INTERNATIONAL BLACK SEA UNIVERSITY

FACULTY OF COMPUTER TECHNOLOGIES AND ENGINEERING

PhD. PROGRAM

DESIGN OF PIECEWISE LINEAR APPROXIMATION METHOD AND CRITERIA FOR ESTIMATING SIMILARITY OF NON STATIONARY TIME SERIES

Daniiar SATYBALDIEV

Extended Abstract of Doctoral Dissertation in Computer Science

Tbilisi, 2016

Scientific Supervisor: Alexander Milnikov

(Professor, Doctor, at International Black Sea University)

____

(supervisor's signature)

Expert:

1. Prof. Dr, Nodar Momtselidze____________________________________

2. Assoc. Prof. Dr. George Mandaria_______________________________

Opponents:

1. Prof. Dr. Guram Lezhava______________________________________

2. Assoc. Prof. Dr. Nikoloz Abzianidze_____________________________

3. Assoc. Prof. Dr. Nurlan Atabaev_______________________________

INTRODUCTION

Big data sets are most useful and carry significant information in the importance of statistics. The question is what we can do with them and how we can analyze. Giant companies such as Apple and Google are masters in analyzing these big data. One of the reasons that make them extremely successful is their ability to discern vital knowledge from the ocean of collected data. How do we define big data? Big data is a result of various observations, bank transactions, scientific research, social network posts, etc. And the outputs of these factors grow exponentially. Current systems for analyzing and storing these large sets of data are no longer sufficient.

The analytics face three V problems related to big data; volume, variability and velocity they are three main aspects of big data that are to be considered before working with it. The reason for exponential growth of data is the evolution of data production, which starts from one progression. Historically, data has been generated and accumulated by employees of the companies, who would manually enter the necessary data into their computer system. The second progression occurred with the invention of the Internet, when ordinary users were able to generate their own data as in millions of users signing up and entering their data into many websites such as Facebook, Twitter, and Instagram. Finally, the third progression occurred with the invention of machines that could accumulate data on their own. For example, there are a lot of devices all over the world that monitor humidity and temperature across given time. As a result, there is a colossal amount of data being generated every second. There are three types of data. In the first one, investigation process of different objects occurs at a particular point in time. In the second one, objects are investigated at different points in time. The last one is the data in the form of time series. Since most of the data generation is intrinsically related to time, it comes in the form of time series. Time series is a set of quantitative observations that are collected at a continuous time interval and are measured successively across that interval using equal spacing. Examples of time series include the continuous monitoring of a person’s blood pressure, hourly readings of humidity, daily closing value of a company stock, monthly measured level of unemployment, and yearly income figures.

Generally, time series analysis is used when there more than 50 data points in a series and it is done in order to discern the fundamental structure and function that produce the observation. Understanding the mechanisms of a time series allows a mathematical model to be developed that explains the data in such a way that prediction, monitoring, or control can occur. Examples include prediction/forecasting, which is widely used in economics and business. Monitoring of ambient conditions, or of an input or an output, is common in science and industry.

There are two main methods for analyzing time series. The first one is analyzing in the time domain when time series is analyzed with respect to time and the second one is analyzing in the frequency domain which refers to analyzing time series with respect to frequency rather than time. These methods are usually suitable if the time series are stationary.

Most business and economic time series are far from stationary when expressed in their original units of measurement, and even after deflation or seasonal adjustment they will typically still exhibit trends, cycles, random-walking, and other non-stationary behavior. For non stationary time series we use other methods of analysis. Especially, for similarity measurement of non stationary time series four well known methods exist. They are:

- Euclidean Distance: directly compares two time series of equal length D, and is usually appropriate for applications that do not directly or necessarily present correlation among distinct features;

- Dynamic Time Warping: is a technique of time series similarity measurement of temporal time sequences, even if they vary in time or speed;

- Longest common subsequence of time series A and B is the longest sequences from A and B that are common between two time series.

- Piecewise Linear Representation(PLR): Representation of time series of any length with the straight lines. PLR is the most frequently used representation. The length of straight lines should be much smaller than the length of time series.

All above methods are used in a variety of fields to find out the similarity of time sequences. However, these techniques and the ones not mentioned here have their advantages and disadvantages when compared with each other. Every method has its drawbacks such as scaling, longer time series, and application specification.

We aim to make a new method that will help machine to improve the accuracy and efficiency of the time series data mining, particularly figuring out the similar time series to pattern time series from the database.

The current study focuses on a new method for estimating similarity of non stationary time series based on Piecewise Linear Approximation. Special techniques for optimization of linear representation by removing unnecessary pivot points are provided. Special comparison criteria were elaborated in order to define the degree of similarity of patterns of time series to analyzed time series.

Structure of the Thesis

Chapter I is literature review. Theory of time series analysis, various types of time series representation and time series data similarity measurement techniques are given in this chapter. As with most computer science problems, representation of the data is a key to efficient and effective solutions. Several high level representations of time series have been proposed, including Fourier Transforms, Wavelets, Symbolic Mappings and Piecewise Linear Representation. The main usage of Piecewise Linear Representation in storage, transmission and computation of the data to make it more efficient is discussed.

In general Chapter II is formal and theoretical definition of the methodology which should be used to effective piecewise linear approximation of time series data. In this chapter the problem is defined, terminology of pattern model and actual model are given, the main essence of this method that is to construct a regression model, which in addition to the time factor also includes seasonal dummy variables is introduced. Before enter dummy or indicator variables that are qualitative characteristic into regression model, they must be assigned digital labels. In other words qualitative variables have to be converted into quantitative ones. So further dummy variables allow to construct and estimate piecewise linear models which can be applied to research of structural changes. Theoretically basis of this problem is explained in details. Since statistical significance of structural change requires the development of a decisive rule that would allow eliminating statistically excessive supporting points. Developed optimization method's theoretical part is also presented in chapter II.

In Chapter III experimental application of developed piecewise linear approximation and optimization methods are presented. Based on methodology in the Chapter II it is provided how using the developed method of piecewise linear approximation increases effectiveness of analyzing and reveals hidden patterns of studied time series. Comparison of pattern model and analyzed model, similarity detection of linear structure of pattern model and estimated piecewise linear approximation of analyzed time series with respects to test of pivot points significance is presented. To optimized linear segments output further applied the process of classification into developed cases. Empirical examples for each of the cases are provided. The example of comparing time series of exchange rates of different countries where exchange rate of US Dollar to Georgian Lari is taken as pattern time series is presented. Another interesting application of developed method for seasonality test of time series sequences shown as last experimental example.

The Problem Actuality:

The world’s data is doubling every 1.2 years. There are 7 billion people in the world. 5.1 of them own a cell phone. Each day we send 11 billion text messages, watch over 2.8 billion YouTube videos and perform almost 5 billion Google searches. People are not just consuming it, they are generating it. As information generation and accumulation grows exponentially, recently, much importance is given to the retrieval of important information from large databases. Clearly, a variety of objects can be a search target: text, images, words, etc. However, most of the data comes in the form of time series because our life is immensely connected with time.

Time series contain large and important information that cannot be verbalized. The information can be about various phenomena of our life: from physics to economics, finance, etc. Therefore extracting vital information from large databases of time series is one of the main tasks of data mining. Looking at the time series representing the dynamics of any stock on the stock exchange, we want to find, among the great number of stocks, the stock with dynamics similar in nature to ours. This is just one example, and there can be many more.

Time series analysis is constantly changing and improving with the development of new technology and new industries. Datasets are growing larger and larger these days. Companies are facing the challenge of dealing with more and more data available in their data storages. As a result of this they are literally diving into the areas where they can apply data mining ideas, data mining algorithms and tasks. Similarity analysis of time series is further used in larger analyses such as time series forecasting, clustering or rule discovery. The need for fast and efficient analysis of time series similarity forces the data mining industry to search for new algorithms for time series similarity measurement. Among numerous methods of similarity matching of time series T1 to time series T2 there should be a new method that will optimize the work of measurement in terms of speed and efficiency.

The current research is addressing the optimization and estimation problem of the similarity measurement of time series by proposing a new approach based on Piecewise Linear

Representation specifically for non stationary time series. The research extended by elaborating comparison criteria for defining degree of similarity of time series.

Methodology

The methodology of the research is based on principles and methods of statistical time series analysis, methods of calculus, linear algebra and mathematical statistics, in particular multidimensional dummy variable regression analysis.

For obtaining piecewise linear representation MatLab Software Language is used. It is also used for figuring out statistically unnecessary pivot points in piecewise linear approximation to optimize whole process of similarity measurement. MatLab is used to demonstrate F-test result of dummy pivot points and their compare with tabular value. Based on obtained results more simple and optimal piecewise linear approximation is depicted. Special cases were developed to define similarity case of time series. Mainly 6 different cases for similarity measurement were elaborated such as structurally identical, structurally identical proportional in slopes, structurally identical with small error, structurally similar, partially similar and structurally different. T-test is used to determine structurally identical case with small error. In the implementation of regression analysis and determination of each of the piecewise linear segments and whole piecewise linear approximation, for the simplicity MatLab code is written and implemented. Outputs of that code are represented as summary statistics within the case studies.

Purpose of the Study

The general purposes of this research are: (i) elaboration of new method that is based on Piecewise Linear Representation of time series segments; (ii) building new comparison criteria for estimation of the similarity degree of query time series to analyzed time series that are taken from the database.

The research objectives are as follows:

- Representation of observed non linear scattered points of data in the form of piecewise linear segments;

- Elaboration of the method for eliminating unnecessary pivot points of the representation. Pivot points are the points of the starting and ending of linear segments;
- Elaboration of the new mathematical method for estimation of time series similarity level;
- Simplifying the process of approximation design for non stationary time series by optimizing piecewise linear approximation;
- To implement comparison method of the optimized piecewise linear approximated result of time series;
- Defining similarity criteria for grouping estimated time series;
- Using MatLab programming language, to write appropriate code for documentation of implementations;

The Novelty and Contributions of Investigation:

1. The time series data were represented in the form of linear segments by using new method that is based on Piecewise Linear Representation.
2. Several cases were elaborated on the basis of the newly introduced concepts of Piecewise Linear Representation. The cases are the ‘structurally identical’, ‘structurally identical with proportional slopes’, ‘structurally identical with small error’, ‘structurally similar’, and ‘partially similar and structurally different’ ones.
3. Based on the usage of the dummy multidimensional linear regression technique, the new statistical method of estimation of similarity of non stationary time series was developed. The capability of the model was demonstrated throughout several numerical examples;
4. The results of the new method of representation of time series with linear segments results were compared to other representation methods. It was shown that the integrated aggregate model of piecewise linear representation easily and rapidly capture the notation of the time series and translate it for machine interpretation.

5. Appropriate software tools (in MATLAB programming language) were created;
6. Based on the elaborated method of efficient representation of time series in a form of piecewise linear approximation the new criteria for estimation of similarity degree of time series were developed.
7. A revised methodology was built based on Fishers method to test the adequacy of the developed model.

Theoretical Value and Practical Importance of the Study:

The current research is a new step in time series data mining, especially in the time series similarity search; it will speed up the process of determining the similarity of time series, which will facilitate the analysis of large databases. The new method of piecewise linear approximation of nonlinear single variable function with relevant type of profile was presented. The benefits of proposed approximation method compared to the conventional subject of building the linear splines is that it uses standard n-dimensional linear regression analysis procedure and it does not require usage of restriction in approximation nodes. The current developed method is applicable to determining similarity between two time series; thereby it can be used in similarity detection of query time series to any time series from database. For example, one can consider two time series X and Y. The question can be: do stocks X and Y have similar movements? The method of similarity measurement introduced in this paper allows for imprecise matches of time series. Further, it is possible to use it in indexing, subsequence similarity, clustering, and rule discovery problems. As the speed of the process is one of the main aspects of data mining in our method, it was attempted to keep the balance between accuracy and efficiency of design solution. The proposed similarity criteria will also provide easier and more efficient definition of the similarity level of the time series. The new method of Piecewise Linear Approximation and set of criteria for estimation similarity of non stationary time series together will greatly simplify the whole process of analyzing time series data sets.

The research can be of particular importance as an alternative instrument in the analysis of time series similarity, indexing and clustering of time series in the fields like business analytics, scientific researches, practitioners, and so on.

Major findings and conclusions of the study have been presented on the various conferences. Moreover, the outputs of the study were neatly highlighted in three articles published in national and international, refereed and peer reviewed academic journals.

Structure and volume of the work

The thesis study is 134 pages and consists of 3 chapters, a list of references and list of figures and list of tables.

CHAPTER 1: LITERATURE REVIEW

Currently, to investigate the properties of complex systems, including experimental studies, an approach of data analyzing of information produced by the variety of systems is widely used. Sometimes, even having a certain characteristic of observed values, it is impossible to make a mathematical description of the process. System analysis especially in experimental research is often realized by processing of the recorded signals. Normally, such signals are called observable, and the research method – the representation of dynamical systems. This part of dynamical system theory is called analysis of time series. Development of time series analysis as a science in recent decades has led to the creation of a variety of methods, procedures, forecasting techniques, and those are unequal in significance. According to researchers, time series analysis has already over one hundred methods. For specialists face the problem of finding the selection methods that would provide adequate analysis of time series, or series of events. To obtain high quality and accurate analysis of time series, researchers have to know applied mathematics, econometrics, and statistics. It is obvious that it is not possible to analyze the complex non-linear multivariate models in their normal form.

Representation of those data is very important for efficient and effective analysis of time series. There are a lot of representations that have been proposed: Fourier Transform, Wavelets, Symbolic Mapping, and Piecewise Linear Representation. Here are just some examples of research that are focused on representation of time series. Representation of time series improves the process of similarity search of time sequences. Similarity measuring is efficient in other computation processes as indexing, subsequence similarity, clustering, rule discovery etc.

Currently, to conduct a research on properties of complex systems, including experimental studies, time series analysis is widely used. Data analysis, especially in experimental research, is usually realized by processing of the recorded observations. For instance, in cardiology electrocardiogram signal is used for this purpose, in seismology - record of fluctuations in the Earth’s crust, in meteorology - data of meteorological observations. Such signals are called observable. Method of investigation of these signals is called reconstruction of dynamical system. This part of the theory of dynamical systems is in turn called time-series analysis. In this context, ‘observed’ means the sequence of values of a variable (or variables) that are recorded continuously or with some intervals over time. Hence, instead of the term ‘observed’ the concept of ‘time series’ is used. A time series is a sequence of numerical data points in successive order, usually occurring in uniform intervals, in other words a time series is simply a sequence of numbers collected at regular intervals over a period of time.

Time series analysis unlike other analyses of random samples is based on the assumption that successive values in the data file represent consecutive measurements taken at equally spaced time intervals. Time series analysis can be useful to see how a given asset, security or economic variable changes over time or how it changes compared to other variables over the same time period. For example, suppose you wanted to analyze a time series of daily closing stock prices for a given stock over a period of one year. You would obtain a list of all the closing prices for the stock over each day for the past year and list them in chronological order. This would be a one-year, daily closing price time series for the stock. Delving a bit deeper, you might be interested to know if a given stock's time series shows any seasonality, meaning it goes through peaks and valleys at regular times each year. Or you might want to know how a stock's share price changes as an economic variable, such as the unemployment rate, changes. Time series analysis has two main objectives; first - identifying the nature of the phenomenon represented by the sequence of observations and second - forecasting (predicting future values of the time series variable). Both of these objectives require that the pattern of observed time series data is identified and more or less formally described. A random sample of 4,000 graphics from 15 of the world’s newspapers published from 1974 to 1989 found that more than 75% of all graphics were time series.

CHAPTER 2: THEORETICAL FOUNDATION

Integral Aggregate model has to be realized by the method of dummies where it relates to the modeling of seasonal components of time series. The main essence of this method is to construct a regression model, which in addition to the time factor also includes seasonal dummy variables. Indeed, the dummies are used when there is a single set of independent significant variable in observation. But in fact, this set contains the subgroups which vary on different qualitative indices. The significant variables indicate a level of the quantitative index taking values from a continuous interval, while the dummy or indicator variable – is a qualitative characteristic. It can be any attributive signs – a profession, gender, education, climate conditions, belonging to a certain region, etc. To enter such variables into a regression model, they must be assigned digital labels, i.e. qualitative variables have to be converted to quantitative ones. It is acceptable to call them dummy variables, structural, or artificial. For instance, 0 – male; 1 – female. Dummy variables allow to construct and estimate piecewise linear models which can be applied to research of structural changes. First dummy variable is additional variable X_0 of the term β_0 in regression model where it is always 1. It is not compulsory to include variable X_0. As it is already comes as the initial point of the first piecewise linear segment. Other dummies should be entered.

For example, the dependency of output y is investigated on the size of enterprise asset x_t.Thus there are bases to consider that restructuring occurred at time point t_0, and the nature of dependency changed. To evaluate this model, it is necessary to enter a binary variable $v_t = \begin{cases} 0 \text{ if } t \le t_0, \\ 1 \text{ if } t \ge t_0 \end{cases}$ and write down the model as: $y = a_0 + a_1 x_t + a_2 (x_t - x_{t-t_0}) v_t$. At $t \le t_0$ the line of regression has a slope a_1 , at $t > t_0$ the slope is equal to $(a_1 + a_2)$ and at point x_t a gap doesn't occur. When a_2=0, it is concluded that the structural change doesn't occur at the time point t_0. Furthermore we will discuss the last circumstance that is absence of gap, but related to the subject in the case of multivariate regression, not containing dummy variables.

Here are the main tasks in which dummy variables method is most effective.

1. Variables - indicators belonging to a particular observation period - for modeling uneven structural displacements. The period boundaries set out of a priori considerations, for instance, 1, if an observation belongs to the period 1941-1945, and 0 - otherwise. This is an example of using dummy variables for modeling temporary structural displacement. Permanent structural displacement is modeled by variable, which is equal to 0 until certain time, and 1 for all observations after this certain time.

2. Seasonal variables - for modeling seasonality. Seasonal variables take different values depending on which month, quarter or day corresponds to observation. For example, the consumption model, which takes into account seasonal variations?

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3,$$

$$x_1 = \begin{cases} 1 \quad \text{for} \quad \text{winter} \quad \text{months}, \\ \qquad\quad \text{else} \quad 0, \end{cases}$$

$$x_2 = \begin{cases} 1 \quad \text{for} \quad \text{spring} \quad \text{months}, \\ \qquad \text{else} \quad 0, \end{cases} \tag{1}$$

$$x_3 = \begin{cases} 1 \quad \text{for} \quad \text{summer} \quad \text{months}, \\ \qquad \text{else} \quad 0. \end{cases}$$

The consumption volume is as follows:

$y = b_0$ - For autumns months;

$y = b_0 + b_1$ - For winter months;

$y = b_0 + b_2$ - For spring months;

$y = b_0 + b_3$ - For summer months

In this case, if the result of the regression analysis turns out to be $b_3 = 0$, it means that consumption difference between the summer and autumn seasons' is insignificant. At $b_1 = b_2$, there is no difference between consumption in winter and spring, etc.

A linear time trend – for modeling of smooth (gradual) structural displacements. The dummy variable shows a period of time passed from a "zero" time point to certain time concerning the observation on the timeline. If the time intervals are identical between consecutive observations, then the trend can be made of the observations' numbers. A time trend differs from binary dummy variables in using the degrees such as t^2, t^3 etc. They help to model a smooth, but non-linear trend. It is meaningless to raise the power of binary variable because as a result the same variable will turn out. Furthermore, the dummy variables' combination of different kinds is possible. It allows you to model the slope inclination of the trend to a certain point. Besides the trend, the following variable is entered into regression: at the beginning of selection until certain time, it is equal to 0, and then it represents a time trend (1, 2, 3, ... in case of identical intervals between observations).

The above method has a number of advantages:

1. The intervals between the observations do not necessarily have to be identical, and in the sample can be missed observations;

2. The coefficients of the fictitious variables can be easily interpreted, they clearly represent the structure of a dynamic process;

3. For model estimation it isn't necessary to go beyond the classical OLS.

Comparing the essence of fictitious variables' method with the task of Integrated Demand, it is possible to see, despite the resemblance that in both cases there is a conversion of one variable to extra multitude of r variables. Considered tasks can't be formalized within the framework of this method. In fact the $(1-\frac{z_i}{x_j})$ expression is a nonlinear function r of variablesx_j (of supporting values):

$$\left(1-\frac{z_i}{x_j}\right)=\left\{\begin{matrix}\left(1-\frac{z_i}{x_j}\right) if\ z_i \le x_j \\ 0\ if\ z_i \rhd x_j\end{matrix}\right\} \qquad (2)$$

Where the data matrix is defined. Unlike the fictitious variables here we don't deal with single series' observations of the independent (significant) variable. Indeed, we deal with r variables independent from each other, which are determined by the function (2). That is received as a result of conversion (2) from the initial series' observations of the independent z_i variable. Thus, we have proved the possibility of solving the inverse task in a model of integrated demand by using multivariate linear regression.

The model presented if the form of linear equation

$$D_{A\Sigma}=d_1\left(1-\frac{z}{x_1}\right)+d_2\left(1-\frac{z}{x_2}\right)+d_3\left(1-\frac{z}{x_3}\right)+...+d_r\left(1-\frac{z}{x_r}\right) \qquad (3)$$

Considering $D_{A\Sigma}$ as a function of the variable z, it is easy to see that:

1. At the points where $z=x_i\left(i=1,2,...,r\right)$ the $D_{A\Sigma}$ function is continuous;
2. The derivative of this function is piecewise constant, underwent a first-order discontinuities (jumps) at the points where $z=x_i\left(i=1,2,...,r\right)$. Indeed, consider $D_{A\Sigma}$ at the points where z=x$_i$, then, as a consequence of (26), it is obvious that

$$\lim_{z\to -x_i} D_{A\Sigma}=\lim_{z\to +x_i} D_{A\Sigma}=\Sigma_{j=i+1}^{r} d_j\left(1-\frac{x_i}{x_j}\right) \qquad (4)$$

The limit of the $D_{A\Sigma}$(z) function is equal to its limit on the right, which shows the continuity of the points where $z=x_i\left(i=1,2,...,r\right)$

This limit at the point where $z=x_i$, particularly equals to

$$\lim_{z\to -x_i} D_{A\Sigma}=\lim_{z\to +x_i} D_{A\Sigma}=\Sigma_{j=i+1}^{r} d_j\left(1-\frac{x_i}{x_j}\right) \qquad (5)$$

Then equality of the left and right limits, and the last equality is consequences of an obvious limit of $\lim_{x \to -x_i} \left(1 - \frac{z}{x_i}\right) = 0$

Similarly, it is easy to see that the derivative at intervals $\left(z_{k,} z_{k+1}\right)\left(k = 0,1,2,\ldots,r-1\right)$ is equal to $\frac{dD_{A\Sigma}}{d_z} = -\Sigma_{i=1}^{k} \frac{d_i}{x_i}$, which is easily verified by direct differentiation (6)

From this it follows that the function is received as a result of regression estimation of Determinants from r-dimensional linear equation (6). Considering the a function of the one z variable is a continuous, piecewise linear function, with a number of linear pieces equal to the number of pivot points.

Let's consider the system of m polynomials of an independent variable$\boldsymbol{t} \geq \boldsymbol{0}$.

$\boldsymbol{P_i(\alpha_1^i, \ldots, \alpha_{r_i}^i, \tau_i, t)}$, (i=1,…,m) (7)

Where $\boldsymbol{\alpha_j^i}(\mathbf{j} = \mathbf{1}, \ldots, \boldsymbol{r_i}) -$ unknown coefficients of $\boldsymbol{it}$□ polynomial; $\boldsymbol{r_i}$ – order of $\boldsymbol{it}$□ polynomial.

Let's assume that each of these polynomials equals to zero at $\boldsymbol{t > \tau_i}$, where $\boldsymbol{\tau_i}$- real numbers such: $\boldsymbol{\tau_1 \leq \tau_2 \leq \cdots \leq \tau_{m-1} \leq \tau_m = T}$.

We call the intervals Ii=[0, $\boldsymbol{\tau_i}$], as supports of $\boldsymbol{it}$□ **polynomial** (6), i.e., $\boldsymbol{P_i\left(\alpha_1^i, \ldots, \alpha_{r_i}^i, \tau_i, t\right) \neq 0}$ at t∈ $[\boldsymbol{0}, \boldsymbol{\tau_i}]$, and $\boldsymbol{P_i\left(\alpha_1^i, \ldots, \alpha_{r_i}^i, t_i, t\right) = 0}$ if not. Let us emphasize that $\boldsymbol{I_i \subseteq I_{i+1}}$. The intervals $\boldsymbol{I_i}$ are called net system of polynomials (1), and points of $\boldsymbol{\tau_i}$ – it's nodes. We simplify the notation of function (7) and write$\boldsymbol{P_i(t)}$, implying that polynomial $\boldsymbol{P_i(t)}$, in addition to the independent variable t, it also depends on $\boldsymbol{r_i}$ parameters and terms $\boldsymbol{\tau_i}$ which determines the upper limit of its supports.

Let's introduce a new function

$\boldsymbol{F(t)} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \boldsymbol{P_i(t)}$

(8)

It is not difficult to see that function $\boldsymbol{F(t)}$ is finite on its support [0,T], and continuous on it, however its first derivative discontinues at net's node $\boldsymbol{\tau_i}$ (i=1,…,m). Indeed, at t<$\boldsymbol{\tau_k}$ $\mathbf{lim}_{t \to \tau_k -} \boldsymbol{F(t)} = \sum_{\boldsymbol{i=k}}^{\boldsymbol{m}} \boldsymbol{P_i(\tau_k)}$ Whereas at t>$\boldsymbol{\tau_k}$ $\mathbf{lim}_{t \to t_k +} \boldsymbol{F(t)} = \sum_{\boldsymbol{i=k+1}}^{\boldsymbol{m}} \boldsymbol{P_i(\tau_k)}$.

These limits are equal at point t=$\boldsymbol{t_k}$, as, by definition, of $\boldsymbol{P_k(\tau_k) = 0}$. At the same time derivative of $\boldsymbol{P'_k(\tau_k) \neq 0}$, therefore it is clear that $\mathbf{lim}_{t\to t_k-}\, \boldsymbol{F'(t)} \neq \mathbf{lim}_{t\to t_k+}\, \boldsymbol{F'(t)}$ at points $\boldsymbol{\tau_i}$ (i=1,…,n).

The function (8) is called Approximation aggregate, and polynomials (7) - Aggregate's components. Let's assume that the net's nodes are defined, such, that the points $\boldsymbol{\tau_i}$ (i=1,…,n) are known, then it is clear that the Aggregate is defined by parameters of $\boldsymbol{r = \sum_{i=1}^{m} r_i}$, which can be estimated via the least squares method by adding constraints, ensuring the continuation in the nodes. The following should be noted. Piecewise structure of the Aggregate is determined by the fact that certain components of polynomials are finite, and their supports Ii=[0, $\boldsymbol{\tau_i}$] form a nested sequence of non-decreasing intervals $\boldsymbol{I_i \subseteq I_{i+1}}$. So, formally recorded function of Aggregate in the form of (8), can be represented as follows

$$\boldsymbol{F(t) = \sum_{i=k}^{m} P_i(t)}, \text{ for } \boldsymbol{t \geq \tau_{k-1}} \qquad (9)$$

In fact, we have constructed m functional of (according to the number of supports Ii=[0, $\boldsymbol{\tau_i}$] $(\mathbf{i} = \mathbf{1}, \dots, \mathbf{m})$)

$$\mathbf{min}_{\alpha_1^k,\dots,\alpha_{r_i}^k} \boldsymbol{S^2} = \sum_{i=\tau_{k-1}}^{\tau_m} (\boldsymbol{x(t_i)} - \sum_{i=k}^{m} \boldsymbol{P(t_i)})^2 \text{ (k=1,2,…,m)} \qquad (10)$$

for piecewise linear approximation of initial time series x($\boldsymbol{t_i}$) (i=1,2,…,n). The last required identification of polynomial components (7) of approximation aggregate (8), that requires an identification of the system's coefficients $\boldsymbol{\alpha_1^i, \dots, \alpha_{r_i}^i}$ (i=1,…,m) of polynomial aggregate (7). The last is quite time consuming, but not too much mathematically complicated the problem of constructing a system of splines, determined by means of polynomial components.

Let's mention the artificial nature of the problem statements (10) as constructing finite functions of approximating aggregate that led to a system of enclosed finite supports $\boldsymbol{I_i \subseteq I_{i+1}}$ $(\mathbf{i} = \mathbf{1}, \dots, \mathbf{m})$ and finite polynomial components determined on them. It is obvious, in fact, that, instead of one could use the construction of the traditional system of splines (Ahlberg, Nilson, Walsh, 1967) on intervals $(\boldsymbol{\tau_i, \tau_{i+1}})$ i=0,1,…,m. Such a problem would not be deserved as separate study, if not one of its special cases - piecewise linear approximation, that is the case when all component polynomials represent the straight lines.

CHAPTER 3: EXPERIMENTAL REALIZATION OF PIECEWISE LINEAR APPROXIMATION

For segmentation and approximation process special Matlab program was developed. The developed program in Matlab programming language is provided in Appendix 1 in the thesis. Input parameters of the program are: the array of sampling moments *x*, the array of time series values *y*. Also as input we have array *d* which is the pivot points set of time series. Number of pivot points can be varying. Values of pivot points *d* are also changeable. It can be taken as fixed interval values or can be defined by user. The output of the Matlab program – the array *f* that is the calculated values of piecewise linear approximation. The result is compared with array observed values. In the graph array *y* is displayed as points and approximated piecewise segments as the lines *f*.

For estimation of time series similarity criteria as similarity criteria following cases were elaborated. There are 6 cases of similarity. They are:

1. Structurally Identical case – In this case number of pivot points in Piecewise Linear Approximation is identical. Slopes of corresponding line segments are also should be identical.
2. Structurally Identical (Proportional in slopes) case – In this case number of pivot points in Piecewise Linear Approximation is identical. Slopes of corresponding line segments are also should be identical or proportional to each other.
3. Structurally Identical (with small error) case – Structurally Identical (Proportional in slopes) case - In this case number of pivot points in Piecewise Linear Approximation is identical. Slopes of corresponding line segments are also should be identical or proportional to each other. For this case we open another case what if slopes are proportional but have some small error. That small error doesn't makes huge difference in linear approximation; furthermore it doesn't make huge difference in time series data observations. All this come to Student's t criteria. It depends what we are analyzing and how we analyzing it. Later we will discuss about this case in more detail.
4. Structurally Similar case – In this case number of pivot points in Piecewise Linear Approximation is the same. Slopes of corresponding lines are not equal but have the same coefficient. In other word direction of the line is the same.
5. Partially Similar case – The number of pivot points it is not necessary to be equal. In this case not all corresponding line's slopes have the same coefficients. In partially similar

case similarity degree defined by how many linear segments out of total have the same coefficient and the percentage of closeness of the corresponding linear segments slopes.

6. Structurally different case – The number of pivot points are different and the coefficients of the line slopes are also totally different.

Above presented 6 cases of similarity can be extended. But we built our approach according to these similarity cases. Below we provided empirical examples for each similarity case. It is very difficult to find times series economics or scientific observations those will be suitable for the first and second cases. For this reason we provided examples with generated data.

In these research comparison examples of US Dollar exchange rate to various currencies including Georgian Lari, Euro, Russian Ruble etc. were provided. By empirical study using our piecewise linear representation method it is found that in last six year Georgian Lari has demonstrated close tendency to Russian Ruble rather than European currency.

Additionally seasonality of time series also can be detected by using proposed method of piecewise linear representation. For seasonality measurement the same techniques like in similarity search of time series are used. The difference is that the whole time series cut into smaller parts, and smaller parts analyzed as different sequences. Optimization algorithm can make the whole process of seasonality measurement more effective. There for it would be more appropriate to apply this algorithm in this case. The divided sequences of time series can be classified into the cases as in previous research part can be. It will help to define the degree of seasonality.

CONCLUSIONS

Following purposes were set. Develop optimization algorithm for piecewise linear representation of time series by eliminating statistically unnecessary pivot points for effectively simplify the complexity of non-stationary time series and speed up the whole process of similarity assessment of time series sets. Further similarity criteria of time series should be elaborated based on this optimization process. Additionally bellow listed objective have been reached;

- A new method for Piecewise Linear Approximation of Non Stationary Time Series is developed.
- An algorithm for optimization of generated Piecewise Linear Representation of Time Series is developed.

- The criteria for evaluation of approximation parameters are developed.
- The criteria for comparing of investigated series with pattern time series are developed.
- The calculations on concrete examples of Time Series based on developed methods are performed.
- The appropriate software tools are created.
- The analysis of time series on the presence of seasonality factor based on developed method and criteria is performed.

Publications

1. A. Milnikov and D. Satybaldiev. Designing Optimal Integral Aggregate Structure. Journal: Journal of Technical Science and Technologies; ISSN 2298-0032; Volume 3, Issue 2, 2014, pp 21-24.
2. A. Milnikov, D. Satybaldiev and C.Mert. A new method of piecewise linear approximation of non-stationary time series. Journal: Journal of Technical Science and Technologies; ISSN 2298-0032; Volume 4, Issue 1, 2015, pp 5-8.
3. D.Satybaldiev A New Method of Piecewise Linear Approximation of Non-stationary Time Series for Similarity Measurement. ICECCO 2015 September, Almaty Kazakhstan, pp 90-93.
4. D.Satybaldiev Non stationary time series similarity measurement based on Piecewise Linear Approximation: Empirical example, Journal of Science, Innovations and Technologies of Kyrgyzstan N) №2. 2016 ISSN 1694-7649, pp 12-16.

.