УДК 004.942:004.75

МОДЕЛЬ ОБНАРУЖЕНИЯ АНОМАЛИЙ В НАБЛЮДЕНИЯХ ПАРАМЕТРОВ ПОЛЕЙ ОКРУЖАЮЩЕЙ СРЕДЫ С ИСПОЛЬЗОВАНИЕМ СИСТЕМ МОНИТОРИНГА

А.В. Скатков[1], Ю.Е. Шишкин[1,2]

[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: yourockpro@gmail.com

Предлагается подход к решению задачи оперативного прогнозирования перехода объекта мониторинга в аномальное состояние за счет использования системы параметрического мониторинга, реализованной в виде комплекса измерительных средств и программной имитационной модели с использованием информационной метрики дивергенции Кульбака-Лейблера. Обсуждаются результаты экспериментов, поставленных на имитационной модели системы мониторинга физико-химических параметров водной среды черноморского региона.
Ключевые слова: мониторинг, имитационное моделирование, система массового обслуживания, Большие Данные, эффект гетероскедастичности, сетевой трафик, критические системы, интеллектуальный анализ данных.

Введение. В настоящее время процесс контроля физико-химических параметров водной среды и критических объектов антропогенной инфраструктуры, связанных с ней, является актуальной задачей [1]. На практике приходится сталкиваться с отсутствием системного подхода при проведении комплексного мониторинга и оперативной реакции на возникающие внештатные ситуации или угрозу их возникновения [2].

Постановка задачи. Моделируется чувствительность системы мониторинга по обнаружению аномалий (управляемых возмущений). Для решения поставленной задачи требуется разработать программную имитационную модель потоков данных системы автономных STD зондов, производящих мониторинг физико-химических параметров в акватории черноморского региона с целью проверки выдвинутой гипотезы об эффективности выявления аномалий (возмущений математического ожидания $w \geq 1{,}3$ закона распределения наблюдаемой величины) в зависимости от порога α, в полях наблюдений с использованием системы параметрического мониторинга, применяющей в качестве информационной метрики дивергенцию Кульбака-Лейблера [3].

Осуществление оперативного мониторинга с последующей интеграцией данных в систему прогнозирования, реализованную в виде имитационной модели, отражающей реакцию объекта мониторинга, позволяет с минимальными финансовыми затратами предотвращать аварийные ситуации [4]. Состояние объекта мониторинга может быть охарактеризовано в виде целой совокупности свойств, например, в виде кортежа, определяющих возможность его нормального функционирования [5].

Мониторинг быстропротекающих процессов критических систем, способных вызвать переход системы в аварийное состояние или другие необратимые воздействия, представляет собой особую сложность. Особенностью работы системы поддержки принятия решений (СППР) по детектированию процессов такого рода является необходимость принятия решений в условиях малых объемов выборок, характеризующих интересующий процесс [6]. С математической точки зрения это приводит к необходимости повысить точность и достоверность принимаемых решений при недостаточном количестве информации за счет использования предварительного обучения на множестве ретроспектив-

ных данных и прогнозирующей параметрической имитационной модели [7].

Материалы и методы. Моделирования процесса динамики распространения водной массы черноморского региона, Основного черноморского течения, основано на данных [8], осуществление замеров и внесение возмущений производилось с использованием агентной, дискретно-событийной и пешеходной моделей (рис. 1) среды имитационного моделирования общего назначения Anylogic Personal Learning Edition 7.2, расчет системы информационных метрик по данным параметрического эксперимента, проводился в Mathcad 15.

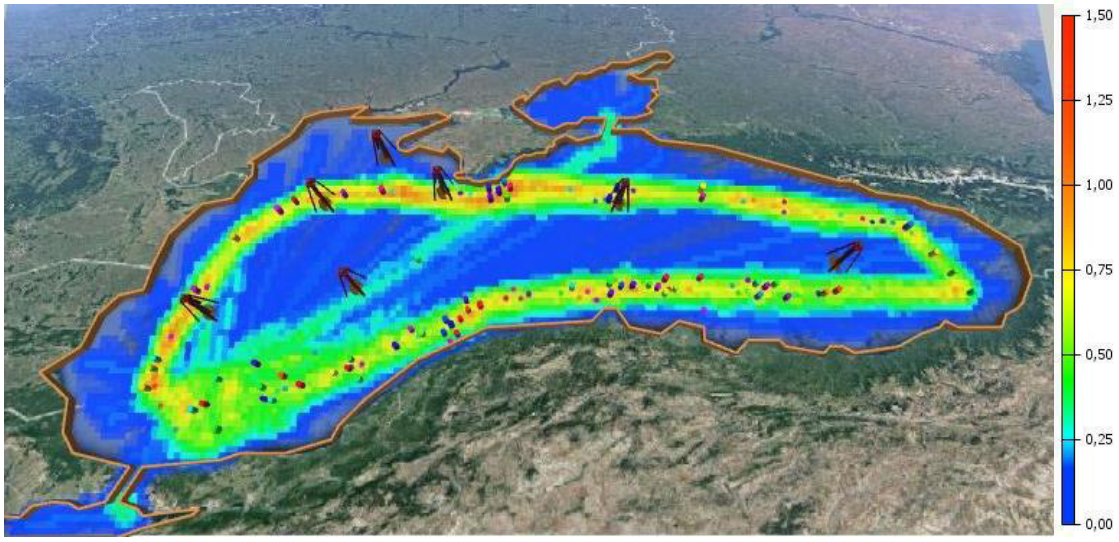

Рис. 1. Тепловая карта плотности распределения агентов состояния водной среды *A* и сеть автономных зондов *C* в программной имитационной модели

Информационной метрикой, позволяющей численно определить расхождение между распределениями параметров мониторинга, служит дивергенция Кульбака-Лейблера:

$$Dif(a \parallel b) = \int a(x) \log \frac{a(x)}{b(x)} dx,$$

где a и b – плотности распределений, мера $Dif(a \parallel b)$ определена в том случае если $\int_{x:a(x)=0} b(x)dx \le 0$, другими словами $a(x)$ полностью покрывает $b(x)$. Известно, что $Dif(a \parallel b) \ge 0$ для любой пары распределений и $Dif(a \parallel b) = 0$ в том случае, если плотности распределений a и b совпадают почти на всей области определения [9]. Докажем данное свойство рассмотрев строго вогнутую функцию $f(x)$:

$$f(\alpha x_1 + (1-\alpha)x_2) \ge \alpha f(x_1) + (1-\alpha) f(x_2), \forall x_1, x_2, \alpha \in [0.1],$$

для произвольного числа точек x имеет место неравенство Йенсена:

$$f(\alpha_1 x_1 + \ldots + \alpha_N x_N) \ge \alpha_1 f(x_1) + \ldots + \alpha_N f(x_N), \sum_{n=1}^{N} \alpha_n = 1, \alpha_n \ge 0,$$

которое может быть доказано по индукции в интегральной форме:

$$f\left(\int \alpha(y) x(y) dy\right) \ge f \int \alpha(y) f(x(y)) dy, \int \alpha(y) dy = 1, \alpha(y) \ge 0, \forall y,$$

подставив значения $\alpha(y) = b(y), x(y) = \frac{a(y)}{b(y)}$, получим

$$0 = \log\left(\int q(y)\frac{a(y)}{b(y)}dy \right) \geq \int q(y)\log\frac{a(y)}{b(y)}dy = -Dif(a \parallel b),$$

таким образом, равенство достигается только в случае $a(y) \equiv b(y)$, в противном случае, если $b(y) \neq 0$, неравенство выполняется всегда.

В литературе понятие дивергенции Кульбака-Лейблера интерпретируется как мера оценки потерянной информации при замене плотности распределения a на b [10].

Построение модели обнаружения аномалий на основе метрики Кульбака-Лейблера. Для формулирования математической модели определим базовую структуру, полностью характеризующую состояние объекта мониторинга в точке S следующим образом:

$$S = \left\langle T_n, K_n, \vec{F}(|m|, d), G, C \right\rangle,$$

где $T, K, \vec{F}, G$ – мгновенные характеристики водной среды, такие как температура, соленость, направление вектора течения и электропроводность, а кластер C представляет набор химических показателей окружающей среды:

$$C = \left\langle O_2, pH, PO_4, NO_3, NO_2 \right\rangle.$$

Вектор течения $\vec{F}(|m|, d)$ задается в виде модуля скорости течения m и направления d.

В терминах теории массового обслуживания систему обработки данных мониторинга можно рассматривать как систему массового обслуживания (СМО) типа G/G1/M/N. Обнаружение аномалий в данных мониторинга производится на основе метрики Кульбака-Лейблера, являющийся несимметричной мерой удалённости друг от друга двух вероятностных распределений, определённых на общем пространстве элементарных событий [11].

В достаточно общем виде модель системы мониторинга, с требуемой точностью отражающей функционал, обеспечиваемый сетью автономных зондов, производящих мониторинг физико-химических параметров в акватории черноморского региона может быть представлена в виде:

$$M_m = \left\langle Z, S, Dif, P \right\rangle,$$

где Z – множество зондов с определенным начальным состоянием и расположением в пространстве моделируемой среды, расположение которых показано на (рис. 2); S – структура полностью характеризующая состояние объекта мониторинга; Dif – информационная метрика, производящая численное сравнение закона распределения фактически поступающего потока данных и ожидаемого согласно имеющейся статистики; P – динамическая модель водной среды, в которой располагается множество Z.

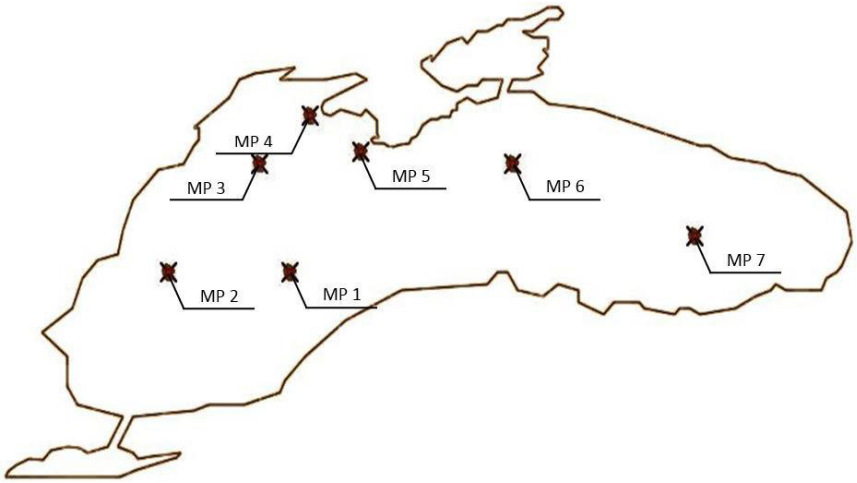

Рис. 2. Расположение гидрологических зондов при проведении имитационного эксперимента

Методика постановки экспериментов. В стационарный поток данных введем возмущение в виде изменения в w раз значения математического ожидания закона распределения наблюдаемой величины. По значению информационной метрики и заданному уровню значимости СППР определяет наличие возмущения в потоке данных, происходит сравнение полученного решения с фактическим состоянием системы и делается вывод об истинности гипотезы эффективности применения информационной метрики дивергенции Кульбака-Лейблера для данного класса задач.

Рассмотрим модель СМО со следующими характеристиками:

– интенсивность генерации заявок M(λ)=1000, σ(λ)=10;

– интенсивность обслуживания заявок M(μ)=[1100; 1300; 1500; 1800; 2000] и σ(μ)=10;

– емкость накопителя ограничена площадью моделируемого пространства;

– диаметр моделируемого атомарного среды агента 5 px.

На рис. 3 представлена динамика потока данных мониторинга в СМО, построенная в системе имитационного моделирования Anylogic.

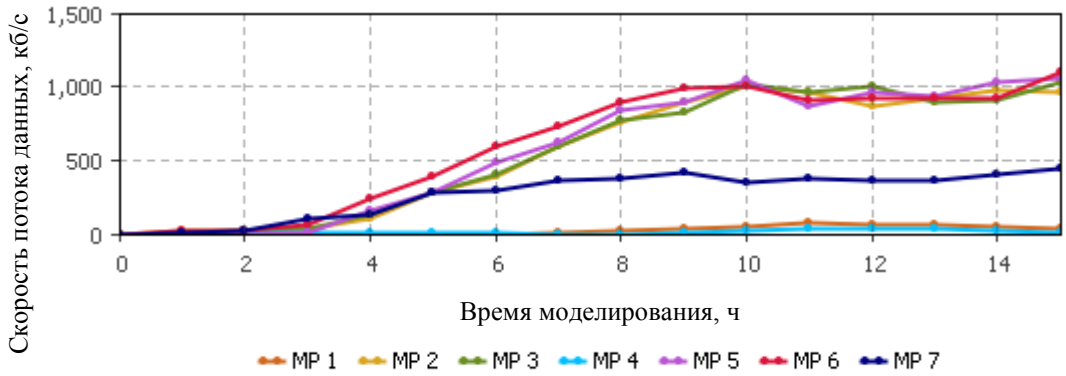

Рис. 3. Диаграмма динамики изменения потока данных мониторинга, передаваемого системой гидрологических зондов, построенная по данным результата имитационного эксперимента

По данным динамики потока данных мониторинга очевидно, что в период моделирования 0 – 10 ед. модельного времени система находится в нестационарном состоянии, а затем происходит переход в стационарный режим, характеризующийся ярко выраженной периодичностью.

Подтвердим данную гипотезу, представив стационарный участок динамики изменения потока данных в виде дискретной функции $f(t)$ длительностью T заданной на отрезке $\{0,T\}$, где t – время моделирования, в виде суммы гармонических функций вида:

$$f(t)=\frac{a_0}{2}+\sum_{k=1}^{+\infty}A_k\cos\left(2\pi\frac{k}{\tau}x+\theta_k\right),$$

где k – номер гармонической составляющей; T – отрезок, где функция определена; A_k – амплитуда k-ой гармонической составляющей; θ_k – начальная фаза k-ой гармонической составляющей.

Для этого выполним дискретное преобразование Фурье с использованием алгоритма быстрого преобразования Фурье для заданного множества динамик изменения потока данных мониторинга зондов (рис. 4). Спектры данных функций на отрезке T совпадают.

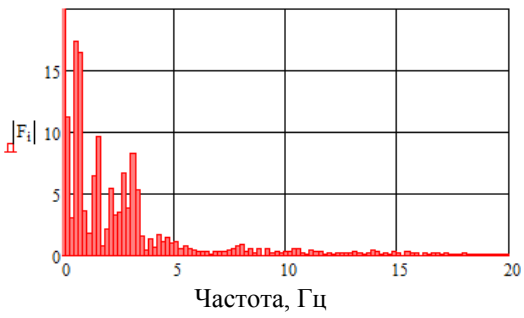

Рис. 4. Спектр функции динамики изменения потока данных мониторинга

Для обеспечения корректности обработки введем требование к частоте взятия отсчетов непрерывного во времени сигнала при его дискретизации в имитационной модели. Согласно теореме Котельникова, если непрерывный сигнал имеет спектр, ограниченный частотой $F_{\max}$, то он может быть полностью и однозначно восстановлен по его дискретным отсчетам, взятым через интервалы времени $T = \frac{1}{2F_{\max}}$, где $F_d = \frac{1}{T}$ – частота дискретизации; $F_{\max}$ – максимальная частота спектра сигнала. Значение частоты дискретизации $F_d = 50$ Гц.

Для детектирующей системы вводится нулевая гипотеза H_0 о том, что поступил сигнал о возникновении события, но при этом событие A не произошло (ложное срабатывание), пусть вероятность этого события ε, и вводится конкурирующая ей гипотеза H_1. Таким образом, рассматривается детектирующая система, характеризующаяся вероятностями обнаружения события A в поле значений (ε, δ). Вероятность обнаружения зависит от относительной величины возмущения (рис. 5) и порога (рис. 6).

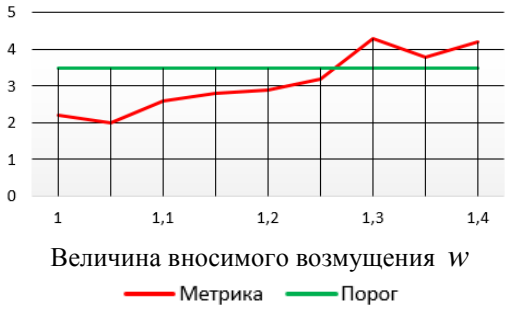

Рис. 5. Диаграмма зависимости информационной метрики дивергенции Кульбака-Лейблера в зависимости от величины искусственно внесенного возмущения w, при пороговой величине $\alpha = 0{,}001$

Получим четыре условных вероятности, описывающие все возможные варианты событий: $P(H_0|A) = \varepsilon$ – событие не произошло, но появился сигнал о возникновении события, ложное срабатывание; $P(H_1|A) = \varepsilon - 1$ – событие произошло и было обнаружено; $P(H_0|\overline{A}) = 1 - \delta$ – событие не произошло и сигнала о нем не поступило; $P(H_1|\overline{A}) = \delta$ – событие произошло, но не было обнаружено.

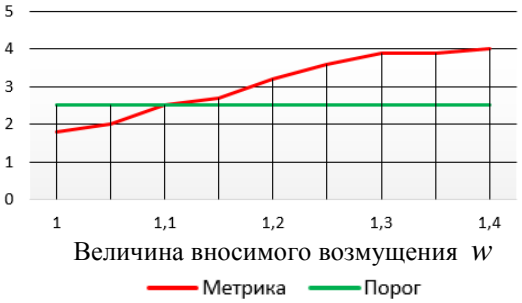

Рис. 6. Диаграмма зависимости информационной метрики дивергенции Кульбака-Лейблера в зависимости от величины искусственно внесенного возмущения w, при пороговой величине $\alpha = 0{,}01$

Заключение. Предложена программная имитационная модель потоков данных системы автономных STD зондов, на которой экспериментально определена чувствительность системы мониторинга по обнаружению управляемых возмущений. На основе модели можно проводить управляемые эксперименты по обнаружению порогов чувствительности и детектирующей способности метода, находить области, где мера Кульбака имеет преимущество при обнаружении аномалий по сравнению с другими мерами, в частности: математического ожидания, критерия Стьюдента, Фишера, коэффициента вариации и др.

Установлены устойчивые области распознавания аномалий $P(H_1|A)$ с достоверностью 0,001 при мониторинге $w \geq 1{,}3$, $w \leq 0{,}8$. Зона уверенного нераспознавания аномалий $P(H_1|\overline{A})$ с достоверностью 0,001 составила $0{,}9 \leq w \leq 1{,}2$. Имитационная модель используется для нахождения пограничных зон неуверенного распознавания аномалий, для работы в системах принятия решений, поскольку целью мониторинга является не только сбор информации об объекте, но и обнаружение критических событий для оперативного принятия неотложных мер по ликвидации критических состояний.

Работа выполнена при частичной поддержке Российского фонда фундаментальных исследований (грант № 15-29-07936/17).

СПИСОК ЛИТЕРАТУРЫ

1. *Брюховецкий А.А., Скатков А.В., Шишкин Ю.Е.* Моделирование процессов обнаружения аномалий в сложноструктурированных данных мониторинга // Системы контроля окружающей среды. Севастополь: ИПТС. 2017. Вып. 9 (29). С. 45–49.

2. *Греков А.Н., Шишкин Ю.Е.* Моделирование трехкомпонентного акустического измерителя скорости течения // Системы контроля окружающей среды. Севастополь: ИПТС. 2016. Вып. 6 (26). С. 33–40.

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

4. *Шишкин Ю.Е., Скатков А.В.* Решение задачи составления расписаний большой размерности с применением технологии Больших Данных // Информационные технологии и информационная безопасность в науке, технике и образовании "ИНФОТЕХ – 2015": материалы междунар. науч.-практ. конф. / под науч. ред. А.В. Скаткова. (г. Севастополь, 7–11 сентября 2015 г.). Севастополь: СевГУ, 2015. С. 103–105.

5. *Боев В.Д.* Концептуальное проектирование систем в Anylogic 7 и GPSS World. М.: НОИ Интуит, 2016. 556 с. ISBN: 978-5-9556-0161-8.

6. *Девятков В.В.* Методология и технология имитационных исследований сложных систем: современное состояние и перспективы развития: монография. СПб.: Вузовский учебник, 2013. 448 с.

7. *Клейнрок Л.* Теория массового обслуживания / пер. с англ. И.И. Грушко / под ред. В.И. Неймана. М.: Машиностроение, 1979. 432 с.

8. *Кузьминская Г.Г.* Черное море. Краснодар. 1988. 95 с.

9. *Шишкин Ю.Е.* Анализ моделей взаимодействия пользователей и провайдеров облачных сервисов // Интеллектуальные системы, управление и мехатроника – 2016: материалы всерос. науч.-техн. конф. молодых ученых, аспирантов и студентов (г. Севастополь, 19–21 мая 2016 г.). Севастополь: СевГУ, 2016. С. 289–293.

10. *Вентцель Е.С., Овчаров Л.А.* Теория вероятностей и ее инженерные приложения. М.: Высшая школа, 2007. 491 с.

11. *Скатков А.В., Брюховецкий А.А., Шишкин Ю.Е.* Сравнительный анализ методов обнаружения изменений состояний сетевого трафика // Автоматизация и приборостроение: проблемы, решения: материалы междунар. науч.-техн. конф. (г. Севастополь, 05–09 сентября 2016 г.). Севастополь: СевГУ, 2016. С. 14–15.

ANOMALY IDENTIFICATION MODEL IN THE OBSERVATION FIELD USING PARAMETRIC MONITORING SYSTEMS

A.V. Skatkov[1], Y.E. Shishkin[1,2]

[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 problem of detecting anomalies in monitoring data of aquatic environment physicochemical parameters using simulation model of autonomous STD probes system is discussed. An approach is proposed to solve the problem of monitoring object transition operative forecasting to an emergency condition by using a parametric monitoring system implemented as a measuring tools complex and a software simulation model using the Kulbak-Leibler divergence information metric. The solution of the sample size influence on the probability of a classification error problem is given. The results of experiments on the simulation model of a monitoring system for physicochemical parameters of the Black Sea region water environment are discussed.

Keywords: monitoring, simulation, queuing system, Big Data, heteroscedasticity effect, network traffic, critical systems, data mining.