ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ

Системный анализ, управление и обработка информации (05.13.01)

УДК 519.27
DOI: 10.24160/1993-6982-2021-3-67-77

Разработка алгоритмов обнаружения разладки временных рядов на основе непараметрических критериев проверки гипотез

Г.Ф. Филаретов, З. Бучаала

Приведено решение задачи обнаружения в реальном масштабе времени спонтанного изменения вероятностных характеристик (разладки) временнóго ряда. Отмечен возрастающий интерес к разработке так называемых непараметрических методов обнаружения разладки, т. е. методов, не требующих для своего использования знания функции распределения вероятностей значений контролируемого процесса. Констатировано, что большинство из известных вариантов такого рода методов базируется на использовании ряда стандартных непараметрических критериев, трансформированных для решения задач обнаружения разладки.

Предложено в качестве базы для построения алгоритмов обнаружения разладки использовать критерии знаков, серий и критерий Рамачандрана–Ранганатана. Рассмотрены методические аспекты исследования статистических свойств и эффективности алгоритмов обнаружения разладки, построенных на их основе. В качестве инструмента использован метод имитационного моделирования. План проведения имитационных экспериментов разработан отдельно для каждого из предложенных алгоритмов с учетом их индивидуальных особенностей, но исходя из общего требования полноценного воспроизведения динамики работы контролирующего алгоритма в реальных условиях, когда разладка может появиться в любой момент, и существует переходной процесс в значениях решающей функции. С помощью имитационного эксперимента для каждого из рассматриваемых алгоритмов получены и систематизированы данные об их статистических характеристиках, достаточные для целей синтеза контролирующей процедуры с заданными свойствами.

Ключевые слова: разладка временнóго ряда, обнаружение разладки, непараметрические методы обнаружения, критерии знаков и серий, критерий Рамачандрана–Ранганатана, вероятностные характеристики алгоритмов обнаружения, синтез контролирующей процедуры.

Для цитирования: Филаретов Г.Ф., Бучаала З. Разработка алгоритмов обнаружения разладки временных рядов на основе непараметрических критериев проверки гипотез // Вестник МЭИ. 2021. № 3. С. 67—77. DOI: 10.24160/1993-6982-2021-3-67-77.

Development of Time Series Disorder Detection Algorithms Based on Nonparametric Hypothesis Testing Criteria

G.F. Filaretov, Z. Bouchaala

The solution of the problem of detecting, in the online mode, a spontaneous change in the probabilistic characteristics (“disorder” or “breakdown”) of a time series is given. It is pointed out that there is a growing interest in the development of so-called nonparametric disorder detection methods, i.e., methods the application of which does not require the knowledge of the probability distribution function of the controlled process values. It is stated that the majority of the known versions of such methods are based on using a number of standard nonparametric criteria transformed for solving disorder detection problems. It is proposed to use the signs criterion, the series criterion, and the Ramachandran–Ranganathan criterion as a basis for construction of disorder detection algorithms. The methodical aspects of studying the statistical properties and efficiency of the disorder detection algorithms built on their basis are considered. The simulation method was used as a study tool. The plan of carrying out simulation experiments was developed separately for each of the proposed algorithms, taking into account their individual characteristics, but based on the general requirement of fully reproducing the monitoring algorithm performance dynamics under real conditions, when a disorder can appear at any time and there is a transient in the values of the decisive function. By using a simulation experiment for each of the algorithms under consideration, data on their statistical characteristics were obtained and systematized in a scope sufficient for synthesizing a monitoring procedure with the specified properties.

Key words: time series disorder, disorder detection, nonparametric detection methods, signs criterion, series criterion, Ramachandran–Ranganathan criterion, detection algorithm probabilistic characteristics, synthesis of a monitoring procedure.

For citation: Filaretov G.F., Bouchaala Z. Development of Time Series Disorder Detection Algorithms Based on Nonparametric Hypothesis Testing Criteria. Bulletin of MPEI. 2021;3:67—77. (in Russian). DOI: 10.24160/1993-6982-2021-3-67-77.

Введение

Задача наискорейшего обнаружения спонтанного изменения свойств (характеристик) контролируемого объекта (процесса) — одна из важнейших. Ее решение возложено на современные автоматизированные системы мониторинга различного назначения (технологические, экологические, медицинские и т. п.).

В статистической трактовке данная задача обычно формулируется следующим образом: пусть x_1, x_2, …, x_i ... — значения наблюдаемого (контролируемого) процесса $X(t)$, зафиксированные в дискретные моменты времени. Предполагается, что существует некоторый заданный (базовый) режим нормальной работы объекта контроля (состояние объекта — «норма»), которому соответствует тот или иной набор вероятностных характеристик процесса $X(i)$. Тогда любое существенное (значимое) отклонение от указанного базового режима и связанного с этим событием изменения вероятностных характеристик $X(i)$ рассматривается как «разладка» процесса. Необходимо установить момент такого изменения (появления разладки) в реальном времени, т. е. в темпе с поступлением значений x_i.

Механизм работы такого рода алгоритмов обнаружения, зачастую называемых последовательными, состоит в рекуррентном вычислении при появлении каждого очередного значения x_i некоторой решающей функции $g_i = g_i(x_i, x_{i-1}, x_{i-2}, \ldots)$ с последующим сопоставлением g_i с порогом H. По результатам сопоставления принимается решение о состоянии объекта («норма» или «разладка»). Желательно, чтобы соответствующий алгоритм обнаружения обладал определенными оптимальными свойствами. Чаще всего в качестве показателя качества работы алгоритма используется среднее время запаздывания в обнаружении разладки $\overline{\tau}_{\text{зап}}$ при фиксированном значении среднего интервала между ложными тревогами $\overline{T}_{\text{лт}}$, когда контролирующий алгоритм подает сигнал о появлении разладки (сигнал тревоги), хотя в действительности объект остается в состоянии «норма».

Исторически первый алгоритм, который в принципе может быть отнесен к алгоритмам обнаружения разладки, был предложен в 1924 г. У. Шухартом [1] как инструмент анализа изменчивости (степени вариабельности) любых процессов. С тех пор интерес к данной тематике только усиливается [2, 3].

Следует отметить возросшее внимание к разработке и использованию так называемых непараметрических методов обнаружения разладки, т. е. методов, не требующих для своего использования знания функции распределения вероятностей значений контролируемого процесса. Это принципиально важно, если речь идет о построении систем автоматизированного мониторинга, когда зачастую исходная информация или малодостоверна, или вообще отсутствует, и тогда необходимо проведение достаточно трудоемкого предварительного исследования процесса, или есть основания считать, что вид функции распределения со временем может постепенно видоизменяться.

Достаточно полное представление о работах, посвященных непараметрическим методам, можно получить из обзорных материалов [4, 5] и монографии [6]. Как показывает анализ, большинство из анализируемых в них вариантов контролирующих алгоритмов базируется на использовании известных непараметрических критериев, изначально предназначенных для проверки гипотез об однородности вероятностных свойств наблюдаемого временного ряда, случайности его значений, наличия трендовой составляющей, трансформированных для решения задач обнаружения разладки в форме контрольных карт Шухарта, алгоритма кумулятивных сумм (АКС) [7] или EWMA-алгоритма [8]. Характерные примеры приведены в публикациях [9 — 18].

Несмотря на обилие имеющихся работ, отметим, что в качестве базы для построения алгоритмов обнаружения разладки рассмотрены далеко не все известные непараметрические критерии. Кроме того, изучение свойств непараметрических алгоритмов очень часто опирается на асимптотические распределения соответствующих решающих функций без учета их дискретного характера. Наконец, результаты многих исследований представлены в форме, затрудняющей решение задачи синтеза контролирующей процедуры с заданными характеристиками и сопоставление различных алгоритмов обнаружения разладки по их эффективности.

С этих позиций в качестве основных целей настоящей работы были выбраны:

● разработка и исследование характеристик оригинальных непараметрических алгоритмов обнаружения разладки на основе специально отобранных непараметрических критериев;

● анализ эффективности разработанных алгоритмов в сопоставлении друг с другом, а также с известными аналогичными по назначению параметрическими алгоритмами;

● получение данных, необходимых для решения задачи синтеза контролирующего алгоритма.

Постановка задачи

Рассмотрим временно́й ряд с некоррелированными отсчетами $x_1, x_2, \ldots, x_n$ и некоторой непрерывной функцией плотности распределения вероятностей $f_X(x;\ \theta)$, где θ — параметр местоположения (сдвига) или параметр рассеяния (масштаба) распределения. Пусть состоянию «норма» соответствует значение $\theta = \theta_0$, а разладка состоит в скачкообразном изменении этого параметра до $\theta = \theta_1$.

В основе последующего построения новых алгоритмов обнаружения положена та же идея использования тех или иных известных непараметрических критериев, общее количество которых весьма велико [19]. Однако, когда речь идет об использовании различных непараметрических критериев, предназначенных для выявления наличия некоторого тренда или неоднородности экспериментальных данных, в качестве основы построения соответствующих непараметрических алгоритмов обнаружения разладки в реальном времени следует установить их пригодность для такого применения. С этой целью проверялось соответствие критериев следующим двум требованиям:

● критерий должен быть чувствителен к изменению параметра местоположения или масштаба стохастической компоненты наблюдаемого процесса в реальном времени, т. е. в ритме с поступлением данных без какого-либо предварительного их группирования;

● использование критерия не должно требовать запоминания некоторого эталонного отрезка реализации процесса, соответствующего состоянию «норма» (без разладки), поскольку формирование такого эталона и его возможного текущего обновления весьма сложно формализовать.

Проверка на соответствие первому требованию сразу позволяет исключить из числа претендентов, в частности, такие популярные критерии, как критерий восходящих-нисходящих серий, критерий инверсий, другие ранговые критерии. Действительно, если вид функции распределения вероятностей стохастической компоненты не меняется (например, гауссовское распределение), то при скачкоообразном изменении ее математического ожидания или среднеквадратического отклонения как параметра масштаба, ни число восходящих-нисходящих серий, ни число инверсий не изменится, кроме краткого участка переходного процесса, в течение которого наличие разладки может и не обнаружиться. Второе требование сразу исключает из числа претендентов так называемые двухвыборочные критерии [20].

В качестве базы для построения алгоритмов обнаружения разладки далее будут использованы только сериальные критерии, а именно: критерии знаков, серий и Рамачандрана–Ранганатана. Именно алгоритмы обнаружения, построенные на их основе, являются предметом последующего изучения.

Методические аспекты проведения исследований

Предлагаемые алгоритмы, несмотря на различия в используемых базовых критериях, строятся с использованием единого подхода и имеют много общих черт.

1. Все алгоритмы относятся к категории последовательных непараметрических алгоритмов, базирующихся на медианах [19]. Это означает, что своего рода опорной точкой, характеризующей состояние «норма» контролируемого процесса $X(i)$, является медиана, т. е. $P\{X(i) < \mathrm{Me}(X)\} = P\{X(i) > \mathrm{Me}(X)\} = p_0 = 1/2$, где $\mathrm{Me}(X)$ — медиана наблюдаемого временного ряда $X(i)$; $P\{U\}$ — вероятность появления некоторого случайного события U. Тогда задача обнаружения разладки может трактоваться как задача обнаружения изменения вероятности $P\{X(i)\}$, причем состоянию «разладка» будет отвечать значение $P\{z_i\} = p_1 \neq 1/2$. Предположим, что разладка ведет к увеличению вероятности $P\{z_i\}$, т. е. $p_1 > 1/2$, что, конечно, непринципиально.

При исследовании свойств рассматриваемых алгоритмов обнаружения разладки значения p_1 выбирают достаточно произвольно. Если иметь в виду последующее сопоставление их эффективности с некоторыми параметрическими алгоритмами, то к такому выбору приходится подходить более ответственно. Используем следующие значения p_1: 0,692; 0,841; 0,933. Они соответствуют разладке по математическому ожиданию m_X гауссовского временно́го ряда $X(t)$ на величину δ_m, равную 0,5; 1,0; 1,5 соответственно, где $\delta_m = \dfrac{|m_1 - m_0|}{\sigma_X}$; m_0, m_1 — значения m_X до и при появлении разладки; σ_X — среднеквадратическое значение процесса $X(t)$.

Данный выбор позволяет сравнить эффективность предлагаемых алгоритмов, например, с классическим АКС для хорошо изученного случая обнаружения разладки гауссовского процесса по математическому ожиданию.

2. Трансформация исходных непараметрических критериев в алгоритмы обнаружения разладки выполнена в форме алгоритма, близкого по своей сути к алгоритму скользящего среднего (Moving Average или MA-алгоритма), ранее практически никогда не применявшегося в задачах обнаружения разладки. Идея указанного алгоритма состоит в фиксации в скользящем

окне (стеке) длиной N для каждого текущего момента времени i значений $x_i, x_{i-1}, x_{i-2}, \ldots, x_{i-N+1}$. Все последующие манипуляции основаны на этих исходных данных, содержащих информацию о текущем состоянии контролируемого объекта и возможном появлении разладки. Каждое новое наблюдение включается в обработку в момент его появления, а самое раннее — исключается из нее, как это всегда делается в скользящем окне или стеке.

3. В качестве основного метода исследования алгоритмов выбран метод имитационного моделирования, поскольку при изучении непараметрических алгоритмов функция распределения случайных величин $X(i)$ может быть любой из класса непрерывных. В качестве источника значений использован стандартный генератор независимых случайных чисел, равномерно распределенных в интервале [0 — 1]. С его помощью легко получить бинарные последовательности $Z(i)$, состоящие из значений +1 и 0 с заданной вероятностью p появления +1 в соответствии с соотношением:

$$z_i = \begin{cases} +1, & x_i \geq p \\ 0, & x_i < p \end{cases}; \qquad i = 1, 2, \ldots. \tag{1}$$

Последующие вычисления, включая определение значения решающей функции g_i, осуществляются с использованием набора такого рода данных, содержащегося в стеке на текущем такте i, причем вычисления идут по формулам, специфичным для каждого из рассматриваемых алгоритмов. Можно только уточнить, что решающие функции g_i всегда носят целочисленный характер. Точно также целочисленными будут и значения решающего порога H.

План имитационного эксперимента предусматривает подразделение процесса исследования каждого из предложенных алгоритмов на два этапа: предварительный и основной.

На первом предварительном этапе исследуются вероятностные свойства решающих функций g_i ($i = 1, 2, \ldots, 10000$): определяются их числовые параметры и строятся гистограммы для различных значений вероятностей p_1 и p_0. Данная информация необходима для выбора подходящих значений решающего порога H в рамках реализации основного этапа.

На втором (основном) этапе имитируется собственно работа алгоритма обнаружения разладки для различных N и выбранных величин H и $P\{z_i\}$. Длина стека N варьируется в широких пределах. Используются следующие варианты значений N: 8; 12; 16; 20; 24; 28. Для каждого значения N выбор порогов H производится с учетом желательности получения значений $\overline{T}_{\text{лт}}$ из наиболее востребованного на практике диапазона от 50 до 5000. Количество повторных запусков контролирующего алгоритма (число опытов) L взято равным 10000, что обеспечивает приемлемую точность результатов имитационного эксперимента. В ходе обработки экспериментальных данных оцениваются значения $\overline{T}_{\text{лт}}$, запаздывание $\overline{\tau}_{\text{зап}}$, показатель эффективности алгоритма обнаружения $E = \overline{T}_{\text{лт}} / \overline{\tau}_{\text{зап}}$, а также некоторые другие характеристики, представляющие интерес для пользователей. Далее приводятся основные результаты исследований каждого из трех анализируемых алгоритмов

Исследование алгоритма обнаружения разладки на основе критерия знаков

Критерий знаков или критерий знаков относительно медианы — один из наиболее известных и хорошо изученных непараметрических критериев для выявления возможной неоднородности данных [20]. Однако в качестве базы для построения алгоритма обнаружения разладки он детальным образом не исследован.

Для данного алгоритма стек всегда заполняется непосредственно значениями z_i, $i = 1, 2, \ldots, N$. Решающая функция g_i на произвольном такте i в данном случае равна:

$$g_i = z_i + z_{i-1} + z_{i-2} + \ldots + z_{i-N+1}.$$

Очевидно, что при $p_0 = 1/2$ количество значений +1 и 0 в стеке длиной N, учитывая соотношение (1), должно быть примерно одинаковым, а при $p_1 > 1/2$ значений +1 должно быть больше, чем нулей, т. е. критерий носит «правосторонний» характер.

При его использовании нет необходимости в проведении этапа предварительных исследований, поскольку распределение решающей функции g хорошо известно. Это биномиальное распределение

$$f(g; p, N) = C_N^g p^g (1-p)^{N-g};$$
$$g = 0, 1, 2, \ldots, N; \quad 0 \leq p \leq 1,$$

ординаты которого легко рассчитываются аналитически.

На основном этапе реализовывалась программа имитационного эксперимента, когда для выбранных значений длины стека N, порога H и $P\{z_i\}$ определялись оценки значений $\overline{T}_{\text{лт}}$ и $\overline{\tau}_{\text{зап}}$, дисперсии $\sigma^2\{T_{\text{лт}}\}$, $\sigma^2\{\tau_{\text{зап}}\}$, а также СКО $\sigma\{\overline{T}_{\text{лт}}\}$, $\sigma\{\overline{\tau}_{\text{зап}}\}$ как показатели точности оценивания.

Основные результаты исследования свойств алгоритма, полученные с помощью имитационного эксперимента, приведены в табл. 1, 2. Для упорядочения вариантов по значениям $\overline{T}_{\text{лт}}$ в табл. 1, 2 введены строки с идентификаторами вариантов Id.

Данные, приведенные в табл. 1, 2, свидетельствуют о том, что эффективность контролирующих алгоритмов увеличивается с ростом N и вероятности p_1.

При решении задачи синтеза контролирующей процедуры пользователь, в первую очередь, задается требуемым значением $\overline{T}_{\text{лт}}$. Однако, в отличие от параметрических методов, когда это вполне возможно для произвольных $\overline{T}_{\text{лт}}$, в непараметрическом случае величины $\overline{T}_{\text{лт}}$ могут выбираться только из вполне определенного дискретного набора. С целью облегчения

Таблица 1

Критерий знаков. Оценки $\overline{T}_{лт}$, $\overline{\tau}_{зап}$ и эффективности E для различных N и H (N = 8...20)

N	8		12			16				20			
H	8	7	11	10	9	15	14	13	12	18	17	16	15
Id	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	*m*
$\overline{T}_{лт}$	510	80	813	178	64	4137	1470	370	132	4440	2412	705	260
$\overline{\tau}_{(1)}$	63	23	70	36	24	211	83	48	35	206	97	62	46
$\overline{\tau}_{(2)}$	24	15	30	21	19	49	35	29	26	51	42	37	33
$\overline{\tau}_{(3)}$	17	13	23	20	18	33	29	26	24	38	35	32	30
E_1	8,1	3,5	12	4,9	2,7	20	18	7,8	3,8	22	25	11	5,7
E_2	21	5,3	27	8,5	3,4	84	42	13	5,1	87	57	19	7,9
E_3	30	6,2	35	8,9	3,6	125	51	14	5,5	117	69	22	8,7

Таблица 2

Критерий знаков. Оценки $\overline{T}_{лт}$, $\overline{\tau}_{зап}$ и эффективности E для различных N и H (N = 24; 28)

N	24					28						
H	21	20	19	18	17	24	23	22	21	20	19	18
Id	*n*	*p*	*q*	*r*	*s*	*t*	*u*	*v*	*w*	*x*	*y*	*z*
$\overline{T}_{лт}$	4688	3339	1349	485	215	4751	4076	2302	910	391	194	144
$\overline{\tau}_{(1)}$	215	116	76	58	47	233	134	92	69	58	50	44
$\overline{\tau}_{(2)}$	56	48	43	40	37	62	55	50	47	43	41	38
$\overline{\tau}_{(3)}$	44	41	39	37	34	50	48	45	43	41	38	36
E_1	22	29	18	8,4	4,6	20	30	22	13	6,7	3,9	3,3
E_2	84	40	31	12	5,8	77	74	41	19	9,1	4,7	3,8
E_3	107	81	35	13	6,3	99	85	45	21	9,5	5,1	4,0

выбора в табл. 3 значения $\overline{T}_{лт}$ упорядочены в порядке возрастания.

Идентификаторы *Id* позволяют легко определить эффективность выбранного варианта алгоритма, используя данные табл. 1 или 2.

Исследование алгоритма обнаружения разладки на основе критерия серий

Критерий серий или, точнее, критерий серий, основанный на медианах, базируется на очень простой идее. Если рассмотреть случайную последовательность значений +1 и 0, то ее некоторые свойства можно охарактеризовать с помощью такого статистического показателя, как общее число серий $\nu(N)$, фиксируемых на интервале наблюдения N. Под «серией» понимается последовательность подряд идущих значений +1 или подряд идущих нулей. В частном случае серия может состоять только из одного значения +1 или 0.

Очевидно, что если анализируемая последовательность состоит из статистически независимых бинарных наблюдений с равновероятным появлением значений +1 и 0 ($p_0 = 1/2$), то в ней не должно быть слишком длинных серий и, следовательно, число серий ν(N) в них не должно быть слишком малым. Наоборот, если вероятность появления значений +1 больше 0,5 ($p_1 > 1/2$), то можно ожидать рождения достаточно длинных серий и уменьшения их числа $\nu(N)$, что может послужить признаком отклонения от $p_0 = 1/2$. Критерий в данном случае носит «левосторонний» характер.

Таблица 3

Критерий знаков. Упорядоченные значения $\overline{T}_{лт}$

$\overline{T}_{лт}$	64	80	132	144	178	194	215	260	370	391	485	510	705
Id	*e*	*b*	*i*	*z*	*d*	*y*	*s*	*m*	*h*	*x*	*r*	*a*	*l*
$\overline{T}_{лт}$	813	910	1349	1470	2302	2412	3339	4076	4137	4440	4688	4751	—
Id	*c*	*w*	*q*	*g*	*v*	*k*	*p*	*u*	*f*	*j*	*n*	*t*	—

При первоначальном заполнении стека на первых N тактах использовано следующее правило получения промежуточных значений s_i:

$$S_1 = 1; \quad S_i = \begin{cases} 1, & \text{если } z_{i-1} \neq z_i; \\ 0, & \text{если } z_{i-1} = z_i, \end{cases} \quad i = 2, 3, ..., N,$$

где z_i, $i = 1, 2, \ldots, N$ определяется по (1) для $p = p_0 = 1/2$.

Значение решающей функции g_N на момент первоначального заполнения стека выглядит как сумма значений в стеке:

$$g_N = s_1 + s_2 + \ldots + s_N.$$

Содержание стека на каждом следующем i-м такте ($i = N + 1$, $N + 2$, ….) обновляется с использованием очередного значения z_i, сформированного в соответствии с (1) для установленного значения вероятности p, по правилу:

$$S_i = S_i^{(1)} + S_i^{(2)},$$

где $S_i^{(1)} = \begin{cases} 0, & \text{если } z_i = z_{i-1}; \\ 1, & \text{если } z_i \neq z_{i-1}; \end{cases}$

$$S_i^{(2)} = \begin{cases} 0, & \text{если } z_{i-N} = z_{i-N+1}; \\ -1, & \text{если } z_{i-N} \neq z_{i-N+1}, \end{cases} \quad i = N+1, N+2, \ldots.$$

Значения решающей функции g_i рассчитываются с помощью рекуррентного соотношения:

$$g_i = g_{i-1} + S_i; \; i = N + 1, N + 2, \ldots.$$

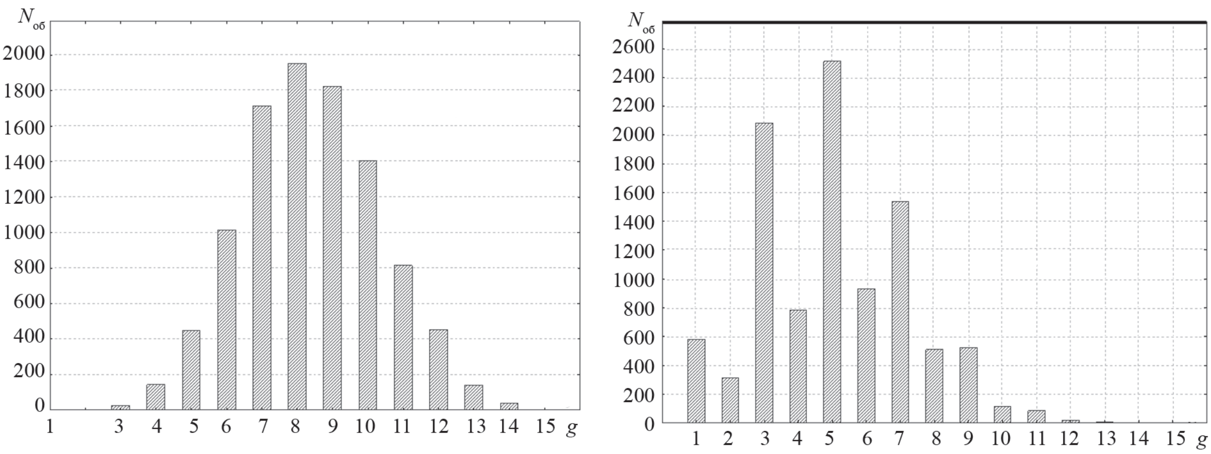
Описанный алгоритм обеспечивает вычисление числа серий $\nu(N)$ значений z_i, содержащихся в стеке, на каждом текущем такте i, т. е. $g_i = \nu_i(N)$.

На предварительном этапе исследований критерия для изучения особенностей функций распределения g_i при различных N и $P\{z_i\}$ построены их гистограммы. В качестве примера на рис. 1 представлены два образца гистограмм для случаев, когда $N = 16$, $p_0 = 0{,}5$, $p_1 = 0{,}842$.

Результаты основного этапа исследования данного алгоритма отображены в табл. 4, 5.

Анализ данных табл. 4, 5 позволяет прийти к выводу, аналогичному сформулированному для предыдущего алгоритма: эффективность, в целом, увеличивается с ростом N и вероятности p_1.

В таблице 6 даны значения $\overline{T}_{лт}$, упорядоченные в порядке возрастания, вместе с идентификаторами *Id*.



а *б*

Рис. 1. Гистограммы функций распределения значений $g_i = \nu(N)$:
а — $p_0 = 0{,}5$; *б* — $p_1 = 0{,}842$

Таблица 4

Критерий серий. Оценки $\overline{T}_{лт}$, $\overline{\tau}_{зап}$ и эффективности E для различных N и H ($N = 8...20$)

N	8		12			16				20			
H	1	2	1	2	3	2	3	4	5	3	4	5	6
Id	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	*m*
$\overline{T}_{лт}$	260	49	3135	455	113	3466	855	231	92	4035	1546	455	179
$\overline{\tau}_{(1)}$	60	29	271	129	43	565	136	81	41	442	251	98	68
$\overline{\tau}_{(2)}$	20	15	46	35	19	75	33	29	21	55	47	30	28
$\overline{\tau}_{(3)}$	12	11	20	18	14	28	20	20	17	26	25	22	22
E_1	4,3	1,7	12	3,5	2,6	6,1	6,3	2,9	2,2	9,1	6,2	4,6	2,6
E_2	13	3,3	120	13	5,9	46	26	8,0	4,4	73	33	15	6,4
E_3	22	4,5	158	25	8,1	124	43	12	5,4	155	62	21	8,1

Таблица 5

Критерий серий. Оценки $\overline{T}_{\text{лт}}$, $\overline{\tau}_{\text{зап}}$ и эффективности E для различных N и H (N = 24; 28)

N	24					28						
H	4	5	6	7	8	5	6	7	8	9	10	11
Id	*n*	*p*	*q*	*r*	*s*	*t*	*u*	*v*	*w*	*x*	*y*	*z*
$\overline{T}_{\text{лт}}$	4377	2536	869	337	158	4673	3532	1577	617	280	152	92
$\overline{\tau}_{(1)}$	838	261	170	88	64	755	459	200	132	81	62	46
$\overline{\tau}_{(2)}$	79	45	40	31	29	68	60	42	39	33	32	30
$\overline{\tau}_{(3)}$	32	27	26	25	25	32	32	30	30	29	29	29
E_1	5,2	9,7	5,1	3,8	2,5	6,2	7,7	7,9	4,7	3,5	2,5	2,0
E_2	55	56	22	11	5,4	68	59	38	16	8,5	4,8	3,1
E_3	136	94	33	13	6,3	146	110	53	21	9,7	5,2	3,2

Таблица 6

Критерий серий. Упорядоченные значения $\overline{T}_{\text{лт}}$

$\overline{T}_{\text{лт}}$	49	92	92	113	152	158	179	231	260	280	337	455	455
Id	*b*	*i*	*z*	*e*	*y*	*s*	*m*	*h*	*a*	*x*	*r*	*d*	*l*
$\overline{T}_{\text{лт}}$	617	855	869	1546	1577	2536	3135	3466	3532	4035	4377	4673	—
Id	*w*	*g*	*q*	*k*	*v*	*p*	*c*	*f*	*u*	*j*	*n*	*t*	—

Исследование алгоритма обнаружения разладки на основе критерия Рамачандрана–Ранганатана (*RR*-критерия)

Данный непараметрический критерий — разновидность критерия серий, но, в отличие от него, в решающей функции учитывается не только количество, но и длины серий [19].

Его статистика имеет следующий вид:

$$RR = \sum_j j^2 n_j,$$

где j — длина серии, состоящей из значений +1; n_j – количество серий длины j на интервале наблюдения длительностью N.

Указанный критерий «правосторонний»: гипотеза о наличии разладки, связанной с увеличением вероятности $P\{z_i\}$, принимается при больших значениях статистики RR.

Стек заполняется как и раньше, значениями z_i, i = = 1, 2, …, т. е. значениями +1 или 0, в соответствии с вероятностью $P\{z_i\}$, а расчет решающей функции алгоритма обнаружения g_i осуществляется с использованием этих данных на каждом i-м шаге контролирующей процедуры. В отличие от случая критерия серий здесь не удается предложить какую-либо итерационную процедуру, а после каждого сдвига информации в стеке, т. е. на каждой очередной итерации, g_i нужно вычислять заново, что и реализуется с помощью специально разработанной для этих целей подпрограммы расчета *RR*-статистики.

Первоначально стек заполняется z_i, i = 1, 2, …, N с учетом условия $P\{z_i\} = p_0 = 1/2$, после чего находится величина g_N, являющаяся начальным значением g_i для основного этапа исследования.

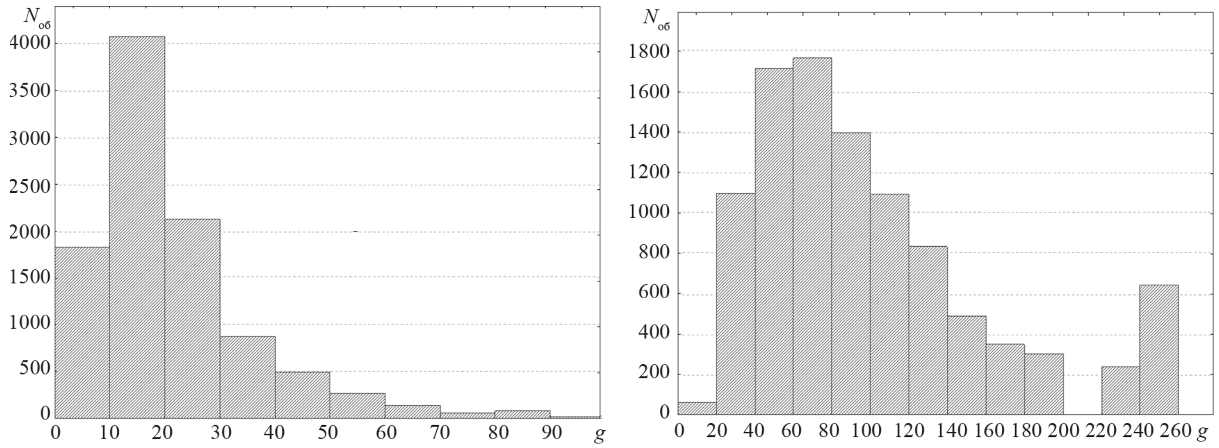
Алгоритм обнаружения разладки на базе указанного критерия, по всей видимости, изучается впервые. Поэтому важно предварительно оценить статистические характеристики g_i, что и производилось на предварительном этапе для различных значений N и $P\{z_i\}$. Типичные результаты для N = 16 приведены в табл. 7 и на рис. 2.

Гистограммы в данном случае имеют вид, характерный для непрерывной случайной величины, что объясняется большим разнообразием значений решающей функции g_i в диапазоне ее изменения. Данное обстоятельство позволяет (для фиксированного N) систематически (с постоянным шагом) менять порог H и обеспечивать вариацию $\overline{T}_{\text{лт}}$ в достаточно широких пределах. Это наглядно видно из результатов основного этапа исследования *RR*-алгоритма для N = 16 и N = 24, представленных в табл. 8.

Таблица 7

***RR*-критерий. Статистические характеристики g_i**

$P\{z_i\}$	$\hat{m}\{g_i\}$	$\hat{\sigma}\{g_i\}$	$\min g_i$	$\max g_i$
$p_0 = 0{,}5$; $\overline{T}_{\text{лт}}$	22	15,3	1	116
$p_1 = 0{,}692$; $\overline{\tau}_{(1)}$	52	33,5	5	256
$p_1 = 0{,}841$; $\overline{\tau}_{(2)}$	101	61,3	10	256
$p_1 = 0{,}933$; $\overline{\tau}_{(3)}$	172	70,6	12	256



а *б*

Рис. 2. Гистограммы функций распределения значений g_i (*RR*-алгоритм)

а — $p_0 = 0{,}5$; *б* — $p_1 = 0{,}841$

Таблица 8

RR-критерий. Оценки $\overline{T}_{лт}$, $\overline{\tau}_{зап}$ и эффективности E для различных N и H

N	16							24						
H	50	60	70	80	90	100	110	100	110	120	130	140	150	160
Id	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	*m*	*n*
$\overline{T}_{лт}$	127	257	457	725	1174	1691	2502	870	1270	1908	2512	3122	3514	3986
$\overline{\tau}_{(1)}$	18	28	39	55	71	93	117	51	65	82	102	127	150	183
$\overline{\tau}_{(2)}$	6,6	8,6	11	13	16	19	21	14	15	18	20,2	22,5	25,3	28,2
$\overline{\tau}_{(3)}$	3,1	3,8	4,5	5,1	5,9	6,6	7,3	3,7	6,2	7,0	7,7	8,44	9,05	9,74
E_1	7,0	9,2	12	13	17	18	21	17	20	23	25	24,6	23,4	21,8
E_2	19	30	43	56	73	89	119	62	85	106	124	139	139	141
E_3	41	68	101	142	199	256	343	235	205	272	326	370	388	409

Очевидно, что отмеченная ранее тенденция увеличения эффективности контролирующего алгоритма с ростом N и вероятности p_1 сохраняется и для *RR*-критерия. Кроме того, оказывается возможным путем обработки данных табл. 8 определить эмпирические зависимости, связывающие значения $\overline{T}_{лт}$ и порога H. Обе зависимости носят линейный характер, но для $N = 16$ в двойном логарифмическом масштабе, а для $N = 24$ — в линейном:

$$N = 16{:}\ \ln H = 2{,}62 + 3{,}83 \ln \overline{T}_{лт};$$

$$N = 24{:}\ H = 84{,}6 + 0{,}0185 \overline{T}_{лт}.$$

Полученные зависимости существенно помогут в решении задачи синтеза контролирующей процедуры с заданными характеристиками.

Сопоставительный анализ алгоритмов

Отсутствие регулярной сетки значений $\overline{T}_{лт}$ существенно затрудняет непосредственное сопоставление эффективности рассматриваемых непараметрических алгоритмов. Поэтому предлагается иной способ решения данной задачи, а именно — путем сравнения для каждого конкретного значения $\overline{T}_{лт}$ эффективностей данного варианта непараметрического алгоритма $E_Н$ и параметрического алгоритма типа АКС, предназначенного для обнаружения разладки гауссовского процесса по математическому ожиданию $E_Г$, при одинаковых значениях величины разладки δ. Тем самым будет решена и другая, быть может, более важная задача сопоставления эффективности параметрического и непараметрического подходов.

Возможность выполнения указанного способа сопоставления связана с использованием программной системы STATCONT [21], предназначенной для анализа гауссовских процессов и позволяющей находить эффективность АКС для произвольных значений $\overline{T}_{лт}$.

Приведем полученные характерные результаты сопоставления алгоритмов. В качестве основного показателя использовано значение относительной эффективности непараметрического алгоритма по сравнению с АКС для гауссовского процесса при прочих равных условиях $Є = E_Н/E_Г$, %.

Алгоритм обнаружения разладки на основе критерия знаков.

Основные результаты сопоставления для варианта $\delta = 0{,}5$ ($p_1 = 0{,}692$) даны в табл. 9. При ее заполнении использованы данные из табл. 1 — 3.

Таблица 9

Относительная эффективность алгоритма обнаружения на основе критерия знаков Є·= ($E_н/E_г$)100% для δ = 0,5 (p_1 = 0,692)

$\bar{T}_{лт}$	132	144	178	194	215	260	370	391	485	510	705	813
Є, %	48	39	51	39	42	46	50	41	44	42	44	44
Id	*i*	*z*	*d*	*y*	*s*	*m*	*h*	*x*	*r*	*a*	*l*	*c*
$\bar{T}_{лт}$	910	1349	1470	2302	2412	3339	4076	4137	4440	4688	4751	—
Є, %	43	40	42	36	39	35	31	21	21	20	19	—
Id	*w*	*q*	*g*	*v*	*k*	*p*	*u*	*f*	*j*	*n*	*t*	—

Как следует из представленных табличных значений, эффективность непараметрического алгоритма с ростом $\bar{T}_{лт}$ постепенно снижается от 48% при $\bar{T}_{лт}$ = 132 до 19% при $\bar{T}_{лт}$ = 4751. Средняя величина показателя Є, установленная по всем приведенным в табл. 9 значениям, оказалась равна $\bar{Є}$ ≈ 37%. Аналогичные расчеты сделаны и для случаев δ = 1,0 (p_1 = 0,841), δ = 1,5 (p_1 = 0,933). Для этих вариантов отмеченная тенденция снижения Є с ростом $\bar{T}_{лт}$ выражена очень слабо, а средние значения Є соответственно составляют $\bar{Є}$ ≈ 25 и 15%.

Алгоритм обнаружения разладки на основе критерия серий.

Основные результаты сопоставления для варианта δ = 0,5 (p_1 = 0,692) даны в табл.10. При ее заполнении использованы значения из табл. 4 — 6.

В данном случае прослеживается тенденция уменьшения Є с ростом $\bar{T}_{лт}$, в особенности четко выраженная при $\bar{T}_{лт}$, бóльших 3000. Средняя величина $\bar{Є}$ показателя Є составляет примерно 20%. Для δ = 1,0 (p_1 = 0,841) и δ = 1,5 (p_1 = 0,933) значения Є практически не меняются с ростом $\bar{T}_{лт}$, при этом средние значения Є равны соответственно $\bar{Є}$ ≈ 25 и 22%. Получается, что алгоритм на основе критерия серий уступает по эффективности алгоритму на основе критерия знаков при малых величинах разладки, но превосходит его при больших.

Алгоритм обнаружения разладки на базе критерия Рамачандрана–Ранганатана.

Для данного алгоритма возможно более полное систематическое сравнение его эффективности с алгоритмом кумулятивных сумм. В таблице 11 продемонстрированы результаты такого сопоставления для *N* = 16 и 24 и различных *H*.

Очевидно, что эффективность указанного алгоритма существенно выше, чем у представленных ранее на основе критериев знаков и серий. Более того, как следует из данных табл. 11, при малых значениях $\bar{T}_{лт}$ и больших разладках он более эффективен, чем классический АКС, что весьма неожиданно.

Заключение

Изучена проблема построения непараметрических алгоритмов обнаружения спонтанного изменения вероятностных характеристик (разладки) временнóго ряда в реальном масштабе времени. Отмечено, что большинство известных вариантов такого рода алгоритмов основано на использовании некоторых стандартных непараметрических критериев, трансформированных для решения задач обнаружения разладки.

Предложено в качестве базы для создания указанных алгоритмов использовать критерии знаков, серий и Рамачандрана–Ранганатана.

Рассмотрены методические аспекты исследования статистических свойств и эффективности алгоритмов обнаружения разладки, построенных на их основе, вопросы планирования имитационных экспериментов как инструмента проведения исследований.

С помощью имитационного эксперимента для каждого из рассматриваемых алгоритмов получены и систематизированы данные об их статистических характеристиках, достаточные для целей синтеза контролирующей процедуры с заданными свойствами.

Проведено сопоставление эффективности предложенных алгоритмов. Доказано, что наибольшей эффективностью обладает алгоритм обнаружения разладки на базе непараметрического критерия Рамачандрана–Ранганатана.

Таблица 10

Относительная эффективность алгоритма обнаружения на основе критерия серий Є·= ($E_н/E_г$)100% для δ = 0,5 (p_1 = 0,692)

$\bar{T}_{лт}$	113	152	158	179	231	260	280	337	455	455	617
Є, %	36	29	28	27	25	35	28	26	20	25	21
Id	*e*	*y*	*s*	*m*	*h*	*a*	*x*	*r*	*d*	*l*	*w*
$\bar{T}_{лт}$	855	869	1546	1577	2536	3135	3466	3532	4035	4377	4673
Є, %	22	18	14	18	15	15	7,2	9,0	9,5	5,7	6,2
Id	*g*	*q*	*k*	*v*	*p*	*c*	*f*	*u*	*j*	*n*	*t*

Таблица 11

Относительная эффективность алгоритма обнаружения на основе непараметрического критерия Рамачандрана–Ранганатана Є·= ($E_н/E_г$)100%

***N* = 16**							
$\overline{T}_{лт}$	127	257	457	725	1174	1691	2502
Id	*a*	*b*	*c*	*d*	*e*	*f*	*g*
Є, %, p_1 = 0,692	91	76	65	52	46	38	33
Є, %, p_1 = 0,841	99	91	82	76	67	61	59
Є, %, p_1 = 0,933	117	112	105	101	95	90	86
***N* = 24**							
$\overline{T}_{лт}$	870	1270	1908	2512	3122	3513	3986
Id	*h*	*i*	*j*	*k*	*l*	*m*	*n*
Є, %, p_1 = 0,692	59	51	44	38	32	28	23
Є, %, p_1 = 0,841	72,8	73,3	65,5	60,9	56,9	51,4	47,0
Є, %, p_1 = 0,933	144	91	86	81	76	72	68

Литература

1. **Shewhart W.A.** Quality Control Charts // Bell Syst. Techn. J. 1926. V. 5(4). Pp. 593—603.

2. **Адлер Ю.П., Максимова О.В., Шпер В.Л.** Контрольные карты Шухарта в России и за рубежом: краткий обзор современного состояния (статистические аспекты) // Стандарты и качество. 2011. № 7. С. 82—87; № 8. С. 82—87.

3. **Shafid A.** Bibliometric Analysis of EWMA and CUSUM Control Chart Schemes // ITEE J. 2018. V. 7(2). Pp. 1—11.

4. **Chakraborti S., Van der Laan P., Bakir S.T.** Nonparametric Statistical Process Control: an Overview and Some Results // J. Quality Technol. 2001. V. 33(3). Pp. 304—315.

5. **Bakir S.** Classification Distribution-free Quality Control Charts // Proc. Annual Meeting of the American Statistical Association. 2001. V. 5(9). Pp. 1—7.

6. **Chakraborti S., Graham M.A.** Nonparametric Statistical Process Control. N.-Y.: John Wiley&Sons, 2019.

7. **Page E.S.** Continuous Inspection Schemes // Biometrika. 1954. V. 41. No. 1. Pp. 100—115.

8. **Roberts, S.W.** Control Chart Tests Based on Geometric Moving Averages // Technometrics. 1959. V. 1(3). Pp. 239—250.

9. **Chakraborti S., Van der Laan P., Van de Wiel M.A.** A Class of Distribution-free Control Charts // J. Royal Statistical Soc.: Series C: Appl. Stat. 2004. V. 53(3). Pp. 443—462.

10. **Bakir S.T., Reynolds Jr.M.R.** A Nonparametric Procedure for Process Control Based on Within-group Ranking // Technometrics. 1979. V. 21(2). Pp. 175—183.

11. **Amin R.W., Reynolds Jr.M.R., Bakir S.T.** Nonparametric Quality Control Charts Based on the Sign Statistic // Communications in Statistics: Theory and Methods. 1995. V. 24(6). Pp. 1597—1623.

12. **Janacek G.J., Meikle S.E.** Control Charts Based on Medians // J. Royal Statistical Soc.: Series D: Statistician. 1997. V. 46(1). Pp. 19—31.

References

1. **Shewhart W.A.** Quality Control Charts. Bell Syst. Techn. J. 1926;5(4):593—603.

2. **Adler Yu.P., Maksimova O.V., Shper V.L.** Kontrol'nye Karty Shukharta v Rossii i za Rubezhom: Kratkiy Obzor Sovremennogo Sostoyaniya (Statisticheskie Aspekty). Standarty i kachestvo. 2011;7:82—87; 8:82—87. (in Russian).

3. **Shafid A.** Bibliometric Analysis of EWMA and CUSUM Control Chart Schemes. ITEE J. 2018;7(2): 1—11.

4. **Chakraborti S., Van der Laan P., Bakir S.T.** Nonparametric Statistical Process Control: an Overview and Some Results. J. Quality Technol. 2001;33(3):304—315.

5. **Bakir S.** Classification Distribution-free Quality Control Charts. Proc. Annual Meeting of the American Statistical Association. 2001;5(9):1—7.

6. **Chakraborti S., Graham M.A.** Nonparametric Statistical Process Control. N.-Y.: John Wiley&Sons, 2019.

7. **Page E.S.** Continuous Inspection Schemes. Biometrika. 1954;41;1:100—115.

8. **Roberts, S.W.** Control Chart Tests Based on Geometric Moving Averages. Technometrics. 1959;1(3): 239—250.

9. **Chakraborti S., Van der Laan P., Van de Wiel M.A.** A Slass of Distribution-free Control Charts. J. Royal Statistical Soc.: Series C: Appl. Stat. 2004;53(3): 443—462.

10. **Bakir S.T., Reynolds Jr.M.R.** A Nonparametric Procedure for Process Control Based on Within-group Ranking. Technometrics. 1979;21(2):175—183.

11. **Amin R.W., Reynolds Jr.M.R., Bakir S.T.** Nonparametric Quality Control Charts Based on the Sign Statistic. Communications in Statistics: Theory and Methods. 1995;24(6):1597—1623.

12. **Janacek G.J., Meikle S.E.** Control Charts Based on Medians. J. Royal Statistical Soc.: Series D: Statistician. 1997;46(1):19—31.

13. **McDonald D.** A CUSUM Procedure Based On Sequential Ranks // Naval Research Logistics. 1999. V. 37. Pp. 627—646.

14. **Bakir S.T.** A Distribution-free Shewhart Quality Control Chart Based on Signed-ranks // Quality Eng. 2004. V. 16(4). Pp. 613—623.

15. **Митрохин И.Н., Орлов А.И.** Обнаружение разладки с помощью контрольных карт // Заводская лаборатория. Диагностика материалов. 2007. № 5. С. 74—78.

16. **Chakraborti S.б Van de Wiel M.A.** A Nonparametric Control Chart Based on the Mann–Whitney Statistic // IMS Collections. 2008. V. 1. Pp. 156—172.

17. **Кузнецов Л.А., Журавлева М.Г.** Построение карт контроля качества с помощью непараметрического критерия Вилкоксона–Манна–Уитни // Заводская лаборатория. Диагностика материалов. 2009. № 1. С. 70—95.

18. **Human S.W., Chakraborti S., Smit C.F.** Nonparametric Shewhart-type Sign Control Charts Based on Runs // Communications in Statistics: Theory and Methods. 2010. V. 39(11). Pp. 2046—2062.

19. **Веретельникова И.В.** Исследование и применение критериев проверки гипотез об отсутствии тренда и критериев однородности: автореферат … дис. канд. техн. наук. Новосибирск: Типография НГТУ, 2019.

20. **Кендалл М., Стьюарт А.** Статистические выводы и связи. М.: Наука, 1973.

21. **Филаретов Г.Ф.** Диалоговая программная система «STATCONT» // Приборы и системы управления. 1998. № 5. С. 16—18.

13. **McDonald D.** A CUSUM Procedure Based On Sequential Ranks. Naval Research Logistics. 1999;37: 627—646.

14. **Bakir S.T.** A Distribution-free Shewhart Quality Control Chart Based on Signed-ranks. Quality Eng. 2004; 16(4):613—623.

15. **Mitrokhin I.N., Orlov A.I.** Obnaruzhenie Razladki s Pomoshch'yu Kontrol'nykh Kart. Zavodskaya Laboratoriya. Diagnostika Materialov. 2007;5:74—78. (in Russian).

16. **Chakraborti S.b Van de Wiel M.A.** A Nonparametric Control Chart Based on the Mann–Whitney Statistic. IMS Collections. 2008;1:156—172.

17. **Kuznetsov L.A., Zhuravleva M.G.** Postroenie Kart Kontrolya Kachestva s Pomoshch'yu Neparametricheskogo Kriteriya Vilkoksona–Manna–Uitni. Zavodskaya Laboratoriya. Diagnostika Materialov. 2009;1:70—95. (in Russian).

18. **Human S.W., Chakraborti S., Smit C.F.** Nonparametric Shewhart-type Sign Control Charts Based on Runs. Communications in Statistics: Theory and Methods. 2010;39(11):2046—2062.

19. **Veretel'nikova I.V.** Issledovanie i Primenenie Kriteriev Proverki Gipotez ob Otsutstvii Trenda i Kriteriev Odnorodnosti: Avtoreferat … Dis. Kand. Tekhn. Nauk. Novosibirsk: Tipografiya NGTU, 2019. (in Russian).

20. **Kendall M., St'yuart A.** Statisticheskie Vyvody i Svyazi. M.: Nauka, 1973. (in Russian).

21. **Filaretov G.F.** Dialogovaya Programmnaya Sistema «STATCONT». Pribory i Sistemy Upravleniya. 1998;5:16—18. (in Russian).

Сведения об авторах:

Филаретов Геннадий Федорович — доктор технических наук, профессор кафедры управления и интеллектуальных технологий НИУ «МЭИ», e-mail: gefefi@yandex.ru
Бучаала Зинеддин — аспирант кафедры управления и интеллектуальных технологий НИУ «МЭИ», e-mail: bouchaala.zinouzin@gmail.com

Information about authors:

Filaretov Gennadiy F. — Dr.Sci. (Techn.), Professor of Control and Intelligent Technologies Dept., NRU MPEI, e-mail: gefefi@yandex.ru
Bouchaala Zineddin — Ph.D-student of Control and Intelligent Technologies Dept., NRU MPEI, e-mail: bouchaala.zinouzin@gmail.com

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов
Conflict of interests: the authors declare no conflict of interest

Статья поступила в редакцию: *27.11.2020*
The article received to the editor: *27.11.2020*