

ПРИМЕНЕНИЕ АППАРАТА МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ ДЛЯ ИССЛЕДОВАНИЯ ДАННЫХ С МНОГОЛУЧЕВОГО ЭХОЛОТА



Н. В. Шашин, студент 2-го курса магистратуры
E-mail: nikita.shashin@digital-klgtu.ru
ФГБОУ ВО «Калининградский государственный
технический университет»

В статье рассматривается методика обработки акустических данных с многолучевого эхолота с применением элементов математической статистики. Проведено сравнение результатов применения методики, основанной на математической статистике, с результатами разметки данных специальным программным обеспечением. Отражены достоинства и недостатки применяемой методики.

Ключевые слова: *обработка данных, разметка данных, Data Mining, кластеризация, математическая статистика, машинное обучение, обратное рассеяние, многолучевой эхолот.*

ВВЕДЕНИЕ

Современные технологии гидроакустического профилирования, в частности многолучевая эхолотация, предоставляют океанологам, геологам и инженерам уникальную возможность получать массивы данных о морском дне. Большой объем информации, генерируемый многолучевыми эхолотами, содержит в себе не только батиметрические измерения, но и параметры акустического обратного рассеяния, которые могут быть использованы для оценки состава и состояния донных отложений.

С ростом объемов собираемой информации возникает необходимость в поиске эффективных методов ее автоматической обработки. Традиционные подходы, основанные на ручной интерпретации и экспертной оценке, оказываются недостаточно эффективными при работе с данными больших объемов (или Big Data), особенно когда требуется быстрое и объективное выявление закономерностей в распределении типов морских осадков. В этой связи применение методов математической статистики, анализа данных и машинного обучения становится неотъемлемой частью современных исследований морского дна.

Актуальность задач анализа данных многолучевых эхолотов обусловлена широким спектром их практических применений – от прокладки подводных коммуникаций и выбора площадок для бурения до мониторинга экологического состояния морских акваторий. Применение алгоритмов анализа данных многолучевой эхолотации представляет собой важное направление современных океанологических исследований, способствующее повышению эффективности изучения морского дна и открывающее новые возможности для комплексного освоения морских ресурсов.

Методы автоматической обработки данных с многолучевого эхолота могут быть интегрированы с системами реального времени для оперативного картографирования морских акваторий и поддержки принятия решений при подводных инженерных работах.

ОБЪЕКТ ИССЛЕДОВАНИЯ

Объектом исследования являются данные, полученные с использованием многолучевого эхолота, на основе которых проводится анализ и последующее картографирование морского дна с целью выделения участков, различающихся по своему гранулометрическому составу.

ЦЕЛЬ И ЗАДАЧИ ИССЛЕДОВАНИЯ

Целью исследования является применение математико-статистического подхода к интерпретации гидроакустических данных, полученных сотрудниками Института океанологии им. П. П. Ширшова, для повышения эффективности и скорости обработки больших массивов поступающей информации с последующим разделением морского дна в отдельных районах Балтийского моря по типам донных отложений.

Задачами текущего исследования является:

- интерпретация данных с многолучевого эхолота с точки зрения математической статистики;
- реализация методики на языке программирования Python;
- сравнение результатов разделения морского дна по типам донных отложений с результатами разметки данных Института океанологии им. П. П. Ширшова.

МЕТОДЫ ИССЛЕДОВАНИЯ

В ходе научных экспедиций в Балтийском море на базе Института океанологии им. П. П. Ширшова добываются огромные массивы данных с применением многолучевого эхолота Reson SeaBat T50-R. Эти данные представляют собой облака точек, которые собраны в грид. Каждая точка содержит в себе информацию о том, в каком месте по широте и долготе было проведено измерение, с какой глубины и с какой интенсивностью акустического обратного рассеяния отражается сигнал от морского дна [1, 2]. На основе этих данных строится модель рельефа морского дна и мозаика обратного рассеяния в оттенках серого. Мозаика обратного рассеяния представлена на рисунке 1.

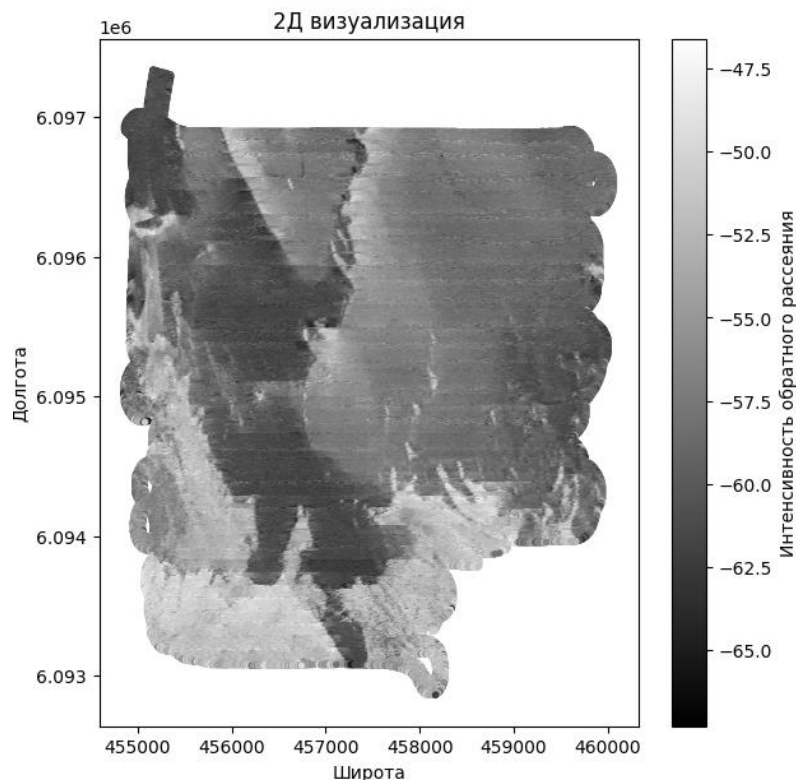


Рисунок 1 – Мозаика интенсивности обратного рассеяния

Изучение модели рельефа морского дна и распределения по нему донных отложений представляет собой серьезный интерес у океанологов. Совместный анализ результатов этих действий помогает в изучении антропогенного влияния на морское дно [2].

Построить цифровую модель морского дна гораздо легче, чем интерпретировать данные акустического обратного рассеяния. Если для построения карты глубин достаточно широты, долготы и глубины, то для интерпретации данных обратного рассеяния необходимо понимать физический смысл его измерения. Теоретически, по значению интенсивности обратного рассеяния можно определить тип поверхности морского дна, так как от каждой поверхности акустический сигнал отражается по-разному. Например, интенсивность обратного рассеяния зависит от твердости и шероховатости грунтов [3].

На текущий момент для работы с данными акустического обратного рассеяния океанологи применяют программное обеспечение QGIS с дополнительным плагином Semi-Automatic Classification Plugin [2]. Результатом работы становится классификация мозаики обратного рассеяния, которая условно делит морское дно по областям с разными типами донных отложений. Алгоритм, который встроен в плагин, основан на одном из способов машинного обучения – обучение с учителем. Экспертом на изображении выделяются эталонные области или регионы, о которых имеется представление с точки зрения того, какой состав поверхности морского дна в них преобладает. Данным областям присваиваются метки или классы, на основе которых алгоритм будет считать эти области эталонными. В результате происходит сравнение пикселей всего изображения мозаики обратного рассеяния с эталонными областями и происходит классификация областей. Данная методика показывает хорошие результаты, но содержит в себе несколько ключевых недостатков:

- 1) разметка мозаики обратного рассеяния проводится экспертом вручную, из-за чего обработка данных может занимать много времени;
- 2) для разметки изображения и последующего применения алгоритма классификации необходимы результаты пробоотбора в разных точках исследуемого участка.

На основе этого можно сделать вывод, что классификация мозаики обратного рассеяния – довольно трудозатратный процесс. Для качественной классификации необходимо как можно больше точек пробоотбора, но проводить пробоотбор с большой частотой сбора образцов нет технической возможности. Сам пробоотбор даже в 10 точках может занимать довольно продолжительное время.

Для экономии ресурсов океанологов и сокращения времени обработки подобных данных предлагается применение альтернативной методики обработки данных акустического обратного рассеяния, основанной на математической их интерпретации. Данная методика была применена в работах Терёхиной Яны Евгеньевны [4, 6, 7]. В этих работах рассматривались данные, собранные в Кандалакшском заливе Белого моря, на основе которых была построена карта областей по разным типам донных отложений. Методика показала отличные результаты на данных гидролокатора бокового обзора, и было принято решение использовать данную методику для данных многолучевого эхолота Института океанологии им. П. П. Ширшова.

Предположим, что сбор данных многолучевым эхолотом – некий статистический эксперимент. Интенсивность обратного рассеяния – результат статистического эксперимента, случайная величина. Предполагается, что данная случайная величина подчиняется некоторому закону распределения. Закон распределения необходимо описать с помощью статистических параметров – атрибутов [4]:

- среднее;
- медиана;
- мода;
- разница между медианой и средним;
- разница между модой и средним;
- коэффициент эксцесса;

- коэффициент асимметрии;
- среднеквадратическое отклонение.

Каждый атрибут имеет свой физический смысл. На основании этого атрибуты можно условно разделить на две группы:

- атрибуты, характеризующие размерность донных отложений: среднее, медиана, мода, разница между модой и средним, разница между медианой и средним;
- атрибуты, характеризующие однородность донных отложений: среднеквадратическое отклонение, коэффициент асимметрии, коэффициент эксцесса.

В общем виде методика выглядит следующим образом [4, 6]:

1. Выбирается окно расчета.
2. По данным батиметрии рассчитываются глубины, угол и азимут наклона поверхностей.
3. По данным гидролокатора бокового обзора рассчитываются атрибуты обратного рассеяния.
4. Выполняется подготовка атрибутов к кластерному анализу методом K-средних.
5. Проводится кластерный анализ.
6. Строятся карты по каждому атрибуту и итоговая карта кластеров.

По текущей методике видно, что для работы с данными применяется концепция Data Mining, в которую входит кластеризация.

Данные, которые использовались в работах [4–7], имеют несколько другую структуру. Гидролокатор бокового обзора собирает данные об интенсивности обратного рассеяния с двух направлений – продольного и поперечного, то есть интенсивность обратного рассеяния в данном случае – двумерный массив [5]. Для данных подобной структуры рассматриваемая методика оказалась подходящей. В работе [6] указано, что данные продольного направления лучше всего описывают рельеф морского дна, а данные поперечного направления дают понимание о типах донных отложений. Поскольку результаты измерения обратного рассеяния собраны многолучевым эхолотом по поперечному направлению, то указанная в [4, 6, 7] методика подлежит проверке на данных Института океанологии им. П. П. Ширшова.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Указанная ранее методика была частично реализована на языке программирования Python. В исследовании используются данные Института океанологии, включающие в себя данные акустического обратного рассеяния и данные батиметрии (глубины), и, поскольку рассматривается интерпретация атрибутов непосредственно обратного рассеяния, было принято решение угол и азимут наклона поверхностей не рассчитывать.

Для расчета атрибутов использовались библиотеки NumPy и SciPy. Атрибуты рассчитывались в рамках окна, которое задается вручную. Оптимальный размер окна варьирует от 5 до 10 значений каждого атрибута в зависимости от размерности исходных данных. Координата точки значения атрибута рассчитывается как среднее значение широты и долготы в рамках рассматриваемого одного окна.

Рассчитанные атрибуты подвергаются нормализации по формуле:

$$normalized = \frac{X - X_{cp}}{\sigma},$$

где X – рассчитанное значение атрибута, X_{cp} – среднее значение атрибута, σ – среднеквадратическое отклонение.

Стоит отметить, что нормализация сугубо индивидуальна для каждого набора данных. Для какого-то участка данных нормализация положительно влияет на результаты кластеризации, а для какого-то участка данных нормализация может быть излишней. Результаты кластеризации могут выдавать менее точную картину кластеров относительно «эталонных» ре-

результатов классификации мозаики обратного рассеяния в QGIS. В частности, ухудшается кластеризация с точки зрения однородности какой-либо области, то есть определенный кластер единой структуры до нормализации атрибутов становится разделенным на множество других кластеров, кластер получается смешанным. Вероятно, так происходит из-за качества исходных данных.

Для кластеризации входными данными являются результаты оценки каждого атрибута методом главных компонент. Физический смысл данного метода применительно к атрибутам обратного рассеяния выглядит следующим образом: в ходе вычислений выяснилось, что вектор-компонента, который объясняет большую часть изменчивости атрибутов между окнами, характеризует размерность донных отложений, а вектор-компонента, который объясняет оставшуюся дисперсию, характеризует однородность морского дна. Соответственно, построение кластеров было основано на различиях между атрибутами по размерности донных отложений, а однородность поверхности морского дна была второстепенным фактором. Не исключено, что на иных наборах данных однородность морского дна может иметь более весомое значение.

Для запуска алгоритма кластеризации методом К-средних задается количество кластеров, определенное методом локтя. Построение карт атрибутов и итоговой карты кластеров реализовано с применением библиотеки Matplotlib. Для атрибутов строится тепловая карта, которая отражает изменение значений определенного атрибута относительно участка морского дна. Пример тепловой карты атрибута «Среднее» представлен на рисунке 2.

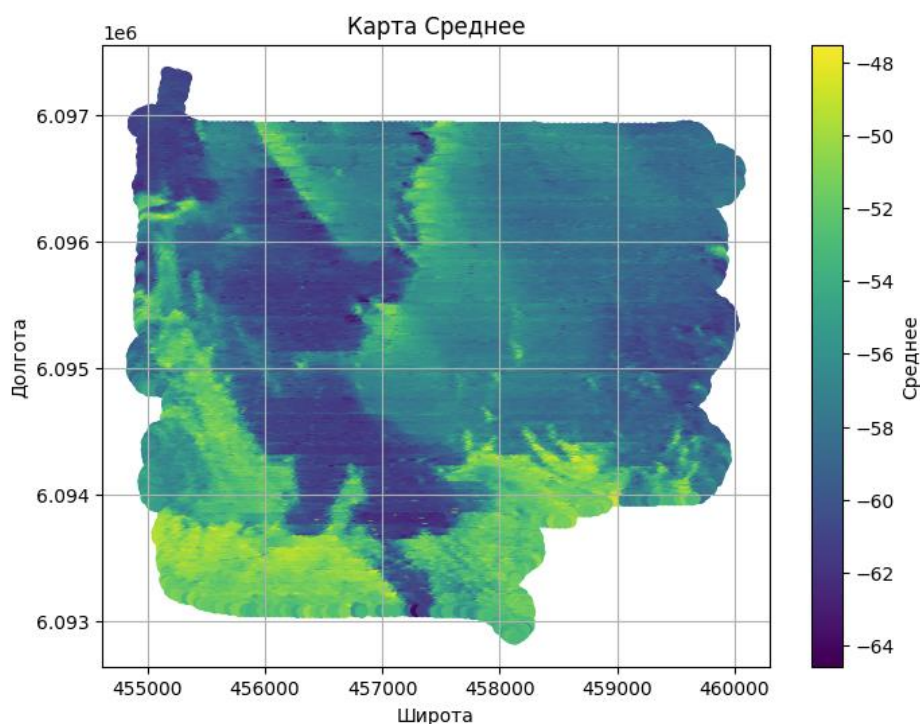


Рисунок 2 – Тепловая карта распределения значений атрибута «Среднее» на участке морского дна

Итоговая карта кластеров представлена на рисунке 3.

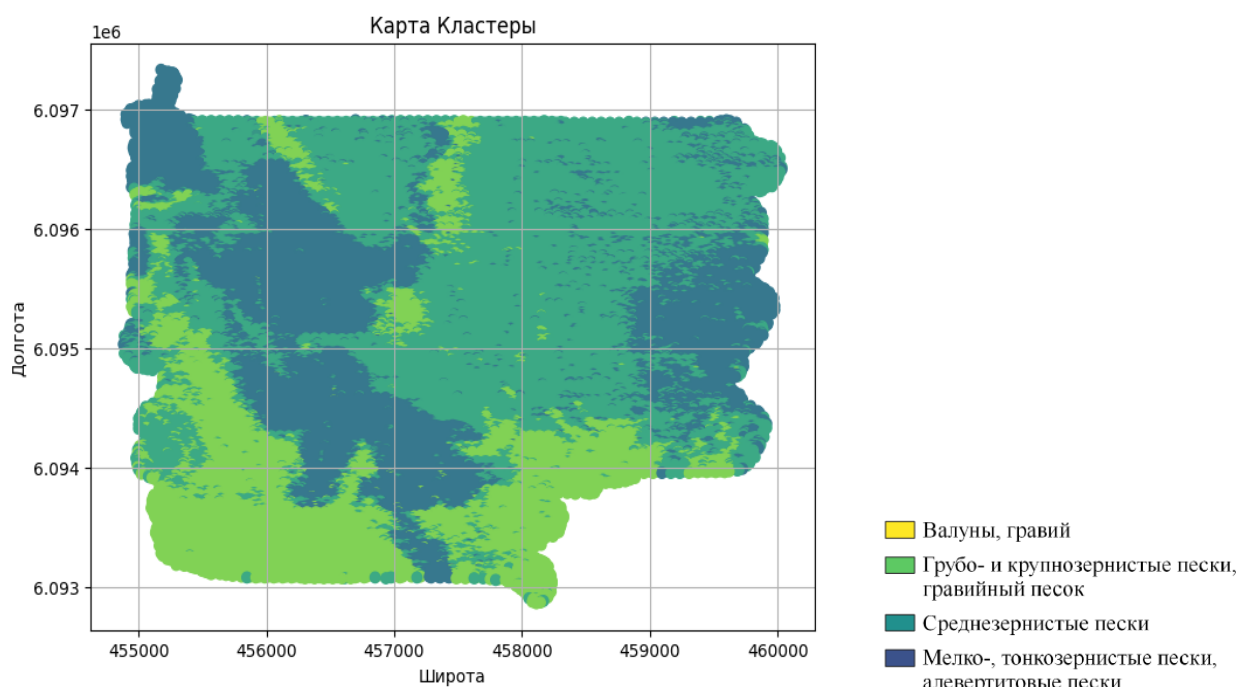


Рисунок 3 – Результат разделения морского дна с помощью кластеризации атрибутов обратного рассеяния методом К-средних

Результаты кластеризации атрибутов обратного рассеяния в большей степени повторяют результаты полуавтоматической разметки данных, полученные сотрудниками Института океанологии им. П. П. Ширшова: паттерны кластеров совпадают с результатами работы плагина в QGIS. Из минусов стоит отметить, что алгоритм кластеризации не смог показать четвертый кластер, который характеризует иной тип донных отложений – валуны. На рисунке 4 представлены результаты классификации мозаики обратного рассеяния, выполненной в Институте океанологии им. П. П. Ширшова.

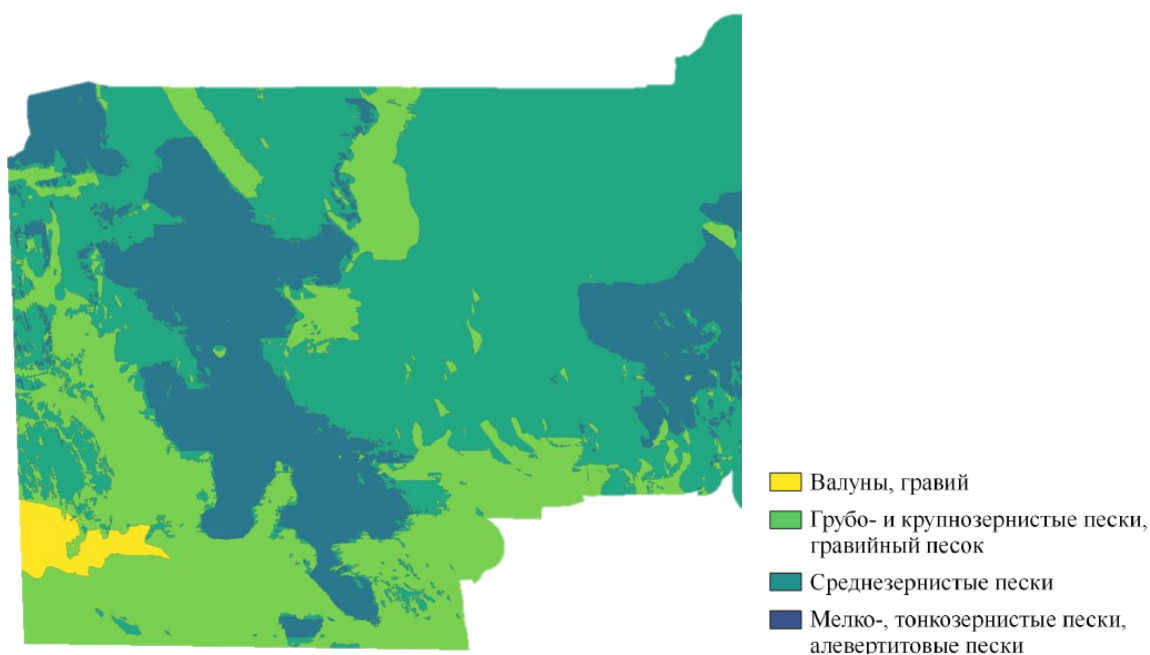


Рисунок 4 – Классификация мозаики обратного рассеяния в программе QGIS

Если на исходной мозаике обратного рассеяния на рисунке 1 обратить внимание на тот участок, где сотрудниками Института океанологии была классифицирована эта область как «валуны», то данный участок визуально не отличается от области, в которой он находится. Тем самым можно предположить, что участок с валунами был размечен либо на основании пробоотбора, который был в этой области, либо с применением экспертной оценки. Тем не менее общее представление об исследуемом участке морского дна, применив алгоритм кластеризации, можно получить без предварительного пробоотбора, что можно отнести к достоинствам данной методики.

ЗАКЛЮЧЕНИЕ

В ходе исследования к данным морского дна Балтийского моря была применена методика, основанная на кластеризации методом К-средних статистических атрибутов интенсивности обратного акустического рассеяния, с целью разделения областей морского дна по типам донных отложений.

Кластерный анализ методом К-средних позволил разделить тестовые участки морского дна на области с различными типами донных отложений. Выделенные кластеры были интерпретированы как области с преобладанием мелкозернистых, среднезернистых и крупнозернистых осадков. Сопоставление карт кластеров с результатами классификации мозаики обратного рассеяния в QGIS подтвердило эффективность применения методики к данным многолучевого эхолота. Применяя методику интерпретации данных методами математической статистики, возможно строить карты морского дна с разделением по типам донных отложений без предварительного пробоотбора.

Результаты применения методики с точки зрения разделения морского дна по зернистости донных отложений можно считать положительными.

СПИСОК ЛИТЕРАТУРЫ

1. Технологии гидроакустических исследований морского дна и донных отложений на Карском шельфе. Предварительные результаты экспедиции TTR-21 / И. Ю. Дудков, С. Н. Понасенко, В. М. Меркурьева [и др.] // Морские исследования и образование (MARESEDU)-2022: Труды XI Международной научно-практической конференции (Москва, 24–28 октября 2022 г.). – Т. IV (IV). – Тверь: ООО «ПолиПРЕСС», 2022. – С. 259–263. – EDN UIRAEE.
2. Морские исследования и образование (MARESEDU-2023): Труды XII Международной научно-практической конференции (Москва, 23–27 октября 2023 г.). – Тверь: ООО «ПолиПРЕСС», 2024. – 505 с. – ISBN 978-5-605-16931-4. – EDN RTVRZB.
3. Yang, Y. Acoustic quantitative analysis of ferromanganese nodules and cobalt-rich crusts distribution areas using EM122 multibeam backscatter data from deep-sea basin to seamount in Western Pacific Ocean / Y. Yang [et al.] // Deep Sea Research Part I: Oceanographic Research Papers. – 2020. – Vol. 161. – P. 103281.
4. Терёхина, Я. Е. Количественный анализ гидроакустических данных для картографирования абiotических компонент подводных ландшафтов / Я. Е. Терёхина, М. Ю. Токарев // Геофизика. – 2018. – № 3. – С. 153–160. – EDN AGVWCM.
5. Атрибутный анализ данных гидролокатора бокового обзора для картографирования подводных ландшафтов / В. Е. Галаев, В. В. Козловский, А. В. Новоявчев [и др.] // Морские исследования и образование (MARESEDU-2017): Труды VI Международной научно-практической конференции (Москва, 30 октября – 2 ноября 2017 г.). – Москва: ООО «ПолиПРЕСС», 2017. – С. 530–534. – EDN JQGVQZ.
6. Терёхина, Я. Е. Геолого-геоморфологические компоненты подводного ландшафта по гидроакустическим данным в Кандалакшском заливе Белого моря: дис. ... канд. геол. наук. – Москва: МГУ, 2024. – 148 с.
7. Терёхина, Я. Е. Технология сбора и анализа геофизических данных для картографирования абiotических компонент подводных ландшафтов / Я. Е. Терёхина, М. Ю. Тока-

рев // Инженерная и рудная геофизика 2021: Материалы 17-й научно-практической конференции и выставки (Геленджик, 26–30 апреля 2021 г.). – Москва: Общество с ограниченной ответственностью «ЕАГЕ ГЕОМОДЕЛЬ», 2021. – С. 167. – DOI 10.3997/2214-4609.202152204. – EDN UAKCQK.

APPLICATION OF MATHEMATICAL STATISTICS APPARATUS FOR THE STUDY OF DATA FROM A MULTIBEAM ECHO SOUNDER

N. V. Shashin, 2nd year Master degree student
E-mail: nikita.shashin@digital-klgtu.ru
Kaliningrad State Technical University

The article discusses the method of processing acoustic data from a multibeam echo sounder using elements of mathematical statistics. The results of applying the method based on mathematical statistics are compared with the results of data marking using special software. Some advantages and disadvantages of the applied method are shown.

Keywords: *data processing, data labeling, Data Mining, clustering, mathematical statistics, machine learning, backscatter, multibeam echosounder.*