УДК 004.56 DOI: 10.33075/2220-5861-2021-2-119-130

МУЛЬТИВАРИАНТНЫЙ МНОГОКАНАЛЬНЫЙ ПРОГРАММНО-ИЗМЕРИТЕЛЬНЫЙ КОМПЛЕКС ОБНАРУЖЕНИЯ АНОМАЛЬНЫХ СОСТОЯНИЙ ПРИРОДНО-ТЕХНИЧЕСКИХ ОБЪЕКТОВ И СИСТЕМ

А.В. Скатков, А.А. Брюховецкий, Д.В. Моисеев

[1]ФГАОУ ВО «Севастопольский государственный университет»,
РФ, г. Севастополь, ул. Университетская, 33
E-mail: dmitriymoiseev@mail.ru

Рассматривается подход к мультивариантной классификации состояний природно-технических объектов (ПТО) и систем (ПТС), основанный на развитии методов динамического обнаружения аномалий в информационных потоках данных. Подход базируется на основе оценки статистического расхождения между распределениями вероятностей случайных величин за вариантно изменяемые временные промежутки, а также оценки вероятностей ошибок первого и второго рода. Предложена структура многоканального программно-измерительного комплекса обнаружения аномальных состояний ПТО и ПТС, приведены результаты модельных расчетов. Применение мультивариантного подхода позволяет оптимизировать процессы обработки, анализа, интеграции гетерогенных данных, повысить чувствительность, достоверность и оперативность принимаемых решений.

Ключевые слова: обнаружение аномалий, мультивариантная модель, статистические оценки, аппроксимирующая функция, ошибки первого и второго рода.

Поступила в редакцию: 11.05.2021. После доработки: 07.06.2021.

Введение. Состояние окружающей среды это неотъемлемый, ключевой компонент обобщенной категории качества жизни населения. В связи с этим возникает объективная потребность разработки методов и средств, предназначенных для реализации системы непрерывного мониторинга ключевых показателей окружающей среды и прогнозирования возникновения аномальных состояний экосистем. В исследованиях антропогенной трансформации природной среды в современных условиях, обусловленных большим числом контролируемых параметров объектов природно-технических систем, предъявляются повышенные требования к системам мониторинга и контроля изменения состояния таких объектов с целью оценки их динамики. Совместное использование средств оперативного мониторинга, имитационного моделирования и вероятностных моделей позволяет прогнозировать динамику изменения состояния экосистемы, предупреждать о возможных аномалиях и превентивно выполнять корректирующие действия, предотвращая тем самым возникновение критических ситуаций.

В настоящее время процесс оперативного выявления аномалий данных мониторинга окружающей среды и критических объектов инфраструктуры является комплексной, трудоемкой и трудно формализуемой задачей. В работе [1] представлены примеры разработки и развития методов повышения достоверности принимаемых решений в ходе мониторинга ключевых показателей природной среды, таких как гидрометеорологические данные об уровне загрязнения и составе воздуха, почвы, предельно допустимых выбросов вредных веществ и др. В статье [2] обращается внимание на сложность многомерных структур контролируемых данных, которые могут содержать множество параметров, значения которых со временем меняются и претерпевают структурные трансформации. Поэтому в комплексных информационных системах контроля окружающей среды и критически важных объектов реализуется механизм поддержки принятия решений по выявлению критического состояния объекта мониторинга

[3, 4]. В [5–8] представлены обзоры методов обнаружения аномалий, приведены их сравнительные характеристики и примеры программных систем моделирования и классификации аномальных данных, отмечаются достоинства и недостатки применяемых подходов и условия их практического использования.

Предлагаемый метод обнаружения аномальных данных рассматривается на примере распознавания изменения значений информационных состояний объектов природно-технических систем. Подход базируется на основе оценки статистического расхождения между распределениями вероятностей случайной величины за различные временные промежутки, а также оценки вероятностей ошибок первого и второго рода, полученных при исследовании критических областей принятия гипотез, построенных на основе функций, аппроксимирующих эмпирическое распределение. На первом этапе генерируются и формируются эталонные варианты состояния объектов исходя из экспертных оценок. На втором этапе определяется достоверность оценок указанных вариантов на основе вероятностного моделирования, при котором состояние объекта оценивается по результатам вычисленных значений отдельных его свойств.

Цель предлагаемого подхода – повышение достоверности классификации информационных состояний объектов ПТС и исследование влияния следующих основных параметров на принятие решений:

- достоверность обнаружения изменений состояний объектов ПТС,
- величина расхождения – статистическое расстояние между двумя распределениями вероятностей случайной величины за различные временные промежутки по критерию Кульбака-Лейблера,
- объемы выборок и число интервалов гистограммы,
- параметры непрерывных функций плотности распределения вероятностей и кумулятивного распределения аппроксимирующих эмпирические значения гистограмм,
- пороговые значения критических областей при оценке гипотез принятия решений,
- вероятность появления ошибок первого и второго рода.

Таким образом, применение мультивариантного подхода позволит оптимизировать процессы обработки, анализа, интеграции гетерогенных данных, повысить достоверность и оперативность принимаемых решений.

Постановка задачи. Будем использовать следующую терминологию и обозначения:

G – генеральная совокупность значений случайной величины X распределенной по заданному закону;

X – непрерывная случайная величина, наблюдаемые значения которой составляют выборку;

n – объем выборки – число реализаций случайной величины X ;

$V = \{x_1, \ldots, x_n\}$ – собственно выборка (экспериментальные данные – реализации X);

$R = (\max x_i - \min x_i)$, $i = 1, n$ – размах варьирования выборки;

m_x – выборочное среднее;

S^2 – выборочная оценка дисперсии;

$\sigma = \sqrt{S^2}$ – стандартное отклонение;

k – число интервалов гистограммы;

$ex(G_{Vi}, G_{Vj}) - G_{Vi}$, G_{Vj} – гистограммы, участвующие в эксперименте;

$j = 1,k$ – номера интервалов гистограммы;

$[x_{i-1}, x_i]$ – границы i –ого интервала гистограммы;

n_i , i =1, k – число элементов выборки x_i , попавших в i -ый интервал;

$p_i = n_i/n$, i=1,k – относительные частоты в интервалах гистограммы;

$f_x(V_x, k_x, m_x, \sigma_x)$ – аппроксимирующая функция эмпирического распределения, построенная для эталонного состояния объекта – $x \in \{0,1,2\}$;

$G_x(V_x, k_x)$ – гистограмма, заданная объемом выборки и числом интервалов, с определенными для каждого интервала соответствующими частотами.

Без потери общности в простейшем случае будем полагать, что контролируемый объект может находиться в одном из трех состояний: S_x – состояние кон-

тролируемого объекта, x ϵ {0,1,2}: S_0 – нормальное, S_1 – предкритическое, S_2 – критическое. Для каждого из трех состояний по соответствующим выборкам из генеральной совокупности G сформированы эталонные гистограммы $G_0(V_0,k_0)$, $G_1(V_1,k_1)$, $G_2(V_2,k_2)$ и вычислены оценки параметров аппроксимирующих функций – $f_0(V_0,k_0,m_0,\sigma_0)$, $f_1(V_1,k_1,m_1,\sigma_1)$ и $f_2(V_2,k_2,m_2,\sigma_2)$.

Известно, что оценка расхождений между распределениями при практическом использовании весьма чувствительна к объемам выборок V и числу интервалов k. С этой целью проводятся модельные эксперименты для повышения достоверности принимаемых гипотез. Эти исследования могут быть выполнены циклически для каждого эталонного варианта, определяя ошибки первого и второго рода.

Таким образом, по построенным эталонным вариантам требуется классифицировать текущее состояние объекта S_x, представленного гистограммой $G_x(V_x,k_x)$, для которой построена аппроксимирующая функция – $f_x(V_x,k_x,m_x,\sigma_x)$, и для заданного уровня достоверности определить вероятности ошибок первого и второго рода.

Методы и результаты. Для вычисления расхождения будем использовать меру (дивергенция) Кульбака-Лейблера. Введем обозначение $D(P,Q)$ между двумя распределениями Q и P. Тогда дивергенция определится как [9, 10]:

$$D(P,Q)=\sum_{i=0}^{n} Pi(x) * (log\ (P_i\ (x) / Q_i(x))).$$

Дивергенция Кульбака распределения P относительно Q может быть оценена как:

- $D(P,Q) \le Z$ – отсутствие расхождения,
- $D(P,Q) > Z$ – наблюдение расхождения выборок,

где Z – предельное значение меры расхождения, зависящее от критичности контролируемого значения параметра объекта, которое задается экспертом. Тогда нулевая гипотеза H_0 имеет место при $D(P,Q) \le Z$ – отсутствие расхождения. В противном случае принимается гипотеза H_1 – качественное изменение информационного состояния объекта.

С целью сравнения мультивариантных оценок расхождений между эталонными $S=\{S_0, S_1, S_2\}$ и исследуемым S_x вероятностными распределениями определим понятие зоны оценки величины расхождения для каждого эталона. Будем для определенности рассматривать следующие зоны классификации: $[Z_1;Z_2)$, $[Z_2;Z_3)$, $[Z_3;Z_4]$. В зависимости от принадлежности текущего значения расхождения $D(P,Q) \in Z_i\ (i=1,k)$ будем классифицировать следующие информационные состояния объекта на примере трех состояний:

$D(P,Q)<Z_1$ – отсутствие расхождения (нормальное состояние),

$Z_1 \le D(P,Q)<Z_2$ – неустойчивая область (предкритическое состояние),

$Z_2 \le D(P,Q)$ – наблюдение расхождения (критическое состояние).

Очевидно, вид гистограммы зависит от того, как построены классовые интервалы принадлежности случайной величины [11]. «Правильность» разбиения подразумевает, что в стационарном случае ошибка аппроксимации предположительно непрерывной плотности функции распределения кусочно-постоянной функцией минимальна. Трудность состоит в том, что оцениваемая плотность неизвестна, поэтому число интервалов сильно сказывается на виде распределения частот конечной выборки. С одной стороны, при фиксированной длине выборки укрупнение интервалов разбиения ведет к уточнению эмпирической вероятности попадания в них, но слишком сильно сглаживает изучаемое распределение. С другой стороны, уменьшение ширины интервалов делает вид распределения неоправданно изрезанным в силу малого количества данных, случайно попадающих в каждый из интервалов. На практике число интервалов распознавания задается экспертом и зависит от назначения объекта, его динамических свойств, требований к точности контроля, возможных рисков принимаемых решений и т.д. На рис. 1 представлена матрица соответствия гипотез и принимаемых решений с учетом мультивариантных оценок эталонных состояний $S=\{S_0, S_1, S_2\}$.

Гипотезы		Нулевая гипотеза $H_0(S, S_x)$	
		Верная	Ложная
Решение о нулевой гипотезе $H_0(S_0,S_x)$	Принять	Правильный вывод – TN (вероятность = $1-\alpha$) $P(H_0(S_0,S_x)\|H_0(S_0,S_x))$	Ошибка II типа – FN (вероятность = β) $P(H_0(S_0,S_x)\|H_1(S_0,S_x))$
	Отклонить	Ошибка I типа – FP (вероятность = α) $P(H_1(S_0,S_x)\|H_0(S_0,S_x))$	Правильный вывод – TP (вероятность = $1-\beta$) $P(H_1(S_0,S_x)\|H_1(S_0,S_x))$
Решение о нулевой гипотезе $H_0(S_1,S_x)$	Принять	Правильный вывод – TN (вероятность = $1-\alpha$) $P(H_0(S_1,S_x)\|H_0(S_1,S_x))$	Ошибка II типа – FN (вероятность = β) $P(H_0(S_1,S_x)\|H_1(S_1,S_x))$
	Отклонить	Ошибка I типа – FP (вероятность = α) $P(H_1(S_1,S_x)\|H_0(S_1,S_x))$	Правильный вывод – TP (вероятность = $1-\beta$) $P(H_1(S_1,S_x)\|H_1(S_1,S_x))$
Решение о нулевой гипотезе $H_0(S_2,S_x)$	Принять	Правильный вывод – TN (вероятность = $1-\alpha$) $P(H_0(S_2,S_x)\|H_0(S_2,S_x))$	Ошибка II типа – FN (вероятность = β) $P(H_0(S_2,S_x)\|H_1(S_2,S_x))$
	Отклонить	Ошибка I типа – FP (вероятность = α) $P(H_1(S_2,S_x)\|H_0(S_2,S_x))$	Правильный вывод – TP (вероятность = $1-\beta$) $P(H_1(S_2,S_x)\|H_1(S_2,S_x))$

Рис. 1. Матрица соответствия гипотез и принимаемых решений с учетом мультивариантных оценок эталонных состояний
Fig. 1. The matrix of correspondence of hypotheses and decisions made taking into account multivariate estimates of reference states

Выдвинутая гипотеза подлежит статистической проверке, при этом в двух случаях имеют место риски принятия неверных решений – ошибки первого и второго рода. Решение о справедливости той или иной гипотезы основывается на знании выборки объема V случайных значений контролируемого параметра. На рис. 2 схематично изображены функции плотности распределения случайной величины и показаны области принятия гипотез для одного эталонного варианта:

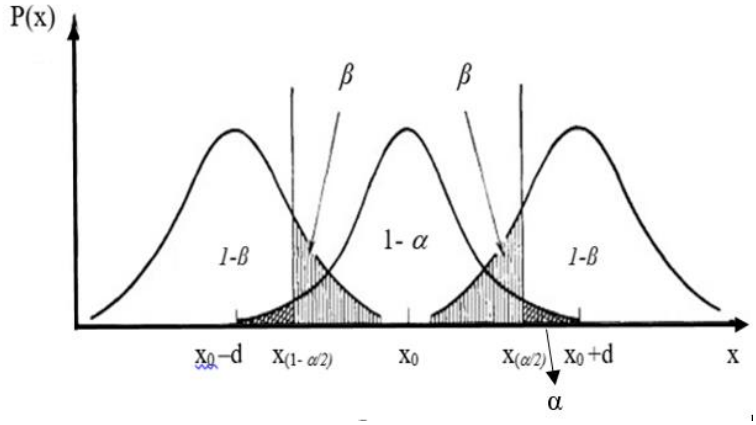

Рис. 2. Определение ошибки второго рода при проверке гипотез одного эталонного варианта
Fig. 2. Determination of the second kind of error when testing hypotheses of one reference variant

Области принятия гипотез, представленные на рис. 2, интерпретируются следующим образом. Если гипотеза состоит в том, что $x=x_0$, тогда как на самом деле $x=x_0\pm d$, то вероятность того, что x попадет в область принятия гипотезы, заключенную между $x_{(1-\alpha/2)}$ и $x_{(\alpha/2)}$ – (квантили нормального распределения

порядка (1 –α/2) и (α/2) соответственно для двусторонней гипотезы) равна β – вероятности шибки второго рода при выявлении отклонений величиной $\pm$d от гипотетического значения.

При решении поставленной задачи возникают проблемы обработки больших данных: вычислительная сложность; дефицит априорной информации о мониторируемых объектах; нестационарность состояния объектов и среды; необходимость получения результатов в реальном времени. Поэтому предлагается использовать интеллектуальную цифровую технологию, реализованную на основе программно-измерительного комплекса, осуществляющую мониторинг и прогнозирование ключевых параметров экосистем посредством имитационного и вероятностного моделирования. Структурно-функциональная схема программно-измерительного комплекса представлена на рис. 3.

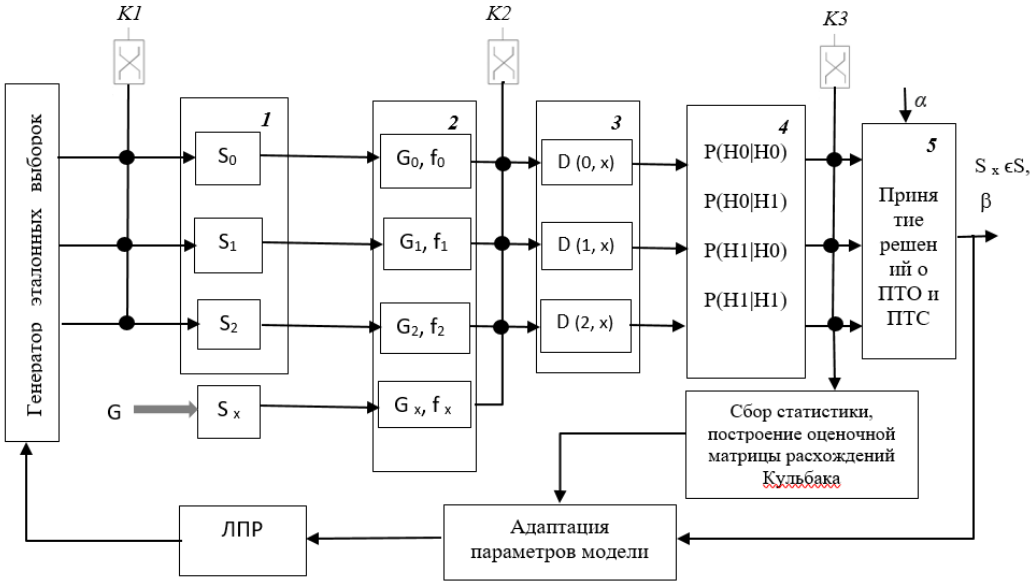

Рис. 3. Структурно-функциональная схема многоканального программно-измерительного комплекса обнаружения аномалий – MDA
Fig. 3. Structural and functional diagram of a multichannel software and measurement complex for detecting anomalies – MDA

В структурной схеме многоканального программно-измерительного комплекса обнаружения аномалий – MDA коммутация информационных потоков осуществляется с помощью управляемых коммутаторов: К1 – коммутатор эталонных выборок, К2 – коммутатор выходных каналов эталонных описаний и исследуемого распределений, К3 – коммутатор подсистемы сбора статистики и построение оценочных матриц расхождений для каждого эталона. На рис. 3 обозначено: 1 – блок параметрической и структурной настройки эталонных состояний ПТО и ПТС (объем анализируемых выборок, число интервалов гистограммы, длины интервалов, уровень значимости); 2 – блок формирования эталонных вариантов гистограмм $G_0(V_0,k_0)$, $G_1(V_1,k_1)$, $G_2(V_2,k_2)$ и параметров соответствующих аппроксимирующих функций – $f_0(V_0,k_0,m_0,\sigma_0)$, $f_1(V_1,k_1,m_1,\sigma_1)$ и $f_2(V_2,k_2,m_2,\sigma_2)$; 3 – блок оценки расхождения между эталонными и исследуемым распределениями по критерию Кульбака; 4 – блок формирования оценок вероятностей принятия гипотез: P(H0|H0), P(H0|H1), P(H1|H0), P(H1|H1) для каждого эталонного варианта; 5 – блок получения результатов классификации информационного состояния ПТО и ПТС, передача данных в блок «адаптации параметров модели» и «ЛПР».

Комплекс может работать в двух режимах: рабочего и тестового имитационного моделирования. Второй режим предназначен для обучения ЛПР обоснованному принятию решений при посту-

лировании событий, связанных с возникновением ситуаций, соответствующих гипотезам H0, H1. Второй режим позволяет, кроме того, определить области устойчивого / неустойчивого распознавания по отношению к используемым критериям и пороговым значениям, области принятия гипотез.

Задача моделирования обнаружения изменения состояния ПТО и ПТС решается по следующей алгоритмической схеме:

1. Генерируются эталонные выборки заданного объема и числа интервалов из генеральной совокупности для заданных эталонных состояний S_x, $x \in \{0,1,2\}$.

2. Формируются эталонные гистограммы $G_0(V_0,k_0)$, $G_1(V_1,k_1)$, $G_2(V_2,k_2)$ и вычисляются оценки параметров аппроксимирующих функций – $f_0(V_0,k_0,m_0,\sigma_0)$, $f_1(V_1,k_1,m_1,\sigma_1)$ и $f_2(V_2,k_2,m_2,\sigma_2)$.

3. Задается величина сдвига параметров функции распределения, характеризующей измененные состояния объектов. Генерируется выборка *Y*.

4. Оценивается расхождение между эталонными и исследуемой выборками по критерию Кульбака.

5. Формулируется нулевая гипотеза об отсутствии существенного различия между распределениями S_x и *Y*;

6. Определяется расчетное значение выбранного непараметрического критерия проверки нулевой гипотезы H_0 при заданном уровне значимости α для;

7. Определяются критические области значений критерия проверки нулевой гипотезы – ($P(H_0|H_0)$, $P(H_1|H_1)$, $P(H_1|H_0)$, $P(H_0|H_1)$ для каждого эталона. Оцениваются ошибки первого и второго рода.

В соответствии с поставленными задачами и мультивариантным подходом были проведены целенаправленные модельные эксперименты, в ходе которых определялось влияние ряда факторов на изменения информационного состояния ПТО и ПТС: объемы выборок – *V;* число интервалов гистограммы; границы *i* –ого интервала гистограммы – *k*; $[x_{i-1};x_i]$; исследуемые распределения – *P*,*Q*. На рис. 4 приведен план проведения экспериментов в виде матрицы для числа интервалов k=3.

№ эксперимента	Факторы			Взаимодействие	Результаты
	K	V	*P*,*Q*	ex(*P*,*Q*)	$D_{k,V}(P,Q)$
1	3	30	1,2	ex(1,2)	$D_{3,30}(1,2)$
2		40		ex(1,2)	$D_{3,40}(1,2)$
3		60		ex(1,2)	$D_{3,60}(1,2)$
4		100		ex(1,2)	$D_{3,100}(1,2)$
5	3	30	2,3	ex(2,3)	$D_{3,30}(2,3)$
6		40		ex(2,3)	$D_{3,40}(2,3)$
7		60		ex(2,3)	$D_{3,60}(2,3)$
8		100		ex(2,3)	$D_{3,100}(2,3)$
9	3	30	1,4	ex(1,4)	$D_{3,30}(1,4)$
10		40		ex(1,4)	$D_{3,40}(1,4)$
11		60		ex(1,4)	$D_{3,60}(1,4)$
12		100		ex(1,4)	$D_{3,100}(1,4)$
13	3	30	2,4	ex(2,4)	$D_{3,30}(2,4)$
14		40		ex(2,4)	$D_{3,40}(2,4)$
15		60		ex(2,4)	$D_{3,60}(2,4)$
16		100		ex(2,4)	$D_{3,100}(2,4)$

Рис. 4. План проведения экспериментов для k=3

Fig. 4. Plan of experiments for k=3

Аналогичные планы построены для k=4,5 и т.д.

В соответствии с планом проведения экспериментов получены следующие результаты исследования процессов обнаружения изменения информационного состояния ПТО и ПТС, а также оценки достоверности принимаемых гипотез.

Исследование влияния объема выборок – *V*. Задано: зоны классификации объектов $[Z_{i-1};Z_i]$ для трех состояний S_x , $x \in \{0,1,2\}$. Число интервалов *k*=3. Сравнивались оценки расхождений для распределений S_x и *Y* определяемых экспертом для случаев, когда ширина интервалов $[x_{i-1};x_i]$ отличалась незначительно (однородные) и существенно (неоднородные). Были заданы следующие интервалы $[x_{i-1};x_i]$ гистограмм G_1,G_2, G_3,G_4 :

– интервалы-1 {0; 0,50; 0,80; 1},

– интервалы-2 {0; 0,40; 0,60; 1},

– интервалы-3 {0; 0,30; 0,70; 1},

– интервалы-4 {0; 0,10; 0,50; 1}.

Ширина интервалов – 1, 2, 3 между собой отличается незначительно, а ширина интервалов – 4 существенно отличается от остальных. Каждый эксперимент повторялся 20 раз. На рис. 5 представлены значения меры Кульбака $D(S_x,Y)$ при сравнении распределений выборок для интервалов: ex(1,2) , ex(2,3), ex(1,4) при *V*=60.

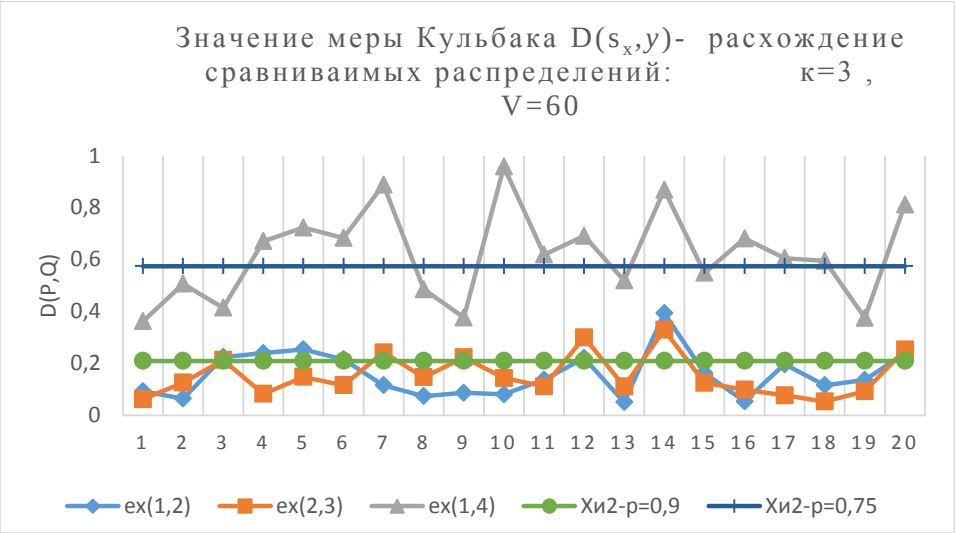

Рис. 5. Расхождение D(S_x,*Y*) при оценке распределений ex(1,2) , ex(2,3), ex(1,4)*: k*=3 , *V*=60
Fig. 5. The discrepancy D (Sx,Y) when estimating the distributions ex (1,2), ex (2,3), ex (1,4): k=3, V=60

На рис. 6 представлены значения меры Кульбака D(S_x,*Y*) при сравнении распределений выборок для интервалов гистограмм: ex(1,2) , ex(2,3), ex(1,4) при *V*=100.

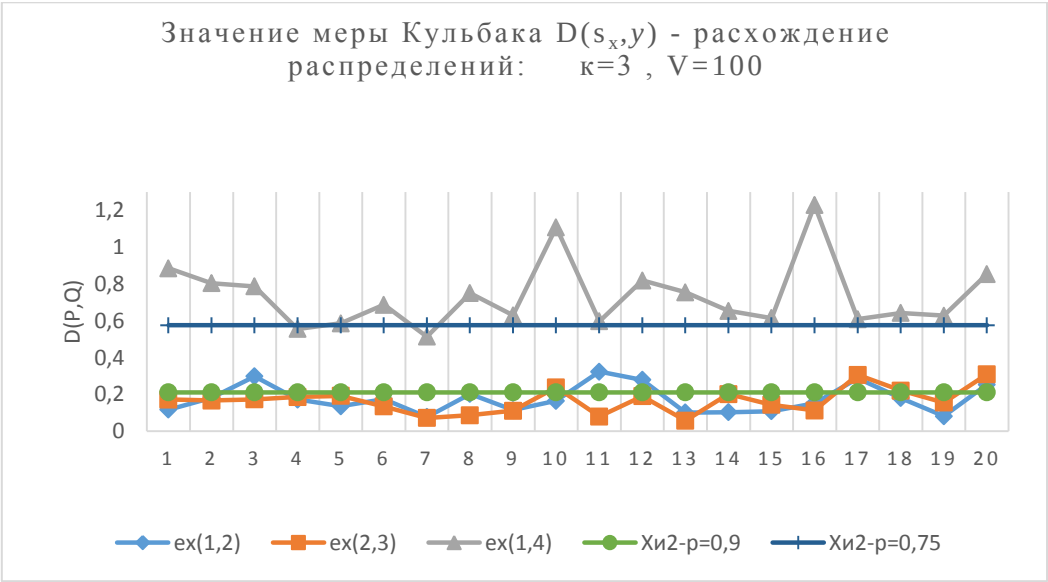

Рис. 6. Расхождение D(S_x,*Y*) при оценке распределений ex(1,2) , ex(2,3), ex(1,4)*: k*=3 , *V*=100
Fig. 6. Discrepancy D (Sx, Y) in the evaluation of the ex(1,2), ex (2,3), ex (1,4) definitions: k=3, V=100

Исследование влияния объема выборок – *V* при измененном числе интервалов *k*=4. Задано: зоны классификации объектов [Z_{i-1};Z_i] для трех состояний S_x , x ϵ {0,1,2}. Сравнивались оценки расхождений для распределений S_x и *Y* определяемых экспертом. Были заданы следующие интервалы [x_{i-1};x_i] гистограмм G_5 ,G_6, G_7:

– интервалы-5 {0; 0,30;0,60; 0,90 ; 1},

– интервалы-6 {0; 0,20;0,40; 0,80 ; 1},

– интервалы-7 {0; 0,10;0,30; 0,50 ; 1}.

Эксперименты проводились по сценарию предыдущего. Первая пара выборок ex(5,6) незначительно отличается по ширине интервалов, в то время как вторая пара ex(5,7) – существенно. На рис. 7 представлены значения меры Кульбака D(S_x,*Y*) при сравнении распределений выборок: ex(5,6) , ex(5,7) при объеме выборки *V*=100.

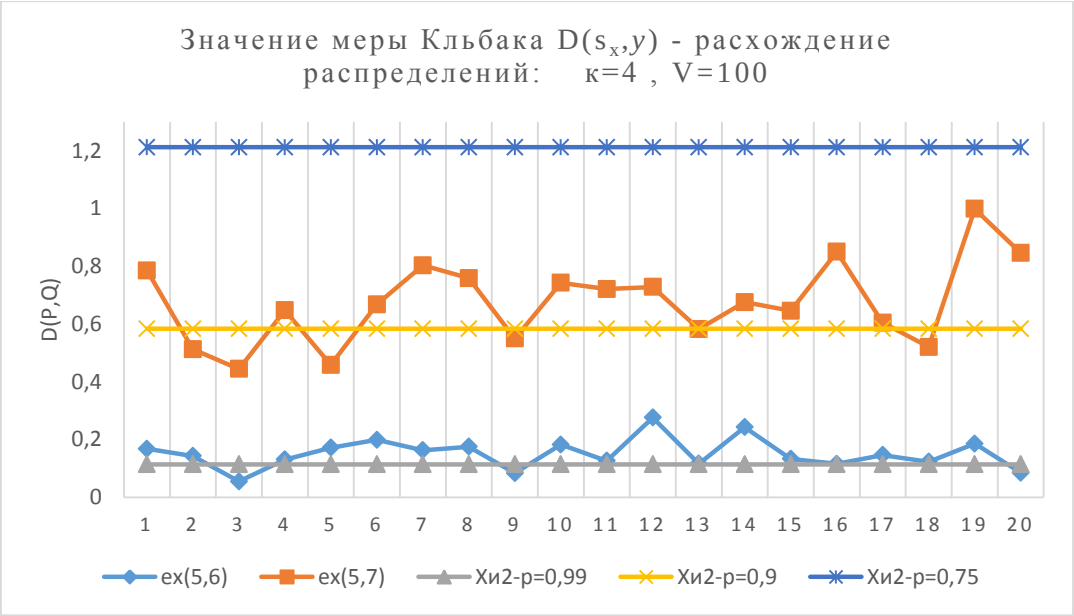

Рис. 7. Расхождение D(S_x,*Y*) при оценке распределений ex(5,6), ex(5,7)*: k*=4 , *V*=100
Fig. 7. The discrepancy D (Sx,Y) when estimating the distributions ex(5,6), ex(5,7)*: k*=4 , *V*=100

В табл. 1 представлены значения расхождения D(S_x,*Y*) для *k*=4 и *V*=30, 100.

Таблица 1. Значения расхождения D(S_x,*Y*) для *k*=4 и *V*=30, 100

Объемы выборок	*V* = 30		*V* = 100	
	ex(5,6)	ex(5,7)	ex(5,6)	ex(5,7)
max	0,41	1,48	0,28	1,01
min	0,06	0,44	0,06	0,45
Ср. знач.	0,18	0,83	0,15	0,68
Δ	0,34	1,04	0,22	0,55

По результатам анализа проведенных экспериментов можно констатировать следующее. Различия величины Δ между максимальными и минимальными значениями расхождения D(S_x,*Y*) при увеличении числа интервалов и объема выборки уменьшился: при сравнении однородных выборок при *V*=100 он составил 0,22, а неоднородных – 0,55, а при *V*=30 соответственно – 0,34 и 1,04. Таким образом, при увеличении числа интервалов и объема выборки мы наблюдаем увеличение величины расхождения между однородными и неоднородными распределениями за счет уменьшения значения СКО. При этом среднее значение расхождения при *V*=30 для однородных распределений ex(5,6) составило 0,18, а для неоднородных ex(5,7) – 0,83. При V=100 соответствующие значения – 0,15 и 0,68. Этот факт свидетельствует о повышении достоверности классификации состояний объектов при увеличении зоны [Z_{i-1};Z_i] и уменьшении числа ошибок 1-го и 2-го рода.

На рис. 8 представлена зависимость величины среднеквадратического отклонения СКО меры Кульбака D(S_x,*Y*) между однородными ex(1,2), ex(2,3) и неоднородными ex(1,4), ex(2,4) распределениями в зависимости от объема *V* при *k*=3.

На рис. 8 видно, что наблюдается тенденция уменьшения величины СКО от увеличения объема *V* во всех экспериментах. Результаты экспериментов являются подтверждением того, что при возрастании объемов *V* значение среднеквадратического отклонения (длина доверительного интервала для заданного уровня достоверности – α) уменьшается пропорционально корню квадратному из объема выборки – *V*.

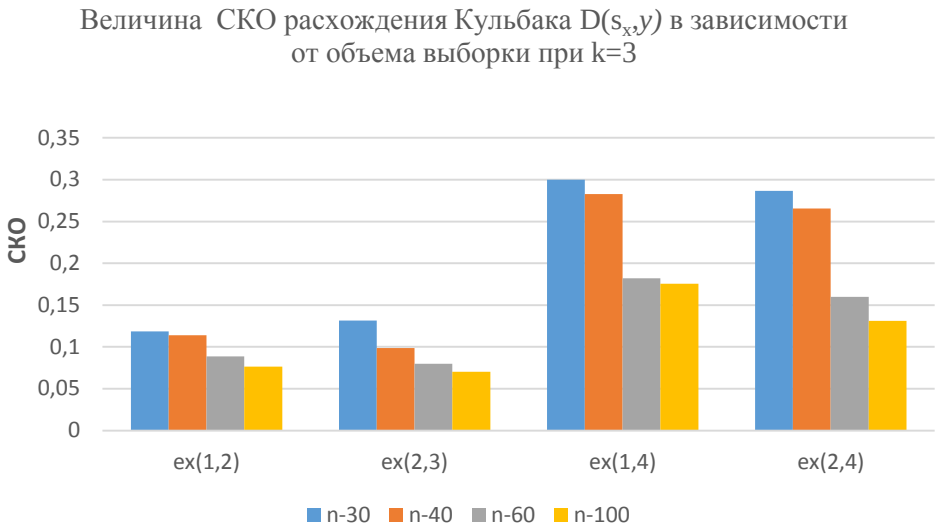

Рис. 8. Величина СКО расхождения $D(S_x,Y)$ между однородными и неоднородными распределениями в зависимости от объема выборки при k=3
Fig. 8. The value of the SKR of the discrepancy D (Sx,Y) between homogeneous and inhomogeneous distributions as a function of the sample size at k=30

Для оценки мощности и чувствительности используемых непараметрических критериев исследуется подход вероятностного моделирования обнаружения сдвигов в малых выборках. Исследуется пара выборок X и Y объемом V, извлеченных из генеральных совокупностей, имеющих один и тот же вид распределения – нормальный. Выборка X рассматривается как эталонная совокупность характеристик объектов до появления внешнего фактора. В результате проведения экспериментов получены оценки:

- вероятностей ошибок первого и второго рода и исследована чувствительность метода в зависимости от объема выборок,
- зависимости вероятности ошибки β при заданной вероятности α и величине сдвига Δ математического ожидания функции распределения, характеризующей измененные состояния объектов при различных объемах выборок.

На рис. 9 представлена зависимость вероятности ошибки β от вероятности α при следующих значениях параметров: объемы выборок V_0 =15 , V_1 = 30, m_0=4, m_1 =5, $\sigma_0 = \sigma_1$ = 1.3, $\alpha \in [0.005;0.03]$ с шагом 0.001.

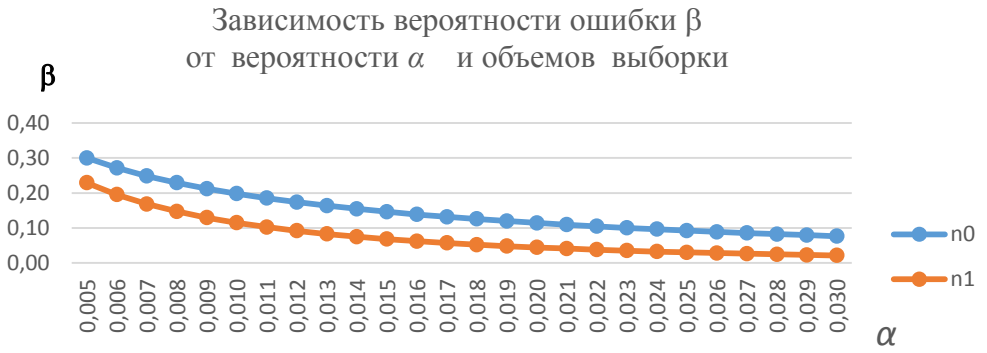

Рис. 9. Зависимость вероятности ошибки β от вероятности α и объема выборки
Fig. 9. Dependence of the error probability β on the probability α and the sample size

На рис. 10 представлена зависимость вероятности ошибки β при величине сдвига Δ математического ожидания функции плотности распределения, характеризующей измененное состояние объектов при различных объемах выборок (при заданной вероятности $\alpha = 0,05$, величине сдвига Δ ϵ [0.05;2.0] с шагом 0.05, d ϵ [0.038;1.54], σ=1.3, V ϵ {15,20,30,40,60}).

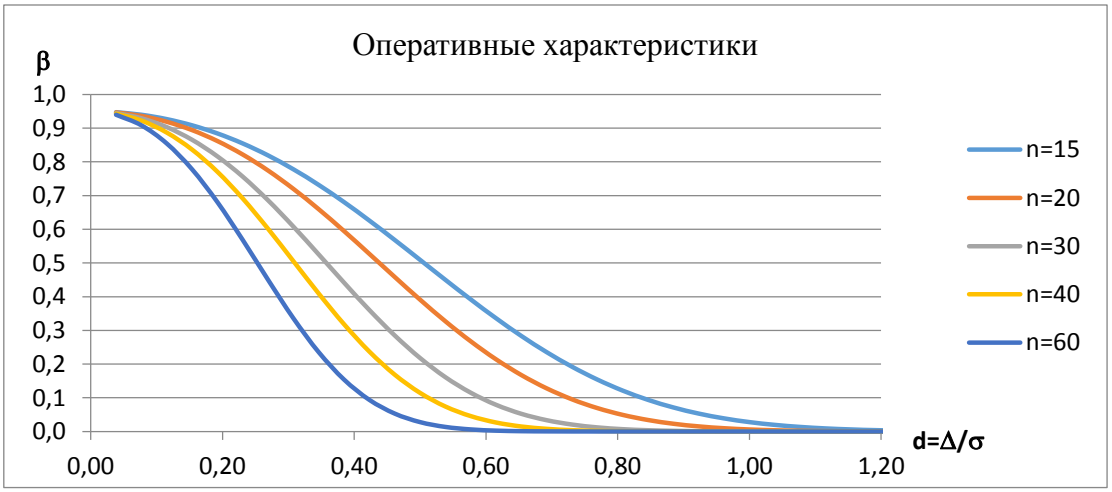

Рис. 10. Зависимость вероятности ошибки β при заданной вероятности α и величины сдвига математического ожидания функции плотности распределения, характеризующей измененные состояния объектов при различных объемах выборок

Fig. 10. The dependence of the error probability β for a given probability α and the value of the shift of the mathematical expectation of the distribution density function that characterizes the changed states of objects at different sample volumes

Увеличить мощность критерия можно двумя основными способами.

1) Увеличить размер выборки (с увеличением числа наблюдений уменьшается стандартная ошибка и увеличивается мощность). Однако на практике из-за ограниченности ресурсов не всегда возможно увеличить выборку.

2) Увеличить область отвержения гипотезы. С увеличением области отвержения увеличивается вероятность отвергнуть верную гипотезу – ошибка первого рода (α). При этом увеличивается вероятность получить статистически значимый результат. Вместе с тем, когда увеличивается ошибка первого рода, ошибка второго рода (β) уменьшается, а значит мощность критерия (1-β) увеличивается.

Заключение. В соответствии с поставленной задачей предложенный в статье подход к мультивариантной классификации состояний природно-технических объектов и систем основан на развитии методов динамического обнаружения аномалий в информационных потоках данных. Подход базируется на основе оценки статистического расхождения между распределениями вероятностей случайных величин за вариантно изменяемые временные промежутки, а также оценки вероятностей ошибок первого и второго рода. Предполагается, что имеется возможность первоначально сформировать эталонные варианты состояний ПТО и ПТС исходя, например, из экспертных оценок или априорной информации. Применение мультивариантного подхода позволяет оптимизировать процессы обработки, анализа, интеграции гетерогенных данных. Полученные результаты исследования подтверждают устойчивость и чувствительность метода при выборе пороговых значений интервалов, определяющих состояния объектов.

В процессе имитационного и вероятностного моделирования становится возможным определение и оценка совокупности следующих характеристик:

- параметрически настраиваемых пороговых значений критических областей;
- соответствие теоретического и эмпирического распределения случайной величины;
- области достоверного распознавания состояния объекта;
- области принятия гипотез;
- определение мощности критерия.

В зависимости от назначения модели, уровня критичности объектов контроля эксперт вправе задавать необходимые пороговые значения настроечных параметров модели, для которых, с одной стороны, будет обеспечена высокая достоверность контролируемых значений характеристик объектов, с другой – достигается допустимое число ошибок первого и второго рода, а значит будут снижены риски при принятии ошибочных решений. Данная модель может применяться в других предметных областях, где требуется оценка динамических параметров контролируемых объектов, например, при обнаружении уязвимостей интерфейсов беспилотных транспортных средств в инфраструктуре умного города.

Работа выполнена при частичной поддержке Российского фонда фундаментальных исследований (грант № 19-29-06015, 19-29-06023, 18-47-920007).

СПИСОК ЛИТЕРАТУРЫ

1. *Гайский В.А., Гайский П.В.* Многомерный гармонический анализ Фурье при измерениях полей морской среды // Системы контроля окружающей среды. 2019. № 4 (38). С. 33–42.

2. *Брюховецкий А.А., Скатков А.В., Шишкин Ю.Е.* Моделирование процессов обнаружения аномалий в сложноструктурированных данных мониторинга // Системы контроля окружающей среды. 2017. № 9 (29). С. 45–49.

3. *Аэрокосмический* мониторинг объектов нефтегазового комплекса / под ред. акад. В.Г. Бондура. М.: Научный мир, 2012. 558 с.

4. *Скатков А.В., Брюховецкий А.А., Моисеев Д.В.* Интеллектуальная система мониторинга для решения крупномасштабных научных задач в облачных вычислительных средах // Информационно-управляющие системы. 2017. № 2 (87). С. 19–25.

5. *Chandola V., Banerjee A., Kumar V.* Anomaly detection: a survey // ACM Computing Surveys, 09 2009. P. 1–72.

6. *Chan P.K., Mahoney M.V.* Modeling multiple time series for anomaly detection // In Proceedings of the Fifth IEEE International Conference on Data Mining. IEEE Computer Society, Washington, USA, 2005. P. 90–97.

7. *Agarwal D.* Detecting anomalies in cross-classified streams: a bayesian approach // Knowledge and Information Systems. 2006. Vol. 11, № 1. P. 29–44.

8. *Charikar M., Chekuri C., Feder T., and Motwani R.* Incremental clustering and dynamic information retrieval // SIAM Journal on Computing. 2004. Vol. 33, No. 6. P. 1417–1440.

9. *Кульбак С.* Теория информации и статистика. М.: Наука, 1967. 408 с.

10. *Скатков А.В., Брюховецкий А.А., Моисеев Д.В.* Мера Кульбака в задачах динамической кластеризации наблюдений состояния окружающей среды // Системы контроля окружающей среды. 2019. № 3 (37). С. 35–38.

11. *Orlov S.P., Vasilchenko A.N.* Intelligent measuring system for testing and failure analysis of electronic devices // 2016 XIX IEEE International Conference on Soft Computing and Measurements (SCM). IEEE Conference Publications. 2016. Vol. 1. P. 401–403.

12. *Скатков А.В., Брюховецкий А.А., Моисеев Д.В.* Адаптивный метод обнаружения уязвимостей интерфейсов беспилотных транспортных средств в инфраструктуре умного города // Инфокоммуникационные технологии. 2020. Т. 18. № 1. С. 45–50.

MULTIVARIATE MULTICHANNEL SOFTWARE AND MEASUREMENT COMPLEX FOR DETECTING ANOMALOUS STATES OF NATURAL AND TECHNICAL OBJECTS AND SYSTEMS

A.V. Skatkov, A.A. Bryukhovetskiy, D.V. Moiseev

Sevastopol State University,
RF, Sevastopol, Universitetskaya St., 33

An approach to the multivariate classification of the states of natural-technical objects and systems is considered, based on the development of methods for dynamic detection of anomalies in information data flows. The approach is based on an estimate of the statistical discrepancy between the probability distributions of random variables over variably changeable time intervals, as well as an estimate of the probabilities of errors of the first and second kind. The structure of a multichannel software and measurement complex for detecting anomalous states of PTO and PTS is proposed, and the results of model calculations are presented. The use of the multivariate approach allows optimizing the processes of processing, analysis and integration of heterogeneous data, as well as increasing the sensitivity, reliability and efficiency of decisions.
Keywords: anomaly detection, multivariate model, statistical estimates, approximation function, errors of the first and second kind.

REFERENCES

1. *Gajskij V.A. and Gajskij P.V.* Mnogomernyj garmonicheskij analiz Fur'e pri izmereniyah polej morskoj sredy (Multidimensional harmonic analysis for measurements of marine environment) *Monitoring systems of environment*, 2019, No. 4 (38), pp. 33–42.

2. *Bryuhoveckij A.A., Skatkov A.V., Shishkin Yu.E.* Modelirovanie processov obnaruzheniya anomalij v slozhnostrukturirovannyh dannyh monitoring (Modeling of anomaly detection processes in complex structured monitoring data). *Sistemy kontrolya okruzhayushchej sredy*, 2017, No. 9 (29), pp. 45–49.

3. *Bondura V.G.* Aerokosmicheskij monitoring ob"ektov neftegazovogo kompleksa (Aerospace monitoring of oil and gas facilities). Moscow: Nauchnyj mir, 2012, 558 p.

4. *Skatkov A.V., Bryuhoveckij A.A., and Moiseev D.V.* Intellektual'naya sistema monitoringa dlya resheniya krupnomasshtabnyh nauchnyh zadach v oblachnyh vychislitel'nyh sredah (Intelligent monitoring system for large-scale tasks in cloud environments). *Informacionno-upravlyayushchie sistemy*, 2017, No. 2 (87), pp. 19–25.

5. *Chandola V., Banerjee A., and Kumar V.* Anomaly detection: a survey. *ACM Computing Surveys*, 09 2009. pp. 1–72.

6. *Chan P.K. and Mahoney M.V.* Modeling multiple time series for anomaly detection. *In Proceedings of the Fifth IEEE International Conference on Data Mining. IEEE Computer Society*, Washington, USA, 2005, pp. 90–97.

7. *Agarwal D.* Detecting anomalies in cross-classified streams: a bayesian approach. *Knowledge and Information Systems*, 2006. Vol. 11, No. 1. pp. 29–44.

8. *Charikar M., Chekuri C., Feder T., and Motwani R.* Incremental clustering and dynamic information retrieval. *SIAM Journal on Computing*, 2004, Vol. 33, No. 6, pp. 1417–1440.

9. *Kul'bak S.* Teoriya informacii i statistika. Moscow: Nauka, 1967, 408 p.

10. *Skatkov A.V., Bryuhoveckij A.A., and Moiseev D.V.* Mera Kul'baka v zadachah dinamicheskoj klasterizacii nablyudenij sostoyaniya okruzhayushchej sredy (Measure of Kulbak in problems of dynamic clustering of observations of the environmental state). *Sistemy kontrolya okruzhayushchej sredy*, 2019, No. 3 (37). pp. 35–38.

11. *Orlov S.P. and Vasilchenko A.N.* Intelligent measuring system for testing and failure analysis of electronic devices. *2016 XIX IEEE International Conference on Soft Computing and Measurements (SCM). IEEE Conference Publications*, 2016, Vol. 1, pp. 401–403.

12. *Skatkov A.V., Bryuhoveckij A.A., and Moiseev D.V.* Adaptivnyj metod obnaruzheniya uyazvimostej interfejsov bespilotnyh transportnyh sredstv v infrastrukture umnogo goroda (An adaptive method for detecting vulnerabilities in the interfaces of unmanned vehicles in the infrastructure of a smart city). *Infokommunikacionnye tekhnologii*, 2020, Vol. 18, No. 1, pp. 45–50.