balancing market to receive the balancing market price. Producers who just happen to be in imbalance in the opposite direction of the system (and thus passively helping the system), will not benefit from balancing market prices, like they would in a single pricing regime such as Spain and Greece [20]. Furthermore, in the Norwegian balancing market, there are price caps and floors stating that balancing prices can never be lower than the day-ahead prices in case of upward regulation, and never higher than the day-ahead prices in the case of downward regulation.

[1] In Norway, there exists an option market for balancing power, RKOM, where producers are paid for reservation of capacity in addition to normal payment for balancing power. Since the turnover in this market is rather marginal, the discussion is omitted for clarity reasons.

Table 1 Characteristics of the Norwegian balancing market

Settlement granularity:	Hourly
Remuneration:	Energy only
Pricing scheme:	Marginal pricing
Activation:	Manual
Singe/Dual pricing:	Dual
Price cap/floor:	Day-ahead price

2.2 The causes of imbalance and demand for balancing power

ENTSO-E [6] identifies three sources of imbalance under normal operation: i)loss of major production, consumption or transmission unit, or ii)stochastic fluctuation of consumption and production, or iii) weaknesses in market design - for instance the failure of hourly program time units to match the continuous changing consumption. In addition to these causes, there is also the issue of whether the market players influence the demand for balancing services through economic incentives in the balancing market. van der Veen et al [20] performed a simulation of the balancing market with random events and found that the optimal balancing strategy (keep imbalances small, opt for surplus rather than shortage) for players is fairly equal regardless of imbalance payment regime. Möller [14] investigated the German balancing markets and concluded that the producers anticipate the imbalances due to day-ahead market design, and act to take advantage of it by producing more when there is an expected need of upward balancing, and less when there is an expected need of downward balancing [2].

If the demand for balancing services and the supply of these services are truly random processes, balancing prices will also be a random process. However, although failure events are hard to predict, it is interesting to see whether the the patterns in consumption and production fluctuation and the anticipation by the market players can be described in any time series model that give more accurate prediction than the forecasts of a purely random process.

3 Data

Balancing prices, day-ahead prices, balancing states and balancing volumes, as well as overall production volumes, are collected for the NO2 price area for the period 19.07.2010-23.12.2012. The selection of estimation period was motivated by the availability of data. The start date marks the day when forecasts of estimated production and consumption for the next day was made available. Except for two lesser adjustments, the price area borders have been stable during the whole period. NO2 covers the southern and western part of

[2] These patterns are probably quite pronounced in the German market, since it has a single-price regime and also settlement time units of 15 minutes in the balancing market.

Norway, with connections to price areas NO1, NO5 and DK1. Earlier, NO2 was a part of the NO1 area analyzed by [19] and [12].

For out-of-sample verification, the balancing market prices and volumes for NO2 in the period 02.01 – 22.03.2013 were selected. These are displayed in Fig. 1 and 2. This period represent the non-holidays of the first quarter of 2013. All the data was downloaded from the Nord Pool ftp server.

In most models for balancing price forecasting, we work with the balancing premium, rather than the balancing price directly. The balancing premium, δ , is defined as:

$$\delta = \rho^{BM} - \rho^{spot} \tag{1}$$

Where possible, we have stayed faithful to the original model formulations which we aim to benchmark. Therefore the data was log transformed and mean differenced in some of the models, but not in others. The only deviation from the original formulation is the EXO model, inspired by Jaehnert et al [12]. We opted to log-transform the prices and exogenous inputs in order to compress the variance and obtain better fit.

As the focus of this article is on the balancing market, we have not attempted to forecast price or turnover in the day-ahead market where this is needed as input for the balancing market forecasts. In the models where day-ahead market prices or volumes were used as input, we have simply used observed data.

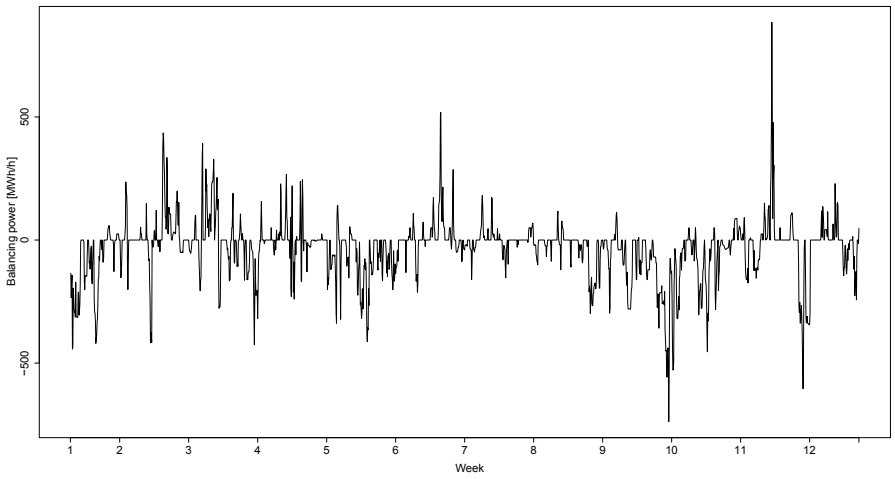

Fig. 1 Realized balancing market volumes in the out-of-sample period. 02.01.2013-22.03.2013

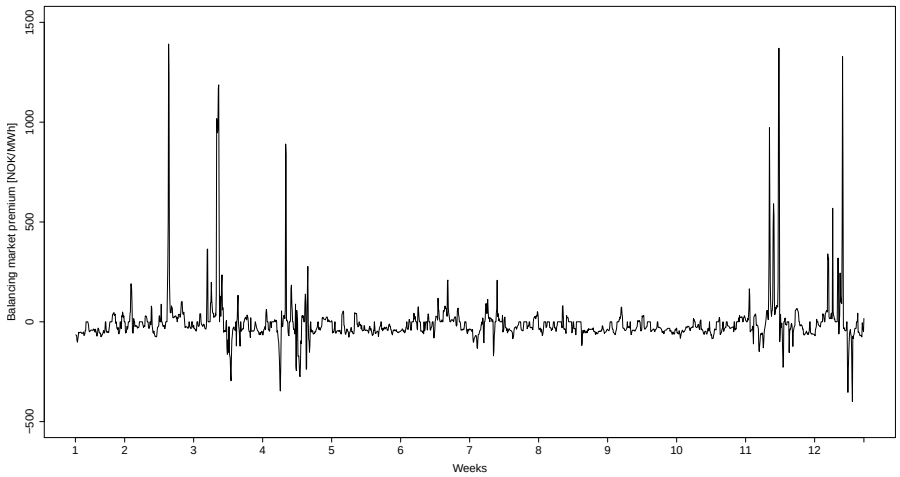

Fig. 2 Observed balancing market premium. [NOK/MWh]

4 Model families

There are generally two families of models for forecasting balancing prices - those which explicitly model the balancing state and those which model it implicitly. Explicitly modeling the balancing state offers some advantages. It enables the analyst to make different assumptions about the balancing prices depending on direction, and gives the opportunity to include an explicit no-balancing state. As described in Section 2, the manual activation of reserves in the Nordic market creates a quite large dead band around zero, causing 50% of the hours in our data set to be in the no-balancing state. If balancing state is modeled implicitly, the balancing state is determined by the sign of the balancing price forecast, and the no-regulation state will only occur if the balancing price is equal to the spot price. Models which explicitly forecast the balancing state include those of Olsson and Söder [17] and Jaehnert et al [12], whereas Boomsma et al [1] and Brolin and Söder [2] use models that forecast the balancing market price without regard to the balancing state.

The other main distinction is whether the model takes in exogenous explanation factors or only relies on current and past price information. A commonly used exogenous explanation factor is the balancing volume, as used by [19], [12] and [2]. Another frequently used explanation factor is the day-ahead market price. The balancing market price is alternatively modelled directly, as in [1], or as the difference to the day-ahead market price, as is done in [12] and [19]. If the balancing market price is modelled as the difference to the day-ahead price, we will not regard this as using the day-ahead market price as exogenous input, but if the day-ahead market price is used as an explanation factor in itself, we will regard it as an exogenous explanation factor. Skytte [19] finds

that the day-ahead market price explains the balancing market price, whereas Jaehnert et al [12] find no correlation. It will therefore be interesting to further test the relation.

4.1 Models for state determination

State determination and forecasts conditional on state are natural topics for regime switching models. However, they are unsuitable for the purpose of determining balancing states and prices due to the fact that states are observable, and that there are no exogenous driving forces that can predict the states. Instead, we turn to Markov models and arrival rate models for predicting the balancing state. Jaehnert et al [12] use a SARIMA model and determine the balancing state the from the price forecasts. We prefer to utilize SARIMA models for price directly, and will come back to it in Section 4.3.

The possible states that we aim to model are summarized in Table 2. As described in [17], there is a fourth possible state which is balancing in both directions within the same hour. Typically there might be a regulation in one direction in the beginning of the hour and regulation in another towards the end. However, this state is so rare that we exclude that possibility. The balancing price and volume for hours with two balancing states were replaced by figures for the dominating direction when estimating parameters.

Table 2 Definition of balancing market states

State	Balancing volume ν	Balancing price ρ
no regulation	$\nu_{up}, \nu_{down} = 0$	$\rho_{up}, \rho_{down} = \rho_{spot}$
up regulation	$\nu_{up} > 0, \nu_{down} = 0$	$\rho_{up} > \rho_{spot}$
down regulation	$\nu_{down} > 0, \nu_{up} = 0$	$\lvert\rho_{down}\rvert < \rho_{spot}$

Markov models for determining the balancing state have been used by Olsson and Söder [17]. They used a non-time-homogenous Markov model, with different transition probabilities depending on the duration of the balancing state. In this article, we benchmark a duration dependent Markov model, with 7 different transition matrices. Balancing incidents with durations 0-5 hours had individual transition matrices, whereas a separate matrix was estimated for incidents lasting 6 hours or more.

Another take on avoiding static transition probability matrices is to include calender information by making the transition matrices dependent on the hour of the day. In this way, we can accommodate the fact that the probability of transition from one state to another is greater in the transition hours from day to night and night to day, whereas states are generally more stable in the middle of the night and during mid day. We applied Pearson's chi-square test to check whether a Markov transition matrix estimated for each individual

Table 3 Summary of balancing state determination models for benchmarking

Name	State determination model
Hour specific Markov	Markov transition matrix dependent on the hour of the day
Duration dependent Markov	Markov transition matrix dependent on the duration of the state
Arrival rate	Time between arrivals is a moving-average process

hour was significantly different from a transition matrix for all hours. As 5 hours (basically the day-night transition hours) were significantly different, and suitable alternative clustering of hours was hard to find, we continued with a Markov model with individual transition matrices for each hour of the day.

The other model that was tested in this article is inspired by inventory control theory, and based on the work of Croston [5] (see also [22]). This model only separates demand from non-demand, but does not discriminate between the balancing directions. In fact, it is a moving-average arrival rate. The time between the occurrence of two events is updated as a moving average every time an event occurs. Thus, in this model the state model discriminates between no regulation and regulation states, whereas the distinction between up- and down regulations is determined by separate price- or volume processes.

The main idea of Croston [5] was to separate the probability of the arrival of demand and the size of the demand into two different stochastic processes. Applying the approach of Willemain et al [22], the time between arrivals is modeled as a moving average in the following way: Let p_t be the (moving) average time between arrivals and let q_t be the specific number of time steps since last event. Let ν_t be the balancing volume in time step t. Then:

$$p_t = \begin{cases} p_{t-1} & \text{if } \nu_t = 0 \\ p_{t-1} + \alpha * (q_{t-1} - p_{t-1}) & \text{if } \nu_t \neq 0 \end{cases} \tag{2}$$

The probability of regulation (the arrival rate) for each time step is expressed as $1/p_t$.

The average time between regulation was calculated as 1.98 from the historical data. In his original article Croston [5] suggested using values of α in the range of 0.05-0.2, based on experience. In this work, α was estimated by minimizing the sum of squared residuals from the empirical arrival rate and the estimated arrival rate (described in (2)) in the historical data. The optimum was found at $\alpha = 0.01$, suggesting a rather slow-moving average.

A summary of state determination models that will be benchmarked can be found in Table 3.

4.2 Models for balancing volume forecasting

In forecasting balancing volume, we test the model from Jaehnert et al [12] with a randomly drawn volume given a balancing state. As in the original article, the general extreme value distribution was found to offer the best fit among all the tested probability distributions. However, since the fit was not particulary good, we also tested with random sampling from historical values.

Additional literature on time series forecasting of balancing volume is rather meagre. Whereas Möller [14] performs advanced analysis of the German market demand for balancing volume, we take a simpler approach and fit an ordinary SARIMA model. Since the augmented Dickey-Fuller test showed that the volume time series is not integrated, a SARMA-model will be sufficient. We found a SARMA(1,2)(1,1) model suitable.

As the Nordic market has many hours with no regulation state, a time series model that does not distinguish between incidents and the size of the incident has been showed to yield too low prediction with too high variability (cf [5]). We therefore also build a new model for balancing volume forecasting, with states determined by a moving average arrival rate model, as described in section 4.1. The model was originally formulated for inventory control problems, where the demand usually has two states: Either there is demand, or there is none. For balancing power, the state is more complicated, as demand either is zero, positive (upward regulation) or negative (downward regulation). However, we choose to discretize the state in two: regulation or no regulation. We could have imagined having two arrival rate processes - one each for upward and one for downward regulation. However, since there is no way of excluding the arrivals of both states in the same time step, we found that approach unsuitable for modelling the Nordic market. Instead, we let the arrival rate model determine the arrivals of balancing incidents, and the sign of the balancing volume forecast determine whether there is an upward or downward regulation. An added advantage, is that we then can take correlation between demands of different signs into account.

The balancing volume itself is modelled as a stationary unevenly spaced autoregressive process of order 1 (AR1), with parameters estimated according to the algorithms in [7]. Ordinary time series analysis techniques will fail, since they require evenly spaced data measurements, which would imply either artificially compressing the time series, or inserting 0 values where there really are no observations, thus distorting the variance (for more on uneven time series, see [13]). Instead, we use algorithms that acknowledge that adjacent observations are more strongly correlated than events further spaced apart in time, and that variance increase over time, and therefore should be scaled by the time between incidents.

A summary of the models benchmarked for volume can be found in Table 4.

Table 4 Summary of volume models for benchmarking

Name	State model	Volume forecast method
RAND	Markov	Random from distribution
HIST	Markov	Random from historical values
SARMA	None	Standard ARMA model
CROST	Arrival	AR1-model for unevenly spaced time series

4.3 Models for balancing premium forecasting

As a reference, a standard SARIMA time series model will be defined and benchmarked. Jaehnert et al [12] find a SARIMA(1,1,2)x(1,1,2)$_{24}$ model suited for short term forecasting of balancing market prices. The analysis of our data set revealed that the balancing premium time series was not integrated, and seasonal effects are so weak that they could be ignored. An ARMA(1,1) model was found to be sufficient and suitable.

Boomsma et al [1] use an autoregressive model with external input in order to make scenarios for the balancing market price (ρ^{BM}), as specified in (3)(where Ł() is the lag operator, ϕ is the autocorrelation coeffcient, β is the coefficient of the external input, and ϵ is the random error). The external input is the current and previous values of the spot market price (ρ^{spot}). These authors forecast the balancing market price directly, rather than as defining it as the difference to the day-ahead market price. The balancing market state is then defined implicitly, depending on whether the balancing market price is higher or lower than the day-ahead market price.

$$(1 - \phi L)(\rho_t^{BM} - \beta\rho_t^{spot}) = \epsilon_t \tag{3}$$

Reconstructing this model for our data set gave a fairly well specified model. An inspection of the residuals revealed thick tails and a slight autocorrelation in the residuals. The model could have been slightly improved through another error specification. However, the fit is well enough, and since our objective is benchmarking rather than tweaking the models to perfection, we leave it as in the original paper.

Olsson and Söder [17] also use a pure time series models to forecast balancing market prices. However, they use two different time series - one for upward regulation and one for downward regulation. A Markov model determines the switch between different balancing states. The continuous upward and downward regulation time series are assumed to be independent, but upward balancing time steps are assumed to be correlated with other upward balancing time steps and vice versa for downward balancing. The time series for upward- and downward balancing market premiums are assumed to be continuous. But we must use techniques for unevenly spaced time series to estimate the parameters, since the observed upward and downward balancing prices are not defined for all time steps. Using similar techniques as those

of Olsson and Söder [17], we estimated the parameters from analysis of the autocovariance function.

Our implementation differs from that of Olsson and Söder [17] in three ways: First, we use a hour-specific Markov model for transition probability, rather than a duration-dependent Markov model, due to better performance on longer-term forecast. Second, we do not find strong evidence for seasonality, and therefore limit our search for suitable models to the ARIMA-family of models. Third, we find that a simpler model with no differencing (i.e., we stick to ARMA-models) and fewer orders for the continuous up- and down processes is sufficient. An ARMA(1,1) process (with no intercept) was chosen for upward regulation, whereas an ARMA(2,1) with intercept was deemed suitable for downward regulation.

Jaehnert et al [12] found that balancing premiums to a large degree can be explained by the balancing volumes. We wanted to include a model that explained balancing market premiums from exogenously given time series. However, upon investigating the data, we found the correlation between the balancing volume and the balancing market premium to have weakened since the publication of Jaehnert et al [12]. Pearson's correlation coefficient had declined from 0.78 in the 2003-2007 NO1 data set to 0.47 in the 2010-2012 NO2 dataset.[3] We tried to include the balancing demand in neighboring price zones, without improved explanation power. In the end, we settled for two models (one for each balancing direction) where the balancing market premiums are determined by the balancing volume, the day-ahead market price and the overall power production in the NO2 price zone. The balancing volume was forecast using the CROST model of Section 4.2

As observed by Conejo et al [4], naive forecasts can be hard to beat when forecasting spot prices, and in industry these practices for predicting balancing prices are common too. For short-term forecasts, we will use the balancing market price from the last hours, but for day-ahead forecasts we will use the price for the same hour in a similar day. Although balancing market prices are less seasonal than day-ahead market electricity prices, we use a similar definition as that of Conejo et al [4].[4]

A summary of the models that will be benchmarked can be found in Table 5.

[3] The current NO2 price zone was formerly a part of NO1

[4] For Mondays, Saturdays and Sundays, we use the balancing market price of the same hour the previous week, whereas for Tuesdays,Wednesdays, Thursdays and Fridays, we use the same hour on the last workday. However, for day-ahead forecasts it must be taken into account that the balancing market price on Monday is not revealed entirely before the bidding for Tuesday closes at Monday noon, so the remaining hours are collected from Tuesday the last week.

Table 5 Description of balancing market price models for benchmarking

Name	State model	Volume model	Price forecast method
ARMA	None	None	ARMA time series model
ARX	None	None	AR1 model with spot price as exogenous input
ARM	Markov	None	ARMA models
EXO	Arrival	AR1 for unevenly spaced time series	Regression on exogenous factor
NAIVE	None	None	Price equal to price in similar hour

5 Test performance measurement

5.1 One step ahead vs multiple steps ahead

In short term forecasting, the one step ahead forecast is often used as a benchmark for how well a model performs. For the power producerd the one step ahead forecast may be relevant for intra-day operations, for instance in the trade-off between trading on a multilateral intra day market or taking part in the balancing market. The most important trade-off is however often between the day-ahead market and the balancing market. If a producer is to make coordinated bids between the day-ahead market and balancing market (see for instance [1]), a price forecast for the balancing market is needed before the day-ahead market closes, 12-36 hours ahead of the operating hour. Thus, both the one-step ahead and the 12-36 step ahead will be tested.

5.2 Point vs interval forecast

Forecasts are often evaluated by how well the forecast mean matches the observed value. Deviations can be measured, for instance, by using the mean absolute error (MAE). We will measure the models' ability to offer a point forecast in this way too. As by Weron and Misiorek [21], we will compare performance by looking at the MAE averaged over the week. Weron and Misiorek [21] calculate a quasi mean average percentage error (MAPE) by introducing a weighed MAE. The weekly average MAE is divided by the average price that week, so that one can avoid trouble calculating MAPE when the simulated prices are close to zero. For balancing prices, the problem is even worse, since balancing market premiums can take both positive and negative values, and the expected values are close to zero. Therefore, we choose to report the MAEs directly, but for the sake of comparison, we also provide the average absolute balancing market premiums of that week, $|\bar{\delta}_t|$.

Models for trade-off between trading in different markets are often based on stochastic optimization and the construction of scenario trees [1]. In these applications, the distribution of the forecast is equally or more important than the forecast mean. Therefore, the models are evaluated for their ability to produce correct probabilistic forecasts as well. We will evaluate the interval

forecasts by their unconditional coverage: Let y_t be an observed value in the out-of-sample period, and let $L_t(p)$ and $U_t(p)$ be the upper and lower limits of the probabilistic forecasts for coverage probability, p, respectively. We then define an indicator variable as follows:

$$I_t = \begin{cases} 1, & \text{if } y_t \in [L_t(p), U_t(p)] \\ 0, & \text{if } y_t \notin [L_t(p), U_t(p)] \end{cases} \tag{4}$$

[3] points out that the unconditional coverage can be a misleading measure if heteroskedasticity is present. Even if the unconditional coverage fits the theoretical percentiles on aggregated level, outliers may come clustered in times with higher variability. However, we find the unconditional coverage measure a sufficient sophisticated measure for benchmarking, and caution future potential users of these models to take a closer look at the conditional coverage before implementing them.

6 Forecasting performance

6.1 Benchmarking of state determination models

In order to determine which state determination model is better, we simulated 3000 out-of-sample scenarios (1920 time steps) for the two variants of each model: one hour ahead forecasts and 12–36 hours ahead forecasts. Then we compared the simulated states to the observed states. Every time the model predicted the correct state, a score of 1 was assigned, otherwise the score was 0. We then compared the mean for all the 3000 scenarios, both for each time step and averaged over all the 1920 time steps.

The two alternative Markov models were benchmarked directly against each other. In order to compare with the arrival rate model, we assessed the Markov models ability to discriminate between a regulation state and a non-regulation state, and compared the results with those of the arrival rate model.

The performance of the two Markov models are illustrated in Figure 6.1. The average score for both models are found in Table 6. The comparison shows that the duration dependent model performs better on short horizon forecasts, whereas the hour dependent Markov model is slightly better in the long run. Neither model has an impressive hit rate for the day-ahead forecasts, the duration dependent model being slightly worse. This is not too surprising, as the duration dependent Markov model uses two state information measures (current state and current duration), and both errors grows larger with the time horizon.

The two Markov models' and the arrival rate models' ability to predict the correct state when considering only two states (regulation and no regulation) can be found in Table 7. For short-horizon forecasts, the duration dependent Markov model is best. However, it performs worst when it comes to the day-ahead horizon. The score of the arrival rate model is remarkably stable,

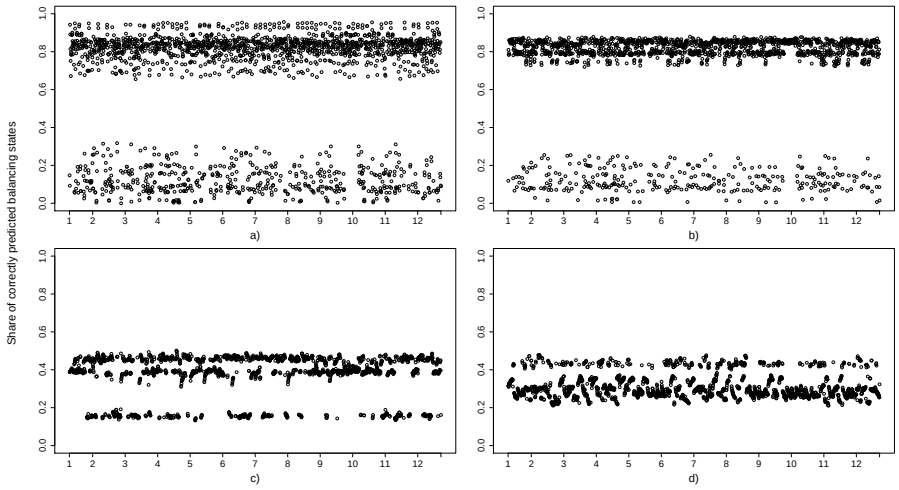

Fig. 3 Share of correct predicted balancing state from the three state Markov models, hour by hour in the out-of-sample period. The week numbers are displayed on the horizontal axis. Forecasts are for a) one hour ahead, hour specific Markov model b) one hour ahead duration dependent Markov model, c) day-ahead forecast, hour specific Markov model, d) day-ahead forecast, duration dependent Markov model

Table 6 Share of correctly predicted balancing states (up, down, no balancing)

Forecast horizon (hours ahead)	1	12–36
Hour specific Markov	0.63	0.37
Duration dependent Markov	0.73	0.35

probably due to the relative stable arrival rate (moving average coefficient α as low as 0.01).

Table 7 Share of correctly predicted balancing state. Two states : balancing, no balancing

Forecast hours ahead	1	12-36
Arrival rate	0.59	0.59
Hour specific Markov	0.67	0.54
Duration dependent Markov	0.73	0.51

For day-ahead forecasts it seems that the hour specific Markov model and the arrival rate model give the best forecast. As they are qualitatively different in the number of states they are able to predict, the analyst must weigh the arrival rate model's precision against the disadvantage of operating with only a binary balancing state.

6.2 Benchmarking of balancing volume forecasting models

The mean average error of the four different models can be found in Table 8. The ranking of the different model is quite clear: for short-term forecasting (one hour ahead), the SARMA-model outperforms all other models in all weeks. For day-ahead forecast, the CROST model with unevenly spaced time series is best for all weeks, except week 1 and 8 where the SARMA-model is better. The task of predicting the day-ahead balancing market volume is difficult; the SARMA model has the worst performance in week 10 and 12, whereas the CROST model is the worst in week 1. Thus, for day-ahead forecasting, no model is unambiguously the best. The models without memory - RAND and HIST - perform badly in times of spikes - for instance in week 1, 9 and 11 (cf Fig 1).

Table 8 Mean absolute error of balancing market volume forecasting for various models. MWh. Weekly average.

		One hour ahead forecast				**Day ahead forecast**			
W	$\lvert\bar{\nu_t}\rvert$	RAND	HIST	SARMA	CROST	RAND	HIST	SARMA	CROST
1	109.80	77.68	76.73	37.99	67.53	108.44	107.25	103.63	110.18
2	59.12	59.74	53.96	33.56	46.73	88.97	86.50	54.64	30.60
3	86.65	65.54	60.36	44.57	68.79	100.28	98.39	72.20	45.17
4	35.05	66.18	54.83	34.13	33.79	63.47	60.97	61.09	20.61
5	90.43	68.02	56.23	41.69	96.00	79.61	80.41	71.47	47.40
6	43.03	61.29	43.23	25.02	33.21	73. 88	70.54	53.46	25.54
7	21.45	49.28	35.97	15.22	19.38	51.82	50.19	27.99	14.69
8	43.02	53.49	44.04	18.97	33.62	56.60	56.10	19.04	25.89
9	121.85	83.81	80.12	35.90	78.31	106.29	107.01	78.69	65.02
10	94.84	89.05	77.77	35.48	58.70	87.94	87.48	145.74	50.86
11	98.37	73.21	71.35	43.86	65.66	117.01	113.50	57.52	55.78
12	39.88	45.85	36.59	30.60	39.85	66.52	64.10	104.29	41.55

The models' ability to create well calibrated probabilistic forecasts can be evaluated by studying Table 9. The table shows how many of the observed values that fall within the limits of four specified interquantile ranges: 50, 75, 90 and 99 %. Generally, the models are too narrow in the middle range, except for the SARMA model, which is too wide. Extreme values are captured quite well for all the models. No model captures the probabilistic structure for one-hour ahead forecast very well, but for day-ahead forecast the CROST model has a better performance.

Conclusively, the models with memory generally perform better than those without memory on both short and day-ahead horizons. No model is unambiguously the best in all respects; however, the CROST model has satisfactory performance on the day-ahead forecasts both in terms of capturing the spread and having a mean average error lower than the other models for all but two weeks. The SARMA model has relatively low mean average errors in forecasting the balancing volume an hour ahead, but the variance is too large. This

Table 9 Unconditional coverage. Percentage of observed balancing volume that fall within the interval specified by the simulated values' median +- range/2

	One hour ahead forecast				**Day ahead forecast**			
Range	RAND	HIST	SARMA	CROST	RAND	HIST	SARMA	CROST
50 %	24.90	30.94	77.76	24.32	15.16	15.36	74.01	43.18
75 %	42.03	43.07	87.66	82.60	71.77	71.98	84.64	77.60
90 %	87.86	82.24	93.75	93.44	90.36	91.82	91.30	88.13
99 %	99.32	98.65	97.55	98.18	99.95	99.69	97.92	97.45

result was anticipated by [5] for models that do not separate the stochastic processes of arrival and size of demand.

6.3 Benchmarking of price forecasts

In Table 10, the weekly average mean absolute error for forecasts of balancing market premium is shown. For the one hour ahead forecast, the naive model is hard to beat. In 8 of the 12 weeks, the naive model has the best short-term performance. The ARM model generally performs well, and is best in 3 of 12 weeks, whereas the pure ARMA model is best in one week. The EXO model seems to be a little less accurate for one hour ahead forecast. This is to be expected, since the linkage from time step to time step in the EXO model is based on balancing state and balancing volume, and not on balancing premium directly. Since the correlation between balancing market volume and price is lower than it historically has been, balancing volume acts as a worse predictor for the balancing premium than the lagged values of the premium itself.

For the day-ahead forecasts, the striking result is how similar the four non-naive models perform. Moreover, the mean average error of the models is very close to the mean absolute balancing market premium of that week. Thus, the accuracy of the model is comparable to a model where the balancing market premium forecast is constant and zero. A practitioner of forecasting might find this a disappointing result, but keeping the structure of the power market in mind, the result is not too surprising. Factors that could influence the balancing market price, such as the outage of plants, weather conditions or production from intermittent sources, will be taken into account when performing day-ahead bidding if known before the closure of the day-ahead market, and thus are reflected in the spot market price rather than the balancing market price. Thus, there is no information basis that can aid the forecasting of next day's balancing prices before the day-ahead market has closed.

The naive model, which uses the balancing prices from the day before (shifted back to account for weekend effects if necessary), performs worse than the other models in all but two weeks.

The evaluation of the methods' probabilistic forecasts are displayed in Table 11. The table shows a clear distinction between the models which includes state information (EXO and ARM) versus those which are purely time series

Table 10 Mean absolute error for various BMpremium models. NOK/MWh. Weekly average.

		One hour ahead forecast					Day ahead forecast				
w	$\lvert\bar{\delta_t}\rvert$	ARMA	ARX	ARM	EXO	NAIVE	ARMA	ARX	ARM	EXO	NAIVE
1	40.82	14.89	19.39	11.51	26.50	9.68	40.92	41.11	40.31	40.76	46.61
2	53.19	29.21	49.61	28.76	45.68	29.41	53.37	52.89	54.55	53.31	64.91
3	93.19	48.63	68.32	50.92	82.98	42.56	93.03	94.77	93.39	94.12	145.76
4	64.23	41.05	72.61	43.64	62.56	42.63	64.81	65.12	64.20	65.69	85.62
5	40.86	14.70	18.02	9.78	26.95	9.57	41.06	41.18	39.17	41.13	42.84
6	30.15	15.32	23.80	15.28	22.81	15.19	30.36	30.28	31.21	30.18	38.52
7	36.98	18.03	25.51	17.02	32.70	16.03	37.66	37.44	37.89	37.63	47.43
8	31.38	12.38	16.09	9.98	25.88	9.10	31.75	31.79	30.57	32.39	35.26
9	32.90	11.90	14.73	8.24	19.44	7.52	33.13	32.97	31.91	33.16	21.19
10	32.66	12.44	14.63	10.26	19.53	8.58	32.69	32.62	31.61	32.06	29.08
11	89.66	49.27	82.98	49.16	78.58	52.48	89.74	90.22	88.40	90.28	131.84
12	87.58	63.25	88.48	57.07	81.80	64.68	87.58	88.39	87.98	87.85	145.47

based (ARMA and ARX). The models without state information generally have forecasts that are too wide. This illustrates the point of Croston [5] - not discriminating between demand and non-demand in forecasting may lead to an overestimation of variance. Although admittedly not perfect, the models that include state information better reflect the distribution of the observed values. For day-ahead forecasts, the EXO-model performs slightly better than the ARM-model. This may be due to the fact that the ARM-model is based on a Markov model for state determination, whereas the EXO-model is based on a balancing volume model that utilises a moving average arrival rate for determining the balancing state. As seen in Section 6.1, the arrival rate model performs better than Markov models for day-ahead forecasts.

Table 11 Unconditional coverage. Percentage of observed balancing market premiums that fall within the interval specified by the simulated values' median +- range/2

	One hour ahead forecast				Day ahead forecast			
range	ARMA	ARX	ARM	EXO	ARMA	ARX	ARM	EXO
50%	86.51	79.01	39.01	58.44	80.36	80.63	37.76	14.27
75%	91.98	89.22	79.84	67.71	92.97	92.76	78.13	77.19
90%	94.48	92.45	88.70	83.07	95.21	95.21	89.32	91.56
99%	96.77	94.69	94.64	96.09	97.08	97.24	95.36	96.41

The results from the evaluation of the probabilistic forecast shown in Table 11 imply that building models for forecasting balancing market prices is not futile, even though the forecasts are outperformed by naive models on an hour-ahead horizon, and even though the average errors for day-ahead forecasts are on the same level as a constant forecast of zero would have delivered. The shape of the simulated distribution is important for making good operational decisions in real bidding situations. Studying the day-ahead forecasts we note

that models without balancing state descriptions severely overestimate variance, and create too wide forecasts. The models with explicit information on balancing state (ARM and EXO) create better interval forecasts.

7 Conclusions

In this paper, we have recreated and developed three models for the prediction of balancing state, four models for the prediction of balancing volume and five models for forecasting of the balancing market price premium. All models have been benchmarked, with special emphasis of the ability of the model to create balancing market premium forecast that can assist the bidding process where a producer has to decide on how to allocate power between the day-ahead market and the balancing market.

Our analysis confirms that it is hard to predict the balancing market before the closure of the day-ahead market. The balancing market is designed to handle unforeseen events and fluctuation, and therefore we are not surprised by concluding that the volume and the premium in the balancing market are random. In fact, it could be interpreted as a sign of an efficient electricity market that it is not possible to predict the balancing market price. Any predictable relation between the information available before the closing of the day-ahead market and the balancing market would open speculative possibilities, as the producers then could make a profit by buying in the day-ahead market and sell in the balancing market (or vice versa).

However, stating that it is impossible to capture the expected balancing market premium precisely, does not mean that balancing market forecasting is futile or that it does not matter which forecasting model that is used. The evaluation of the interval forecasts clearly shows that models which include balancing state describe the distribution of the forecasted premium or volume far better than models without balancing state information. Thus, we have shown that the observations of Croston [5] also applies to the balancing market: Separating between demand and non-demand is important for estimating the variance correctly. Getting the distribution of scenarios right is crucial for stochastic optimization models, thus for that purpose we strongly recommend using models with balancing state information for scenario generation.

References

1. Boomsma TK, Juul N, Fleten SE (2013) Bidding in sequential electricity markets: The Nordic case, submitted to European Journal of Operational Research
2. Brolin MO, Söder L (2010) Modeling Swedish real-time balancing power prices using nonlinear time series models. In: IEEE 11th International Conference on Probabilistic Methods Applied to Power Systems, pp 358–363

3. Christoffersen PF (1998) Evaluating interval forecasts. International Economic Review 39(4):841–862
4. Conejo AJ, Contreras J, Espìnola R, Plazas M (2005) Forecasting electricity prices for a day-ahead pool-based electric energy market. International Journal of Forecasting 21:435–462
5. Croston J (1972) Forecasting and stock control for intermittent demands. Operational Research Quarterly 23(3):289–303
6. ENTSO-E (2012) Operational reserve ad hoc team report final version. Tech. rep., URL http://www.eprg.group.cam.ac.uk/wp-content/uploads/2008/11/eprg0711.pdf
7. Erdogan E, Ma S, Beygelzimer A, Rish I (2005) Statistical models for unequally spaced time series. In: Proceedings of the 2005 SIAM International Conference on Data Mining
8. European Wind Energy Association (2012) Wind in power: 2012 European statistics, URL http://www.ewea.org/fileadmin/files/library/publications/statistics/Wind_in_power_annual_statistics_2012.pdf
9. Fleten SE, Pettersen E (2005) Constructing bidding curves for a price-taking retailer in the Norwegian electricity market. IEEE Transactions on Power Systems 20(2):701–708
10. Glachant JM, Saguan M (2007) An institutional frame to compare alternative market designs in EU electricity balancing. Tech. Rep. EPRG 0711, Electricity Policy Research Group, University of Cambrigde
11. Holttinen H (2008) Estimating the impacts of wind power on power systems - summary of IEA wind collaboration. Environmental Research Letters 3(2)
12. Jaehnert S, Farahmand H, Doorman GL (2009) Modelling of prices using the volume in the Norwegian regulating power market. In: IEEE PowerTech, Bucharest
13. Jones RH (1986) Time series regression with unequally spaced data. Journal of Applied Probability 23:89–98
14. Möller C (2010) Balancing energy in the German market design. PhD thesis, Universität Karlsruhe
15. Möller C, Rachev ST, Fabozzi FJ (2011) Balancing energy strategies in electricity portfolio management. Energy Economics 22(1):2–11
16. Nogales FJ, Contreras J, Conejo AJ, Espnola R (2002) Forecasting next-day electricity prices by time series models. IEEE Transactions on Power Systems 17(2):342–348
17. Olsson M, Söder L (2008) Modeling real-time balancing power market prices using combined SARIMA and Markov processes. IEEE Transactions on Power Systems 23(2):443–450
18. Rivero E, Barqun J, Rouco L (2011) European balancing markets. In: 8th International Conference on the European Energy Market (EEM), pp 333–338
19. Skytte K (1999) The regulating power market on the Nordic power exchange Nord Pool: an econometric analysis. Energy Economics 21(4):295–

308
20. van der Veen RA, Abbasy A, Hakvoort RA (2012) Agent-based analysis of the impact of the imbalance pricing mechanism on market behavior in electricity balancing markets. Energy Economics 34(4):874–881
21. Weron R, Misiorek A (2008) Forecasting spot electricity prices: A comparison of parametric and semiparametric time series models. International Journal of Forecasting 24:744–763
22. Willemain TR, Smart CN, Shockor JH, DeSautels PA (1994) Forecasting intermittent demand in manufacturing: a comparative evaluation of Croston's method. International Journal of Forecasting 10(4):529–538