UDK 007.681
Original scientific paper

TIME SERIES ANALYSIS USING A UNIQUE MODEL OF TRANSFORMATION

Goran Klepac
Raiffeisen Consulting, Zagreb, Croatia
goran.klepac@rba.hr.com

Abstract: *REFII[1] model is an authorial mathematical model for time series data mining. The main purpose of that model is to automate time series analysis, through a unique transformation model of time series. An advantage of this approach of time series analysis is the linkage of different methods for time series analysis, linking traditional data mining tools in time series, and constructing new algorithms for analyzing time series.*

It is worth mentioning that REFII model is not a closed system, which means that we have a finite set of methods. At first, this is a model for transformation of values of time series, which prepares data used by different sets of methods based on the same model of transformation in a domain of problem space. REFII model gives a new approach in time series analysis based on a unique model of transformation, which is a base for all kind of time series analysis.

The advantage of REFII model is its possible application in many different areas such as finance, medicine, voice recognition, face recognition and text mining.

Keywords: *Time series, data mining, transformation model, automated series data preparation and data preparation.*

1. INTRODUCTION

In traditional time series data mining analysis, there is a lot of different methods which solve a particular kind of problem.

As a result of this approach, we have a situation in which if we want to solve a problem of discovering patterns in time series we could use methods which are described by Xsniaping [Xsniaping, 1998], Han [Han, 1998], or Prat [Pratt, 2001]. If we would like to solve problems in domain of events and episodes in time series, then we could use a method which is described by Manilla [Manilla, 1997].

Many different methods for different kind of problems exist in scientific works. There are many different methods for solving seasonal oscillations, recognition time segments, similarity search etc. A mutual characteristic of announced future work in those scientific

[1] Acronym of "Rise, Equal, Fall, second generation"

papers is to discover methods for time series analysis, which are already discovered as a result of research done by other authors. Authors usually develop a qualitative model which solves some kind of problem, but it is usually impossible to connect their model with already discovered models which solve other problems. The reason why authors could not link other methods with their own methods can be found in the model of transformation.

Authors are, at first, focused on solving problems, and they do not care much about the model of transformation. Every developed model has its own model of transformation which implicates incompatibility among methods. Different model of transformation does not allow linking of different methods. Other big problem which is based on transformation models is linkage of time series with traditional data mining methods, automation in series data preparation, and construction of new analytical algorithms for time series analysis based on contingency. For example, we could not connect time series with decision trees using traditional way of analysis. This could be interesting if we would like to know the main characteristic and attributes of clients which show the rising trend of using Automatic Teller Machines (ATM) during some period of time.

Another problem which resulted from this approach, and which is significant in time series analysis in the domain of marketing, is a problem of chaining methods for time series analysis. For example, if we would like to know whether clients which show seasonal oscillation in behavior of using ATMs during summer, have some similar patterns in using ATMs during the rest of the year which are not seasonal oriented, we have to use more than one different method.

The main problem is how to connect these two methods. In practice we have to connect two different models of transformation, and sometimes it is not so easy and elegant to be realized. Some methods use discrete Fourier transformation as a model of transformation. Some methods use symbolic transformation, leg transformation etc.

Those approaches implicate that:

"Series data has features that require more involvement by the miner in the preparation process than for non-series data. Where miner involvement is required, fully automated preparation tools cannot be used. The miner just has to be involved in the preparation and exercise judgment and experience." [Pyle, 2001] . Pyle as a solution of mentioned problem see in more powerful, low-cost computer systems.

Suggested solution for all mentioned problems are REFII model, as a fully automated preparation tool which gives solution for problems such as:

- Discover seasonal oscillation;
- Discover cyclic oscillation;
- Discover rules from time series;
- Discover episodes from time series;
- Discover similarity of time segments;
- Discover correlation between time segments;
- Discover rules from in domain of finances from time series;
- Connect time series and standard data mining methods;
- Analyze time series with the help of data mining methods (clustering of time segments, classification of time segments).

This kind of approach of analyzing time series is different from traditional approach which is based on using different model of transformation for each type of analysis. In case of using traditional approach of analysis, there are many problems in chaining different sets of methods, using traditional data mining methods in time series analysis and automation in series data preparation. The task of REFII model is linking and chaining different methods for time series analysis as it is shown in Figure 1.

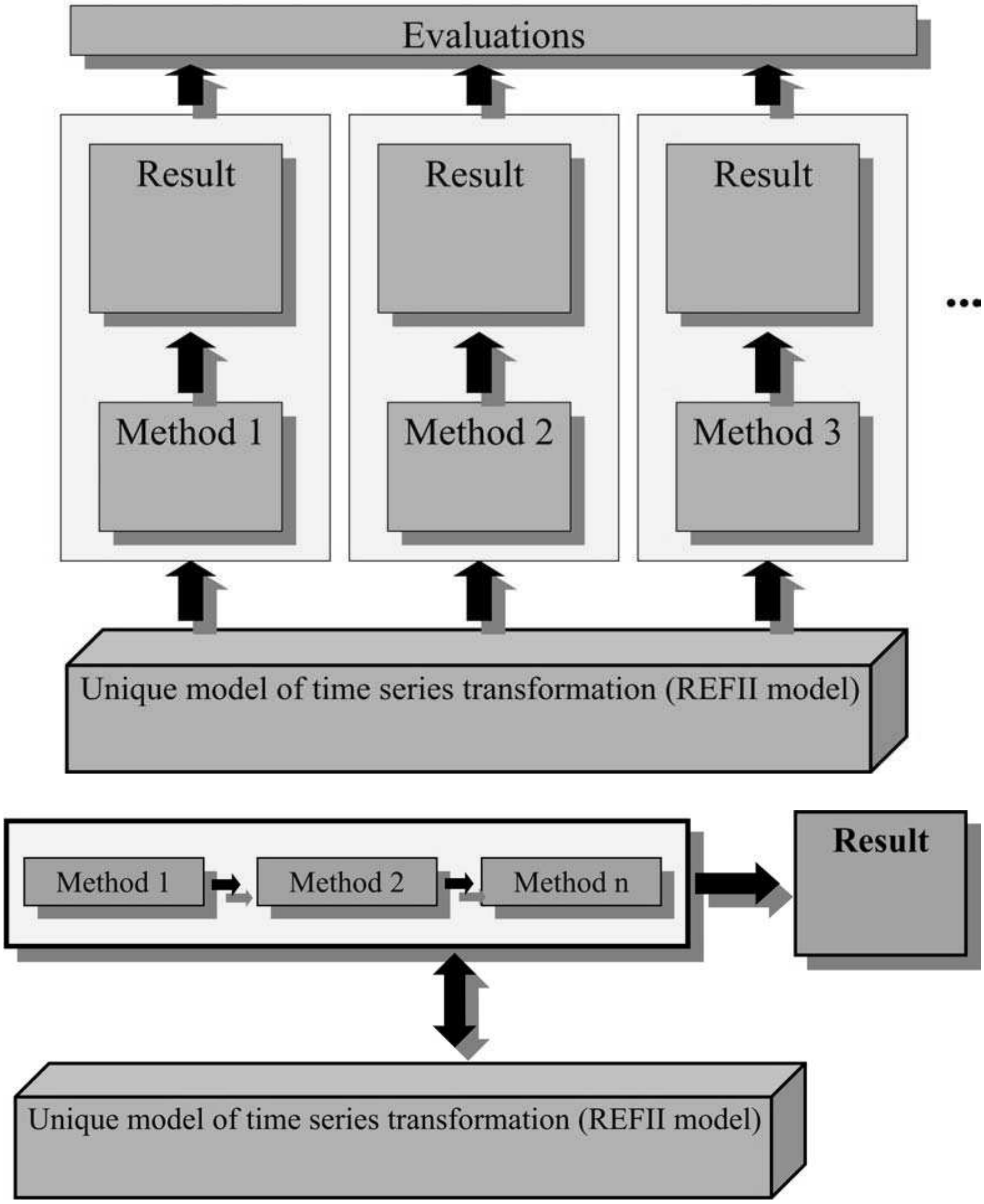

Evaluations
Result
Result
Result
Method 1
Method 2
Method 3
...
Unique model of time series transformation (REFII model)
Method 1
Method 2
Method n
Result
Unique model of time series transformation (REFII model)

Figure 1. Linking and chaining different methods with REFII model

2. REFII MODEL

The mathematical background of REFII model is focused on shape and area beneath the curve. REFII model is able to analyze every time series, which is represented by values. In the first step of the analysis, after preparing the original data, we make a transformation of time series into REFII model values. The next step is to select an appropriate method inside to analyze data. The selection of the method is determined within the scope of the analysis. Methods could be focused on discovering: seasonal oscillation, cyclic oscillation, hidden rules, episodes, similarity of time segments, and correlation of time segments, clusters or links between time segments. The basic characteristic of the REFII model is a uniformity of describing time series with the model parameters. Mathematical uniformity implies the possibility of executing the basic operations of comparison, which is of a great significance for the observed field. The concept aimed to meet another criterion, which is connectedness with algorithms applied in data mining. Until now, the known methods for analysis of time series yielded certain ratios which could not be processed later through one of the known algorithms to extract additional knowledge. The REFII model aims at openness, i.e. its mathematical instruments serve to describe, as well as to generate knowledge hidden in time series, provide solution models of connecting with other data mining algorithms. Thus, we can use the strength of verified algorithms in the field of time series, within standard software solutions.

The classical data mining algorithms include neural networks, clustering, decision trees, market basket analysis, link analysis and the like, as well as all mutations of algorithms. It is easy to presume the strength lying in a system which is capable of clustering time segments, or in a system which uses decision tree algorithms to classify time segments or complete time series. In addition, we can execute link analysis on time series or their segments, as well as study similarities of time series on the basis of distance function. The concept opens a whole new field offering more detailed and precise analytical instruments in the time series domain.

The REFII model focuses on three basic segments which uniformly could describe the curve, concentrated on:

- Shape of curve (how time series looks);
- Area beneath the curve (quantification of time series);
- Angle inside time segment (“strongest of trend").

Relations of the three elements making the REFII model are shown in Figure 2.

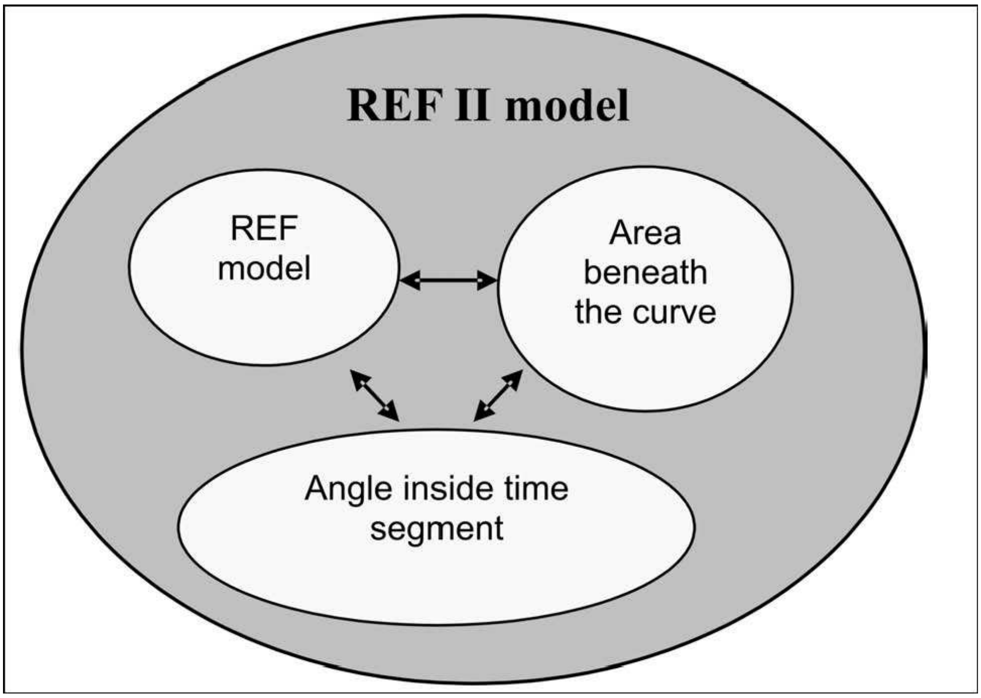


Figure 2. REF II model

Within the REF II model, REF model is what determines shape which will be described in details on further pages. Its characteristic is diagnostics and modeling of shape definition. Unfortunately, this model cannot define the curve uniformly, therefore, the other two mentioned elements of the model should be used

Area beneath the curve gives a quantitative dimension to a REFII model. The curve can be shaped equally as some other curve, but this does not mean it has the same quantitative value, which explains the concept of curve inequality. Area beneath the curve can help us to obtain the value of that inequality. In order to completely define time series uniformly by a mathematical model, it is necessary to introduce a third element, the angle within the time segments. Simple schema of REFII model is shown in Figure 3.

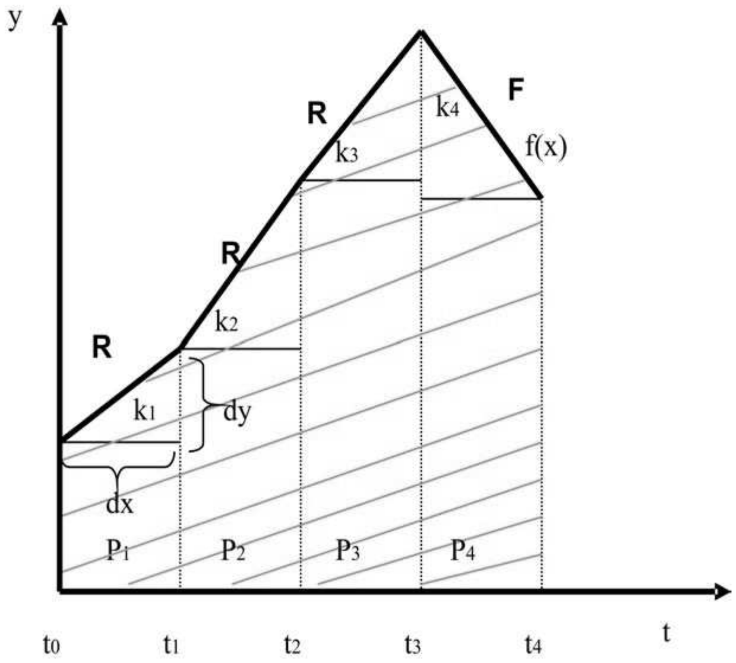


Figure 3. Schema of REFII model

"R","E", or "F" marks represent the time series trend inside of observed segment, P_n represents time series area beneath the observed segment, and k_n represents angle coefficient inside of observed segment.

The REF II model, as shown above, consists of three basic components: REF model, calculation of area under the curve, and angles of the observed time segment. The characteristic of the model is a possible selective approach to its components, which depends on the nature of the problem we wish to solve by applying it. The main characteristic and aim of the model, when we observe it through the prism of all the components of which it consists, is its possibility of uniform description of curve, so that the concept can be applied to time series analysis. There is a whole range of conceptual solutions on the basis of which this model can be fitted into a standardized nomenclature readable to algorithms.

Prior to embarking on the analysis process, it is necessary to reduce time series to equal time segments at the point of pre-processing. This means that we aggregate events within equal time distances, mostly using the method of summarization, or average values.

2.1. THE REF SUB MODEL

Within the REF II model, the REF sub model is a segment which has a task to detect curve shape. However, it does not estimate the strength of trends of which the curve consists, or details of quantitative aspects of time segments, which are in the domain of other components of the REF II model. The task of the REF model is primarily the diagnostics of curve shape in certain segments of time series. Mathematical set of instruments used for developing this method can be reduced to the theory of function monotony, which is also the core of the whole method.

The basic idea of the model stems from the mathematical theory on function monotonicity [Fanchi, 2000], [Javor, 1988] .

Function

$f : D \to R, D \subseteq R$

monotonously grows on D if for each $x_1, x_2 \in D$ we have

$(x_1 < x_2) \Rightarrow (f(x_1) \leq f(x_2))$

or monotonously falls on D if for each $x_1, x_2 \in D$ we have

$(x_1 < x_2) \Rightarrow (f(x_1) \geq f(x_2))$

where D is the domain of *R*, and *R* is a set of real numbers.

In situations when function is in the form of time series $(e_1, e_2, ... e_n)$, and is obtained from the database, we can calculate the partial trend from formula:

$t = e_{i-1} - e_{i-2}$

in which

t - presents the trend of time segment

e - Presents the value of time series (hour, day, month...)

i – time index

If $t > 0$, this means the function is growing. If $t < 0$, this means the function is falling. If $t = 0$, the function is neither growing nor falling, but retains the same value of the time trend.

The growing part $(t > 0)$ of the trend function resulting from this formula in a certain segment can be presented by "*R*" mark.

The falling part $(t < 0)$ of the trend function resulting from this formula in a certain segment can be presented by "*F*" mark.

The neutral part $(t = 0)$ of the trend function resulting from this formula in a certain segment can be presented by "*E*" mark.

Using this notation, we can present the complete time series by three marks. For example, for the function (R,R,F,F,F,R) we can say to be growing for two time units, to be falling for three time units afterwards, and to be growing for onetime unit. The presented model makes REF model to look like a redundant element, because we can describe the shape of the curve with angles between time segments. The reason for using REF model as a system reference segment comes from the fact that time used by the machine to estimate curve shape by REF model is much shorter than in a model which would be based on elements of angles among time segments. Certainly, REF model is also much more imprecise than a model based on angles between time segments, but it is also much faster in processing.

2.2. AREA BENEATH THE CURVE

The depicted REF model can help us in time series analysis when we are not interested in the quantitative aspect of the time series. If we want to analyze time segments with regard to quantitative component, calculation of area beneath the curve will be helpful. The basic idea of calculating area beneath the curve lies in the integration, in using a certain integral.

Area beneath the curve can be calculated by the method of numeric integration for certain segments partially, as well as globally for a given interval, which of course depends on the aim of the analysis. When we speak of compatibility within a model, looking from the perspective of the REF model and calculation of area beneath the curve, it is important to establish time intervals to which we shall divide the curve in order to make the model transparent. As we have shown time series by a number of broken lines, the method of numeric integration based on the rectangle is the most convenient for area beneath the curve. The result of calculating area beneath the curve or its segments is a numeric value representing value of the area beneath the curve, which is calculated by expression:

$$p=\frac{(x_n*\Delta t)-(x_{n+1}*\Delta t)}{2}$$

Areas beneath the curve give us information about the volume of same event. Some time segment could have equal shape, but if they do not have equal values of areas beneath the curve that means that curve segments are not equal.

2.3. ANGLE COEFFICIENT BETWEEN LINES WITHIN TIME SEGMENTS

The final element completing the REFII model is the calculation of angles between lines within time segments. When we compare the REF model with this model, the degree of expected precision of this model is very high. In the REF model, the growth of a trend in a time segment could be defined by R, but with calculating angle we could precisely calculate the value of angle.

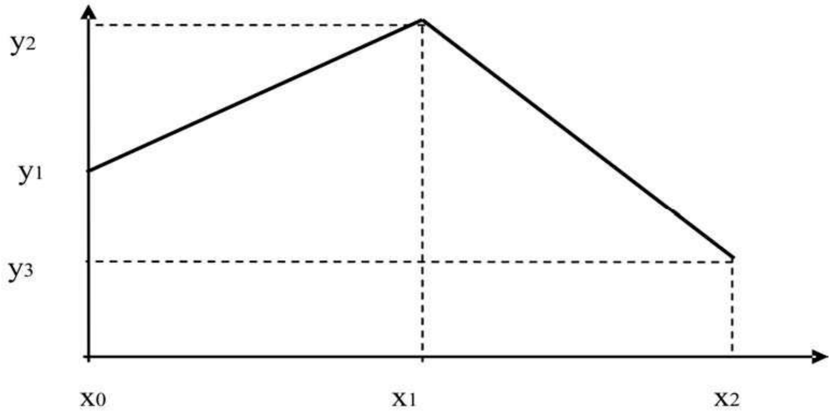


Figure 4. Model of calculating angle coefficient

As the figure shows, one of the prerequisites is the process of data preprocessing which divides the time series into equal segments, and presents it by a number of lines, be it for the possibility of applying the concrete model as well as for the possibility of achieving compatibility. The presented model yields the following rules:

In the first step, time series $S=(y_1,...,y_n)$ is transformed into time series $T=(t_1,...,t_n)$ with formula

$$t = \frac{y_i - min}{max - min}$$

where t_i is an element of $T = (t_1,...,t_n)$, min is a minimal value of *S*, max is a maximal value of *S*.

The values of angle coefficient k_j are calculated as follows:

$k_j = t_{i+1} - t_i$ for the $t_{i+1} - t_i > 0;$

$k_j = t_i - t_{i+1}$ for the $t_{i+1} - t_i < 0;$

$k_j = 0$ for the $t_{i+1} - t_i = 0$

Expected values of angle coefficient are in interval $< 0 - 1 >$.

2.4. TRANSFORMATION MATRIX

The result of REF II model is the transformation matrix of time series which represents time series as it is shown in Figure 5.

Time segment index	I1	I2	I3	In
REF notation	REF(I1)	REF(I2)	REF(I3)	REF(I n)
Angle coefficient	angle(I1)	angle (I2)	angle (I3)	angle (In)
Area of time segment	P(I1)	P(I2)	P(I3)	P(In)

Figure 5. Transformation matrix

Such a structure, viewed from the perspective of dynamic memory, can be a matrix $4 \times t_{n-1}$, or a table with four attributes of length t_0, viewed from the perspective of disc recording.

Indicators of time segment were calculated on the basis of co-ordinates of two neighboring values in the time series. Thus, for example, the segment with index 1 was formed on the basis of the value of time series co-ordinates t_0 and t_1. Time segment index serves for uniform identification of the time segment in order to analyze it. Elements included into the above table are the basic elements of the REFII model, which can be used to describe the curve uniformly and make all analysis for which the model was developed in the first place. Apart from the described indicators, it is possible to include also deducted indicators, shown throughout the model development, but that is an optional approach which depends on the character of analysis. Indicators from the above table are the basis for all analysis for which the model was developed. After transforming time series by the REF II model, we obtain a transformed time series in the form shown in the table. Data which are thus transformed are processed by algorithmic methods in order to solve concrete problems in the domain of time series. The character of analysis determines the approach to a transformed time series regarding the division of a structure thus formed into smaller logical units. If, for example, we observe weeks, and time series has data for each day of the week throughout the year, then the series should be logically divided into weeks and analyze time segments by a concrete algorithm. This procedure of dividing a time series to smaller, analytically comparable, logical segments is related to the concept of time complexity. This measure determines the point of logical division of time series regarding the aim of analysis, and it is a part of the algorithmic process of the analysis. Thus, for example, the coefficient for week can have value 7 or 5 (working days), and one should have in mind the non-existence of a value at a certain point of time, with which the measure also has to deal with. The basic conclusion is that the REF II model in a narrow sense is the model of transforming time series, whereas in its broader sense it is a number of algorithmic procedures of thus transformed data.

This transformation model is the universal starting point for all possible types of analysis which are conducted on time series and which the concept successfully solves. The elements shown in the table, when procedurally processed through various algorithm types, can discover different types of knowledge in time series. And this is where the strength of the REF II concept lies.

This concept can be successfully applied on development of the query language for time series. In order to make the picture as complete as possible, we should not forget the deducted indicators belonging to the specific indicators of sub-models. Their presence can provide a higher quality approach to algorithmic solutions for certain problem types in the observed problem space. As the range of the indicators is extensive, we shall not describe them each separately but within the conceptual application of the model and algorithmic solutions for extracting knowledge from time series.

We could describe generating matrix of transformation in algorithmic way:

1. Create auxiliary time series on interval $<1,n>$ (days, weeks, years…) with the value 0 for each time point; Interpolation existing values of origin time series into existing auxiliary time series;
2. Time granulation (Compressing days to weeks, weeks to months) using statistical function AVG(), SUM(), MOD();
3. Data normalization $N_s(Y_1,..,Y_n) = ((X_i - min(V_s)) / (max(V_s) - min(V_s)), min/max$ normalization. Defining measure of time complexity $d(Y_i, Y_{i+1}) = a$
4. Angle coefficient

   $T_r > 0$ (R) $Coefficient = Y_{i+1} - Y_i$

   $T_r < 0$ (F) $Coefficient = Y_i - Y_{i+1}$ 1

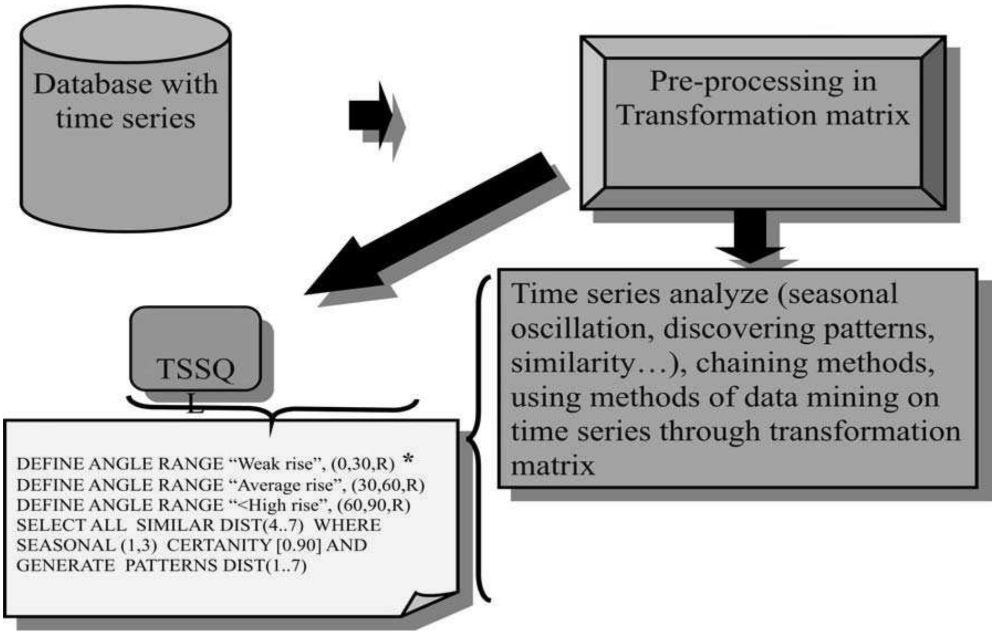
   $T_r = 0$ (E) $Coefficient = 0$ Tr = 0 (E)

   Area beneath the curve $p = ((X_i * a) + (X_{i+1} * a)) / 2$
5. Creating time indexes
6. Creating new categories based on transformation matrix (IF angle =R and angle coefficient > 0.80 THAN category = SHARP RISING)

3. USING REFII MODEL

REFII model as a unique model of transformation give us an opportunity to use different concepts in time series analysis. As result of that concept we are able to make more complex analysis which demand chaining of methods for time series analysis, using traditional data mining methods on time series, and making new algorithms for solving contingency problem in domain of time series analysis.

Figure 6. shows a preposition of system for time series analyse based on REFII model.


Database with time series
Pre-processing in Transformation matrix
TSSQ L
Time series analyze (seasonal oscillation, discovering patterns, similarity…), chaining methods, using methods of data mining on time series through transformation matrix

```
DEFINE ANGLE RANGE "Weak rise", (0,30,R) *
DEFINE ANGLE RANGE "Average rise", (30,60,R)
DEFINE ANGLE RANGE "<High rise", (60,90,R)
SELECT ALL  SIMILAR DIST(4..7)  WHERE
SEASONAL (1,3)  CERTANITY [0.90] AND
GENERATE  PATTERNS DIST(1..7)
```

Figure 6. Time series analytical system based on REFII model

As it is shown in Figure 6, REFII model could be a base for time series query language, or it could be used like a base for using different set of methods. REFII model proves that condition for automated pre-processing and automated analyzing of time series are not powerful computers as it was mentioned by Pyle [Pyle, 2001]. Condition for automated pre-processing and automated analyzing of time series which integrate variety of methods for time series analysis, and which allow possibility for constructing new methods in domain of time series analysis, as REFII model does.

4. CONCLUSION

With REFII model as transformation model for time series, it is possible to create systems for automated series data preparation and systems for time series data mining based on different types of algorithms.

Instead of using many different incompatible methods for knowledge extraction from time series, with REFII model as a base it is possible to chain different conceptions, and using traditional data mining methods in time series. Main advantage of this approach is a possibility for constructing powerful time series query languages which enables a unification of different methods, and construction of new algorithms for time series analysis based on contingency. REFII has possible applications in many different areas such as finance, medicine, voice recognition, face recognition and text mining. Future work will be focused on constructing systems for solving complex problems in domain of marketing time series analysis which demand chaining of methods, constructing of new algorithms, and using of traditional data mining methods in time series analysis through usage of REFII model of transformation. Within these systems, for example, it is possible to solve problems of :

- market segmentation based on customer behavior
- discovering client's characteristics, such as an intensive use of ATMs on Friday noon
- prediction of usage of a new service in an expected period of time within different market segments based on time series similarity
- forecasting of events and discovering of conditions in which events are occurring in specific period of time
- discovering of client's characteristics which show cyclic oscillations in behavior with probabilities of occurring events

It is not possible to solve the mentioned problems, or it is very hard to solve them by traditional approach in time series analysis.

REFERENCES

[Agrawal, 1995] Agrawal Rakhes, Lin K., Sawhney H. , Shim K.: Fast similarity search in the presence of noise, scaling and translation in time-series databases, Department of computer science university of Maryland, Paper presented on 21st VLDB conference Zurich, 1995.

[Bradley, 1997] Bradley Elizabeth: Time- series analysis, University of Colorado, Computer science technical report, 1997.

[Cheng, 1997] Cheng-Jian Lin: SISO nonlinear system identification using a fuzzy-neural hybrid system, International journal of neural systems, Vol. 8, No 3 (June,1997.) 325-337

[Fanchi, 2000] Fanchi, John ; Math Refresher for Scientists and Engineers, 2nd Edition, Wiley-IEEE Press; 2 edition (May 11, 2000)

[Han, 2000] Han Jiawei, Kamber Micheline: Data mining-concepts and techniques, Morgan Kaufmann publishers, 2001.

[Javor, 1988] Javor Petar, Uvod u matematičku analizu, Školska knjiga- Zagreb, 1988.

[Manilla, 1997] Heikki Mannila, Hanu Toivonen, Verkamo Inkeri: Discovery of frequent episodes in event sequences, University of Helsinki Finland, Report C-1997-15

[Manilla, 2001] Heikki Mannila, Johan Himberg, Kalle Korpiaho, Johanna Tikanmäki, Hannu T.T. Toivonen: Time series segmentation for context recognition in mobile devices, Nokia research centre, ICDM http://citeseer.nj.nec.com/himberg01time.html , 2001.

[Pyle, 2001] Pyle Dorian: Data preparation for data mining, Morgan Kaufmann publishers, Inc, 1999.

[Pratt, 2001] Pratt Kevin: Locating patterns in discrete time series, University of south Florida, M.Sc. these, 2001.

[Xsniaping, 1998] Xsniaping Ge: Pattern matching in financial time series data, http://www.datalab.uci.edu/people/xge/chart/, 1998.

Received: 5 July 2006

Accepted: 6 November 2007