ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/283513131

A survey on data mining techniques applied to energy time series forecasting

ARTICLE *in* ENERGIES · NOVEMBER 2015

Impact Factor: 2.07

READS

60

4 AUTHORS, INCLUDING:

José C Riquelme

Universidad de Sevilla

225 PUBLICATIONS **1,401** CITATIONS

SEE PROFILE

Available from: Francisco Martínez-Álvarez

Retrieved on: 06 April 2016

View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by idUS. Depósito de Investigación Universidad de Sevilla

Submitted to *Energies*. Pages 1 - 41.

OPEN ACCESS

Article

A SURVEY ON DATA MINING TECHNIQUES APPLIED TO ENERGY TIME SERIES FORECASTING

F. Martínez-Álvarez [1,†]*, A. Troncoso [1,†], G. Asencio-Cortés [1] and J. C. Riquelme [2]

[1] Department of Computer Science, Pablo de Olavide University, Seville, Spain
[2] Department of Computer Science, University of Seville, Seville, Spain
[†] These authors contributed equally to this work.

Version November 6, 2015 submitted to Energies. *Typeset by LaTeX using class file mdpi.cls*

Abstract: Data mining has become an essential tool during the last decade to analyze large sets of data. The variety of techniques it includes and the successful results obtained in many application fields, make this family of approaches powerful and widely used. In particular, this work explores the application of these techniques to time series forecasting. Although classical statistical-based methods provides reasonably good results, the result of the application of data mining outperforms those of classical ones. Hence, this work faces two main challenges: i) to provide a compact mathematical formulation of the mainly used techniques, ii) to review the latest works of time series forecasting and, as case study, those related to electricity price and demand markets.

Keywords: Energy; time series; forecasting; data mining

1. Introduction

The prediction of the future has fascinated the human being since its early existence. Actually, many of these efforts can be noticed in everyday events such as energy management [1], telecommunications [2], pollution [3], bioinformatics [4], earthquakes [5], and so forth. Accurate predictions are essential in economical activities as remarkable forecasting errors in certain areas may involve large loss of money.

Given this situation, the successful analysis of temporal data has been a challenging task for many researchers during the last decades and, indeed, it is difficult to figure out any scientific branch with no time-dependant variables.

A thorough review of the existing techniques devoted to forecast time series is provided in this survey. Although a description of classical Box-Jenkins methodology is also discussed, this text is particularly focused on those methodologies that make use of data mining techniques. Moreover, a

family of energy-related time series are examined due to the scientific relevance exhibited during the last decade: electricity price and demand time series. These series have been chosen since they present some peculiarities such as nonconstant mean and variance, high volatility or presence of outliers, that turns the forecasting process into a particularly difficult task to fulfil.

Actually, the electric power markets have become competitive markets due to the deregulation carried out in the last years, allowing the participation of all producers, investors, traders or qualified buyers. Thus, the price of the electricity is determined on the basis of this buying/selling system. Consequently, a will of obtaining optimized bidding strategies has arisen in the electricity-producer companies [6], needing both insight into future electricity prices and assessment of the risk of trusting in predicted prices.

On the other hand, the process of forecasting the quantity of electricity required for a specific geographical area during a time period is called load forecasting or demand forecasting. This process is key since current technology allows to store only little amount of electricity in batteries. Therefore, the demand forecasting plays an important role for electricity power suppliers because both excess and insufficient energy production may lead to large costs and significative reduction of benefits.

Some works have already reviewed electricity price time series forecasting techniques. For instance, [7] collates a massive review of artificial neural networks, but it barely reviews other data mining techniques. Also, Weron [8] presented an excellent review, describing many different approaches for several markets. However, none of them are focused on the whole data mining paradigm. Moreover, they do not provide mathematical foundations for all the methods they evaluated. And this is maybe the most significative strength of the paper, since information relating to underlying mathematics is provided, as well as an exhaustive description of the measures typically used to evaluate the performance. In short, this survey is to provide the reader with a general overview of current data mining techniques used in time series analysis and to highlight all the skills these techniques are exhibiting nowadays. As case study, their application to a real-world energy-related set of series is reported.

As it will be shown in subsequent sections, the majority of the techniques have been applied to Pennsylvania-New Jersey-Maryland (PJM) [9], New York (NYSIO) [10] and Spain (OMEL) [11] electricity markets. This is due to their standard market design structure, which is basically a two-settlement market comprising a day-ahead market and a real-time intraday market. By contrast, both Australian National Electricity Market (ANEM) [12] and Ontario [13] follow a single settlement real-time structure and few researchers have dealt with such markets. ANEM is also well-known for its volatility and its frequent appearance of outliers, turning this market into a perfect target for robust forecasting. Additionally, the Californian electricity market (CAISO) [14] has also been widely analyzed because of the well-known problems that it experienced in the second half of 2000's. Some other markets appear in this work, given the relevance of the model applied. Such are the cases for the UK, India, Malaysia, Finland, Turkey, Egypt, Nord Pool, Brazil, Jordan, China, Taiwan or Greece. Note that most of them provide public access to data.

The remainder of this work is structured as follows. Section 2 provides a formal description of a time series and describes its main features.

Section 3 describes statistical indicators and errors typically used in this field. Also, the concept of persistence model and forecasting skill is here described.

In particular, Section 4 describes the approaches based on linear methods. Classical Box and Jenkins-based methods such as AR, MA, ARMA, ARIMA, ARCH, GARCH or VAR are thus reviewed. Note that from this section on, all sections consist of a brief mathematical description of the technique analyzed and a review of the most representative works.

As for Section 5, it is a compendium of the non-linear forecasting techniques currently in use in the data mining domain. In particular, these methods are divided into global (neural networks, support vector machines, genetic programming) and local (nearest neighbors).

In Section 6, rule-based forecasting methods are analyzed, providing a brief explanation of what a decision rule is, and revisiting the latest and most relevant works in this domain.

The use of wavelets, as relevant method for hybridization, is detailed in Section 7 as well as discussing the most relevant improvements achieved by means of these techniques.

A compilation of several works that cannot be classified in none of the aforementioned groups is described in Section 8. Thus, forecasting approaches based on Markov processes, on Grey models, on Pattern-Sequence similarity or on manifold dimensionality reduction, are there detailed.

Due to the large amount of ensemble models that are being used nowadays, Section 9 is devoted to cover these methods.

Finally, the conclusions drawn from the exploration of all existing techniques are summarized in Section 10.

2. Time series description

This section is to describe temporal data features as well as to provide mathematical description for such a kind of data. Thus, a time series can be understood as a sequence of values observed over time and chronologically ordered. Time is a continuous variable, however, samples are recorded at constant intervals in practice. When the time is considered as a continuous variable, the discipline is commonly referred as *functional data analysis* [15]. The description of this category is out of scope in this survey.

Let y_t, $t = 1, 2, ..., T$ be the historical data of a given time series. This series is thus formed by T samples, where each y_i represents the recorded value of the variable y at time i. Therefore, the forecasting process consists in estimating the value of y_{T+1} ($\hat{y}_{T+1}$) and, the goal, to minimize the error, which is typically represented as a function of $y_{T+1} - \hat{y}_{T+1}$. This estimation can be extended when the horizon of prediction is greater than one, that is, when the objective is to predict a sample at a time $T + h$ ($\hat{y}_{T+h}$). In this situation, the best prediction is reached when a function of $\sum_{i=1}^{h}(y_{T+i} - \hat{y}_{T+i})$ is minimized.

Time series can be graphically represented. In particular, the x–axis identifies the time ($t = 1, 2, ..., T$) whereas the y–axis the values recorded at punctual time stamps (y_t). This representation allows the visual detection of the most highlighting features of a series, such as oscillations amplitude, existing seasons and cycles or the existence of anomalous data or outliers. Figure 1 illustrates, as example, the price evolution for a particular period of 2006 in the Spanish electricity market.

An usual strategy to analyze time series is to decompose them in three main components [16,17]: trend, seasonality and irregular components, also known as residuals.

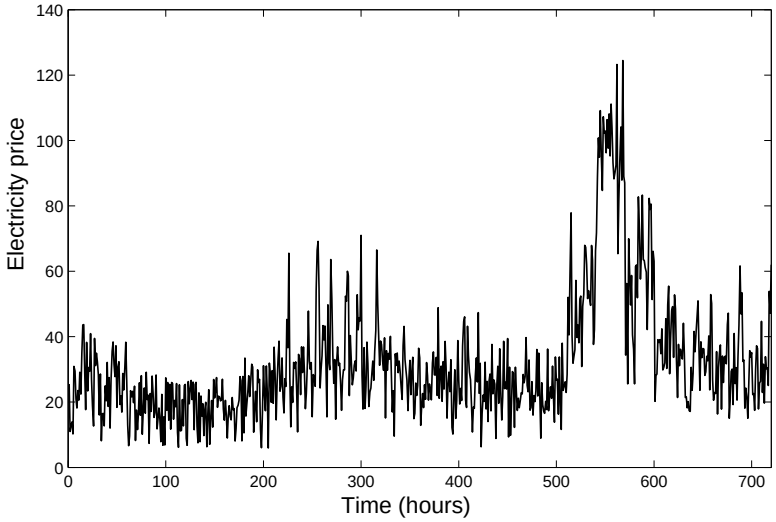

Figure 1. Time series example.

1. **Trend.** It is the general movement that the variable exhibits during the observation period, without considering seasonality and irregulars. Some authors prefer to refer the trend as the long–term movement that a time series shows. Trends can present different profiles such as linear, exponential or parabolic.

2. **Seasonality.** This component typically represents periodical fluctuations of the variable subjected to analysis. It consists of the effects reasonably stable along with the time, magnitude and direction. It can arise from several factors such as weather conditions, economical cycles or holidays.

3. **Residuals.** Once the trend and cyclic oscillations have been calculated and removed, some residual values remain. These values can be, sometimes, high enough to mask the trend and the seasonality. In this case, the term *outlier* is used to refer these residuals, and robust statistics are usually applied to cope with them [18]. These fluctuations can be of diverse origin, which makes the prediction almost impossible. However, if by any chance, this origin can be detected or modeled, they can be thought of precursors in trend changes.

Figure 2 depicts how a time series can be decomposed in the variables above described.

Obviously, real-world time series present a meaningful irregular component, which makes their prediction a especially hard task to fulfil. Some forecasting techniques are focused on detecting trend and seasonality (especially traditional classical methods), however, residuals are the most challenging component to be predicted. The effectiveness of one technique or another is assessed according to its capability of forecasting this particular component. It is for the analysis of this component where data mining-based techniques has been shown to be particularly powerful, as this survey will attempt to show in next sections.

3. Accuracy measures

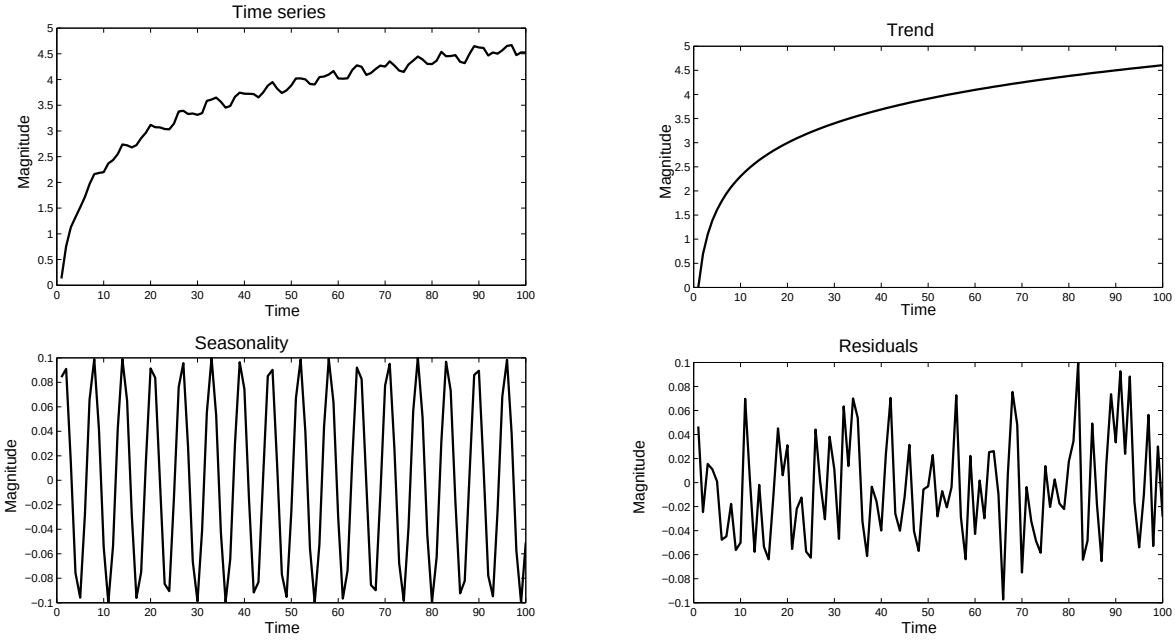

Figure 2. Time series main components decomposition.

The purpose of error measures is to obtain a clear and robust summary of the error distribution. It is common practice to calculate error measures by first calculating a loss function (usually eliminating the sign of the single errors) and then computing an average. Let in the following y_t be the observed value at time t, also called the reference value, and let $\hat{y}_t$ be the forecast for y_t. The error E_t is then computed by $y_t - \hat{y}_t$. Hyndman and Koehler [19] give a detailed review of different accuracy measures used in forecasting and classify the measures into the groups detailed in subsequent sections.

3.1. Scale-dependent measures

There are some commonly used accuracy measures whose scale depends on the scale of the data. These are useful when comparing different methods on the same set of data, but should not be used, for example, when comparing across data sets that have different scales.

The most commonly used scale-dependent measures are based on the absolute error $AE_t = |y_t - \hat{y}_t|$ or squared error $SE_t = (y_t - \hat{y}_t)^2$. These errors are averaged by arithmetic mean or median, leading to the mean absolute error (MAE, Eq. (1)), the median absolute error (MDAE, Eq. (2)), the mean squared error (MSE, Eq. (3)) or the root mean squared error (RMSE, Eq. (4)).

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t| \tag{1}$$

$$MDAE = median(|y_t - \hat{y}_t|) \tag{2}$$

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2 \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \tag{4}$$

When comparing forecast methods on a single data set, the MAE is popular as it is easy to understand and compute. While MAE do not penalize extreme forecast errors, MSE and RMSE emphasize the fact that the total forecast error is in fact much affected by large individual errors, i.e. large errors are much expensive than small errors. Often, the RMSE is preferred to the MSE as it is on the same scale as the data. However, MSE and RMSE are more sensitive to outliers than MAE or MDAE.

3.2. Percentage errors

To address the scale-dependency, the error can be divided by the reference value. Thus, the percentage error (PE) is given by $100(y_t - \hat{y}_t)/(y_t)$. Percentage errors have the advantage of being scale-independent and, therefore, they are frequently used to compare forecast performance across different data sets. The most commonly used measure is the Mean Absolute Percentage Error (MAPE, Eq. (5)).

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|100\frac{y_t - \hat{y}_t}{y_t}\right| \tag{5}$$

These measures have the disadvantage of being infinite or undefined if $y_t = 0$ for any t in the period of interest, and having an extremely skewed distribution when any y_t is close to zero. Where the data involves small counts (which is common with intermittent demand data) it is impossible to use these measures as occurrences of zero values of y_t occur frequently.

By using the median for averaging these problems are easier to deal with, as single infinite or undefined values do not necessarily result in an infinite or undefined measure. However, they also have the disadvantage that they put a heavier penalty on positive errors than on negative errors. This observation led to the use of the so-called symmetric measures sMAPE and sMdAPE, defined in Eq. (6) and Eq. (7).

$$sMAPE = \frac{1}{n}\sum_{t=1}^{n}200\frac{|y_t - \hat{y}_t|}{|y_t| + |\hat{y}_t|} \tag{6}$$

$$sMdAPE = median\left(200\frac{|y_t - \hat{y}_t|}{|y_t| + |\hat{y}_t|}\right) \tag{7}$$

3.3. Relative errors

An alternative way of scaling is to divide each error by the error obtained using another standard method of forecasting as benchmark. Let $r_t = e_t/e_t^*$ denote the relative error where e_t^* is the forecast error obtained from the benchmark method. Usually, the benchmark method is the random walk where $\hat{y}_t$ is equal to the last observation. Then we can define Mean Relative Absolute Error (MRAE, Eq. (8)) and Median Relative Absolute Error (MdRAE, Eq. (9)).

$$MRAE = mean(|r_t|) \tag{8}$$

$$MdRAE = median(|r_t|) \tag{9}$$

A serious deficiency in relative error measures is that e_t^* can be small. In fact, r_t has infinite variance because e_t^* has positive probability density at 0. One common special case is when e_t and e_t^* are normally distributed, in which case r_t has a Cauchy distribution.

3.4. Relative measures

Rather than use relative errors, one can use relative measures. For example, let MAE_b denote the MAE from the benchmark method. Then, a relative MAE is given by:

$$RelMAE = MAE/MAE_b \tag{10}$$

Similar measures can be defined using RMSE, MDAE or MAPE. An advantage of these methods is their interpretability. For example relative MAE measures the possible improvement from the proposed forecast method relative to the benchmark forecast method. When $RelMAE < 1$, the proposed method is better than the benchmark method and when $RelMAE > 1$, the proposed method is worse than the benchmark method.

When the benchmark method is a random walk, and the forecasts are all one-step forecasts, the relative RMSE is the Theil's U statistic, as defined in Eq. (11). The random walk (where $\hat{y}_t$ is equal to the last observation) is the most common benchmark method for such calculations.

$$U = \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}}{\sqrt{\frac{1}{n}\sum_{t=1}^{n} y_t^2}\sqrt{\frac{1}{n}\sum_{t=1}^{n} \hat{y}_t^2}} \tag{11}$$

The Theil's U statistic is a normalized measure of total forecasting error and $0 \leq U \leq 1$. This measure is affected by change of scale and data transformations. For assessing good forecast accuracy, it is desirable that the Theil's U statistic is close to zero. $U = 0$ means a perfect fit.

3.5. Persistence model

The persistence model is an important dynamic property of any time series and usually related to memory properties. Specifically, a time series is a persistent process if the effect of infinitesimally small shock will influence future predictions of the time series for a very long time. Thus the longer the influence time the longer is the persistence.

If a series suffers an external shock, the persistence degree provides information about the impact of the shock on such series, whether it will soon revert to its mean path or it will be further pushed away from the mean path. In case of a highly persistence series, a shock to the series tends to persist for long and the series drifts away from its historical mean path. On the contrary, for the case of a time series with low persistence degree after a shock, the time series tends to get back to its historical mean path.

The persistence of a time series model has been measured by different ways in literature [20].

3.6. Forecasting skill

The forecasting skill is a type of measures that scores the ability of a forecasting method to predict future values of a time series with respect to a reference model as benchmark. The forecasting skill is a scaled representation of the relative forecasting error and its purpose is the same of the relative measures introduced in subsection 3.4.

The most commonly used forecasting skill measure is shown in Eq. (12) and it is based on the previously introduced mean squared error (MSE, see Eq. (3)). MSE is the error of the tested forecasting method and MSE_b is the error of the reference benchmark.

$$SS = 1 - \frac{MSE}{MSE_b} \tag{12}$$

A perfect forecast skill implies $SS = 1$, a forecast with similar skill to the benchmark forecast produces a SS close to 0, and a forecast which is less skillful than the benchmark would produce a negative SS value.

4. Forecasting based on linear methods

There exist real complex phenomena that cannot be represented by means of linear difference equations since they are not fully deterministic. Therefore, it may be desirable to insert a random component in order to allow a higher flexibility on its analysis.

Linear forecasting methods are those that try to model a time series behavior by means of a linear function. From all the existing techniques, seven of them are quite popular: AR, VAR, MA, ARMA, ARIMA, ARCH and GARCH. These models follow a common methodology, whose application to time series analysis was first introduced by Box and Jenkins. The original work has been extended and published many times since its first apparition in 1970, but the newest version can be found in [21].

Autoregressive –$AR(p)$–, moving average –$MA(q)$–, mixed –$ARMA(p,q)$– autoregressive integrated moving average –$ARIMA(p,d,q)$– autoregressive conditional heteroskedastic –$ARCH(q)$– and generalized autoregressive conditional heteroskedastic –$GARCH(p,q)$– models were described following this idea, where p is the number of autoregressive parameters, q is the number of moving average parameters and d is the number of differentiations for the series to be stationary. Vector autoregressive models –$VAR(p)$– are the natural extension for AR models to multivariate time series, where p denotes the number of lags considered in the system.

4.1. Autoregressive processes

An autoregressive process (AR) is denoted by $AR(p)$, where p is the order of the AR process. This process assumes that every y_t can be expressed as a linear combination of some past values. It is a simple model but that adequately describes many real complex phenomena. The generalized AR model of order p is described by:

$$y_t = \sum_{i=1}^{p} \alpha_i y_{t-i} + \epsilon_t \tag{13}$$

where α_i are the coefficients that models the linear combination, ϵ_t the adjustment error, and p the order of the model.

When the error is small compared to the actual values, a future value can be estimated as follows:

$$\begin{aligned} \hat{y}_t &= y_t + \epsilon_t \\ &= \sum_{i=1}^{p} w_i y_{t-i} \end{aligned} \tag{14}$$

4.2. Vector autoregressive models

Vector autoregressive models (VAR) are the natural extension of the univariate AR to multivariate time series. VAR models have shown to be especially useful to describe dynamic behaviors in time series and therefore to forecast. In a VAR process of order p with N variables –$VAR(p)$–, N different equations are estimated. In each equation a regression of the target variable over p lags is carried.

Unlike the univariate case, VAR allow that each series to be related with its own lag and the lag of the other series that form the system. For instance, in two time series systems, there are two equations, one for each variable. This two-series system ($VAR(1)$, $N = 2$) can be mathematically expressed as follows:

$$y_{1,t} = \alpha_{11} y_{1,t-1} + \alpha_{12} y_{2,t-1} + \epsilon_{1,t} \tag{15}$$

$$y_{2,t} = \alpha_{21} y_{1,t-1} + \alpha_{22} y_{2,t-1} + \epsilon_{2,t} \tag{16}$$

$$\tag{17}$$

where $y_{i,t}$ for $i = 1, 2$ are the series to be modeled, and α's the coefficients to be estimated.

Note that the selection of an optimum length of the lag is a critical task for VAR processes and, for this reason, has been widely discussed in literature [22].

4.3. Moving average processes

When the error ϵ_t cannot be assumed as negligible, AR processes are not valid. In this situation it is practical to use the moving average (MA) process, where the series is represented as linear combination of the error values:

$$y_t = \sum_{i=1}^{q} \beta_i \epsilon_{t-i} \tag{18}$$

where q is the order of the MA model and β_i the coefficients of the linear combination. As observed, it is not necessary to make explicit use of past values of y_t to estimate its future value. Finally, MA processes are seldom used alone in practice.

4.4. Autoregressive moving average processes

Autoregressive and moving average models are combined in order to generate better approximations than that of Wold's representation [23]. This hybrid model is called autoregressive moving average process (ARMA) and denoted by $ARMA(p, q)$. Formally:

$$y_t = \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{i=1}^{q} \beta_i \epsilon_{t-i} + \epsilon_t \tag{19}$$

Again, ARMA assumes that ϵ_t is small compared to y_t to estimate future values of y_t. The estimates of ϵ_t past values at time $t - i$ can be obtained from past actual values of y_t and past estimated values of $\hat{y}_t$:

$$\hat{e}_{t-i} = y_{t-i} - \hat{y}_{t-i} \tag{20}$$

Therefore, the estimate for $\hat{y}_t$ is calculated as follows:

$$\hat{y}_t = \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{i=1}^{q} \beta_i \hat{\epsilon}_{t-i} \tag{21}$$

4.5. Generalized Autoregressive Conditional Heteroskedastic processes

Autoregressive conditional heteroskedastic processes (ARCH), firstly presented in [24], or extended ARCH models, called generalized autoregressive conditional heteroskedastic processes (GARCH), introduced in [25], are especially designed to deal with volatile time series, that is, with series that exhibit high volatility and outlying data (for detailed information refer to [26,27]). The ARCH model considers that the conditional variance is dependent of the time, namely, a MA process of order q of the square error values:

$$\sigma(\epsilon_t|\epsilon_{t-1}) = \sum_{i=1}^{q} \beta_i \epsilon_{t-i}^2 \tag{22}$$

The extension of an ARCH model to a GARCH model is similar to the extension of AR models to ARMA models. The conditional variance depends on their own past values in addition to the past values of the square errors:

$$\sigma(\epsilon_t|\epsilon_{t-1}) = \sum_{i=1}^{p} \alpha_i \sigma(\epsilon_{t-i}|\epsilon_{t-i-1}) + \sum_{i=1}^{q} \beta_i \epsilon_{t-i}^2 \tag{23}$$

4.6. Autoregressive integrated moving average processes

Autoregressive integrated moving average processes (ARIMA) are the most general methods and are the result of combining AR and MA processes. ARIMA models are denoted as $ARIMA(p, d, q)$, where p is the number of autoregressive terms, d the number of nonseasonal differences, and q the number of lagged forecast errors in the prediction equation. These models follows a common methodology, whose application to time series analysis was first introduced by Box and Jenkins [21]. Thus, this methodology proposes an iterative process formed by four main steps as illustrated in Figure 3.

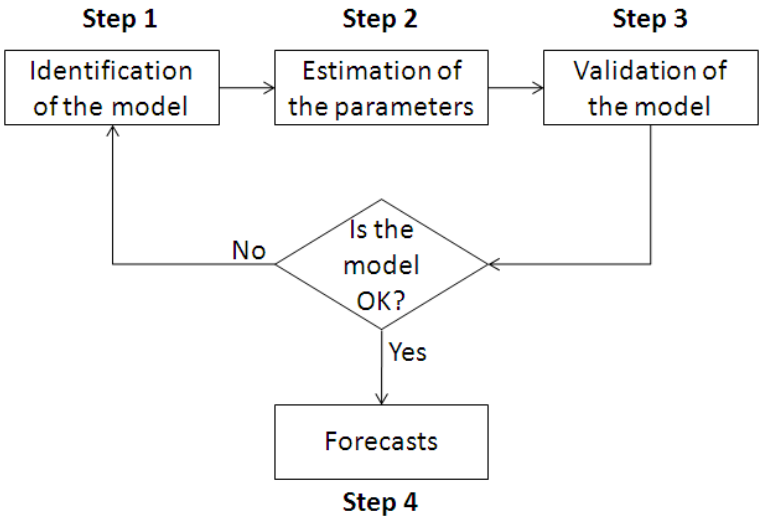

Figure 3. The Box-Jenkins methodology.

1. **Identification of the model.** The first task to be fulfilled is to determine wether the time series is stationary or not, that is, to determine if the mean and variance of a stochastic process do not vary along with time. If the time series does not satisfy this constraint, a transformation has to be applied and the time series has to be differentiated until reaching stationarity. The number of times that the series has to be differentiated is denoted by d and is one of the parameters to be determined in ARIMA models.

2. **Estimation of the parameters.** Once d is determined, the process is reduced to an ARMA model with parameters p and q. These parameters can be estimated by following non-linear strategies. From all of them, three stand out: the evolutionary algorithms, the least squares (LS) minimization and the maximum likelihood (ML). Evolutionary algorithms and LS consist in minimizing the square error of forecasting for a training set while the ML consists in maximizing the likehood function, which is proportional to the probability of obtaining the data given the model.

Comparisons between different Box-Jenkins time series models can be easily found in the literature [28–31], but there are very few works comparing the results of different parameter estimation methods. ML and LS were compared in [32] to obtain an ARIMA model to predict the gold price. The results reported an error of 0.81% and 2.86% when using a LS and a ML, respectively. A comparative analysis between autocorrelation function, conditional likelihood, unconditional likelihood and genetic algorithms in the context of streamflow forecasting was made in [33]. Although similar results were obtained by the four methods, the autocorrelation function and the methods based on ML were the most computationally cost, especially when increased the order of the model. For that, the authors finally recommended the use of evolutionary algorithms.

The good performance of several metaheuristics to solve optimization problems along with the limitations of the classical methods, such as the low precision and poor convergence, has motivated the appearance of recent works comparing evolutionary algorithms and traditional methods for parameter estimation in time series models [34,35]. In general, evolutionary algorithms obtain better results due to the likelihood function is highly nonlinear, and therefore, conventional methods usually converge to a local maxima contrarily to genetic algorithms, which tend to find the global maxima [36].

3. **Validation of the model.** Once the ARIMA model has been estimated several hypotheses have to be validated. Thus, the fitness of the model, the residual values or the significance of the coefficients forming the model are forced to agree with some requirements. In cases in which this step is not fulfilled, the process begins again and the parameters are recalculated.

   In particular, an ARIMA model is validated if estimated residuals behave as white noise, that is, if they exhibit normal distribution as well as constant variance and null mean and covariance. To determine if they are white noise, autocorrelation and partial autocorrelation functions are calculated. These values must be significatively small.

   Additionally, to assess different models' performance, Akaike information criterion (AIC) and Bayesian information criterion (BIC) measures are typically used (instead of classical error measures, such as MAE or RMSE) given their ability to avoid the overfitting that overparameterization causes.

   A problem with the AIC is that it tends to overestimate the number of parameters in the model and this effect can be important in small samples. If AIC and BIC are compared, it can be seen that the BIC penalizes the introduction of new parameters more than the AIC does, hence it tends to choose more parsimonious models [37].

4. **Forecasts.** Finally, if the parameters have been properly determined and validated, the system is ready to perform forecasts.

4.7. Related work

The authors in [38] used the GARCH method to forecast the electricity prices in two regions of New York. The obtained results were compared to different techniques such as dynamic regression (DR),

transfer function models (TFM) and exponential smoothing. They also showed that accounting for the spike values and the heteroscedastic variance in these time series could improve the forecasting, reaching error rates lesser than 2.5%.

García et al. [39] proposed a forecasting technique based on a GARCH model. Hence, this paper focused on day-ahead forecast of electricity prices with high volatility periods. The proposal was tested on both mainland Spanish and California deregulated markets.

Also related with electricity prices time series, the approach proposed by Malo et al. in [40] was equally noticeable. In it, the authors considered a variety of specification tests for multivariate GARCH models that were used in dynamic hedging in the Nordic electricity markets. Moreover, hedging performance comparison were conducted in terms of unconditional and conditional ex-post variance.

An application of ARMA models to electricity prices can be found in [41], where the exogenous variable is the electricity demand. The study was carried out with data of California. The average error verges on 10%.

In [42] ARIMA models, selected by means of Bayesian Information Criteria, were proposed to obtain the forecasts of electricity prices in the Spanish market. In addition, the work analyzed the optimal number of samples used to build the prediction models.

Weron et al. [43] presented twelve parametric and semi-parametric time series models to predict electricity prices for the next day. Moreover, in this work forecasting intervals were provided and evaluated taking into account the conditional and unconditional coverage. They concluded that the intervals obtained by semi-parametric models are better than that of parametric models.

Table 1 summarizes the content of this section. Note that *5+ models* means that the approach has been compared to five or more models. As it can be appreciated, linear methods were very popular at the beginning of 2000's as main methods to make predictions. However, nowadays, these kind of methods have turned into baselines for other methods to be compared to.

Table 1. Summary on linear methods.

Reference	Technique	Outperforms	Metrics	Horizon	Year	Market
[38]	GARCH	DR/TFM/Smoothing	RMSE/MAPE	1 day	2002	NYISO
[39]	GARCH	ARIMA	RMSE	1 day	2000	CAISO/OMEL
[40]	GARCH	5+ models	MAPE/MAE	1 day	2004	Northern Europe
[41]	ARMA	5+ models	RMSE	1 day	2000	CAISO
[42]	Mixed ARIMA	ARIMA	RMSE/MAPE	1 day	2000-2002	OMEL
[43]	ARIMA	5+ models	MAE/MAPE	1 day	2004	CAISO/Nord Pool

5. Forecasting based on non-linear methods

Non linear forecasting methods are those that try to model a time series behavior by means of a non linear function. This function is often generated by lineally combining non-linear functions whose parameters have to be determined. Moreover, the non linear methods can be classified in global or local methods depending on the characteristics required for the function to find.

5.1. Global methods

On the other hand, global methods are based on finding a linear function able to model the output data from the input ones. Several techniques form this family of methods, among which the most important are: artificial neural networks, whose main advantage is that they do not need to know the input data distribution; the support-vector machines, which are very powerful classifiers that follow a philosophy similar to that of the artificial neural networks; and genetic programming, where the type of non-linear function that models the data behavior can be selected.

5.1.1. Artificial neural networks

This section is devoted to artificial neural networks (ANN) which have widely applied for forecasting energy time series. In particular, a general description is presented in Section 5.1.1.1 and two specific ANN, namely extreme learning machine (ELM) and self-organizing Kohonen's maps (SOM) are introduced in Sections 5.1.1.2 and 5.1.1.3, respectively. Finally, Section 5.1.1.4 presents a review of recently published literature related to ANN.

5.1.1.1. Fundamentals

ANNs were originally conceived by McCulloch and Pitts in [44]. These mechanisms search for solving problems by using systems inspired in the human brain and not by applying step by step as usually happens in most techniques. Therefore, these systems own a certain intelligence resulting from the combination of simple interconnected units –neurons– that work in parallel in order to solve several tasks, such as prediction, optimization, pattern recognition or control.

Neural networks are inspired in the structure and running of nervous systems, in which the neuron is the key element due to its communication ability. The existing analogies between ANN and the synaptic activity are now explained. Signals that arrive to the synapse are the neuron's inputs and can be whether attenuated or amplified by means of an associated *weight.* These input signals can excite the neuron if a positive weighted synapsis is carried out or, on the contrary, they can inhibit it if the weight is negative. Finally, if the sum of the weighted inputs is equal or greater than a certain threshold, the neuron is activated. Neurons present, consequently, binary results: activation or not activation. Figure 4 illustrates an usual structure of an ANN.

There are three main features that characterize a neural network: topology, learning paradigm and the representation of the information. A brief description of them are now provided.

1. **Topology of the ANN.** Neural networks architecture consists in the organization and position of the neurons with regard to the input or output of the network. In this sense, the fundamental parameters of the network are the number of layers, the number of neurons per layer, the connection grade and the type of connections among neurons. With reference to the *number of layers,* ANN can be classified into monolayer or multilayer networks (MLP). The first ones only have one input layer and one output layer, whereas the multilayer networks [45] are a generalization of the monolayer ones, which add intermediate or hidden layers between the input and the output. When discussing about the *connection type,* the ANN can be feedforward if the signal propagation is produced in just one way and, therefore, they do not have a memory or

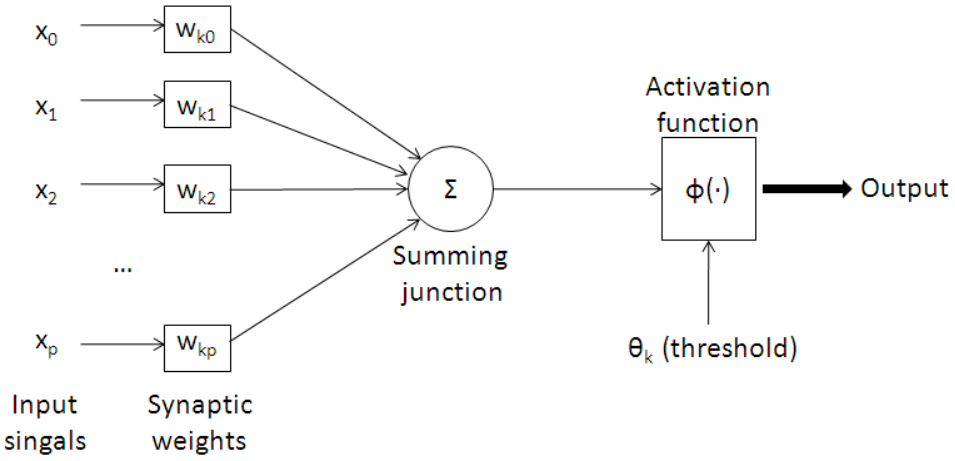

Figure 4. Mathematical model of an ANN.

recurrent if they keep feedback links between neurons in different layers, neurons in the same layer or in the same neuron. Finally, the *connection grade* can be totally connected if all neurons in a layer are connected with the neurons in the next layer (feedforward networks) or with the neurons in the last layer (recurrent networks) and, otherwise, partially connected networks in cases where there is not total connection among neurons from different layers.

2. **Learning paradigm.** The learning is a process that consists in modifying the weights of the ANN, according to the input information. The changes that can be carried out during the learning process are removing (the weight is set to zero), adding (conversion of a weight equal to zero to a weight different to zero) or modifying neurons connections. The learning process is said to be finished or, in other words, the network has learnt when the values assigned to the weights remain unchanged.

3. **Representation of the input/output information.** ANN can be also classified according to the way in which information relative to both input and output data is represented. Thus, in a great number of networks input and output data are analog which entails activation functions also analogs, either linear or sigmoidal. In contrast, there are some networks that only allow discrete or even binary values as input data. In this situation, the neurons are activated by means of an echelon function. Finally, hybrid ANNs can be found in which input data may accept continuous values and output data would provide discrete values or viceversa.

5.1.1.2. Extreme Learning Machine

Extreme Learning Machine (ELM) [46] is a feedforward neural network with an only hidden layer that uses a method for the training faster than the classical ANNs. Namely, the ELM randomly generates the weights W^1 that connect the input layer with the hidden layer and computes the weights W^2 that connect the hidden layer with the output using a simple matrix computation. Thus, the output y is defined by the following model:

$$y = W^2 \phi(W^1 x) \tag{24}$$

where ϕ is the activation function and x is the input vector.

The training consists in computing the weights W^2 as follows:

$$H = \phi(W^1 x_i) \quad (25)$$

$$W^2 = H^+ y_i \quad (26)$$

where (x_i, y_i) are the points of the training set and H^+ represents the pseudoinverse of the matrix H.

5.1.1.3. Self Organizing Maps

The learning in ANN can be either supervised (perceptron and backpropagation [47] techniques) or unsupervised, from which the self-organizing Kohonen's maps (SOM) [48] stands out.

SOM have been mainly applied to discover patterns in data. The learning paradigm is based on a competitive learning, that is the neurons compete among them and win the neuron with the nearest weights to the input vector. Then, all neurons near to the win neuron update their weights according to a specific rule defined by:

$$w^j_{n+1} = w^j_n + \mu_n(x - w^j_n) \quad (27)$$

where w^j_n is the weight associated to the neuron j at the n-th iteration, μ_n is the learning factor and x is the input vector.

The neurons that are not neighbors to the win neuron do not update their weights. Finally, a clustering of the data is obtained when the training phase ends.

5.1.1.4. Related work

Many references proposing the use of ANNs, or a variation of them, as a powerful tool to forecast time series, can be found in the literature. The most important works are detailed below. Furthermore, the creation of hybrid methods that highlight most of the strengths of each technique is currently the most popular work among the researchers. However, from all of them, the combination of ANN and fuzzy set theory has become a new tool to be explored.

Rodríguez and Anders [49] presented a method to predict electricity prices by means of an ANN and fuzzy logic, as well as a combination of both. The basic selected network configuration consisted of a back propagation neural network with one hidden layer that used a sigmoid transfer function and a one-neuron output layer with a linear transfer function. They also reported the results of applying different regression-based techniques over the Ontario market.

A hybrid model which used ANNs and fuzzy logic was introduced in [50]. As regards the neural network presented, it had a feed-forward architecture and three layers, where the hidden nodes of the proposed fuzzy neural network performed the fuzzyfication process. The approach was tested over the Spanish electricity price market and showed to be better than many other techniques such as ARIMA or MLP.

Taylor et al. [51] compared six univariate time series methods to forecast electricity load for Rio de Janeiro and England and Wales markets. These methods were an ARIMA model and an exponential smoothing (both for double seasonality), an artificial neural network, a regression model with a previous principal component analysis and two naive approaches as reference methods. The best method was the

proposed exponential smoothing and the regression model showed a good performance for the England and Wales demand.

Another neural network-based approach was introduced in [52] in which multiple combinations were considered. These combinations consisted of networks with different number of hidden layers, different number of units in each layer and several types of transfer functions. The authors evaluated the accuracy of the approach reporting the results from the electricity markets of mainland Spain and California.

The use of ANN for forecasting electricity prices in the Spanish market was also proposed in [53]. The main novelty of this work lies on the proposed training method for ANN, which is based on making a previous selection for the MLP training samples, using an ART-type [54] neural network.

In [55], the authors discussed and presented results by using an ANN to forecast the Jordanian electricity demand, which is trained by a particle swarm optimization technique. They also showed the performance obtained by using a back propagation algorithm (BP) and autoregressive moving average models.

Neupane et al. [56] used an ANN model with carefully selected inputs. Such inputs were selected by means of a wrapper method for feature selection. The proposal was applied to data from Australia, New York and Spain electricity markets, outperforming the PSF algorithm performance.

The feature selection problem to obtain optimal inputs for load forecasting has also been addressed by means of ANN [57]. The authors evaluated the performance of four feature selection methods in conjunction with state-of-the-art prediction algorithms, using two years of Australian data. The results outperformed those of exponential smoothing prediction models.

In spite of the widespread use of the ANNs, the ELM has not been too explored to predict energy time series. An ELM and bootstrapping to predict probabilistic intervals for Australian electricity market was proposed in [58]. First, an ELM was applied to obtain point forecasts, and later, a bootstrap method was used for uncertainty estimations. The results were compared with two ANNs, namely a back-propagation ANN and a radial basis function neural network, showing that ELM outperforms other methods in most of the test sets. For the same market, prediction intervals (PI) were also obtained in [59]. In this case, a maximum likelihood method was used to estimate the noise variance indeed of a bootstrap method. The results were compared to a random walk (RW), and both traditional ANN and ELM with a bootstrap method. The proposed method provided the best training time and errors.

In [60] five recent methods to train radial-basis function (RBF) networks were applied to obtain the short-term load forecasting in New England. These method were SVR, ELM, decay RBF neural networks, improved second order and error correction. The best results regarding the training, errors, network size, and computational time were obtained with the error correction.

Li et al [61] presented a wavelet transform to deal with the nonstationary of the load time series and an ELM with weights initially computed by an artificial bee colony algorithm to predict the load time series in New England and North American from the wavelet series. The authors showed that the use of an optimization algorithm to set the weights in ELM improves the forecasting errors.

Most approaches based on SOMs published in the literature for forecasting tasks, use the SOM to group the data in an initial stage, and later obtain a prediction model for each group. In [62] the authors propose to combine SOM and support vector machines to predict hourly electricity prices for next-day. First, they applied a SOM to split the data into groups, and then, a support vector machine model for each

group is used to obtain the prediction of the prices in the New England electricity market. In this work, two months were used to validate the method, which provided errors of 7% approximately. Likewise, a SOM along with an ANN was applied to forecast the prices for Australian and New York electricity markets [63]. In this case, the ANN predicted the nearest cluster and the prediction was obtained by the centroid of the cluster. The errors reported for the year 2006 were around a 1.76% and 2.88% for Australian and New York markets, respectively. A SOM without combining with another technique was presented in [64] to predict the prices for the Spanish electricity market. A preprocessing to select the input variables was proposed as a previous step to the prediction, which was obtained from the prices of the nearest centroid to the input data. The proposed SOM obtained forecasts with an error of 2.32% for the daily market.

Table 2 summarizes the content of this section.

Table 2. Summary on ANN, SOM and ELM for electricity forecasting.

Reference	Technique	Outperforms	Metrics	Horizon	Year	Market
[49]	Hybrid ANN	5+ models	MAPE	1 day	2002	Ontario
[50]	Hybrid ANN	MLP/ARIMA/RBF	MRE	1 day	2002	OMEL
[51]	ANN	5+ models	RMSE/MAE	1 day	2003	Brazil
[52]	ANN	ARIMA/Naive	MAPE	1 day	2000/2002	CAISO/OMEL
[53]	ART-NN	ARIMA/ANN	MAPE	1 day	2003	OMEL
[55]	ANN	ARMA/BP	RMSE/MAPE	1 day	2004	Jordan
[56]	ANN	PSF	MRE/MAPE	1 day	2006	NYISO/ANEM/OMEL
[57]	ANN	Smoothing	MAE/MAPE	1 day	2007	ANEM
[58]	ELM	5+ models	MAE/MAPE/RMSE	1 day	2006/07	ANEM
[59]	ELM	RW/ANN	PI	1 day	2007/09	ANEM
[60]	ELM	RBF/SVR	MAPE	1 day	2011	ANEM
[61]	ELM	5+ models	MAPE	1 day	2006	NYISO/ANEM
[62]	SOM	SVM	MAE/MAPE	1 day	2005	ANEM
[63]	SOM	PSF	MRE/MAPE	1 day	2006	NYISO/ANEM/OMEL
[64]	SOM	5+ models	MAPE	1 day	2011	OMEL

5.1.2. Genetic programming

5.1.2.1. Fundamentals

A genetic algorithm (GA) [65] is a kind of searching stochastic algorithm based on natural selecting procedures. Such algorithms try to imitate the biological evolutive process since they combine the survival of the best individuals in a set, by means of an structured and random process of information exchange.

Every time the process iterates, a new set of data structures is generated gathering just the best individuals of older generations. Thus, the GA are evolutionary algorithms due to their capacity to efficiently exploit the information relating to past generations. This fact allows the speculation about new searching points in the solution space, trying to obtain better models thanks to its evolution.

Many genetic operators can be defined. However, selection, crossover and mutation are the most relevant and used and are now going to be briefly described.

1. *Selection.* During each successive generation, a proportion of the existing population is selected to breed a new generation. Individual solutions are selected through a fitness-based process, where fitter solutions (as measured by a fitness function) are typically more likely to be selected. Certain selection methods rate the fitness of each solution and preferentially select the best solutions. Other methods rate only a random sample of the population, as this process may be very time-consuming. Most functions are stochastic and designed so that a small proportion of less fit solutions are selected. This helps keep the diversity of the population large, preventing premature convergence on poor solutions. Popular and well-studied selection methods include roulette wheel selection and tournament selection.

2. *Crossover.* Just after two parents are selected by any selection method, crossover takes place. Crossover is an operator that mates these two parents to produce offspring. The newborn individuals may be better than their parents and the evolution process may continue. In most crossover operators, two individuals are randomly selected and recombined with a crossover probability, p_c. That is, an uniform number r is generated and if $r \leq p_c$ the two randomly selected individuals undergo recombination. Otherwise, the offspring can be sheer copies of their parents. The value of p_c can either be set experimentally or set based on schema-theorem principles [65].

3. *Mutation.* Mutation is the genetic operator that randomly changes one or more of the individuals' genes. The purpose of the mutation operator is to prevent the genetic population from converging to a local minimum and to introduce to the population new possible solutions.

Genetic programming (GP) is a natural evolution of GA and its first apparition in the literature dates of 1992 [66]. It is a specialization of genetic algorithms where each individual is a computer program. Therefore it is used to optimize a population of computer programs according to a fitness landscape determined by a program's ability to perform a given computational task. Hence, specialized genetic operator that generalize crossover and mutation are used for tree-structured programs.

The main steps to be followed when using GP are now summarized. Obviously, depending on the type of the application, these steps may change in order to be adapted to the particular problem to be dealt with.

1. Random generation of an initial population, that is, programs.

2. Iterative execution until the stop condition –to be determined in each situation– is fulfilled:

   (a) To execute each program of the population and to assign an aptitude value, according to their behavior in relation with the problem.

   (b) To create new programs by applying different primary operations to the programs.

      i. To copy an existing program in the new generation.

      ii. To create two programs from two existing ones, genetically and randomly recombining some chosen parts of both programs, making use of the crossover operator, which will also be randomly chosen for each program.

      iii. To create a program from another randomly chosen by randomly changing a gene.

3. The program identified as possessing the best aptitude (the best for the last generation) is the designed result of the GP running.

5.1.2.2. Related work

The viability of forecasting the electricity demand via linear GP is analyzed in [67]. Hence, the authors considered load demand patterns for ten consecutive months, observed every thirty minutes for the Victoria State of Australia. The performance was compared with an ANN and a neuro-fuzzy system (EFuNN) and the system delivered best results in terms of accuracy and computational cost.

An evolutionary technique applied to the optimal short-term scheduling of the electric energy production was presented in [68]. The equations that define the problem led to a nonlinear mixed-integer programming problem with a high number of real and integer variables. The required heuristics, introduced to assure the feasibility of the constraints, are analyzed, along with a brief description of the proposed GA. Results from the Spanish power system were reported and compared to dynamic regression (DR).

Another price forecasting strategy was proposed in [69]. In fact the authors presented a mutual information-based feature selection technique (MI) in which the prediction part was a cascaded neuro-evolutionary algorithm. The accuracy was largely evaluated since they compared their results –obtained from Pennsylvania-New Jersey-Maryland and Spanish electricity markets– with seven different models.

The electricity energy consumption is forecasted by using genetic algorithms in Turkey [70]. The results were compared with conventional regression techniques, and the estimated values of the Turkish Ministry of Energy and Natural Resources (TMENR). An estimation for the electricity demand in the year 2020 is also provided.

A variant of genetic programming, Multi-Gene Genetic Programming (MGGP), was introduced in [71] and applied to Egypt load forecasting. The method was compared with RBF network and the standard genetic programming.

A variant of genetic programming, improved by incorporating semantic awareness in algorithm, for short term load forecasting is described in [72]. The authors analyzed South Italy data and outperformed standard GP and some other machine learning methods.

Finally, Table 3 summarizes all the methods reviewed in this section.

Table 3. Summary on GP for electricity forecasting.

Reference	Technique	Outperforms	Metrics	Horizon	Year	Market
[67]	Linear GP	ANN/EFuNN	RMSE	2 days	1995	ANEM
[68]	GP	DR	MRE/MAE	1 day	2002	OMEL
[69]	MI GP	5+ models	MAE/MSRE	1 day	2007	PJM/OMEL
[70]	GP	TMENR	MSE	1 day	2020	Turkey
[71]	MGGP	RBF/GP	MAPE	1 day	2012	Egypt
[72]	Semantic GP	5+ models	MAE/MSRE	1 day	2009/10	Italy

572 5.1.3. Support vector machines

573 5.1.3.1. Fundamentals

574 The support vector machine (SVM) model the way is nowadays understood, initially appeared in
575 1992 in the Computational Learning Theory (COLT) Conference and it has been subsequently studied
576 and extended [73,74]. The interest for this learning model is continuously increasing and it is considered
577 an emerging and successful technique nowadays. Thus, it has become to a widely accepted standard in
578 machine learning and data mining disciplines.

579 The learning process in SVM represents an optimization problem under constraints that can be solved
580 by means of quadratic programming. The convexity guarantees a single solution which is an advantage
581 with regard to the classical model of ANN. Furthermore, current implementations provide moderate
582 efficiency for real-world problems with thousands of samples and attributes.

Support vector machines aims at separating points by means of what they defined as hyperplane, which are just linear separators with a high dimensionality whose functions are defined according to different kernels. Formally, a hyperplane in a D-dimensional space is defined as follows:

$$h(x) =< w, x > +b \tag{28}$$

583 where x is the sample, $w \epsilon \mathbf{R}^D$ is the orthogonal vector to the hyperplane, $b \epsilon \mathbf{R}$, w is the weight vector, b
584 is the bias or threshold decision and $< w, x >$ expresses the scalar product in $\mathbf{R}^D$.

In case of a binary classifier is required, the equation can be reformulated as:

$$f(x) = sign\left(h(x)\right) \tag{29}$$

where the *sign* function is defined as:

$$sign(x) = \begin{cases} +1, & if \quad x \geq 0 \\ -1, & if \quad x < 0 \end{cases} \tag{30}$$

585 There exist many algorithms directed to create hyperplanes (w, b) given a dataset linearly separable.
586 These algorithms guarantee the convergency to a solution hyperplane although particularities of all of
587 them will lead to slightly different solutions. Note that there can be infinity hyperplanes that perform
588 adequate separations. So the key problem for the SVM is to choose the best hyperplane, in other words,
589 the hyperplane that maximizes the minimum distance (or geometric margin) between the samples in the
590 dataset and the hyperplane itself.

591 Another peculiarity of SVM is that only take into consideration those points belonging to the frontiers
592 of the region of decision, which are the points that do not clearly belong to a class or to another. Such
593 points are named support vectors. Figure 5 illustrates a bidimensional representation of an hyperplane
594 equidistant to two classes, as well as showing the support vectors and the existing margin.

595 If non linear transformation is carried out from the input space to the feature space, non linear
596 separators-based learning is reached with SVM. Kernel functions are used thus in order to estimate
597 the scalar product of two vectors in the features space. Consequently the election of an adequate kernel
598 function is crucial and a priori knowledge of problem is required for a proper application of SVM.
599 Nevertheless, the samples may not be linearly separable (see Figure 6) even in the features space.

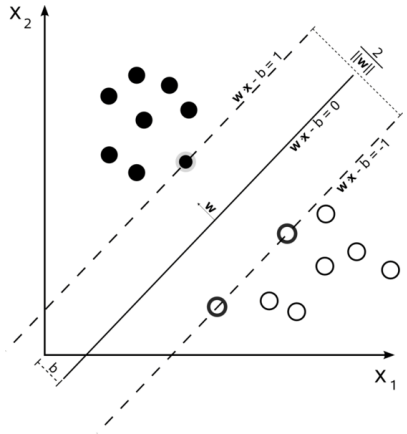

Figure 5. Hyperplane (w, b) equidistant to two classes, margin and support vectors.

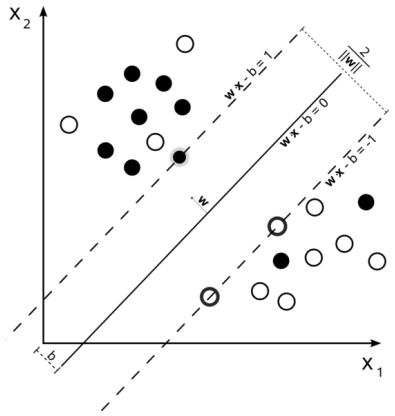

Figure 6. Non linearly separable dataset.

Trying to classify properly all the samples can seriously compromise the generalization of the classifier. This problem is known as *overfitting*. In such situations it is desirable to admit that some samples will be misclassified in exchange for having more promising and general separators. This behavior is reached by inserting soft margin in the model, whose objective function is composed by the addition of two terms: the geometric margin and the regularization term. The importance of both terms is pondered by means of a typically called parameter C. This model appeared in 1999 [75], and it was the model that really allowed the practical use that SVMs have nowadays, since it provided robustness against the noise.

On the other hand, SVMs can be easily adapted to solve regression problems by means of the introduction of a loss function. SVMs are commonly called Support Vector Regression (SVR) for time series forecasting. Now, the problem consists in finding a non linear function f that minimizes the forecasting error for the training set. The ϵ-insensitive loss function L_ϵ defined by Eq. (31) is typically used due to a reduced number of support vectors is obtained. The ϵ parameter represents the error allowed for each point of the training set.

$$L_\epsilon(y) = \begin{cases} 0 & if \quad |y - f(x)| \leq \epsilon \\ |y - f(x)| - \epsilon & otherwise \end{cases} \tag{31}$$

To approximate all data of the training set with an error less than ϵ is not always possible in practice. For this reason, slack variables ξ_i and ξ_i^* are inserted to allow errors greater. Thus, the SVR model consists in solving the following problem:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}||w||^2 + C\sum_i(\xi_i + \xi_i^*) \\ \text{subject to} \quad & y_i - f(x_i) \leq \epsilon + \xi_i \\ & f(x_i) - y_i \leq \epsilon + \xi_i^* \end{aligned} \tag{32}$$

where (x_i, y_i) are the points of the training set, w is the margin and C is the regularization parameter.

Once the optimization problem has been solved, the following function is obtained:

$$f(x) = \sum_{i=1}^{n}(\alpha_i^+ - \alpha_i^-)K(x, x_i) \tag{33}$$

where α_i^+ and α_i^- are the multipliers of Lagrange of the dual optimization problem and K is the kernel function.

5.1.3.2. Related work

Many works have been focussed on forecasting time series by applying SVM. Hence, the study carried out in [76] analyzed the suitability of applying SVM to forecast the electric load for the Taiwanese market. The results were compared to that of linear regressions and ANN. The same time series type, but related to the Chinese market, was forecasted in [77], in which the authors reached a globally optimized prediction by applying a SVM.

The occurrence of outliers (also called spike prices) or prices significantly larger than the expected values is an usual feature found in these time series. With the aim of dealing with this feature, the authors in [78] proposed a data mining framework based on both SVM and probability classifiers.

The research published in [79] proposed a new prediction approach based on SVM and rough sets techniques (RS) with a previous selection of features from data sets by using an evolutionary method. The approach improved the forecasting quality, reduced the speed of convergence and the computational cost as regards a conventional SVM and a hybrid model formed by a SVM and simulated annealing algorithms (SAA).

The Taiwanese electricity market was forecasted by means of SVR in [80]. The author proposed a novel initialization of the SVR by using particle swarm optimization. The results were compared to other SVR but with different initialization strategies, mainly, the least-squares (LS) method.

A two-stage multiple SVM based model for midterm electricity price forecasting was proposed in [81].The first stage was used to separate input data into different price zones, and was carried out by means of a single SVM. Then, four parallel designed SVM were applied to forecast the electricity price. The method was applied to PJM market and the results compared to the standard SVM.

Finally, Table 4 summarizes all the methods reviewed in this section. Note the GRNN stands for general regression neural networks.

Table 4. Summary on SVM for electricity forecasting.

Reference	Technique	Outperforms	Metrics	Horizon	Year	Market
[76]	SAA-SVM	ARIMA/GRNN	MAE/MSRE	1 day	2004	China
[77]	SVM	ANN	MAPE	1 day	2005	China
[78]	M-SVM	SVM	MAE/MSRE	1 day	2006	ANEM
[79]	RS-SVM	SAA-SVM	MAE/MSRE	1 day	2007	NYISO
[80]	PSO-SVM	LS-SVM	MSE	1 day	2009	Taiwan
[81]	M-SVM	SVM	MAE/MSRE	1 day	2009/10	PJM

5.2. Forecasting based on local methods

Due to the complexity to find a global function that models the whole system, the local models emerge as learning methods for time series forecasting. Conversely to global methods, a local model does not use the input data to predict the output but only the points close to the point to forecast. In general, global models have a lower computational cost than local models, since the latter have to be rebuilt for each point of the test set. But, the accuracy achieved by local methods is usually better than that of global methods. The main local methods for prediction tasks are the methods based on nearest neighbors.

5.2.1. Forecasting based on nearest neighbors

5.2.1.1. Fundamentals

One of the most popular way of either predicting or classifying a new data, based on past and known observations, it the nearest neighbors technique (NN), that was first formulated by Cover and Hart in 1967 [82]. The classical example to illustrate the application of NN refers to a doctor that tries to predict the result of a surgical procedure by comparing it with the obtained result from the most similar patient subjected to the same operation. However, a single case in which surgery had failed may have an excessive influence over other slightly different cases in which the operation had successfully carried out. For this reason, the NN algorithm is generalized with the k nearest neighbors, kNN. Thus, a simple election of the k nearest neighbors generates a prediction for every cases. Moreover, this rule can be extended by weighting the importance of the neighbors, giving a larger weight to the really nearest neighbors.

The search of the nearest neighbor process can be defined as follows:

Definition. Given a dataset $\mathbf{P} = p_1, ..., p_n$ in a metric espace X of distance d, two different type of queries are wanted to be answered:

- Nearest neighbor: find the point in P nearest to $q \epsilon X$
- Range: given a point $q \epsilon X$ and $r > 0$, return all the points $p \epsilon P$ that satisfy $d(p, q) \leq r$

Figure 7 illustrates an example in which k is set to three (three nearest neighbors are searched for) and an Euclidean metric is used.

Formally, the classification rule is formulated as follows:

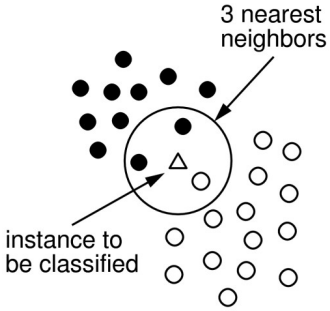

Figure 7. Three nearest neighbors of an instance to be classified.

Definition. Let $\mathcal{D} = \{e_1, \ldots, e_N\}$ be a dataset with N labeled examples, in which each example e_i has m attributes $(e_{i1}, \ldots, e_{im})$ belonging to the metric space $\mathcal{E}^m$ and a class $\mathcal{C}_i \in \{\mathcal{C}_1, \ldots, \mathcal{C}_d\}$. The classification of each new example e' fulfils that:

$$e' \dashv \mathcal{C}_i \Leftrightarrow \forall j \neq i \cdot d(e', e_i) < d(e', e_j) \tag{34}$$

where $e' \dashv \mathcal{C}_i$ indicates the assignation of the class label $\mathcal{C}_i$ to the example e'; and d expresses a distance defined in the m-dimensional space, $\mathcal{E}^m$.

5.2.1.2. Related works

One example is thus labeled according to the nearest neighbor's class. This closeness is defined by means of the distance d which turns the election of this metric essential, since different metrics will most likely generate different classifications. As a consequence the election of the metric is widely discussed in the literature, as shown in [83]. Note that the other main drawback that this technique presents is the selection of the number of neighbors to consider [84].

In [85] a forecasting algorithm based on nearest neighbors was introduced. The selected metric was the weighted Euclidean distance and the weights were calculated by means of a GA. The authors forecasted electricity demand time series in the Spanish market and the reported results were compared to those of an ANN. The same algorithm was tested on electricity price time series in [86] in which the authors proposed a methodology based on weighted nearest neighbors (WNN) techniques. The proposed approach was applied to the 24-hour load forecasting problem and they built an alternative model by means of a conventional dynamic regression (DR) technique, where the parameters are estimated by solving a least squares problem, to perform a comparative analysis.

A modification of the WNN (mWNN) methodology was proposed in [87]. To be precise, they explained how the relevant parameters –the window length of the time series and the number of neighbors to be chosen– are adopted. Then, the approach weighted the nearest neighbors in order to improve the prediction accuracy. The methodology was evaluated with the Spanish electricity prices time series.

Later, WNN was also applied to the California electricity market (CAISO) [88]. This time, the authors reported results for year 2000 and compared the approach to ARIMA-based models.

A multivariate KNN (mKNN) regression method for forecasting the electricity demand in the UK market was presented in [89]. They reported results date from 2004 and were compared to several benchmarks, as well as to univariate KNN (uKNN).

A work reporting short term load forecasting results for India, years 2012 and 2013, can be found in [90]. This paper evaluates the accuracy of Holt-winter model and K-NN algorithm. Their performance

is compared to SARIMA, ANN and SVM, showing that K-NN is the method with better results in terms of MAPE.

Finally, Table 5 summarizes all the methods reviewed in this section. It can be concluded that there exist few works based on KNN to forecast time series, which have mainly been assessed by means of diverse distance metrics in order to identify univariate time series motifs or episodes in the historical data [91].

Table 5. Summary on KNN methods for electricity forecasting.

Reference	Technique	Outperforms	Metrics	Horizon	Year	Market
[85]	KNN	ANN	MRE/MAE	1 day	2002	OMEL
[86]	WNN	DR	MRE/MAE	1 day	2002	OMEL
[87]	mWNN	ANN/GARCH	MRE/MAE	1 day	2002	OMEL
[88]	WNN	ARIMA	MAE/MAPE	1 day	2000	CAISO
[89]	mKNN	uKNN/Benchmarks	MAPE	1 day	2004	UK
[90]	KNN/Holt	SARIMA/ANN/SVM	MAPE	1 day	2012/13	India

6. Rule-based forecasting

6.1. Fundamentals

Prediction based on decision rules usually makes reference to the expert system developed by Collopy and Armstrong in 1992 [92]. The initial approach consisted of 99 rules that combined four extrapolation-based forecasting methods: linear regression, Holt-Winter's exponential smoothing, Brown's exponential smoothing and random walk. During the prediction process, 28 features were extracted in order to characterize the time series. Consequently, this strategy assumed that a time series can be reliably identified by some features. Nevertheless, just eight features were obtained by the system itself since the remaining ones were selected by means of experts' inspections. This fact implies high inefficiency insofar as too much time is taken, the ability of the analyst plays an important (and subjective) role and it shows a medium reliability.

Formally, an association rule (AR) can be expressed as a sentence such that: *If* A *Then* B, with A a logic predicate over the attributes whose fulfillment involves to classify the elements with a label B. The learning based on rules tries to find rules involving the highest number of attributes and samples.

ARs were first defined by Agrawal et al. [93] as follows. Let $I = \{i_1, i_2, ..., i_n\}$ be a set of n items, and $D = \{tr_1, tr_2, ..., tr_N\}$ a set of N transactions, where each tr_j contains a subset of items. Thus, a rule can be defined as $X \Rightarrow Y$, where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. Finally, X and Y are called antecedent (or left side of the rule) and consequent (or right side of the rule), respectively.

When the domain is continuous, the association rules are known as quantitative association rules (QAR). In this context, let $F = \{F_1, ..., F_n\}$ be a set of features, with values in $\mathbf{R}$. Let A and C be two

disjunct subsets of F, that is, $A \subset F$, $C \subset F$, and $A \cap C = \emptyset$. A QAR is a rule $X \Rightarrow Y$, in which features in A belong to the antecedent X, and features in C belong to the consequent Y, such that:

$$X = \bigwedge_{F_i \in A} F_i \in [l_i, u_i] \tag{35}$$

$$Y = \bigwedge_{F_j \in C} F_j \in [l_j, u_j] \tag{36}$$

where l_i and l_j represent the lower limits of the intervals for F_i and F_j respectively, and the couple u_i and u_j the upper ones. For instance, QAR could be numerically expressed as:

$$F_1 \in [12, 25] \wedge F_3 \in [5, 9] \Rightarrow F_2 \in [3, 7] \wedge F_5 \in [2, 8] \tag{37}$$

where F_1 and F_3 constitute the features appearing in the antecedent and F_2 and F_5 the ones in the consequent.

6.2. Related work

Ismail et al. [94] presented a mathematical model for forecasting electricity peak load demand using a rule-based approach. The method was applied to data from Malaysia. The results were compared to SARIMA and regression models.

A data association mining-based rule extraction mechanism to extract the patterns in consumers' reaction to price forecasts can be found in [95]. The resulting rules were then employed to fine-tune the initially generated demand and price forecasts of a multi-input multi-output (MIMO) engine. The methodology was tested on Australia's and New England's electricity data.

A rule-based approach to forecast anomalous load conditions for Great Britain data was introduced in [96]. The authors used Holt-Winters-Taylor exponential smoothing, ARMA, ANN, and singular value decomposition based exponential smoothing to demonstrate how these methods can be adapted to discover outliers, when used together with a rule-based approach.

By contrast, not all the rule-based system provides crisp decisions. Hence, fuzzy rule-based systems are usually used when the available data presents missing values. In these systems, each element can belong to different groups with different grade of membership, not providing thus strict rules for every sample. Due to its flexibility for dealing with incomplete, imprecise or uncertain data, fuzzy rule-based strategies are often applied to prediction purposes. Hence a fuzzy association rule can be expressed as: *If* X is A *Then* Y is B, where X, Y are disjoint subsets of attributes that forms the database and A, B contain the fuzzy sets that are associated with X and Y.

A fuzzy rule based approach is presented to generate a crisp estimate for system load in [97]. To get this done, historical load, temperature, and time information were converted into fuzzy information. The method was applied to the European Energy Exchange (EEE) and the prediction results were compared to the conventional method (CM).

A novel fuzzy logic methodology for short term load forecasting was introduced in [98]. It was concluded that using time, temperature and similar previous day load as the inputs and by formulating rule base of fuzzy logic using available data where enough to obtain reliable fuzzy rules for some particular days. Data from Indian market were analyzed.

A paper focused on improving the performance of fuzzy rules-based forecasters through application of FCM algorithm can be found in [99]. The approach was evaluated by using data of certain region of the USA.

In general, the search of rule-based works to forecast electricity led to the conclusion that this kind of works is scarce. That is, there could be an interesting starting point for those researchers wanting to develop new algorithms.

Finally, Table 6 summarizes all the methods reviewed in this section, where NP means not provided (the authors did not compared their approach to any other).

Table 6. Summary on rule-based methods for electricity forecasting.

Reference	Technique	Outperforms	Metrics	Horizon	Year	Market
[94]	Rules	MA/Smoothing	MAE	1 day	2001-2005	Malaysia
[95]	MIMO	NP	MAPE	1 day	2009	ANEM
[96]	Holt/Rules	SARMA/ANN	MAPE	1 day	2007	UK
[97]	Fuzzy rules	CM	MAPE	1 day	2002-2005	EEE
[98]	Fuzzy rules	NP	MRE	1 day	2013	India
[99]	Fuzzy rules	Holt/ARIMA	MSE/MAPE	1 day	2005	Brazil

7. Wavelet transform methods

7.1. Fundamentals

All the methods described are applied in the time domain. However, time series can also be analyzed in the frequency domain by means of several techniques. Fourier transform –and different Fourier-related transforms such as short-time Fourier transform (STFT), fast Fourier transform (FFT) or discrete Fourier transform– is the most widely used tool to extract the spectral components from temporal data. However, there is another technique derived from this analysis which is more suitable to time series analysis in the frequency domain: the wavelet transform.

There are two different types of wavelet transforms. The discrete wavelet transform (DWT) performance is similar to that of low and high-pass filters, since it divides the time series in high and low frequencies. On the other hand, the continuous wavelet transform (CWT) works as if it was a band-pass filter, isolating just the frequency band of interest. Although both strategies can be used to perform spectral analysis, only the CWT is going to be described in this Section because it is much more useful –and, consequently, used– in time series analysis. DWT is usually used in data that present great variations and discontinuities, which is not the case of time series that frequently as modeled by smooth variations.

Hence, the CWT is a convolution of a time series and the wavelet function [100]. That is, the time series is filtered by a function that plays the same role of the window in the STFT. Nevertheless, in wavelet transform this window has a variable length according to the frequency band to be studied.

Formally, the N points-CWT of a time series x_n, sampled each Δt units of time, is defined as the convolution of such series with an extended and delayed wavelet function $\Psi(t)$:

$$CWT_x(n,s) = \frac{1}{\sqrt{s}} \sum_{n'=0}^{N-1} x_{n'} \Psi^* \left(\frac{n'-n}{s} \Delta t \right) \text{ with } n = 0 \dots N-1 \tag{38}$$

As this product has to be done N times for the scale s considered, if N is too large it is faster to estimate the result by using the FFT than by means of the definition. From the convolution theorem [101], the CWT can be obtained from the inverse fast Fourier transform (IFFT) of time series and the wavelet's direct transform:

$$CWT_x(n,s) = IFFT \left(\frac{1}{\sqrt{s}} FFT(x(n,\Delta t) FFT(\Psi(n,\Delta t,s) \right) \tag{39}$$

Since s is the single parameter from which the transform depends on, the estimation of the CWT can be carried out by means of FFT algorithms for each scale as well as simultaneously for all the points forming the series.

7.2. Related work

Conejo et al. [102] proposed a new approach to predict day-ahead electricity prices based on the wavelet transform and ARIMA models. Thus, they decomposed the time series in a set of better-behaved constitutive series by applying the wavelet transform. Then, the future values of these new series were forecast using ARIMA models, with a prior application of the inverse wavelet transform. This approach improved former strategies that they had also published [103–105].

Aggarwal et al. [106] also forecasted electricity prices. For this purpose, they divided each day into segments and they applied a multiple linear regression (MLR) to the original series or the constitutive series obtained by the wavelet transform depending on the segment. Moreover, the regression model used different input variables for each segment.

Pindoriya et al. [107] proposed an adaptive wavelet-based neural network (AWNN) for short-term electricity price time series forecasting for Spanish and California markets. As for the neural network, the output of the hidden layer neurons was based on wavelets that adapted their shape to training data. The authors concluded that their approach converged with higher rate and outperformed in the forecasting the electricity prices compared to other methods due to the ability for modeling the non-stationary and high frequency signals. The target market was PJM.

An approach based on non-decimated multilevel wavelet (ML-WL) transform, combined with feature selection and machine learning prediction algorithm was presented in [108]. The feature selection integrated autocorrelation and ranking-based methods. The method was applied to Australian electricity data, outperforming exponential smoothing with single and double seasonality, the industry model and all other baselines.

A methodology to forecast normal and spike prices was proposed in [109]. Normal price module was forecasted as a mixture of wavelet transform, ARIMA and ANN models. Price spike occurrences were generated by a three classifiers ensemble. The forecasting accuracy of the proposed method is evaluated with real data from Finland energy market.

The work presented in [110] used Local Linear Wavelet Neural Network (LLWNN) trained by a special adaptive version of the PSO algorithm, with parallel implementation. Experiments for short term load and price forecasting were conducted for Greece and the USA energy markets and were compared to a classic PSO algorithm.

Finally, Table 7 summarizes all the methods reviewed in this section, where WL stands for wavelets.

Table 7. Summary on wavelets for electricity forecasting.

Reference	Technique	Outperforms	Metrics	Horizon	Year	Market
[102]	WL-ARIMA	ARIMA	MRE	1 day	2002	OMEL
[106]	WL-MLR	GARCH	RMSE/MAPE	1 day	2003-2005	ANEM
[107]	AWNN	ANN/MLP/RBF	MAPE/MSE	1 day	2002/2004	OMEL/PJM
[108]	ML-WL-FS	Smoothing	MSE	1 day	2010	ANEM
[109]	WL-ARIMA-ANN	ARIMA	MAPE	1 day	2010	Finland
[110]	LLWNN	PSO	RMSE/MAPE	1 day	2012	Greece/NYISO

8. Other models

Despite of the vast description of methods provided in prior sections, some authors proposed new forecasting approaches that cannot be classified into any of the aforementioned categories. For this reason, this section is describe to introduce all these works.

Hence, transfer functions models (TFM) –known as dynamic econometric models in the economics literature– based on past electricity prices and demand were proposed to forecast day-ahead electricity prices by Nogales et al. in [111], but the prices of all 24 hours of the previous day were not known. They used the median as measure due to the presence of outliers and they stated that the model in which the demand was considered presented better forecasts.

The authors in [112] focussed on the one year-ahead electricity demand prediction for winter seasons by defining a new Bayesian hierarchical model (BH). They provided the marginal posterior distributions of demand peaks. The results for one year-ahead were compared to those of the National Grid Trasc (NGT) group in the United Kingdom.

A fuzzy inference system (FIS) –adopted due to its transparency and interpretability– combined with traditional time series methods was proposed for day-ahead electricity price forecasting [113].

A novel non-parametric model using the manifold learning (MFL) methodology was proposed in [114] in order to predict electricity price time series. For this purpose, the authors used cluster analysis based on the embedded manifold of the original dataset. To be precise, they applied manifold-based dimensionality reduction to curve modeling, showing that the day-ahead curve can be represented by a low-dimensional manifold.

Another different proposal can be found in [115], where a forecasting algorithm based on Grey Models was introduced to predict the load of Shanghai. In the Grey model the original data series was transformed to reduce the noise of the data series and the accuracy was improved by using Markov chains techniques.

The use of clustering as an initial step to forecast electrical time series has been used. For instance, the authors in [116,117] evaluated the performance of both K-means and Fuzzy C-Means in detecting

patterns in the Spanish market. Later, these patterns were used to transform the time series into a sequence of labels showing the benefits of using this information as previous step in time series forecasting [118]. Finally, an extended and improved approach, PSF, was introduced in [119], where New York, Australian and Spanish electricity and demand time series were successfully forecasted, showing remarkable performance compared to classical methods. The same method was adapted to forecast outliers (o-PSF) for the same markets in [120].

A method using a principal component analysis (PCA) network was introduced in [121] to forecast day-ahead prices. The PCA network extracts essential features from periodic information in the market. Later, these features are used as inputs in a multilayer feedforward network. PJM market was used to test the proposed method and the results compared to ARIMA models.

Finally, Table 8 summarizes all the methods reviewed in this section.

Table 8. Summary on other models for electricity forecasting.

Reference	Technique	Outperforms	Metrics	Horizon	Year	Market
[111]	TFM	ARIMA	RMSE/MAPE	1 day	2003	PJM
[112]	BH	NGT	RMSE	1 year	2002/03	UK
[113]	FIS	ARMA/GARCH	RMSE/MAPE	1 day	2003/04	PJM
[114]	MFL	ARIMA/Holt	MSE	up to 1 month	2010	NYISO
[115]	Grey-Markov	Grey	MRE	1 day	2005/06	Shangai
[119]	PSF	5+ methods	MRE/MAPE	1 day	2006	NYISO/ANEM/OMEL
[120]	o-PSF	5+ methods	MRE/MAPE	1 day	2006	NYISO/ANEM/OMEL
[121]	PCA	ANN	MAE	1 day	2008	PJM

9. Ensemble models

Recently, ensemble models are beginning to receive attention from the research community due to the good performance obtained for classification problems [122,123]. In general, ensemble models consists in combining different models in order to improve the accuracy of the individual models. In most of works, the combination is usually based on a system of majority votes (bagging) or weighted majority votes (boosting).

In the last years, ensemble techniques have been also applied to the prediction of energy time series. Fan et al. [124] proposed a machine learning model based on Bayesian Clustering by Dynamics (BCD) and SVM. First, Bayesian clustering techniques were used to split the input data into 24 subsets. Then, SVM methods were applied to each subset to obtain the forecasts of the hourly electricity load for the city of New York.

The work in [125] introduced a price forecasting method based on wavelet transform combined with ARIMA and GARCH models. The method was assessed on Spanish and PJM electricity markets and compared to some other forecasting methods.

An ensemble of RBF neural networks for short-term load forecasting in seven buildings from Italy can be found in [126]. The main novelty of this work is the introduction of a new term in the objective function to minimize the correlation between the error of a network with the errors of the rest of

networks of the ensemble. In this case, the results were compared to SARIMA, which proved to be more competitive in most of the buildings.

An ensemble of ELM was presented in [127] to short-term load forecasting of Australian electricity market. Both the weights of the input layer and the number of nodes in hidden layer for each ELM were randomly set. The median of the outputs generated for each ELM was the final prediction. The results reported an error of 1.82% for the year 2010 versus 2.89%, 2.93%, and 2.86% obtained by a single ELM, a back-propagation ANN and a RBF neural network, respectively.

Many ensembles of ANN have been recently published in the literature with the purpose of electricity prices or load forecasting. In fact, most of the proposed ensemble techniques for regression tasks have been ensembles of ANN. For instance, the authors in [128] proposed the hybrid method PSF-NN, which combines pattern sequence similarity with neural networks. The results show that the use of ensemble of NNs instead of a single NN in the NN component of the PSF-NN prediction method is beneficial considering that it produces better accuracy at acceptable computational cost.

Another ensemble based on PSF was introduced in [129]. In this case, five forecasting models using different clustering techniques: K-means, SOM, Hierarchical Clustering, K-medoids model, and Fuzzy C-means were used. The ensemble model was implemented with an iterative prediction procedure. The method was applied to New York, Australia and Spain markets, and the results compared to those of the original PSF algorithm.

The performance of an ensemble of ANN was compared with a Seasonal Autoregressive Integrated Moving Average (SARIMA) model, a Seasonal Autoregressive Moving Average (SARMA), a Random Forest, a Double Exponential Smoothing and Multiple Regression in [130], providing the best results. The ANNs composing of the ensemble were trained with different subsets provided by a previous clustering.

An ensemble was proposed in [131] to predict the load in California for the next day. The authors used a reference forecast made by the system operator as input variable of the proposed method, and this prediction was improved by means of two Box-Jenkins time series models. Then, the forecasts provided by these two models were combined to obtain the final prediction. The weights of the combination were optimized by means of least square method, and moreover, the authors built different ensembles considering global weights or weights depending on the hour or the day.

Finally, Table 9 summarizes all the methods reviewed in this section.

Table 9. Summary on ensembles for electricity forecasting.

Reference	Technique	Outperforms	Metrics	Horizon	Year	Market
[124]	BCD+SVM	SVR	MAPE	1 day	2001-2003	NYISO
[125]	WL+GARCH	5+ models	RMSE/MAPE	1 day	2002	OMEL/PJM
[126]	ANN	SARIMA	MSE/MAE/MAPE	1 day	2010	Italy
[127]	ELM	ANN/RBF	MAE/MAPE	1 day	2010	ANEM
[128]	PSF+ANN	5+ models	MAE/MAPE	1 day	2010	ANEM
[129]	PSF+Clust	PSF	MRE/MAPE	1 day	2006	NYISO/ANEM/OMEL
[130]	ANN	SARIMA	MAPE	1 day	2012	C&I
[131]	ARIMA	5+ models	RMSE/MAE/MAPE	1 day	2013	CAISO/ERCOT

10. Conclusions

It is expected that this work serve as initial guide for those researchers interested in time series forecasting and, in particular, in forecasting based on data mining approaches. Thus, a brief but rigorous mathematical description of the main existing data mining techniques that have been applied to forecast time series is reported. Due to the wide variety of application of such techniques, one case study has been selected: The analysis of energy-related time series (electricity price and demand). The large amount of works carried out during the last decade in this topic highlights the strengths that data mining had already exhibit in other fields.

With reference to the type of prediction, it can be concluded that almost all methods use a horizon of prediction equals to one day. There are few works forecasting recent years since, for comparative purposes, they prefer to use older data. Moreover, there are several techniques that have been rarely used so far in this research areas: nearest-neighbors and genetic programming. This fact suggests that much work is still remaining for such models. On the contrary, ANN and SVM have been extensively used for this forecasting task. Linear models are still being used, but mainly to be used as baselines, since most of the data mining approaches outperform them in terms of accuracy. Wavelets and rule-based methods are mainly used in hybrid approaches and are causing significative accuracy improvement when properly combined. The accuracy measures mainly used are MAPE and RMSE. Finally, the current trend in electricity forecasting points to the development of ensembles, thus highlighting single strengths of every method.

Acknowledgments

The authors would like to thank Spanish Ministry of Economy and Competitiveness, Junta de Andalucía and Pablo de Olavide University for the support under projects TIN2014-55894-C2-R, P12-TIC-1728 and APPB813097, respectively.

Author Contributions

F. Martínez-Álvarez and A. Troncoso conceived the paper. J. C. Riquelme and G. Asencio-Cortés proposed the paper structure. All authors contributed to the writing of the paper.

The authors declare no conflict of interest

References

1. Sabo, K.; Scitovski, R.; Vazler, I.; Zekić-Sušac, M. Mathematical models of natural gas consumption. *Energy Conversion and Management* **2011**, *52*, 1721–1727.
2. Ye, L.; Qiuru, C.; Haixu, X.; Yijun, L.; Guangping, Z. Customer segmentation for telecom with the k-means clustering method. *Information Technology Journal* **2013**, *12*, 409–413.
3. Aznarte-Mellado, J.L.; Benítez-Sánchez, J.M.; Nieto, D.; Fernández, C.L.; Díaz, C.; Alba-Sánchez, F. Forecasting airborne pollen concentration time series with neural and neuro-fuzzy models. *Expert Systems with Applications* **2007**, *32*, 1218–1225.

4. Záliz, R.R.; Rubio-Escudero, C.; Zwir, I.; del Val, C. Optimization of Multi-classifiers for Computational Biology: Application to gene finding and gene expression. *Theoretical Chemistry Accounts* **2010**, *125*, 599–611.
5. Martínez-Álvarez, F.; Reyes, J.; Morales-Esteban, A.; Rubio-Escudero, C. Determining the best set of seismicity indicators to predict earthquakes. Two case studies: Chile and the Iberian Peninsula. *Knowledge-Based Systems* **2013**, *50*, 198–210.
6. Plazas, M.A.; Conejo, A.J.; Prieto, F.J. Multimarket Optimal Bidding for a Power Producer. *IEEE Transactions on Power Systems* **2005**, *20*, 2041–2050.
7. Aggarwal, S.K.; Saini, L.M.; Kumar, A. Electricity Price Forecasting in Deregulated Markets: A Review and Evaluation. *International Journal of Electrical Power and Energy Systems* **2009**, *31*, 13–22.
8. Weron, R. Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting* **2014**, *30*, 1030–1081.
9. Pennsylvania-New Jersey-Maryland Electricity Market. *http://www.pjm.com*.
10. The New York Independent System Operator. *http://www.nyiso.com*.
11. Spanish Electricity Price Market Operator. *http://www.omel.es*.
12. Australia's National Electricity Market. *http://www.nemmco.com.au*.
13. Independent Electricity System Operator of Ontario. *http://www.ieso.com*.
14. California Independent System Operator. *http://www.caiso.com*.
15. Ramsay, J.O.; Silverman, B.W. *Functional data analysis*; Springer, 2005.
16. Brockwell, P.J.; Davis, R.A. *Introduction to time series and forecasting*; Springer, 2002.
17. Shumway, R.H.; Stoffer, D.S. *Time series analysis and its applications (with R examples)*; Springer, 2011.
18. Maronna, R.A.; Martin, R.D.; Yohai, V.J. *Robust Statistics: Theory and Methods*; Wiley, 2007.
19. Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. *International journal of forecasting* **2006**, *22*, 679–688.
20. Kapetanios, G. Measuring Conditional Persistence in Time Series. *University of London Queen Mary Economics Working Paper* **2002**, *474*.
21. Box, G.; Jenkins, G. *Time series analysis: forecasting and control*; John Wiley and Sons, 2008.
22. Yang, M. Lag Length and Mean Break in Stationary VAR Models. *The Econometrics Journal* **2002**, *5*, 374–386.
23. Wold, H. *A Study in the Analisis of Stationary Time Series*; Almquist and Wicksell, 1954.
24. Kohonen, T. Autoregressive Conditional Heteroskedasticity With Estimates of the Variance of UK. *Inflation Econometrica* **1982**, *50*, 987–1008.
25. Bollerslev, T. Modelling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH model. *Review of Economics and Statistics* **1986**, *72*, 498–505.
26. Xekalaki, E.; Degiannakis, S. *ARCH Models for Financial Applications*; Wiley, 2010.
27. Francq, C.; Zakoian, J.M. *GARCH Models: Structure, Statistical Inference and Financial Applications*; Wiley, 2010.

28. Valipour, M.; Banihabib, M.E.; Behbahani, S.M.R. Parameters Estimate of Autoregressive Moving Average and Autoregressive Integrated Moving Average Models and Compare Their Ability for Inflow Forecasting. *Journal of Mathematics and Statistics* **2012**, *8*, 330–338.
29. Dashora, I.; Singal, S.; Srivastav, D. Streamflow prediction for estimation of hydropower potential. *Water and Energy International* **2015**, *57*, 54–60.
30. Pfeffermann, D.; Sverchkov, M. Estimation of Mean Squared Error of X-11-ARIMA and Other Estimators of Time Series Components. *Journal of Official Statistics* **2014**, *30*, 811–838.
31. Suhartono, S. Time series forecasting by using seasonal autoregressive integrated moving average: Subset, multiplicative or additive model. *Journal of Mathematics and Statistics* **2011**, *7*, 20–27.
32. Miswan, N.H.; Ping, P.Y.; Ahmad, M.H. On parameter estimation for Malaysian gold prices modelling and forecasting. *International Journal of Mathematics Analysis* **2013**, *7*, 1059–1068.
33. Wu, B.; Chang, C.L. Using genetic algorithms to parameters (d; r) estimation for threshold autoregressive models. *Computational Statistics & Data Analysis* **2002**, *38*, 315–330.
34. Wei, S.; Lei, L.; Qun, H. Research on weighted iterative stage parameter estimation algorithm of time series model. *Applied Mechanics and Materials* **2014**, *687-691*, 3968–3971.
35. Hassan, S.; Jaafar, J.; Belhaouari, B.; Khosravi, A. A new genetic fuzzy system approach for parameter estimation of ARIMA model. Proceedings of the International Conference on Fundamental and Applied Sciences, 2012, Vol. 1482, pp. 455–459.
36. Chen, B.S.; Lee, B.K.; Peng, S.C. Maximum Likelihood Parameter Estimation of F-ARIMA Processes Using the Genetic Algorithm in the Frequency Domain. *IEEE Transactions on Signal Processing* **2002**, *50*, 2208–2220.
37. Peña, D.; Tiao, G.C.; Tsay, R.S. *A Course in Time Series Analysis*; Wiley, 2001.
38. Guirguis, H.S.; Felder, F.A. Further Advances in Forecasting Day-Ahead Electricity Prices Using Time Series Models. *KIEE International Transactions on PE* **2004**, *4-A*, 159–166.
39. García, R.C.; Contreras, J.; van Akkeren, M.; García, J.B. A GARCH Forecasting Model to Predict Day-Ahead Electricity Prices. *IEEE Transactions on Power Systems* **2005**, *20*.
40. Malo, P.; Kanto, A. Evaluating Multivariate GARCH Models in the Nordic Electricity Markets. *Communications in Statistics: Simulation and Computation* **2006**, *35*, 117–148.
41. Weron, R.; Misiorek, A. Forecasting Spot Electricity Prices with Time Series Models. *International Conference: The European Electricity Market* **2005**, pp. 52–60.
42. García-Martos, C.; Rodríguez, J.; Sánchez, M.J. Mixed Models for Short-Run Forecasting of Electricity Prices: Application for the Spanish Market. *IEEE Transactions on Power Systems* **2007**, *22*, 544–552.
43. Weron, R.; Misiorek, A. Forecasting Spot Electricity Prices: a Comparison of Parametric and Semiparametric Time Series Models. *International Journal of Forecasting* **2008**, *24*, 744–763.
44. McCulloch, W.S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* **1943**, *5*, 115–133.
45. Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review* **1958**, *65*, 386–408.

1006 46. Huang, G.B.; Zhu, Q.Y.; Siew, C.K. Extreme Learning Machine: Theory and Applications.
1007 *Neurocomputing* **2006**, *70*, 489–501.
1008 47. Rumelhart, D.E.; Hinton, G.E.; Willians, R.J. *Learning internal representations by error*
1009 *propagation*; MIT Press, 1986; pp. 673–695.
1010 48. Kohonen, T. Self-organized formation of topologically correct feature maps. *Biological*
1011 *Cybernetics* **1986**, *43*, 59–69.
1012 49. Rodríguez, C.P.; Anders, G.J. Energy price forecasting in the Ontario Competitive Power System
1013 Market. *IEEE Transactions on Power Systems* **2004**, *19*, 366–374.
1014 50. Amjady, N. Day-Ahead Price Forecasting of Electricity Markets by a New Fuzzy Neural
1015 Network. *IEEE Transactions on Power Systems* **2006**, *21*.
1016 51. Taylor, J. Density forecasting for the efficient balancing of the generation and consumption of
1017 electricity. *International Journal of Forecasting* **2006**, *22*, 707–724.
1018 52. Catalao, J.P.S.; Mariano, S.J.P.S.; Mendes, V.M.F.; Ferreira, L.A.F.M. Short-term electricity
1019 prices forecasting in a competitive market: a neural network approach. *Electric Power Systems*
1020 *Research* **2007**, *77*, 1297–1304.
1021 53. Pino, R.; Parreno, J.; Gómez, A.; Priore, P. Forecasting next-day price of electricity in the
1022 Spanish energy market using artificial neural networks. *Engineering Applications of Artificial*
1023 *Intelligence* **2008**, *21*, 53–62.
1024 54. Zurada, J.M. *An introduction to artificial neural systems*; St. Paul: West Publishing Company,
1025 1992.
1026 55. El-Telbany, M.; El-Karmi, F. Short-term forecasting of Jordanian electricity demand using
1027 particle swarm optimization. *Electric Power Systems Research* **2008**, *78*, 425–433.
1028 56. Neupane, B.; Perera, K.S.; Aung, Z.; Woon, W.L. Artificial Neural Network-based Electricity
1029 Price Forecasting for Smart Grid Deployment. *Proceedings of the IEEE International Conference*
1030 *on Computer Systems and Industrial Informatics* **2012**, pp. 103–114.
1031 57. Koprinska, I.; Rana, M.; Agelidis, V.G. Correlation and instance based feature selection for
1032 electricity load forecasting. *Knowledge-Based Systems* **2015**, *82*, 29–40.
1033 58. Chen, X.; Dong, Z.Y.; Meng, K.; Xu, Y.; Wong, K.P.; Ngan, H. Electricity Price Forecasting
1034 With Extreme Learning Machine and Bootstrapping. *IEEE Transactions on Power Systems*
1035 **2012**, *27*, 2055–2062.
1036 59. Wan, C.; Xu, Z.; Wang, Y.; Dong, Z.Y.; Wong, K.P. A Hybrid Approach for Probabilistic
1037 Forecasting of Electricity Price. *IEEE Transactions on Smart Grid* **2014**, *5*, 463–470.
1038 60. Cecati, C.; Kolbusz, J.; Rozycki, P.; Siano, P.; Wilamowski, B. A Novel RBF Training Algorithm
1039 for Short-Term Electric Load Forecasting and Comparative Studies. *IEEE Transactions on*
1040 *Industrial Electronics* **2015**, *62*, 6519–6529.
1041 61. Li, S.; Wang, P.; Goel, L. Short-term load forecasting by wavelet transform and evolutionary
1042 extreme learning machine. *Electric Power Systems Research* **2015**, *122*, 96–103.
1043 62. Fan, S.; Mao, C.; Chen, L. Next day electricity-price forecasting using a hybrid network. *IET*
1044 *Generation, Transmission and Distribution* **2007**, *1*, 176–182.

63. Jin, C.H.; Pok, G.; Lee, Y.; Park, H.W.; Kim, K.D.; Yun, U.; Ryu, K.H. A SOM clustering pattern sequence-based next symbol prediction method for day-ahead direct electricity load and price forecasting. *Energy Conversion and Management* **2015**, *90*, 84 – 92.
64. López, M.; Valero, S.; Senabre, C.; Aparicio, J.; Gabaldon, A. Application of SOM neural networks to short-term load forecasting: The Spanish electricity market case study. *Electric Power Systems Research* **2012**, *91*, 18–27.
65. Goldberg, D.E. *Genetic algorithms in search, optimization and machine learning*; Addison-Wesley: Massachusetts, USA, 1989.
66. Koza, J.R. *Genetic programming: on the programming of computers by means of natural selection*; MA: MIT Press: Cambridge, 1992.
67. Bhattacharya, M.; Abraham, A.; Nath, B. A Linear Genetic Programming Approach for modelling Electricity Demand Prediction in Victoria. Proceedings of International Workshop on Hybrid Intelligent Systems, 2002, pp. 379–394.
68. Troncoso, A.; Riquelme, J.M.; Riquelme, J.C.; Gómez, A.; Martínez, J.L. Time-Series Prediction: Application to the Short Term Electric Energy Demand. *Lecture Notes in Artificial Intelligence* **2004**, *3040*, 577–586.
69. Amjady, N.; Keynia, F. Day-Ahead Price Forecasting of Electricity Markets by Mutual Information and Cascaded Neuro-Evolutionary Algorithm. *IEEE Transactions on Power Systems* **2009**, *24*, 306–318.
70. Cunkas, M.; Taskiran, U. Turkey's Electricity Consumption Forecasting Using Genetic Programming. *Energy Sources, Part B: Economics, Planning, and Policy* **2011**, *6*, 406–416.
71. Ghareeb, W.T.; El-Saadany, E.F. Multi-Gene Genetic Programming for Short Term Load Forecasting. Proceedings of the International Conference on Electric Power and Energy Conversion Systems, 2013, pp. 1–5.
72. Castelli, M.; Vanneschi, L.; Felice, M.D. Forecasting short-term electricity consumption using a semantics-based genetic programming framework: The South Italy case. *Energy Economics* **2015**, *47*, 37–41.
73. Cortes, C.; Vapnik, V. Support-Vector Networks. *Machine Learning* **1995**, *20*, 273–297.
74. Vapnik, V. *Statistical learning theory*; Wiley, 1998.
75. Vapnik, V. An Overview of Statistical Learning Theory. *IEEE Transactions on Neural Networks* **1999**, *10*, 988–999.
76. Hong, W.C. Electricity Load Forecasting by using SVM with Simulated Annealing Algorithm. Proceedings of World Congress of Scientific Computation, Applied Mathematics and Simulation, 2005, pp. 113–120.
77. Guo, Y.; Niu, D.; Chen, Y. Support-Vector Machine Model in Electricity Load Forecasting. Proceedings of the International Conference on Machine Learning and Cybernetics, 2006, pp. 2892–2896.
78. Zhao, J.H.; Dong, Z.Y.; Li, X.; Wong, K.P. A Framework for Electricity Price Spike Analysis with Advanced Data Mining Methods. *IEEE Transactions on Power Systems* **2007**, *22*, 376–385.
79. Wang, J.; Wang, L. A new method for short-term electricity load forecasting. *Transactions of the Institute of Measurement and Control* **2008**, *30*, 331–344.

80. Qiu, Z. Electricity Consumption Prediction based on Data Mining Techniques with Particle Swarm Optimization. *International Journal of Database Theory and Application* **2013**, *6*, 153–164.
81. Yan, X.; Chowdhury, N.A. Midterm Electricity Market Clearing Price Forecasting Using Two-Stage Multiple Support Vector Machine. *Journal of Energy* **2015**, *ID384528*, 1–11.
82. Cover, T.M.; Hart, P.E. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* **1967**, *13*, 21–27.
83. Wang, J.; Neskovic, P.; Cooper, L.N. Improving nearest neighbor rule with a simple adaptive distance measure. *Pattern Recognition Letters* **2007**, *28*, 207–213.
84. Wang, J.; Neskovic, P.; Cooper, L.N. Neighborhood selection in the k-nearest neighbor rule using statistical confidence. *Pattern Recognition* **2006**, *39*, 417–423.
85. Troncoso, A.; Riquelme, J.C.; Riquelme, J.M.; Martínez, J.L.; Gómez, A. Electricity Market Price Forecasting: Neural Networks versus Weighted-Distance k Nearest Neighbours. *Lecture Notes in Computer Science* **2002**, *2453*, 321–330.
86. Troncoso, A.; Riquelme, J.M.; Riquelme, J.C.; Gómez, A.; Martínez, J.L. A Comparison of Two Techniques for Next-Day Electricity Price Forecasting. *Lecture Notes in Computer Science* **2002**, *2412*, 384–390.
87. Troncoso, A.; Riquelme, J.C.; Riquelme, J.M.; Martínez, J.L.; Gómez, A. Electricity Market Price Forecasting Based on Weighted Nearest Neighbours Techniques. *IEEE Transactions on Power Systems* **2007**, *22*, 1294–1301.
88. Bhanu, C.V.K.; Sudheer, G.; Radhakrishn, C.; Phanikanth, V. Day-ahead Electricity Price forecasting using Wavelets and Weighted Nearest Neighborhood. Proceedings of the International Conference on Power System Technology, 2008, pp. 422–425.
89. Al-Qahtani, F.H.; Crone, S.F. Multivariate k-Nearest Neighbour Regression for Time Series data –a novel Algorithm for Forecasting UK Electricity Demand. Proceedings of the International Joint Conference on Neural Networks, 2013, pp. 1–8.
90. Shelke, M.; Thakare, P.D. Short Term Load Forecasting by Using Data Mining Techniques. *International Journal of Science and Research* **2014**, *3*, 1363–1367.
91. Martínez-Álvarez, F.; Troncoso, A.; Riquelme, J.C. Improving time series forecasting by discovering frequent episodes in sequences. *Lecture Notes in Computer Science* **2009**, *5772*, 357–368.
92. Collopy, F.; Armstrong, J.S. Rule-based forecasting: Development and validation of an expert systems approach to combining time series extrapolations. *Management Science* **1992**, *38*, 1392–1414.
93. Agrawal, R.; Imielinski, T.; Swami, A. Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 1993, pp. 207–216.
94. Ismail, Z.; A.Yahya.; Mahpol, K. Forecasting Peak Load Electricity Demand Using Statistics and Rule Based Approach. *American Journal of Applied Sciences* **2009**, *6*, 1618–1625.

95. Motamedi, A.; Zareipour, H.; Rosehart, W.D. Short-Term Forecasting of Anomalous Load Using Rule-Based Triple Seasonal Methods. *Electricity Price and Demand Forecasting in Smart Grids* **2012**, *3*, 664–674.
96. Arora, S.; Taylor, J.W. Short-Term Forecasting of Anomalous Load Using Rule-Based Triple Seasonal Methods. *IEEE Transactions on Power Systems* **2013**, *28*, 3235–3242.
97. Aggarwal, S.K.; Kumar, M.; Saini, L.M.; Kumar, A. Short-Term Load Forecasting in Deregulated Electricity Markets using Fuzzy Approach. *Journal of Engineering and Technology* **2011**, *1*, 24–30.
98. Manoj, P.P.; Shah, A.P. Fuzzy logic methodology for short term load forecasting. *International Journal of Research in Engineering and Technology* **2010**, *3*, 322–328.
99. Faustino, C.P.; Novaes, C.P.; Pinheiro, C.A.M.; Carpinteiro, O.A. Improving the performance of fuzzy rules-based forecasters through application of FCM algorithm. *Artificial Intelligence Review* **2014**, *41*, 287–300.
100. Daubechies, I. *Ten lectures on wavelets*; Society of Industrial in Applied Mathematics, 1992.
101. Proakis, J.G.; Manolakis, D.G. *Digital Signal Processing*; Prentice Hall, 1998.
102. Conejo, A.J.; Plazas, M.A.; Espínola, R.; Molina, B. Day-Ahead Electricity Price Forecasting using the Wavelet Transform and ARIMA Models. *IEEE Transactions on Power Systems* **2005**, *20*, 1035–1042.
103. Jiménez, N.; Conejo, A.J. Short-Term Hydro-Thermal Coordination by Lagrangian Relaxation: Solution of the Dual Problem. *IEEE Transactions on Power System* **1999**, *14*, 89–95.
104. Nogales, F.J.; Contreras, J.; Conejo, A.J.; Espínola, R. Forecasting Next-Day Electricity Prices by Time Series Models. *IEEE Transactions on Power System* **2002**, *17*, 342–348.
105. Contreras, J.; Espínola, R.; Nogales, F.J.; Conejo, A.J. ARIMA Models to Predict Next-Day Electricity Prices. *IEEE Transactions on Power System* **2003**, *18*, 1014–1020.
106. Aggarwal, S.K.; Saini, L.M.; Kumar, A. Price forecasting using wavelet transform and LSE based mixed model in Australian Electricity Market. *International Journal of Energy Sector Management* **2008**, *2*, 521–546.
107. Pindoriya, N.M.; Singh, S.N.; Singh, S.K. An Adaptative Wavelet Neural Network-Based Energy Price Forecasting in Electricity Markets. *IEEE Transactions on Power Systems* **2008**, *23*, 1423–1432.
108. Rana, M.; Koprinska, I. Electricity Load Forecasting Using Non-decimated Wavelet Prediction Methods With Two-Stage Feature Selection. Proceedings of the International Joint Conference on Neural Networks, 2012, pp. 10–15.
109. Voronin, S.; Partanen, J. Price Forecasting in the Day-Ahead Energy Market by an Iterative Method with Separate Normal Price and Price Spike Frameworks. *Energies* **2013**, *6*, 5897–5920.
110. Kintsakis, A.M.; Chrysopoulos, A.; Mitkas, P.A. Agent-based Short-Term Load and Price Forecasting Using a Parallel Implementation of an Adaptive PSO Trained Local Linear Wavelet Neural Network. Proceedings of the International Conference on the European Energy Market, 2015, pp. 1–5.
111. Nogales, F.J.; Conejo, A.J. Electricity Price Forecasting Through Transfer Function Models. *Journal of the Operational Research Society* **2006**, *57*, 350–356.

112. Pezzulli, S.; Frederic, P.; Majithia, S.; Sabbagh, S.; Black, E.; Sutton, R.; Stephenson, D. The seasonal forecast of electricity demand: a hierchical Bayesian model with climatological weather generator. *Applied Stochastic Models in Business and Industry* **2006**, *22*, 113–125.
113. Li, G.; Liu, C.C.; Mattson, C.; Lawarrée, J. Day-ahead electricity price forecasting in a grid environment. *IEEE Transactions on Power Systems* **2007**, *22*, 266–274.
114. Chen, J.; Deng, S.J.; Huo, X. Electricity Price Curve Modeling by Manifold Learning. *IEEE Transactions on Power Systems* **2008**, *23*, 877–888.
115. Wang, X.; Meng, M. Forecasting electricity demand using Grey-Markov model. Proceedings of the International Conference on Machine Learning and Cybernetics, 2008, pp. 1244–1248.
116. Martínez-Álvarez, F.; Troncoso, A.; Riquelme, J.C.; Riquelme, J.M. Partitioning-clustering techniques applied to the electricity price time series. *Lecture Notes in Computer Science* **2007**, *4881*, 990–991.
117. Martínez-Álvarez, F.; Troncoso, A.; Riquelme, J.C.; Riquelme, J.M. Discovering patterns in electricity price using clustering techniques. Proceedings of the International Conference on Renewable Energy and Power Quality, 2007, pp. 245–252.
118. Martínez-Álvarez, F.; Troncoso, A.; Riquelme, J.C.; Aguilar, J.S. LBF: A Labeled-Based Forecasting Algorithm and its Application to Electricity Price Time Series. Proceedings of IEEE International Conference on Data Mining, 2008, pp. 453–461.
119. Martínez-Álvarez, F.; Troncoso, A.; Riquelme, J.C.; Aguilar, J.S. Energy time series forecasting based on pattern sequence similarity. *IEEE Transactions on Knowledge and Data Engineering* **2011**, *23*, 1230–1243.
120. Martínez-Álvarez, F.; Troncoso, A.; Riquelme, J.C.; Aguilar-Ruiz, J.S. Discovery of motifs to forecast outlier occurrence in time series. *Pattern Recognition Letters* **2011**, *32*, 1652–1665.
121. Hong, Y.Y.; Wu, C.P. Day-Ahead Electricity Price Forecasting Using a Hybrid Principal Component Analysis Network. *Energies* **2012**, *5*, 4711–4725.
122. Galar, M.; Fernandez, A.; Barrenechea, E.; Herrera, F. EUSBoost: Enhancing Ensembles for Highly Imbalanced Data-sets by Evolutionary Undersampling. *Pattern Recognition* **2013**, *46*, 3460–3471.
123. Galar, M.; Derrac, J.; Peralta, D.; Triguero, I.; Paternain, D.; Lopez-Molina, C.; García, S.; Benítez, J.; Pagola, M.; Barrenechea, E.; Bustince, H.; Herrera, F. A Survey of Fingerprint Classification Part II: Experimental Analysis and Ensemble Proposal. *Knowledge-Based Systems* **2015**, *81*, 98–116.
124. Fan, S.; Mao, C.; Zhang, J.; Chen, L. Forecasting Electricity Demand by Hybrid Machine Learning Model. *Lecture Notes in Computer Science* **2006**, *4233*, 952–963.
125. Tan, Z.; Zhang, J.; Wang, J.; Xu, J. Day-Ahead Electricity Price Forecasting Using Wavelet Transform Combined with ARIMA and GARCH Models. *Applied Energy* **2010**, *87*, 3606–3610.
126. De Felice, M.; Yao, X. Short-Term Load Forecasting with Neural Network Ensembles: A Comparative Study [Application Notes]. *IEEE Computational Intelligence Magazine* **2011**, *6*, 47–56.

127. Zhang, R.; Dong, Z.Y.; Xu, Y.; Meng, K.; Wong, K.P. Short-term load forecasting of Australian National Electricity Market by an ensemble model of extreme learning machine. *IET Generation, Transmission and Distribution* **2013**, *7*, 391–397.
128. Koprinska, I.; Rana, M.; Troncoso, A.; Martínez-Álvarez, F. Combining Pattern Sequence Similarity with Neural Networks for Forecasting Electricity Demand Time Series. Proceedings of the International Joint Conference on Neural Networks, 2013, pp. 940–947.
129. Shen, W.; Babushkin, V.; Aung, Z.; Woon, W. An ensemble model for day-ahead electricity demand time series forecasting. Proceedings of the ACM Conference on Future Energy Systems, 2013, pp. 51–62.
130. Jetcheva, J.G.; Majidpour, M.; Chen, W.P. Neural network model ensembles for building-level electricity load forecasts. *Energy and Buildings* **2014**, *84*, 214 – 223.
131. Kaur, A.; Pedro, H.T.; Coimbra, C.F. Ensemble re-forecasting methods for enhanced power load prediction. *Energy Conversion and Management* **2014**, *80*, 582 – 590.

© November 6, 2015 by the author; submitted to *Energies* for possible open access publication under the terms and conditions of the Creative Commons Attribution license http://creativecommons.org/licenses/by/4.0/.