ПОВЫШЕНИЕ ДОСТОВЕРНОСТИ ОЦЕНОК РИСКОВ В ПРОЦЕССАХ МОНИТОРИНГА ПРИ РАСПРЕДЕЛЕНИЯХ ОБЩЕГО ВИДА

Ю.Е. Шишкин^{1,2}, А.В. Скатков¹

¹ ФГАОУ ВО «Севастопольский государственный университет»,
РФ, г. Севастополь, ул. Университетская, 33

² Институт природно-технических систем,
РФ, г. Севастополь, ул. Ленина, 28
E-mail: iurii.e.shishkin@gmail.com

Выполнен анализ влияния выбора пороговых значений стратегии принятия решений на возникновение ошибок первого и второго рода при выявлении аномалий в процессе мониторинга параметров окружающей среды, которые на практике описываются распределениями общего вида. В ходе целенаправленного эксперимента рассмотрены аномалии двух классов: сдвиг распределения наблюдения и повышение дисперсионности. Выполнена оценка чувствительности детектирующего статистического метода и сформулированы практические рекомендации по выбору пороговых значений системы поддержки принятия решений по выявлению аномалий, учитывающие закон распределения наблюдаемой случайной величины.

Ключевые слова: мониторинг, математическое моделирование, детектирование аномалий, кластеризация, критические системы, интеллектуальный анализ данных, распределение общего вида.

Поступила в редакцию: 06.01.2019.

Введение. Мониторинг окружающей среды представляет собой важный компонент системы рационального природопользования и в контексте непрерывно возрастающей антропогенной нагрузки на природные комплексы повышение достоверности выявления аномалий при мониторинге представляет собой актуальную научно-практическую задачу [1, 2].

Существующие модели выявления аномалий при мониторинге, как правило, строятся в предположении того, что исходные данные распределены по нормальному закону, что обусловлено удобством выполнения аналитических расчетов [3, 4], в то время как в практике мониторинга физических величин часто имеют дело с аномальными моделями распределения, у которых параметры асимметрии и эксцесса значительно отличаются от нуля [5].

Повышение качества мониторинга параметров окружающей среды как деятельности оперативного характера в общем случае может быть описано циклом PDCA, включающим ряд этапов [6].

1. Планирование в широком смысле: определение параметров, имеющих наибольшую ценность при мониторинге,

и выбор для каждого из них пороговых значений, полученных в результате измерений, оценок или прогнозов; выбор методов и средств измерений для осуществления сбора, анализа, интерпретации и прогнозирования объекта наблюдения; определение оптимальных пространственных и временных интервалов между точками проведения замеров [7].

2. Осуществление прямых или косвенных измерений интересующих параметров объектов или процессов с заданными пространственно-временными интервалами по определенной методике [8].

3. Проверка результатов размеров с применением установленных методов, выявление аномальных значений и определение их природы [9, 10].

4. Корректировка пороговых значений показателей с учетом поступающих в реальном времени новых и имеющихся ретроспективных данных [11, 12].

Постановка задачи. В рамках статьи изучается влияние выбора пороговых значений стратегии принятия решений на риски возникновения ошибок первого и второго рода при выявлении аномалий, вызванных сдвигом наблюдения и повышением дисперсионности. На моделирующем стенде проводится целена-

правленный эксперимент, позволяющий определить насколько изменится достоверность функции определения аномалий при замене закона распределения входного параметра с нормального на распределение общего вида, в частности, рассматриваются экспоненциальное, Вейбулла и гамма распределения.

Материалы и методы. При распознавании аномалий определим их математическую природу как искажение поля исходного параметра с применением линейных и нелинейных операций.

Примером линейной операции может служить преобразование поля T^0 факторного пространства Fs вида $L: T_{i,j}^0 \rightarrow T_{i,j}^1 + C$, где C – некоторая константа. Нелинейное преобразование подразумевает изменение значений, составляющих поле факторного пространства с применением квадратичных, тригонометрических и других нелинейных преобразований вида $T_{i,j}^1 = N(T_{i,j}^0)^M$

$\forall T_{i,j}^0 \in Fs$ [13].

С использованием моделирующего стенда осуществляется контроль факта выхода интересующего параметра за контрольные границы, будем считать аномальными те значения T , которые выходят за диапазон $\Delta_1 - \Delta_2$. Функция выявления аномалий, в таком случае,

$$\text{примет вид } F(T_{i,j}) = \begin{cases} 0, & \Delta_1 \leq T_{i,j} \leq \Delta_2 \\ 1, & \text{иначе} \end{cases}.$$

Для любого распределения всегда найдется критический уровень, при котором будет совершена ошибка первого или второго рода. В случае нормального распределения вероятность выхода значения множества T за пределы заданных границ a, b можно выразить из функции распределения случайной величины, которая определяется в виде

$$F(T_i \leq a) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^a \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt,$$

что в общем виде можно представить в следующей форме [14]:

$$P(a \geq T_i | b < T_i) = F(a) + 1 - F(b). \quad (1)$$

Решив обратную задачу, получим границы, соответствующие обратному кумулятивному нормальному распределению с заданным процентилем. Получим аналитическое решение и представим его в виде прямых 5–95% (синяя), 10–90% (зеленая) и 25–75% (красная) (рис. 1). Для нормального распределения с параметрами $\mu = 0$, $\sigma = 1$, представим графически экспериментальные точки моделирования и их аппроксимацию, распределение вероятностей (синяя кривая) и кумулятивное распределение вероятностей (черная кривая).

Рис. 1. График эмпирического распределения мониторируемой величины и его аппроксимация нормальным распределением

Соотношение (1) справедливо в общем случае и для законов распределения общего вида.

Если рассматривать множество T как набор данных температуры, то такая случайная величина, на практике, не

подчинена нормальному закону распределения по причине того, что температура является характеристикой объекта в статике и достаточно медленно реагирует на наличие сиюминутных возмущений [15].

Проведем целенаправленный машинный эксперимент для нескольких распределений, подчиненных ряду законов распределения (нормальному, экспоненциальному, Вейбулла, гамма), с различными дисперсиями, но одинаковым математическим ожиданием, для того, чтобы показать, как изменяется достоверность определения аномалий на моделирующих алгоритмах экспериментально.

Пусть аномалией считается изменение математического ожидания рассматриваемой случайной величины на значение ε , рассчитаем вероятность ошибок первого и второго рода в зависимости от ε при заданных границах U и D .

Под ошибкой первого рода будем понимать случай, когда аномалия не произошла, но система ошибочно признала ситуацию аномальной, вероятность данной ошибки при известном законе распределения случайной величины, параметров его формы, сдвига и масштаба, определяется границами U и D при этом не зависит от ε .

Под ошибкой второго рода будем понимать случай, когда аномалия имела место быть, а система не выявила аномалию, вероятность такой ошибки при неизменном законе распределения, но при изменении математического ожидания на значение ε , определяется параметрами сдвига ε и границами U и D .

Произведем целенаправленный эксперимент для определения относительного числа ошибок первого и второго рода при выявлении аномалий в данных, имеющих нормальное распределение (рис. 2) $\mu = 3 + \varepsilon$, $\sigma = 1$, при изменении параметра ε с неизменными границами U и D , красная кривая $\mu = 3$, оранжевая $\mu = 3,5$, желтая $\mu = 4$, зеленая $\mu = 4,5$, синяя $\mu = 5$.

Вычислим значения относительного числа ошибок первого и второго рода и сведем их в (табл. 1). Для мониторируемых процессов в этой ситуации установлено, что число ошибок второго рода β существенно зависит от параметра сдвига μ , так например, увеличение μ в 5 раз приводит к уменьшению β в 10–30 раз.

Рис. 2. Графики эмпирических распределений мониторируемой величины и их аппроксимация нормальными распределениями

Таблица 1. Сводная таблица относительного числа ошибок первого и второго рода при аппроксимации мониторируемой величины экспоненциальными распределениями

Границы доверительного интервала	Относительное число ошибок первого рода, %	Относительное число ошибок второго рода, %				
	$\mu = 3$	$\mu = 3,5$	$\mu = 4$	$\mu = 4,5$	$\mu = 5$	$\mu = 5,5$
$D = 1,6 ; U = 4,6$	10	86	74	56	20	9
$D = 1,7 ; U = 4,3$	20	75	60	41	11	4
$D = 2,3 ; U = 3,7$	50	45	33	19	3	1

Ситуация совершенно изменяется, если аномалия вызвана изменением дисперсионности рассматриваемой случайной величины, когда распределение становится отличным от нормального, вследствие чего резко возрастает риск ошибок первого и второго рода. Для чистоты эксперимента пусть аномалия будет вызвана изменением закона распределения при неизменном математическом ожидании случайной величины $M=3$. Для этого рассмотрим три варианта аномалий, описываемых следующими законами распределения: экспоненциальным, Вейбулла, гамма, и произведем расчет границ для их процентилей 25% и 75%, 10% и 90 %, 5% и 95 %.

Рассмотрим первый случай, когда аномалия вызвана изменением закона распределения на экспоненциальное распределение с коэффициентом затухания $r=1/3$ и математическим ожиданием $M=3$, представленный на рис. 3. Пусть распределение случайной величины x задается плотностью вероятности $f_{\Xi}(x)$, имеющей вид

$$f_{\Xi}(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases},$$

где $\lambda > 0$ – обратный коэффициент масштаба.

Результаты расчета относительного числа ошибок первого и второго рода при возникновении аномалии изменения закона распределения с нормального на экспоненциальное, для заданных границ доверительного интервала представлены в табл. 2. Из данных таблицы видно, что β в зависимости от выбранных U и D изменяется в диапазоне 42–17%.

Таблица 2. Сводная таблица относительного числа ошибок первого и второго рода при аппроксимации мониторируемой величины экспоненциальным распределением

Границы доверительного интервала	Отн. число ошибок первого рода, %	Отн. число ошибок второго рода, %
$D=1,6; U=4,6$	10	42
$D=1,7; U=4,3$	20	32
$D=2,3; U=3,7$	50	17

Рис. 3. График эмпирического распределения мониторируемой величины и его аппроксимация экспоненциальным распределением

Рассмотрим второй случай, когда аномалия вызвана изменением закона распределения изучаемой величины на распределение Вейбулла с параметром формы $w=1,5$, коэффициентом масштаба $s \approx 3,39$ и математическим ожиданием $M=3$, представленный на рис. 4. Пусть распределение случайной величины x задается плотностью вероятности $f_B(x)$, имеющей вид

$$f_B(x) = \begin{cases} \frac{k}{\lambda} \left(\frac{x}{\lambda} \right)^{k-1} e^{-\left(\frac{x}{\lambda} \right)^k}, & x \geq 0 \\ 0, & x < 0 \end{cases},$$

где $\lambda > 0$ – коэффициент масштаба, $k > 0$ – коэффициент нормы.

Результаты расчета относительного числа ошибок первого и второго рода для распределения Вейбулла, представлены в табл. 3.

Таблица 3. Сводная таблица относительного числа ошибок первого и второго рода при аппроксимации мониторируемой величины распределением Вейбулла

Границы доверительного интервала	Отн. число ошибок первого рода, %	Отн. число ошибок второго рода, %
$D = 1,6; U = 4,6$	10	58
$D = 1,7; U = 4,3$	20	46
$D = 2,3; U = 3,7$	50	24

Рассмотрим третий случай, когда аномалия вызвана изменением закона распределения на гамма распределение, с параметром формы $s = 3$ и математическим ожиданием $M = 3$, представленный на рис. 5.

Пусть распределение случайной величины x задается плотностью вероятности $f_r(x)$, имеющей вид

$$f_r(x) = \begin{cases} x^{k-1} \frac{e^{-x/\theta}}{\theta^k \Gamma(k)}, & x \geq 0, \\ 0, & x < 0 \end{cases}$$

где $\Gamma(k)$ – гамма-функция Эйлера, $k > 0$, $\theta > 0$ – коэффициенты масштаба.

Результаты расчета относительного числа ошибок первого и второго рода для Гамма распределения представлены в табл. 4.

Анализ данных по табл. 2–4 показал, что в рассмотренных примерах наибольшую чувствительность на достоверность функции выявления аномалий $1 - \beta$ оказали аномалии, характеризующиеся изменением закона распределения входных данных на гамма распределение 67–30%, вторым по влиянию на уровень достоверности является аномалия с распределением Вейбулла 54–24% и наименьшее влияние имеют аномалии, имеющие экспоненциальное распределение 42–17%.

Таблица 4. Сводная таблица относительного числа ошибок первого и второго рода при аппроксимации мониторируемой величины гамма распределением

Границы доверительного интервала	Отн. число ошибок первого рода, %	Отн. число ошибок второго рода, %
$D = 1,6; U = 4,6$	10	67
$D = 1,7; U = 4,3$	20	55
$D = 2,3; U = 3,7$	50	30

Рис. 4. График эмпирического распределения мониторируемой величины и его аппроксимация распределением Вейбулла

Рис. 5. График эмпирического распределения мониторируемой величины и его аппроксимация гамма распределением

Заключение. В результате проведения целенаправленного эксперимента на моделирующем стенде с целью определения степени достоверности для оценки рисков возникновения аномалий при мониторинге, имеет смысл выделить следующие результаты.

1. Выбор оптимального доверительного интервала U и D происходит в пространстве компромиссов. При малых значениях сдвига ε оптимальным решением будет выбор узкого доверительного интервала U и D , а при больших, соответственно, широкого.

2. Выбор границ доверительного интервала U и D однозначно определяет коэффициент ошибок первого рода α и поэтому может быть рассчитан заранее.

3. При детерминированном законе распределения и границах доверительного интервала U и D параметр сдвига ε определяет коэффициент ошибок второго рода β . В рассмотренном примере $\alpha = 10\% \rightarrow \beta = (86...9)\% \mid \varepsilon = (0,5...2,5)\sigma$.

4. Сужение доверительного интервала U и D приводит к значительному увеличению числа ошибок первого рода α до 25% и 50% и практически пропорциональному снижению числа ошибок второго рода 74–4% и 45–1% соответственно.

5. При определении аномальностей, вызванных изменением дисперсионности входных данных, для того, чтобы коэффициент ошибок α и β был одного порядка, следует сужать границы доверительного интервала U и D так, чтобы вызвать двукратное увеличение коэффициента ошибок первого рода α . Эта рекомендация актуальна для всех рассмотренных законов распределения.

Учет ошибок α и β при различных значениях сдвига и дисперсионности позволяет повысить достоверность при обнаружении аномальностей, таким образом, значительно повышается качество принятия решений, если принять меры по компенсации этих ошибок, вследствие чего надежность принятия решений увеличивается.

Исследование выполнено при финансовой поддержке РФФИ и города Севастополь в рамках научных проектов № 18-47-920005\18, № 18-47-920007\18.

СПИСОК ЛИТЕРАТУРЫ

1. *Anomalies of the upper water column in the Mediterranean Sea* / I. Rivetti, F. Boero, S. Frascetti [et al.] // *Global and Planetary Change*. 2017. Vol. 151. P. 68–79. DOI: 10.1016/j.gloplacha.2016.03.001

2. Шишкин Ю.Е., Скатков А.В. Метод распознавания в интерактивном режиме аномалий градиентов скалярных полей наблюдений // *Системы контроля окружающей среды*. 2018. № 12 (32). С. 30–37.

3. Лепихин А.П., Возняк А.А. Статистические функции распределения гидрохимических показателей качества воды поверхностных водных объектов // *Водное хозяйство России: проблемы, технологии, управление*. 2012. № 4. С. 21–32.

4. Zinchenko V., Falko V. Mathematical model and solution method elaboration for the task of ecological risk predictive assessment // *Екологічна безпека*. 2013. № 2 (16). С. 36–39.

5. Скатков А.В., Шишкин Ю.Е. Модель обнаружения аномалий в наблюдениях параметров полей окружающей среды с использованием систем мониторинга // *Системы контроля окружающей среды*. 2017. № 10 (30). С. 48–53.

6. A PDCA-based approach to Environmental Value Stream Mapping (E-VSM) / J.A. Garza-Reyes, J.T. Romero, K. Govindan [et al.] // *Journal of Cleaner Production*. 2018. Vol. 180. P. 335–348. DOI: 10.1016/j.jclepro.2018.01.121.

7. Шишкин Ю.Е., Скатков А.В. Акторная модель мониторинга с использованием мобильных облачных микросервисов // *Системы контроля окружающей среды*. 2018. № 14 (34). С. 56–62. DOI: 10.33075/2220-5861-2018-4-56-62.

8. Шишкин Ю.Е., Греков А.Н. Концепция интеллектуальной системы автоматизированного экологического мониторинга севастопольской бухты на базе малогабаритных автономных роботов // Системы контроля окружающей среды. 2018. № 14 (34). С. 63–69. DOI: 10.33075/2220-5861-2018-4-63–69.
9. Скатков А.В., Шишкин Ю.Е. Модель обнаружения аномалий в наблюдениях параметров полей окружающей среды с использованием систем мониторинга // Системы контроля окружающей среды. 2017. № 10 (30). С. 48–53.
10. Vychuzhanin P., Hvatov A., Kalyuzhnaya A.V. Anomalies Detection in Metocean Simulation Results Using Convolutional Neural Networks // Procedia Computer Science. 2018. Vol. 136. P. 321–330. DOI: 10.1016/j.procs.2018.08.282
11. Comparison of atmospheric particle concentration measurements using different optical detectors: Potentiality and limits for air quality applications / A. Dinoi, A. Donato, F. Belosi [et al.] // Measurement. 2017. Vol. 106. P. 274–282. DOI: 10.1016/j.measurement.2016.02.019
12. Брюховецкий А.А., Скатков А.В., Шишкин Ю.Е. Моделирование процессов обнаружения аномалий в сложно-структурированных данных мониторинга // Системы контроля окружающей среды. 2017. № 9 (29). С. 45–49.
13. Харман Г. Современный факторный анализ. М.: Статистика. 1972. 489 с.
14. Семенов В.А. Теория вероятностей и математическая статистика: Учебное пособие. Стандарт третьего поколения. СПб.: Питер. 2013. 192 с.
15. Shishkin Y.E. Big Data visualization in decision making // Science in Progress: тез. Всерос. науч.-практ. конф. магистрантов и аспирантов. Новосибирск, 20 октября 2016 г. Новосибирск: НГТУ, 2016. С. 203–205.

INCREASING THE RELIABILITY OF RISK ASSESSMENTS IN THE MONITORING PROCESSES DESCRIBED BY GENERAL DISTRIBUTIONS

Y.E. Shishkin^{1,2}, A.V. Skatkov¹

¹ Federal State Educational Institution of Higher Education «Sevastopol State University», Russian Federation, Sevastopol, Universitetskaya St., 33

² Institute of Natural and Technical Systems, Russian Federation, Sevastopol, Lenin St., 28

The influence of the decision-making strategy of threshold values on the occurrence of errors of the first and second kind when identifying anomalies in the process of monitoring of environmental parameters, which are described in practice by general type distributions, was studied. During the purposeful experiment, anomalies of two classes were considered: the observation shift and the increase in dispersion. The sensitivity of the detecting statistical method was assessed. Practical recommendations on choosing the threshold values of the decision support system for anomalies that take into account the distribution of the observed random variable were formulated

Keywords: monitoring, mathematical modeling, anomaly detection, clustering, critical systems, data mining, general distribution.