Discovery of motifs to forecast outlier occurrence in time series

F. Martínez–Álvarez , A. Troncoso, J.C. Riquelme , J.S. Aguilar–Ruiz

A B S T R A C T

The forecasting process of real-world time series has to deal with especially unexpected values, commonly known as outliers. Outliers in time series can lead to unreliable modeling and poor forecasts. Therefore, the identification of future outlier occurrence is an essential task in time series analysis to reduce the average forecasting error. The main goal of this work is to predict the occurrence of outliers in time series, based on the discovery of motifs. In this sense, motifs will be those pattern sequences preceding certain data marked as anomalous by the proposed metaheuristic in a training set. Once the motifs are discovered, if data to be predicted are preceded by any of them, such data are identified as outliers, and treated separately from the rest of regular data. The forecasting of outlier occurrence has been added as an additional step in an existing time series forecasting algorithm (PSF), which was based on pattern sequence similarities. Robust statistical methods have been used to evaluate the accuracy of the proposed approach regarding the forecasting of both occurrence of outliers and their corresponding values. Finally, the methodology has been tested on six electricity-related time series, in which most of the outliers were properly found and forecasted.

Keywords:
Time series forecasting
Pattern recognition
Motifs
Outliers

1. Introduction

This work proposes a new strategy to predict the occurrence of outlying data in time series, as well as providing accurate forecasts for them. It is worth highlighting that the goal of this methodology is to forecast their appearance, instead of detecting them in an already known set of values, which is a common goal in robust statistics (Maronna et al., 2007). The majority of robust statistical techniques perform a posteriori detection, that is, they determine whether a datum is an outlier or not, but once it has already occurred. However, a comparison with these techniques will be of the utmost importance in order to evaluate the accuracy of the proposed metaheuristic.

A general-purpose forecasting algorithm, called PSF, was presented in Martínez-Álvarez et al. (in press). Its main feature lied in performing a discretization of the time series by means of certain clustering technique. Then, it only used the generated labels to make predictions. Based on that previous discretization, this work attempts to discover pattern sequences (henceforth called motifs) in the historical data to forecast the occurrence of outliers and their associated values. This novel methodology is inserted in the general scheme of PSF.

The prediction of outliers plays an important role as wrong models and poor forecasts are obtained when ignoring outliers. This work presents a metaheuristic to discover motifs in time series and, then, if data to be predicted are preceded by any of these discovered motifs, consider these data as outlier. The motifs are determined during the training phase as those pattern sequences that precede data with remarkable forecasting error. Thus, the existing PSF algorithm is modified by adding a new motif extraction step.

The enhanced version is capable of predicting the appearance of such outliers with great reliability when the motifs extraction step is added. In fact, the approach has been successfully tested on six real-world electricity-related time series, in particular, on energy prices and demand of three different markets, reaching sensitivity values greater than 82%, and specificity values greater than 95%. Furthermore, results about the effect of outliers on the average forecasting errors are reported for all the six time series, exhibiting remarkable forecasting error reduction.

Despite the vast variety of works related to outliers detection and motifs discovery in time series, there is no approach in time series in order to forecast the occurrence of outliers, to the authors' knowledge.

The remaining of the paper is organized as follows. A review of the most recently published works regarding energy time series forecasting, motifs discovery and outliers detection can be found in Section 2. Section 3 provides formal description for sensitive terms, and presents a brief explanation of the original algorithm. As for Section 4, it introduces the proposed methodology, showing

how to insert the outlier occurrence forecasting in the original algorithm's general scheme. The results obtained for the six electricity prices and demand time series are reported and discussed in Section 5. Finally, Section 6 summarizes the main conclusions achieved.

2. Related work

This section provides useful and recent references about the three main topics involved in this paper: Energy time series forecasting, motifs discovery in temporal data and robust statistical methods to detect outliers. For the sake of clarity, these topics have been separated in three different sections.

2.1. Energy time series forecasting

The interest of analyzing electricity price time series resides in the progressive deregulation of electric power markets. Furthermore, electricity price time series possess certain features that turn the prediction into a difficult task: non-constant mean/variance and frequently outlier occurrences. For this reason, electricity-producer companies want optimized bidding strategies as well as needing assessment about the risk of trusting forecasts (Plazas et al., 2005).

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 a limited 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 increased costs and a significant reduction of profits.

The pursuit of accurate forecasting in electricity price time series has motivated research works by many authors (Aggarwal et al., 2009). Thus, the use of mixed models was proposed in García-Martos et al. (2007) to forecast prices for different horizons of prediction. Also remarkable was the work introduced in Troncoso et al. (2007) that, by means of weighted nearest neighbors methodology, forecasted next-day electricity prices. Also, the use of an artificial neural network to fulfil the same goal can be found in Pino et al. (2008). Even the use of classical autoregressive models has been recently used to forecast prices in several markets (Weron and Misiorek, 2008).

On the contrary, the authors in Troncoso et al. (2004) proposed a weighted nearest neighbors-based methodology to forecast electricity demand. The proposed approach was tested over the next-day Spanish load forecasting. Also, the authors in El-Telbany and El-Karmi (2008) forecasted the Jordanian electricity demand with an artificial neural network, which was trained by means of particle swarm optimization techniques. This market was also studied in Badran et al. (2008) but, this time, the authors preferred to concentrate on short and medium-term load forecasting by using regression models. Finally, Wang and Wang (2008) proposed a new prediction approach based on support vector machines (SVM) techniques with a previous selection of features from data sets by using an evolutionary method.

The discovery of outliers in electricity price time series has also been widely discussed in literature. Thus, the authors in Lu et al. (2005) proposed a model based on the analysis of several variables (among which the electricity demand is highlighted) by means of Bayesian classification (BC) and similarity searching techniques. A hybrid methodology that combined SVMs and BC was developed in Wu et al. (2006) to classify both spikes and normal electricity prices. Alternatively, a data mining framework based on SVM and probability classifiers was described in Zhao et al. (2007) with the aim of forecasting spikes in prices accurately.

By contrast, the prediction of peaks in electricity demand was addressed in Saini (2008). This work forecasted demand peaks up to seven days ahead using feed forward neural network and adaptive backpropagation learning methods. In the work introduced in Ismail et al. (2009), the authors developed a rule-based method that combined regression models and fuzzy systems to analyze daily electricity peak load demands in Malaysia. Also, the authors in Hyndman and Fan (2010) described a semi-parametric additive model to discover relationships between the demand and exogen variables. The approach was applied to long-term peaks for the South Australian market.

2.2. Motifs discovery

The discovery of motifs in continuous data, also known as functional data in many works (Valderrama, 2008), was originally formalized in Lin et al. (2002), in which the authors introduced several algorithms to mine motifs in time series, among which the k-motif algorithm highlights. However, the main drawback of this algorithm is its dependence on a pre-fixed pattern length. Later, the authors in Tang and Liao (2008) proposed a modified version that improved, precisely, this feature. Furthermore, they generated original patterns by considering the discovered motifs.

The detection of on-line motifs in continuous data has also been addressed. Particularly, a new methodology to detect on-line motifs in time series by combining probabilistic models and polynomial least-squares approximations was proposed in Fuchs et al. (2009). This topic was also studied in Mueen and Keogh (2010) in which the authors found and maintained time series motifs from robotics, online compression and wildlife management domains.

Besides these goals, different objectives have been fulfilled by discovering motifs recently. Hence, the work presented in Tanaka et al. (2005) proposed an algorithm devoted to discover motifs based on the minimum description length principle. The approach also allowed to obtain motifs from multi-dimensional time series data by using principal component analysis. Mueen et al. (2009) proposed in 2009 an exact algorithm to find time series motifs much faster than brute-force searching strategies do. Also, an approach based on tree-construction search to discover motifs in multivariate time series was proposed in Wang et al. (2010) and applied to sensory datasets.

Since Stormo (2000) first reviewed strategies to find DNA motifs (meaningful base sequence patterns that identify binding sites responsible for transcription factors) in 2000, a large amount of algorithms have been developed. Thus, an ensemble algorithm attempting to discover regulatory motifs in DNA sequences was proposed in Hu and Kihara (2006). Another algorithm was proposed in Wijaya et al. (2007) which, given a set of sequences, executes m different motif finders, each of them reporting n motifs. Finally, Sharov and Minoru (2009) presented CisFinder, a software that generates a comprehensive list of motifs enriched in a set of DNA sequences and describes them with position frequency matrices.

2.3. Robust statistical methods to detect outliers

The problem of a posteriori outliers detection in time series has been widely studied in the literature, and faced from many different points of view. In fact, the existence of even few outliers usually leads to inaccurate models and not satisfactory forecasts (Galeano et al., 2006), since they may deeply influence the estimates that classical methods propose (Carnero et al., 2007).

For this reason, there is a large family of robust statistical methods (Rousseeuw and Hubert, 2011) that deal with outliers and, particularly, propose approaches to detect their existence in the datasets subjected to analysis. Gelper et al. proposed an

adapted version of the classical exponential and Holt-Winters smoothing methodologies, providing them with robustness (Gelper et al., 2010). Another version of a robust multivariate exponential smoothing applied to time series can be found in Croux et al. (2010). Following with classical methods, a work that enhanced ARMA by adding robustness can be found in Muler et al. (2009), in which the authors succeeded in limiting the effect of outlying data to the time stamp in which they happen.

Support vector machines (SVM) have also been adapted to deal with outliers. Actually an approach to model time series using robust SVM was proposed in Camps-Valls et al. (2004). In particular, the authors claimed that their proposal provides stable models and allows the analysis of models' memory depth. Recently, a two-steps methodology was proposed in Chuang and Lee (2011) that combined the use of robust SVM to remove anomalous observations, and non-robust SVM to obtain estimates from that reduced dataset.

Many of these proposals have been implemented and freely distributed in software packages. But from all of them there are two that highlight. LIBRA (Verboven and Hubert, 2005) is a Matlab library for robust analysis, that contains (among others) robust covariance estimation, regression, principal component analysis, principal component regression or partial least squares, as well as methodologies to detect outlying observations in datasets. The TOMCAT toolbox Daszykowski et al. (2007), also developed in the Matlab environment, includes almost the same methods that LIBRA does, but also includes a graphical interface.

3. Fundamentals

This Section first defines some terms in order to prevent possible misinterpretations in sensitive terms. Since the proposed methodology is based on an existing algorithm, this Section also provides a brief summary of the mathematical fundamentals underlying the PSF algorithm. Note that a more detailed explanation can be found in Martínez-Álvarez et al. (in press).

3.1. Definitions

This work uses certain concepts –such as outlier or motif– that can be interpreted in many different senses, depending on the application or even the author. For this reason, this Section provides a formal definition for these sensitive terms.

Definition 1 (*Hourly time series*). An hourly time series T is a set of real-valued data in successive order, occurring every hour. In this work, $T = [t_1, \ldots, t_p]$, where p is the length of the time series and usually a multiple of 24.

Definition 2 (*Daily time series*). From an hourly time series, a daily time series D is formed by tuples in R^{24}, $D = [d_1, \ldots, d_{p/24}]$, where $d_i = [t_{24(i-1)+1}, \ldots, t_{24i}]$.

Definition 3 (*Label*). In this work, the term label is used to identify a set of possible categorical values. Thus, $\mathcal{L} = \{l_1, \ldots, l_K\}$, where K is a pre-fixed number.

Definition 4 (*Sequence*). A sequence S is a set of labels occurring in successive order. In this work, $S = [s_1, \ldots, s_q]$, where q is the length of the sequence and $s_i \in \mathcal{L}$.

Definition 5 (*Outlier*). Given a test set, an outlier is an observation which appears to be inconsistent with the rest of the data, relative to an assumed model (Everitt, 2006). Outliers are usually represented by a binary random variable v_i, for $i = 1, \ldots, p$ that models their occurrence ($v_i = 1$ if it occurs, and $v_i = 0$ otherwise), and by another real random variable z_i, for $i = 1, \ldots, p$ that models their magnitude (Maronna et al., 2007). Although different outlier types can be found in the literature, only the additive outlier model is considered in this work, due to the nature of the studied data:

$$t_i = x_i + v_i z_i, \tag{1}$$

where t_i the observed value, and x_i the $i - th$ cleaned data modeled by any approach.

Definition 6 (*Motif*). A motif M_W is a sequence of W consecutive labels considered to occur just before an outlier, where W is the pre-fixed length of the sequence. In addition, M_W is a subsequence found in S and, consequently: $M_W = [s'_1, \ldots, s'_W]$, where $s'_i \in \mathcal{L}$.

3.2. Time series forecasting: the PSF algorithm

The PSF algorithm is a general-purpose time series forecasting algorithm whose main feature is that it only makes use of certain labels –obtained by means of a clustering process– to forecast arbitrary horizons of prediction. However, the output is not composed by labels but by real values.

PSF can deal with an arbitrary number of samples per day. However, specifically in this paper, the time series considered consists of twenty-four samples per day. That is, given the hourly values up to day i for a time series, the PSF algorithm provides the 24 hourly values corresponding to day $i + 1$. Formally, let $d_i \in \boldsymbol{R}^{24}$ be a vector that comprises the 24 hourly values of a certain day, i.

First, PSF applies clustering techniques to such data in order to assign a label to each day. Formally, it uses a function F_K that assigns a label $l_i \in \mathcal{L}$ to the values $d_i \in D$ of each day by means of a clustering process, $F_K : D \rightarrow \mathcal{L}$, that is, every 24 h are identified by a label. Once K is fixed, this process transforms the daily time series D into a sequence of labels S, thus discretizing the original data. Let l_i be the label assigned to the day i obtained by means of the application of a clustering technique. Let S^i_W be the labels' subsequence of W consecutive days, from day i backward:

$$S^i_W = [l_{i-(W-1)}, l_{i-(W-2)}, \ldots, l_{i-1}, l_i] \tag{2}$$

where the length of the window, W, is a parameter to be determined.

Let W^* be the length of the window determined by PSF. For a day i and length of window W^*, the PSF algorithm searches for the subsequences of labels which are exactly equals to $S^i_{W^*}$ in the dataset, providing the equal subsequences set, ES, defined by the equation,

$$ES(i, W^*) = \left\{ \text{days } j \in D \text{ such that } S^j_{W^*} = S^i_{W^*} \right\} \tag{3}$$

It is worth remarking that if no subsequence equal to $S^i_{W^*}$ was found in the dataset, that is, $ES(i, W^*) = \emptyset$, the length of the window would decrease by one unit, $W' = W^* - 1$, and the PSF would search for subsequences equal to $S^i_{W'}$. This process may be repeated until any subsequence is found, that is, $ES(i, W) \neq \emptyset$.

Therefore, the W^* consecutive labels that precede the day to be predicted are extracted and searched for in the historical data. Once all occurrences of $S^i_{W^*}$ are found, the 24 hourly values of the day $i + 1$ are predicted by averaging the real values found immediately after each $S^i_{W^*}$ match. Mathematically,

$$\hat{d}_{i+1}(W^*) = \frac{1}{\#ES(i, W^*)} \sum_{j \in ES(i, W^*)} d_{j+1} \tag{4}$$

Finally, the daily error for any day i is defined by:

$$e_{day}(i, W^*) = |\hat{d}_i(W^*) - d_i| \tag{5}$$

4. Outlier forecasting in time series

This section explains the methodology proposed to improve the forecasting process provided by the PSF algorithm. The discovery of motifs is included in the aforementioned algorithm as a crucial step for forecasting the occurrence of outliers and, then, providing accurate estimates for such anomalous observations.

The value of two parameters had to be determined in the PSF process: The number of clusters K and the length of the window W. With regard to K, the new approach acts exactly the same as what was proposed in the original PSF, that is, it applies three well-known validity indices –Silhouette, Dunn and Davies–Bouldin– and determines the optimal number of clusters by means of a majority vote system.

On the other hand, the $n-$ fold cross-validation is used to obtain the optimal value of W. Twelve folds have been created in this work ($n = 12$) for all the datasets, where each fold represents a month. Therefore the training set consists of one year. The $12-$ fold cross–validation is then evaluated. The forecasting errors are calculated in every fold by varying the length of W. For each window size W, the monthly errors are denoted by $e_{month}(W)$ and are calculated as follows:

$$e_{month}(W) = \frac{1}{\#month} \sum_{i \in month} e_{day}(i, W) \tag{6}$$

for $W = 1, \ldots, W_{max}$ and $W_{max} = 10$, since no longer sequences were found in daily time series. Then, the average errors are calculated for each window size as follows,

$$\bar{\mathbf{e}}(W) = \frac{1}{n} \sum_{month} e_{month}(W) \tag{7}$$

where $n = 12$ and $month = \{Jan, \ldots, Dec\}$.

The W^* selected is the one that minimizes the average error corresponding to the 12 folds (months) evaluated.

$$W^* = \arg\min\{\bar{\mathbf{e}}(W)\} \text{ with } W = 1, \ldots, W_{max} \tag{8}$$

It is now –just after the training step and before the prediction process– that the discovery of motifs plays a crucial role, as it attempts at forecasting the occurrence of anomalous days in time series.

Specifically, the motifs to be found are those which generate a prediction error greater than the average error in the cross–validation process. Therefore, a set of days i belonging to the training set (TS) that satisfies $e_{day}(i, W^*) > \bar{\mathbf{e}}(W^*)$ is constructed. This set, CS or candidates set, gathers all the candidate days to be preceded by a sequence that will eventually be a motif. Formally,

$$CS = \{i \in TS \text{ such that } e_{day}(i, W^*) > \bar{\mathbf{e}}(W^*)\} \tag{9}$$

Nevertheless, not all the sequences that precede these candidates have the same probability to be eventually considered as outlier precursors, since the associated errors range from values close to the mean error (these candidate sequences should be eventually discarded by the approach) to significantly high values. For this reason, each candidate is co-labeled by using clustering techniques, more specifically, the K-means algorithm. The decision on how many clusters have to be created is always an open question and many indices could be used. However, it is worthless to have a large number of clusters and therefore, only three conceptual classes will be created: A class that gathers days with low errors (C_l) or nearest errors to the $\bar{\mathbf{e}}(W^*)$, a class containing medium errors (C_m) and, finally, a class devoted to identify high errors (C_h) or farthest errors to the $\bar{\mathbf{e}}(W^*)$. Thus, for all days $i \in C_l, j \in C_m$ and $k \in C_h$ the following inequalities are fulfilled:

$$e_{day}(k, W^*) > e_{day}(j, W^*) > e_{day}(i, W^*) > \bar{\mathbf{e}}(W^*) \tag{10}$$

Fig. 1 illustrates an imaginary error distribution in the TS, determining the candidates that will eventually form CS as the union of candidates in C_l, C_m and C_h. In other words, $CS = C_l \bigcup C_m \bigcup C_h$.

A priori reasoning reveals that those candidates belonging to C_h must be more probable to be preceded by motifs preceding outliers than the candidates in C_l or C_m. Results corresponding to each cluster of data will be separately analyzed in Section 5.

The next step consists of computing the sequences of labels occurring before the candidates in order to determine which sequences will be considered outlier precursors, since not all these sequences will be motifs. Hence, the approach has to decide whether the sequences preceding the candidates are motifs responsible for outliers or not. In particular, the sequences that only appear before the candidates will be considered motifs preceding an outlier occurrence. That is, if a sequence of labels preceding a candidate is found prior to any other day that does not belong to the CS, the sequence is discarded and not considered to be a motif. Thus, a set of motifs MS is defined by:

$$MS = \left\{ S^i_{W^*} \text{ such that } i \in CS \text{ and } ES(i, W) \subseteq CS \right\} \tag{11}$$

Note that MS can be also expressed as follows: $MS = M_l \bigcup M_m \bigcup M_h$, where M_l, M_m and M_h are the discovered motifs associated to the sequences found in the classes C_l, C_m, and C_h respectively.

Fig. 4 depicts the process of discovering the sequences of labels that represent the motifs preceding outliers. This figure shows an illustrative example in which six clusters were created, $K = 6$ (labels are digits 1 to 6). Therefore, the time series appears discretized, making only use of six different labels to be assigned one to each day. The labels in bold refer to those days initially included in CS, that is, those that obtained forecasting error greater than the average. By contrast, the labels followed by a bullet are those days that do not belong to CS but are preceded by a sequence equal to another that precedes a candidate. Then, the three labels preceding each day in CS ($W^* = 3$ in this example) are extracted. Finally, the cases in which the sequences before the candidates also appeared before any day not belonging to CS were discarded. Otherwise, these sequences are considered motifs or pattern sequences preceding an especially unexpected value. For this particular example, note that only the sequence $\{1, 5, 4\}$ would have been considered as motif.

The goals are now: (i) to forecast the occurrence of an outlier and (ii) provide accurate estimation for it. Thus, once MS is

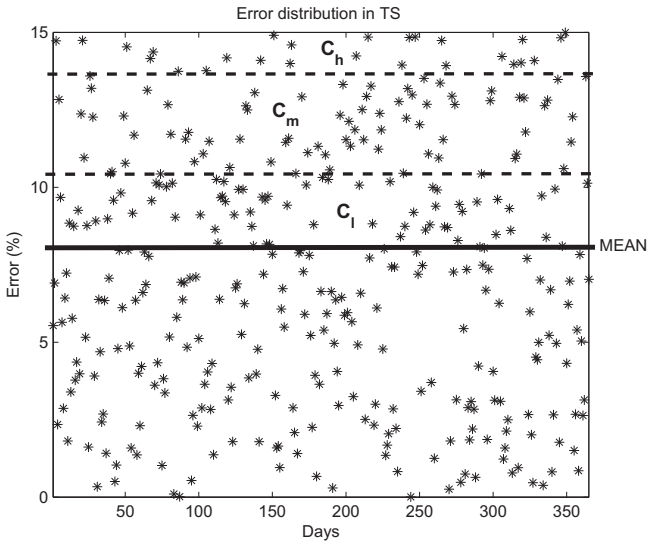


Fig. 1. Illustrative distribution of candidates in C_l, C_m and C_h.

constructed, the general scheme of forecasting is as follows. Original PSF just extracted S_{W^*} and searched for it in the historical data. But now, before its search, it has to be determined if this pattern sequence matches any of the motifs forming *MS*. Given this situation, two cases may arise:

(1) S_{W^*} does not match any of the motifs in *MS*. The approach determines that the day to be forecasted is not an outlier and it would continue with normal PSF procedure. That is, $\hat{d}_{i+1}(W^*)$ is calculated as PSF does from Eq. (4).
(2) S_{W^*} matches any of the motifs in *MS*. The approach determines that the day to be forecasted is an outlier. In this case the challenge is to provide accurate estimations for these observations, i.e. to provide accurate values for z_i. To fulfill this goal, a simple strategy is proposed: To average the value of the outliers found by the proposed approach in the historical data. Formally, the set of outlying days from the training set is defined by:

$$OS = \{i \in TS \text{ such that } S^i_{W^*} \in MS\} \tag{12}$$

Then, the forecast for the outlier is based on the a posteriori detected outliers in the training set:

$$\hat{d}_{i+1}(W^*) = \frac{1}{\#OS} \sum_{i \in OS} d_{i+1} \tag{13}$$

where $\#OS$ is the number of outliers detected by the approach in the historical data (the number of elements in *OS*), and $\hat{d}_i$ the values of the time series for these outlying days that form *OS*.

Fig. 2 illustrates the entire process of prediction when the discovery of motifs is included in the PSF algorithm. Note that this step has to be performed immediately after the clustering (creation of the sequence of labels) and before the forecasting. In addition, the steps corresponding to discovery of motifs and prediction are further detailed in Fig. 3.

Finally, the pseudocode of the proposed methodology is presented in Fig. 5, and that of the discovery of motifs process in Fig. 6.

5. Results

This section presents the results obtained by the application of the proposed methodology to six time series. The motifs discovery process to six real-world time series is described in Section 5.1. Then, a statistical analysis has been carried out to determine the validity of the assumptions made when forecasting the occurrence of outliers in the time series. This analysis can be found in Section 5.2. Finally, to compare the results obtained, Section 5.3 reports average forecasting errors of the new methodology and other techniques.

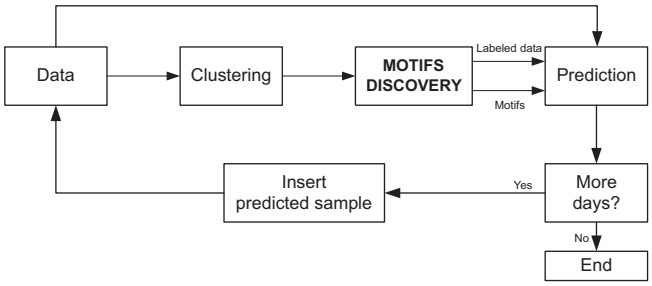


Fig. 2. Illustration of the proposed methodology.

5.1. Motifs discovery in real-world time series

The discovery of motifs on real-world time series is now described. In particular, six public electricity-related (three of prices and three of demand) time series have been considered to show that the proposed methodology properly works on different datasets. Thus, the new approach has been applied to the Spanish (OMEL), New Yorker (NYISO) and Australian (ANEM) markets, whose data are available on-line in Spanish Electricity Price Market Operator (http://www.omel.es), The New York Independent System Operator (http://www.nyiso.com) and Australia's National Electricity Market (http://www.nemmco.com.au), respectively.

The forecasting process is applied to the year 2006 for the three markets, with a historical data of one year and with a horizon of prediction of one month. As twelve months are going to be evaluated for each market, the methodology is going to be tested on 72 datasets. Given this situation, every time a month is forecasted the training set changes. For instance, when January 2006 is forecasted, the training set comprises the whole year of 2005. However, when February 2006 is forecasted the historical data ranges from February 2005 to January 2006, and so on.

These changes in the training set involve changes in the configuration of PSF. First of all, both *K* and *W* have to be determined according to the methodology presented in Section 4. Table 1 summarizes the values of these parameters for the six markets in the year 2006.

The results after the motifs extraction step for the three markets are summarized in Tables 2–4. That is, a summary of all encountered classes, sequences and motifs can be found in theses Tables for the Spanish, New Yorker and Australian markets, respectively. However, only electricity prices results of January 2006 for the Australian market are now described as the explanation for the remaining eleven months for each year and market is similar. Therefore, all the comments about the results provided below refer to prices shown in Table 4. First, the parameters to be set in the PSF are equal to: $(K,W) \leftarrow (3,6)$, according to Table 1. The *CS* can be now constructed. For this purpose, the $\bar{\mathbf{e}}(W)$ (see Eq. (7)) has to be considered since the candidates are those days belonging to the training set (January to December 2005) that obtained an error greater than $\bar{\mathbf{e}}(W)$. The value of the mean error, calculated according to the methodology in Section 4 is $\bar{\mathbf{e}}(6) = 5.81\%$. Therefore, *CS* would be formed by all days in 2005 with forecasting error greater than 5.81%.

Now the three classes are constructed by applying *K*-means, with $K = 3$ as mentioned in Section 4. The three clusters are defined as: C_l is the class that contains the candidates days with error from 5.81% to 7.13%, C_m the one that gathers the candidates with errors ranging from 7.13% to 9.47% and C_h the class that contains the candidates with errors greater than 9.47%. The error distribution, according to these three clusters, is shown in Fig. 7.

From the 365 days of 2005 that comprise the training set, 137 (see Table 4, row 1: $\#C_l + \#C_m + \#C_h = 101 + 32 + 4 = 137$) had an error greater than 5.81% so the constructed *CS* contains 137 candidates days. Once the candidates are selected, the number of different sequences that generated them are considered. From the candidates in C_l, 5 different sequences were found ($S_l = 5$); from the candidates in C_m, 3 ($S_m = 3$) and from the candidates in C_h, 2 ($S_h = 2$). This fact involves that from all the $K^W = 729$ possible sequences, only 10 caused errors greater than the average.

Note that there were situations in which a particular sequence appeared before different candidates that belong to different classes. For these cases, only the sequence that appeared in the class with higher associated error (C_l was devoted to include days with lower errors, C_m to medium errors and C_h to higher ones) was counted.

Finally, the number of motifs that identify outliers are determined. From the sequences C_l, only one appeared exclusively as

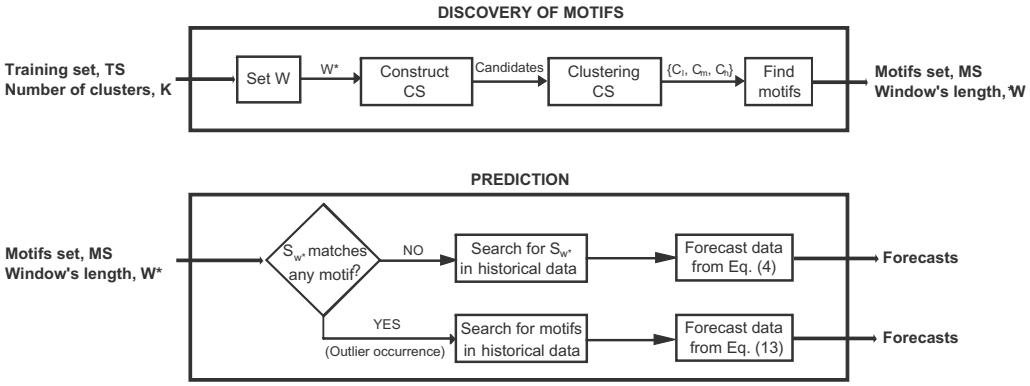


Fig. 3. Detail of discovery of motifs and prediction steps.

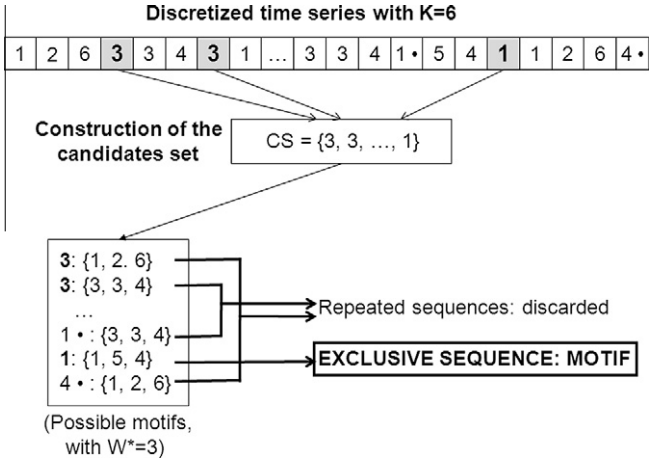


Fig. 4. Illustrative example of motifs discovery.

an outlier precursor, $M_l = 1$. With reference to the sequences in C_m, one out of three, $M_m = 1$. Last, both sequences in C_h were exclusive, $M_h = 2$.

Fig. 8 is provided to determine the usefulness of dividing the *CS* into three groups. These histograms show the motifs distribution along with C_l, C_m and C_h and, to be precise, the relation between different sequences and motifs for each class and market, expressed as %. Thus, each bar is calculated by dividing the number of different sequences that precede the candidates and the number of motifs that are finally selected. As it is possible to observe, the probability that a sequence becomes a motif is directly related with the error associated to the candidate to which it precedes. For instance, the percentage of motifs for the ANEM's electricity price time series are 20.26% for the sequences in C_l, 29.82% for the sequences in C_m and 52.94% for the sequences in C_h.

The motifs found in January 2006, represented as a numerical sequence of labels, are shown in Tables 5 and 6 corresponding to prices and demand, respectively. For instance, note that the first motif found in ANEM's prices is $M_l^1 = \{1, 3, 2, 2, 3, 1\}$. Each label is represented by one of the K clusters generated during the training of the PSF (where $K = 3$ in this case) and identifies 24 h. As the length of the window was set to $W = 6$, these six labels in fact represent 144 h.

Figs. 9 and 10 illustrate the most representative motifs found in all the markets when forecasting January 2006. Actually, these motifs represent the time series values that will precede an outlier. As each motif was represented by W consecutive labels (see Definition 6), these figures depict the values associated to every label, which have been obtained by means of clustering techniques.

Input: Dataset D, number of clusters K and test set P
Output: Forecasts $\widehat{d}_i$ for all days of P

PSF + motifs()
  $ES \leftarrow \{\}$
  $\widehat{d}_i \leftarrow 0$
  $[MS, W^*] \leftarrow$ **Motifs discovery()**
  for each day $i \in P$
    if $(S_{W^*}^{i-1} \notin MS)$
      //**Prediction according to PSF**
      Calculate the set ES from Eq. (3)
      Calculate the prediction of the day i from Eq. (4)
    else
      Calculate the set of outliers OS from Eq. (12)
      Calculate the prediction of the day i from Eq. (13)
    Insert $\widehat{d}_i$ in D
    clustering(D,K)
  return $\widehat{d}_i$

Fig. 5. A general scheme of the proposed methodology.

Input: Training set TS and number of clusters K
Output: Motifs set MS and length of window W^*

Motifs discovery()
  $MS \leftarrow \{\}$
  $CS \leftarrow \{\}$
  $\overline{\mathbf{e}}(W) \leftarrow 0$
  For each $W \in [1, 10]$
    For each day $i \in TS$
      Calculate $e_{day}(i, W)$
    Calculate $e_{month}(W)$
    $\overline{\mathbf{e}}(W) \leftarrow mean(e_{month}(W))$
  $W^* \leftarrow \arg\min\{\overline{\mathbf{e}}(W)\}$
  For each day $i \in TS$
    If $e_{day}(i, W^*) > \overline{\mathbf{e}}(W^*)$
      $CS \leftarrow$ add (i, CS)
  For each day $i \in CS$
    Obtain $S^i_{W^*}$
    If $ES(i, W^*) \subseteq CS$
      $MS \leftarrow$ add $(S^i_{W^*}, MS)$
  return MS and W^*

Fig. 6. Pseudocode of the motifs discovery.

Table 1
Setting the PSF. OMEL refers to the Spanish market, NYISO to the New Yorker market and ANEM corresponds to the Australian market.

Month	Prices						Demand					
	OMEL		NYISO		ANEM		OMEL		NYISO		ANEM	
	K	*W*	*K*	*W*	*K*	*W*	*K*	*W*	*K*	*W*	*K*	*W*
January	4	5	5	4	3	6	7	3	4	3	6	4
February	4	5	5	3	3	6	7	4	3	5	6	5
March	4	5	5	4	3	6	7	3	4	4	4	3
April	4	5	5	4	4	6	6	4	4	5	4	4
May	4	5	6	3	4	6	5	4	4	4	6	4
June	6	4	5	3	3	6	6	3	4	5	5	5
July	5	5	6	4	3	5	6	4	3	5	6	4
August	6	4	6	3	4	6	5	4	4	3	5	4
September	6	4	5	3	3	6	7	3	5	4	6	5
October	6	4	5	4	3	6	5	4	5	4	5	4
November	6	4	5	4	3	6	5	3	4	4	6	4
December	5	5	5	3	3	5	6	3	5	3	4	6

Table 2
Motifs distribution for Spanish markets.

Month	Prices			Demand		
	$\#C_l(S_l)[M_l]$	$\#C_m(S_m)[M_m]$	$\#C_h(S_h)[M_h]$	$\#C_l(S_l)[M_l]$	$\#C_m(S_m)[M_m]$	$\#C_h(S_h)[M_h]$
January	98(7)[1]	25(4)[3]	8(2)[2]	103(6)[2]	35(6)[2]	6(3)[3]
February	87(6)[0]	31(7)[2]	5(2)[2]	121(8)[1]	22(5)[3]	3(2)[1]
March	73(5)[2]	16(3)[1]	8(1)[1]	99(7)[1]	30(7)[4]	6(1)[0]
April	103(9)[1]	30(6)[1]	6(3)[2]	83(4)[0]	17(5)[2]	5(3)[3]
May	65(4)[0]	51(6)[0]	10(4)[2]	101(6)[2]	26(8)[3]	5(3)[3]
June	97(6)[0]	38(5)[2]	4(0)[0]	124(7)[3]	27(5)[1]	7(1)[1]
July	180(8)[2]	27(3)[1]	12(5)[3]	97(8)[2]	13(3)[0]	8(4)[3]
August	101(8)[3]	26(5)[0]	9(4)[2]	113(11)[4]	29(6)[3]	11(3)[2]
September	110(8)[1]	25(5)[3]	5(0)[0]	105(9)[2]	19(5)[4]	3(2)[2]
October	108(7)[1]	23(4)[2]	6(1)[0]	98(8)[3]	32(3)[2]	4(1)[1]
November	120(9)[1]	40(6)[0]	6(3)[1]	99(9)[2]	20(5)[3]	9(3)[3]
December	169(10)[2]	38(9)[3]	10(3)[1]	114(8)[2]	24(3)[2]	12(2)[0]

Regarding the electricity prices, note that for the Spanish market, six motifs were found; six for New York, and four for the Australian market. As for the electricity demand, the number of motifs found were seven, five and four for the Spanish, New York and Australian markets, respectively. In actual fact, the curves in these figures represent the average evolution of the five (OMEL), four (NYISO) and six (ANEM) days prior to an outlier forecast in prices time series and the average evolution of the three (OMEL and NYISO) and four (ANEM) days prior to an outlier forecast in demand time series.

Table 3
Motifs distribution for New York markets.

Month	Prices			Demand		
	$\#C_l(S_l)[M_l]$	$\#C_m(S_m)[M_m]$	$\#C_h(S_h)[M_h]$	$\#C_l(S_l)[M_l]$	$\#C_m(S_m)[M_m]$	$\#C_h(S_h)[M_h]$
January	101(8)[3]	34(3)[1]	12(2)[2]	96(9)[1]	28(3)[1]	14(4)[3]
February	92(11)[2]	36(4)[1]	14(3)[2]	88(9)[3]	14(2)[1]	7(2)[2]
March	89(7)[2]	45(5)[1]	11(2)[1]	101(9)[2]	33(2)[0]	9(3)[2]
April	110(13)[3]	21(4)[1]	9(5)[4]	93(8)[2]	29(3)[2]	11(4)[2]
May	121(7)[2]	31(2)[1]	6(3)[1]	114(9)[3]	30(4)[2]	13(6)[3]
June	142(5)[1]	32(0)[0]	7(0)[0]	103(7)[0]	26(3)[1]	10(4)[2]
July	92(10)[3]	41(5)[1]	18(4)[4]	86(5)[1]	29(4)[2]	8(3)[2]
August	84(7)[2]	39(6)[2]	9(6)[5]	76(7)[4]	14(2)[2]	4(1)[1]
September	107(7)[0]	40(4)[0]	10(1)[1]	84(8)[1]	21(4)[1]	8(3)[3]
October	141(9)[2]	28(3)[0]	4(0)[0]	95(6)[0]	32(4)[2]	9(3)[1]
November	99(12)[3]	32(8)[2]	15(4)[3]	115(8)[3]	38(6)[2]	15(6)[2]
December	87(8)[1]	44(3)[1]	8(0)[0]	109(8)[3]	29(4)[2]	12(4)[3]

Table 4
Motifs distribution for Australian markets.

Month	Prices			Demand		
	$\#C_l(S_l)[M_l]$	$\#C_m(S_m)[M_m]$	$\#C_h(S_h)[M_h]$	$\#C_l(S_l)[M_l]$	$\#C_m(S_m)[M_m]$	$\#C_h(S_h)[M_h]$
January	101(5)[1]	32(3)[1]	4(2)[2]	131(7)[0]	46(7)[2]	5(3)[2]
February	165(5)[0]	25(6)[1]	9(1)[1]	115(6)[1]	56(7)[3]	6(3)[3]
March	133(8)[2]	13(2)[0]	3(0)[0]	92(6)[0]	72(9)[1]	7(3)[3]
April	190(13)[3]	8(2)[0]	11(4)[2]	102(7)[2]	64(5)[0]	5(4)[3]
May	187(17)[5]	13(4)[2]	6(3)[1]	123(7)[2]	41(3)[1]	11(4)[0]
June	169(12)[5]	22(5)[1]	3(3)[1]	183(16)[3]	33(3)[0]	4(4)[2]
July	172(22)[3]	9(0)[0]	2(2)[2]	117(9)[1]	40(5)[0]	8(5)[3]
August	142(20)[2]	34(3)[2]	6(4)[2]	139(11)[2]	41(4)[3]	7(3)[1]
September	102(18)[4]	20(8)[3]	12(2)[1]	110(8)[2]	27(3)[2]	9(3)[3]
October	81(13)[2]	53(13)[4]	9(2)[0]	142(12)[3]	42(8)[3]	5(1)[1]
November	112(9)[1]	43(7)[1]	14(5)[4]	137(12)[2]	38(7)[3]	7(3)[2]
December	121(11)[3]	39(4)[2]	13(6)[2]	134(10)[4]	29(5)[2]	8(4)[2]

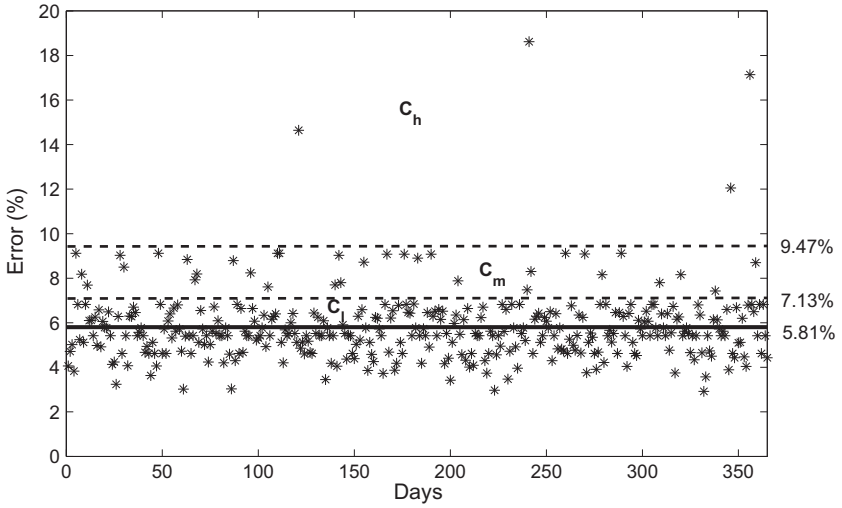
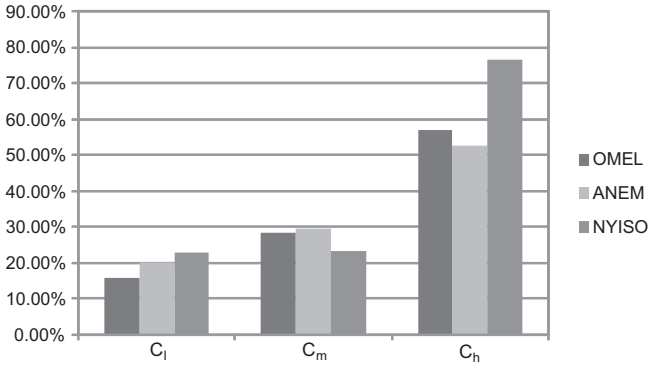


Fig. 7. Forecasting errors in TS distributed in C_l, C_m and C_h obtained by applying the K-means to the candidate set CS with $K = 3$.

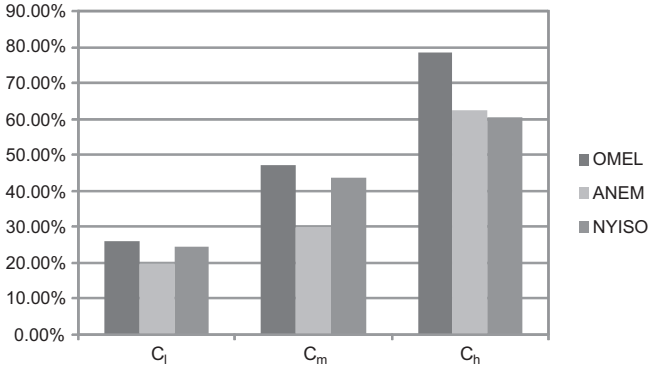
Fig. 11 shows ANEM's electricity prices for July 2006. This month is illustrated since it presents large outliers. As for this month $W = 5$ (see Table 1), the number of labels considered for outlier occurrence forecasting is five. Also, the number of motifs found in TS were five (see Table 4): $M_l^1 = \{1,2,3,3,1\}, M_l^2 = \{2,2,3,1,3\}, M_l^3 = \{3,1,1,1,1\}, M_h^4 = \{2,1,3,3,1\}$ and $M_h^5 = \{2,1,1,1,3\}$ Thus, grey bars represent outliers a posteriori detected by means of the robust statistical method presented in Gelper et al. (2010), that is, days 3, 13, 18 and 22 July were the outliers identified. It can be appreciated that sequences M_l^3, M_h^4 and M_l^1 are three of the motifs found in TS that eventually preceded days 13, 18 and 22 July, respectively. Furthermore, 3rd July was preceded by the motif M_h^5. Only the last two labels of this motif (1,3) are depicted in Fig. 11 because the first three labels (2,1,1) correspond to days in June. Finally, note that one of the motifs found in TS, M_l^2, did not occur when forecasting July 2006.

5.2. Evaluation of outlier occurrence forecasting

Once all the *MS* have been constructed, the proposed method predict a priori if the day to be forecasted will be an outlier. This Section is devoted to statistically quantify the outlier occurrences



(a) Price time series



(b) Demand time series

Fig. 8. Motifs distribution in C_l, C_m and C_h.

Table 5
Motifs found in TS for the electricity price when forecasting January 2006.

Market	Motif
OMEL	$M_l^1 = \{1,1,3,4,4\}$
	$M_m^2 = \{3,4,4,1,3\}$
	$M_m^3 = \{3,4,4,2,1\}$
	$M_m^4 = \{2,3,4,4,1\}$
	$M_h^5 = \{1,1,1,2,3\}$
	$M_h^6 = \{3,2,3,3,4\}$
NYISO	$M_l^1 = \{1,3,5,4\}$
	$M_l^2 = \{2,2,3,1\}$
	$M_l^3 = \{5,1,1,3\}$
	$M_m^4 = \{3,4,2,5\}$
	$M_h^5 = \{3,2,4,1\}$
	$M_h^6 = \{3,1,5,1\}$
ANEM	$M_l^1 = \{1,3,2,2,3,1\}$
	$M_m^2 = \{2,1,3,3,3,1\}$
	$M_h^3 = \{3,2,1,1,2,3\}$
	$M_h^4 = \{3,3,3,1,1,2\}$

properly forecasted by the proposed methodology. Thus, the quality parameters used to evaluate the accuracy of the approach are first introduced in Section 5.2.1 and, then, the conducted statistical analysis is reported in Section 5.2.2.

5.2.1. Parameters of quality

The parameters used to assess the accuracy of the approach are now introduced. A posteriori analysis has been carried out to

Table 6
Motifs found in TS for the electricity demand when forecasting January 2006.

Market	Motif
OMEL	$M_l^1 = \{6,4,4\}$
	$M_l^2 = \{7,1,7\}$
	$M_m^3 = \{5,5,4\}$
	$M_m^4 = \{4,7,5\}$
	$M_h^5 = \{4,4,7\}$
	$M_h^6 = \{6,5,6\}$
	$M_h^7 = \{7,7,2\}$
NYISO	$M_l^1 = \{2,3,3\}$
	$M_m^2 = \{4,3,4\}$
	$M_h^3 = \{1,3,2\}$
	$M_h^4 = \{3,4,1\}$
	$M_h^5 = \{2,2,3\}$
ANEM	$M_m^1 = \{6,5,3,4\}$
	$M_m^2 = \{2,6,3,4\}$
	$M_h^3 = \{2,3,1,4\}$
	$M_h^4 = \{4,6,6,5\}$

determine the existence of outliers in all examined series. In particular, the robust method proposed in Gelper et al. (2010) (hereafter called RHW for simplicity) to detect outliers in time series has been considered. Hence, a forecast of an outlier occurrence is said to be properly made by the proposed approach, if RHW also points the observation as anomalous. Thus, a priori forecasting (the proposed approach in this work) is compared to a posteriori detection (the method proposed in Gelper et al. (2010)).

Note that the authors in Gelper et al. (2010) determined that outliers are those data that do not fulfil any of the two bounds they define:

$$UB_t = \hat{y}_t + 2\hat{\sigma} \tag{14}$$

$$LB_t = \hat{y}_t - 2\hat{\sigma} \tag{15}$$

where UB_t and LB_t are the upper and lower bounds respectively, $\hat{y}_t$ is the fitted value, and $\hat{\sigma}$ is the standard deviation of the regression residuals they obtain.

Hence, in subsequent equations, true positives or *TP* is the number outlier occurrences properly forecasted, that is, the number of days preceded by a motif in *MS* that are outliers according to RHW; true negatives or *TN* is the number of days that were not preceded by a motif in *MS* and were not considered outlier by RHW either; false positives or *FP* is the number of days preceded by a motif in *MS* that were not considered outlier by RHW; and false negatives or *FN* is the number days not considered outliers (not preceded by any motif in *MS*) and eventually considered outliers by RHW.

According to these definitions, the sensitivity is the probability that a motif discovered precedes a real outliers. Its formula is defined as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{16}$$

Another relevant parameter is the specificity, or the ratio of sequences preceding the day to be forecasted properly discarded by the approach. The mathematical expression is:

$$Specificity = \frac{TN}{TN + FP} \tag{17}$$

The positive predictive value (PPV) is the probability that a forecasted outlier is indeed a real one. Its formula is:

$$PPV = \frac{TP}{TP + FP} \tag{18}$$

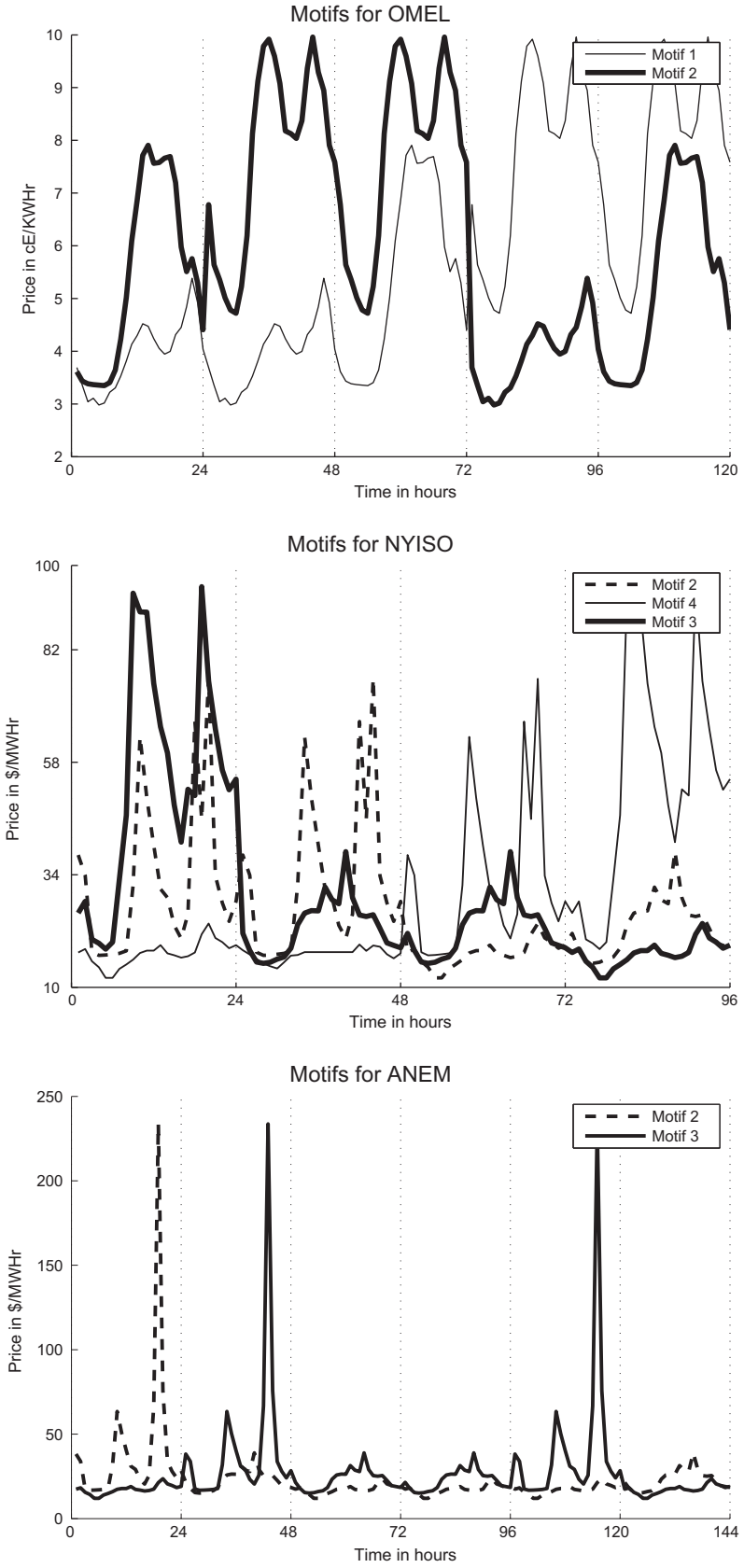


Fig. 9. Motifs found in January 2006 for electricity prices.

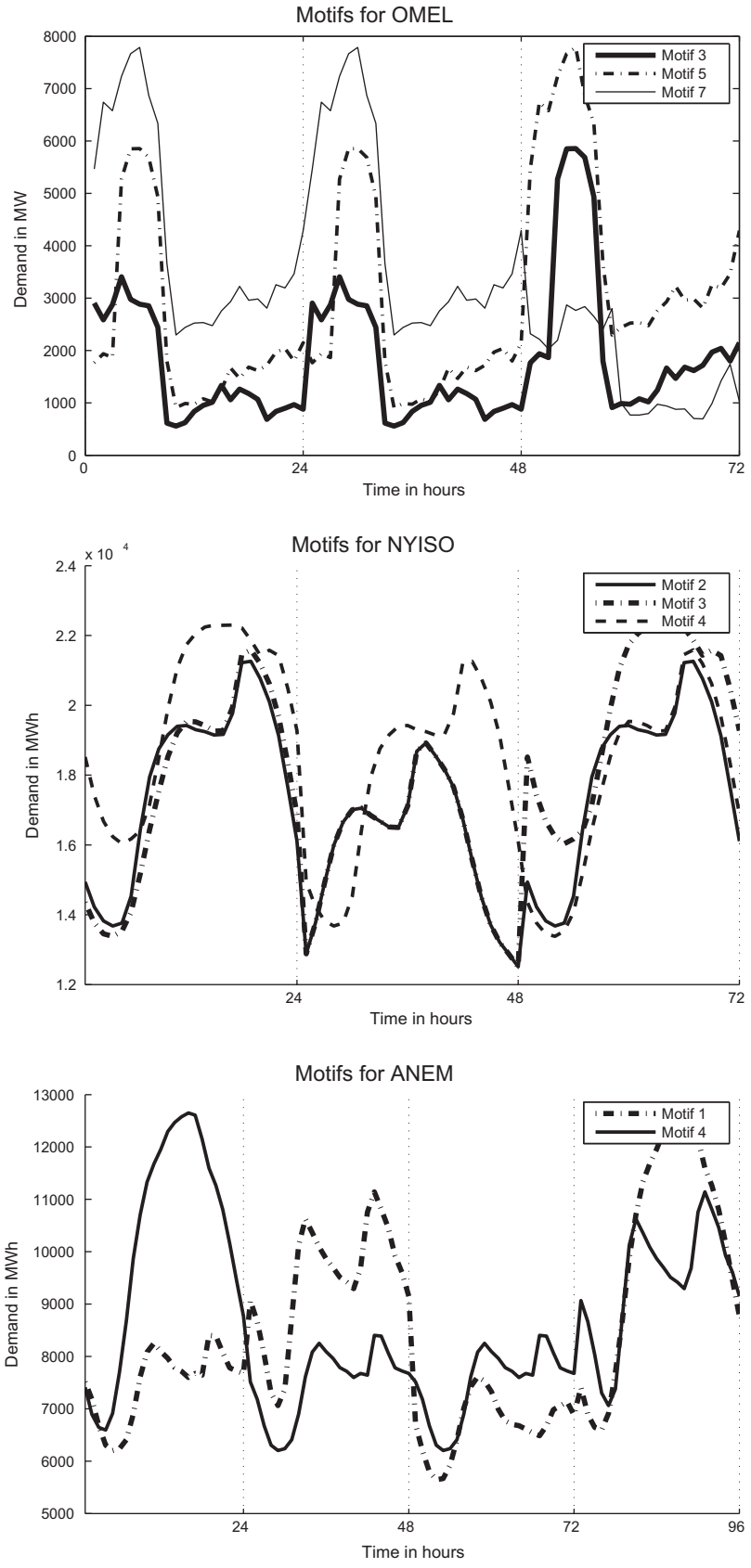


Fig. 10. Motifs found in January 2006 for electricity demand.

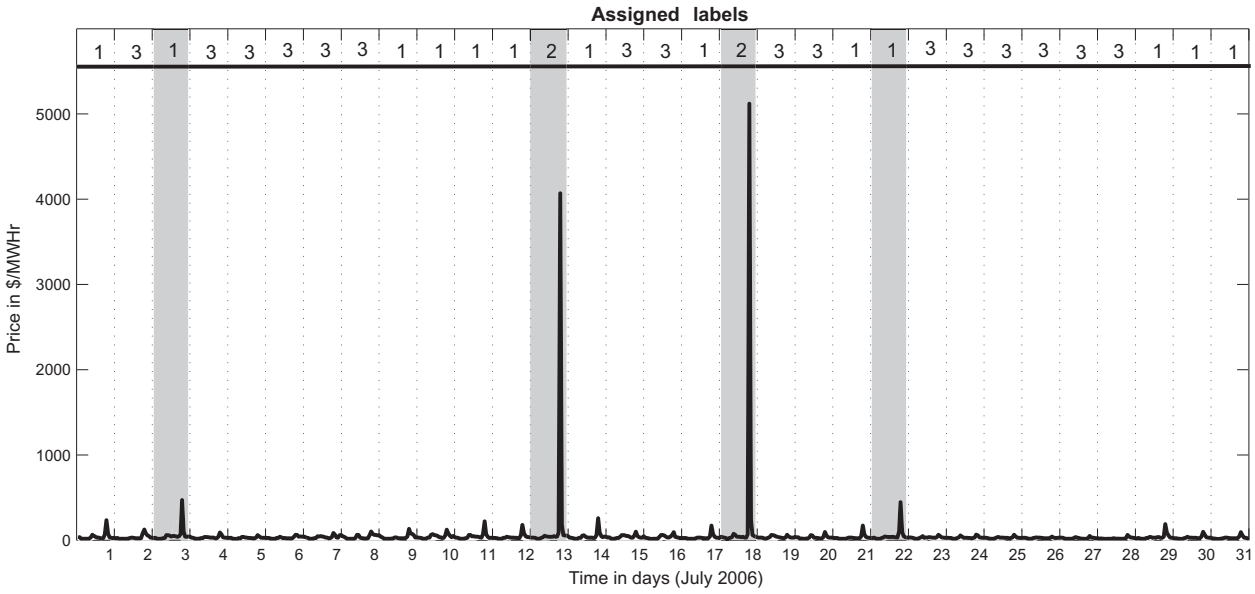


Fig. 11. Motifs discovery for outliers detection in ANEM for July 2006.

Finally, the negative predictive value (NPV) is the probability that a discarded sequence was not indeed a real motif preceding an outlier. Its formula is:

$$NPV = \frac{TN}{TN + FN} \tag{19}$$

5.2.2. Quantifying the forecast of outlier occurrences

A statistical measure of the accuracy is provided in this Section. The parameters used are the ones described in Section 5.2.1 and summarized in Table 7. Note that all parameters refer to the whole year of 2006 for the three markets, that is, the numbers gather the twelve forecasted sets or months for both prices and demand. Only the explanation for the prices in the year 2006 of the Australian market is provided, since the reasoning for the remaining markets is analogous.

During the forecasting process, there were 75 outliers properly detected, that is, $TP = 75$. On the other hand, there were thirteen occasions in which a motif appeared and did not precede a day considered an outlier by RHW. Therefore, $FP = 13$. Furthermore, there were seventeen more days which were outliers according to RHW and were not preceded by a motif, that is, $FN = 17$. And the remaining 260 days were not preceded by motifs and were not outliers by RHW. Thus, TN=260 (75 + 13 + 17 + 260 = 365 forecasted days in 2006).

For instance, in the Fig. 11 (ANEM's electricity prices in July 2006), four TP were found (days 3,13,18,22), that is, RHW determined that these days were outliers, as well as the new approach did. On the contrary, no FP were found, since there were no motifs followed by a non-outlier day (FP = 0). As for FN, RHW determined that day 14 was an outlier, contrarily to what the new approach determined, that it considered it a regular day as this day not was preceded by any of the motifs found during the training phase (FN = 1). Finally, the remaining 26 days are TN.

The results show great accuracy for all markets. Especially remarkable are the values reached by the specificity. In particular, it exceeds 95% in all markets (except for OMEL prices, in which it reaches 93.55%), obtaining 95.27% on average. These values mean that when the approach classifies the day to be predicted as outlier, it does it with high reliability.

As for the sensitivity, all the markets reached values greater than 80% (except for demand in NYISO that almost reached it: 79.78%), obtaining 82.13% on average. This fact implies that when a motif is found before the day to be forecasted, it is highly probable that such a day will be an outlier.

The PPV reached values similar to those of the sensitivity, in particular, slightly greater (83.67% on average). Therefore, it can be stated that when the proposed method determines that there is an upcoming outlier, it is highly reliable.

Finally, NPV provides similar values for all the six series time, reaching 94.76% on average, that is, the rate of real outliers not found by the approach cannot be considered significant.

Table 7
Statistical analysis of the method.

Parameters	Prices			Demand			Mean
	OMEL	NYISO	ANEM	OMEL	NYISO	ANEM	
TP	73	65	75	67	71	58	68.17
TN	261	271	260	272	267	281	268.67
FP	18	14	13	14	9	12	13.33
FN	13	15	17	12	18	14	14.83
Sensitivity	84.88%	81.25%	81.52%	84.81%	79.78%	80.56%	82.13%
Specificity	93.55%	95.09%	95.24%	95.10%	96.74%	95.90%	95.27%
PPV	80.22%	82.28%	85.23%	82.72%	88.75%	82.86%	83.67%
NPV	95.26%	94.76%	93.86%	95.77%	93.68%	95.25%	94.76%

5.3. Time series forecasting in presence of outliers

This Section shows the prediction errors obtained by the proposed approach. Moreover, the results obtained are compared to those of the original PSF, in which the motifs discovery step was not performed, and to those of the RHW robust statistical method (Gelper et al., 2010). Further comparison can also be found in Martínez-Álvarez et al. (in press), where comparisons between the original PSF and other methods were reported.

Note that RHW only provides one sample-ahead estimations. Instead, the new approach provides one day-ahead estimations, i.e. 24 samples-ahead. In this sense, to compare both techniques would not be fair as the latter has a larger horizon of prediction and, therefore, a more difficult task to fulfil. However, as robust

statistical methods are devoted to forecast time series under the presence of outliers, the results of RHW are provided in order to compare to the proposed methodology.

To evaluate the accuracy of the forecasting, different criteria may be taken into consideration. However, two parameters widely used –the mean relative error (*MRE*) and its standard deviation (σ_{MRE})– are considered to evaluate the results of prediction.

Table 8 reports the results for the OMEL market. Compared to the former PSF version, note that the new approach improves the forecasting in all the datasets considered except in April for prices and in January, June, July and October for demand. This fact is due to the absence of motifs found when these months were forecasted and, therefore, the prediction was exactly equal to the original PSF. As for RHW, the proposed methodology outperforms its predictions for every month, again, in terms of MRE and σ_{MRE}.

Tables 9 and 10 are analogous to Table 8 and report the results for the New Yorker and Australian markets, respectively. It can be observed that during certain months no motifs were found and the

Table 8
Forecast for the OMEL time series.

Month	Prices						Demand					
	RHW		PSF		PSF + motifs		RHW		PSF		PSF + motifs	
	MRE (%)	σ_{MRE}	MRE (%)	σ_{MRE}	MRE (%)	σ_{MRE}	MRE (%)	σ_{MRE}	MRE (%)	σ_{MRE}	MRE (%)	σ_{MRE}
January	8.82	0.20	7.26	0.25	6.31	0.21	4.02	2.23	3.12	1.86	3.12	1.86
February	7.67	0.28	4.93	0.19	4.28	0.16	4.89	3.07	4.21	2.26	4.09	2.12
March	7.34	0.23	5.88	0.22	5.47	0.19	4.87	3.82	5.07	4.17	3.03	3.25
April	4.08	0.14	3.62	0.18	3.62	0.18	5.71	2.74	4.18	1.28	3.86	1.19
May	10.12	0.34	8.11	0.21	5.57	0.19	5.60	1.87	5.90	2.33	2.61	1.62
June	4.46	0.24	3.76	0.24	3.11	0.24	4.23	2.36	2.89	1.81	2.89	1.81
July	5.91	0.22	4.30	0.23	4.25	0.23	4.87	2.40	2.34	1.19	2.34	1.19
August	5.03	0.28	5.37	0.34	4.62	0.27	6.91	2.21	3.61	2.17	2.93	1.76
September	6.60	0.33	6.41	0.31	6.38	0.30	5.37	2.91	3.15	1.55	3.12	1.54
October	7.14	0.39	7.89	0.29	6.02	0.22	4.56	2.99	2.89	3.40	2.89	3.40
November	11.49	0.41	8.30	0.40	5.72	0.29	5.12	3.08	4.72	2.39	3.27	1.43
December	9.71	0.37	8.02	0.36	5.19	0.31	7.94	3.43	6.21	3.82	4.36	2.51
Average	7.36	0.29	6.15	0.27	5.05	0.23	5.34	2.76	4.02	2.35	3.21	1.97

Table 9
Forecast for the NYISO time series.

Month	Prices						Demand					
	RHW		PSF		PSF + motifs		RHW		PSF		PSF + motifs	
	MRE (%)	σ_{MRE}	MRE	σ_{MRE}	MRE	σ_{MRE}	MRE (%)	σ_{MRE}	MRE (%)	σ_{MRE}	MRE (%)	σ_{MRE}
January	6.21	3.17	4.45	2.07	4.45	2.07	7.12	2.40	5.05	1.95	4.98	2.21
February	7.37	5.56	5.53	1.52	5.21	1.50	6.63	2.38	6.88	2.62	4.28	2.16
March	7.81	6.12	6.30	2.52	6.30	2.48	7.17	2.42	5.31	2.42	5.07	2.19
April	4.74	2.03	4.94	1.47	3.95	1.46	5.09	2.33	4.97	2.22	4.97	2.22
May	9.06	4.61	7.59	2.13	5.01	1.84	6.41	2.39	6.18	2.39	6.16	2.19
June	3.66	1.88	3.34	1.92	3.34	1.92	3.24	2.28	3.75	2.66	3.67	2.24
July	4.18	2.51	3.93 -	1.68	3.62	1.65	4.68	2.29	3.41	1.78	3.40	2.23
August	8.27	5.87	5.37	1.87	3.75	1.41	6.30	2.37	3.99	2.13	3.80	2.27
September	7.20	4.54	6.24	1.74	5.49	1.70	7.47	2.44	4.83	2.16	4.41	2.30
October	8.97	3.97	7.43	2.33	4.87	1.69	8.95	2.57	5.37	2.25	4.97	2.22
November	6.24	2.77	5.19	2.09	5.19	2.09	10.21	2.53	4.86	1.99	4.12	2.29
December	6.00	3.63	6.04	1.99	5.31	1.86	8.84	2.55	6.80	2.40	5.74	2.31
Average	6.64	3.89	5.53	1.94	4.71	1.81	6.84	2.41	5.12	2.25	4.63	2.24

Table 10
Forecast for the ANEM time series.

Month	Prices						Demand					
	RHW		PSF		PSF + motifs		RHW		PSF		PSF + motifs	
	MRE (%)	σ_{MRE}	MRE (%)	σ_{MRE}	MRE (%)	σ_{MRE}	MRE (%)	σ_{MRE}	MRE (%)	σ_{MRE}	MRE (%)	σ_{MRE}
January	6.42	3.38	5.58	1.34	5.21	1.31	5.53	4.28	4.74	3.54	3.88	3.31
February	8.32	4.82	8.59	3.24	6.27	2.81	6.12	3.09	4.98	2.98	4.34	2.63
March	8.67	3.01	7.84	2.98	5.01	2.34	6.47	6.82	5.02	5.27	4.12	4.99
April	13.54	4.64	9.92	3.90	9.92	3.90	6.82	5.73	6.03	7.46	4.29	3.79
May	15.04	6.39	12.85	4.03	9.16	3.17	5.01	3.44	4.17	2.72	4.17	2.72
June	26.31	11.08	22.04	12.34	12.32	9.32	5.79	4.91	5.67	3.84	5.67	3.84
July	19.22	9.77	17.11	10.58	13.91	7.33	5.26	5.83	4.91	5.84	4.05	5.02
August	11.65	4.83	11.71	5.08	8.49	4.68	5.80	6.37	5.88	6.01	5.15	5.51
September	10.83	3.91	8.23	2.45	8.23	2.45	6.63	3.62	3.99	2.74	3.99	2.74
October	9.76	3.90	7.66	2.89	7.66	2.89	4.91	3.51	4.04	3.34	4.02	3.34
November	6.79	2.86	6.76	1.94	5.73	1.61	5.33	5.74	6.12	5.90	4.87	3.92
December	7.31	3.16	6.42	2.01	6.42	2.01	5.87	4.67	3.91	3.22	3.91	3.22
Average	11.99	5.15	10.39	4.40	8.19	3.65	5.80	4.83	4.96	4.41	4.37	3.75

results are equal to those of PSF. These months are: January, March, June and November for NYISO prices; April for NYISO demand; April, September, October and December for the ANEM prices; and May, June, September and December for the ANEM demand.

Furthermore, the analysis of these tables leads to two main considerations. First, the greater the average error is in both PSF and RHW, the better the methodology works since outliers are usually involved in high rates of error. This fact can be appreciated, for instance, from May to July in ANEM prices, where RHW reported *MRE* = 20.19% and PSF 17.33%, while the new methodology obtained *MRE* = 11.80%, on average.

Second, the reduction of the σ_{MRE} is equally remarkable for all the markets (from 0.27 to 0.23, from 1.94 to 1.81 and from 4.40 to 3.65 for the prices in the Spanish, New Yorker and Australian markets, respectively and from 2.35 to 1.97 for the demand in the Spanish market, from 2.25 to 2.24 for the demand in the New Yorker one, and from 4.41 to 3.75 for the Australian demand market) which leads to robust forecasts.

6. Conclusions

The improvement of an existing technique has been used in order to, first, forecast the occurrence of outliers in time series and, second, to provide accurate estimations for these outlying data. The original approach –the PSF algorithm– was based on finding similar patterns in time series. However, its application to any kind of time series revealed that there were some samples that cannot be properly forecasted.

In particular, a step devoted to discover motifs in sequences has been included. The discovery of motifs has been carried out not only for providing accurate predictions for these samples, but for indicating that it is highly probable that an outlier occurs. The method has been successfully tested on 72 sets of the Spanish, Australian and New York electricity price and demand time series (36 of price and 36 of demand).

Future work is directed towards finding not only the days that are going to present anomalous behavior, but also the days whose prediction is going to be especially accurate. In addition, a relaxation for the rule that decides if a given sequence is a motif or not is intended to be created with the aim of creating larger candidates sets that might bring more information.

Acknowledgements

The authors want to acknowledge the invaluable help provided by Prof. R. A. Maronna regarding robust statistical methods. The authors want also to thank Prof. S. Gelper for having provided useful material to make robust predictions. The research was partially supported by the Spanish Ministry of Science and Technology under project TIN2007-68084-C-00, and Junta de Andalucía under project P07-TIC-02611.

References

Aggarwal, S.K., Saini, L.M., Kumar, A., 2009. Electricity price forecasting in deregulated markets: A review and evaluation. Internat. J. Electr. Power Energy Syst. 31 (1), 13–22.

Badran, I., El-Zayyat, H., Halasa, G., 2008. Short-term and medium-term load forecasting for Jordan's power system. Am. J. Appl. Sci. 5 (7), 763–768.

Camps-Valls, G., Martínez-Ramón, M., Rojo-Álvarez, J.L., Soria-Oliva, E., 2004. Robust gamma-filter using support vector machines. Neurocomputing 62, 493–499.

Carnero, M.A., Peña, D., Ruiz, E., 2007. Effects of outliers on the identification and estimation of GARCH models. J. Time Series Anal. 28 (4), 471–497.

Chuang, C.C., Lee, Z.J., 2011. Hybrid robust support vector machines for regression with outliers. Appl. Soft Comput. 11 (1), 64–72.

Croux, C., Gelper, S., Mahieu, K., 2010. Robust exponential smoothing of multivariate time series. Comput. Stat. Data Anal. 54, 2999–3006.

Daszykowski, M., Serneels, S., Kaczmarek, K., Espen, P.V., Croux, C., Walczak, B., 2007. TOMCAT: A MATLAB toolbox for multivariate calibration techniques. Chemomet. Intel Labor. Syst. 85 (2), 269–277.

El-Telbany, M., El-Karmi, F., 2008. Short-term forecasting of jordanian electricity demand using particle swarm optimization. Elect. Power Syst. Res. 78, 425–433.

Everitt, B.S., 2006. The Cambridge Dictionary of Statistics. Cambridge University Press.

Fuchs, E., Gruber, T., Nitschke, J., Sick, B., 2009. On-line motif detection in time series with Swift Motif. Pattern Recognit. 42 (11), 3015–3031.

Galeano, P., Peña, D., Tsay, R.S., 2006. Outlier detection in multivariate time series by projection pursuit. J. Amer. Statist. Assoc. 101 (474), 645–669.

García-Martos, C., Rodríguez, J., Sánchez, M.J., 2007. Mixed models for short-run forecasting of electricity prices: application for the Spanish market. IEEE Trans. Power Syst. 22 (2), 544–552.

Gelper, S., Fried, R., Croux, C., 2010. Robust forecasting with exponential and holt-winters smoothing. J. Forecast. 29, 285–300.

Hu, J., Kihara, D., 2006. EMD: An ensemble algorithm for discovering regulatory motifs in DNA sequences. BMC Bioinformat. 7 (342), 4899–4913.

Hyndman, R.J., Fan, S., 2010. Density forecasting for long-term peak electricity demand. IEEE Trans. Power Syst. 25 (2), 1142–1153.

Ismail, Z., Yahya, A., Mahpol, K.A., 2009. Forecasting peak load electricity demand using statistics and rule based approach. Am. J. Appl. Sci. 6 (8), 1618–1625.

Lin, J., Keogh, E., Lonardi, S., Patel, P., 2002. Finding motifs in time series. In: Proc. 2nd Workshop on Temporal Data, pp. 493–498.

Lu, X., Dong, Z.Y., Li, X., 2005. Electricity market price spike forecast with data mining techniques. Electr. Power Syst. Res. 73 (1), 19–29.

Maronna, R.A., Martin, R.D., Yohai, V.J., 2007. Robust Statistics: Theory and Methods. Wiley.

Martínez-Álvarez, F., Troncoso, A., Riquelme, J.C., Aguilar-Ruiz, J.S., Energy time series forecasting based on pattern sequence similarity, IEEE Trans. Knowledge Data Eng., in press doi:10.1109/TKDE.2010.227.

Mueen, A., Keogh, E., 2010. Online discovery and maintenance of time series motifs. In: Proc. ACM SIGKDD Internat. Conf. Knowledge Discovery and Data Mining, pp. 1089–1098.

Mueen, A., Keogh, E., Zhu, Q., Cash, S., Westover, B., 2009. Exact discovery of time series motifs. In: Proc. SIAM Internat. Conf. Data Mining, pp. 453–461.

Muler, N., Peña, D., Yohai, V.J., 2009. Robust estimation for ARMA models. Ann. Statist. 37 (2), 816–840.

Pino, R., Parreno, J., Gómez, A., Priore, P., 2008. Forecasting next-day price of electricity in the Spanish energy market using artificial neural networks. Eng. Appl. Artif. Intel. 21 (1), 53–62.

Plazas, M.A., Conejo, A.J., Prieto, F.J., 2005. Multimarket optimal bidding for a power producer. IEEE Trans. Power Syst. 20 (4), 2041–2050.

Rousseeuw, P.J., Hubert, M., 2011. Robust statistics for outlier detection. WIREs Data Min. Knowl. Disc. 1 (1), 73–79.

Saini, L.H., 2008. Peak load forecasting using bayesian regularization, resilient and adaptive backpropagation learning based artificial neural networks. Electr. Power Syst. Res. 78 (7), 1302–1310.

Sharov, A.A., Minoru, S.H., 2009. Exhaustive search for over-represented DNA sequence motifs with C is Finder. DNA Res. 16, 261–273.

Stormo, G.D., 2000. DNA binding sites: Representation and discovery. Bioinformatics 16 (1), 16–23.

Tanaka, Y., Iwamoto, K., Uehara, K., 2005. Discovery of time-series motif from multi-dimensional data based on mdl principle. Mach. Learn. 58 (3), 269–300.

Tang, H., Liao, S.S., 2008. Discovering original motifs with different lengths from time series. Knowl.-Based Syst. 21, 666–671.

Troncoso, A., Riquelme, J.M., Riquelme, J.C., Gómez, A., Martínez, J.L., 2004. Time-series prediction: Application to the short-term electric energy demand. Lect. Notes Artif. Intel. 3040, 577–586.

Troncoso, A., Riquelme, J.C., Riquelme, J.M., Martínez, J.L., Gómez, A., 2007. Electricity market price forecasting based on weighted nearest neighbours techniques. IEEE Trans. Power Syst. 22 (3), 1294–1301.

Valderrama, M.J., 2008. An overview to modelling functional data. Comput. Statist. 22 (3), 331–334.

Verboven, S., Hubert, M., 2005. LIBRA: A MATLAB library for robust analysis. Chemometr. Intel. Labor. Syst. 75 (2), 127–136.

Wang, J., Wang, L., 2008. A new method for short-term electricity load forecasting. Trans. Inst. Measure. Control 30 (3), 331–344.

Wang, L., Chng, E.S., Li, H., 2010. A tree-construction search approach for multivariate time series motifs discovery. Pattern Recognition Lett. 31, 869–875.

Weron, R., Misiorek, A., 2008. Forecasting spot electricity prices: A comparison of parametric and semiparametric time series models. Internat. J. Forecast. 24 (4), 744–763.

Wijaya, E., Rajaraman, K., Yiu, S.M., Sung, W.K., 2007. Detection of generic spaced motifs using submotif pattern mining. Bioinformatics 23 (12), 1476–1485.

Wu, W., Zhou, J., Mo, L., Zhu, C., 2006. Forecasting electricity market price spikes based on bayesian expert with support vector machines. Lect. Notes Comput. Sci. 4093, 205–212.

Zhao, J.H., Dong, Z.Y., Li, X., Wong, K.P., 2007. A framework for electricity price spike analysis with advanced data mining methods. IEEE Trans. Power Syst. 22 (1), 376–385.