EKONOMSKI HORIZONTI, 2011, 13, (1) str. 5-25

**Originalni naučni članak**
**005.94;**
**005.511:519.246.8**

Mr Marina Milanović*
Milan Stamenković*

# DATA MINING IN TIME SERIES

***Abstract:*** *In modern market economies, knowledge has become a key economic resource, and knowledge management by applying the concept of business intelligence has become the infrastructure constituent of modern management. Essential ingredient of the process of knowledge discovery from databases is data mining, which is successfully applied in various business and scientific research areas. Adaptive and innovative application of the principles and techniques of classic data mining in the analysis of time series resulted in the concept called time series data mining (TSDM). Starting from the above mentioned this paper presents conceptual determination of TSDM as a relatively new field of research in which the methodological context of data mining is adjusted to the temporal nature of data. It also emphasizes the complexity of mining in large time series data sets, as well as the importance and usefulness of research results in the form of extracted knowledge in decision–making process.*

***Key words:*** *knowledge discovery, data mining, time series, large datasets, business decision–making*

# DATA MINING U VREMENSKIM SERIJAMA

***Apstrakt:*** *U savremenim tržišnim uslovima privređivanja znanje je postalo ključni ekonomski resurs, a upravljanje znanjem primenom koncepta poslovne inteligencije infrastrukturni konstituent savremenog menadžmenta. Esencijalni ingredijent procesa otkrivanja znanja iz baza podataka je data mining, koji se uspešno primenjuje u različitim poslovnim i naučno-istraživačkim područjima. Adaptivna i inovativna aplikacija principa i klasičnih tehnika data mining-a u analizi vremenskih serija rezultirala je konceptom koji se naziva data mining vremenskih serija (eng. Time Series Data Mining – TSDM). Polazeći od navedenog, u Radu se prezentuju koncepcijska određenja TSDM, kao relativno nove oblasti istraživanja, u kojoj je metodološki kontekst data mining-a prilagođen vremenskoj prirodi podataka. Takođe, ukazano je na kompleksnost mining-a u*

* Faculty of Economics, University of Kragujevac

*velikim setovima vremenskih serija, kao i na značaj i korisnost rezultata istraživanja u formi ekstrahovanog znanja u procesu poslovnog odlučivanja.*

***Ključne reči:*** *otkrivanje znanja, data mining, vremenske serije, veliki setovi podataka, poslovno odlučivanje*

**JEL Classification:** C22, C81, C82

## 1. INTRODUCTION

The most significant change in the intellectual history of human society, and thus in the world of modern economy, concerns the role of knowledge, which is becoming a key business resource for survival and creation of the competitive advantage of the organizational entities, with implications on each industry, region and state. Therefore, knowledge management is an integral and an unavoidable component of modern management.

Satisfaction of the strategic need for knowledge of any organization is achieved by the synergy generated from the extraction and processing of information (including the information from databases) using advanced information and communication technologies, and innovative and creative capacities and talents of people, as the most useful organizational (corporate) resource. Accordingly, knowledge management based on the application of the concept of business intelligence is a prerequisite for gaining competitive advantage and survival in today's market, because the application of this concept contributes to a simpler solving of management problems and making quality business decisions.

The term ***K***nowledge ***D***iscovery in ***D***atabases (*KDD*), refers to *"the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data"* [Fayyad et al., 1996, pp 40-41]. The key steps of the interactive and iterative *KDD* process are presented in Figure 1. The essential ingredient in the process of knowledge discovery from databases is data mining. As an interdisciplinary research area, data mining is a connection between many disciplines, such as statistics, computer science and artificial intelligence, machine learning, database management, etc. Therefore, as a result of different research focus on different aspects of data mining, data mining can be defined in different ways. Comprehensive definition, often cited in literature, is:*"Data mining is the discovery of interesting, unexpected, or valuable structures in large data sets."* [Hand, 1999, p.433]. Thus understood, the term *data mining* is the basis for the perception of data mining as the process of identifying the significant patterns or models in data, for making, among other, crucial business decisions.

Data mining as a component of *KDD* process can be successfully applied in various fields. In this context, we can talk about business, scientific research and other data mining applications. In addition, many data mining problems include temporal aspects, and the most frequent form of presenting temporal data is time series. Adaptive and innovative application of the principles and techniques of classic data mining in the analysis of time series resulted in the concept called ***T***ime ***S***eries ***D***ata ***M***ining (*TSDM).* Unlike the traditional techniques for the time series analysis, and limiting assumptions, that they are based on, the methods in the *TSDM* network can be successfully applied in identification of complex characteristics, and prediction of non-periodic, non-linear, irregular, and chaotic time series.

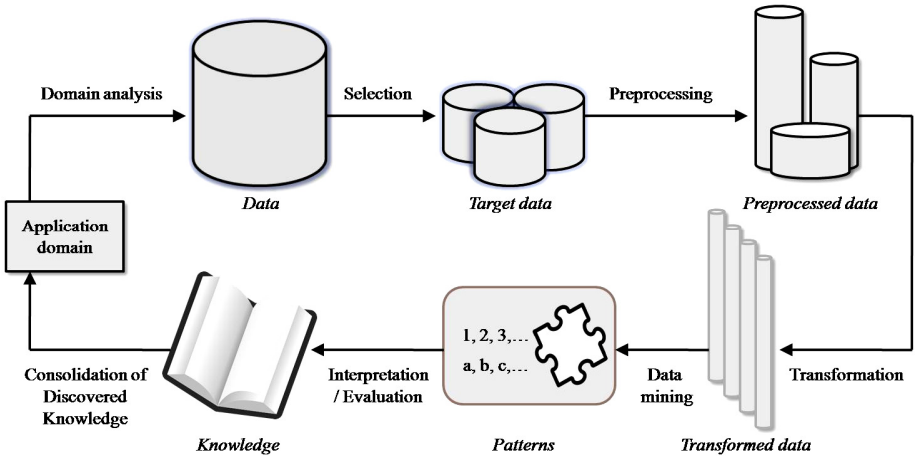



***Figure 1.*** The process of knowledge discovery in databases (*KDD*)
*Source:* authors' representation (adapted from Fayyad et al. 1996)

Starting from the above presented statements, the aim of this Paper is to stress out the importance and complexity of data mining methodology in the extraction of relevant information and knowledge discovery from large datasets of time series. After the referent literature review in Section 2, the concept and brief overview of an extended list of tasks for time series data mining are presented in Section 3. In Section 4, the importance of data preprocessing in data mining process is described. Similarity search and time series representation are presented in Section 5. In Section 6, considerations are focused on the fundamentals of the segmentation process and its role in reduction of dimensionality of high-dimensional time series. A special review of the importance of *TSDM* in economic research is the content of Section 7. In the last Section, the conclusions and outlines of further research are presented.

## 2. LITERATURE REVIEW

*TSDM* is a relatively new area of research in which the application of data mining techniques and methods is adapted to the temporal nature of the data. Increased interest in this research area has led to the publication of a significant number of scientific papers dealing with various aspects of the *TSDM*. One of the indicators of the increased interest in the field of time series analysis is the number of publications, containing the term *"time series"* in their title. Many of this papers are linked, directly or indirectly, with various aspects of data mining in datasets of time series. Based on the information available on the DBLP[1] site, Figure 2 illustrates the exponential growth of the above mentioned indicator until 2008. After this period, there has been a slight decrease in the number of published papers, which, in terms of research interest in this area, is not alarming, because the recorded decrease is minor compared to the indicative rapid growth recorded in the last fifteen years.

It is known that the analysis of time series is associated with the discovery of useful patterns and rules in the structure of time series and forecasting of the future values of the observed phenomena. Compared to the traditional analysis [Box et al. 1994], *TSDM* is implemented on a much larger amount of time series data, and / or much larger number of time series. Therefore, in *TSDM* applications, automated modeling is the only possible approach that can be applied to identify the complex characteristics of the high-dimensional datasets presented in the form of time series. To the best of the authors' knowledge after a thorough examination of the referent literature there is no paper that comprehensively describes all aspects of the *TSDM*. Basic definitions of the *TSDM* concept are presented in Povinelli (1999), Antunes and Oliveira, (2001), Keogh (2003), Keogh and Kasetty (2002), Faloutsos (2003), Last et al., (2004), Mörchen (2006), Keogh (2010).

[1] DBLP is a server that provides bibliographic information on major computer science journals and proceedings, and contains more than 10.000 links to home pages of computer scientists, (*http://www.informatik.uni-trier.de/~ley/db/).*

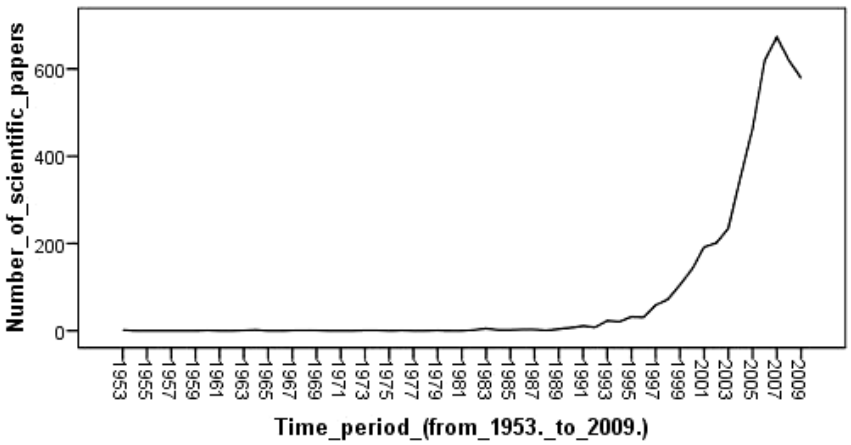



***Figure 2.*** Number of publications containing the term *"time series"* in their title
*Source of data: http://www.informatik.uni-trier.de/~ley/db/*

In addition, in the referent literature [Lin et al., 2005; Mörchen, 2006; Lin et al., 2007; Chundi and Rosenkrantz, 2009; Keogh, 2010], which refers to the study of time series using the data mining methodology, the following, essential, and methodologically linked typical tasks of the *TSDM* are emphasized: preprocessing, similarity search, clustering, classification, segmentation, visualization, anomaly detection, rule and motif discovery, and prediction (*forecasting* is the term commonly used in the analysis of time series). Listed tasks have a number of similarities with corresponding data mining tasks. However, the temporal aspect of data opens up a range of specificities and constraints in concrete applications, which are, with a detailed theoretical description and adequate empirical evaluation of numerous algorithmic methods for the realization of the *TSDM* tasks, a content component of the above listed literature. Panian and Klepac (2003), Keogh (2010), and Mörchen (2006) describe a number of procedures for preprocessing (temporal) data, emphasizing their crucial role in ensuring the quality of the results of the implemented time series data mining analysis. Also, the literature comprehensively presents and discusses, incorporating temporal dimension, the other tasks as well, as some of the major time series data mining areas. In addition, the role of segmentation of time series is especially emphasized (as a preprocessing step in time series analysis applications) in dimensionality reduction and extraction of patterns and rules in the behavior of the observed phenomena [Keogh et al., 2004; Gionis and Mannila, 2005; Bingham et al., 2006; Hiisilä, 2007; Chundi and

Rosenkrantz, 2009]. In close connection with the time series segmentation are time series representations, and an excellent review of many methods for creating time series representations can be found in Mörchen, (2006), and Lin et al., (2007).

In an attempt to popularize the valid application of the *TSDM* analysis results in solving real world problems, Keogh and Kasetty (2002), indicate on the presence of certain flaws in the empirical evaluation of the approaches and methods for realization of various *TSDM* tasks, that are proposed in the literature. In addition, they emphasize that flaws not only negatively affect the objectivity of the empirical evaluations, but also reduce the general usefulness of these papers. Making the distinction between the implementation bias and data bias, mentioned authors suggest the necessity of more careful and detailed benchmark analysis of the real time series, so that the formulated conclusions and analysis results could be generalized.

Therefore, the exploration of time series in data mining context is a relatively new and ever changing research field [Keogh, 2010]. The main source of information about achievements and developments in this research area are top-tier scientific conferences such as *ACM Knowledge Discovery in Data and Data Mining, IEEE International Conference on Data Mining and the IEEE International Conference on Data Engineering*. In addition to already listed papers, scientific and technical papers published in proceedings of these conferences are recommended to the readers interested in conceptual and methodological aspects of the *TSDM*. In fact, taking into account the long time it takes to publish the articles in scientific journals and natural tendency of researchers to present and submit their results to "public judgment" in the shortest possible time, in terms of types of publications, it is not surprising that the structure of published works in this field is dominated by conference proceedings.

For example, the percentage structure of published papers which in their title contain the term *"time series data mining"*[2] illustrates that approximately 76% of the papers have been published in conference proceedings, 18% in journals, and 6% were other types of publication. In addition, the attractiveness of knowledge discovery in high-dimensional time series, should not be seen only in the context of the research challenges in scientific community, but also in terms of usefulness of the obtained results in function of supporting the process of business decision–making, because, as a nugget of a gold is hidden beneath the earth or water, the nugget of (business) information is hidden in the data.

[2] Source of data for the presented percentage structure: *http://www.informatik.uni-trier.de/~ley/db/*

## 3. *TSDM* CONCEPT

Generally, time series is defined as set of real values of the observed variable that are arranged by chronological order in successive time periods. In the context of the discussion that follows, it is useful to define the time series as a sequence of time dependent values of the observed variable. Symbolically, the time series, as a set of $n$ pairs of data (data points) is represented as follows: $Q = \{(q_1, t_1), (q_2, t_2), \ldots, (q_n, t_n)\}$, where, (for $i = 1,2,...,n$), $q_i$ - value of observed variable, $t_i$ - time, $(q_i, t_i)$ - pair of data or data point. Consecutive parts (pieces) of time series, which are, in a way, time series in themselves, are called segments of time series, where each segment is composed of a certain number of pairs of data. Different classifications of time series are listed in the literature, depending on the applied criteria. One of them, based on the type of measurement scale, is division into numeric and symbolic time series [Mörchen, 2006, p.22]. A numeric time series is defined as time series with numerical values for each time point, and a symbolic time series as a time series with nominal or ordinal values for each time point. For the purposes of analysis, different procedures are used for conversion of a numeric into a symbolic time series, and *vice versa*.

In recent years, there has been a growing interest of data miners for mining of temporal data and application of data mining techniques in the analysis of time series datasets. This raises the question of essential differences between the classic and the data mining approach to the analysis of time series. The key difference primarily refers to a huge amount of data, which is the result of information – technological revolution and its implications on literally every area of life and business, and with which the time series data miners routinely encounter. Furthermore, complying with the new circumstances, data miners no longer focus exclusively on analysis and / or determination of the statistical properties of time series data, but insist on effective and efficient discovering of useful information from massive time series datasets.

Hence, the application of specific algorithmic data mining methods for extraction of patterns and rules from time series datasets, identification of their components, and prediction of future trends and movements, is the essence of time series data mining. In Figure 3, the examples of high–dimensional time series datasets from different areas are illustrated. It is meaningless to discuss manual implementation of the analysis process on such time series datasets. In other words, automated modeling based on entire range of algorithmic methods and procedures is the integral part of every mining process in temporal data. In general, crucial aspects of mining time series data focus on the goal of identifying movements and / or components which exist within the data through:

a) automated detection (discovery) of previously, or until now, unknown rules and patterns from large time series datasets;
b) automated detection of outliers;

c) automated prediction of trends and behavior of structural components.

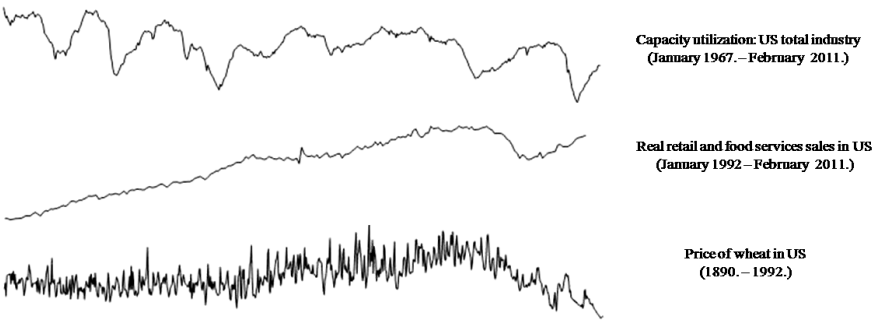



***Figure 3.*** The examples of time series
*Sources of data: (respectively):*
*http://research.stlouisfed.org/fred2/series/TCU/downloaddata?cid=3*
*http://research.stlouisfed.org/fred2/series/RRSFS*
*http://robjhyndman.com/tsdldata/data/9-9.dat*

Holistic approach to *TSDM* concept, in addition to these basic definitions, includes the consideration and discussion of *TSDM* tasks framework. Brief description of the *TSDM* primary tasks, as the most commonly used for discovering hidden relations between the observed time series data in function of forecasting of the future values, can be presented as follows.

***Preprocessing*** of data is a key component in providing data quality, and their preparation, according to the purpose of the analysis, for the application of data mining methods and techniques.

One of the unavoidable functions that follows literally every step in the data mining process is data ***visualization***, with which the miner, in a simple and efficient way, acquire the necessary guidelines, critical for the selection of direction in the further analysis.

***Anomaly detection*** refers to an identification of those parts of time series that are characterized by a specific, different behavior pattern that does not fit the model of data „normal“ behavior (i.e. expected behavior pattern).

Based on the use of some similarity/dissmilarity measure, ***similarity search*** refers to finding (identifying) the most similar subsequences in time series or time series in a large database to a given query subsequence or query time series.

***Classification*** is a task that consists of creating a classification model (or a function) for positioning of any new entity (i.e. time series) according to its features into one of two or more predefined classes.

The essence of ***clustering*** problem, as an optimization problem, can be described as follows: If *C* is a collection of *n* entities (time series or subsequences), perform the division of *C* into *m* independent groups of entities (clusters) *Cj*, (where, *j = 1, 2,…, m*), but in such way that entities within the same group are similar to each other while the entities that belong to different groups differ.

***Segmentation*** of time series provides a significant dimensionality reduction by dividing (i.e. splitting) the time series into appropriate, internally homogenous parts (pieces, or segments), but in such way that the number of segments, *k*, is considerably smaller than the number of data points in original time series, *n*.

Unlike the typical similarity search tasks and mining through a set of time series in order to find the presence of the predefined pattern, ***motif discovery*** relates to a detection of previously unknown, frequently occurring patterns (called motifs).

***Rule discovery*** task can be defined as a problem of finding rules relating the behavior of patterns in a time series to other patterns in that series, or patterns in one series to patterns in another time series [Das et al., 1998].

***Prediction*** includes forecasting of the future values of time series, using different methods depending on the type of time series (numeric or symbolic).

In order to achieve and ensure the sophisticated results of conducted data mining application in solving the real problems, mainly, it is generally necessary to use and rely on the various combinations of these tasks. Additionally, in the realization of the basic idea of the methodological aspects of *TSDM*, preprocessing, similarity search and segmentation have a primary role. Accordingly, it can be said that these tasks have the status of "full–time members" in almost all *TSDM* applications.

## 4. PREPROCESSING

Preprocessing of data whose analysis will be carried out is of crucial importance for the usefulness and validity of the derived conclusions in the context of defined research objectives. Data, absorbed from existing databases, data warehouses and data marts (i.e. internal and external sources), can be, in terms of mathematical and logical correctness, incomplete, unsystematic and inconsistent. Therefore, the importance of the preprocessing activities, considering the fact that the data quality is crucial factor for successful analysis, is reflected in identification, elimination, or reduction of deficiencies of the original (source) data. Generally, in data mining process, preprocessing activities, where each of them is based on the application of appropriate methodological procedures for incorporating the temporal dimension, include the following [Kamel, 2009, pp 538-543]: ▸ data selection (data sampling); ▸ data reorganization (data summarization,

complex data manipulation); ▸ data exploration (summary statistics, data visualization, OLAP[3]); ▸ data cleansing (anomaly detection, noise reduction, missing values analysis, determination of data consistency); ▸ data transformation (data recording, data smoothing, data aggregation, data generalization, functional transformation, grouping values, data normalization, feature construction, feature selection).

When preprocessing activities are conducted, a strict care should be taken of their association with the area of research, the objectives of analysis, and assumptions underlying data mining methods and techniques. Otherwise, badly preprocessed inputs cause poor quality of output (*GIGO effect*). Therefore, it is quite understandable, as many research studies point out, that the implementation of preprocessing activities engages between 60% and 90% of data miners' total working time in particular data mining project, [Pyle, 1999].

The component of data preprocessing that attracts great attention of the researchers from different fields is anomaly detection. Conceptually, anomalies represent the cases that do not fit the model of data "normal" behavior, so that, the anomaly detection refers to detection of irregularities in data and identification of previously unknown patterns, so called anomalous (or surprising, outlying, novel) patterns. In addition, anomaly detection task refers not only to the detection of extremely different values in dataset, but also to detection of more moderate and milder forms of deviations or differences in behavior. The above mentioned is confirmed by the terms that are used as a synonym for anomaly detection, such as interestingness / deviation / surprise / novelty detection, etc.

In the preprocessing of data, often, some essential features of the analyzed datasets can be identified with appropriate and adequate graphical displays. Visualization methods increase the level of intelligibility of the analyzed data and ensure or facilitate that some hidden or unnoticed features become apparent, because people have a remarkable ability to detect hidden patterns, and a human eye is unsurpassed data mining tool. These methods are successfully used in the detection of outliers in the data. There are many visual displays of data, such as: box-plots, scatter-plots, 3D-cubes, data distribution charts, curves, volume visualization, surface or link graphs, etc. [Viktor et al., 2009, p.2057]. The simplest form of visualization of time series is a line diagram, and one, novel visual display is Viz-Tree [Lin et al., 2005, p.61], useful for different *TSDM* tasks, and especially for the anomaly detection. The essence of the application of visualization techniques is revealed the following data mining phrase: *"A picture is worth thousands of numbers!"*

Preprocessing of time series data takes much more time and knowledge than that of non-temporal data. In fact, there are no fully automated procedures for

[3] OLAP – On-Line Analytical Processing

either the act of preprocessing or the selection of the optimal way of conducting the preprocessing activities. Therefore, the crucial factor for the selection of the optimal solution is an expert knowledge and experience of a data miner.

## 5. SIMILARITY SEARCH AND TIME SERIES REPRESENTATIONS

***Similarity search***, as a special task in *TSDM*, refers to finding similarities in the behavior of various time series. The other term, which is frequently used as a synonym for similarity, is distance (big distance – small similarity). Essentially, the similarity problem can be interpreted as follows: Given two time series, $Q$ and $C$, similarity search between them includes defining and determining a similarity function $Sim(Q, C)$ or, equivalently, a distance function $Dist(Q, C)$. Similarity search is a complex task. There are different types of similarity, depending on whether its identification is based on the observation of entire time series or parts of time series. In fact, this task can be divided into two categories [Lin et al., 2005, p.62]: ▸ whole matching (find a sequence that is similar to the query sequence), and ▸ subsequence matching (find all pairs of similar sequences).

Similarity search is based on similarity measures. Literature offers a variety of measures and their classifications stemming from different criteria. Starting from the mentioned classification of time series into numeric and symbolic series, one of the possible classifications of similarity measures is shown in Figure 4. In practice, the selection of specific methodological basis and measures for similarity search is based on characteristics of time series that are compared (length of time series, difference between their lengths, presence of outliers or noisy regions, data miner's prior knowledge about the structure of time series data, etc.).

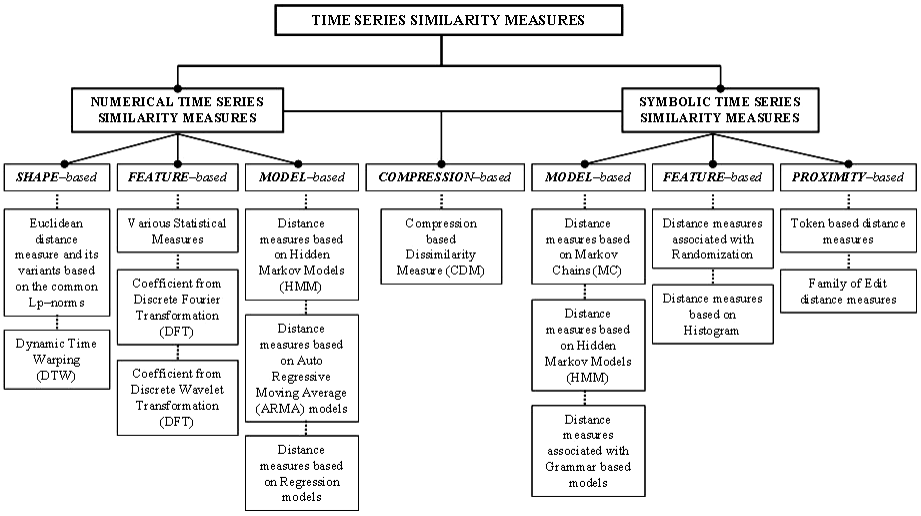



***Figure 4.*** Classification of time series similarity measures
*Source:* authors' representation (adapted from Mörchen, 2006, pp 27-28]

The similarity measure most commonly used to test the similarity between original time series, or evaluate the accuracy of the modeled forms of time series, is Euclidean distance (or, some of its derived form). Euclidean distance is based on comparison and determination of a difference (distance) between actual values in the $i^{th}$ point of one time series and modeled values or actual values in the $i^{th}$ point of another time series. Given two time series $Q$ and $C$ of length $n$, the Euclidean distance between them is defined as [Keogh et al., 2005, p.227]:

$$Dist(Q,C) = \sqrt{\sum_{i=1}^{n}(q_i - c_i)^2},$$

where $C = (c_1, c_2 ..., c_n)$, and $Q = (q_1, q_2, ..., q_n)$.

In order to perform comparison of the time series datasets with different characteristics, Euclidean distance is usually calculated after the transformation of raw data by the Z–normalization process (normalization to zero mean and unit standard deviation). The sequences $C$ and $Q$ are replaced by the Z–normalized sequences $C'$ and $Q'$, respectively:

$$c_i' = \frac{c_i - \overline{X}_C}{S_C}, \text{ and } q_i' = \frac{q_i - \overline{X}_Q}{S_Q}.$$

The relevance of choosing the supreme similarity measure is also directly connected with the relevance of the effects of other *TSDM* tasks and procedures, where a strong and unambiguous conceptual connection with the representations of time series should be pointed out.

The basic idea of time series ***representation*** in the form different than the original datasets is to provide concise display and clear notion of their basic characteristics through the appropriate approximation forms with minimal loss of relevant information. Typically, time series are characterized by high dimensional data, and therefore, working directly with raw data is very wasteful in terms of efficiency of storage, transmission, and processing of data. Therefore, it is necessary to develop, on an adequate methodological basis, representations that will reduce dimensionality and eliminate the noise from the original time series. In addition, for the same dataset, different representations can be used. In the literature, many different methods have been proposed for representation of time series. Figure 5, inspired by the researches of Eamonn Keogh and Jessica Lin[4] illustrates their detailed classification. For example, the essence of ***P***iecewise ***L***inear ***A***pproximation (*PLA*), as one of the most commonly used representation, is reflected in the approximation of the original time series of length *n*, over *k*, non-overlapping pieces (i.e. segments) presented in a form of *k* straight lines.

The ways for achieving such an approximation are *linear interpolation*, and *linear regression*. In case of linear interpolation, approximated straight line of a certain segment is a straight line which connects the starting and ending point of the segment, where the ending point of a previous segment is, at the same time, a starting point for the next segment, i.e. the starting point of approximated straight line of the next segment. Unlike the linear interpolation which results in connected segments, the result of implemented linear regression are disconnected segments, represented by straight lines, which are determined as the best fitting lines in least square sense. It is often pointed out, that the aesthetic superiority of linear interpolation along with the simplicity of its computation makes it very convenient and attractive to use. However, the quality of approximated straight lines (the approximation quality), from the view point of Euclidean distance, is fully on the side of the approach that is based on linear regression [Keogh et al., 2004].

[4] For more, see: http://www.cs.ucr.edu/~eamonn/; and *http://www.ise.gmu.edu/~jessica/*

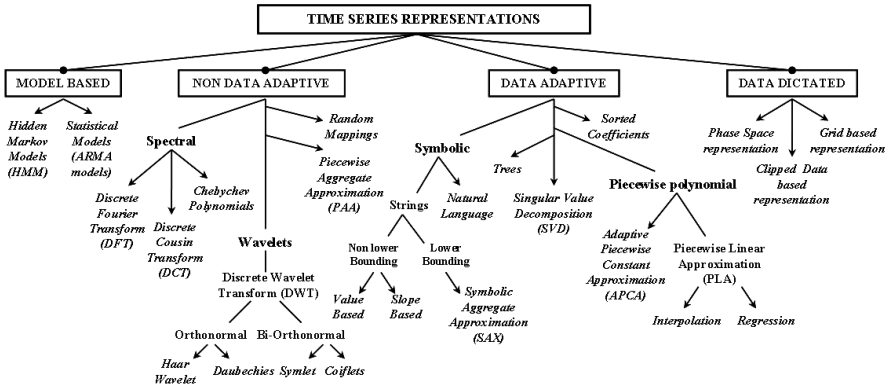



***Figure 5.*** Categorization of time series representations
*Source:* authors' representation (adapted from Mörchen, 2006, and Keogh et al. 2004)

## 6. SEGMENTATION

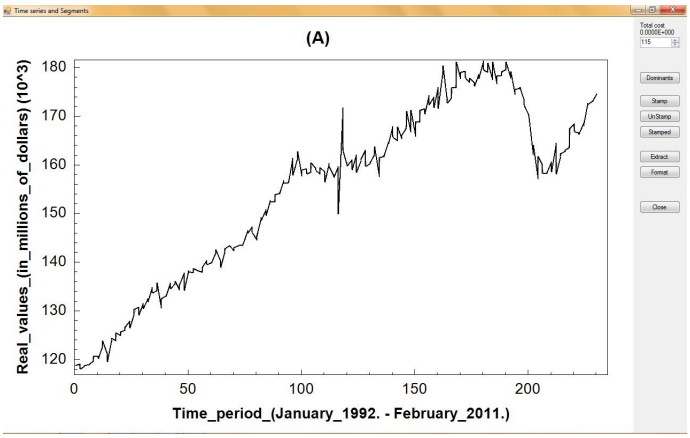
Time series segmentation is a fundamental component in the process of analysis and research of time series data. As a data mining research problem, segmentation focuses on dividing the time series into appropriate, internally homogenous segments, so that the structure of time series, through pattern and / or rule discovery in the behavior of the observed variable, could be revealed.

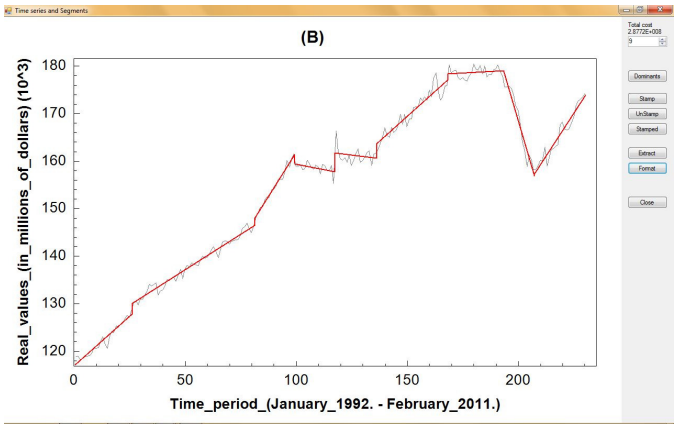
It has been mentioned already that adjacent parts of time series represent its segments, which consist of a certain number of adjacent pairs of data (i.e. data points). The process of conversion of the actual time series into its segmented version leads to reduction of the dimensionality of original values, in a way that the data points within the particular segment are presented, either with: (a) one specific value that represents them (e.g. mean, or median of a segment), or (b) model that is suited to data points within the segment. The representation of time series as a sequence of segments presented in the form of straight lines is illustrated in Figure 6. Straight lines can be determined in a different ways, as shown in Figure 6, where curve (A) refers to the original time series – *real retail and food services sales in USA, within the time period from January 1992 to February 2011*, curve (B) refers to its segmented version, where segments are represented by a model that is suited to data, and curve (C) refers to its segmented version also, but where segments are represented by one specific value.

However, this process should also be viewed in terms of achieved level of accuracy of approximate representation of original time series. Accuracy of approximate representation of time series is measured by using the appropriate error function. The most commonly used measure of accuracy is Euclidean

distance. Therefore, the essence of segmentation problem is reflected in finding the optimal approximation for which the error function is minimal. In fact, the optimal segmentation of time series, for the defined parameters is defined as the segmentation that results in the lowest segmentation error in relation to other possible combinations of segmentation. Finding the optimal solution is a complex iterative process composed of a series of sub-processes, phases, and activities that transform inputs into output elements. Therefore, the time series segmentation process consists of the following phases:

- ▸ feature reduction of input time series;
- ▸ segmentation of extracted time series;
- ▸ selection of the optimal segmentation.



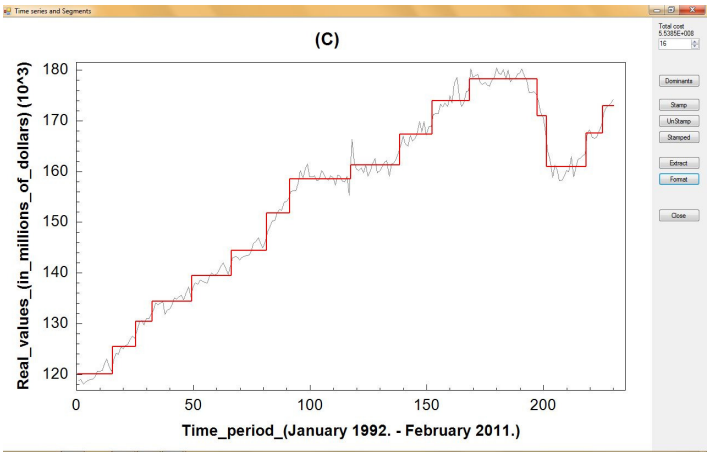



***Figure 6.*** The original time series (A) and its segmented approximations (B, and C)
*Source:* authors` representation, Segmenter 2.1 software output
*Source of data:* http://research.stlouisfed.org/fred2/series/RRSFS.

The realization of the mentioned process is based on entire range of algorithmic methods and procedures. However, given the variety of ways of formulating the segmentation problem in terms of defining the key parameters (number of segments, segmentation starting point, length of segments, error function, user-specified threshold, etc.), it is obvious that the universal algorithm, which in all cases results in optimal solution, does not exist[5]. Independently from the applied algorithm, through reduction of dimensionality, the process of segmentation provides a compact representation of the underlying time series data, which is more suitable for discovering relevant and interesting information in the data.

## 7. THE IMPORTANCE OF *TSDM* IN ECONOMIC RESEARCHES

It is generally known that the time series are relevant for the analysis of the phenomena in different research areas, such as medicine, chemistry, geology, meteorology, economics, etc. Therefore, all aspects of mining time series data are focused on identification of rules and patterns in the movement of the components

[5] In spite of their differences in terms of efficiency, application complexity, and specific implementation details, as well as the names under which they have been listed in the referent literature, the most of segmentation algorithms can be classified into one of the following categories of algorithms: ▸ Top–Down algorithm; ▸ Bottom–Up algorithm; and ▸ Sliding Window algorithm. For details see, Keogh et al., 2004.

that characterize the time series (including trend movements, seasonal variations, cyclical variations, and random movements), in order to, on the basis of understanding the past, predict the future behavior of the observed variable. In most cases, for the time series, that contains *n* data points, the prediction of the value at time *n+h* (where *h* is the length of the step in prediction) is based on combinations of outcomes of other *TSDM* tasks.

Different prediction models were known even before the appearance of data mining and computers. The earliest prediction models are related to the classical decomposition models. Over time, the level of methodological complexity and consequently the level of accuracy in prediction, have increased from simple moving averages and exponential smoothing methods to sophisticated ARIMA models[6] (and their variants, such as seasonal ARIMA models, vector ARIMA models using multivariate time series, etc.), with the inclusion of elements of regression analysis. Today, at the time of information–technological revolution, the entire arsenal of highly sophisticated algorithmic methods and procedures for the analysis of time series has appeared (from segmentation techniques for reduction of dimensionality to neural networks[7] and fuzzy logic[8]). However, extraction of relevant information and knowledge from large time series datasets, regardless of the research area to which they relate, is practically impossible without the application of appropriate software solutions.

There are many companies that offer software solutions for data mining. Popular software packages (as a collection of data mining algorithms), which include tool-boxes or modules for time series analysis are: *WEKA* (*W*aikato *E*nvironment for *K*nowledge *A*nalysis), *Rapid Miner* (formerly Yale), *IBM PASW*[9]

[6] ARIMA models – Autoregressive Integrated Moving Average models

[7] Artificial neural networks are the techniques that have been formed on the basis of the superior learning processes in the human brain. In addition, the principles of functioning of the human brain and its structure are used, so that, through the learning process, and based on the preprocessing of historical data, conducted modeling and testing of the created model, the future value of some observed phenomenon could be predicted.

[8] There is an entire range of algorithms that are based on the concept of *fuzzy* logic, where the principles of fuzzy logic are incorporating in the models of time series, usually with the intention of creating a valid model for prediction.

[9] According to the research conducted by *Gartner*, leading consulting company on the field of information technology, SPSS's predictive analytics, is used by 95% of the 1000 largest companies from the list of the largest companies of the business magazine *Fortune*. Among the users of the SPSS's solutions is a top 10 global banks, 8 of the 10 largest telecommunications companies on the world, 21 of the 25 largest trading companies, 24 of the 25 leading research agencies. Also, according to the research *„2008 Data Mining Survey"*, conducted by the company *Rexer Analitics* in 44 states and in which participated 348 analysts from different industries, SPSS's tools are first on the list of the most commonly used analytical tools for the needs of data mining. As one of the main reasons of such positioning of the SPSS's tools, analysts state their help in improvement of the

(formerly SPSS-Clementine), *SAS* (Enterprise Miner), and *MATLAB* (*mat*rix *lab*oratory). Mentioned software solutions support only some aspects of statistical time series analysis, and this is considered to be their main fault. Therefore, the discovery of the structure of time series datasets must be based on a compilation of available software solutions and, of course (by default), expert knowledge of data miners.

In fact, the inherent property of the economic phenomena is high dimensionality. Reduction of the actual dimensionality of huge amount of raw data stored in databases to a much lower level through high quality representations and abstractions helps and contributes to the identification of the key and relevant patterns and rules. Hence, data mining application, with the elements of the time series analysis, can be applied in the field of economic research. In many economic domains, such as, stock market analysis, budgetary analysis, analysis and fraud detection in banking and finance, analysis of fluctuations in the level of product quality, analysis of market segmentation and customer retention, inventory forecasting, sale forecasting analysis, insurance, retail, etc., hidden business information of strategic importance is being revealed thanks to the analysis process based on the conceptual framework of data mining. In addition, practice shows that for the high–quality analysis of economic phenomena, which should result in the useful information it is not sufficient to use only one method. Linking different methods of *TSDM*, as well as the methods of *TSDM* and classical data mining, results in a synergic effect embodied in the high–quality input parameters in the process of decision–making.

Additional complexity of the presented issues faced by the business entities in modern market conditions can also be described, metaphorically, by the following data mining phrase: *"swim in a sea of data and be thirsty for useful information"*. The gap, between the amount of collected data and their usefulness as a input in the process of decision–making makes data mining a key source of knowledge in the function of acquiring distinctive competitive advantage. The necessity of applying data mining has led to the creation of a completely new profession in today's business world, called "*data miner*". Expertise in the fields of information systems and statistics, i.e. the science of data, is the competence that each data miner must have. Many companies in solving real data mining problems have difficulties in finding employees who possess the required multidisciplinary knowledge, and therefore, in spite of job vacancies published in the business ads, they hire specialized agencies, which provide the data mining consulting services, as well as the expert companies, engaged in development of software intended for data mining researches.

operational processes, and solving of key business problems, to support in making decisions of the strategic character. (*Source: http://www.economy.rs*)

## 8. CONCLUSION

Knowledge discovery in large datasets of time series by idenfying the interpretable, novel, and useful temporal patterns, is practically not possible without the application of data mining methodology. Mining in time series data provides insight into the hidden, unknown, unexpected and useful information and knowledge, whether it is about the relationships between data, patterns of behavior, correlations, or rules, which improve the process of making strategic decisions, based on the clear and specific interpretation of business results.

In the analysis of time series, research interest focuses on the following analyses: trend analysis, analysis of cyclical and seasonal variations, analysis of similarity within the segments of the series, analysis of the correlation relationships between both time series and its parts, analysis and discovery of links (relationships) between time series and its parts with the respective corresponding market trends, autocorrelation analysis, etc. In general, all aspects of mining time series data are focused on identification of rules and patterns in the movement of the components that characterize the time series, in order to, understanding the past behavior, predict the future behavior of observed variable and future business outcomes. Decision makers who have the access to these predictions are able to make highly informed proactive decisions.

On the basis of available literature, this Paper presents a conceptual framework of *TSDM*, emphasizes the role of data preprocessing, points out to the significance of the similarity search, and reviews segmentation problem in function of dimensionality reduction of time series data. Future research will be based on the empirical evaluations and discovery of hidden information in the structures of the real economic time series datasets through appropriate software applications.

***References:***

1. **Antunes, C. and Oliveira, A.**, "Temporal Data Mining: An Overview", *Proceedings of the Workshop on Temporal Data Mining at the 7th International Conference on Knowledge Discovery and Data Mining* (KDD'01), ACM Press, 2001, pp 1–13.
2. **Bingham, E., Gionis, A., Haiminen, N., Hiisilä, H., Mannila, H. and Terzi E.**, "Segmentation and Dimensionality Reduction", Proceedings of the *SIAM International Conference on Data Mining 2006*, 2006, pp 370–381. Retrieved from: http://www.cs.helsinki.fi/u/mannila/
3. **Box, G., Jenkins, G. M. and Reinsel, G.**, *Time Series Analysis: Forecasting and Control*, 1994, Prentice Hall, 3rd edition.

4. **Chundi, P. and Rosenkrantz, D.**, "Segmentation of Time Series Data", *Encyclopedia of Data Warehousing and Mining*, J. Wang (Ed.), Information Science Reference, New York, USA, 2009, pp 1753–1758.
5. **Das G., Lin I. K., Mannila H., Renganathan G. and Smyth P.**, "Rule Discovery From Time Series", *International Conference of Knowledge Discovery and Data Mining*, New York, USA, 1998.
6. **Faloutsos, C.**, "Mining Time Series Data", Tutorial ICML 2003, Washington DC, USA, 2003.
7. **Fayyad, U., Shapiro, G. P. and Smyth, P.**, "From Data Mining to Knowledge Discovery in Databases", AI Magazine, 17(3), 1996, pp 37–54.
8. **Gionis, A. and Mannila, H.,** "Segmentation Algorithms for Time Series and Sequence Data", Tutorial at 5th SIAM International Conference on Data Mining 2005, (2005).
9. **Hand, D. J.**, "Why Data Mining is more than Statistics Writ Large", *Bulletin of the International Statistical Institute*, 52nd Session, Vol. 1, 1999, pp 433–436.
10. **Hiisilä, H.**, *Segmentation of Time Series and Sequences Using Basic Representations*, Ph.D. Thesis, 2007, Helsinki University of Technology, Laboratory of Computer and Information Science, Helsinki, Finland. Retrieved from: http://cis.legacy.ics.tkk.fi/heli/lisuri.pdf
11. **Kamel M.**, "Data Preparation for Data Mining", Encyclopedia *of Data Warehousing and Mining*, Wang J. (ed.), Information Science Reference, New York, USA, 2009.
12. **Keogh, E. and Kasetty, S.**, "On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration", *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, 2002, pp 102–111.
13. **Keogh, E.**, "Data Mining and Machine Learning in Time Series Databases", *Tutorial ECML/PKDD-2003*: Fourteenth European Conference on Machine Learning and Seventh European Conference on Principles and Practice of Knowledge Discovery in Databases, 2003, Cavtat-Dubrovnik, Croatia.
14. **Keogh, E., Chu, S., Hart, D. and Pazzani, M.**, "Segmenting Time Series: A Survey and Novel Approach", *Data Mining in Time Series Databases*, M. Last, A. Kandel, and H. Bunke (Eds.), World Scientific Publishing Co. Pte. Ltd., Singapore, 2004, pp. 1–21.
15. **Keogh, E., Lin, J. and Fu, A.**, "HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence", *Proceedings of the 5th IEEE International Conference on Data Mining* (ICDM 2005), Houston, Texas, 2005, pp. 226–233.
16. **Keogh, E.**, "Data Mining Time Series Data", I*nternational Encyclopedia of Statistical Science*, M. Lovrić (Ed.), Springer, New York, USA, 2010.
17. **Last, M., Kandel, A. and Bunke, H.**, editors, *Data Mining in Time Series Databases*, 2004, Singapore: World Scientific Publishing Co. Pte. Ltd.

18. **Lin, J., Keogh, E. and Lonardi, S.**, "Visualizing and Discovering Non-Trivial Patterns in Large Time Series Databases", *Information Visualization Journal*, 4, 2, 2005, pp 61–82.
19. **Lin, J., Keogh, E., Wei, L. and Lonardi, S.**, "Experiencing SAX: a Novel Symbolic Representation of Time Series", *Data Mining and Knowledge Discovery*, 2, 2007, pp 107–144.
20. Retrieved from: http://www.citeulike.org/user/lenov/article/2821475
21. Mörchen, F., Time Series Knowledge Mining, Ph.D. Thesis, 2006, Philipps University, Marburg, Germany. Retrieved from: http://www.mybytes.de/papers/moerchen06tskm.pdf
22. **Panian Ž., & Klepac G.**, *Poslovna inteligencija*, 2003, Masmedia, Zagreb.
23. **Povinelli, R. J.**, *Time Series Data Mining: Identifying Temporal Patterns for Characterization and Prediction of Time Series Events*, Ph.D. Thesis, 1999, Marquette University, Faculty of the Graduate School, Milwaukee, Wisconsin. Retrieved from: http://povinelli.eece.mu.edu/publications/papers/dissertation.pdf
24. **Pyle, D.**, *Data Preparation for Data Mining*, 1999, Morgan Kaufmann, San Francisco, USA.
25. **Viktor, H. L. and Paquet, E.**, "Visual Data Mining from Visualization to Visual Information Mining", *Encyclopedia of Data Warehousing and Mining*, Wang J. (ed.), Information Science Reference, New York, 2009.
26. http://www.cs.ucr.edu/~eamonn/ (date of visit: almost every day)
27. http://www.ise.gmu.edu/~jessica/ (date of visit: almost every day)
28. http://www.economy.rs (date of visit: almost every day)
29. http://www.informatik.uni-trier.de (date of visit: January 15, 2011.)
30. *Time Series Data Library* (October 8, 2010)
31. Retrieved October 8, 2010, from: http://robjhyndman.com/tsdldata/data/9-9.dat
32. St. Louis Fed: Economic Research, (2011).
33. Retrieved March 3, 2011, from: http://research.stlouisfed.org/fred2/series/TCU/downloaddata?cid=3
34. St. Louis Fed: Economic Research, (2011).
35. Retrieved March 3, 2011, from: http://research.stlouisfed.org/fred2/series/RRSFS