DOI: 10.25558/VOSTNII.2025.21.80.009
УДК 004.032.26

Д. Н. ПАТРИКЕЕВ
аспирант
Пензенский государственный технологический университет, Пенза
e–mail: patrikeevdn@list.ru

К. Р. ТАРАНЦЕВА
доктор технических наук, профессор,
заведующая кафедрой
Пензенский государственный технологический университет, Пенза

АНАЛИЗ ПРЕИМУЩЕСТВ И НЕДОСТАТКОВ SOM-ФИЛЬТРА ДЛЯ ОЦЕНКИ ЭКОЛОГИЧЕСКОГО СОСТОЯНИЯ ВОДНОЙ СРЕДЫ

Актуальность данной работы обусловлена необходимостью совершенствования систем экологического мониторинга, особенно в контексте анализа загрязнения водных объектов. Самоорганизующиеся карты Кохонена (SOM) представляют собой перспективный инструмент кластеризации и визуализации многомерных данных, однако их потенциал в экологическом мониторинге до конца не раскрыт и требует дополнительного изучения.

Цель исследования заключается в оценке границ применимости SOM-фильтра как составного модуля в комплексных системах экологического мониторинга.

Полученные результаты подтверждают эффективность SOM для решения задач экологического анализа, способность нейронной сети Кохонена к сжатию многомерных данных может быть использована для выбора входных параметров для предсказания показателей загрязнения окружающей среды.

Ключевые слова: ЭКОЛОГИЧЕСКИЙ МОНИТОРИНГ, ОКРУЖАЮЩАЯ СРЕДА, МАШИННОЕ ОБУЧЕНИЕ, НЕЙРОННЫЕ СЕТИ, НЕЙРОННАЯ СЕТЬ КОХОНЕНА, SOM.

ВВЕДЕНИЕ

Самоорганизующаяся карта Кохонена (SOM) — разновидность нейронной сети, предназначенная для кластеризации и визуализации многомерных данных. SOM преобразует сложные входные данные в упрощённое представление, сохраняя при этом их топологическую структуру, позволяя пользователю лучше понять структуру данных, что может способствовать принятию решений.

Проблемы, для решения которых необходимо совершенствование систем экологического мониторинга, актуальны и регулярно обсуждаются в государственных докладах [1–2]. В связи с этим исследование преимуществ и недостатков SOM актуально для задач мониторинга экологического состояния водных объектов.

Авторы Габдрахманова Г. Н., Кремлева Э. Ш. и Байбакова Е. В. [3–5] рассматривали SOM в виде одного из элементов разработанных систем для оценки экологического состояния водной среды, используя 2-уровневый каскадный SOM-фильтр в качестве модуля

кластеризации. Однако в данном случае потенциал SOM раскрыт не полностью, кроме того, его ограничения могут помешать совершенствовать рассмотренные системы. **Цель настоящей статьи** — оценить границы применимости SOM фильтра в качестве составного модуля в комплексных системах экологического мониторинга [3–6].

Для достижения поставленной цели выделены следующие задачи: рассмотреть преимущества и возможности SOM, провести эксперимент на тестовых данных, выявить недостатки архитектуры SOM.

МАТЕРИАЛЫ И МЕТОДЫ

Поставленная цель может быть достигнута несколькими способами. Структура набора данных может сильно влиять на полученный вывод, поэтому она не должна быть слишком простой. Поскольку ранее [3, 5] было рассмотрено не более 20 параметров, необходимо исследовать набор данных с числом поданных на вход SOM параметров более 20.

В эксперимент включена обработка данных с помощью программных библиотек для языка программирования Python (Pandas и Numpy), создание моделей на основе самоорганизующихся карт Кохонена и её обучение на данных о загрязнении.

Обучение этой нейронной сети происходит по принципу «без учителя», кластеры образуются вокруг нейронов произвольно, с каждой итерацией уточняя данные о соседних нейронах и кластерах.

Одной из ключевых характеристик этих сетей является способность сохранять топологическую структуру входных данных, упрощая многомерное пространство до двухмерного для наглядного представления и анализа. Анализ весов нейронов SOM может дать данные об их корреляции. На основании вышеперечисленного был проведён эксперимент для выявления закономерностей загрязнения поверхностных водных объектов России за период с 2008 по 2021 гг. [7]. Поиск зависимостей был осуществлён с помощью вычисления попарной корреляции весов нейронов в обученной модели. Ниже рассмотрен принцип работы SOM и алгоритм получения данных зависимостей [8–9].

На начальном этапе происходит инициализация весового вектора W и установка начальной скорости обучения η_0. Все входные векторы X (1) и весовые векторы W нормализуются (2).

$$X' = \frac{X}{||X||} = \frac{(x_1, x_2, x_n)T}{[x_1^2 + x_2^2 + \cdots + x_n^2]^{\frac{1}{2}}}, \quad (1)$$

где $||X|| = (\sum_{j=1}^{n}(x_j)^2)^{\frac{1}{2}}$ — норма входного вектора.

$$W_i'(0) = \frac{W_i(0)}{||W_i(0)||}, \quad (2)$$

где $||W_i(0)|| = (\sum_{j=1}^{n}[w_{ij}(0)]^2)^{\frac{1}{2}}$ — норма весового вектора на этапе инициализации.

Далее вычисляется евклидово расстояние (3) между нормализованными весовыми векторами W_i' и входным вектором X′.

$$d_i = [\sum_{j=1}^{n}(x_i - w_{ij})^2]^{\frac{1}{2}}, \text{ i} = 1,2,\ldots,\text{m}, \quad (3)$$

На основании минимального евклидова расстояния (4) определяется победивший в конкурентном отборе, наиболее близкий к входному вектору в пространстве признаков, нейрон.

$$\left||X' - W_C'|\right| = min_i \left||X' - W_i'|\right| = \\ = min_i[d_i], i = 1,2, \ldots, m, \quad (4)$$

где C — нейрон, для которого выполняется условие, проходящий через процесс конкурентного обучения.

Весовые векторы обновляются для победившего нейрона и нейронов в его топологической окрестности N_i(X, n) (5).

$$\begin{cases} w_j(n+1) = w_j(n) + \eta(n)h_{j,i(x)}(n)\left(X - w_j(n)\right) \; j \in N_{i(x)}(n) \\ w_j(n+1) = w_j(n+1) \; j \notin N_{i(x)}(n) \end{cases}, \quad (5)$$

где η(n) — скорость обучения на nn-м шаге; h_j (n) — функция соседства, зависящая от расстояния до победившего нейрона.

Скорость обучения, выраженная функцией затухания коэффициента обучения (6) и радиуса функции соседства (7), адаптируются для каждой итерации.

$$\eta(n) = \eta_0 \exp\left(-\frac{n}{\tau_2}\right), n = 0,1,2, \dots, N, \quad (6)$$

$$\sigma(n) = \sigma_0 \exp\left(-\frac{n}{\tau_2}\right), n = 0,1,2, \dots, N, \quad (7)$$

После обновления веса нейронов нормализуются (8).

$$W'(n+1) = w_i(n+1)/||w_i(n+1)||, \quad (8)$$

В ходе обучения модели выбирается наилучшая соответствующая единица (BMU) для каждого входного вектора путём минимизации евклидовых расстояний между входными элементами x_i и узлами самоорганизующейся карты m_c. Ошибка квантования (Quantization Error, QE) (9) вычисляется как среднее расстояние между каждой точкой данных x_i и её ближайшей BMU m_c.

$$QE = \frac{1}{N}\sum_i \| m_c - x_i \| =$$

$$= \frac{1}{N}\sum_i \sqrt{(m_c^2 + x_i^2 - 2m_c x_i)} \quad (9)$$

где n — общее число точек данных; $||m_c - x_i||$ — евклидово расстояние между точкой данных x_i и её ближайшей BMU.

Далее измеряется степень сохранения топологии данных на карте SOM с помощью топологической ошибки. Ошибка топографической точности (Topographic Error, TE) (10) вычисляется как доля точек данных x_i, для которых лучшие соответствующие BMU не являются ближайшими соседями. При этом, если сетка не прямоугольная, а шестиугольная, показатель TE может быть завышен из-за увеличения количества соседей.

$$TE = \frac{1}{N}\sum_{i=1}^{N} u_{x_i}, \quad (10)$$

где u_{Xi}=1, если первый и второй BMU x_i являются соседями, и u_{Xi}=0 в противном случае.

Ввиду неизбежности искажения SOM элементами, присутствует система контроля меры искажения (Distortion Measure, DM) (11). Для расчета используется функция соседства, которая учитывает расстояния между каждым элементом карты и каждой точкой данных. DM учитывает взвешенные квадратные расстояния, где вес задается функцией соседства, в отличие от QE.

$$DM = \sum_{i=1}^{N}\sum_{j=1}^{M} h_{ij} m_j - x_i^2, \quad (11)$$

где h_{ij} — ядро соседства, центрированное на BMU в точке x_i; m_j — координаты узла карты; x_i — координаты точки данных.

На рис. 1 изображена блок-схема модуля кластеризации, где N — количество уровней кластеризации. Для эксперимента было выбрано N = 2.

Параметры, которые использовались для обучения модели, представленной ниже: скорость обучения = 0,01; радиус обнаружения соседних кластеров = 3; количество итераций = 2.

Показатели степени загрязнения водных объектов: лигнин сульфатный, дихлорфенол, вольфрам, дихлордифенилдихлорэтилен (ДДЭ), растворенный кислород, бериллий, марганец, запах, ХПК, аммоний-ион, алюминий, аспав, БПК5, железо общее, фенол, свинец, метанол, ртуть, ДДТ, нефть и нефтепродукты, изопропанол, ацетон, взвешенные вещества, мышьяк, хром (VI), ванадий, никель, нитрит-ионы, формальдегид, бензол, нафталин, дитиофосфат крезиловый, сульфиды и сероводород, этилацетат, лигносульфонаты, водородный показатель (pH), молибден, медь, фосфор элементарный, ГХЦГ, бенз(а)пирен, кадмий, циклогексанол, цинк, фосфаты.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Было выявлено 4 кластера при первой итерации кластеризации, один из которых, содержащий в себе наибольшее количество

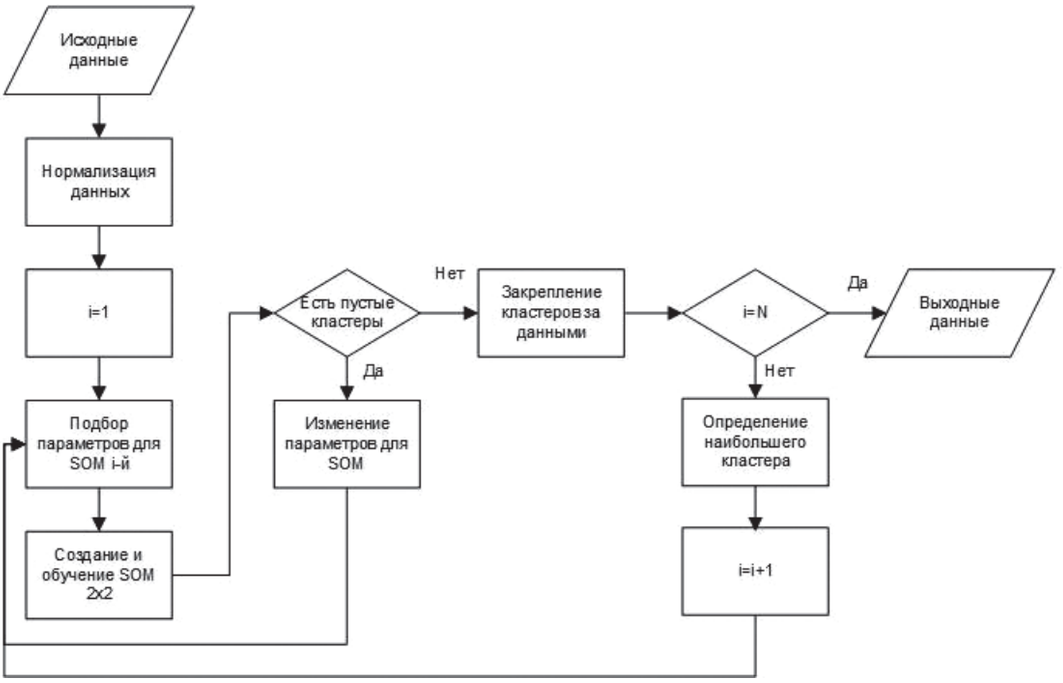

Рис. 1. Блок-схема модуля кластеризации

записей, был разбит ещё на 4 кластера, образуя второй уровень кластеризации. Результат кластеризации представлен в табл. 1.

Также была создана и обучена модель одноуровневая модель с размером сетки 10 х 10 для оценки границы применимости архитектуры Кохонена при поиске скрытых зависимостей в данных.

Для визуализации и дальнейшего анализа удобно использовать тепловые карты. Тепловая карта представляет собой визуальное представление данных, в котором различные значения параметров отображаются с использованием цветовой шкалы, где каждому значению соответствует определённый цвет. Визуализация весов SOM на нескольких тепловых картах для всех показателей может указать на данные, которые похожим образом изменяют веса, следовательно, предполагается корреляция данных (рис. 2).

Ввиду того, что между нормализованными значениями параметров в исходных данных может быть корреляция, необходимо построить тепловую карту их корреляции (рис. 3).

Заметно, что корреляция большинства показателей стремится к 0, при этом у растворённого кислорода наблюдается отрицательная корреляция с остальными показателями. При

Таблица 1

Кластеры, выявленные в ходе эксперимента

Уровень кластера	Номер кластера (нейрон)	Число записей в кластере
1	(0,0)	7631
1	(0,1)	21
1	(1,0)	10
1	(1,1)	8
2	(0,0)	6490
2	(0,1)	1102
2	(1,0)	38
2	(1,1)	1

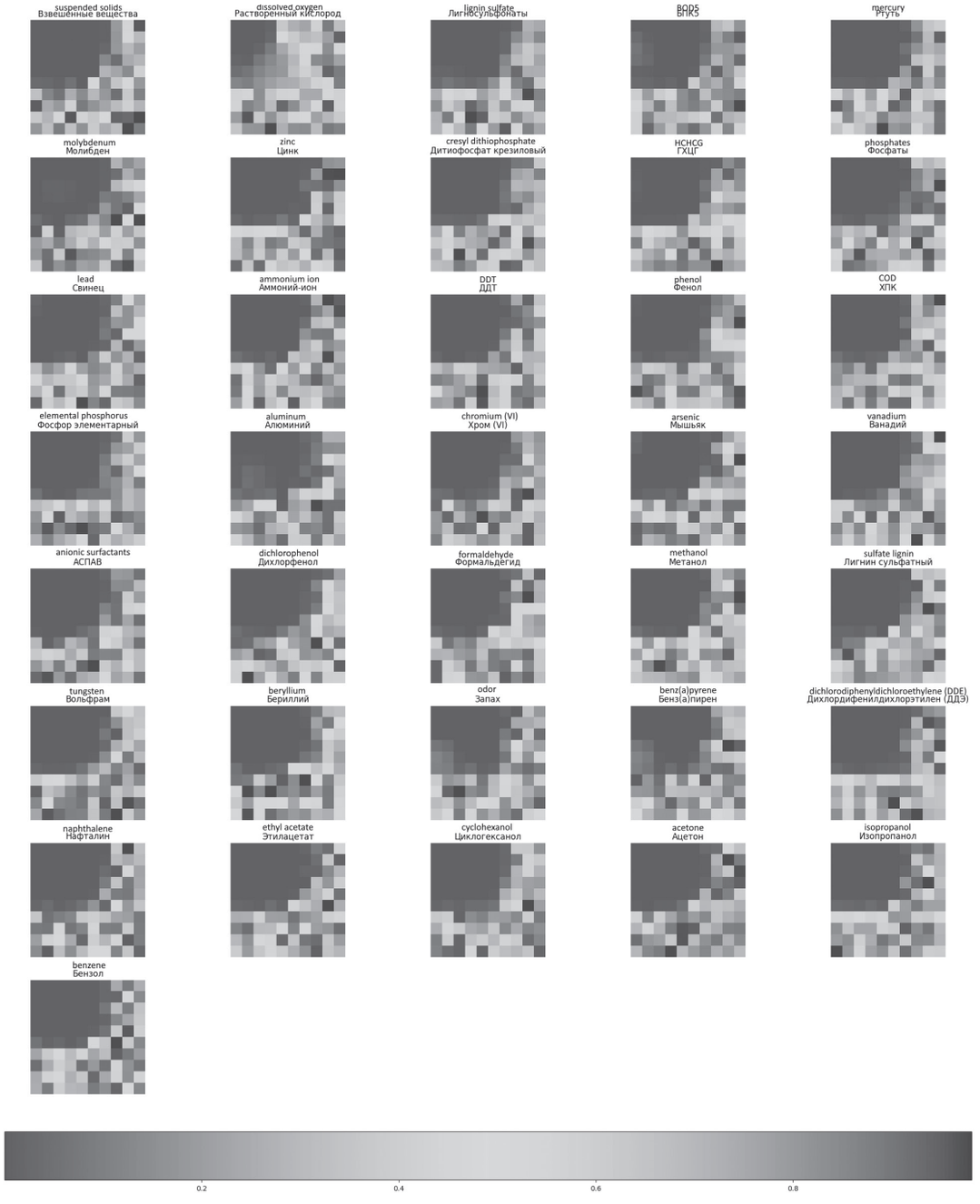

Рис. 2. Визуализация весов

этом данные на тепловой карте отражены относительно диагонали.

Получив данные об отсутствии видимой корреляции данных, необходимо построить тепловую карту корреляции между весами (рис. 4).

Наблюдается явная корреляция весов между собой, при этом растворённый кислород показывает крайне низкий уровень корреляции с остальными параметрами, что не противоречит тепловой карте корреляции нормализованных данных до обучения.

Возможно, низкий уровень корреляции растворённого кислорода с другими параметрами связан с тем, что для эксперимента был использован набор данных высокого и экстремально высокого загрязнения водных объектов [7]. Концентрация растворённого кислорода для большинства измерений находилась в диапазоне от 2,2 до 2,9 мг/л.

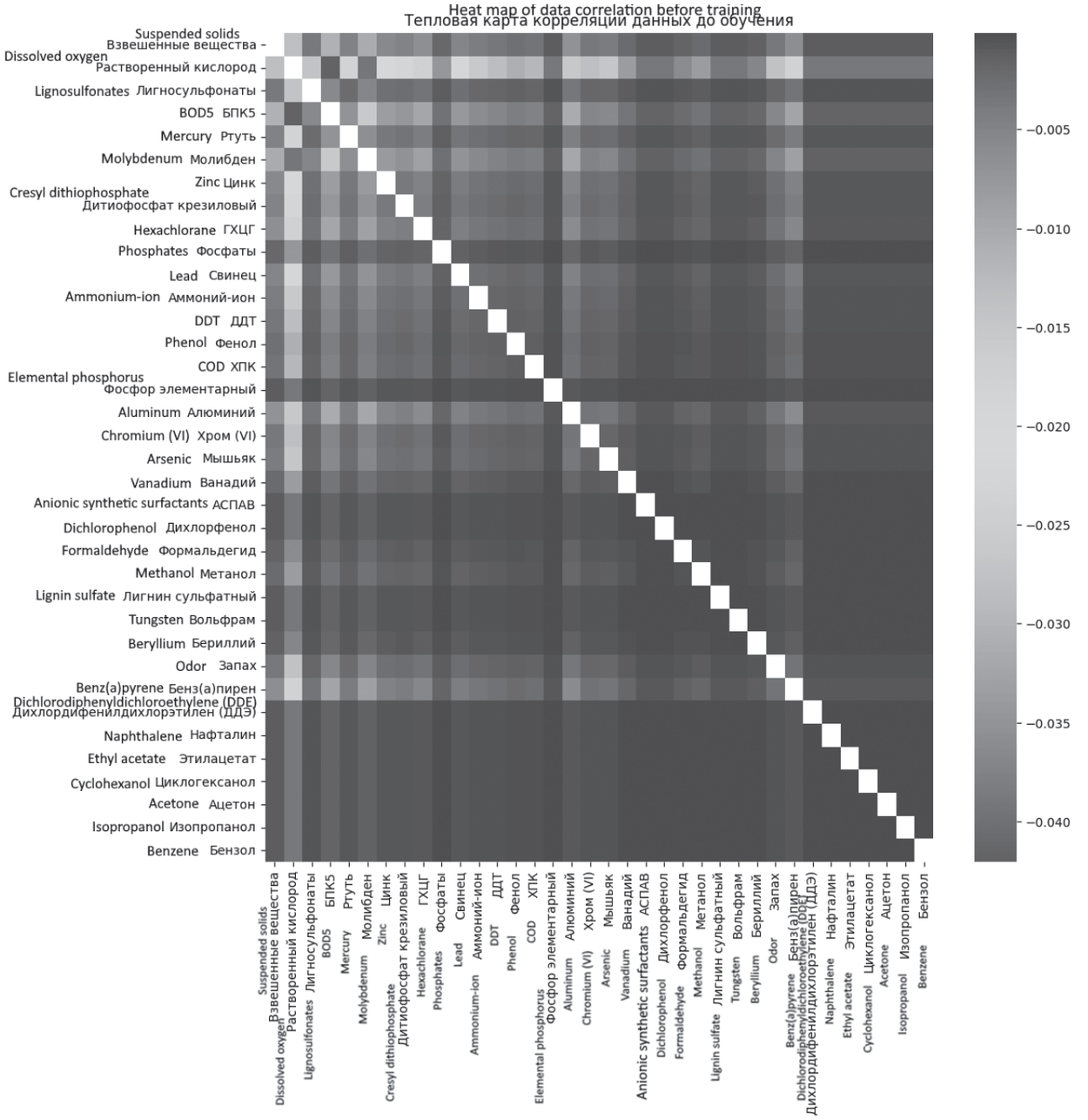

Рис. 3. Тепловая карта корреляции данных до обучения

В ходе эксперимента были выявлены следующие недостатки самоорганизующихся карт Кохонена как архитектуры нейронной сети: для крупных наборов данных с множеством параметров выходные данные модели в числовом или визуализированном виде трудно интерпретировать, закономерности визуально неразличимы; нельзя прибегнуть к обучению с учителем, поэтому точная классификация с заранее заданными метриками невозможна; настройка модели производится эмпирическим путём, неправильный подбор параметров настройки модели (скорость обучения, коэффициент поиска соседних нейронов и кластеров, размер карты) может привести к неэффективной кластеризации, потери топологической структуры или к ошибкам отображения при последующем анализе; SOM не имеет встроенного механизма для учёта временных зависимостей

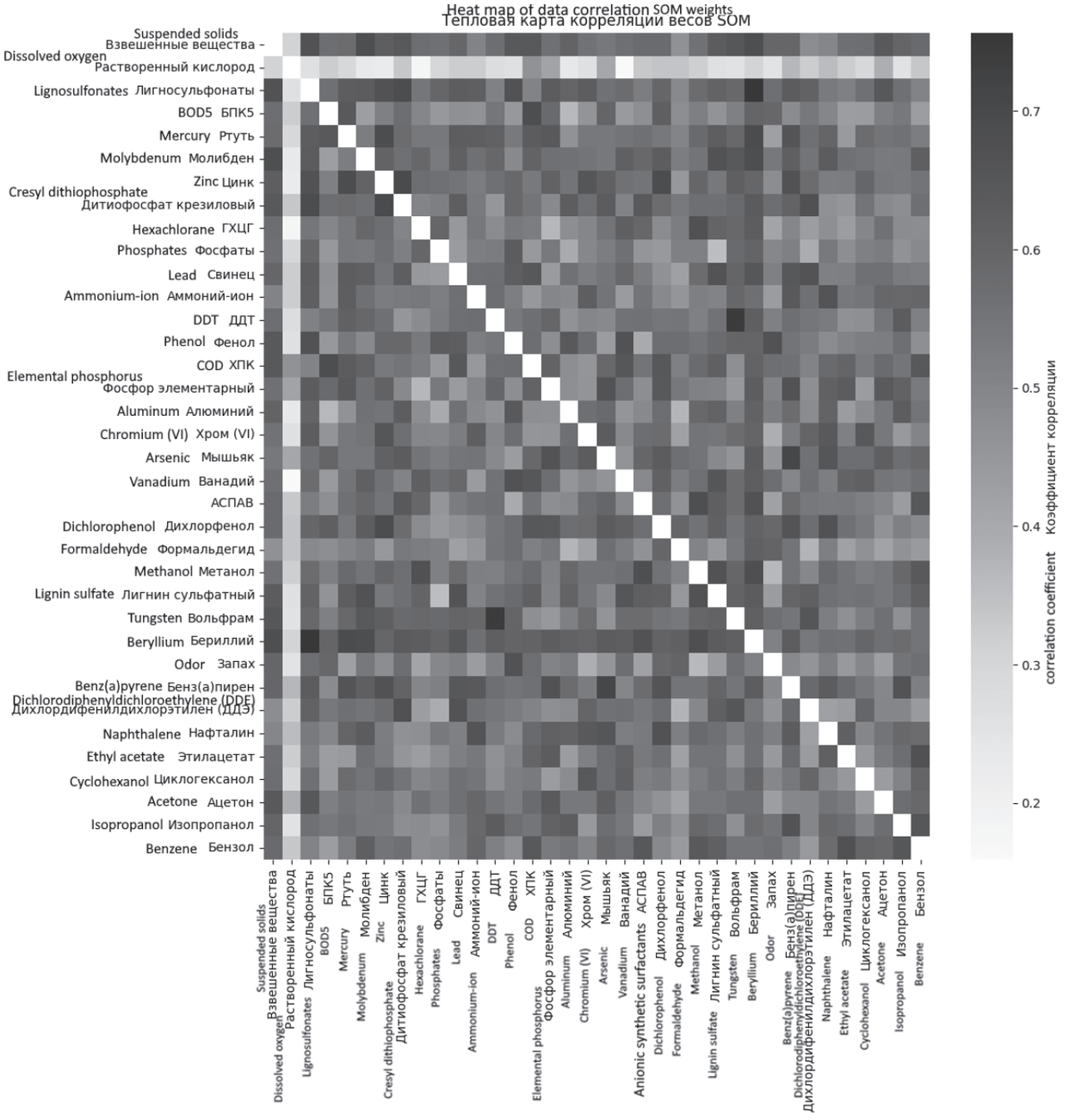

Рис. 4. Тепловая карта корреляции весов SOM

между последовательными событиями: вычленяет общие паттерны и кластеры данных, но не учитывает временную зависимость — параметр времени может быть учтён в расчётах только косвенно; при увеличении размерности сетки существенно увеличиваются временные затраты на обучение.

Наиболее влияющая на результат поиска скрытых зависимостей характеристика: радиус обнаружения соседних кластеров.

На рис. 5–6 представлены веса и их тепловая карта корреляции для радиуса 1.

ВЫВОДЫ

Самоорганизующиеся карты Кохонена — функциональный и полезный инструмент для экологической оценки. Однако границы применимости n-уровневых фильтров, ограничены архитектурой SOM, как описано

Рис. 5. Визуализация весов для радиуса обнаружения соседних кластеров = 1

выше. Необходимо, используя архитектуру SOM, разработать методику градации водных объектов питьевого, рыбохозяйственного, рекреационного назначения по степени загрязнения относительно ПДК и предсказания значения показателей, характеризующих антропогенную загрязнённость, что является целью дальнейших исследований.

Исследование скрытой корреляции концентрации растворённого кислорода и других показателей в условиях искусственно ограниченного набора данных может стать важным этапом в совершенствовании методов оценки экологического состояния водной среды.

Основные возможности архитектуры были подтверждены экспериментальным путём.

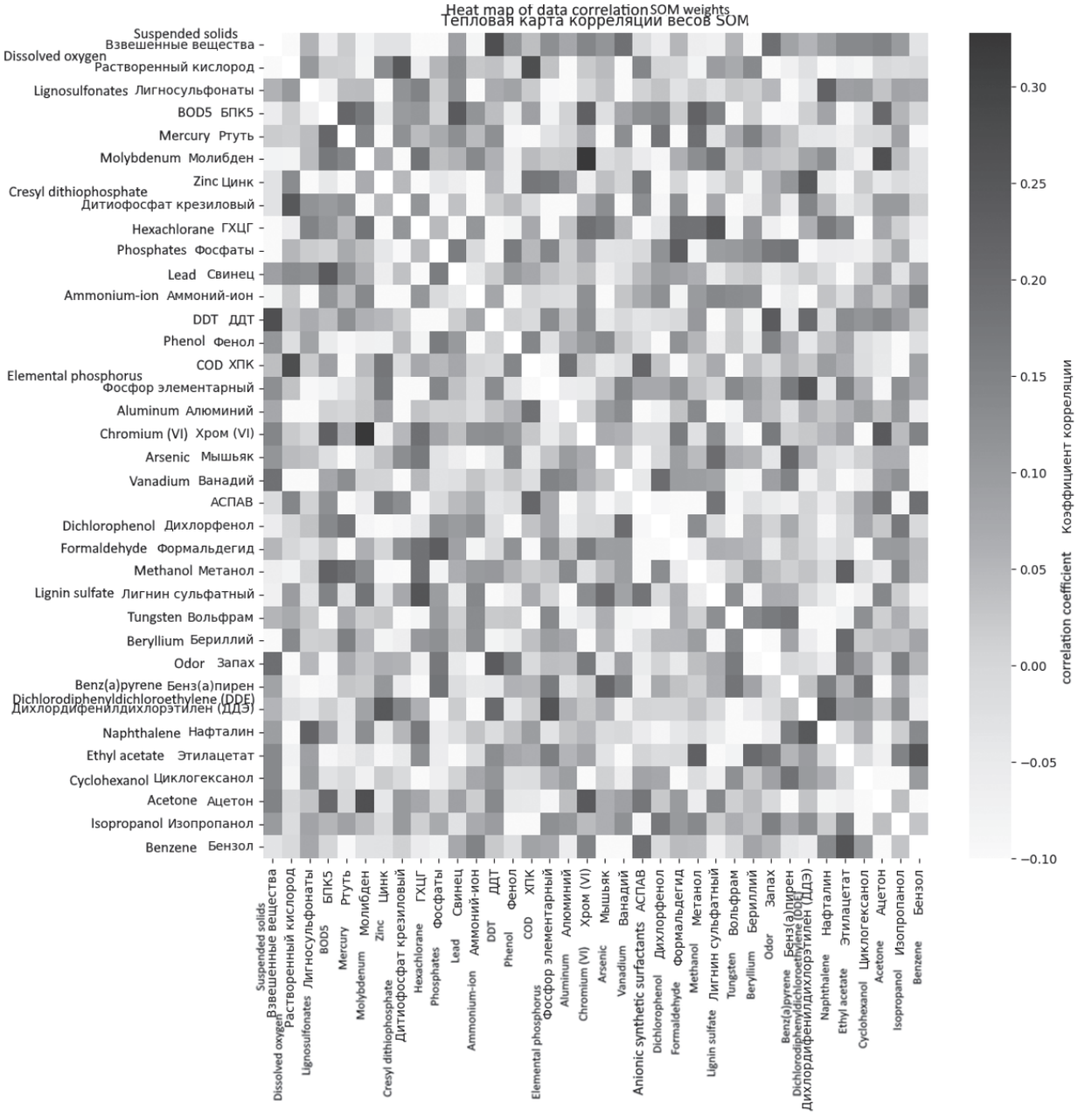

Рис. 6. Тепловая карта корреляций для радиуса обнаружения соседних кластеров = 1

СПИСОК ЛИТЕРАТУРЫ

1. О состоянии природных ресурсов и об охране окружающей среды Пензенской области в 2023 году: государственный доклад. Пенза, 2024. 127 с.

2. О состоянии окружающей среды Пензенской области (данные открытых источников) в рамках реализации проекта «Экологический патруль «Зелёной волны». Пенза: государственный доклад. Пенза, 2018.

3. Габдрахманова, Г. Н. Комплексная региональная оценка качества вод в урбоэкосистеме: автореф. дис. на соиск. учен. степ. канд. техн. наук / Габдрахманова Гульнара Наилевна. Казань, 2019.

4. Кремлева, Э. Ш. Модели и методы интеллектуальной обработки данных для систем поддержки принятия решений (на примере систем экологической безопасности): автореф. дис. на соиск. учен. степ. канд. техн. наук / Кремлева Эльмира Шамильевна. Казань, 2021.

5. Байбакова, Е. В. Хемометрический подход к региональному нормированию природных вод в урбоэкосистеме: автореф. дис. на соиск. учен. степ. канд. хим. наук / Байбакова Евгения Васильевна. Казань, 2024.

6. Царькова, Е. Г. Самоорганизующиеся карты Кохонена как инструмент анализа данных в ведомственных научных исследованиях // Научно-технический вестник Поволжья. 2024. № 11. С. 458–460.

7. Загрязнение поверхностных вод в России: ежемесячные данные о высоком и экстремально высоком загрязнении водных объектов за 2008–2021 гг. / Росгидромет; обработка: Гостева И. И., Сёмин П. О., Инфраструктура научно-исследовательских данных, АНО «ЦПУР», 2021. Доступ: Лицензия CC BY-SA. Размещено: 23.09.2021. (Ссылка на набор данных: http://data.rcsi.science/data-catalog/datasets/176/).

8. Clark, S., Sisson, S. A., Sharma, A. Tools for enhancing the application of self-organizing maps in water resources research and engineering // Advances in Water Resources. 2020. Vol. 143.

9. Xiang, Q. et al. The potential ecological risk assessment of soil heavy metals using self-organizing map // Science of the Total Environment. 2022. Vol. 843.

10. Больщиков В. А. Использование самоорганизующихся карт Кохонена для метрологически обоснованной кластеризации результатов измерений // Системный анализ в проектировании и управлении. 2024. Т. 27. № 2. С. 504–511.

DOI: 10.25558/VOSTNII.2025.21.80.009
UDC 004.032.26

D. N. PATRIKEEV
Postgraduate Student
Penza State Technological University, Penza
e–mail: patrikeevdn@list.ru

K. R. TARANTSEVA
Doctor of Engineering Sciences, Professor,
Head of the Department
Penza State Technological University, Penza
e–mail: krtar2018@bk.ru

ANALYSIS OF ADVANTAGES AND DISADVANTAGES OF SOM-FILTER FOR ASSESSMENT OF ECOLOGICAL STATE OF AQUATIC ENVIRONMENT

The relevance of this work is due to the need to improve environmental monitoring systems, especially in the context of analyzing the pollution of water bodies. Self-Organizing Kohonen Maps (SOM) represent a promising tool for clustering and visualization of multidimensional data, but their potential in environmental monitoring has not been fully disclosed and requires additional study.

The aim of the study is to assess the limits of applicability of the SOM filter as a composite module in integrated environmental monitoring systems.

The results obtained confirm the effectiveness of SOM for solving problems of environmental analysis, the ability of Kohonen neural network to compress multidimensional data can be used to select input parameters for prediction of environmental pollution indicators.

Keywords: ENVIRONMENTAL MONITORING, ENVIRONMENT, NEURAL NETWORKS, KOHONEN NETWORK, SOM

REFERENCES

1. On the state of natural resources and environmental protection of the Penza Region in 2023: a state report. Penza, 2024. 127 p. [In Russ.].
2. On the state of the environment of the Penza Region (open-source data) within the framework of the Green Wave Environmental Patrol project. Penza: State report. Penza, 2018. [In Russ.].
3. Gabdrakhmanova, G. N. Comprehensive regional assessment of water quality in the urban ecosystem: abstract of the dissertation. for the job. learned. step. Candidate of Technical Sciences / Gabdrakhmanova Gulnara Nailevna. Kazan, 2019. [In Russ.].
4. Kremleva, E. S. Models and methods of intelligent data processing for decision support systems (on the example of environmental safety systems): abstract of the dissertation for the job. learned. step. Candidate of Technical Sciences / Kremleva Elmira Shamilyevna. Kazan, 2021. [In Russ.].
5. Baibakova, E. V. Chemometric approach to regional regulation of natural waters in the urban ecosystem: abstract of the dissertation. for the job. learned. step. Candidate of Chemical Sciences / Baibakova Evgeniya Vasilyevna. Kazan, 2024. [In Russ.].
6. Tsarkova, E. G. Self-organizing Kohonen maps as a data analysis tool in departmental scientific research // Scientific and Technical Bulletin of the Volga Region. 2024. No. 11. P. 458–460. [In Russ.].
7. Surface water pollution in Russia: monthly data on high and extremely high pollution of water bodies for 2008-2021 / Roshydromet; processing: Gosteva I. I., Semin P. O., Infrastructure of scientific research data, ANO «CPU», 2021. Access: CC BY-SA license. Posted: 09/23/2021. (Link to the dataset: http://data.rcsi.science/data-catalog/datasets/176/). [In Russ.].
8. Clark, S., Sisson, S. A., Sharma, A. Tools for enhancing the application of self-organizing maps in water resources research and engineering // Advances in Water Resources. 2020. Vol. 143.
9. Xiang, Q. et al. The potential ecological risk assessment of soil heavy metals using self-organizing map // Science of the Total Environment. 2022. Vol. 843.
10. Bolshchikov V. A. The use of self-organizing Kohonen maps for metrologically based clustering of measurement results // System analysis in design and management. 2024. Vol. 27. No. 2. P. 504–511. [In Russ.].