Time series data analysis for long term forecasting and scheduling of organizational resources – few cases

Sunil Bhaskaran
Indian Military Academy, Dehradun, India

ABSTRACT

Our society is increasingly influenced by modern information and communication technology (ICT), Data warehouse, data mining and time series data mining etc. Time series data mining can be treated as a subset of data mining domain. The nature of time series data is large data size, high dimensionality and necessary to update continuously. Time series data mining (TSDM) is a rapidly evolving research area in Computer Science. While processing data stored in a data base, if we consider the time at which the event happened, the information technology professional can generate more reliable and dependable information in comparison with conventional methods. Potentially, today, every stake holders has got the opportunity for time series data mining. In this paper I am introducing a methodology and a strategy for the effective planning of various organizational resources for different stake holders in the form of cases. Few of them are Information Technology professionals planning their hardware, software and network (bandwidth) requirement for the organizations. Another category of user is top level business executives, they are responsible for the long term strategic decision making of their business. The next category of users I am planing to cover in this paper is medical professionals or biological researchers and share traders (equity market).

General terms

Patterns, Linear Interpolation, Current through resistor, Spikes in stock price, Trend analysis, general algorithm.

Keywords

Future and Option (F&O) segment; day trading, Time Series Data Mining, swap area, equity market filtering, smoothening.

INTRODUCTION

Data mining is also called knowledge discovery, the discovery of knowledge from an available source of data set. The traditional method of turning data into knowledge relies on manual analysis and interpretation. A time series data is a set of observations made in the chronological order. Data stored in financial, medical, weather forecasting and scientific databases are few examples of time series data. Such data changes with aspect to the change in time as given in figure 1, figure 2 and figure 3. A discrete time series is one in which the set T0 of time at which observations are made is a discrete set, as is the case for example when observations are made at fixed time intervals[1,2]. Continuous time series is obtained when observations are recorded continuously over some time interval, e.g. when T0 [0, 1]. We shall use the notation x(t) rather than x, if we wish to indicate specifically that observations are recorded continuously. Its techniques permit exploring large amounts of time series data in search of consistent patterns and/or interesting relationships between variables. Merging of human decision making process with machine process is still continuing as an open research problem. One of the big challenges of mining time series data is their size and dimensionality[3].

A time series X is "a sequence of observed data, usually ordered in time"

$$X = \{ x\,t, t=1,....N \} \quad (1)$$

where t is a time index, and N is the number of observations.

APPLICATION AREAS

Time series analysis is fundamental to engineering, scientific, health care research, manufacturing and business endeavors. Researchers study systems as they evolve through time, hoping to discern their underlying principles and develop models useful for predicting or controlling them. The increasing use of time series data has initiated a great deal of research and development attempts in the field of data mining.

Case 1 : Current through a resistor

If a sinusoidal voltage

$$v(t) = a \cos(vt + \theta) \quad (2)$$

is applied to a resistor of resistance r and the current is recorded continuously, we obtain a continuous time series as given bellow

$$x(t) = r^{-1} a \cos(vt+\theta) \quad (3)$$

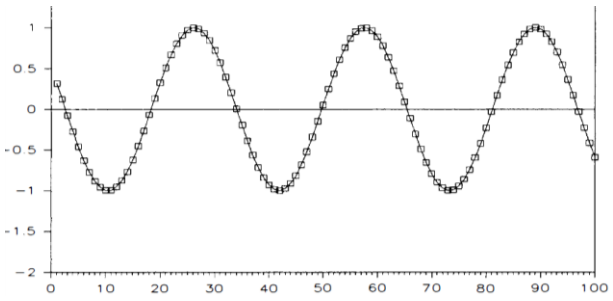


Figure 1 Time series of current through a resistor

1. 1 00 observations of the series

$$x(t)=\cos(2t+p/3) \quad (4)$$

It is of course but a few of the multitude of time series

to be found in the fields of engineering, science, sociology and economics.

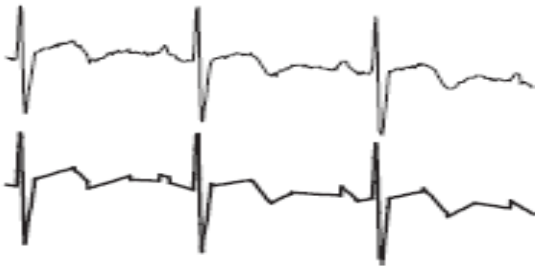


Figure 2 – A typical time series data and their piecewise liner representation

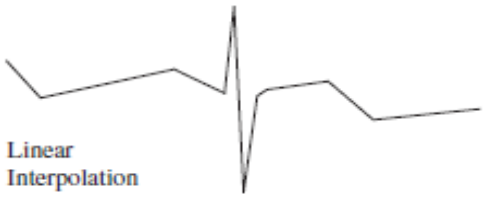


Figure 3 -Segment approximations of the above time series

Classification of Time Series

Depending on the character of the data that they carry, the time series could be
(a). Stationary and non-stationary.
(b). Seasonal and non-seasonal.
(c). Linear and nonlinear.
(d).Univariate and multivariate.

Time Series Modeling

In engineering, modeling of dynamic phenomena has long been seen as a valuable support tool for winning a deep insight into the structure and behavior of dynamic systems. Much research and development efforts have been made in development and application of system models.
The most commonly used approach in modeling stock market rate is to model the univariate time-series with autoregressive (AR) and moving average (MA) models. A trader can determine an appropriate number of lags for AR and ARMA by analyzing the time series data[4]. By analyzing the stock market, we know there are some combinations of the parameters, which can produce a near-max profit and give some reasonable buy/sell suggestions.

(a) Deterministic models

It is one in which every set of variable states is uniquely determined by parameters in the model and by sets of previous state of these variables. Therefore, deterministic models perform is the same way for a given set of initial conditions. Mathematically viewed as analytical models represented by

$$x(_t)= f(_t), \quad (5)$$

$$x_t=f(x_{t-1},x_{t-2}...) \quad (6)$$

(b) Stochastic models

Randomness is the main property of stochastic model. In it, variable states are not described by unique values, but by probability distributions. It is statistically viewed as functions of random variables[5]. If the starting points are determined, there are many combinations in which the process might progress. In such citations, some paths may be more probable and others less.

Multivariate Time Series Models

The observation values of some time series are multivariate, made up of components which themselves are observations of, some time series. Such multivariate values are presented as vector values x [x1 , x2 ,...x_n]T , and the entire set of multiple values as a matrix made up of individual observation vectors. This method can be used for forecasting, but not for structural analysis and policy evaluation. For a structural model to be estimated, certain restrictions are required about which variables are allowed to affect each other. It is not possible to estimate a simultaneous model where all variables are considered endogenous[6].

$$x= \begin{pmatrix} X_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots. \\ \dots & \dots & . & \dots. \\ x_{n-1} & x_{n2} & \dots. & xn_m \end{pmatrix} \quad (7)$$

The State-Space Model

The state variables can be considered as the smallest possible subset of the system variables that can represent the entire states of the system in any given time. The phrase system state transition function would be the transfer function representing the interaction between observed variables occurring in any given time that maps present values of the state variables (system state) to their future values. The strongest feature of state-space models is the existence of very general algorithm for filtering, smoothening, and predicting (Bay 1999). To apply the state-space methodology, a model must be expressed in the following format, called the discrete time state-space representation[5].

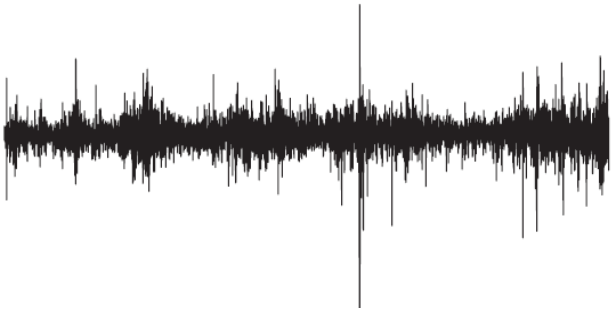
Case 2 : Spikes in stock price

Getting some knowledge about the probable spikes in stock market in advance is important to take a decision of the trading edge of the particular stock or the index as such in future and Option (F&O) segment, which will give a small advantage that allows greater than expeed gains to be realized. The stock will be bought at the open of the day and sold at the closing of the same day (trading). On observing the movements of a particular share, it has got four sets of ranges associated with its index value at opening, index value at closing, highest index value, lowest index value and trading volume.

$$X=\{x_t,t=1,...150\} \quad (8)$$

In business and financial engineering, and power distribution systems, mathematical models have been used for a long time. Time-series analysis plays a very important role in todays financial market and financial risk management. It forces the financial experts to make decisions based on observations. Financial time series consists of data collected from markets as given in figure 4. However, figure 5 gives the composite log range return of Bobbay stock exchange index, SENSEX from Jan 1980 to Jun 2010. On 24th Nov 2011 the index moved up to 15,858.49, up by 158.52 points, which is 1.01% of the previous days closing.

Figure 4. Performance of Bombay stock exchange index (SENSEX- India) 30 share index on 24/11/11. (Source, http://www.bseindia.com – accessed on 25/11/11).



1980 1990 2000 2010

Figure 5 : SENSEX (India) composite log range return from Jan 1980 to Jun 2010.

Time series analysis tool – The R language [11]

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms. It can be installed free of charge from www.r-project.org. R has many features in common with both functional and object oriented programming languages. In particular, functions in R are treated as objects that can be manipulated or used recursively. In common with functional languages, assignments in R can be avoided, but they are useful for clarity and convenience and hence, will be used in the examples that follow. In addition, R runs faster when 'loops' are avoided, which can often be achieved using matrix calculations as shown below. It is an expression language with simple syntax, however, it is case sensitive. R is particularly strong in plotting scientific graphs.

Array operation in R language

More generally, subsections of an array may be specified by giving a sequence of index vectors in place of subscripts; however if any index position is given an empty index vector, then the full range of that subscript is taken.

> x <- array(1:20, dim=c(4,5)) (Generate a 4 by 5 array).
>x

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	1	5	9	13	17
[2,]	2	6	10	14	18
[3,]	3	7	11	15	19
[4,]	4	8	12	16	20

i <- array(c(1:3,3:1), dim=c(3,2))

Reading data from files

Large data objects will usually be read as values from external files rather than entered during an R session at the keyboard. R input facilities are simple and their requirements are fairly strict and rather inflexible.

Generating a time series in R for air passenger booking and its trend analysis.

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
112	118	132	131	130	111	127	125	150	104	105	118
110	120	140	125	127	110	120	123	145	108	132	140
112	111	122	121	120	115	120	124	145	103	130	166
112	118	132	131	130	111	127	125	150	104	135	194
196	196	236	235	229	243	264	272	237	211	180	201
204	188	235	227	234	264	302	293	259	229	203	229
242	233	267	269	270	315	364	347	312	274	237	278
254	277	317	313	318	374	413	405	355	306	271	306
315	301	356	348	355	422	465	467	404	347	305	336
340	318	362	348	363	435	491	505	404	359	310	337
360	342	406	396	420	472	548	559	463	407	362	405
417	391	419	461	472	535	622	606	508	461	190	432

> plot (air, ylab = "Passangers (1000's)"

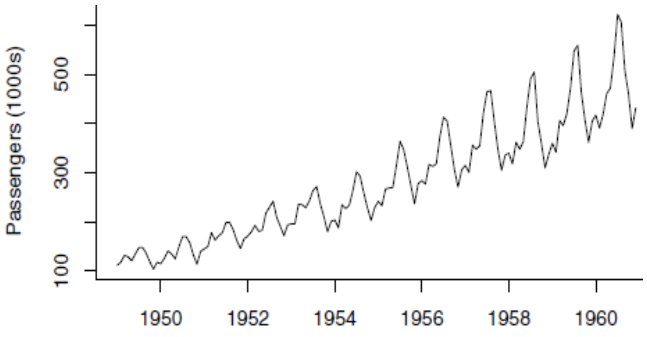


Figure 6 -The time series generated from air passenger data using R language.

A repeating pattern within each year is known as seasonal variation, although the term is applied more generally to repeating patterns within any fixed period, such as restaurant bookings on different days of the week. There is clear seasonal variation in the air passenger time series. An understanding of the reasons for the features in the plot helps us formulate an appropriate time series model. In the given case of air passenger the possible causes of the increasing trend may be rising prosperity among population, availability of aircraft, availability of low cost flights due to competition among flight operating companies, an increasing population and attitude of the people[7].

If we closely watch the given time series it many exhibits a unique property called trend. A a systematic change in a time series that does not appear to be periodic is known as a trend. Random, (stochastic) trends are common in economic and financial time series. A time series graph not only emphasises patterns and features of the data but can also expose outliers and erroneous values. The simplest model for a trend is a linear

increase or decrease, and this is often an adequate approximation.

Time series forecasting is based on extrapolation and forecasts are generally based on an assumption that the under consideration is going to continue in the future. Though this assumption is little unrealistic to trust. Yet, if we can identify likely causes for a trend under consideration, we can justify extrapolating it, for a few time steps at least.

Higher-order polynomials would give us more realistic result. But they should not be used for extrapolation. It is better to use linear extrapolation from the more recent values. Forecasts based on extrapolation beyond a year can described the scenarios in a better way. These days airline companies are using various trend analysis techniques for attracting passengers to its flight stream. They are also using such techniques for planning suitable aircrafts for crowded, less crowded and routes actively considered for introducing new flight services..

Case 3 : Computer system resource planning

Planning, scheduling, estimation and forecasting the computer systems, networks, and other hardware and software requirements is a big challenging operation for system engineers in any organization. The following example, about the anthers system utilization while preparing this paper is analyzed by applying time series based mechanism as a case of resource planning. The graph in figure 7 is the system resource utilization.

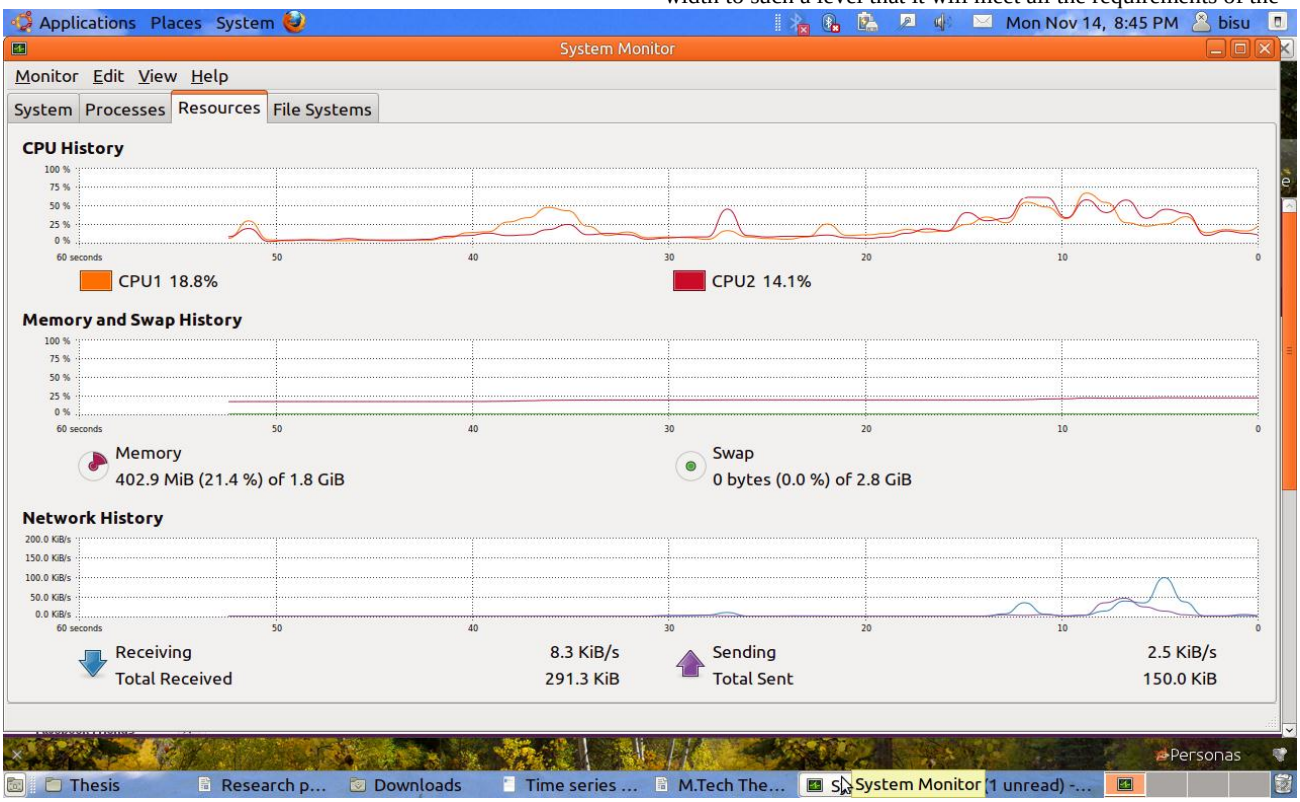


Figure 7 - Resource requirement and allocation of anthers computer stem (Screen shot)

Time series analysis of figure 7

Figure 5 is the screen shot of authors, computer running on Ubuntu (GNU/Linux system), Gnome desktop. It shows the allocation of resources such as the two cores of the Dual-core (1.7 Mhz), Swap area (2.8 GB), Memory (DDR3 - 700 Mhz) and the network activity. When the screen shot was observed CPU no.1 (Orange) was performing at a rate of 18.8 %, while the CPU no.2 (Red) was performing at a rate of 14.1 %. However the memory requirement was 21.4 % and the swap area requirement was 0%. The network activity remains steady for some time but increased sporadicly to 2.5 KiB/s (Send) and 150 KiB/s (received). Under the test condition there were three active documents, one browser, two folders and the system monitor running in the background. The above indicates the following

(i) Both processors at the given point of time remain under utilized – Hence more process can be successfully executed in the local processor as and when the need arise.

(ii) No need of increasing the capacity of the RAM or Swap as the RAM utilization is just bellow 25 %.

(iii) To get better performance of the network band width needs to be increased as it has a tendency of jamming the bandwidth through over utilization, sporadically.

Fixing the bandwidth is a different kind of problem than the other two. it's not just a technical problem, which could be resolved through a better method instead, it's a fundamental problem that relates to the data set itself. If the data follows a smooth distribu- tion, then a wider bandwidth is appropriate, but if the data follows a very wiggly distribution, then we need a smaller bandwidth to retain all relevant detail. In other words, the optimal bandwidth is a property of the data set and tells us something about the nature of the data. We need to fix the band width to such a level that it will meet all the requirements of the organisation, however the cost of the band is minimum, so that, the expenditure will be as low as possible.

Case 4 : Time series data mining in Biological data sets

Hospital databases have accumulated large quantity of data about patients and there status and response to treatments with respect to time. The mathematical understanding of estimation and

hypothesis formation in medical data may be a different activity. It is primarily treated as a patient care activity, and secondarily as a research area[8]. The justification for collecting medical data is to benefit the individual patient. Proteins and nucleic acids are the two major biological macromolecules forms the living organisms. Most important aspect of protein is its three dimensional structure. Modern biologists has become an information scientists. Since the invention of a DNA sequencing method by Sanger, public interest in ergonomics sequences have been growing. Today biological data mining is simply an analysis of data sets generated from biological objects like DNA, RNA, Chromosomes etc. If we consider the Protein Data Bank (http://www.rcsb.org), the database that collects all the 3D structures of biological macromolecules till now experimentally determined, at present, there are about 47,000 structures of proteins available. In biological objects like protein structure a term "fold" is used to categorize and classify the 3D structured object.[9]

FMRI and brain activity

The brain is a very complex system in which several regions are functioning through collaboration. Such cooperation among regions are known as functional integration[10]. It has been studied using fMRI techniques that leads to connectivity maps, figure 8. Mapping brain activity through maps generated from fMRI (Functional Magnetic Resonance Imaging or functional MRI fMRI) data requires knowledge and techniques from cognitive neuropsychology, physics, engineering and mathematics, particularly statistics. A pulse sequence program is required to acquire image data in the MRI's system. The most common approach to activation map computation involves a univariate analysis of the time-series associated with each (3D) voxel.

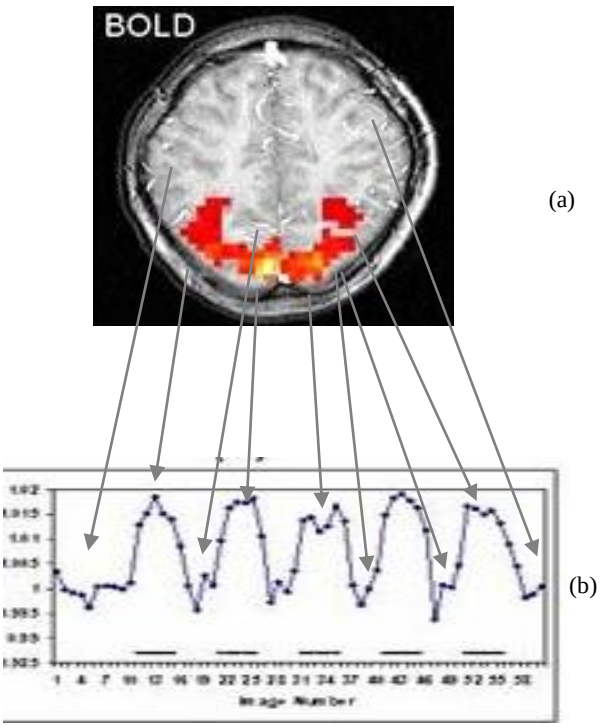


Figure 8 Anatomical image processing through fMRI
(a) Anatomical image under consideration.
(b) Corresponding time series generated through fMRI.

fMRI data is analyzed using multivariate techniques so that correlations between the time courses of the individual voxels are taken into account .

CONCLUSION

In this paper it has been discussed, how time series data analysis and mining techniques can be successfully used to plan the essential resources of an organization in details. The main advantage of the present research work, is the fact, that it analyses different cases related to time series data mining proposes and derives some conclusion for the prediction of key organizational resource so that they can be incorporated on a practical basis. Four cases such as current through a resistor, spikes in stock open price, Computer system resource planning, Time series data mining in Biological data sets are analyzed and discussed in detail. These case studies can effectively be used in other domains, such as voltage variations, fluctuations in ECG, fMRI, of a patient under observation in a clinic, in predicting climatic variation, flood forecasting, computer hardware, network, operating systems, etc. But all other areas other than what has been elaborated here requires data from their respective domains. It is of great importance for the policy makers to have reliable integrated information about the resources they are planning and forecasting every day. My present research has got the main advantage that researcher and the developer of the system has a vast number of training data at his disposal. As the continuation of this research, I am planning to consider few more domains like climate change, melting of snow in the Himalayan glacier etc.

ACKNOWLEDGMENTS

I would like to record my sincere gratitude to Dr. K.V Arya of ABV-IITM, Gwalior, India for sparing his valuable time and helping me in the successful completion of the research, compilation of this paper and its publication.

REFERENCES

[1] Sheng Chang, Wynne Hsu and Mong Li Lee . 2006. Mining Dense Periodic Patterns in Time Series Data, Proceedings of the 22nd International conference on data engineering (ICDE'06), 8-7695-2570-9/06, IEEE.

[2] Jose Zubcoff , Jesús Pardillo and Juan Trujillo . 2009. A UML profile for the conceptual modeling of data-mining with time-series in data warehouses. Information and Software Technology 51(2009) 977–992, Science Direct, ELSEVIER.

[3] Juan Trujillo. 2011. A review on time series data mining - Engineering Applications of Artificial Intelligence, 24 (2011) 164–181. ELSEVIER.

[4] Xiao Hu, Peng Xu, Shaozhi Wu, Shadnaz Asgari and Marvin Bergsneider . 2010. "A data mining framework for time series estimation. Journal of Biomedical Informatics" 43 (2010) 190–199. ELSEVIER.

[5] Das P. K, Maya Nayak , Senapati. M.R and Lee I.W.C. 2007. Mining for similarities in time series data using wavelet-based feature vectors and neural networks. Engineering Applications of Artificial Intelligence, 20 (2007) 185–201. Science Direct, ELSEVIER.

[6] Zhe Song, Xiulin Geng, Andrew Kusiak, and Chang Xu. 2011. Mining Markov chain transition matrix from wind speed time series data. Expert Systems with Applications xxx (2011) xxx–xxx. Science Direct, ELSEVIER.

[7] Chun-Hao Chen, Tzung-Pei Hong and Vincent S. Tseng . 2009. Mining fuzzy frequent trends from time series. Expert Systems with Applications, 36 (2009) 4147–4153. Science Direct, ELSEVIER.

[8] Huei-Wen Wu and Anthony J.T. Lee. 2009. Mining closed patterns in multi-sequence time-series databases. Data & Knowledge Engineering, 68 (2009) 1071–1090, Science Direct, ELSEVIER.

[9] Dash P.K, Behera H.S and Lee I.W.C 2009. Time sequence data mining using time–frequency analysis and soft computing techniques. Applied Soft Computing, 8 (2008) 202–215. Science Direct, ELSEVIER.

[10] Hailin Li and Chonghui Guo . 2011. Piecewise cloud approximation for time series mining. Knowledge-Based Systems, 24 (2011) 492–500. IEEE, Science Direct, ELSEVIER.

[11] W. N. Venables, D. M. Smith and the R Development Core Team , "An Introduction to R. Manual of R language" , Institute for Statistics and Mathematics, 2007.