DOI: 10.17323/2587-814X.2021.4.22.35

To the question of restoring symbol sequences encoding noisy periodic functions

Galina N. Zhukova [a]
E-mail: galinanzhukova@gmail.com

Mikhail V. Ulyanov [b,c]
E-mail: muljanov@mail.ru

[a] National Research University Higher School of Economics
Address: 20, Myasnitskaya Street, Moscow 101000, Russia

[b] Trapeznikov Institute of Control Sciences, Russian Academy of Sciences
Address: 65, Profsoyuznaya Street, Moscow 117997, Russia

[c] Lomonosov Moscow State University
Address: 1, Leninskie Gory, Moscow 119991, Russia

Abstract

In business informatics, one of the research subjects is the analysis of data on processes in applied subject areas; here problems of qualitative analysis arise. Such problems arise, for example, in the qualitative study of log files of business processes, in the analysis and prediction of time series and other processes of a different nature. Quite often, to represent information about the processes under study, the methods of qualitative analysis use symbolic coding, which makes it possible to remove unnecessary detailing of numerical descriptions. The relevance of this study is due to the fact that when working with the raw data, researchers often face the presence of noise and distortions of the data, which significantly complicates the solution of the problems of qualitative analysis. When working with symbolic representations of the processes under study, which quite often have a periodic nature, we observe noise of deletion, insertion and replacement of symbols, which complicate the solution of the problem of revealing and analyzing the periodicity. This article deals with the problem of recovering periodic symbolic sequences obtained by coding from samples of continuous periodic functions and distorted by noise of insertion, replacement and deletion of symbols. Trigonometric functions are considered as a specific example of synthetic time series data. To encode trigonometric functions, alphabets of various cardinalities are used. The article presents an experimental study of the dependence of the quality characteristics of the method of period and a periodically repeating fragment recovery, previously proposed by the authors and improved in this study. For alphabets of different cardinalities at fixed sampling intervals, the fraction of sequences with a satisfactorily reconstructed period and the relative error in determining the period are given. The quality of reconstruction of a periodically repeating fragment is estimated by the edit distance from the reconstructed periodic sequence to the original sequence distorted by noise.

Key words: symbolic sequence; cardinality of an alphabet; periodic sequence; sequence with noise; noise of insertion; noise of deletion; noise of change.

Citation: Zhukova G.N., Ulyanov M.V. (2021) To the question of restoring symbol sequences encoding noisy periodic functions. *Business Informatics*, vol. 15, no 4, pp. 22–35. DOI: 10.17323/2587-814X.2021.4.22.35

Introduction

One of the subjects of scientific research in business informatics is the analysis of data on processes recorded in applied subject areas [1]. The variety of emerging data analysis problems includes the tasks of qualitative analysis. They arise, for example, during a qualitative study of the log files of business processes [1] and are associated, for example, with determining the correspondence of the log file to the process model [2]. One of the ways to present information about the processes being studied in scientific research is to represent them in the form of time series. At the same time, a significant proportion of the tasks of qualitative analysis relates precisely to research in the field of time series, and is associated both with the analysis of their periodicity and, in general, with the problems of their complex analysis and forecasting [3–8]. Similar qualitative problems arise in the analysis of other processes of a different nature, represented by time series, for example, in environmental monitoring and forecasting environmental changes [9].

The observed values of the process under study, which are elements of the time series, are quite often exposed to random distortions caused by external factors. “The values compared to the elements of the resulting series also contain measurement errors and, in the general case, are subject to random external influences. Further, such measurement errors and the results of external influences are interpreted as noise” [10].

Let us point out some publications from the field of business informatics and management in which one way or another the influence of noise on the results of forecasting time series is discussed and/or neutralized. The authors in [11–13] note that when creating models and predicting the production and consumption of electricity, noise in the raw data affects the predictive power of models and the quality of forecasts. Stochastic forecasting of risks in business [14], including the risk of operating profit for firms, relies heavily on the assumption of incomplete and noisy data. The authors of [15] point out that when studying the behavior of bank clients by the method of clustering time series, errors and range of the raw data are taken into account. In [16], when assessing the efficiency of forecasting passenger air traffic flows using multiple error indicators, in particular, it was shown that the noise in the raw data affects the forecasting quality. The authors in [17] try to improve the quality of tourism demand forecasting using deep learning methods in conjunction with image processing of time series visualization, and thereby reduce the influence of errors in the raw data on the forecasting results.

Among other subject areas, we note, for example, microelectronics, where image blurring caused by noise strongly affects the quality of prediction [18] and observations of the biosphere. In [19], the authors note that biometric data (in a broad sense, a kind of data observations of the biosphere) often have observation gaps, outliers and breaks. For one more subject area – remote sensing of the Earth from space, we quote from [20]: "The use of time series of satellite data for monitoring the earth's surface is associated with the problem of taking into account all sorts of interfering factors leading to partial loss or distortion of information on the dynamics of spectral-reflective characteristics of objects of observation. Such factors include mist and cloudiness, which are opaque in the visible and near-infrared range, shadows from it, as well as measurement errors."

In order to solve the problems of qualitative analysis, time series data are subjected to symbolic coding, which allows us to remove unnecessary detailing of numerical descriptions [21–23]. In this case, the description of the elements of the time series or steps of the business process is encoded with a word over a finite alphabet, which is the object of further research. Moreover, such coding is relevant for the qualitative analysis of big data. This is due to the fact that the high accuracy of numerical representations of time series elements leads not only to unjustifiably large amounts of information, but also laborious calculations that do not improve the quality of the results obtained [22, 23].

Obviously, when studying time series, working with noisy data causes significant difficulties. This leads to the formulation of the problem of noise elimination. For noise reduction in numerical time series data, various smoothing methods are used, such as moving average, exponential smoothing, etc. [10]. However, these methods are not applicable when dealing with noisy symbolic sequences.

In the aspect of symbolic coding, the arising errors, interpreted as noise, lead to the fact that insertion, deletion and replacement of symbols occur in symbolic sequences. Therefore, for example, errors associated with the adjustment of measuring instruments, inaccuracies in manual data entry and accidental mistakes or deliberate distortion of the values of individual indicators, lead to replacement noise [19]. Registration errors are a source of deletion and insertion noise, and these noises can also occur during data preparation. We also note that the methods for detecting periodicity used for numerical sequences are not applicable when working with symbolic representations [24].

The relevance of this study is due to the fact that noise reduces the efficiency of time series analysis. In this regard, the article discusses the problem of recovering periodic symbolic sequences obtained by coding from samples of periodic functions and distorted by noise of insertion, replacement and deletion of symbols. Trigonometric functions are considered as a specific example of synthetic time series data. To encode trigonometric functions, alphabets of various cardinalities with different granularity of sampling intervals by model time are used.

The article presents an experimental study of the quality characteristics of the method of period and a periodically repeating fragment recovery, previously proposed by the authors in [25] and improved in this study. For alphabets of different cardinalities at fixed sampling intervals according to the model time, the fraction of sequences with a satisfactorily reconstructed value of the period and the relative error in determining the period are given. The quality of reconstruction of a periodically repeating fragment is measured by the edit distance from the reconstructed periodic sequence to the original sequence distorted by noise.

1. Terminology and notation

Further, we will use the notation introduced by us in article [25], which describes a method for recovering a periodic symbolic sequence.

Let Σ^σ be the alphabet of cardinality $\sigma = |\Sigma| \geq 2$. We will call a word of length n the symbolic sequence $q^\sigma = s_1, s_2, ..., s_n$ over a finite alphabet Σ^σ, where s is an arbitrary symbol of Σ^σ, and a subword or a fragment of the word q^σ is any sequence of symbols $s_k, s_{k+1}, ..., s_{l-1}, s_l$, $1 \leq k \leq l \leq n$.

We consider periodic symbolic sequences with period p. To avoid ambiguity in understanding, we will call the period p the length of a repeating subword (fragment), and call a part of a periodic sequence of length p a periodically repeating fragment. Unless otherwise stated, a periodically repeating fragment is a fragment of length p that begins with the first symbol of the periodic word.

Let $q^\sigma(m, p)$ be a periodic word containing $m \geq 8$ repeating fragments of length p; $\tilde{q}^\sigma(m, p)$ – a word over the same alphabet as $q^\sigma(m, p)$, but with introduced noise; $\bar{q}^\sigma(m, p)$ – a periodic word obtained by analysing $q^\sigma(m, p)$ using the algorithm from [25] with the improvements proposed in this article.

Note that the word $\bar{q}^\sigma(m, p)$ has the same length as $\tilde{q}^\sigma(m, p)$ and serves as an approximation of the periodic word $\tilde{q}^\sigma(m, p)$.

2. Statement of the problem

Let there be some continuous-time periodic process $g(t)$ from time t_0 to time $t_0 + n\Delta t$ during which the value $g(t_i)$, $i = 1, n$, is measured at regular intervals Δt. Partitioning the range of measured values of $g(t)$ into σ equal half-segments and encoding the values of $g(t_i)$ with symbols of alphabet Σ^σ, we get the symbolic sequence $q^\sigma = s_1, s_2, ..., s_n$ over this alphabet. Note that some symbols of alphabet Σ^σ may not be present in the resulting symbolic sequence. We will assume that the period p of the observed function $g$is a multiple of the length of interval $r\Delta t$ between successive measurements, i. e. $p = r\Delta t$, where r is an integer, due to which, when encoding the periodic function $g(t)$ on a segment of m periods, a periodic symbolic sequence is obtained, also containing m periods, while $n = mp = mr\,\Delta t$.

Now let us introduce random distortions into the measured values of the function $g(t)$. We will consider the noises of insertion, replacement, and deletion. Insertion of a new value corresponds to some failure in measurements, when an extraordinary measurement occurred between the scheduled measurements; deletion means the loss of a value when entering or transferring results of measurements; and replacement is considered an incorrect measurement or deliberate distortion of the data.

After introducing all the distortions, we get a noisy sequence, which we encode in the same alphabet Σ^σ and consider the problem of recovering a periodic symbolic sequence from a given noisy sequence.

In this article, we consider the symbolic sequences obtained by encoding values of a continuous periodic function (in particular, $\sin(t)$) in points with step Δt on a segment with a length of at least eight full periods ($m \geq 8$) when they are distorted by different types of noise.

When coding a continuous function, an essential role is played by the choice of the cardinality of the alphabet. We encode each function under consideration with symbols of alphabets of cardinality from 10 to 60 with a step of 10. Despite the fact that when encoding by partitioning the range of values of the encoded function into equal half-segments, some symbols of the alphabet may not occur in an undistorted periodic symbolic sequence. Such symbols can be observed in the sequence obtained by encoding the distorted sequence of measured values of the function.

Statement of the problem: to study the influence of the cardinality of the alphabet, the type of function and the noise level on the quality of reconstruction of the periodic symbolic sequence obtained by coding the values of the periodic function under conditions of its distortion by the noise of insertion, replacement and deletion of values.

3. Continuous periodic function encoding

Consider function $\sin(t)$ on the segment $[0 - 16\pi]$, which contains 8 full periods of function $\sin(t)$. To construct a numerical sequence with period p, we partition the interval $[0 - 16\pi]$ into $8p$ equal half-segments of length $\Delta t = \frac{2\pi}{p}$ and evaluate the value of function $\sin(t)$ in the middle of each half-segment, getting a sequence of $8p$ real numbers $y_1, y_2, ..., y_{8p}$.

For the purpose of symbolic coding of the values obtained in the alphabet of cardinality σ, we partition the range of sine values – segment $[-1, 1]$ into σ consecutive half-segments $I_1, I_2, ..., I_\sigma$ of equal length; in this case, the number of half-segments is equal to the cardinality of the alphabet. We assign the semi-segments $I_1, I_2, ..., I_\sigma$ to the symbols of the alphabet Σ^σ. Each number in the sequence $y_1, y_2, ..., y_{8p}$ is encoded by the symbol corresponding to the half-segment I_j, in which this number falls. As a result, we get the symbolic sequence $q^\sigma(m, p)$. Note that some symbols of alphabet Σ^σ may not appear in this sequence.

4. Method for determining the period from a periodic sequence with noise

To understand the proposed improvements to the method for constructing a periodically repeating fragment, we present a brief description of the method for determining the period [25, 26].

The method assumes counting the number of all subwords of length $k = 10$ in a noisy sequence $\tilde{q}^\sigma = s_1, s_2, ..., s_n$. Subwords of length 10 in $\tilde{q}^\sigma$ are taken with a shift by one symbol, i. e. the subwords $s_1, s_2, ..., s_{10}, s_2, s_3, ..., s_{11}$, etc, are considered. Those subwords that have met at least 3 times compose the set R. Each subword from R is associated with a list of position numbers of symbols of the sequence $\tilde{q}^\sigma$, starting from which this subword is included in $\tilde{q}^\sigma$. So, in the sequence "*abcabcdabcdeabcabcdabcde*" the subword "*abcabcdabc*" is included starting from position numbers 1 and 13.

Further, for each subword of length $k = 10$, the differences between the numbers of consecutive occurrences are calculated, after which the set Ω is constructed from such differences. Each element of this set is compared the number of times when such a difference in the numbers of consecutive occurrences of the first symbols of subwords of length 10 was observed. For example, in a word "*abcabcdabcdeabcabcdabcde*" the difference $r = 13 - 1 = 12$ is observed 3 times (for the words "*abcabcdabc*", "*bcabcdabcd*", "*cabcdabcde*"). Due to the introduced noise, the differences not always equal the period or its multiples, but most of the differences take values close to the period or its multiple. In this regard, for each value of the difference, it is calculated how many times a difference close to it has occurred in Ω. In our experiments, the proximity was defined as falling within the interval ±20%. The analysis of the obtained differences allows one to obtain an estimate of the period for an unknown strictly periodic sequence [25].

5. Improved method for constructing a periodically repeating fragment

The solution that the algorithm from [25] delivers in terms of constructing a periodically repeating fragment is the subword $\tilde{f}^{\sigma} = s_1, s_2, \ldots, s_{\tilde{p}}$ of the analyzed word $\tilde{q}^{\sigma}(m, p)$, minimizing (on the set of obtained variants of fragments) the edit distance from the symbolic sequence of length $|\tilde{q}^{\sigma}(m, p)|$ constructed from this fragment to the sequence $\tilde{q}^{\sigma}(m, p)$.

The construction of an approximating periodically repeating fragment $\tilde{f}^{\sigma}$ in [25] is carried out by splitting the distorted sequence into successive subwords of length $\tilde{p}$ (the last subword of length less than $\tilde{p}$ was not taken into account) and choosing such one from them, in which the edit distance to one of the remaining subwords is minimal. If there are several such subwords, the first subword with the minimum edit distance to another subword is selected.

The method proposed by us and described below attempts to improve the $\tilde{f}^{\sigma}$ fragment in order to obtain a smaller value of the edit distance between the distorted and approximating periodic sequences. The proposed improvement is achieved by using information about the previously determined frequencies of subwords of length $k = 10$ observed in $\tilde{q}^{\sigma}(m, p)$.

First, based on the fragment $\tilde{f}^{\sigma}$, the word $\tilde{f}_4^{\sigma}$ is built, containing the fragment $\tilde{f}^{\sigma}$, written four times in a row. Then each subword of the fragmen $\tilde{f}_4^{\sigma}$ of length $k = 10$ (($s_1, s_2, \ldots, s_{10}$; $s_2, s_3, \ldots, s_{11}$ etc.) is checked for occurrence in the word $\tilde{q}^{\sigma}(m, p)$ at least 3 times, i.e. to belong to the set R. The first of the checked subwords $\omega_0 = s_{t+1}, s_{t+2}, \ldots, s_{t+10}$, belonging to the set R, becomes a start subword, the improved periodic fragment will begin with it. If no such subword was found, the periodic fragment remains unimproved.

After the subword ω_0 is found, the subwords ω from $\tilde{f}_4^{\sigma}$ are sequentially scanned with a shift by one character, i. e. $\omega = s_{t+h}, s_{t+h+1}, \ldots, s_{t+h+9}$, $h = 2, 3, \ldots$, membership check is performed of these subwords to the set R. If the subword belongs to R, then the corresponding subword of the word $\tilde{f}_4^{\sigma}$ remains unchanged, otherwise we start counting consecutive subwords (with a shift by one character) not included in R.

Since the length of ω is 10, until there are 10 consecutive subwords not included in R, the word $\tilde{f}_4^{\sigma}$ remains unchanged. If, after less than 10 consecutive ω not included in R, the next ω turned out to be included in R, then the counter of consecutive subwords not included in R is reset to zero. If in R there are no 10 consecutive subwords ω but the eleventh ω is included in R, then the search for a subword ω^R in R begins, such that the first m symbols of ω^R coincide with the last m symbols of ω^- — the last of the considered subwords in R, $3 \leq m \leq 9$. The values of m are taken from 9 to 3, i. e. first we try to find ω^R, in which the first 9 symbols coincide with the last 9 symbols ω^-, if such ω^R is found, then in $\tilde{f}_4^{\sigma}$ replace the symbol following the last by the subword symbol ω^-, to the last symbol ω^R.

If for $m = 9$ the subword ω^R is not found, we continue with $m = 8$, and so on up to $m = 3$. Let ω^R be found for some m from 3 to 9, then in$\tilde{f}_4^{\sigma}$ we replace the symbol following the last symbol of the subword ω^- by the $(m + 1)$-st symbol ω^R. If no ω^R was found for any m from 3 to 9, then the fragment $\tilde{f}^{\sigma}$ remains unimproved.

At some point, either $2\tilde{p}$ consecutive subwords ω will be scanned and the fragment ω_0 will never be encountered, or the next ω subword will match ω_0. In the first case, as an improved fragment $\tilde{f}^{\sigma}$ we take the first $\tilde{p}$ symbols of $\tilde{f}_4^{\sigma}$ starting from the first symbol ω_0 (i.e. from $\tilde{f}_4^{\sigma}$ cut out ω_0 and the following symbols in $\tilde{f}_4^{\sigma}$, $\tilde{p}$ symbols in total). In the second case, $\tilde{f}^{\sigma}$ consists of symbols of $\tilde{f}_4^{\sigma}$, starting from the first occurrence of ω_0 to the second ω_0, while checking the possibility that at the end of the periodic fragment due to insertion or deletion noise, one symbol is lost or added.

The check is done as follows. If after the next ω included in R in $\tilde{f}_4^\sigma$ there were 9 consecutive subwords not included in R, after which a word from R was found and this word coincided with ω_0, then perhaps there was a noise of deletion. To check, in the set R, we look for ω^R, in which the first m symbols coincide with the last m symbols of ω. If such a word ω^R is found and its symbols starting from $m + 2$ coincide with the first symbols ω_0, then at the end of $\tilde{f}^\sigma$ we add the $(m + 1)$-st symbol of ω^R.

In addition, if after the next ω that is included in R, in $\tilde{f}_4^\sigma$ there were 9 consecutive subwords not included in R, after which a word from R was found and this word coincided with ω_0, it is possible that there was a noise of insertion, and we try to eliminate it like that. If the first symbol ω_0 matches the last symbol ω, then we do not include in $\tilde{f}_4^\sigma$ the last symbol before the second occurrence of ω_0 in $\tilde{f}_4^\sigma$.

In order to obtain a more accurate value of a period simultaneously with the search for the improved fragment $\tilde{f}^\sigma$, $2\tilde{p}$, consecutive subwords ω of the word $\tilde{f}_4^\sigma$, are searched, until on the next step among the ω we meet the word ω_0. Then as the improved fragment we regard the subword of the improved $\tilde{f}_4^\sigma$, starting from the first occurrence of ω_0 to the second one, if the period is at least 3. If changes in $\tilde{f}_4^\sigma$ during the improvement process did not yield another ω_0, then after the end of the improvement process, we perform another scan of $2\tilde{p}$ consecutive subwords ω of the word $\tilde{f}_4^\sigma$ starting from the first occurrence of ω_0, and if there is a second occurrence of ω_0, then the improved fragment is the subword of the improved $\tilde{f}_4^\sigma$, starting from the first occurrence of ω_0 to the second. If the second occurrence of ω_0 is not found, then $\tilde{p}$ consecutive symbols of improved $\tilde{f}_4^\sigma$, starting from the first occurrence of ω_0, are taken as the improved fragment.

If it is possible to improve a periodic fragment, its cyclic shifts are considered, and the final fragment is taken as the shift, which has the minimum edit distance to the beginning of the noisy sequence, i.e. its first $\tilde{p}^*$ symbols, where $\tilde{p}^*$ is the length of the improved periodic fragment.

If, as a result of improving a fragment, a subword of length more than three is obtained, and the edit distance from the periodic sequence constructed by repeating the improved fragment to the noisy one is less than in the case $\tilde{f}^\sigma$ before improvement, we use improved $\tilde{f}^\sigma$ as an approximation of the periodically repeating fragment, otherwise original.

6. Evaluation of the quality of the period recovery

The estimation of the accuracy of determining the period and the quality of reconstruction of a periodic fragment will be carried out separately. Let the period of the sequence before introducing distortions be p, and our algorithm determined that the period is $\overline{p} = \overline{p}\left(\overline{q}^\sigma\right)$. Then the precision δ of determining the period is defined as

$$\delta = \frac{\left|p - \overline{p}\right|}{p}. \qquad (1)$$

For a periodic sequence obtained by encoding a periodic function on eight periods, we get a series of 100 random noisy sequences. For each of them we determine the period and find the value δ, then we calculate the average value δ and the median of the sample over 100 noisy sequences.

The quality of reconstruction of a periodic fragment will be estimated by evaluating the ratio of the edit distance $d\left(\overline{q}^\sigma, q^\sigma(m,p)\right)$ between the reconstructed and original periodic symbolic sequences, to the length of the original periodic sequence [27]. Let's denote this ratio $\varepsilon\left(\overline{q}^\sigma, q^\sigma\right)$:

$$\varepsilon\left(\overline{q}^\sigma, q^\sigma\right) = \frac{d\left(\overline{q}^\sigma, q^\sigma(m,p)\right)}{mp}. \qquad (2)$$

With this approach, a good estimate will be given by the case when the period obtained by the algorithm is two to three times larger than the initial one, but at the same time the periodic fragment is close to the original, repeated the required number of times.

In addition, the original and reconstructed sequences were compared with the noisy one; in this case, the first n symbols were taken from each periodic sequence for comparison, where n is the length of the noisy sequence.

7. Computational experiment scheme

In a computational experiment, a study was carried out of the method proposed in [25, 26] and improved in this article on the following functions:

- $\sin(t)$on the segment $[0 - 16\pi]$;
- $\left(t-8\left\lfloor\frac{t}{8}\right\rfloor+1\right)\sin(\pi t)$ on the segment [0 – 64];
- $\left(t-16\left\lfloor\frac{t}{16}\right\rfloor+1\right)\sin(\pi t)$ on the segment [0 – 128];
- $\left|7-t+16\left\lfloor\frac{t}{16}\right\rfloor\right|\sin(\pi t)$ on the segment [0 – 128], where $\lfloor\ldots\rfloor$ stands for the integer part of the number.

Thus, for all functions, the segment of argument values containing eight full periods was considered. The functions were encoded in the alphabets Σ^{σ} of cardinality δ from 10 to 60 with a step of 10. The range of function values was partitioned into δ intervals, with each interval I_j being encoded by the symbol s_j of the alphabet Σ^{σ}.

The values of the period p were chosen equal to 20, 30 and 50, while one period of the function was partitioned into p equal half-segments. Then for each half-segment Δt_i the value of the function in the middle of the half-segment was evaluated, and this value was used to determine the half-segment I_j to which this value belongs, after which the symbol s_j, encoding the half-segment I_j, was written to the i-th position of the coding word. Thus, periodic symbolic sequences were obtained in alphabets of cardinality from 10 to 60.

Based on the periodic sequences thus constructed, random noisy sequences were obtained, 100 sequences for each purely periodic one. The noise was introduced in accordance with our earlier proposed probabilistic noise model for periodic symbolic sequences [28].

In the first series of experiments, noise was introduced with a total level of 5%, with the levels of insertion, replacement, and deletion noise taking values from 1 to 5% with a step of 1%, so that the sum of the noise levels is 5%. In the second series, noise was introduced uniformly distributed among the types, i. e. the level of insertion, replacement and deletion noise was the same in a separate experiment, taking values from 1 to 4% in 1% increments.

8. Results and discussion

The results of experimental studies for all four functions are shown in *Figures 1–4* and in *Tables 1* and *2*. In *Figures 1–4*, we show the numerical sequences of the values of functions on two periods containing 50 samples per period, with the coding alphabet cardinality of 20. In the figures, the numbers of the samples are plotted along the horizontal axis, and the numbers of the half-segments of the coding corresponding to the symbols of alphabet of cardinality 20 are shown along the vertical axis.

For all functions, as an example, one variant of noise injection is shown from the various variants investigated. Namely, a variant with a total noise level of 5%, consisting of a deletion noise of 2% and an insertion noise of 3%.

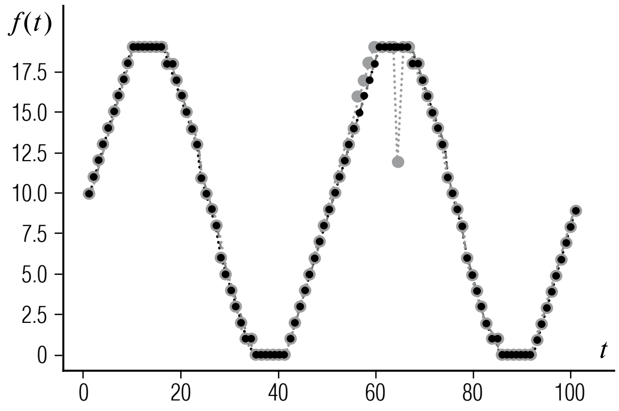

Fig. 1. Two periods of the function $\sin(t)$, total noise 5%, deletion noise 2%, insertion noise 3%, 50 samples per period, coding alphabet cardinality 20 symbols

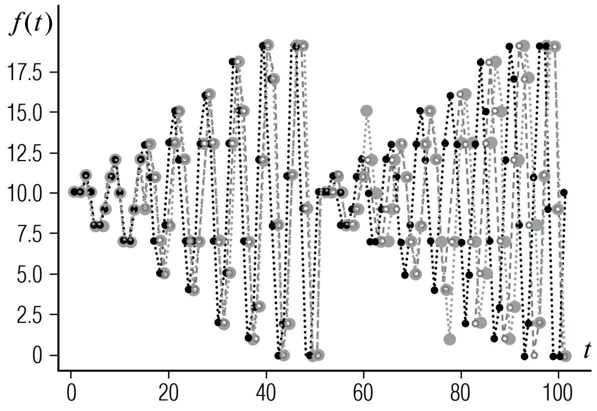

Fig. 3. Two periods of the function $\left(t-16\left\lfloor\frac{t}{16}\right\rfloor+1\right)\sin(\pi t)$, total noise 5%, deletion noise 2%, insertion noise 3%, 50 samples per period, coding alphabet cardinality 20 symbols

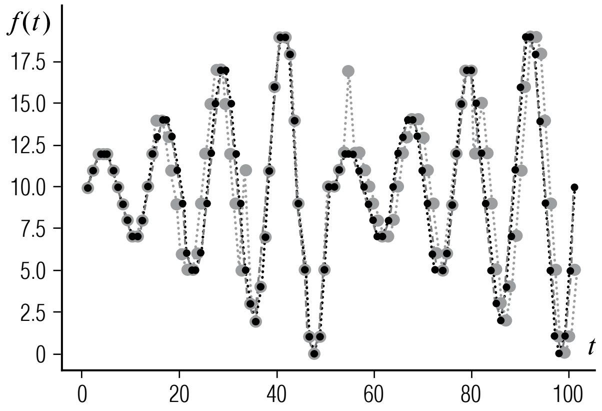

Fig. 2. Two periods of the function $\left(t-8\left\lfloor\frac{t}{8}\right\rfloor+1\right)\sin(\pi t)$, total noise 5%, deletion noise 2%, insertion noise 3%, 50 samples per period, coding alphabet cardinality 20 symbols

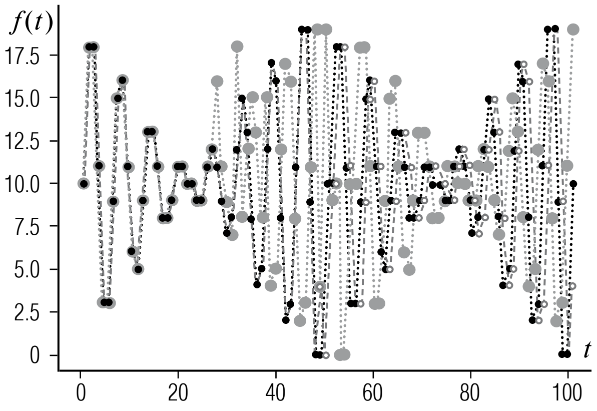

Fig. 4. Two periods of the function $\left|7-t+16\left\lfloor\frac{t}{16}\right\rfloor\right|\sin(\pi t)$, total noise 5%, deletion noise 2%, insertion noise 3%, 50 samples per period, coding alphabet cardinality 20 symbols

In all the figures, the black dots correspond to the original periodic sequence obtained from the specified function, the gray dots are the noisy sequence of function values in samples, and the white ones with an outline are the periodic sequence restored by the algorithm under study. The dotted line shows the sequence of points by count.

The results of the study on evaluating the accuracy δ of determining the period showed that for all functions the results obtained do not differ much from each other, i. e. the method under investigation is weakly sensitive to the form of a periodic function (at least for these four functions). Therefore, we present the results for only one function.

Table 1.

Fraction of symbolic sequences with a recovered period, not more than 2% different from the original, the function is $\left(t-8\left\lfloor\frac{t}{8}\right\rfloor+1\right)\sin\left(\pi t\right)$

Noise level, %			Alphabet cardinality					
deletion	change	insertion	10	20	30	40	50	60
0	0	5	25	26	33	24	26	31
0	1	4	25	16	22	26	24	26
0	2	3	98	97	94	95	96	93
0	3	2	99	100	98	99	99	100
0	4	1	100	98	100	100	100	100
0	5	0	100	100	100	100	100	100
1	0	4	94	93	94	90	95	96
1	1	3	100	99	96	100	97	97
1	2	2	99	100	99	100	98	97
1	3	1	99	97	97	97	98	99
1	4	0	94	96	97	96	97	95
2	0	3	99	96	99	99	96	97
2	1	2	98	97	99	100	99	98
2	2	1	99	100	99	97	100	97
2	3	0	97	98	97	98	98	98
3	0	2	100	100	98	97	99	100
3	1	1	98	97	98	96	98	98
3	2	0	85	94	91	91	87	90
4	0	1	88	85	88	83	90	88
4	1	0	22	34	28	26	35	31
5	0	0	8	9	7	7	17	12

For better clarity, we give in *Table 1* not the average value δ averaged over 100 experiments by introducing random noise with a given level, but the fraction of sequences (out of 100 noisy ones) with a reconstructed period, the value of which differs by no more than 2% from the initial period.

The results of studies on the quality of reconstruction of a periodically repeating fragment by the improved algorithm are shown in

Table 2.

Influence of the noise level and cardinality of the alphabet on the median $\varepsilon\left(\overline{q}^{\sigma}, q^{\sigma}\right)\cdot 100\%$

Total noise %	Alphabet cardinality	Median of values $\varepsilon\left(\overline{q}^{\sigma}, q^{\sigma}\right)\cdot 100\%$ of different functions			
		$\sin(t)$	$\left(t-8\left\lfloor\frac{t}{8}\right\rfloor+1\right)\sin(\pi t)$	$\left(t-16\left\lfloor\frac{t}{16}\right\rfloor+1\right)\sin(\pi t)$	$\left\lvert 7-t+16\left\lfloor\frac{t}{16}\right\rfloor\right\rvert\sin(\pi t)$
3%	10	0	0	0	0
	20	0	0	0	0
	30	0	0	0	0
	40	0	0	0	0
	50	0	0	0	0
6%	10	2	2	2	2
	20	2	2	2	2
	30	2	2	2	2
	40	2	2	2	2
	50	2	2	2	2
9%	10	4	4	4	4
	20	6	4	4	4
	30	4	4	4	4
	40	4	4	5	4
	50	4	4	5	4
12%	10	7	6	8	6
	20	8	6	8	6
	30	6	6	6	6
	40	8	8	6	8
	50	8	8	7	6

Table 2. The values of the median $\varepsilon\left(\overline{q}^{\sigma}, q^{\sigma}\right)$ are given for all four functions under study with a noise of uniform structures. In this case, the same noise level of each type (insertion, deletion, replacement) was randomly introduced into each of the 100 initial sequences. This level varied from 1 to 4% with a step of 1%. The experiments were carried out for all cardinalities of the alphabet – 10, 20, 30, 40 and 50 and all values of the period. *Table 2* shows the results for a period value of $p = 50$.

Based on the experimental data obtained, the improved method for reconstructing the period and periodically repeating fragment shows generally satisfactory results. The data in *Tables 1* and *2* show that the method has weak sensitivity in terms of the cardinality of the coding alphabet and in the form of the function, both in determining the period and in recovering a periodically repeating fragment.

When determining the period (see *Table 1*), the method is susceptible to the difference between the levels of insertion and deletion noise, since it is this difference that affects the distances between repeated subwords of length 10 in a noisy sequence. Note that the presence of only 5% replacement noise leads to the best experimentally observed result. When determining a periodically repeating fragment with a noise of uniform structure, neither the form of the function, nor the cardinality of the alphabet has a noticeable effect on the results. The only influencing factor in this case is the overall noise level.

Conclusion

The use of models of symbolic cycles with noise makes it possible to solve problems of probabilistic forecasting of symbolic noisy sequences and allows you to develop effective methods for forecasting, reconstruction and approximation of data in the form of symbolic codes based on fragmentary, incomplete and distorted information.

This article proposes an improved method for solving the problem of recovering a periodic symbolic sequence based on the original sequence obtained by introducing insertion, deletion and replacement noise into an unknown periodic sequence. The method is based on the study of frequency occurrence and distances between coinciding subwords of fixed length. Symbolic sequences in alphabets of different cardinality encoding noisy periodic functions are considered as synthetic data. In addition to describing the improved method for finding a periodically repeating fragment, the article contains the results of an experimental study of the dependence of the quality characteristics of the method for restoring the period and periodically repeating fragment on the cardinality of the coding alphabet and noise levels of various types. The study was carried out for noisy symbolic codes of periodic functions, which are models of noisy (quasiperiodic) time series. This kind of input data often occurs in the problems of analysis and forecasting of time series in business informatics and management.

A computational experiment has shown that the quality of the method depends not only on the general noise level, but also on the ratio of the noise level. The proposed method has weak sensitivity in terms of the cardinality of the coding alphabet and in the form of a periodic function, both in determining the period and in recovering a periodically repeating fragment. A study for noise with a uniform structure showed that the only factor affecting quality is the noise level, while neither the type of function, nor the cardinality of the alphabet has a noticeable effect on the results.

The results obtained in the article make it possible to give recommendations on the possible application of the method when solving problems of analyzing symbolic codes of noisy periodic continuous functions in alphabets of small cardinality, with a noise level not exceeding 10–12%. Such problems arise in the analysis of both dynamic processes and time series in business informatics and management, and in the analysis of business processes in conditions of incomplete and fragmentary information. ■

Acknowledgments

This work was supported by RFBR grants 19-07-00150 and 19-07-00151.

References

1. Andersen B. (2007) *Business process improvement toolbox.* Milwaukee, Wisconsin: ASQ Quality Press.
2. Mitsyuk A.A., Lomazova I.A., van der Aalst W. (2017) Using event logs for local correction of process models. *Automatic Control and Computer Sciences*, vol. 51, no 7, pp. 709–723. DOI: 10.3103/S0146411617070306.
3. Keogh E.J., Pazzani M.J. (1998) An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. Proceedings of the *Fourth International Conference on Knowledge Discovery and Data Mining (KDD '98), New York, USA, 27–31 August 1998, pp. 239–241.* Available at: https://www.aaai.org/Papers/KDD/1998/KDD98-041.pdf (accessed 16 November 2021).
4. Bemdt D.J., Clifford J. (1994) Using dynamic time warping to find patterns in time series. *AAAI Technical Report. Workshop on Knowledge Discovery in Databases (KDD '94), Seattle, Washington, USA, 31 July – 1 August 1994, pp. 359–370.* Available at: https://www.aaai.org/Papers/Workshops/1994/WS-94-03/WS94-03-031.pdf (accessed 16 November 2021).
5. Wu Y.-L., Agrawal D., el Abbadi A. (2000) A comparison of DFT and DWT based similarity search in time-series databases. Proceedings of the *Ninth International Conference on Information and Knowledge Management (CIKM' 00), McLean, Virginia, USA, 6–11 November*, pp. 488–495. DOI: 10.1145/354756.354857.
6. Ding H., Trajcevski G., Scheuermann P., Wang X., Keogh E. (2008) Querying and mining of time series data: Experimental comparison of representations and distance measures. Proceedings of the *VLDB Endowment*, Auckland, New Zealand, 23–28 August 2008, vol. 1, no 2, pp. 1542–1552. DOI: 10.14778/1454159.1454226.
7. Kurbalija V., Radovanović M., Geler Z., Ivanović M. (2011) The influence of global constraints on DTW and LCS similarity measures for time-series databases. *Advances in Intelligent and Soft Computing*, vol. 101, pp. 67–74. DOI: 10.1007/978-3-642-23163-6_10.
8. Dreyer W., Dittrich A.K., Schmidt D. (1994) Research perspectives for time series management systems. *ACM SIGMOD Record*, vol. 23, no 1, pp. 10–15. Available at: https://dl.acm.org/doi/abs/10.1145/181550.181553 (accessed 16 November 2021).
9. Rozenberg G.S., Shitikov V.K., Brusilovskij P.M. (1994) *Environmental forecasting (time series functional predictors)*. Tolyatti: IEVB RAS (in Russian).
10. Sklyar A.Ya. (2019) Analysis and elimination of noise component in time series with variable step. *Cybernetics and Programming*, no 1, pp. 51–59 (in Russian). DOI: 10.25136/2306-4196.2019.1.27031.
11. Mauleón I. (2021) Aggregated world energy demand projections: Statistical assessment. *Energies*, vol. 14, no 15, pp. 1–13. Available at: https://www.mdpi.com/1996-1073/14/15/4657 (accessed 16 November 2021). DOI: 10.3390/en14154657.
12. Suganthi L., Samuel A.A. (2012) Energy models for demand forecasting – A review. *Renewable and Sustainable Energy Reviews*, vol. 16, no 2, pp. 1223–1240. DOI: 10.1016/j.rser.2011.08.014.
13. Boßmann T., Staffell I. (2015) The shape of future electricity demand: Exploring load curves in 2050s Germany and Britain. *Energy*, vol. 90, no 2, pp. 1317–1333. DOI: 10.1016/j.energy.2015.06.082.
14. Akca A., Canakoğlu E. (2021) Adaptive stochastic risk estimation of firm operating profit. *Journal of Industrial and Business Economics*, vol. 48, no 3, pp. 463–504. DOI: 10.1007/s40812-021-00184-z.
15. Abbasimehr H., Shabani M. (2021) A new methodology for customer behavior analysis using time series clustering. A case study on a bank's customers. *Kybernetes,* vol. 50, no 2, pp. 221–242. DOI: 10.1108/K-09-2018-0506.
16. Fildes R., Wei Y., Ismail S. (2011) Evaluating the forecasting performance of econometric models of air passenger traffic flows using multiple error measures. *International Journal of Forecasting*, vol. 27, no 3, pp. 902–922. DOI: 10.1016/j.ijforecast.2009.06.002.
17. Bi J.-W., Li H., Fan Zh.-P. (2021) Tourism demand forecasting with time series imaging: A deep learning model. *Annals of Tourism Research,* vol. 90, article no 103255. DOI: 10.1016/j.annals.2021.103255.
18. Meszmer P., Majd M., Prisacaru A., Gromala P.J., Wunderle B. (2021) Neural networks for enhanced stress prognostics for encapsulated electronic packages – A comparison. *Microelectronics Reliability,* vol. 123, article no 114181. DOI: 10.1016/j.microrel.2021.114181.

19. Deshcherevsky A.V., Zhuravlev V.I., Nikolskij A.N., Sidorin A.Ya. (2016) Time series analysis problems with gaps and methods for solving them in the program WINABD. *Geophysical processes and the biosphere*, vol. 15, no 3, pp. 5–34 (in Russian).

20. Plotnikov D.E., Miklashevich T.S., Bartalyov S.A. (2014) Reconstruction of time series of remote sensing data by the method of polynomial approximation in a sliding window of variable size. *Current Problems in Remote Sensing of the Earth from Space*, vol. 11, no 2, pp. 103–110 (in Russian).

21. Lin J., Keogh E., Wei L., Lonardi S. (2007) Experiencing SAX: A novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, vol. 15, no 2, pp. 107–144. DOI: 10.1007/s10618-007-0064-z.

22. Zhukova G., Smetanin Y., Uljanov M. (2019) Informative symbolic representations as a way to qualitatively analyses time series. Proceedings of the *2019 International Conference on Engineering Technologies and Computer Science: Innovation & Application*, Moscow, Russia, 26–27 March 2019, pp. 43–47.

23. Lin J., Keogh E., Lonardi S., Chiu B. (2003) A symbolic representation of time series, with implications for streaming algorithms. Proceedings of the *8th ACM SIGMOD Workshop on Research Issues in Data mining and Knowledge Discovery, San Diego, California, USA, 13 June 2003*, pp. 2–11. DOI: 10.1145/882082.882086.

24. Nesterenko A.Yu. (2010) Algorithms for finding the lengths of cycles in sequences and their applications. *Fundamental and Applied Mathematics*, vol. 16, no 6, pp. 109–122 (in Russian).

25. Zhukova G.N., Zhukov A.V., Smetanin Yu.G., Ulyanov M.V. (2020) The method of estimating the period of a symbolic periodic sequence with noise, based on the sub-words positions in the sequence. *Modern information technologies and IT education*, vol. 16, no 1, pp. 23–32 (in Russian). DOI: 10.25559/SITITO.16.202001.23-32.

26. Ulyanov M.V. (2020) An approach to identifying the cycle length in noisy character sequences based on the entropy of words. Proceedings of the *III International Scientific and Technical Forum "Modern Technologies in Science and Education", Ryazan, 4–6 March 2020*, vol. 4, pp. 120–124 (in Russian).

27. Levenshtejn V.I. (1965) Binary codes with corrected dropouts, insertions and character replacements. *Proceedings of the Academy of Sciences*, vol. 163, pp. 707–710 (in Russian).

28. Zhukova G.N., Zhukov A.V., Smetanin Yu.G., Ulyanov M.V. (2019) Stochastic model of noises for periodic symbol sequences. *Modern information technologies and IT education*, vol. 15, no 2, pp. 431–440. DOI: 10.25559/SITITO.15.201902.431-440 (in Russian).

About the authors

Galina N. Zhukova

Cand. Sci. (Phys.-Math.);

Associate Professor, School of Software Engineering, Faculty of Computer Science, National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: galinanzhukova@gmail.com

ORCID: 0000-0003-1835-7422

Mikhail V. Ulyanov

Dr. Sci. (Tech.);

Leading Researcher, Laboratory of Scheduling Theory and Discrete Optimization, V.A. Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, 65, Profsoyuznaya Street, Moscow 117997, Russia;

Professor, Department of Algorithmic Languages, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, 1, Leninskie Gory, Moscow 119991, Russia;

E-mail: muljanov@mail.ru

ORCID: 0000-0002-5784-9836