https://elibrary.ru/wqyfbc
УДК 004.89; 519.246.8
DOI: 10.33075/2220-5861-2023-3-141-147

ПРИМЕНЕНИЕ МОДЕЛИ ARIMA ДЛЯ ОБНАРУЖЕНИЯ АНОМАЛИЙ В РЯДАХ АКТИВНОСТИ ДВУСТВОРЧАТЫХ МОЛЛЮСКОВ

Е.В. Вышкваркова, А.Н. Греков, А.С. Маврин, В.В. Трусевич

Институт природно-технических систем,
РФ, г. Севастополь, ул. Ленина, 28
E-mail: aveiro_7@mail.ru

Использование двустворчатых моллюсков в качестве биоиндикаторов в системах автоматизированного мониторинга водной среды позволяет в режиме реального времени обнаружить чрезвычайную ситуацию, связанную с загрязнением водной среды. Применение алгоритмов машинного обучения позволяет быстро обнаружить аномалию для последующего формирования сигнала тревоги. Модель ARIMA с сезонной составляющей применена для прогнозирования данных активности моллюсков и обнаружения аномалий. Результаты показали возможность использования модели ARIMA с сезонной составляющей для обнаружения аномалий, что позволяет в будущем интегрировать разработанный алгоритмический подход в программное обеспечение биологических систем раннего обнаружения.
Ключевые слова: аномалии, биомониторинг, прогноз, ARIMA.

Поступила в редакцию: 15.08.2023. После доработки: 01.09.2023.

Введение. Надежная оценка качества воды и точный прогноз показателей загрязнения воды являются ключевыми звеньями в управлении водными ресурсами и борьбе с загрязнением воды. В современных условиях в водную среду постоянно поступают сотни тысяч веществ, и большинство из них оказывает токсическое воздействие на водные экосистемы и человека. Существующие системы мониторинга воды, основанные в основном на физико-химическом анализе, в принципе не способны обнаруживать весь спектр загрязняющих веществ. Только живые организмы способны мгновенно дать интегральную оценку состояния водной среды [1, 2]. Именно биологические методы мониторинга вод, так называемые биологические системы раннего оповещения (Biological Early Warning Systems – BEWS) наиболее перспективны для оценки состояния качества водной среды. BEWS используются по всему миру для постоянного наблюдения за состоянием водной среды. В качестве биосенсоров используют аборигенные виды двустворчатых моллюсков [3], ракообразных [4], рыб [5], водоросли [6]. В качестве реакций живых организмов на неблагоприятное воздействие окружающей среды используется широкий спектр биомаркеров: молекулярные, физиологические, поведенческие реакции и др. [2]. Поведенческие маркеры представляют собой реакции организмов на внутренние (физиологические) и внешние (окружающие и социальные) факторы.

В 2008 г. был разработан [7], а затем модернизирован [8] комплекс автоматизированного биомониторинга водной среды. Работа комплекса основана на фиксации и анализе поведенческих реакций моллюсков и формировании сигнала тревоги при обнаружении аномалий. Величина открытия створок двустворчатых моллюсков, особенности ритма их движений характеризуют фильтрационную активность, а, следовательно, и уровень их жизнедеятельности в нормальных и токсических средах. Аномалии в данных активности моллюсков (или других организмов, используемых в системах мониторинга вод) возникают при реакции на загрязнения или по техническим причинам. Основными причинами аномалий в данных могут быть: неисправная система (например, отсутствие связи с сервером, вычисли-

тельные ошибки или небрежная запись); плохое состояние водной среды.

Системы обнаружения аномалий широко используются в самых разных приложениях, таких, как обнаружение мошенничества, обнаружение вторжений для обеспечения кибербезопасности, анализ производительности и обнаружение неисправностей.

Существуют различные типы аномалий, поэтому правильный способ их распознавания сильно зависит от приложения. В нашем проекте аномалии очень тесно связаны с задачей прогнозирования временных рядов, поскольку аномалии выявляются на основе отклонений от прогнозируемого значения.

В нашей работе для построения прогноза выбрана модель авторегрессионного интегрированного скользящего среднего (autoregressive integrated moving average – ARIMA) с сезонной составляющей (seasonal ARIMA – SARIMA). Это хорошо известный метод прогнозирования временных рядов, также называемый моделью Бокса-Дженкинса или методом Бокса-Дженкинса [9]. Модели ARIMA основаны на теории стохастических процессов и характеризуются меньшими требованиями к данным, простой структурой и быстрым моделированием. Прогнозирование временных рядов моделями ARIMA используется в различных областях, включая прогнозы, связанные с биологическими процессами и оценкой качества воды в естественных водоемах. Примером является использование модели ARIMA для прогнозирования суточной концентрации хлорофилла *a* в озере Тайху (Китай) для прогнозирования цветения водорослей с помощью онлайн-датчиков [10], или прогнозирование качества речной воды и гидрологических переменных в реке Джохор (Малайзия) [11].

Цель работы – обнаружение аномалий в данных активности двустворчатых моллюсков алгоритмами прогнозирования машинного обучения для последующего включения в программное обеспечение комплекса автоматизированного биомониторинга водной среды. Новизна исследования заключается в использовании модели ARIMA для анализа и прогнозирования активности двустворчатых моллюсков комплекса биомониторинга водной среды.

Материалы и методы. В работе использованы данные активности пресноводных двустворчатых моллюсков *Unio pictorum* (Linnaeus, 1758) за период с 26 февраля по 24 апреля 2017 г. Размер моллюсков 40–45 мм, отловлен в районе работ. Комплекс биомониторинга был установлен на гидроузле № 14 реки Черной (г. Севастополь). Данные активности 16 моллюсков одновременно фиксируются и передаются на сервер.

Модель SARIMA использована для прогнозирования временных рядов активности двустворчатых моллюсков. Модель ARIMA (p, d, q) имеет 3 компонента: «p» – порядок авторегрессионной части, «q» – порядок части скользящего среднего, а «d» – порядок взятия последовательной разности, необходимый для того, чтобы сделать ряд стационарным [12].

В параметрах моделей SARIMA необходимо указать два типа параметров. Первая аналогична модели ARIMA (p, d, q), а вторая предназначена для уточнения влияния сезонности (сезонного порядка): P – порядок сезонной составляющей SAR(P); D – порядок интегрирования сезонной составляющей; Q – порядок сезонного компонента SMA(Q), а m – размерность сезонности (месяц, квартал и т. д.) [13].

Для оценки качества прогностических моделей использованы две метрики:

– MAPE (mean absolute percentage error) – средняя абсолютная процентная ошибка используется в качестве статистического индекса для измерения точности прогноза:

$$MAPE = \frac{1}{n}\sum\left(\left|\frac{\widehat{y_i} - y_i}{y_i}\right|\right) * 100\%$$

– RMSE (Root mean squared error) представляет собой квадратный корень из среднеквадратической ошибки (MSE) и масштабирует значения MSE до диапа-

зонов наблюдаемых значений. Оценивается по уравнению

$$RMSE = \sqrt{\frac{\sum(|\hat{y}_i - y_i|)^2}{n}},$$

где y_i и $\hat{y}_i$ являются фактическими и прогнозируемыми значениями, а n – количество выборок.

Анализ данных проводился на языке программирования Python (V3.9.12) с использованием пакета машинного обучения scikit-learn (V 1.2.2) [14] и пакета статистических моделей statsmodels (V 0.14.0) [15].

Результаты. Для разработки модели использовано среднее арифметическое значение величины раскрытия створок всех функционирующих мидий (в нашем случае 14, две мидии в ходе эксплуатации вышли из строя). Выбор подходящих параметров модели крайне важен для обеспечения оптимальной настройки параметров модели ARIMA. Для этого весь набор данных (за исключением аномалий) был разбит на двухдневные интервалы со сдвигом в один час. В пределах каждого интервала, за исключением последнего часа, модели обучались с использованием различных комбинаций параметров, как показано в таблице 1. Параметр «m», который соответствует количеству точек данных за период (сезон), в нашем случае установлен равным 144. Это значение представляет собой день наблюдений с 10-минутным усреднением, учитывающим четкий суточный характер активности моллюсков [16].

Для уменьшения количества возможных вариантов параметров нашей модели оценим наш ряд на стационарность. Это позволит определить минимальное значение параметра d модели ARIMA. Один из способов проверить, является ли временной ряд стационарным, — это выполнить расширенный тест Дики-Фуллера (ADF), в котором используются следующие нулевая и альтернативная гипотезы:

H0 – временной ряд является нестационарным. Другими словами, он имеет некоторую структуру, зависящую от времени, и не имеет постоянной дисперсии во времени;

H1 – временной ряд является стационарным.

Тестирование наших данных показало следующие результаты:

Критерий ADF: -1.42. P-значение: 0.57. Критические значения: 1%: -3.45, 5%: -2.87, 10%: -2.57.

Так как p-значение >0,05 и критерий ADF больше критических значений, то нулевая гипотеза не может быть отвергнута и наш ряд не является стационарным. Таким образом, параметр d нашей модели должен быть минимум первого порядка.

Для оценки качества модели рассчитаны показатели RMSE и MAPE путем сравнения прогнозируемых значений с фактическими значениями за последний час каждого двухдневного интервала. Средние арифметические показателей RMSE и MAPE для каждого набора параметров были рассчитаны для облегчения сравнения моделей по всему набору данных. Наименьшие значения ошибок и соответствующие им оптимальные значения параметров ARIMA отмечены красным цветом в таблице 1.

Оценим возможность обнаружения аномалий на основе поиска относительно больших отклонений нашей целевой переменной от прогнозируемых значений. Наши данные содержат 3 дня с аномалиями, выявленных экспертами по анализу данных. Результат моделирования показан на примере аномалии 24 апреля 2017 года. На рис. 1а приведены графики движения створок всех работающих мидий в этот период, на которых четко прослеживается синхронная реакция мидий на стрессовую для них ситуацию. Точное время (минуты и секунды) возникновения аномалии неизвестны [17]. С использованием оптимальной модели SARIMA построен прогноз целевой переменной (рис. 1б) и рассчитаны показатели MAPE (0,5634%) и RMSE (0,1468). На рис. 1б видно, что во время аномального случая фактическое значение нашей переменной значительно отклоняется от прогнозируемого значения. Более того, это отклонение превышает

95% доверительный интервал прогноза нашей модели. Кроме того, значения метрик RMSE и MAPE в несколько раз превышают средние значения, полученные при настройке модели.

Таблица 1. Параметры модели SARIMA и соответствующие ошибки

№	order(p,d,q)			seasonal_order(P,D,Q,m)				RMSE	MAPE (%)
	p	d	q	P	D	Q	m		
0	0	1	0	0	1	0		0,225993	0,050045
1	0	1	0	0	1	1		0,225993	0,050045
2	0	1	0	0	1	2		0,225993	0,050045
3	0	1	0	0	2	0		0,65936	0,12864
4	0	1	0	0	2	1		0,659361	0,12864
5	0	1	0	0	2	2		0,659361	0,12864
6	0	1	0	1	1	0		0,131594	0,023813
7	0	1	0	1	1	1		0,131594	0,023813
8	0	1	0	1	1	2		0,131594	0,023813
9	0	1	0	1	2	1		0,576546	0,110556
10	0	1	0	1	2	2		0,573239	0,10991
11	0	1	0	2	1	0		0,133435	0,024139
12	0	1	0	2	1	1		0,133435	0,024139
13	0	1	0	2	1	2		0,133435	0,024139
14	0	1	0	2	2	1		0,449988	0,084233
15	0	1	0	2	2	2	144	0,449988	0,084233
16	0	1	1	0	1	0		0,130097	0,023523
17	0	1	1	0	1	1		0,130075	0,023518
18	0	1	1	0	1	2		0,131594	0,023813
19	0	1	1	0	2	0		0,722864	0,143018
20	0	1	1	0	2	1		1,362772	0,321766
21	0	1	1	0	2	2		1,362772	0,321766
22	0	1	1	1	1	0		0,130088	0,023512
23	0	1	1	1	1	1		0,130064	0,023507
24	0	1	1	1	1	2		0,133435	0,024139
25	0	1	1	1	2	0		0,619253	0,119797
26	0	1	1	1	2	1		0,575843	0,110401
27	0	1	1	1	2	2		0,573239	0,10991
28	0	1	1	2	1	0		0,133435	0,024139
29	0	1	1	2	1	1		0,133435	0,024139
30	0	1	1	2	1	2		0,133435	0,024139
31	0	1	1	2	2	0		0,449988	0,084233

Заключение. Обнаружение аномалий в ряду данных активности двустворчатых моллюсков является ключевым моментом для формирования тревожного сигнала для автоматизированной системы контроля водной среды. В данном исследовании мы применили модель ARIMA с сезонной составляющей для выявления аномалий в данных об активности моллюсков. Данные получены с помощью разработанного авторами автоматизированного комплекса биомониторинга водной среды в 2017 г. на гидроузле №14 р. Черная (г. Севастополь). Наименьшие ошибки RMSE (0,130064) и MAPE (0,023506%) были получены для модели ARIMA порядка (p, d, q) = (0, 1, 1) и сезонного порядка season_order (P, D, Q, m) = (1, 1, 1). Результаты исследования показывают практичность использования модели ARIMA с сезонной составлявшей для прогнозирования активности двустворчатых моллюсков, что позволяет получать сигналы тревоги в режиме реального времени. Этот алгоритмический подход может быть легко интегрирован в программное обеспече-

ние биологических систем раннего обнаружения. Исследование представляет собой важный вклад в область биомониторинга и разработки алгоритмов обнаружения аномалий, способствуя более эффективной оценке состояния водных экосистем и обеспечению экологической устойчивости.

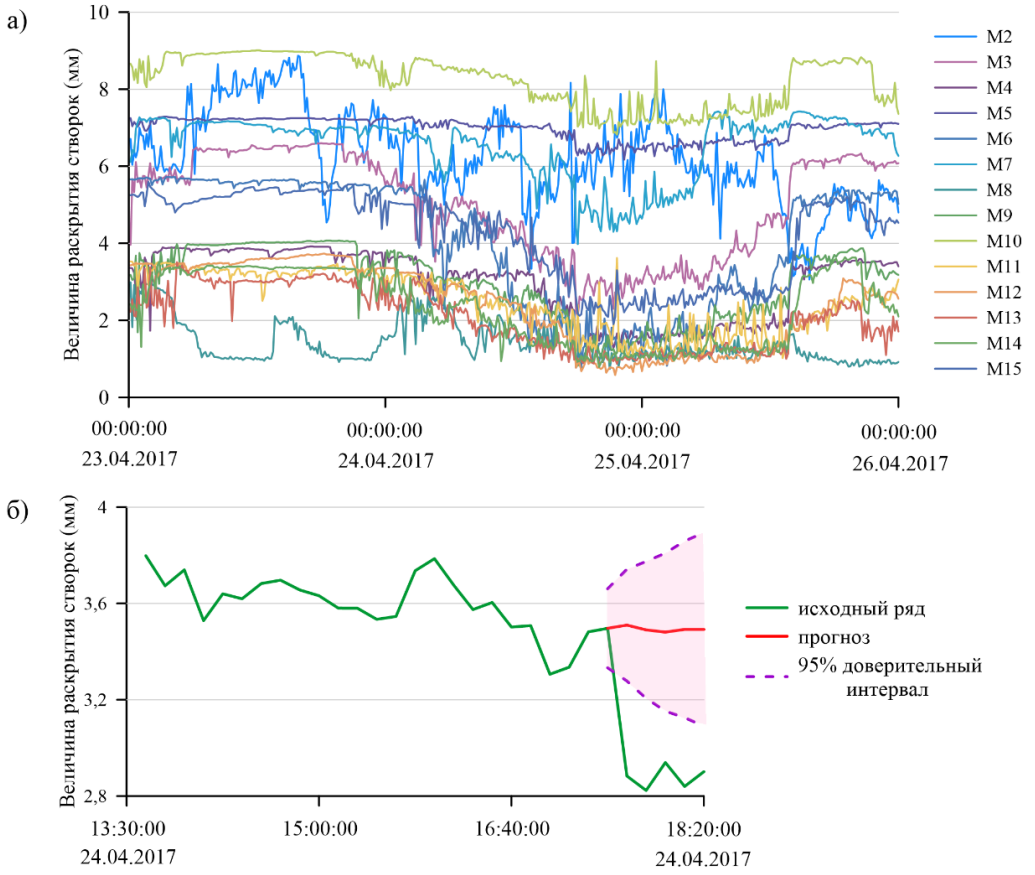

Рис. 1. Данные активности моллюсков 23–25 апреля 2017 года (а) и результат прогнозирования аномалии моделью SARIMA (б)
Fig. 1. Mollusk activity data for April 23–25, 2017 (a) and the result of the anomaly prediction by the SARIMA model (b)

Исследование выполнено за счет гранта Российского научного фонда № 23-29-00558, https://rscf.ru/project/23-29-00558/.

СПИСОК ЛИТЕРАТУРЫ

1. *Borcherding J.* Ten years of practical experience with the Dreissena-Monitor, a biological early warning system for continuous water quality monitoring // Hydrobiologia. 2006. 556. P. 417–426.

2. *Dvoretsky A.G., Dvoretsky V.G.* Shellfish as Biosensors in Online Monitoring of Aquatic Ecosystems: A Review of Russian Studies // Fishes. 2023. 8. 102.

3. *Gnyubkin V.F.* An early warning system for aquatic environment state monitoring based on an analysis of mussel valve movement // Russ. J. Mar. Biol. 2009. 35. P. 431–436.

4. *Kholodkevich S.V., Ivanov A.V., Kurakin A.S., Kornienko E.L., Fedotov V.P.* Real time biomonitoring of surface water toxicity level at water supply stations // J. Environ. Bioindic. 2008. 3. P. 23–34.

5. *Kane A.S., Salierno J.D., Gipson G.T., Molteno T.C.A., Hunter C.* A video-based movement analysis system to

quantify behavioral stress responses of fish // Water Res. 2004. 38. P. 3993–4001.

6. *Koch C.W., Cooper L.W., Lalande C., Brown T.A., Frey K.E., Grebmeier J.M.* Seasonal and latitudinal variations in sea ice algae deposition in the Northern Bering and Chukchi Seas determined by algal biomarkers // PLoS ONE. 2020. 15(4). e0231178.

7. *Трусевич В.В., Гайский П.В., Кузьмин К.А.* Автоматизированный биомониторинг водной среды с использованием реакций двустворчатых моллюсков // Морской гидрофизический журнал. 2010. № 3. С. 75–83.

8. *Grekov A.N., Kuzmin K.A., Mishurov V.Z.* Automated early warning system for water environment based on behavioral reactions of bivalves. 2019 International Russian Automation Conference (RusAutoCon) IEEE. 2019. P. 1–5.

9. *Box G.E., Jenkins G.M.* Time series analysis: forecasting and control, revised ed: Holden-Day. 1976. P. 575.

10. *Chen Q., Guan T., Yun L., Li R., Recknage F.* Online forecasting chlorophyll a concentrations by an auto-regressive integrated moving average model: Feasibilities and potentials // Harmful Algae. 2015. 43. P. 58–65.

11. *Katimon A., Shahid S., Mohsenipour M.* Modeling water quality and hydrological variables using ARIMA: a case study of Johor River, Malaysia. Sustain // Water Resour. Manag. 2018. 4. P. 991–998.

12. *Kumar U., Jain V.K.* ARIMA forecasting of ambient air pollutants (O_3, NO, NO_2 and CO) // Stochastic Environmental Research and Risk Assessment. 2009. 24 (5). P. 751–760.

13. *Siami-Namini S., Tavakoli N., Namin A.S.* A comparison of ARIMA and LSTM in forecasting time series. In international conference on machine learning and applications (ICMLA) // IEEE. 2018. P. 1394–1401.

14. *Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Müller A., Nothman J., Louppe G., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É.* Scikit-learn: machine learning in python // J. Mach. Learn. Res. 2011. 12. P. 2825–2830.

15. *Skipper S., Perktold J.* Statsmodels: Econometric and statistical modeling with python. In Proceedings of the 9th Python in Science Conference. 2010. P. 92–96.

16. *Трусевич В.В., Кузьмин К.А., Мишуров В.Ж., Журавский В.Ю., Вышкваркова Е.В.* Особенности поведенческих реакций черноморской мидии Mytilus galloprovicialis в естественных условиях обитания // Биология внутренних вод. 2021. № 1. С. 12–22.

17. *Grekov A.N., Kabanov A.A., Vyshkvarkova E.V., Trusevich V.V.* Anomaly Detection in Biological Early Warning Systems Using Unsupervised Machine Learning // *Sensors.* 2023. № 23. P. 2687.

APPLICATION OF THE ARIMA MODEL TO ANOMALY DETECTION IN THE BIVALVE ACTIVITY DATA

E.V. Vyshkvarkova, A.N. Grekov, A.S. Mavrin, V.V. Trusevich

Institute of Natural and Technical Systems, RF, Sevastopol, Lenin St., 28

The use of bivalve mollusks as bioindicators in automated monitoring systems of the aquatic environment allows real-time detection of an emergency situation associated with pollution of the aquatic environment. The use of machine learning algorithms allows quick anomaly detection for the subsequent generation of an alarm. The seasonal ARIMA model is used to predict mollusk activity data and detect an anomaly. The results show the possibility of using the seasonal ARIMA model for anomaly detection, which allows integrating the developed algorithmic approach into the software of biological systems of early detection in the future.
Keywords: anomalies, biomonitoring, forecast, ARIMA.

REFERENCES

1. *Borcherding J.* Ten years of practical experience with the Dreissena-Monitor, a biological early warning system for continuous water quality monitoring. *Hydrobiologia*, 2006, No. 556, pp. 417–426. https://doi.org/10.1007/s10750-005-1203-4.
2. *Dvoretsky A.G. and Dvoretsky V.G*. Shellfish as Biosensors in Online Monitoring of Aquatic Ecosystems: A Review of Russian Studies. *Fishes*, 2023, No. 8. 102. https://doi.org/10.3390/fishes8020102.
3. *Gnyubkin V.F.* An early warning system for aquatic environment state monitoring based on an analysis of mussel valve movement. *Russ. J. Mar. Biol.*, 2009, No. 35, pp. 431–436.
4. *Kholodkevich S.V., Ivanov A.V., Kurakin A.S., Kornienko E.L., and Fedotov V.P.* Real time biomonitoring of surface water toxicity level at water supply stations. *J. Environ. Bioindic*, 2008, No. 3, pp. 23–34.
5. *Kane A.S., Salierno J.D., Gipson G.T., Molteno T.C.A., and Hunter C.* A video-based movement analysis system to quantify behavioral stress responses of fish. *Water Res*., 2004, No. 38, pp. 3993–4001. https://doi.org/10.1016/j.watres.2004.06.028.
6. *Koch C.W., Cooper L.W., Lalande C., Brown T.A., Frey K.E., and Grebmeier J.M.* Seasonal and latitudinal variations in sea ice algae deposition in the Northern Bering and Chukchi Seas determined by algal biomarkers. *PLoS ONE,* 2020, No. 15(4), e0231178.
7. *Trusevich V.V., Gaiskii P.V., and Kuz'min K.A.* Avtomatizirovannyj biomonitoring vodnoj sredy s ispol'zovaniem reakcij dvustvorchatyh molljuskov (Automatic biomonitoring of aqueous media based on the response of bivalves). *Morskoj gidrofizicheskij zhurnal*, 2010, No. 20, pp. 231–238. https://doi.org/10.1007/s11110-010-9080-4.
8. *Grekov A.N., Kuzmin K.A., and Mishurov V.Z.* Automated early warning system for water environment based on behavioral reactions of bivalves. 2019 International Russian Automation Conference (RusAutoCon) IEEE, 2019, pp. 1–5.
9. *Box G.E. and Jenkins G.M.* Time series analysis: forecasting and control, revised ed: Holden-Day, 1976, p. 575.
10. *Chen Q., Guan T., Yun L., Li R., and Recknage F.* Online forecasting chlorophyll a concentrations by an auto-regressive integrated moving average model: Feasibilities and potentials. *Harmful Algae*, 2015, No. 43, pp. 58–65.
11. *Katimon A., Shahid S., and Mohsenipour M.* Modeling water quality and hydrological variables using ARIMA: a case study of Johor River, Malaysia. Sustain. *Water Resour, Manag*, 2018, No. 4, pp. 991–998.
12. *Kumar U. and Jain V.K.* ARIMA forecasting of ambient air pollutants (O_3, NO, NO_2 and CO). *Stochastic Environmental Research and Risk Assessment*, 2009, No. 24 (5), pp. 751–760.
13. *Siami-Namini S., Tavakoli N., and Namin A.S.* A comparison of ARIMA and LSTM in forecasting time series. In international conference on machine learning and applications (ICMLA). *IEEE*, 2018, pp. 1394–1401.
14. *Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Müller A., Nothman J., Louppe G., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., and Duchesnay É.* Scikit-learn: machine learning in python. *J. Mach. Learn. Res*., 2011, No. 12, pp. 2825–2830.
15. *Skipper S. and Perktold J.* Statsmodels: Econometric and statistical modeling with python. In Proceedings of the 9th Python in Science Conference, 2010, pp. 92–96.
16. *Trusevich V.V., Kuz'min K.A., Mishurov V.Zh., Zhuravsky V.Yu., and Vyshkvarkova E.V.* Osobennosti povedencheskih reakcij chernomorskoj midii Mytilus galloprovicialis v estestvennyh uslovijah obitanija (Features of behavioral responses of the Mediterranean mussel in its natural habitat of the Black Sea). *Inland Water Biology,* 2021, No. 14 (1), pp. 10–19.
17. *Grekov A.N., Kabanov A.A., Vyshkvarkova E.V., and Trusevich V.V.* Anomaly detection in biological early warning systems using unsupervised machine learning. *Sensors,* 2023, No. 23, p. 2687. https://doi.org/10.3390/s23052687.