EXTRACTING FORMATIONS FROM LONG FINANCIAL TIME SERIES USING DATA MINING

STELLA KARAGIANNI¹, THANASIS SFETSOS² AND COSTAS SIRIOPOULOS³

ABSTRACT:
Technical analysis has become a custom decision support tool for traders and analysts, though not widely accepted by the academic community. It is based on the identification of a series of well-defined formations appearing over irregular intervals. The same principle forms the basis for the application of data mining methodologies as a tool to discover hidden patterns that exist in a time series, which is achieved by a detailed breakdown of historic information. This paper introduces a methodology for the discovery of formations that exist within a time series and have high probability of reoccurrence. The methodology was developed in an efficient manner requiring only a small number of user-specified parameters. Its two main stages are (a) a modified bottom-up segmentation algorithm with an optimization stage to reach the optimal number of segments, and (b) a rule extraction algorithm. The developed methodology is tested on two major financial series, the daily closing values of the SP500 Index and the GB Pound to US Dollar exchange rates.

JEL-CLASSIFICATION: C22, C63, F31.

KEYWORDS: technical analysis, data mining, exchange rates, stock market, pattern recognition, rule extraction.

¹ Department of Economics, University of Macedonia, 156, Egnatia st., 54006 Thessaloniki, Greece.
² EREL, INTRP, NCSR Demokritos, 15310 Aghia Paraskevi, Greece. Email: ts@ipta.demokritos.gr
³ Dept. of Business Administration, University of Patras, 26500, Patra, Greece. Email: siriopoulos@eap.gr
INTRODUCTION

Recent empirical studies in the international literature have documented substantial evidence on the predictability of asset returns and foreign exchange markets using technical analysis trading rules (Brock et al, 1992; Levich & Thomas, 1993; Chan et al, 1996; Bessembinder & Chan, 1998; Allen & Karjalainen, 1999; Lo et al, 2000; Fang & Xu, 2003). These studies report that asset returns are correlated, and hence, some degree of predictability can be captured, by technical trading rules, by time series models or their combination. In this work a data-driven methodology is developed to discover formations, of non-constant length, that appear in a univariate financial data set. The methodology was developed to require only a limited number of parameters from a potential future user.

Technical analysis is defined as the attempt to identify regularities in the time series of price and volume information from a financial market, by extracting patterns from noisy data (Lo, 2000). Analysts use price charts to study market action as a means to predict future price trends. They believe that prices move in trends, and that history (price patterns) repeats itself (Murphy, 1986). Price patterns are pictures or formations, that can be classified into different categories, and have predictive value. Although there are potentially an infinite number of price formations, the technical analysis literature (Bulkowski, 2000) suggests that certain price formations are widely identified in stock and foreign exchange markets.

A step beyond technical analysis is the application of data mining to identify previously unknown formations in a time series. This technique also provides the opportunity to generate association rules that linguistically describe the examined process. Weiss and Indurkhya (1998) define data mining as “the search for valuable information in large volumes of data. Predictive data mining is a search for very strong patterns in big data sets that can generalize to take accurate future decisions”. There are two main types of data mining applications in finance. The first is that some predefined structure of the series exists and an indexing approach aims to identify similar behaviour within the dataset. The alternative is to apply theses methodologies to discover previously unknown patterns within the series and examine reoccurrence characteristics for prediction purposes.

In the literature there are several techniques developed to analyse financial data and extract information about their future behaviour. Zeng et al (2001) proposed a hybrid approach for the prediction of stock market trend using pattern recognition and classification. A set of vectors describing future patterns were introduced which were classified based with a probabilistic relaxation algorithm. The approach was tested on the Air-Quantas stock price and a 68% success was reported. Leigh (2002) introduced a method for template matching to predict future trends of a series. The nature of the produced results resembled those produced from technical analysis. On the NYSE composite index he reported similar success in the prediction of a 5-day pattern.

Das et al (1998) proposed an approach for the discovery of rules from sequential observations, using discrete representation. The rules included local relationships in the series using a finite number of primitive shapes, initially found from a greedy
clustering algorithm. Experimental results were presented for ten companies of the NASDAQ stock market for a period of 19 months. Lu et al. (1998) studied stock series for extracting inter-transaction association rules. They defined events on the time scale and tried to extract relations between similar patterns. The algorithm was limited to a fixed "sliding window" time interval. Only associations among the events within the same window were extracted. A demonstration of a single dimensional database of a stock market was given, but without prediction results.

Last et al (2001) introduced a general methodology for knowledge discovery in time series databases. The process included cleaning of the data, identifying the most important predicting attributes, and extracting a set of association rules to predict the series future behaviour. The computational theory of perception was used to reduce the set of extracted rules by fuzzification and aggregation. Povinelli and Feng (2003) proposed a method that identified predictive temporal structures in reconstructed phase spaces. A genetic algorithm was employed to search such phase spaces for optimal heterogeneous clusters that are predictive of the desired events. Their approach was able to classify time series events more effectively compared to Time Delay Neural Networks and the C4.5 algorithm.

This works introduces a two-stages methodology to identify formations of non-constant length appearing within a financial time series. Initially, the series is represented in linear segments using a modified bottom-up approach. Furthermore, one of the most vague issues in similar studies, the determination of the optimal number of segments is addressed. Then, a rule extraction algorithm is applied to estimate those series formations with a higher probability of reoccurrence. The proposed approach is applied on two major financial series, the daily closing values of the SP500 Index and the GB Pound to US Dollar exchange rate.

I. METHODOLOGY

In the literature there are a number of authors discussing the issue of similarity for time series segments. The problem is defined in Hetland (2004) as: “For a sequence q, find the nearest ones from a set x. These are defined as those whose distance, using any selected measure, from the original is below a predefined tolerance threshold.” The problem gets further complicated when the original segments are not previously known and the examined segments are of non-constant length.

The target of the developed methodology is to identify patterns or formations, of a time series that have high probability of reoccurrence, without any previous knowledge about their characteristics. It is designed to produce fast and easily interpretable results with only a small number of parameters that need to be selected. Furthermore, as will be discussed in the following paragraphs, all parts of the methodology present an improvement over previous research (Kim et al, 2000; Park et al, 2000; Caraca-Valente & Lopez-Chavarrias, 2000; Last et al, 2001; Lin et al, 2002; Povinelli & Feng, 2003; Sarker et al, 2003). In this work the developed methodology is tested to financial data, although its implementation to type of data is straightforward. The various stages of the process are:

- **Data Cleaning.** A suitably selected low pass filter is used to reduce the amount of noise present in the series.
• **Piecewise Linear Segmentation.** The bottom up segmentation algorithm is applied with the modification of joining nearest segments if they have similar trends. An optimization phase is included to identify the optimum number of segments.

• **Symbolic Representation.** The resulting segments of the series are then classified into a finite alphabet.

• **Rule Extraction.** Finally, the sequential characteristics of the alphabet are extracted in the form of symbolic rules.

### A. DATA CLEANING

Initially, the original time series was smoothed using a low-pass digital FIR filter, (1). The signal contains lower amount of noise which brings to light characteristics such as trend. The coefficients, $a_k$ and $b_k$, are estimated so that the resulting signal has the desired behaviour in terms of frequency response. The low-pass filter cut-off frequency was determined using Welch’s (1967) power spectral density algorithm.

$$y_t = \sum_{k=0}^{N} a_k x_{t-k} - \sum_{k=0}^{M} b_k y_{t-k}$$  \hspace{1cm} (1)

### B. LINEAR SEGMENTATION WITH MATCHING TRENDS

The epicentre of all subsequent matching approaches is the efficient representation of the time series, which substantially speeds up the rule extraction algorithm. Argawal (1993) gathered information about the characteristics of a series using the F-index methodology, later evolved by Faloutsos (1994) into the ST-index to account for subsequence matching. Yi and Faloutsos (2000) and Keogh et al (2001) derived independently a piecewise constant approximation approach to divide each sequence into equally spaced segments, using the average value of each as a signature.

Keogh and Smyth (1997) applied a probabilistic landmark-based technique, where the extracted features are characteristic parts of the sequence. Perng et al (2000) applied a similar approach that identifies points of “great importance” as landmark features. The specific form used defined landmarks as points where the approximation function’s $n$th derivative is zero. Kim et al (2001) used as landmarks the boundaries of the segments as well as minimum and maximum points.

Many researchers attempt to capture the main characteristics of the data using piecewise linear approximation. There are three main algorithms applied for this task: sliding windows, top-down and bottom-up, which are combined with two alternatives approximations: interpolation and regression. For an excellent review and performance comparison of these approaches the reader is directed to the work of Keogh at al (2004). The authors also proposed the SWAB algorithm, a combination of sliding windows and bottom-up, which was empirically shown to be superior to any other examined.
The proposed algorithm is a modified bottom up algorithm with a linear interpolation between end-points, also including an extra phase that forces the merging of nearest segments if their trends are similar, (Figure 1). Two trends are presumed similar if their difference is below a predefined threshold. Furthermore, we address one of the most vague issues in similar studies, the selection of an optimum number of segments. Thus, the only parameters that need to be selected are the minimum number of data contained in each segment and the trend similarity threshold.

This algorithm requires the use of a cost function. It assists in the process of selecting the appropriate segments to merge, and is the critical measure for the selection of the optimal number of segments. The selected cost function in this work is the squared sum of the Euclidean distance of all members of a segment from the linear interpolant between the end points, i.e.

\[
E = \sum_{j=1}^{k} \left( \sum_{j=1}^{k} \text{dist}(y_j - I_k) \right)^2
\]

In case of total series cost

\[
E = \sum_{\text{all segments}} \left( \sum_{j=1}^{k} \text{dist}(y_j - I_k) \right)^2
\]

**Figure 1. Bottom-up Algorithm with Matching Trends**

<table>
<thead>
<tr>
<th>Step 1</th>
<th>Split series into segments with length (d_{min}), Merge nearest segments if their trend is similar</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Assign any remain data to the first or last segment</td>
</tr>
<tr>
<td></td>
<td>Store number of total segments (n_{seg}(1)), and total cost of series (E(1))</td>
</tr>
<tr>
<td>Step 2</td>
<td>Loop all segments</td>
</tr>
<tr>
<td></td>
<td>Estimate the merging cost of two nearby segments, (e)</td>
</tr>
<tr>
<td></td>
<td>Find segments with minimum (e) and merge</td>
</tr>
<tr>
<td>Step 3</td>
<td>Estimate trend of new segment and compare with neighboring trends</td>
</tr>
<tr>
<td></td>
<td>If trends similar then merge</td>
</tr>
<tr>
<td></td>
<td>Store total cost of series (E(k)) and number segments (n_{seg}(k))</td>
</tr>
<tr>
<td>Step 4</td>
<td>Go to Step 2, unless a predefined number of segments are reached</td>
</tr>
<tr>
<td>Step 5</td>
<td>Selection of the optimal number of segments</td>
</tr>
</tbody>
</table>
The looping stage of the algorithm stops when the total number of segments reaches below 10% of the number segments estimated in Step 1 \((n_{\text{seg}(1)})\). This was set to speed up the algorithm, since it was empirically found that financial series require a relatively high number of segments due to their complex characteristics. The optimal number of segments is found from a trade-off the number of segments and the total cost of the series, using their combinatory expression. These two indices are scaled to vary between \([0,1]\), which ensures the validity of the scheme to any sort of data. The finally selected number is the one that minimizes (3).

\[
\begin{align*}
    n &= \frac{\max(n_{\text{seg}}) - n_{\text{seg}}}{\max(n_{\text{seg}}) - \min(n_{\text{seg}})} \\
    E - n &= \frac{\max(E) - E}{\max(E) - \min(E)} \\
    Comb &= n + E - n
\end{align*}
\]  

(3)

**C. SYMBOLIC REPRESENTATION**

The finally determined segments, \(s\), are then classified into a finite number of distinct classes or alphabet, \(C_s\), which facilitates the process of rule extraction. There are two main procedures for achieving this, heuristically and clustering (Das 1998). The size of alphabet should be selected taking into consideration the total number of series segments. In this work, the alphabet size was directly related to the financial nature of the examined series. Heuristically, five classes were chosen to represent different market states. Additional information included was the duration of the segments, classified into three types (Table 1).

<table>
<thead>
<tr>
<th>Series Values</th>
<th>Segment Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class</td>
<td>Symbol</td>
</tr>
<tr>
<td>Very Negative</td>
<td>VN</td>
</tr>
<tr>
<td>Negative</td>
<td>N</td>
</tr>
<tr>
<td>Constant</td>
<td>C</td>
</tr>
<tr>
<td>Positive</td>
<td>P</td>
</tr>
<tr>
<td>Very Positive</td>
<td>VP</td>
</tr>
</tbody>
</table>

**D. RULE EXTRACTION**

A rule extraction algorithm is derived to estimate association rules that describe sequential characteristics of the estimated classes, as selected in the previous step. The association rules are of the generalised form.
The antecedent part or Left Hand Side (LHS) contains historic information about the classes, C, ordered in a sequential manner. The consequent part or Right Hand Side (RHS) contains future information of the series. Two measures are used to determine the strength of the rule, using information contained in the data set:

♦ Support is the number of instances that the rule appears within the data set.
♦ Confidence is the accuracy with which the rule predicts correctly, i.e. that the LHS leads to the RHS.

These measures, which are user-defined parameters, are a measure of the frequency that a formation appears in the data set and its predictability. The total number of possible rules is the number of different classes to the power of the combined LHS and RHS dimensions.

**Figure 2. Rule Extraction Algorithm**

```
Define: minimum support and confidence levels

Step 1
Select LHS as C(s-1)

Step 2
Set flag (for all segments) = true
Loop all segments (s)
   Find first true flag and add as a new rule
   Find remaining data following specified rule
   Set flag(data) = false

Step 3
Remove rules < support threshold
Remove rules < confidence threshold
Save final set of rules

Step 4
If no rules are found
   End algorithm
Else
   Add previous class to LHS
   Go to Step 2
```

Previous attempts of solving the problem of mining predictive rules from time series can be classified into two main types. Supervised methods, where the target-rule form is known in advance and used as input to the rule extraction algorithm. Thus the objective of the analysis becomes to generate rules for predicting these events based on the data available before the event occurred (Weiss, 1998, Hetland, 2002). In the second type, unsupervised methods, the inputs to the algorithms are only the data. The goal is to automatically extract informative rules from the series. In most cases this means that the rules should have some level of preciseness, be representative of the data, easy to interpret, and interesting to the financial analyst (Freitas, 2002). The most celebrated approach in the literature (e.g. Agrawal, 1995; Mannila, 1997; Keogh, 2002) is to sequentially scan all the available data and add up every occurrence of a rule, and of every antecedent and consequent part. This counting makes it possible to calculate the frequency of appearance and confidence
of each rule, in the cost of limiting the rule format. The developed rule extraction algorithm is based on a sequential database scan, but the search in each loop is done with decreasing number of samples. The entire process is iterated with an increasing adjacent part until the support of all estimated rules are below a predefined threshold value. The finally presented rules are those above a specified support and confidence threshold value. These values are selected to extract only frequent patterns with high probability that the consequent part will appear in the future, enabling profitable trading decisions.

FIGURE 3. (A) DETAIL OF SEGMENTED SP500 SERIES AND (B) THE OPTIMIZATION PHASE
2. APPLICATION TO THE SP500 DAILY INDEX

The first series that the methodology was applied to is the daily closing values of the SP500 index. The data covered a period of 19 years including only working days, from the first trading day of 1985 to January 27, 2004. Initially, a low-pass filter was applied with a cut-off frequency value of 0.1 (1/day), which is approximately a period of two working days. The minimum segment length, dmin, was set to three, including end-points, and the trend-matching threshold, ε, to 0.3%. A detail of the resulting segmented series is presented in Figure 3a, alongside the graph showing the scaled values of the total cost, E, and the number of segments (3b).

The derived segments were characterised by their trend defined as the percentage change between the start and end points as:

\[ tr = 100 \times \frac{x_{\text{end}} - x_{\text{start}}}{x_{\text{start}}} \]  

The trends of each segment were classified into five different classes, Table 2, as

<table>
<thead>
<tr>
<th>Trend Classification for the SP500 Series</th>
</tr>
</thead>
<tbody>
<tr>
<td>VN ≤ -4%</td>
</tr>
</tbody>
</table>

These values were selected so that the classes contain roughly the same number of data. The threshold values for the rule extraction algorithm were set as 5 values for support and 70% for confidence. The later ensured that the extracted rules would lead to profitable decisions. Two different types of rules were extracted. Initially, the antecedent part contained only the trend type class, whilst in the second it additionally contained the duration class of the segment. The results from both cases are summarized in Tables 3 and 4.

<table>
<thead>
<tr>
<th>Table 3. SP500 Series Rules Using Trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
</tr>
<tr>
<td>----</td>
</tr>
<tr>
<td>R1</td>
</tr>
<tr>
<td>R2</td>
</tr>
<tr>
<td>R3</td>
</tr>
<tr>
<td>R4</td>
</tr>
</tbody>
</table>

The first rule (R1) is translated, as segments with strong increase followed by relatively milder decrease and increase respectively and then a period of stable market behaviour will be followed by increasing values. This rule appeared...
collectively six times in the examined series, all in periods with upward trend, and all of them resulted in the same consequent characteristics. Rule 2 shows a stable market followed by strong decrease is and then by a succession of positive and negative behaviour. This rule mostly appears as the first half of a peak starting cyclical behaviour. Rule 3 shows a succession of segments with negative and positive trends ending in stable market behaviour when two positive periods, with different trend, appear consecutively. Rule 4 is partly the continuation of the previous rule exhibiting a decrease in the future value of the trend, but also showed up independently in more than half of appearances.

**Figure 4. Examples of R1-R4 in the SP500 series**
TABLE 4. SP500 SERIES RULES USING TREND AND DURATION

<table>
<thead>
<tr>
<th>ID</th>
<th>Rule</th>
<th>Support</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>R5</td>
<td>seg(s-2): VP_B &amp; seg(s-1): N_S ⇒ seg(s): P</td>
<td>10</td>
<td>71.4%</td>
</tr>
<tr>
<td>R6</td>
<td>seg (s-2): N_B &amp; seg (s-1): P_B ⇒ seg (s): N</td>
<td>6</td>
<td>75%</td>
</tr>
<tr>
<td>R7</td>
<td>seg (s-3): N_B &amp; seg (s-2): P_B &amp; seg (s-1): N_S ⇒ seg (s): P</td>
<td>5</td>
<td>71.4%</td>
</tr>
<tr>
<td>R8</td>
<td>seg (s-3): VN_S &amp; seg (s-2): C_S &amp; seg (s-1): P_S ⇒ seg (s): N</td>
<td>5</td>
<td>71.4%</td>
</tr>
<tr>
<td>R9</td>
<td>seg (s-3): C_S &amp; seg (s-3): N_S &amp; seg (s-2): C_S &amp; seg (s-1): P_S ⇒ seg (s): N</td>
<td>5</td>
<td>71.4%</td>
</tr>
</tbody>
</table>

In the next stage, the analysis included the duration of the antecedent part as additional information, in attempt to gather more information the behaviour and transitional characteristics of the series. Rule five (R5) shows that biweekly periods with strong increase that are followed by short periods of decrease are more likely to results in increasing future behaviour. This rules is associated mainly with bullish market periods. The next rule, R6, is a succession of negative and positive periods lasting biweekly periods, which leads to a decrease. Rule 7 represents periods where biweekly negative and positive periods followed by short decrease will mostly be followed by increasing behaviour. The last two rules appear on the first 12 years the examined series. Rule 8, R8, shows a succession of bearish, stable and positive market characteristics, each lasting less than a week, is followed by negative trend. On the majority of the appearances of this rule the antecedent part forms a complete cycle, and the consequent the start of a second one.
A short decrease is followed by biweekly stable periods are more likely to be followed by an increase in the series price. The antecedent part of R9 is a full cycle last more than three weeks in total, which is followed by the start of a new cycle.

**Figure 6. Examples from R5, R6, R7 and R8**
3. APPLICATION TO THE DAILY POUND/DOLLAR EXCHANGE RATE

The developed methodology is applied to the daily GB pound to US dollar exchange rate series. The data covered a period of approximately 32 years from the 4th January 1971 to December 23, 2003. The exchange rates series is influenced by a number of factors, due to its significance as a policy tool for dealing with macroeconomic issues. Furthermore, there are many ways that central banks intervene in the foreign exchange, which induces additional features in the series (Taylor). Initially, a low-pass filter was applied with a cut-off frequency value of 0.07 (1/day), which is approximately a period of two weeks. The minimum segment length, dmin, was set to five and the trend-matching threshold, εt, to 0.25%. The optimization stage returns 763 segments (Figure 7).
The derived segments were characterised by their trend again defined in percentage terms. A five letters alphabet was chosen, Table 5. For the trend classification, the three classes described in Table 1 were selected.

**TABLE 5. TREND CLASSIFICATION FOR THE GBP/USD SERIES**

<table>
<thead>
<tr>
<th>VN ≤ -2%</th>
<th>-2% &lt; N ≤ -0.5%</th>
<th>-0.5% &lt; C ≤0.5%</th>
<th>0.5% &lt; P ≤ 2%</th>
<th>VP &gt; 2%</th>
</tr>
</thead>
</table>

The threshold values for the rule extraction algorithm were set as 3 values for the support and 70% for the confidence parameters. Again the same two types of rules were extracted, with the antecedent part contained trend type class and the duration of the segments. Results from both cases are summarized in Tables 6 and 7.

**TABLE 6. GBP/USD RULES USING SEGMENT TREND**

<table>
<thead>
<tr>
<th>ID</th>
<th>Rule</th>
<th>Support</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>TR(s-4): VN &amp; TR(s-3): N &amp; TR(s-2): P &amp; TR(s-1): VN ⇒ TR(s): VN</td>
<td>3</td>
<td>100%</td>
</tr>
<tr>
<td>R2</td>
<td>TR(s-4): P &amp; TR(s-3): N &amp; TR(s-2): VN &amp; TR(s-1): P ⇒ TR(s): VN</td>
<td>3</td>
<td>75%</td>
</tr>
<tr>
<td>R3</td>
<td>TR(s-4): VN &amp; TR(s-3): P &amp; TR(s-2): VN &amp; TR(s-1): P ⇒ TR(s): VN</td>
<td>3</td>
<td>75%</td>
</tr>
<tr>
<td>R4</td>
<td>TR(s-4): C &amp; TR(s-3): VP &amp; TR(s-2): C &amp; TR(s-1): N ⇒ TR(s): P</td>
<td>3</td>
<td>75%</td>
</tr>
</tbody>
</table>
Rules 1 to 3 represent periods of decreasing trend, ending in very negative market characteristics (Figure 8). Rules 1 appears at periods with strong increasing trends, whilst R2 the latter mainly appears as the falling part of cycles lasting approximately three months. R3 appears at non well-defined periods of the data. Rule 4 shows a general upward trend with stable and mildly decreasing characteristics. Rule 5 shows a succession of positive and negative periods ending in low fluctuating markets.

**Figure 8. Examples from R1, R2 and R3**
In the next stage, the analysis included the duration of the antecedent part as additional information, in attempt to gather more information the behaviour and transitional characteristics of the series (Table 7). A set of three rules (R6-R8) shows periods ending in strong increasing characteristics (Figure 9). R6 shows the formation of a cycle that starts with long periods of negative and stable characteristics, followed by sharp increase. R7 captures the duration of a whole cycle whose antecedent part is approximately three weeks. Rule 8 shows strong increasing behaviour interrupted by a very short decreasing period. Another set of rules with bigger LHS is associated with decreasing consequent behaviour (Figure 10). Rule 9 represents the behaviour of the series period after a cycle peak. Rule 10 is a succession of decreasing and increasing periods followed by decreasing behaviour. This rule appears at irregular instances in the period. Finally, R11 is a succession of biweekly trends that appears mostly on periods of low fluctuations.

**Table 7. GBP/USD extracted rules using trend and duration information**

<table>
<thead>
<tr>
<th>ID</th>
<th>Rule</th>
<th>Support</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>R6</td>
<td>( \text{seg}(t-2): N _L &amp; \text{seg}(t-1): C _L \Rightarrow \text{seg}(t): VP )</td>
<td>3</td>
<td>75%</td>
</tr>
<tr>
<td>R7</td>
<td>( \text{seg}(t-2): VN _S &amp; \text{seg}(t-1): C _B \Rightarrow \text{seg}(t): VP )</td>
<td>3</td>
<td>100%</td>
</tr>
<tr>
<td>R8</td>
<td>( \text{seg}(t-2): VP _L &amp; \text{seg}(t-1): N _S \Rightarrow \text{seg}(t): VP )</td>
<td>3</td>
<td>75%</td>
</tr>
<tr>
<td>R9</td>
<td>( \text{seg}(t-3): P _L &amp; \text{seg}(t-2): VP _B &amp; \text{seg}(t-1): C _B \Rightarrow \text{seg}(t): N )</td>
<td>3</td>
<td>75%</td>
</tr>
<tr>
<td>R10</td>
<td>( \text{seg}(t-3): N _B &amp; \text{seg}(t-2): P _B &amp; \text{seg}(t-1): N _L \Rightarrow \text{seg}(t): VN )</td>
<td>5</td>
<td>71.4%</td>
</tr>
<tr>
<td>R11</td>
<td>( \text{seg}(t-3): P _B &amp; \text{seg}(t-2): N _B &amp; \text{seg}(t-1): P _B \Rightarrow \text{seg}(t): N )</td>
<td>3</td>
<td>100%</td>
</tr>
</tbody>
</table>
FIGURE 9. EXAMPLES OF R6, R7 AND R8
4. Conclusions

In this paper we developed a methodology for the detection and extraction of similar subsequences of the data that appear in a time series. The methodology is designed to produce easily interpretable results with only a small number of parameters that needed to be defined. The main stages of the algorithm are the series representation using a bottom up segmentation algorithm extended to account for matching joining segments with similar trends, and a modified sequential rule extraction algorithm. Furthermore, it addresses one of the most overlooked issues in similar studies, that is the optimal selection of the number of segments.
The approach is tested on two financial series the daily closing values of the SP500 index and the GB Pound to US Dollar exchange rate. Adding to trend type representation of the series, the duration of the trend was included. Setting a confidence threshold parameter as 70%, three and five rules were identified for the two series respectively using only trend type information. The number of rules increased to five and six respectively when duration was used.

REFERENCES


Kim, S.W., S. Park and W.W. Chu, 2001. An index based approach for similarity search supporting time warping in large sequence databases. In ICDE.


