Григорьев Я. Ю., Альхименко И. Н.
Ya. Yu. Grigoriev, I. N. Alkhimenko

ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ДАННЫХ ДЛЯ РЕАЛИЗАЦИИ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В ЗАДАЧАХ ЭКОЛОГИЧЕСКОГО КОНТРОЛЯ

DATA PREPROCESSING FOR MACHINE LEARNING METHODS IMPLEMENTATION IN ENVIRONMENTAL CONTROL TASKS

Григорьев Ян Юрьевич – кандидат физико-математических наук, доцент, проректор по учебной работе Комсомольского-на-Амуре государственного университета (Россия, Комсомольск-на-Амуре); тел. 8(914)211-73-00. E-mail: alhimenko12345@inbox.ru.
Yan Yu. Grigoriev – PhD in Physics and Mathematics, Associate Professor, Vice-Rector for Academic Affairs of Komsomolsk-na-Amure State University (Russia, Komsomolsk-on-Amur); tel. 8(914)211-73-00. E-mail: alhimenko12345@inbox.ru.

Альхименко Игорь Николаевич – студент Комсомольского-на-Амуре государственного университета (Россия, Комсомольск-на-Амуре); тел. 8(914)211-73-00. E-mail: alhimenko12345@inbox.ru.
Igor N. Alkhimenko – Student, Komsomolsk-na-Amure State University (Russia, Komsomolsk-on-Amur); tel. 8(914)211-73-00. E-mail: alhimenko12345@inbox.ru.

Аннотация. В работе рассматриваются проблемы, связанные с формированием набора данных для реализации методов машинного обучения, предлагаются математическая модель предварительной обработки данных и её программная реализация. В исследовании рассматриваются задачи оценки состояния различных видов земной поверхности, базирующиеся на моделях компьютерного зрения. Применение предлагаемых подходов к формированию набора данных позволяет повысить точность моделей машинного обучения и выявить наиболее значимые спектральные характеристики для различных поверхностей в задачах дистанционного зондирования.

Summary. The paper discusses problems associated with the formation of a data set for the implementation of machine learning methods, and offers a mathematical model of data preprocessing and its software implementation. The study considers the tasks of assessing the state of various types of the Earth's surface, based on computer vision models. The application of the proposed approaches to the formation of a data set makes it possible to increase the accuracy of machine learning models and identify the most significant spectral characteristics for various surfaces in remote sensing tasks.

Ключевые слова: данные, предварительная обработка, машинное обучение, экологический контроль, математическая модель, алгоритм, спектральный анализ, дистанционное зондирование.

Key words: data, preprocessing, machine learning, environmental control, mathematical model, algorithm, spectral analysis, remote sensing.

Исследование выполнено за счёт средств гранта Министерства образования и науки Хабаровского края № 41С/2023.

УДК 004.92

Введение. Современное общество сталкивается с рядом экологических проблем, которые оказывают губительное влияние на окружающую среду, здоровье людей и функционирование экономических систем. Возникает необходимость непрерывного мониторинга экологического состояния земной поверхности для своевременного выявления нештатных ситуаций. В настоящее время применяются различные методы мониторинга состояния земной поверхности, к которым относятся геоинформационные системы, методы биомониторинга, химического анализа.

Данные методы имеют ряд существенных недостатков: высокая стоимость применяемого оборудования; необходимость постоянной экспертной оценки высококвалифицированными специалистами обрабатываемой информации; длительное время, требуемое для получения результатов.

Использование методов дистанционного зондирования Земли (ДЗЗ) минимизирует вышеперечисленные проблемы. Методы ДЗЗ, основанные на спектральном анализе, имеют широкую область применения и позволяют своевременно определять области с возможными отклонениями от штатного состояния. Указанный подход основан на применении спектральных индексов и не позволяет определять различные состояния поверхностей без постоянной экспертной оценки, а также имеет низкую устойчивость к влиянию атмосферных явлений.

Применение ДЗЗ с использованием технологии машинного обучения не требует определения спектральных индексов, позволяет минимизировать обозначенные проблемы и повысить эффективность применяемого подхода. Ключевой задачей эффективной реализации методов машинного обучения является подготовка набора данных. Качественные данные требуют предварительной обработки, реализация которой основана на разработке и применении отдельных моделей.

Целью данной работы является построение алгоритма предварительной обработки данных. В качестве примера реализации рассматривается задача бинарной классификации нефтяных загрязнений, решаемая с применением методов машинного обучения.

Методы и материалы. Исходные данные, рассматриваемые в работе, содержат информацию о значениях 12 спектральных каналов. Изображения получаются с сенсоров спутника Sentinel-2 L2A. Схема работы с данными каналами приводится на рис. 1.

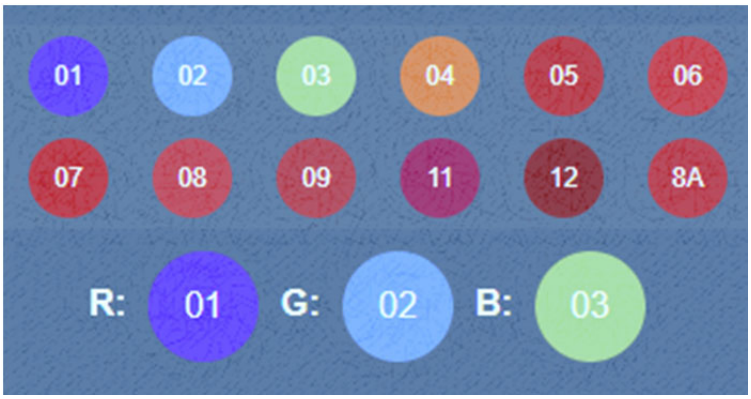

Рис. 1. Двенадцать спектральных каналов спутника Sentinel-2 L2A

Пиксель любого RGB-изображения представляется набором из 3 значений в диапазоне от 0 до 255. Таким образом, стандартный пиксель представляется в виде (x, y, z), где $x, y, z \in [0, 255]$. Число из указанного диапазона показывает влияние каждого канала на цвет пикселя на снимке. Проводится анализ данных, представленных в виде файла формата csv (см. рис. 2). Осуществляется предварительная обработка отсортированных данных.

Предварительная обработка данных реализуется методами математической статистики на основе корреляционного анализа. Исследование направлено на оценку статистических свойств данных для последующего формирования обучающего набора. Определяются закономерности и связи, обеспечивающие возможность выявления параметров модели. На основе корреляционной матрицы оценивается влияние одного спектрального канала на другой. В исследовании выделяется пять промежутков для определения тесноты связи: $R[0{,}70; 1]$ – тесная, R[0,50;0,69) – средняя, R[0,30;0,49) – умеренная; R[0,20;0,29) – слабая, R[0;19) – очень слабая.

	image	class	Aerosol	Blue	Green	Red	IR1	IR2	IR3	IR4	IR5	IR6	IR7	IR8
2	A22	1	8	8	6	7	8	3	10	15	8	20	15	12
3	A22	1	8	8	6	7	8	3	10	15	8	20	15	12
4	A22	1	8	8	6	7	8	3	10	15	8	20	15	12
5	A22	1	8	8	6	7	8	3	11	16	9	19	15	12
6	A22	1	8	8	6	7	8	3	11	16	9	19	15	12
7	A22	1	8	8	6	7	8	3	11	17	7	19	15	12
8	A22	1	8	8	6	7	8	3	11	17	7	19	15	12
9	A22	1	8	8	6	8	7	3	13	16	9	19	16	11
10	A22	1	8	8	6	8	7	3	13	16	9	19	16	11
11	A22	1	8	8	6	7	6	2	13	16	9	19	16	11
12	A22	1	8	8	6	7	6	2	14	17	10	19	16	11
13	A22	1	8	8	6	7	6	2	12	15	8	19	16	11
14	A22	1	8	8	6	8	7	3	9	12	5	19	16	11
15	A22	1	8	8	6	7	6	2	8	10	5	19	16	11

Рис. 2. Представление обрабатываемых данных

Коэффициент корреляции для матрицы Пирсона вычисляется по формуле

$$r_{xy} = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\sigma(x) \cdot \sigma(y)},$$

где x_i – значения, принимаемые первым спектральным каналом (в нашем случае); y_i – значения, принимаемые вторым спектральным каналом; $\overline{x}$ – среднее значение первого канала; $\overline{y}$ – среднее значение второго канала; σ – стандартное отклонение, определяемое по формуле

$$\sigma = \sqrt{\frac{\sum(t_i - \mu)^2}{N}},$$

здесь t_i – значение элемента в выборке; μ – среднее значение выборки; N – количество элементов в выборке.

Основная часть. При сборе информации в полуавтоматическом режиме возникают проблемы, связанные с формированием набора данных.

Первая проблема заключается в несбалансированности классов, количество пикселей для объектов разных типов поверхности имеет разный объём, что негативно влияет на обучение модели.

Вторая проблема выражается в некорректном присвоении метки значению чужого класса. Сформированный набор данных с такими ошибками не позволяет провести обучение модели.

Третья проблема связана с учётом схожих характеристик при формировании набора данных, дублирование характеристик является критичным для задач машинного обучения.

Для устранения обозначенных проблем используются модели предварительной обработки данных.

Алгоритм оптимизации обучающего набора данных разрабатываемой интеллектуальной системы состоит из следующих этапов (см. рис. 3):

1. Сортировка данных. Первоначально данные, поступающие в неупорядоченном виде, не пригодны для обучения модели. Корректный набор данных описывается множеством X, исследуемые классы поверхностей представляются множеством $D\{y_1, y_2, \dots, y_n\}$, где y_n – конкретный класс данных $D \in X$. Порядок множества D может быть нарушен, т. е. пиксели различных классов поверхности не упорядочены. Проблема решается перебором значений класса принадлежности и их

группировкой, т. е. любой пиксель $P_i \in D$ и соседний пиксель $P_{i+1} \in D$ сравниваются, и в случае если один из пикселей принадлежит первому по счёту классу, то он убирается в начало таблицы, сразу после крайнего пикселя того же класса.

2. Удаление повторяющихся значений. В случае некорректного изображения, сформированного для набора данных при совпадающих значениях соседних компонентов палитры, т. е. дублировании одних и тех же спектральных каналов, соответствующие пиксели удаляются.

Пусть в каком-либо классе $y_i \in D$ находится такой пиксель $P_{l,m,k} \in y_i$, что $l = m \cup m = k$, (l, m, k – значение цветовой RGB-палитры), тогда он удаляется из класса y_i.

3. Нахождение среднего значения. Для дальнейшей работы и подсчёта параметров находятся средние значения для каждого спектрального канала.

В задаче определения нефтяных загрязнений используется 2 класса разметки (y_1 и y_2), каждый класс имеет набор из 12 спектральных характеристик $\{x_1, x_2, \dots, x_{12}\}$, подсчитывается среднее значение для каждой спектральной характеристики по формуле

$$mean_{j1,j2} = \frac{\sum x_i}{n_i},$$

где $mean_{j1,j2}$ – средние значения спектральных характеристик $j1$ и $j2$ для первого и второго класса соответственно; x_i – значение столбца i-го класса m-й строки; n – количество строк i-го класса.

4. Нахождение стандартного отклонения. После загрузки двумерного массива значений спектральных характеристик осуществляется перебор значений для всех столбцов таблицы и вычисляются средние значения для каждого. Подсчитывается стандартное отклонение для каждой спектральной характеристики по формуле $standart_{otklonenie} = \sqrt{\left(\frac{sum}{n}\right)}$, где sum – сумма квадратов отклонений; n – количество элементов в столбце.

5. Фильтрация по стандартному отклонению. Каждый пиксель P_i класса разметки y_i определяется набором спектральных характеристик x_i, $i = \overline{1,12}$. Для каждой спектральной характеристики x_i производится подсчёт стандартных отклонений $standart_{otklonenie_i}$, проверяется условие $x_i - standart_{otklonenie_i} < x_i < x_i + standart_{otklonenie_i}$. Если значения спектральной характеристики $x_i \in P_i$ выходит за допустимый интервал, то пиксель P_i из множества D удаляется.

6. Балансировка значений. Имеется пространство классов D $\{y_1, y_2\}$, где y_1, y_2 – классы данных разметки. Для реализации задачи машинного обучения необходимо сбалансировать количество пикселей каждого класса, т. е. количество пикселей $\sum P_{y_1}$ первого класса разметки y_1 не должно превышать количество пикселей $\sum P_{y_2}$ второго класса разметки y_2 более чем на 10 %. Чтобы обеспечить балансировку в установленной границе, осуществляется подсчёт пикселей каждого класса, пиксели, не входящие в допустимый интервал, удаляются из разметки.

На рис. 3 представляется схема работы алгоритма фильтрации данных.

Общий модуль предварительной обработки данных представлен на рис. 4 и реализуется на языке C#.

Тестирование алгоритма производится на данных, гарантированно содержащих информацию с мест реальных разливов нефти. Тестовая разметка предполагает следующие типы поверхностей: «чистая» вода, нефтяная плёнка. С помощью модуля предварительной обработки осуществляется балансировка и сортировка данных, результаты реализации программного модуля демонстрируются на рис. 5.

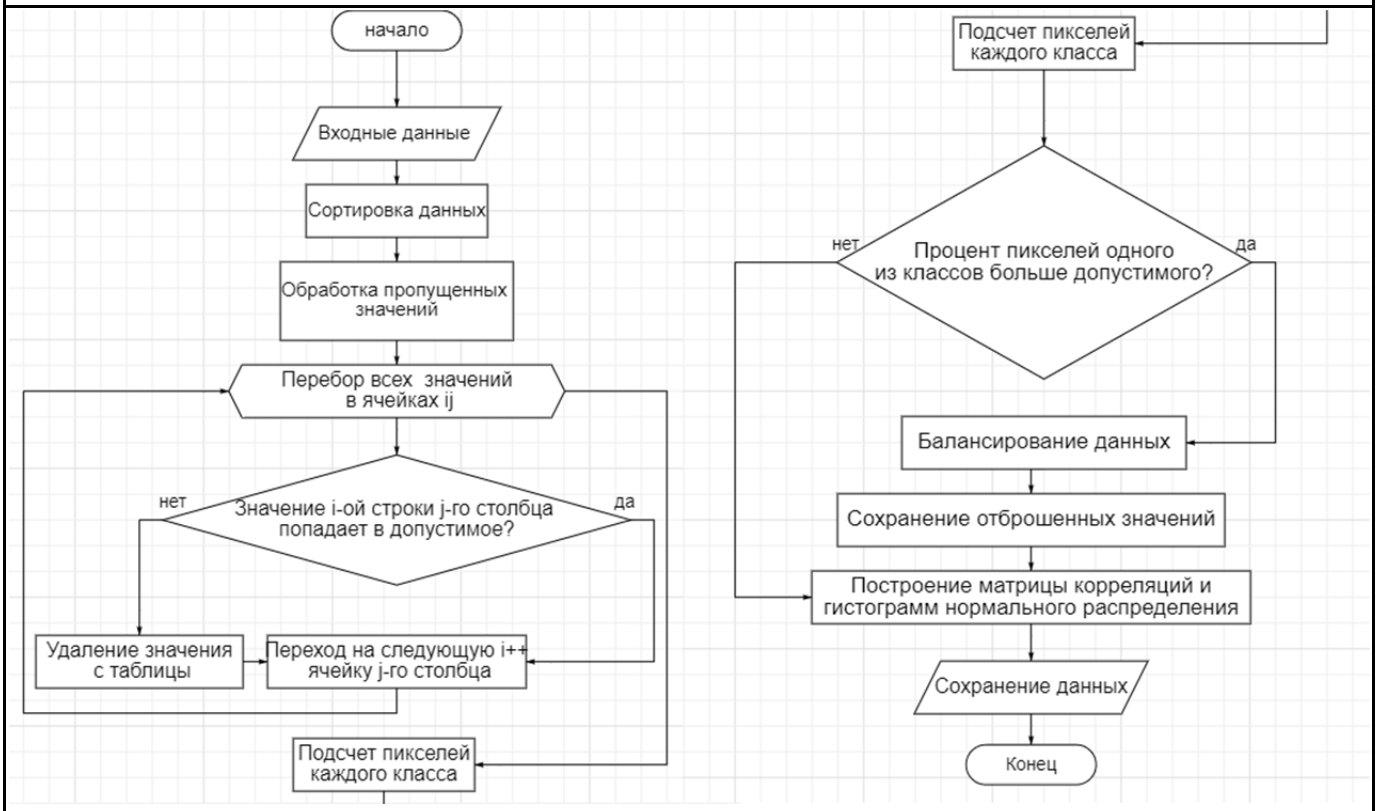

Рис. 3. Блок-схема предварительной обработки данных

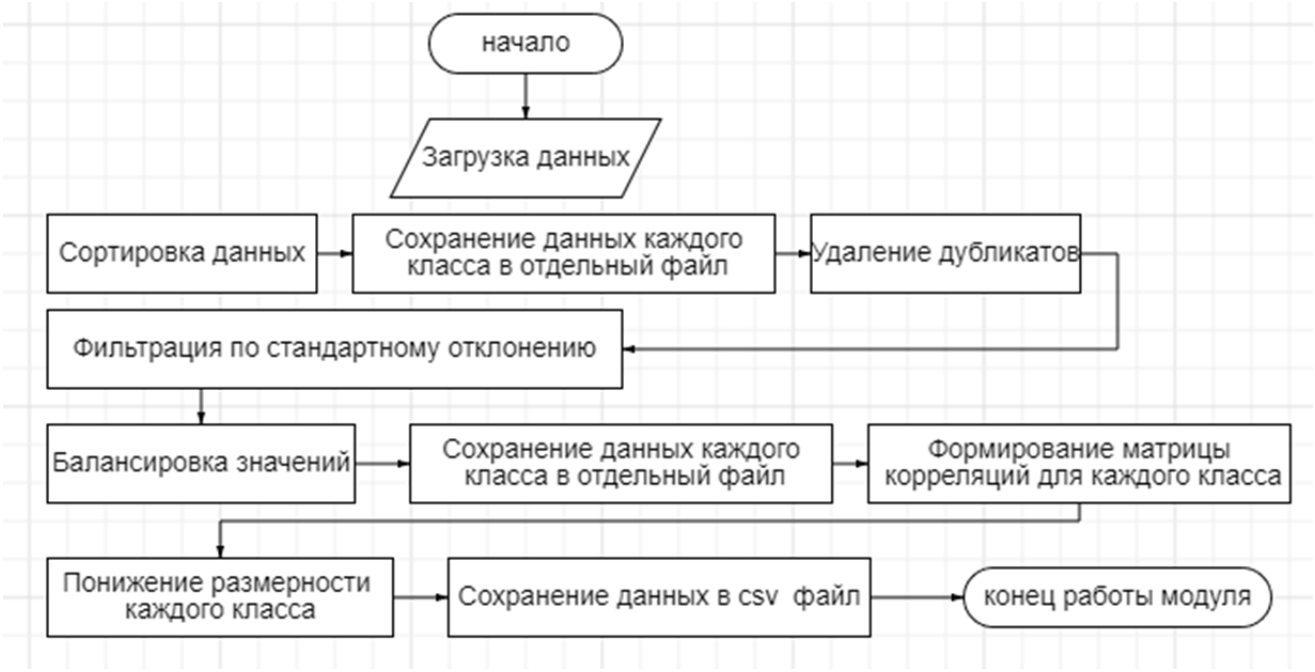

Рис. 4. Общая блок-схема работы модуля

```
На обработку поступило: 228435 пикселей
        Начальное количество пикселей первого класса: 185896
        Начальное количество пикселей второго класса: 42539


                По итогам проверки по стандартному  отклонению :

Удалено пикселей первого класса: 72678  Удалено пикселей второго класса: 16055  Всего удалено пикселей: 88733

      Колисество пикселей первого класса после обработки (удаления) 110026
      Колисество пикселей второго класса после обработки (удаления) 25253


      Колисество пикселей первого класса после обработки (балансирование) 27779
      Колисество пикселей второго класса после обработки (балансирование) 25253
                Всего пикселей в разметке 53032
```

Рис. 5. Предварительная обработка данных

Следующий этап предполагает понижение размерности исходных данных путём исключения набора спектральных характеристик (см. табл. 1), определяемого на основе корреляционных оценок (см. рис. 6).

Таблица 1

Условные обозначения спектральных каналов

Название спектрального канала	Условное обозначение	Название спектрального канала	Условное обозначение
Band 1 – Coastal aerosol	X1	Band 7 – Vegetation red edge	X7
Band 2 – Blue	X2	Band 8 – Vegetation red edge	X8
Band 3 – Green	X3	Band 9 – Water vapour	X9
Band 4 – Red	X4	Band 11 –SWIR	X10
Band 5 – Vegetation red edge	X5	Band 12 –SWIR	X11
Band 6 – Vegetation red edge	X6	Band 8A – NIR	X12

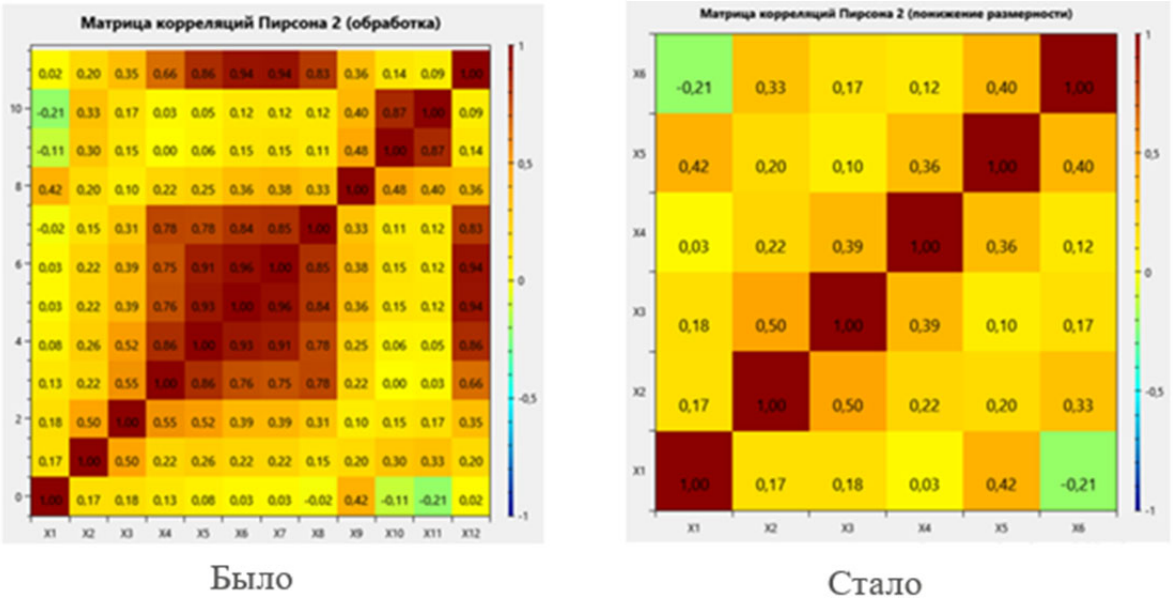

Рис. 6. Матрицы корреляций до и после понижения размерности

Анализ указывает на сильную связь между каналами X4, X5, X6, X7, X8, X12, а также X10, X11. Для понижения размерности исходных данных исключаются каналы X4, X5, X6, X8, X10, X12. Каналы X1, X2, X3, X7, X9, X11 являются базовыми для обучения интеллектуальной системы. Экспертная предметная оценка может влиять на набор оставляемых характеристик, скорректировать выбор.

Заключение. В результате исследования построена модель предварительной обработки данных, разработан и программно реализован алгоритм, апробирующийся на тестовых данных. Результаты работы модуля позволяют понизить размерность первоначального набора данных, формируемого для оценки наличия нефтяных загрязнений водных поверхностей, исключить из рассмотрения 6 спектральных каналов. Полученные данные обеспечивают эффективное обучение реализуемых моделей машинного обучения. Предлагаемый подход применим для оптимизации различных типов данных.

ЛИТЕРАТУРА

1. Детектирование состояния поверхности / Е. П. Жарикова, И. А. Трещев, Я. Ю. Григорьев, А. Л. Григорьева // Учёные записки Комсомольского-на-Амуре государственного технического университета. Науки о природе и технике. – 2019. – № III-1 (39). – С. 58-63.
2. Zharikova, E. P. Surface state detection / E. P. Zharikova, J. U. Grigoriev, A. L. Grigoryeva // 2019 International Multi-Conference on Indus-trial Engineering and Modern Technologies, FarEastCon 2019. – 2019. – P. 8934205.
3. Zharikova, E. P. Methods of remote sensing in forest fund assessment problems / E. P. Zharikova, J. U. Grigoryev, A. L. Grigoryeva // 2019 International Science and Technology Conference «EastConf», EastConf 2019. – 2019. – С. 8725343.
4. Жарикова, Е. П. Применение искусственного интеллекта в задачах анализа состояния акваторий / Е. П. Жарикова, Я. Ю. Григорьев, А. Л. Григорьева // Морские интеллектуальные технологии. – 2021. – Т. 2. – № 2 (52). – С. 129-133.
5. Амосов, О. С. Моделирование обнаружения и распознавания аномального поведения динамических систем / О. С. Амосов, С. Г. Амосова // Управление развитием крупномасштабных систем MLSD'2020. Труды тринадцатой международной конференции / Под общ. ред. С. Н. Васильева, А. Д. Цвиркуна. – М.: Институт проблем управления им. В. А. Трапезникова РАН, 2020. – С. 1151-1158.
6. Zharikova, E. P Applications of computer vision in cross-sectoral tasks / Zharikova E. P., Grigoriev Y. Y., Grigorieva A. L. // Current Problems and Ways of Industry Development: Equipment and Technologies / Warsaw, 2021. – P. 415-426.
7. Жарикова, Е. П. Применение методов машинного обучения в задачах мониторинга мирового океана и континентальных поверхностных вод / Е. П. Жарикова, Я. Ю. Григорьев, А. Л. Григорьева // Учёные записки Комсомольского-на-Амуре государственного технического университета. Науки о природе и технике. – 2022. – № VII (63). – С. 33-40.
8. Жбанов, В. А. Проектирование и разработка модели нейронной сети для определения сходства двух образцов неструктурированных данных / В. А. Жбанов, Е. Б. Абарникова // Учёные записки Комсомольского-на-Амуре государственного технического университета. Науки о природе и технике. – 2023. – № I (65). – С. 47-53.
9. Остриков, В. Н. Влияние предварительной обработки данных гиперспектральной съёмки на качество их тематического анализа / В. Н. Остриков, О. В. Плахотников // Исследование Земли из космоса. – 2014. – № 1. – С. 29.