HSC Research Report

HSC/06/01

Short-term electricity price forecasting with time series models: A review and evaluation

Rafał Weron*
Adam Misiorek**

* Hugo Steinhaus Center, Wrocław University of Technology, Poland
** IASE – Institute of Power Systems Automation, Wrocław, Poland

Hugo Steinhaus Center
Wrocław University of Technology
Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland
http://www.im.pwr.wroc.pl/~hugo/

Rafał Weron
Hugo Steinhaus Center, Wrocław

Adam Misiorek
IASE, Wrocław

1. Short-term electricity price forecasting with time series models: A review and evaluation

1.1. Introduction

In the last decades, with deregulation and introduction of competition a new challenge has emerged for power market participants. Extreme price volatility, which can be even two orders of magnitude higher than for other commodities or financial instruments, has forced producers and wholesale consumers to hedge not only against volume risk but also against price movements. Price forecasts have become a fundamental input to an energy company's decision-making and strategy development. This in turn has propelled research in electricity price modeling and forecasting [5][35].

The proposed solutions can be classified both in terms of the planning horizon's duration and in terms of the applied methodology. In the first case it is customary to talk about short-term (STPF), medium-term (MTPF) and long-term price forecasting (LTPF). The main objective of LTPF is investment profitability analysis and planning, such as determining the future sites or fuel sources of power plants. Lead times are typically measured in years. Medium-term or monthly time horizons are generally preferred for balance sheet calculations, risk management and derivatives pricing. In many cases not the actual point forecasts but the distributions of future prices over certain time periods are evaluated. As this type of modeling has a long-dated tradition in finance, inflow of “finance solutions” is readily observed [6][16][35].

However, not only monthly or annual time horizons are interesting for generators, utilities and power marketers. When bidding for spot electricity in a power exchange or a pool-type market, actors are requested to express their bids in terms of prices and quantities. Buy (sell) orders are accepted in order of increasing (decreasing) prices until total demand (supply) is met. A power plant that is able to forecast spot prices can adjust its own production schedule accordingly and hence maximize its profits. Since the day-ahead spot market typically consists of 24 hourly (or 48 half-hourly) auctions that take place simultaneously one day in advance, STPF with lead times from a few hours to a few days is of prime importance in day-to-day market operations [4][8][35]. It is also the topic of this study, which concentrates on proposing several time series models and comparing their short-term forecasting performance during normal as well as extremely volatile periods in the history of the deregulated California power market. An assumption is made that only publicly available information is used to predict spot prices, i.e. generation constraints, line capacity limits or other fundamental variables are not considered.

The chapter is structured as follows. In Section 1.2 we review modeling approaches for spot electricity prices. In the following section we present relevant time series models. We start with linear time series, followed by their extensions that allow for incorporating exogenous (fundamental) factors. Since the residuals of the linear models typically exhibit heteroscedasticity, next we discuss implementations of AR-GARCH type models. Finally, we introduce threshold autoregressive models that, by construction, should be well suited for modeling the non-linear nature of electricity prices. In Section 1.4 we describe the dataset, introduce our models, present calibration details and define error measures. In Section 1.5 we provide empirical forecasting results for the studied models and compare the results with those of other authors. Finally, in Section 1.6 we summarize the results and draw conclusions.

1.2. Overview of modeling approaches

Electricity spot price modeling and forecasting techniques generally can be traced back to models that originate either in electrical engineering or in finance. For some time power engineers have been familiar with both scheduling and dispatching units in the system and load forecasting. With the restructuring of the electric power industry, it has been very natural for the engineers to adapt these models to price forecasting under the new economic conditions. Production-cost models were directly transferred or amended with strategic bidding considerations, while load forecasting techniques were additionally supplied with past price data to yield price forecasts.

On the other hand, price modeling and forecasting has long been at the center of intense studies in other commodity and financial markets. Depending on the objectives of the analysis, a number of methods for modeling price dynamics have been proposed, ranging from parsimonious stochastic models to fundamental and game theoretic approaches. It was only a question of time before these methods were put into use in the power markets.

The various approaches that have been developed to analyze and predict power markets' behavior and the resulting electricity prices may be broadly divided into six classes [35]:

- ❑ **Production-cost** (or **cost-based**) models, which simulate the operation of generating units aiming to satisfy demand at minimum cost. They have the capability to forecast prices on an hour-by-hour, bus-by-bus level. However, they ignore the strategic bidding practices, including execution of market power [22]. They were appropriate for the regulated markets with little price uncertainty, stable structure and no gaming, but are not well suited for the recently established competitive markets.
- ❑ **Equilibrium** (or **game theoretic**) approaches, which may be viewed as generalizations of cost-based models amended with strategic bidding considerations. They may give good insight into whether prices will be above marginal costs and how this might influence the players' outcomes. But they pose problems if more quantitative conclusions have to be drawn. Furthermore, a number of components have to be defined: the players, their potential strategies, the ways they interact and the set of payoffs. Obviously, a substantial modeling risk is present. In general, two types of approaches have been proposed [13][34]: the Cournot-Nash framework (see eg. [7][38]), which tends to provide higher prices than those observed in reality, and the supply function equilibrium framework (see eg. [19][37]), which requires considerable numerical computations and, consequently, has limited applicability in day-to-day market operations.
- ❑ **Fundamental** (or **structural**) methods, which describe price dynamics by modeling the impact of important physical and economic factors on the price of electricity [4][32]. The functional associations between fundamental drivers – loads, weather conditions, system parameters, etc. – are postulated (consequently, there exists a significant modeling risk) and the fundamental inputs are independently modeled and predicted, often via statistical, econometric or non-parametric techniques. Some recent examples include [14][33]. Because of the nature of fundamental data (which is typically collected over longer time intervals; data availability is a separate issue), pure fundamental models are better suited for medium-term rather than short-term predictions.
- ❑ **Quantitative** (or **stochastic**, **econometric**, **reduced-form**) models, which characterize the statistical properties of electricity prices over time, with the ultimate objective of derivatives evaluation and risk management. Consequently, these models are not required to accurately forecast hourly prices but to recover the main characteristics of electricity prices (in particular, seasonality, mean-reversion, high volatility and the occurrence of spikes), typically at the daily time scale. The tools and approaches used are generally adopted from methods developed for modeling interest rates or other commodities [6][16]. Based on the type of the market in focus, the stochastic techniques can be divided into two main classes: spot and forward price models. The former have problems with pricing derivatives, i.e. with identifying the risk premium linking spot and forward prices. The latter lack the data that can be used for calibration and are often unable to yield the properties of spot prices from the analysis of forward curves.

- **Statistical** (or **technical analysis**) approaches, which either are direct applications of the statistical techniques of load forecasting or power market implementations of econometric models. While the efficiency and usefulness of technical analysis in financial markets is often questioned, in power markets these methods do stand a better change. The reason for this is the seasonality prevailing in electricity price processes during normal, non-spiky periods. It makes the electricity prices more predictable than those of "very randomly" fluctuating financial assets. Moreover, to enhance their efficiency many of the statistical approaches considered in the literature often incorporate one or two fundamental factors, like loads or fuel prices. Examples of statistical models are further discussed in Section 1.3.
- **Artificial intelligence-based** (or **non-parametric**) techniques, which model price processes via non-parametric tools such as neural networks, fuzzy logic, etc. AI-based models tend to be flexible and can handle complexity and non-linearity. This makes them promising for short-term predictions and a number of authors have reported their excellent performance in STPF. Like in load forecasting, artificial neural networks (ANNs) have probably received the most attention [30][39]. Other non-parametric techniques have been also applied, however, typically in hybrid constructions [29].

Of the six above mentioned approaches, statistical and AI-based models are best suited for STPF, in particular at the hourly time horizon. It would be interesting to compare representatives from both groups, however, because of limited space here we will only review one group. The choice is backed by results of a recent study by Conejo et al. [8], who compared different methods of STPF: three time series specifications (transfer function, dynamic regression and ARIMA), a wavelet multivariate regression technique and a multilayer perceptron with one hidden layer. For a dataset comprising PJM prices from year 2002, the ANN technique was the worst out of the five tested models! Surely more research is needed but this report already indicates that there might be serious problems with the efficiency of ANNs and AI-based methods in general. Consequently, in this chapter we will utilize statistical approaches.

1.3. Time series models

1.3.1. ARMA models and their extensions

In the engineering context the standard model that takes into account the random nature and time correlations of the phenomenon under study is the autoregressive moving average (ARMA) model. In the ARMA model the current value of the process (say, the price) is expressed linearly in terms of its past values (autoregressive part) and in terms of previous values of the noise (moving average part) [2]:

$$\phi(B)P_t = \theta(B)\varepsilon_t \, , \tag{1}$$

where B is the backward shift operator, i.e. $B^h P_t \equiv P_{t-h}$, $\phi(B)$ is a shorthand notation for $\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$ and $\theta(B)$ is a shorthand notation for $\theta(B) = 1 + \theta_1 B + \ldots + \theta_q B^q$. Furthermore, $\phi_1, \ldots, \phi_p$ and $\theta_1, \ldots, \theta_q$ are the coefficients of autoregressive and moving average polynomials, respectively, and ε_t is independent and identically distributed (iid) noise with zero mean and finite variance. For $q = 0$ we obtain the well known autoregressive AR(p) model.

The ARMA modeling approach assumes that the time series under study is (weakly) stationary. If it is not, then a transformation of the series to the stationary form has to be done first. This can be performed by differencing. The resulting ARIMA model contains autoregressive as well as moving average parts, and explicitly includes differencing in the formulation [2]. If differencing is performed at a larger lag than 1 then the obtained model is known as seasonal ARIMA or SARIMA.

Cuaresma et al. [11] applied variants of AR(1) and general ARMA processes (including ARMA with jumps) to STPF in the German market. They concluded that specifications, where each hour of the day was modeled separately presented uniformly better forecasting properties than specifications for the whole time-series, and that the inclusion of simple probabilistic processes for the arrival of extreme price events (jumps) could lead to improvements in the forecasting abilities of univariate models for electricity spot prices. In a related study, Weron and Misiorek [36] used various autoregression schemes for modeling and forecasting prices in California. They observed that an AR model where each hour of the day was modeled separately performed better than a single for all hours, but large (S)ARIMA specification proposed by Contreras et al. [10]. The reduction in MWE reached even 30% for a normal, non-spiky out-of-sample test period (first week of April 2000).

Conejo et al. [9] proposed a wavelet-ARIMA technique. It consists of a level 3 decomposition of the price series using a discrete wavelet transform, modeling the resulting detail and approximation series with ARIMA processes to obtain 24 hourly predicted values and applying the inverse wavelet transform to yield the forecasted prices for the next 24 hours. The performance of the wavelet-ARIMA technique is generally better than that of a standard ARIMA process. In all four weekly test samples (Spanish market, year 2002) the MWE's were reduced; for the winter week the error dropped even by 25%.

1.3.2. Time series models with exogenous variables

ARIMA-type models relate the signal under study to its own past and do not explicitly use the information contained in other pertinent time series. However, electricity prices may also be influenced by the present and past values of various exogenous factors, most notably the load profiles and ambient weather conditions. To accurately capture these relationships, time series models with *exogenous* or *input* variables can be used. These *hybrid* models do not constitute a new class, rather they can be viewed as generalizations of the existing ones.

The autoregressive moving average model with exogenous variables $v^1,\ldots,v^k$, or ARMAX($p,q,r_1,\ldots,r_k$), can be compactly written as [21]:

$$\phi(B)P_t = \theta(B)\varepsilon_t + \sum_{i=1}^{k} \psi^i(B)v_t^i \ , \qquad (2)$$

where r_i 's are the orders of the exogenous factors (e.g. system load, temperature, power plant availability) and ψ_j^i 's are the relevant coefficients. Alternatively, the ARMAX model is often defined in a “transfer function” form:

$$P_t = \frac{\theta(B)}{\phi(B)}\varepsilon_t + \sum_{i=1}^{k} \tilde{\psi}^i(B)v_t^i \ , \qquad (3)$$

where $\tilde{\psi}_j^i$'s are the appropriate coefficients. Additionally, the differencing transformation can be imposed leading to ARIMAX and seasonal ARIMAX models. Models with input variables are also known as *transfer function*, *dynamic regression*, *Box-Tiao*, *intervention* or *interrupted* time series models [2]. Some authors distinguish among them, others use the names interchangeably causing a lot of confusion in the literature.

Time series models with exogenous variables have been extensively applied to STPF. Nogales et al. [26] utilized ARMAX and ARX models (which they called “transfer function” and “dynamic regression”, respectively) for predicting hourly prices in California and Spain. Both models performed comparably, with the weekly MAPE (note that Nogales et al. called it the “Mean Weekly Error”; consult also Section 1.4.4) just below 3% for the first week of April 2000 in

California and around 5% for the third weeks of August and November 2000 in Spain. The results were significantly better than for the ARIMA and ARIMA-E (ARIMA with load as an explanatory variable) models proposed by Contreras et al. [10]. It is somewhat surprising that the “transfer function” and “dynamic regression” models – which also utilized one common multi-parameter specification for all hours – outperformed by over 40% the ARIMA-E model. After all, “transfer function” and ARIMA-E are more or less equivalent in terms of variables used. Possibly this is related to the way the load data is included in both methods. In ARIMA-E it is just an explanatory variable, but in the “transfer function” specification it is bundled with the autoregressive part of the model. What is even more surprising, ARIMA performed comparably to ARIMA-E, even though the latter additionally used an important exogenous variable [10].

In a review study, Conejo et al. [8] compared different methods of STPF: three time series specifications (“transfer function”, “dynamic regression” and ARIMA), a wavelet multivariate regression technique and a multilayer perceptron ANN with one hidden layer. For a dataset comprising PJM prices from year 2002, the time series models with exogenous variables yielded the best performance; for the last week of July 2002 better by over 75% (!) than the ARIMA predictions. Interestingly, the worst results were obtained for the ANN.

Weron and Misiorek [36] and Misiorek et al. [23] took a different line of approach. They used a set of 24 relatively small ARX models, one for each hour of the day, with the CAISO day-ahead load forecast as the exogenous variable and three dummies for recovering the weekly seasonality. They concluded that these models performed much better than a single for all hours, but large (S)ARIMA specification proposed by Contreras et al. [10] and slightly worse than the “transfer function” and “dynamic regression” models of Nogales et al. [26]. However, only the results for the first week of April 2000 in the California power market could be compared as this was the only common test sample used in all four papers. Consequently, the question whether the common for all hours, multi-parameter specification is also superior for other periods (and other data sets) remains open.

Knittel and Roberts [20] considered various econometric models for modeling and STPF in the California market, including mean-reverting diffusions and jump diffusions, a seasonal ARMA process (called “ARMAX”), an AR-EGARCH (Autoregressive Exponential GARCH) and a seasonal ARMA model with temperature, squared temperature and cubed temperature as explanatory variables. They found all temperature variables to be highly statistically significant during the pre-crisis period (April 1, 1998 to April 30, 2000). The WRMSE (Weekly Root Mean Square Error) was also the lowest of all models examined, though the difference from the seasonal ARMA process was small.

1.3.3. Autoregressive GARCH models

The linear ARMA-type models assume homoscedasticity, i.e. a constant variance and covariance function. From an empirical point of view, financial time series – and electricity spot prices in particular – present various forms of non-linear dynamics, the crucial one being the strong dependence of the variability of the series on its own past. Some non-linearities of these series are a non-constant conditional variance and, generally, they are characterized by the clustering of large shocks or heteroskedastity [35].

The AutoRegressive Conditional Heteroskedastic (ARCH) model of Engle [15] was the first formal model which successfully addressed the problem of heteroskedastity. In this model the conditional variance of the time series is represented by an autoregressive process, namely a weighted sum of squared preceding observations. In practical applications it turns out that the order of the calibrated model is rather large. Surprisingly, if we let the conditional variance depend not only on the past values of the time series but also on a moving average of past conditional variances the resulting model allows for a more parsimonious representation of the data. This model, the Generalized AutoRegressive Conditional Heteroskedastic GARCH(p,q) model put forward by Bollerslev [1] is defined as:

$$h_t = \varepsilon_t \sigma_t \text{ with } \sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i h_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-j}^2 , \qquad (4)$$

where ε_t are as before and the coefficients have to satisfy $\alpha_i, \beta_j \geq 0, \alpha > 0$ to ensure that the conditional variance is strictly positive. Identification and estimation of GARCH models is performed analogously to that of (S)AR(I)MA models; maximum likelihood is the preferred algorithm [35].

The GARCH model by itself is not attractive for STPF, however, coupled with autoregression (or a more general (S)AR(I)MA model) presents an interesting alternative – the AR-GARCH model, where the residuals of the regression part are further modeled with a GARCH process. Nevertheless, the general experience with GARCH-type components in statistical or econometric STPF models is mixed. There are cases when modeling heteroscedasticity is advantageous, but there are at least as many examples of poor performance of such models.

Knittel and Roberts [20] evaluated an AR-EGARCH specification and found it superior to five other models during the crisis period (May 1, 2000 to August 31, 2000) in California. However, the AR-EGARCH process yielded the worst forecasts of all models examined during the pre-crisis period (April 1, 1998 to April 30, 2000). A similar result was obtained by Garcia et al. [18] who studied ARIMA models with GARCH residuals and concluded that ARIMA-GARCH outperforms a generic ARIMA model, but only when high volatility and price spikes are present. In a related study Mugele et al. [25] proposed ARMA-GARCH time series with α-stable innovations for modeling the asymmetric and heavy-tailed nature of electricity spot price returns from the Nordic and German power markets.

1.3.4. Regime-switching models

The "spiky" character of spot electricity prices suggests that there exists a non-linear mechanism switching between normal and high-price states or regimes. As such these processes should be prone to modeling with the so-called *regime switching* models. The available specifications of regime switching models differ in the way the regime evolves over time. Roughly speaking, two main classes can be distinguished [35]: those where the regime can be determined by an observable variable (and, consequently, the regimes that have occurred in the past and present are known with certainty) and those where the regime is determined by an unobservable, latent variable (which implies that we can never be certain that a particular regime has occurred at a particular point in time, but can only assign probabilities to their occurrences).

The most prominent member of the first class is the Threshold AutoRegressive (TAR) model, which assumes that the regime is specified by the value of an observable (threshold) variable v_t relative to a threshold value/level T :

$$\begin{cases} \phi_1(B) P_t = \varepsilon_t & v_t \geq T, \\ \phi_2(B) P_t = \varepsilon_t & v_t < T, \end{cases} \qquad (5)$$

where $\phi_i(B)$ is a shorthand notation for $\phi_i(B) = 1 - \phi_{i,1} B - \ldots - \phi_{i,p} B^p$. To simplify the exposition, in this study we have specified a two-regime model only, however, generalization to multi-regime models is straightforward. The Self Exciting TAR (SETAR) model arises when the threshold variable is taken as the lagged value of the price series itself, i.e. $v_t = P_{t-d}$. It can be further modified by allowing for a gradual transition between the regimes, leading to the Smooth Transition AR (STAR) model [17]. A popular choice for the transition function is the logistic function; the resulting model is known as the Logistic STAR (LSTAR) model.

There are only a few documented applications of regime-switching TAR-type models to electricity prices. Robinson [28] fitted an LSTAR model to prices in the English and Welsh wholesale electricity Pool and showed that it performed superior to a linear autoregressive alternative. Stevenson [31] calibrated AR and TAR processes to wavelet filtered half-hourly data from the New South Wales (Australia) market. He concluded that the TAR specification (with the observable threshold variable being the change in demand and the threshold value $T = 0$) outperformed the AR alternative in forecasting performance.

Recently Rambharat et al. [27] introduced a SETAR-type model with an exogenous variable (temperature recorded at the same time as the maximum price of the day) and a gamma distributed jump component. A common threshold level was used both for determining the autoregression coefficients and the jump intensities. Rambharat et al. estimated the model by using a Markov chain Monte Carlo approach with 3 years of daily data from Allegheny County, Pennsylvania, and found it superior (both in-sample and out-of-sample) to a jump-diffusion model (i.e. an AR(1)-type process).

These examples show that non-linear regime-switching time series models might provide us with good models of electricity price dynamics. However, is the regime-switching mechanism simply governed by a fundamental variable or the price process itself? Perhaps not. The spot electricity price is the outcome of a vast number of variables including fundamentals (like loads and network constraints) but also the unquantifiable psycho- and sociological factors that can cause an unexpected and irrational buyout of certain commodities or contracts leading to pronounced price spikes. In this context the Markov regime-switching (or simply regime-switching) models, where the regime is determined by an unobservable, latent variable, seem interesting. However, the adequacy of these models for forecasting has been questioned [12][23]. We will leave them out from this study and concentrate only on models with observable threshold variables.

1.4. Data and models

1.4.1. The data

Like in [23] and [36], we forecast hourly CalPX market clearing prices from the period preceding and including the California market crash. This lets us evaluate the performance of the models during normal (calm) weeks, as well as during highly volatile periods. Moreover, the out-of-sample interval spans over half a year and allows for a more thorough analysis of the forecasting results than typically used in the literature single week test samples.

The time series of hourly system prices, system-wide loads, and day-ahead load forecasts was constructed using data obtained from the UCEI institute (www.ucei.berkeley.edu) and the California independent system operator CAISO (oasis.caiso.com). The missing and “doubled” data values corresponding to the changes to and from the daylight saving time (summer time) were treated in the usual way. The former were substituted by the arithmetic average of the two neighboring values, while the latter by the arithmetic average of the two values for the “doubled” hour. Likewise, the few outliers (but not the spikes; spike preprocessing is addressed in Section 1.4.3) were substituted by the arithmetic average of the two neighboring values. The obtained time series are depicted in **Figure 1**. The day-ahead load forecasts (i.e. the official forecasts of the system operator CAISO) are indistinguishable from the actual loads at this resolution; only the latter have been plotted.

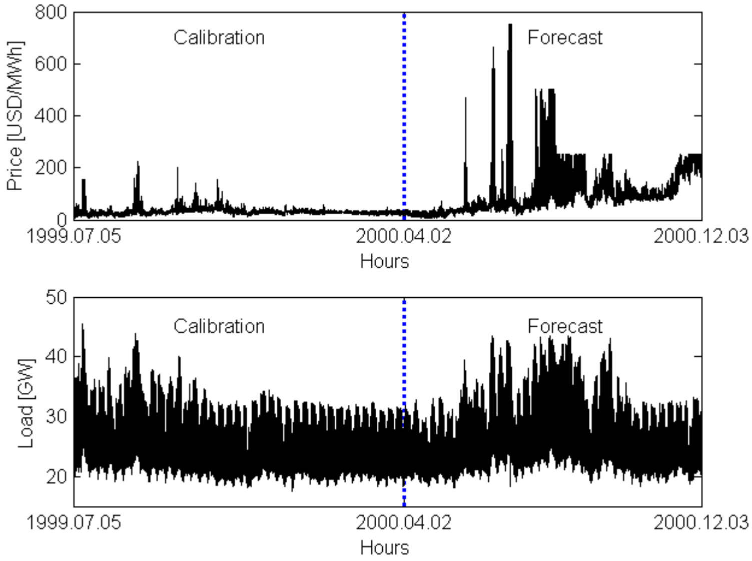

Figure 1. Hourly system prices (*top panel*) and hourly system loads (*bottom panel*) in California for the period July 5, 1999 – December 3, 2000. The changing price cap (750 → 500 → 250 USD/MWh) is clearly visible in the top panel.

We used the data from the period July 5, 1999 – April 2, 2000 solely for the purpose of calibration. Such a relatively long period of data was needed to achieve high accuracy. For example, limiting the calibration period to data coming only from the year 2000, like in [10] and [26], led to a decrease in forecasting performance by up to 70%. Consequently, the period April 3 – December 3, 2000 was used for out-of-sample testing. Since in practice the market-clearing price forecasts for a given day are required on the day before, we used the following testing scheme. To compute price forecasts for hour 1 to 24 of a given day, data available to all procedures included price and demand historical data up to hour 24 of the previous day plus day-ahead load predictions for the 24 hours of that day.

1.4.2. Model specifications

The models considered in this study comprised simple time series specifications with and without exogenous variables, namely ARMAX and ARMA processes, more elaborate autoregression models with GARCH residuals and regime-switching models. The calibration was performed in Matlab (prediction error estimate) and SAS (maximum likelihood and conditional least squares estimates) computing environments.

The logarithmic transformation was applied to price, $p_t = \log(P_t)$, and load, $z_t = \log(Z_t)$, data to attain a more stable variance. Furthermore, the mean (the median for loads) was removed to center the data around 0. Since each hour displays a rather distinct price profile reflecting the daily variation of demand, costs and operational constraints the modeling was implemented separately across the hours, leading to 24 sets of parameters. This approach was also inspired by the extensive research on demand forecasting, which has generally favored the multi-model specification for short-term predictions [4][30].

Seasonal market conditions were captured by the autoregressive structure of the models: the log-price was made dependent on the log-prices for the same hour on the previous days, and the previous weeks, as well as a certain function (e.g. mean, minimum) of all prices on the previous day. The latter created the desired link between bidding and price signals from the entire day.

Since the system load partly explains the price behavior (especially on the daily scale) it was used as the fundamental variable, see **Figure 2**. In the calm period (till mid-May 2000) the dependence between the log-price and the log-system load is almost linear with a slight downward bend for small values of the load. Later that year the prices tend to jump during high load hours, leading to an S-shaped curvilinear dependence. This observation suggests that non-linear regression models should outperform the linear approaches during the spiky periods.

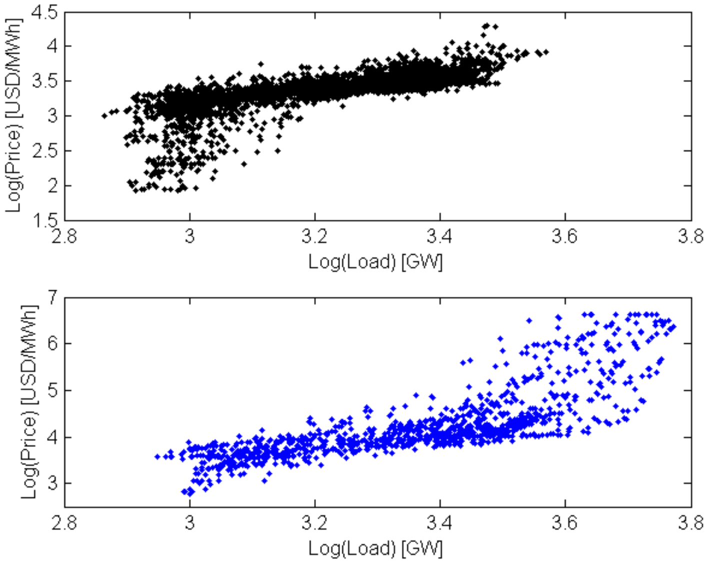

Figure 2. The dependence between hourly log-prices and hourly log system loads in California for the period January 1 – May 21, 2000 (*top panel*) and May 22 – July 2, 2000 (*bottom panel*). The relation changes from a nearly linear (except for a few very low loads) in the calm period to an S-shaped curvilinear dependence during the spiky period.

In our time series models we used only one exogenous variable: the hourly values of the system-wide load. At lag 0 the CAISO day-ahead load forecast for a given hour was used, while for larger lags the actual system load was used. Interestingly, the best models turned out to be the ones with only lag 0 dependence. Using the actual load at lag 0, in general, did not improve the forecasts either. This phenomenon can be explained by the fact that the prices are an outcome of the bids, which in turn are placed with the knowledge of the CAISO load forecasts but not actual future loads.

Furthermore, a large moving average part typically decreased the performance, despite the fact that in many cases it was suggested by Akaike's Final Prediction-Error (FPE) criterion [21]. The best results were obtained for pure ARX models, i.e. with $\theta(B)\varepsilon_t = \varepsilon_t$. Likewise, a large autoregression part generally led to overfitting and worse out-of-sample forecasts. The optimal AR structure was found to be of the form:

$$\phi(B)p_t = p_t - a_1 p_{t-24} - a_2 p_{t-48} - a_3 p_{t-168} - a_4 mp_t, \tag{6}$$

where mp_t was a function of all prices on the previous day. The best results were obtained for the minimum of the 24 hourly prices. Note, that we have simplified the notation: the coefficients are now numbered consecutively and their indices are not directly related to the indices of the corresponding variables as in eqn. (1).

This very simple structure was unable to cope with the weekly seasonality, the results for Mondays, Saturdays, and Sundays were significantly worse than for the other days. Separate modeling of each hour of the week (leading to 168 ARX models) was not satisfactory, probably due to a much smaller calibration set. Incorporation of 7 dummy variables (one for each day of the week) did not improve the results significantly. However, inclusion of 3 dummy variables (for Monday, Saturday and Sunday) helped a lot. The best model structure, in terms of forecasting performance for the first week of the test period (April 3-9, 2000), turned out to be (denoted later in the text as **ARX**):

$$\phi(B) p_t = \psi_1 z_t + d_1 D_{Mon} + d_2 D_{Sat} + d_3 D_{Sun} + \varepsilon_t, \tag{7}$$

where d_1, d_2, d_3 denote the coefficients of the dummies $D_{Mon}, D_{Sat}, D_{Sun}$, respectively. Its simplified version without the exogenous variable (**AR**):

$$\phi(B) p_t = d_1 D_{Mon} + d_2 D_{Sat} + d_3 D_{Sun} + \varepsilon_t, \tag{8}$$

also performed relatively well.

The residuals obtained from the fitted **ARX** and **AR** models seemed to exhibit a non-constant variance. Indeed, when tested with the Lagrange multiplier “ARCH” test statistics [15] the heteroscedastic effects were significant at the 5% level. This motivated us to calibrate **ARX-G** and **AR-G** models to the data (“**G**” stands for **GARCH**). They differ from **ARX** and **AR** models in that the noise terms in eqns. (7) and (8), respectively, are not just $iid\left(0, \sigma^2\right)$ but are given by h_t from eqn. (4) with $p = q = 1$.

Because of the non-linear nature of electricity prices, we also calibrated regime-switching TAR-type models to the spot price time series. They are natural generalizations of the **ARX** and **AR** models defined above. Namely, the **TARX** model is given by:

$$\begin{cases} \phi_1(B) p_t = \psi_{1,1} z_t + d_{1,1} D_{Mon} + d_{1,2} D_{Sat} + d_{1,3} D_{Sun} + \varepsilon_t & \text{when } v_t \geq T, \\ \phi_2(B) p_t = \psi_{2,1} z_t + d_{2,1} D_{Mon} + d_{2,2} D_{Sat} + d_{2,3} D_{Sun} + \varepsilon_t & \text{when } v_t < T, \end{cases} \tag{9}$$

where v_t and T are the threshold variable and the threshold level, respectively. We tried different threshold variables (including various combinations of past prices and loads) and threshold values (constant and variable). The best results – in terms of forecast errors during the first week of the test period – were obtained for v_t equal to the price for hour 24 on the previous day and T estimated for every hour in a multi-step optimization procedure with ten equally spaced starting points spanning the entire parameter space. The simpler **TAR** model was obtained when $\psi_{1,1} = \psi_{2,1} = 0$, i.e. when no exogenous variables were used.

Finally, note that all models were estimated using an adaptive scheme, i.e. instead of using a single model for the whole sample, for every day (and hour) in the test period we calibrated the model (given its structure) to the previous values of prices and loads and obtained a forecasted value for that day (and hour). Originally, at each time step also the model structure was optimized by minimizing the FPE criterion [3] for a given set of model structures. However, this procedure, apart from being time consuming, did not produce satisfactory results. The models were apparently overfitted. Hence, we decided to use only one model structure for all hours and all days.

1.4.3. Spike preprocessing

Price spikes pose a serious problem for linear time series models, which assume stationarity of the signal. Possible solutions involve excluding or limiting price spikes [30][35]. In the first case we treat the abnormal prices as outliers and substitute them with the average of the neighboring observations or with “similar-day” prices. However, price spikes are inherent in electricity prices, so we do not want to delete them completely from the calibration process. Instead of excluding them, we can limit their severity or damp all observations above a certain threshold.

We have evaluated three preprocessing schemes (applied to prices, not log-prices):

- ❑ In the first one we treated the spikes, i.e. prices exceeding a certain threshold T, as outliers and substituted them with “similar-day” (see Section 1.5) prices. Note that we could not substitute them with the average of the neighboring observations since very often consecutive hourly prices exceeded the specified threshold.
- ❑ In the second scheme we set an upper limit on prices – if the price was higher than the specified threshold T, it was set to T.
- ❑ In the third scheme we damped the spikes. Like before we set an upper limit, T, and if the price P_t was higher than T, it was set to $T + T\log_{10}(P_t/T)$. This scheme allowed to differentiate between “regular” and “extreme” spikes.

The optimal (in terms of forecast errors during the first week of the test period) preprocessing scheme turned out to be the latter one with T equal to the mean plus three standard deviations of the calibration sample prices. Spike preprocessing was used only in combination with **ARX** and **AR** models. The resulting models (calibrated to spike-damped data) are denoted by **p-ARX** and **p-AR**, respectively.

1.4.4. Forecast error measures

To assess the prediction performance of the models, different statistical measures can be utilized. The most widely used measures are those based on absolute errors, i.e. absolute values of differences between the actual, P_h, and predicted, $\hat{P}_h$, prices for a given hour, h. The Mean Absolute Error (MAE) is a typical example. For hourly prices it is given by:

$$MAE_{daily} = \frac{1}{24}\sum_{h=1}^{24}\left|P_h - \hat{P}_h\right|. \qquad (10)$$

Sometimes not the absolute, but the relative or percentage difference is more informative. For instance, when comparing results for two distinct data sets. In such cases the Mean Absolute Percentage Error (MAPE) is preferred. For hourly prices the daily MAPE takes the form:

$$MAPE_{daily} = \frac{1}{24}\sum_{h=1}^{24}\frac{\left|P_h - \hat{P}_h\right|}{P_h}. \qquad (11)$$

The MAPE measure works well in load forecasting, since actual load values are rather large. However, when applied to electricity prices, MAPE values could be misleading. In particular, when electricity prices drop to zero, MAPE values become very large regardless of the actual absolute differences $\left|P_h - \hat{P}_h\right|$. The reason for this is the normalization by the current (close to zero, and hence very small) price P_h.

Alternative normalizations have been proposed in the literature [30][35]. For instance, the absolute error $\left|P_h - \hat{P}_h\right|$ can be normalized by the average price attained during the day. The resulting measure, also known as the Mean Daily Error [8][23], is given by:

$$MDE = \frac{1}{24}\sum_{h=1}^{24}\frac{\left|P_h - \hat{P}_h\right|}{\overline{P}_{24}} = \frac{1}{\overline{P}_{24}}MAE_{daily}, \tag{12}$$

where $\overline{P}_{24} = \frac{1}{24}\sum_{h=1}^{24}P_h$. In general, MDE compared to MAPE puts more weight to errors in the high-price range. Analogously to MDE, the Mean Weekly Error can be computed as:

$$MWE = \frac{1}{168}\sum_{h=1}^{168}\frac{\left|P_h - \hat{P}_h\right|}{\overline{P}_{168}} = \frac{1}{\overline{P}_{168}}MAE_{weekly}, \tag{13}$$

where $\overline{P}_{168}$ is the mean price for a given week. Note that instead of the mean, the median price could be used for normalization. As the median is more robust to outliers (or spikes), the resulting measures – Median Daily Error (MeDE) and Median Weekly Error (MeWE) – in some cases exhibit yet better performance. However, they are not as popular as MDE and MWE.

Apart from absolute value-type norms, square-type norms are also often used, even exclusively [20]. Perhaps the most popular are the Daily Root Mean Square Error (DRMSE) and the Weekly Root Mean Square Error (WRMSE), calculated as the square root of the average of 24 and 168, respectively, square differences between the predicted and the actual prices:

$$DRMSE = \sqrt{\frac{1}{24}\sum_{h=1}^{24}\left(P_h - \hat{P}_h\right)^2}, \quad WRMSE = \sqrt{\frac{1}{168}\sum_{h=1}^{168}\left(P_h - \hat{P}_h\right)^2}. \tag{14}$$

Like in the absolute error-based measures, the square differences $\left(P_h - \hat{P}_h\right)^2$ in the above two formulas can be normalized by (the square of) the current actual price, the mean daily (weekly) price or the median daily (weekly) price.

Finally, we have to note that there is no “industry standard” and the error benchmarks used in the literature vary a lot. What is worse, they cause a lot of confusion as the names are not used consistently either. As a result, the forecasts are not comparable from paper to paper even if the same data sets are used. For instance, Nogales et al. [26], Contreras et al. [10] and Garcia et al. [18] defined the “Mean Weekly Error” as the weekly MAPE (literally as the average of the seven daily “average prediction errors”, i.e. daily MAPE values) while Conejo et al. [8] used formula (13). Likewise, in the latter two papers the WRMSE, denoted by $\sqrt{FMSE}$, was computed using formula (14), while in the former two articles the normalization by $\sqrt{1/168}$ was missing.

1.5. Forecasting results

The forecast accuracy was checked afterwards, once the true market prices were available. For all weeks under study, three types of average prediction errors were computed: one corresponding to the 24 hours of each day (MDE) and two to the 168 hours of each week (MWE and WRMSE). Following Conejo et al. [8] and Misiorek et al. [23] a naïve but challenging test was used as a benchmark for all forecasting procedures. The forecasts were compared to the 24 prices of a day similar to the one to be forecast. A “similar day” is characterized as follows. A Monday is similar to the Monday of the previous week and the same rule applies for Saturdays and Sundays; analogously, a Tuesday is similar to the previous Monday, and the same rule applies for Wednesdays, Thursdays, and Fridays. The naïve test is passed if errors for the estimates are smaller than for the prices of the similar day. Quite often the forecasting procedures did not pass this test. See for instance **Table 1**, where Mean Daily Errors for the first week of the test period (April 3-9, 2000) are given; the results for some of the models are also plotted in **Figure 3**. The **p-ARX** model performed best in terms of the MDE criterion: it yielded the smallest errors for Tuesday, Wednesday and Thursday. For all seven days it was better than the naïve forecast; the only other model that passed the naïve test was **AR-GARCH** (surprisingly not **ARX-GARCH**). The **p-AR** model also performed reasonably well with smallest errors for Monday and Sunday. Note also that the performance of the non-linear regime-switching models was only moderate and they produced large errors either for Saturday (**TAR**) or for Sunday (**TARX**). On the weekly scale both criteria (MWE and WRMSE; see **Table 2** and **Table 3**) favored the **p-ARX** model, with **ARX** being second best; both **AR-GARCH** and **p-AR** failed to predict Saturday's prices.

Table 1. Mean Daily Errors (MDE) for the first week of the test period (April 3-9, 2000). Best results are emphasized in bold. Results not passing the naïve test are underlined.

Day	AR	ARX	AR-G	ARX-G	TAR	TARX	p-AR	p-ARX	Naïve
Mon	3,73	3,91	3,32	3,86	3,31	3,60	**3,01**	3,64	5,68
Tue	3,01	2,33	2,35	2,79	3,25	2,64	2,68	**2,21**	3,77
Wed	2,30	2,06	2,05	2,53	2,31	1,93	1,99	**1,89**	2,19
Thu	1,96	1,58	2,10	2,05	2,21	1,79	1,87	**1,49**	2,97
Fri	3,63	2,92	**2,54**	3,48	3,94	3,04	3,41	2,79	2,89
Sat	5,43	**3,96**	7,60	6,86	6,91	4,85	6,15	4,32	8,72
Sun	3,94	4,85	4,17	4,20	4,20	6,14	**3,87**	4,37	10,11

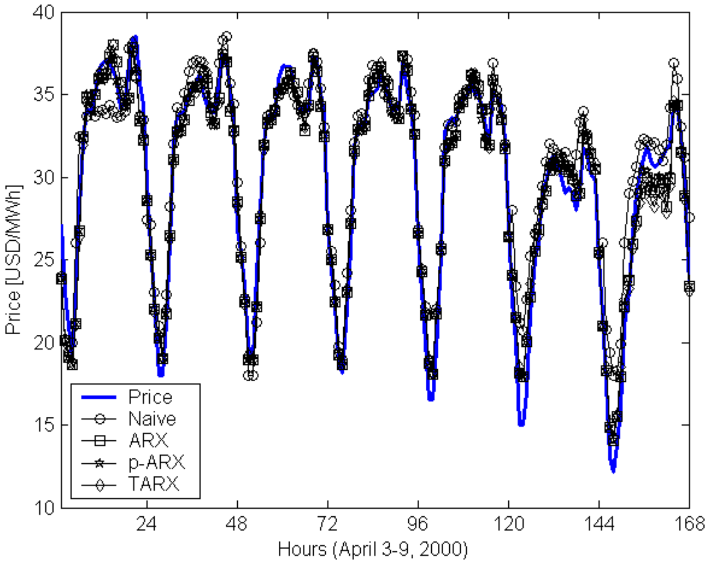

Figure 3. Prediction results for the first week of the test period (April 3-9, 2000) for the naïve, **ARX**, **p-ARX** (preprocessed ARX) and **TARX** models.

Nogales et al. [26] and Contreras et al. [10] fitted and evaluated transfer function (TF), dynamic regression (DR) and ARIMA (also with explanatory variables) models on exactly the same out-of-sample test period. Interestingly, they used single models (though very large) for all 24 hours of a day. Their conclusion was that TF (equivalent to ARMAX with system load as the exogenous variable) was best for the first week of April, followed closely by DR (equivalent to ARX, again with system load as the exogenous variable). The WRMSE's for these two models were 1.04 and 1.05, respectively, which is better than that of our **p-ARX** specification (1.12; see

Table 3). Unfortunately, we were not able to obtain as good results with the ARMAX models we tried. Perhaps, different software implementations of the calibration schemes (Matlab and SAS vs. SCA) prevented us from converging to the same model. Since only the results for the first week of April were reported by Nogales et al., the question whether this common for all hours, multi-parameter TF specification is also superior for other periods (and other data sets) remains open.

Mean Weekly Errors and Weekly Root Mean Square Errors for all 35 weeks of the test period are given in **Table 2** and **Table 3**; see also **Figure 4**. The overall winner was the relatively simple **ARX** model – it yielded the best forecasts for 8 (or 9 in terms of the WRMSE criterion) weeks and was only 6 times worse than the naïve approach. Compared to other models, its performance deteriorated during highly volatile, yet not very spiky periods towards the end of the year. However, during spiky weeks and even spiky price-capped periods it was the best or next to the best model. Not surprisingly, the preprocessing scheme was optimal only for the relatively calm periods. During the first seven weeks **p-ARX** was best 5 times (4 according to the WRMSE criterion) and very close to being the best for the remaining weeks. The simpler **p-AR** model followed closely. However, during spiky and abnormal periods both models had problems with forecasting future prices and failed to pass the naïve test 13 (12 in terms of the WRMSE criterion) times in the whole test sample.

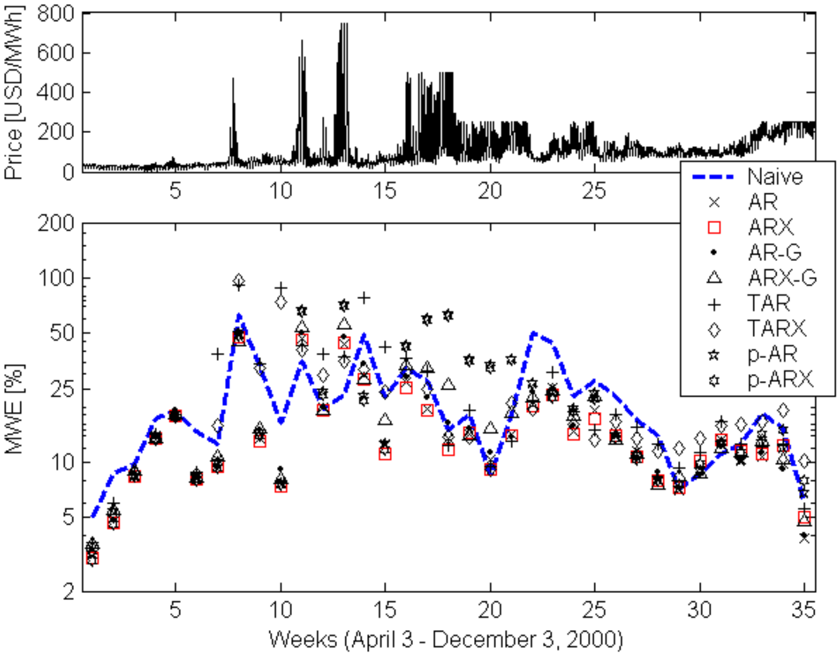

Figure 4. Hourly system prices (*top panel*) and Mean Weekly Errors for all forecasting methods (*bottom panel*; note the logarithmic scale) during the whole test period: April 3 – December 3, 2000.

Table 2. Mean Weekly Errors (MWE) in percent for all weeks of the test period. Best results are emphasized in bold. Results not passing the naïve test are underlined.

Week	AR	ARX	AR-G	ARX-G	TAR	TARX	p-AR	p-ARX	Naïve
1	3,37	3,03	3,34	3,60	3,65	3,33	3,20	**2,90**	5,00
2	5,29	4,71	4,84	5,46	6,05	5,10	5,36	**4,63**	8,62
3	8,41	**8,37**	8,67	8,92	8,62	8,63	8,43	8,39	9,74
4	13,99	13,51	14,10	13,48	14,13	**13,45**	13,92	13,54	17,14
5	18,26	17,82	19,12	18,22	18,16	17,73	18,05	**17,62**	19,31
6	8,40	8,04	8,24	8,26	8,76	8,50	8,31	**7,94**	14,70
7	10,32	9,43	9,32	10,72	38,81	15,64	9,98	**9,25**	12,56
8	50,35	48,15	51,40	**45,55**	91,06	96,92	49,41	48,76	62,97
9	13,44	**13,11**	14,93	15,19	34,23	32,43	14,33	13,95	33,22
10	7,81	**7,39**	9,23	8,10	87,92	73,65	7,71	7,52	16,23
11	46,82	46,23	50,04	53,64	42,75	41,09	67,00	65,48	**35,59**
12	19,77	19,23	19,78	**19,18**	38,52	29,92	24,01	23,19	19,41
13	43,88	44,19	47,90	56,00	37,33	35,69	71,82	70,69	**23,31**
14	29,53	28,01	34,45	28,22	78,21	31,62	22,83	**21,76**	49,47
15	12,61	**11,11**	12,53	16,99	41,99	23,87	12,46	11,66	22,37
16	27,07	**25,46**	29,22	33,45	36,70	30,59	43,02	42,07	32,35
17	19,34	**19,24**	22,61	32,49	31,46	24,79	60,36	59,26	27,74
18	13,58	**11,71**	16,29	26,47	12,34	13,64	63,15	61,78	15,00
19	14,10	14,46	15,15	14,02	19,29	**13,71**	35,75	35,24	18,20
20	10,43	9,18	11,25	15,19	9,55	9,10	33,43	32,52	**8,60**
21	14,13	13,90	13,60	18,51	**12,94**	21,02	35,87	36,13	18,22
22	20,71	20,28	24,26	22,40	21,52	**19,57**	26,70	26,93	50,33
23	25,21	23,28	24,88	24,64	30,96	23,11	23,76	**22,25**	44,17
24	14,80	**14,30**	15,77	17,83	18,92	16,08	19,64	18,96	22,86
25	19,03	17,27	22,60	22,92	14,84	**13,17**	23,49	21,86	27,90
26	14,50	13,98	13,94	**13,30**	18,20	16,54	13,87	13,43	22,99
27	11,57	10,65	10,34	11,13	15,55	13,39	10,94	**10,31**	16,98
28	8,09	7,95	8,76	**7,57**	12,53	11,52	8,15	7,92	13,96
29	**6,97**	7,34	7,22	8,41	9,28	11,88	7,17	7,56	7,11
30	9,24	10,21	**8,48**	8,73	11,34	13,38	9,02	9,90	8,66
31	13,12	13,35	12,19	11,94	16,68	15,90	12,58	12,82	**11,12**
32	10,38	11,41	**10,13**	11,29	12,61	15,92	10,19	11,37	12,62
33	**10,65**	11,07	11,33	12,92	12,10	16,54	12,68	14,23	18,57
34	9,80	12,39	**9,22**	10,30	12,54	19,02	11,94	14,98	15,15
35	**3,87**	5,06	4,00	4,74	5,64	10,20	6,80	7,94	6,09
# best	3	8	3	4	1	4	0	8	4
# better than naïve	29	29	28	25	19	20	22	22	-

Table 3. Weekly Root Mean Square Errors (WRMSE) for all weeks of the test period. Best results are emphasized in bold. Results not passing the naïve test are underlined.

Week	AR	ARX	AR-G	ARX-G	TAR	TARX	p-AR	p-ARX	Naïve
1	1,29	1,17	1,32	1,37	1,41	1,29	1,26	**1,12**	2,06
2	1,76	1,60	1,64	1,78	1,98	1,71	1,76	**1,56**	2,93
3	2,56	**2,51**	2,57	2,65	2,60	2,55	2,56	2,52	3,20
4	4,70	4,51	4,77	**4,46**	4,73	4,60	4,65	4,52	5,59
5	7,46	7,35	7,78	7,54	7,35	7,23	**7,16**	7,18	8,55
6	3,48	3,37	3,50	3,46	3,67	3,58	3,45	**3,34**	6,15
7	4,85	4,60	4,68	5,06	77,62	8,39	4,75	**4,57**	6,41
8	87,89	**85,53**	88,30	88,72	163,08	160,17	98,18	97,24	97,98
9	10,04	**9,78**	10,67	11,27	75,34	44,54	10,79	10,58	30,35
10	5,35	**5,14**	6,33	5,57	141,44	124,08	5,35	5,45	12,95
11	126,97	125,59	133,52	148,33	115,77	106,86	196,75	193,66	**99,88**
12	28,11	26,55	**26,13**	30,60	59,23	35,81	39,02	37,90	27,66
13	154,07	151,05	162,93	196,8	123,69	127,79	275,72	273,27	**93,17**
14	23,42	21,05	26,69	20,84	127,80	38,81	15,44	**14,39**	37,34
15	9,42	**8,58**	9,81	12,77	49,53	22,51	9,57	9,06	18,58
16	68,40	**64,60**	74,87	82,65	72,66	64,78	103,55	101,76	69,83
17	70,53	**68,30**	79,78	103,16	95,87	74,79	176,79	174,52	96,73
18	48,31	**42,45**	55,67	98,19	58,09	51,01	222,49	218,46	61,97
19	27,29	27,03	29,67	26,98	35,26	**25,55**	67,54	66,61	33,73
20	21,96	19,85	22,53	30,32	18,89	17,38	69,89	67,92	**16,70**
21	32,53	32,71	**30,52**	41,70	34,59	47,49	79,22	79,71	45,10
22	34,38	33,39	38,38	34,73	36,30	**27,79**	42,41	42,67	77,40
23	31,83	29,80	32,19	33,04	41,29	29,95	30,41	**29,55**	60,34
24	30,11	**27,96**	31,46	35,30	39,54	30,96	39,45	37,52	41,54
25	34,80	33,92	38,17	37,62	29,34	**25,51**	39,60	36,39	50,21
26	19,88	19,97	19,70	19,88	26,29	23,37	**19,65**	19,91	34,64
27	15,97	14,41	14,30	15,88	23,45	19,80	14,90	**13,90**	25,39
28	9,45	9,28	10,47	**9,16**	15,31	13,29	9,61	9,29	20,11
29	**8,76**	9,28	9,12	10,66	12,52	14,21	8,94	9,52	9,12
30	11,25	12,54	**10,79**	11,31	15,29	16,28	11,07	12,36	11,01
31	16,03	15,90	14,51	14,43	23,02	19,65	15,07	15,19	**13,41**
32	16,14	17,56	16,24	18,24	19,69	23,60	**15,83**	17,59	19,66
33	**25,58**	26,87	27,37	31,01	28,06	39,24	29,81	32,67	42,13
34	27,09	34,27	**25,00**	27,56	35,03	54,56	30,92	39,19	41,09
35	14,82	16,67	**14,81**	15,99	18,96	30,57	19,16	22,05	26,24
# best	2	9	5	2	0	3	3	7	4
# better than naïve	29	29	29	25	19	18	23	23	-

Surprisingly, inclusion of the system load as a fundamental variable was not always optimal (a similar observation was made by Contreras et al. [10] who calibrated (seasonal) ARIMA models to California and Spanish data). While for the first 28 weeks of the test period **ARX** and **p-ARX** were better than or roughly the same as **AR** and **p-AR**, the situation changed in favor of the latter in late 2000 when the minimum daily price increased above 70 USD/MWh. For the relatively calm periods a ca. 10% decrease in MWE was observed, however, during the spiky weeks the improvement was negligible. For the autoregressive models with GARCH noise this effect was even more striking. There was no clear winner among the two considered models, perhaps **AR-GARCH** was even slightly better. In terms of WRMSE (see **Table 3**) it yielded the best predictions for 5 weeks (compared to 2 weeks for **ARX-GARCH**) and was better than the naïve method 29 times (compared to 25).

Despite the heteroscedastic nature of the residuals, addition of a GARCH component in the specification, in general, did not improve the forecasts. **ARX-GARCH** performed considerably worse than **ARX** while **AR-GARCH** was comparable to **AR**, the simplest of all autoregressive models. These results somewhat contradict the reports of Garcia et al. [18] who concluded that (seasonal) ARIMA-GARCH models outperformed simpler (seasonal) ARIMA models fitted to California (!) and Spanish data.

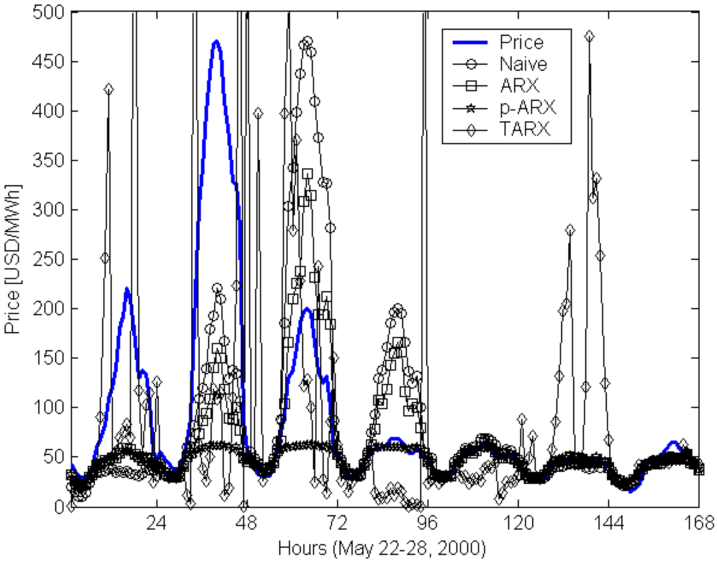

Figure 5. Prediction results for the 8th week of the test period (May 22-28, 2000) for the naïve, **ARX**, **p-ARX** (preprocessed ARX) and **TARX** models. Note that the vertical scale is limited to 500 USD/MWh. For some hours, price forecasts of the **TARX** model (and **TAR** as well) were well above 750 USD/MWh.

Finally, we found the performance of the regime-switching models to be disappointing. Despite the fact that these models are able to cope with the nonlinear nature of the signal, the out-of-sample forecasting results were well below acceptable levels. They were the worst of the studied techniques with over 15 violations (in 35 weeks) of the naïve benchmark! The problems these models had can be very well seen in **Figure 5** where the 8th week of the test period is depicted. This is the first week where extremely high prices (exceeding 450 USD/MWh) were experienced in California. The calibration algorithm classifies some hours as belonging to the spiky regime and most as belonging to the normal regime. This classification, however, has two shortcomings. Firstly, it tends to be “chaotic” – quite often the regime switches from hour to hour leading to a jagged plot. Secondly, since the parameters of the spiky autoregression are estimated from the very few hours that were earlier classified as being in the spike regime, the parameter estimates are highly susceptible to outliers or single extreme observations. Later in the test sample, when the regime switches are more common and the price stays in the spiky regime for longer periods of time, the models (**TARX** in particular) yield much better forecasts and for 3 weeks even outperform the competitors.

Interestingly, if the threshold variable and level are not chosen to minimize MWE for the first week of the test period but the former one is set *a priori* to the difference in mean prices for yesterday and eight days ago and the latter to zero, then the results for the whole test period improve significantly. In fact, such a choice of v_t and T leads to a TARX model that on average beats the ARX model during volatile periods and performs equally well for the calm weeks of the test period [23]. Moreover, as Misiorek and Weron [24] have shown, the TAR/TARX models yield much more accurate interval forecasts than their linear counterparts, especially for the volatile periods.

1.6. Conclusions

In this chapter we investigated the forecasting power of different time series models for electricity spot prices. The models included different specifications of linear autoregressive time series with heteroscedastic noise and/or additional fundamental variables. Further, non-linear regime-switching TAR-type models were considered. The models were tested on a time series of hourly system prices and loads from the California power market. Data from the period July 5, 1999 – April 2, 2000 was used for calibration and from the period April 3 – December 3, 2000 for out-of-sample testing.

Our findings support the adequacy of the tested linear models for forecasting electricity spot prices, also in comparison to earlier empirical studies. The best results in normal (calm) periods were obtained using an ARX model combined with a spike preprocessing scheme that damped prices exceeding a certain threshold with a logarithmic function (**p-ARX**). During the first seven weeks **p-ARX** was best 5 times (4 according to the WRMSE criterion) and very close to being the best for the remaining weeks. However, during volatile periods the preprocessing scheme was not optimal, as obviously it considerably changed input data. In practical applications spike preprocessing should be used, but only as long as the price series is relatively normal and does not exhibit too many (consecutive) spikes. When the nature of the process changes the ARX model becomes the best choice.

Somewhat surprisingly, we found the performance of the non-linear models – autoregressions with GARCH residuals and regime-switching threshold autoregressions – to be disappointing. Despite the heteroscedastic nature of the residuals, addition of a GARCH component in the specification, in general, did not improve the forecasts. **ARX-GARCH** performed considerably worse than **ARX** while **AR-GARCH** was comparable to **AR**. Also the out-of-sample forecasting results for the threshold models were not satisfactory. During normal (calm) periods the regime-switching approach provided only moderate results. Interestingly, during spiky periods TAR-type models performed well below acceptable levels as well. However, as Misiorek et al. [23] have shown, other choices of the threshold variable and level (i.e. ones that do not minimize MWE for the first week of the test period) can lead to a significantly better overall performance.

1.7. References

[1] Bollerslev, T., Generalized autoregressive conditional heteroscedasticity, Journal of Econometrics 31, 1986, 307-327.

[2] Box, G.E.P., Jenkins, G.M., Time Series Analysis: Forecasting and Control, Holden-Day, San Francisco, 1976.

[3] Brockwell, P.J., Davis, R.A., Introduction to Time Series and Forecasting, Springer-Verlag, New York, 1996.

[4] Bunn, D.W., Forecasting loads and prices in competitive power markets, Proceedings of the IEEE 88(2), 2000, 163-169.

[5] Bunn, D.W. (ed.), Modelling Prices in Competitive Electricity Markets, Wiley, 2004

[6] Bunn, D.W., Karakatsani, N., Forecasting electricity prices, EMG Working Paper, London Business School, 2003.

[7] Cabero, J., Baillo, A., Cerisola, S., Ventosa, M., Garcia-Alcalde, A., Peran, F., Relano, G., A medium-term integrated risk management model for a hydrothermal generation company, IEEE Transactions on Power Systems 20(3), 2005, 1379-1388.

[8] Conejo, A.J., Contreras, J., Espinola, R., Plazas, M.A., Forecasting electricity prices for a day-ahead pool-based electric energy market, International Journal of Forecasting 21(3), 2005, 435-462.

[9] Conejo, A.J., Contreras, J., Espinola, R., Plazas, M.A., Day-ahead electricity price forecasting using the wavelet transform and ARIMA models, IEEE Transactions on Power Systems 20(2), 2005, 1035-1042.

[10] Contreras, J., Espinola, R., Nogales, F.J., Conejo, A.J., ARIMA models to predict next-day electricity prices, IEEE Transactions on Power Systems 18(3), 2003, 1014-1020.

[11] Cuaresma, J.C., Hlouskova, J., Kossmeier, S., Obersteiner, M., Forecasting electricity spot prices using linear univariate time-series models, Applied Energy 77, 2004, 87-106.

[12] Dacco R., Satchell C., Why do Regime-Switching Forecast So Badly?, Journal of Forecasting 18, 1999, 1-16.

[13] Day, C.J., Hobbs, B.F., Pang, J.-S., Oligopolistic competition in power networks: A conjectured supply function approach, IEEE Transactions on Power Systems 17(3), 2002, 597-607.

[14] Dueholm, L., Ravn, H.F., Modelling of short term electricity prices, hydro inflow and water values in the Norwegian hydro system, Proceedings of the 6th IAEE European Conference, Zurich, 2004.

[15] Engle, R.F., Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, Econometrica 50, 1982, 987-1007.

[16] Eydeland, A.,Wolyniec, K., Energy and Power Risk Management, Wiley, Hoboken, NJ, 2003.

[17] Franses, P.H., van Dijk, D., Non-Linear Time Series Models in Empirical Finance, Cambridge University Press, Cambridge, 2000.

[18] Garcia, R.C., Contreras, J., van Akkeren, M., Garcia, J.B.C., A GARCH forecasting model to predict day-ahead electricity prices, IEEE Transactions on Power Systems 20(2), 2005, 867-874.

[19] Hinz, J., Modeling day-ahead electricity prices, Applied Mathematical Finance 10(2), 2003, 149-161.

[20] Knittel, C.R., Roberts, M.R., An empirical examination of restructured electricity prices, Energy Economics 27(5), 2005, 791-817.

[21] Ljung, L., System Identification -- Theory for the User, 2nd ed., Prentice Hall, Upper Saddle River, 1999.

[22] Mielczarski, W., Michalik, G., Widjaja, M., Bidding strategies in electricity markets, Proceedings of the 21st IEEE International Conference Power Industry Computer Applications PICA'99, 1999, 71-76.

[23] Misiorek, A., Trueck, S., Weron, R., Point and Interval Forecasting of Spot Electricity Prices: Linear vs. Non-Linear Time Series Models, Studies in Nonlinear Dynamics and Econometrics, 2006, *forthcoming*.

[24] Misiorek, A., Weron, R., Interval forecasting of spot electricity prices, Proceedings of the European Electricity Market EEM-06 Conference, Warszawa, 2006.

[25] Mugele, Ch., Rachev, S.T., Trueck, S., Stable modeling of different European power markets, Investment Management and Financial Innovations 3, 2005.

[26] Nogales, F.J., Contreras, J., Conejo, A.J., Espinola, R., Forecasting next-day electricity prices by time series models, IEEE Transactions on Power Systems 17, 2002, 342-348.

[27] Rambharat, B.R., Brockwell, A.E., Seppi, D.J., A threshold autoregressive model for wholesale electricity prices, Journal of the Royal Statistical Society Series C 54(2), 2005, 287-300.

[28] Robinson, T.A., Electricity pool prices: a case study in nonlinear time-series modelling, Applied Economics 32(5), 2000, 527-532.

[29] Rodriguez, C.P., Anders, G.J., Energy price forecasting in the Ontario competitive power system market, IEEE Transactions on Power Systems 19(1), 2004, 336-374.

[30] Shahidehpour, M., Yamin, H., Li, Z., Market Operations in Electric Power Systems: Forecasting, Scheduling, and Risk Management, Wiley, 2002.

[31] Stevenson, M., Filtering and forecasting spot electricity prices in the increasingly deregulated Australian electricity market, Research Paper No 63, Quantitative Finance Research Centre, University of Technology, Sydney, 2001.

[32] Stoft, S., Power System Economics: Designing Markets for Electricity, Wiley-IEEE Press, 2002.

[33] Vahvilainen, I., Pyykkonen, T., Stochastic factor model for electricity spot price – the case of the Nordic market, Energy Economics 27(2), 2005, 351-367.

[34] Ventosa, M., Baillo, A., Ramos, A., Rivier, M., Electricity market modeling trends, Energy Policy 33(7), 2005, 897-913.

[35] Weron, R., Modeling and forecasting electricity loads and prices: A statistical approach, Wiley, 2006, *forthcoming*.

[36] Weron, R., Misiorek, A., Forecasting spot electricity prices with time series models, Proceedings of the European Electricity Market EEM-05 Conference, Łódź, 2005, 133-141.

[37] Wyłomańska, A., Borgosz-Koczwara, M., The equilibrium models in oligopoly electricity market, Proceedings of the European Electricity Market EEM-04 Conference, Łódź, 2004, 67-75.

[38] Wyłomańska, A., Borgosz-Koczwara, M., Optimal bidding strategies on energy market under imperfect information, Proceedings of the European Electricity Market EEM-05 Conference, Łódź, 2005, 67-73.

[39] Zhang, L., Luh, P.B., Neural network-based market clearing price prediction and confidence interval estimation with an improved extended Kalman filter method, IEEE Transactions on Power Systems 20(1), 2005, 59-66.

HSC Research Report Series 2006

For a complete list please visit http://ideas.repec.org/s/wuu/wpaper.html

01 *Short-term electricity price forecasting with time series models: A review and evaluation* by Rafał Weron and Adam Misiorek