International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)
Volume No.3 Issue No.9, pp : 1217-1224 1 Sep 2014

Forecasting Non-Stationary Time Series Method of Allocation Patterns

Perminov G.I., Batueva M.S.
Department of Business Intelligence, National Research University - Higher School of Economics (Russia)
giperminov@gmail.com, batueva.marija@gmail.com

Abstract: Importance. This paper proposes a method of forecasting of essentially non-stationary time series, i.e. of series that changing structure, variables or model, the coefficients for the variables. Because of nonstationarity of processes that generate economic indicators, most economic time series falls in this category. Objective. Currently, structural analysis of essentially non-stationary time series can be represented in two directions: 1) a breakdown of such series on the segments in which the properties of the component little changed, after which the analysis of patterns, one approach which is consistent allocation component of the time series for each segment and 2) identify patterns. Found in the history of areas with similar dynamic characteristics allow to receive quite accurate forecasts of future values of the series. In the work in the second direction was studied algorithms of the method of k-nearest neighbor with constant and variable-length pattern and different metrics vicinity.Methods. To study the variations of the algorithm was built prediction model data with varying period and forecasts of real cyclic series production of electricity from coal and natural gas.Results. The most accurate forecast model data was built using the algorithm with the conversion of the abscissa, the algorithm with fixed-length pattern gave high errors. Conclusions and Relevance. When dealing with economic time series change as the model coefficients, so and its structure. Reduces the number of variables and their influence disappears from the model, while others appear. For such series forecasting patterns should be used with transformations on the abscissa.

Key words: *variable pattern length, forecasting non-stationary time-series.*

Introduction

As has been noted in many papers on economic theory and mathematical modeling, most of the economic indicators are characterized by significant non-stationary processes of generating and variability of its structure. When working with multi-dimensional and one-dimensional time series of economic change as the model coefficients, so, and its structure. Reduces the number of variables and their influence disappears from the model, while others appear (Perminov G. 2013), (Ovchinnikov AA, Hramov AE, Lyuttehann A. & Koronovskii AA 2011). Currently, the structural analysis of essentially non-stationary time series can be represented as two directions: 1) a breakdown of such series into segments on which properties of the components do not change much, and then analyzes the structure, one of the approaches that are consistent allocation component time series for each segment (Perminov GI & Leonova NV. 2012) and 2) identifying patterns. Found in the history of sites with similar dynamic characteristics allow to obtain fairly accurate predictions of future values of the series (Alesgerov FT, Belousov VU, LG Egorova & Mirkin BG. 2013).

Research Questions

The purpose of this paper is to examine the prediction algorithm of nonstationary time series by allocation patterns with constant and variable length.
In accordance with this objective were as follows:
• the method of k nearest neighbors to find the most suitable metric proximity,
• the need to make the transformation patterns along the horizontal axis, or enough to search for patterns of fixed length.

Research Methods

The work was carried out practical calculation of real time series algorithms using k nearest neighbors method with constant and variable-length pattern and different metrics vicinity.

Research results

The most accurate forecast model data was built using the algorithm with the conversion of the abscissa and the Minkowski metric is not noisy series and weighted Euclidean for very noisy series. It has been found that selection of the most suitable metric is also strongly dependent on the type of series. For less noisy series, reflecting the production of electricity from coal, better suited Minkowski metric, and for more noisy production of natural gas - the weighted Euclidean, as often it uses minimal number of nearest neighbors and thus less susceptible to noise.

State of the question

One of the most urgent tasks is the prediction of financial time series, which is impossible without investment activity. However, the practical application of these methods is difficult because of the nonstationarity of the time series.' Patterns of development of the financial markets are constantly changing, and these changes occur very quickly. Accordingly, the quality of the forecast will depend on the correct choice of the training set. In this regard, in an attempt to find specific patterns of movement of financial time series has been described a variety of market patterns (Perminov G. 2013). Market pattern is a set of features, allowing it to determine the market. These symptoms can be absolutely or relatively formalized (Exchange information-analytical portal ClusterDelta.com: http://clusterdelta.com/patterns/).

1. Methods for determining patterns

In (MQL5.community: http://www.mql5.com/ru/code/133) discussed in detail the various algorithms for computing patterns.
Another aspect of the calculation is the choice patterns of source data. For example, in (Exchange information-analytical portal ClusterDelta.com: http://clusterdelta.com/patterns/) analyzes different approaches.
Methods for determining patterns of the most frequently used method is the k-nearest neighbors. Method of k-nearest neighbor

(k-NN, k-Nearest Neighbor algorithm) k looking past patterns (neighbors), most similar to the current pattern and calculates a price based on the average of any suspended neighbors.

On the basis of this method was developed indicator written in MetaQuotes Language 5 (MQL5) - built-in language for programming trading strategies developed by MetaQuotes Software Corp. The indicator is designed for MetaTrader 5 trading platform, allowing to carry out brokerage services in Forex, CFD, Futures and exchange markets (MQL5.community: http://www.mql5.com/ru/code/133).

Fig.1. Predicting patterns in the packet in the MetaTrader 5

This indicator is based on the algorithm of 1-NN, ie to find one nearest neighbor. To estimate the distance between the current and past patterns indicator uses the correlation coefficient. The indicator draws two curves calculated using the nearest neighbor found: solid curve reflects the dynamics of the process in the past, and the dot-dash - in the future (Fig. 1). Baseline data are shown with tears. In constructing these curves nearest neighbor is scaled obtained from the linear regression between patterns. The meter also displays information about the date of the nearest neighbor and the correlation coefficient with this pattern. For example: "Date 2003.08.27 7:00:00 nearest neighbor, the correlation coefficient is equal to the current pattern of 0.9434264228359904."

The main drawback of this algorithm is that it measures only one nearest neighbor, whereas the more accurate prediction can be obtained by incorporating consideration of k neighbors.

2. Dynamic analysis of patterns

Another way of finding the nearest neighbor is to use DTW (dynamic time warping) algorithm. This algorithm builds a path of least cost, minimizing the distance between the elements of the series and its subsequence (Romanenko AA. 2011).

The idea of the algorithm is to use the first step of the classical DTW algorithm to weed out a subsequence, which can not satisfy the constraint expressed by the condition that the step of the way W participated only adjacent elements of the matrix (including diagonally adjacent). And on the next steps in the usual enumeration of the remaining subsequences selected by the classical optimal DTW algorithm.

3. Prediction of univariate time series by the k-nearest neighbors

Problem of forecasting univariate time series by the k-nearest neighbors adequately studied. Unresolved issue is the length of the patterns. In most studies the pattern length is constant and equal history (Varfolomeeva AA 2011).

But lately there have been other proposals - row length is not fixed, as the patterns are compared to the background after transformation on the x-axis (Tsiganova SV. 2011). We can assume that for series with varying frequency is more efficient algorithm, implying transformation along the horizontal axis, and for a time series with a constant period such conversion does not necessarily lead.

Statement of the Problem

In this paper, first, we study the problem of finding the most suitable metric proximity, and secondly - the need to make the transformation patterns along the horizontal axis, or enough to search for patterns of fixed length. To determine the closest segments in this paper we study the linear transformation (compression, tension and shear) invariants of transformation and function of "closeness" of segments of time series. The latter will be one of the criteria for an adequate prediction of the constructed algorithm.

Brief theoretical information

1. Statement of the Problem

`In this paper we consider the one-dimensional time series - series in which each point in time compared real number $\{x_{t1}, x_{t2}, \ldots, x_{tn}\}$

Required to predict the following sequence of values L $\{x_{tn+1}, x_{tn+2}, \ldots, x_{tn+L}\}$, which will be determined by the value of history $\{x_{tn-L+1}, x_{tn-L+2}, \ldots, x_{tn}\}$ length L.

2. Algorithm for finding patterns with transformations in *X* and *Y*

Algorithm has the form (Tsiganova SV 2011):

1. Throughout the time series are allocated vector of length r: $r_{min}, \ldots, r_{max}$, after which linear transformations (compression, tension and shear) are similar to the background $\{x_{tn-L+1}, x_{tn-L+2}, \ldots, x_{tn}\}$.
2. Are invariants and study transformations between two close vectors of time series and using the criteria defined by proximity most "similar" vector.
3. Proximity criterion is the proximity function of two vectors a, b - in this paper are:

standard Euclidean metric:

$$D_E(a.b) = \sqrt{(a-b)^T(a-b)},$$

weighted Euclidean metric:

$$D_{WE}(a.b) = \sqrt{(a-b)^T \Lambda^2 (a-b)},$$

Minkowski metric Lp:

$$D_{Lp}(a.b) = (\sum_i |a_i - b_i|^p)^{1/p}.$$

4. Selected proximity function is minimized for each potential neighbor.
5. Для отыскания k наиболее близких векторов используется метод k ближайших соседей.

International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)
Volume No.3 Issue No.9, pp : 1217-1224 1 Sep 2014

6. Прогноз вычисляется как взвешенное среднее арифметическое k векторов

$$(x_{ti+1},\dots,x_{ti+L}L = \frac{\sum_{j=1}^{k}\omega_j A_j(x_{tj+1},\dots,x_{tj+L}}{\sum_{j=1}^{k}\omega_j},$$

$$\omega_0 = \left(1 - \frac{d_{ij}^{t}}{d_{ik+1}^{2}}\right),$$

d_{ik+1}^{2} – расстояние до k+1 ближайшего соседа.

7. Для нахождение преобразования φ по оси OX и выбор потенциальных соседей – векторов $a_1, a_2, \dots, a_m$ во всем временном ряде X находятся точки экстремумов.

Further, such vectors are to the point of extreme values of these vectors to the points of extreme values prehistory had a minimum distance of the chosen metric. The maximum deviation allowed by the algorithm - specified parameter ε. From this condition, for each i-th neighbor is a potential factor a_{0i} stretching on OX and selected vectors $a_1, a_2, \dots, a_m$ become potential neighbors. Subsequent values $b_1, b_2, \dots, b_m$ - potential prognosis depending on the proximity neighbor. The number of potential neighbors is directly proportional to the parameter ε - less than the value, the less potential neighbors algorithm allocates.

Thus, there are neighboring vectors $a_1, a_2, \dots, a_m$ for prehistory $\{x_{tn-L+1}, x_{tn-L+2}, \dots, x_{tn}\}$ in time series not only constant, but with varying period.

8. For each potential neighbor minimized proximity function and the coefficients are b0i stretching along the axis OY for each vector close. Let Y - the vector of the time series, and X - vector time series to be converted to the axis OY, to get closest to the Y vector. This requires a minimum of the following function:

$$F = \sum_{t=1}^{L}(Y_t - (a + bX_t))^2$$

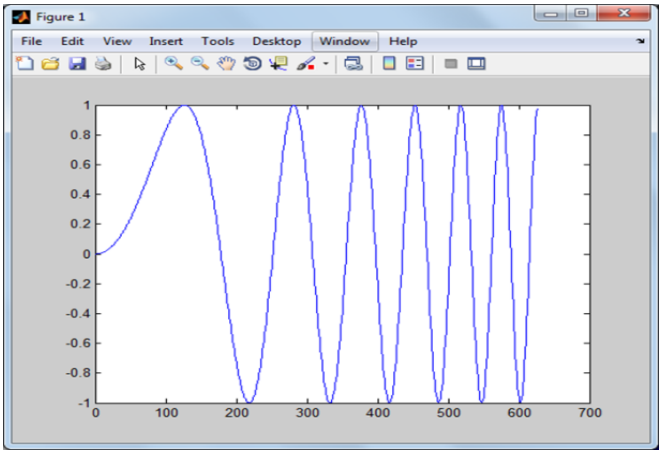

Fig. 2. Graph of the function f1

9. Neighbors are sorted by the value of the proximity function. Next selected k nearest neighbors (k - a specific number). Their potential averaged forecast (the lower the value of the proximity function for a neighbor, the greater contribution is the k-th potential prognosis averaging).

3. Quality Criteria

And finally, you need to choose a quality criterion algorithm. You can evaluate it using various error functions such as (Tsiganova SV 2011):

$$E = \frac{1}{l}\sum_{j=1}^{l}|f_{n+j} - \hat{f}_{n+j}|,$$

or indicators such as SMAPE (Symmetric Mean Absolute Percent):

$$SMAPE = \frac{1}{l}\left(\sum_{j=1}^{l}\frac{|f_{n+j} - \hat{f}_{n+j}|}{|f_{n+j} - \hat{f}_{n+j}|/2}\right) * 100\%$$

Parameters of the model and the final form of the algorithm depends on the test time series. Thus, for a series of varying periodicity is more efficient algorithm involving converting the abscissa, and time series for a constant period to conduct such a transformation is not required.

Selecting metrics is largely dependent on the number of noise, as it allows to limit the number of nearest neighbors, which will take place prediction.

Practical implementation of the proposed method

To confirm the aforementioned assumptions for the forecast of: 1) simulated data with varying period and 2) the actual data generating electricity from coal and natural gas in the U.S. in the period from 1973 to 2013.

1. Prediction lineup with patterns of fixed length

As adopted lineup function $f_1 = \sin(x)^2$ (Fig. 2).

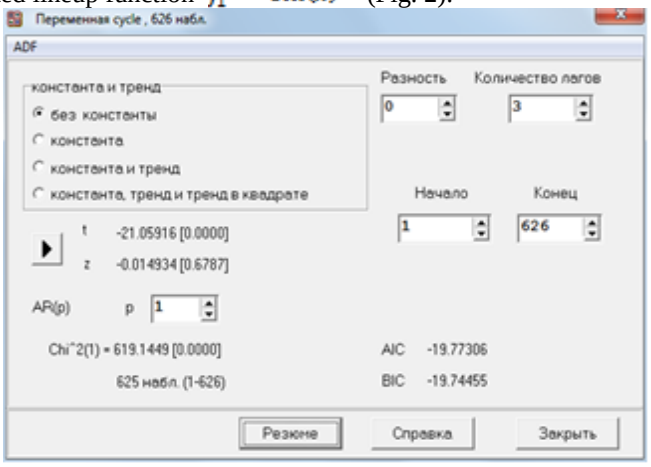

Fig.3. Dickey-Fuller test for the function f1

Despite the fact that this process is stationary (prove this below), it has a variable frequency, which will allow us to demonstrate the algorithm. Calculation of the Dickey-Fuller test, revealed the stationarity of the process. The null hypothesis of this test is non-stationary processes.

In our case, was selected test version without constants (as the process moves with equal amplitude near zero). Since the

calculated statistics lies to the left of the critical value, the null hypothesis is rejected and the process is stationary (Fig. 3).

1.1. To begin to build a forecast, taking as a function of proximity to the standard Euclidean metric (see Fig. 4):

Algorithm identified five nearest neighbors. Of Figure 4 shows that the algorithm is not able to take into account the changing frequency range and built a forecast based on previous patterns periods. Here and further evidence by a solid line, the forecast - dotted and error - the dot-dash.

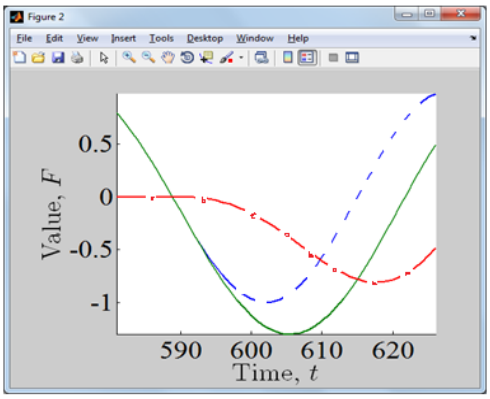

Fig. 4. Forecast f1 using Euclidean metric

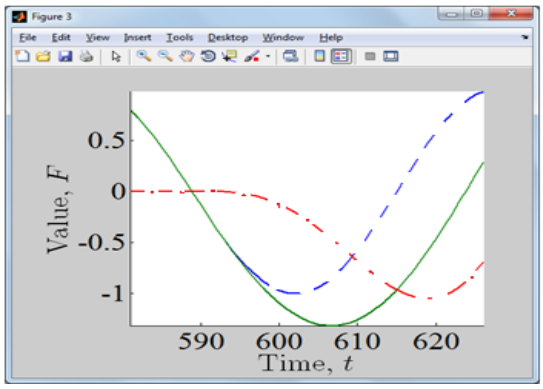

Fig. 5. Forecast f1 using a weighted Euclidean metric

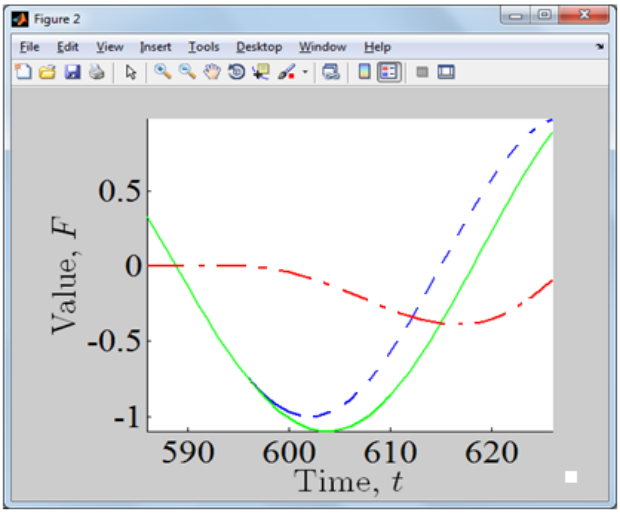

Fig. 6. Forecast f1 using the Minkowski metric

1.2. Similar results showed a weighted Euclidean metric (Fig. 5). This metric is allocated only one nearest neighbor, so the error became even more than the usual Euclidean metric.

1.3. More accurate forecast managed to build using the Minkowski metric Lp (Fig. 6):

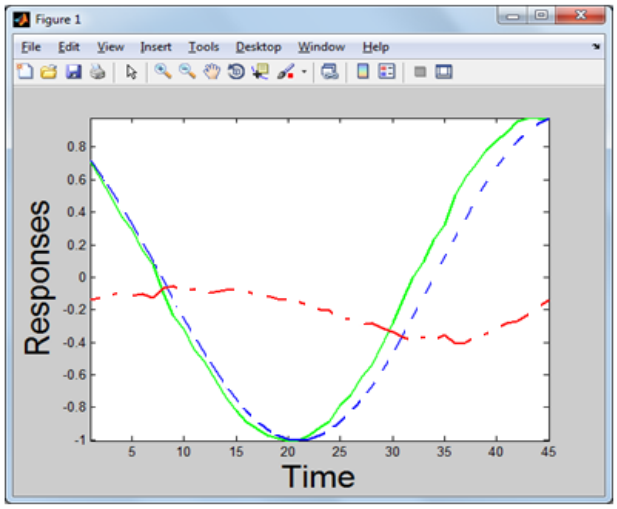

Figure 7. F1 Prediction using the conversion on the abscissa

2. Prediction lineup with patterns transformed length

We apply the algorithm seeking potential neighbors, spending over them transform along the horizontal axis and comparing them with the current pattern (Fig. 7). Of Figure 7 shows that for processes with varying intervals, this algorithm gives the smallest error.

3. Estimates of production of electricity from coal

Construct a forecast for real data describing the production of electricity from coal. These examples are cyclical with a constant period of 12 months, so the conversion on the abscissa for these series is not necessary.

The first row reflects the dynamics of the total electricity generated from coal in the United States from January 1973 to December 2013 (Fig. 8) (U.S. Energy Information Administration (EIA):

International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)
Volume No.3 Issue No.9, pp : 1217-1224 1 Sep 2014

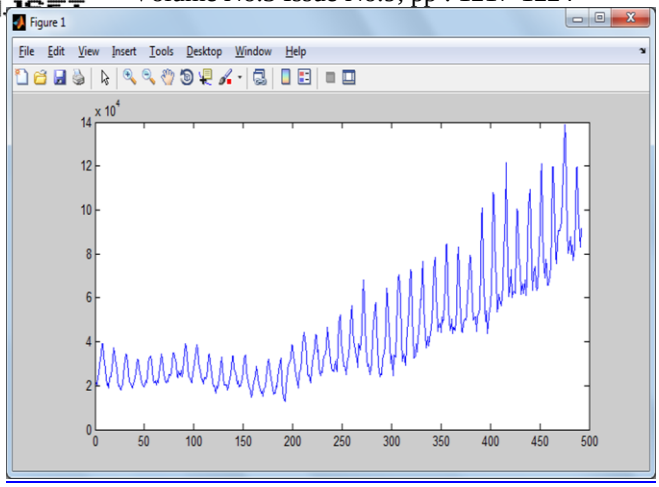

Fig. 8. Electricity production from coal in the U.S.

Calculate the forecast of production of electricity from coal in 2013 and compare it with actual data. Estimated value of the Dickey-Fuller test is the right of the critical value (significance 63%), which also speaks of non-stationary processes (Fig. 9).

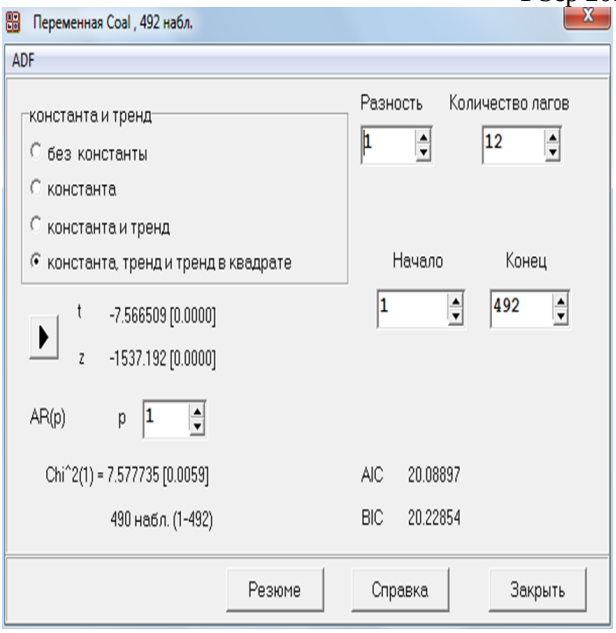

Fig. 9. Dickey-Fuller test for the production of electricity from coal

After taking first differences of significance criterion became zero, indicating that their stationary. From this it can be concluded that the process is an integrated first order.

Construct forecasts of electricity production from coal to that observed previously metrics with patterns of variable length. All three criteria give a fairly accurate prediction (Fig. 10, 11, 12).

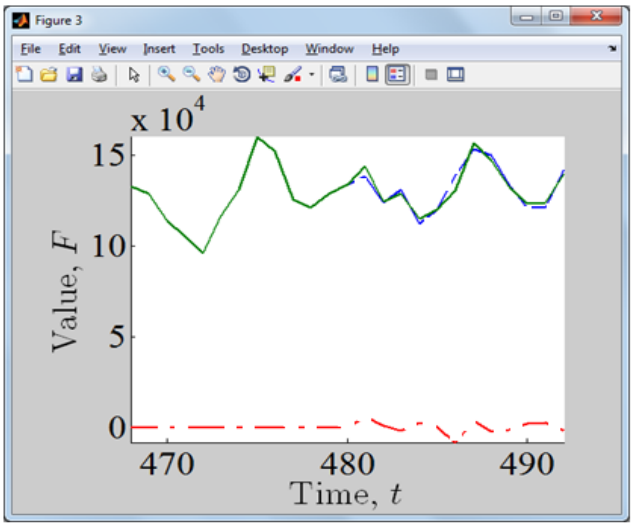

Fig. 10. Forecast of production of electricity from coal using the Euclidean metric

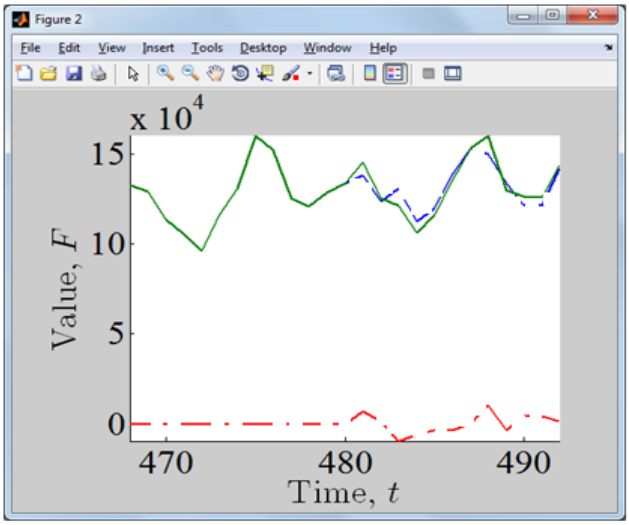

Fig. 11. Forecast of production of electricity from coal using a weighted Euclidean metric

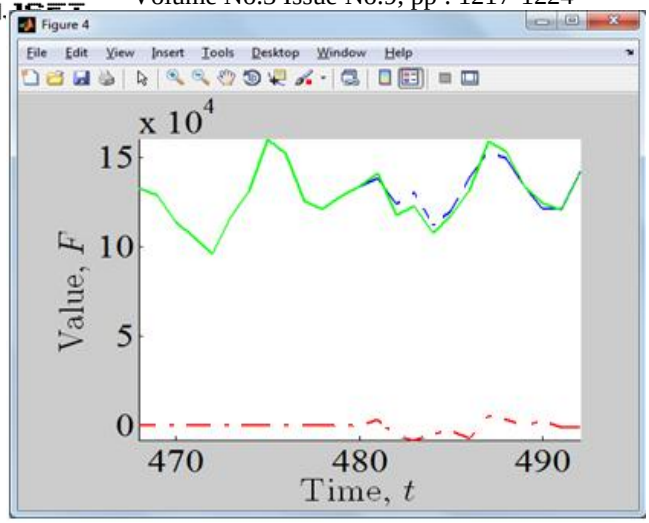

Fig. 12. Forecast of production of electricity from coal using the Minkowski metric

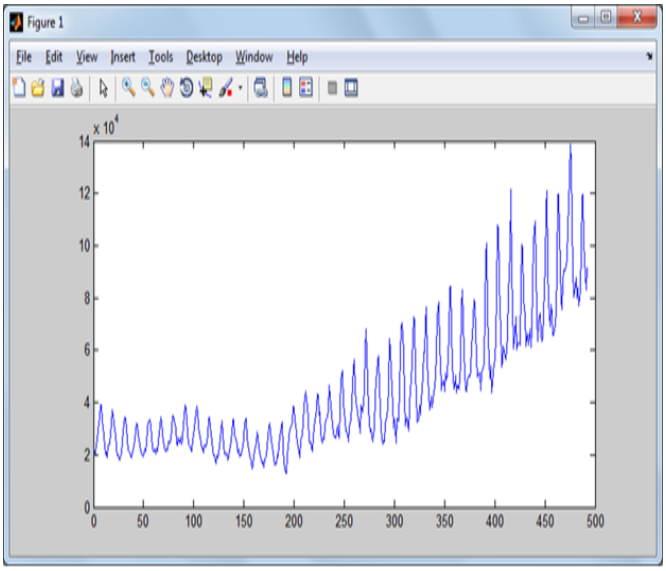

Figure 13. Production of electricity from natural gas in the U.S.

Estimated value of the Dickey-Fuller test is the right of the critical value (significance of 84.8%), indicating that the non-stationary processes (Figure 14)

After taking first differences of new relevance criterion became zero, indicating that their stationary. From this it can be concluded that this process is also the first order integrated.

To assess which of the metrics gave the most accurate result obtained are comparable errors. Table 1 shows that the most accurate was the Minkowski metric.

Table 1. Comparison of metrics for the production of electricity from coal

Metric	Euclidean	Weighted Euclidean	Minkowski
Number of nearest neighbors	2	5	3
Error	43.7374	3.63	2.8834

4. Estimates of production of electricity from natural gas

The second row reflects the dynamics of the total electricity generated from natural gas in the United States from January 1973 to December 2013 (Fig. 13) (U.S. Energy Information Administration (EIA): http://www.eia.gov/electricity/).

Calculate the forecast electricity production from natural gas in 2013 and compare it with actual data.

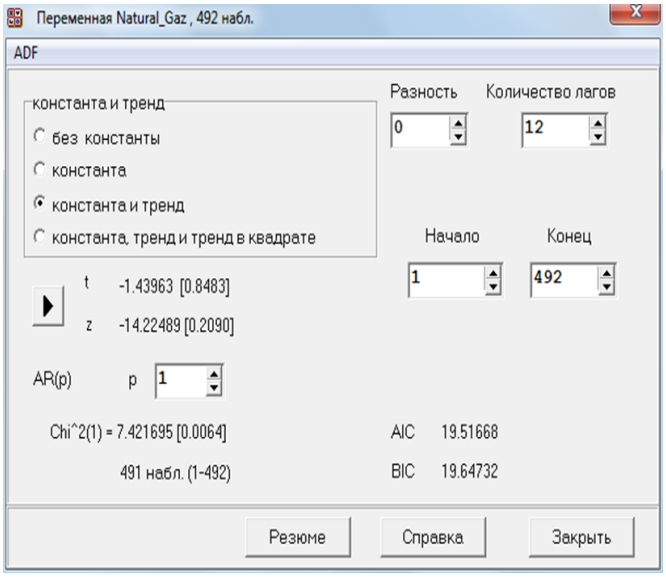

Figure 14. Dickey-Fuller test for the production of electricity from natural gas

Forecasts constructed using all three metrics given enough small errors (Fig. 15, 16, 17):

International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)
Volume No.3 Issue No.9, pp : 1217-1224 1 Sep 2014

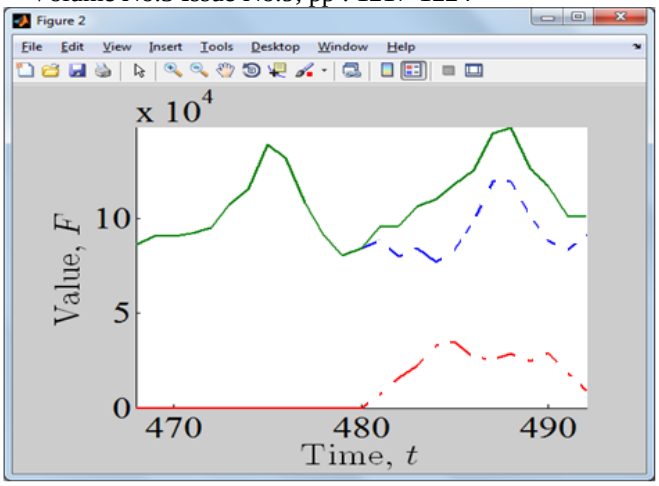

Fig. 15. Forecast of production of electricity from natural gas using the Euclidean metric

Fig. 15. Forecast of production of electricity from natural gas using the Euclidean metric

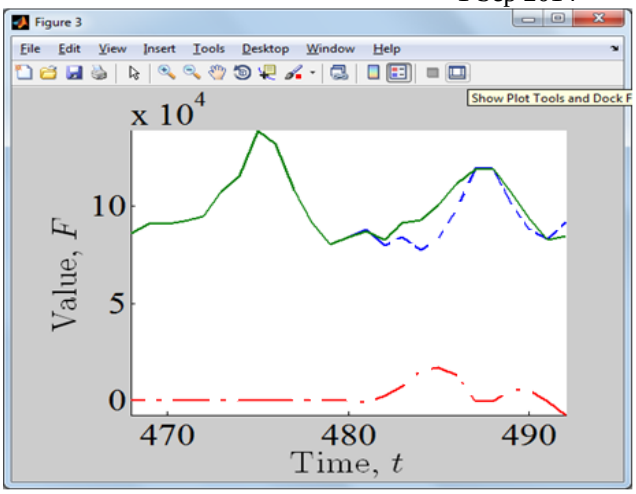

Fig. 16. Forecast of production of electricity from natural gas using a weighted Euclidean metric

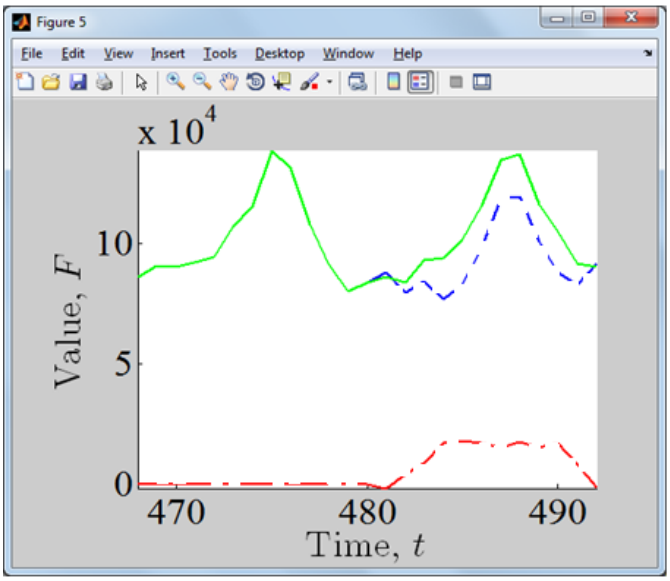

Fig 17. Forecast of production of electricity from natural gas using the Minkowski metric

To assess which of the metrics gave the most accurate result obtained are comparable to the error in Table 2. When forecasting the production of electricity from natural gas was the most accurate weighted Euclidean metric.

Table 2.

Comparing metrics for the production of electricity from natural gas

Metric	Euclidean	Weighted Euclidean	Minkowski
Number of nearest neighbors	10	3	30
Error	21.9496	6.9177	11.8183

Thus, from the above calculations show that for the series with a variable period, better suited algorithm k nearest neighbors, allowing to carry out the transformation patterns along the horizontal axis. For quite a time series of cyclic algorithm used with a constant pattern length equal to the length of the cycle.

Selection of a metric is also highly dependent on a number. For less noisy series, reflecting the production of electricity from coal, better suited Minkowski metric, and for more noisy electricity production from natural gas - weighted Euclidean, as often it uses minimal number of nearest neighbors and thus less susceptible to noise.

Conclusion

In this paper we analyzed several promising areas of forecasting non-stationary time series. Urgent task is to predict financial time series, which greatly impeded their nonstationarity. Dealt with the problem of forecasting dimensional non-stationary time series by the *k*-nearest neighbors. Describes various methods for determining the patterns (fixed-length transforms and abscissa), the proximity metrics, determining the number of neighbor method (fixed number or dynamic determination), and finally, the quality criteria of the algorithm.

To study the variations of the algorithm was built prediction model data with varying period and forecasts of real cyclic series production of electricity from coal and natural gas. The most accurate forecast model data was built using the algorithm with the conversion of the abscissa, the algorithm with fixed-length pattern gave high errors. Forecast for cyclic series production of electricity produced by the algorithm with fixed and variable-length pattern (12 months), proved to be quite accurate.

It has been found that selection of the most suitable metric is also highly dependent on a number. For less noisy series, reflecting the production of electricity from coal, better suited Minkowski metric, and for more noisy production of natural gas - the weighted Euclidean, as often it uses minimal number of nearest neighbors and thus less susceptible to noise.

References

i. *Alesgerov FT, Belousov VU, LG Egorova, Mirkin BG. (2013). Pattern analysis in statics and dynamics, Part 1: a literature review and refinement of the concept//Business Informatics, № 3 (25),. Pp. 3-18.*

ii. *Banavas G. N., Denham S., Denham M. J. (2000). Fast nonlinear deterministic forecasting of segmented stock indices using pattern matching and embedding techniques: Computing in Economics and Finance, 2000, 64: Society for Computational Economics.*

iii. *Demin AV, Vityaev E.E. (2009). Financial time series: forecasting and detection of violations dynamics Reports Vseross. conf. Umbrella-09 "Knowledge-Ontology-Theory", October 22-24, , Novosibirsk Univ IM SB RAS, 2009. - P. 79-86.*

iv. *Exchange information-analytical portal ClusterDelta.com: http://clusterdelta.com/patterns/*

v. *Filipenkov NV. (2006). On problems of time-series analysis of beams with varying laws. - Ukraine. Boxed intelligence..*

vi. *Lukashin YP. (2003). Adaptive methods of short-term time series prediction. - Moscow: Finance and statistics.*

vii. *MQL5.community: http://www.mql5.com/ru/code/133*

viii. *Ovchinnikov AA, Hramov AE, Lyuttehann A. Koronovskii AA. (2011). The diagnostic method is characteristic patterns on observed time series and its experimental implementation in real relation to neurophysiological signals. - St. Petersburg: Technical Physics.*

ix. *Perminov G. (2013). Prediction of Non-stationary Time Series With Replacement Variables//China-USA Business Review. July, Vol. 12, No. 7*

x. *Perminov GI, Leonova NV. (2012). Prediction of essentially non-stationary multivariate time series as an example indicator Russian investment nonbank corporations abroad//Culture of the Black Sea region, the scientific journal, , № 231. Pp. 74-78.*

xi. *Piatnitski MA. Pattern Recognition and Bioinformatics: http://bioinformatics.ru/Data-Analysis/patrecog_bioinf.html*

xii. *Romanenko AA. (2011). Alignment of time series: forecasting using DTW//Machine learning and data analysis. -. - № 1. - S. 77-85. - ISSN 2223-3792.*

xiii. *Suslov VI, Ibragimov NM, Talysheva LP Zyplakov AA. (2006). Time series analysis: studies. Benefit/NSU. - Novosibirsk,. - 207.*

xiv. *Trofimov AG, Skrugin VI. (2010). Adaptive classifier multidimensional non-stationary signals based on the analysis of dynamic patterns//Science and education.*

xv. *Tsiganova SV. (2011). Local prediction methods with a choice of conversion//"Intelligent Systems" FUPM MIPT. Machine Learning and Data Analysis, T. 1, № 2.*

xvi. *U.S. Energy Information Administration (EIA): http://www.eia.gov/electricity/*

xvii. *Varfolomeeva AA. (2011). Local prediction methods with a choice of metric.//"Intelligent Systems" FUPM MIPT. Machine Learning and Data Analysis,. T. 1, № 2.*

xviii. *Wolpaw J.R., Birbaumer N., McFarland D.J., Pfurtscheller G., Vaughan T.M. (2002). Brain–computer interfaces for communication and control//Clinical Neurophysiology, V.113, P.767–791.*