Правительство Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего профессионального образования

«Национальный исследовательский университет

«Высшая школа экономики»

Факультет Бизнес-информатики

Отделение Прикладной математики и информатики

Кафедра Анализа данных и искусственного интеллекта

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА

на тему

**«Разработка данных геоинформационных систем»**

Выполнил студент группы 471

Шеменков Никита Эдуардович

Научный руководитель:

доцент

Игнатов Дмитрий Игоревич

Консультант:

к.т.н.

Жукова Наталья Александровна

Москва 2014

**Аннотация**

Наш мир состоит из различных областей, которые содержат огромное количество данных и взаимодействуют друг с другом. Для обнаружения различных закономерностей в объемных базах данных ученые применяют научные автоматические методы, которые являются результатом взаимодействия таких областей науки как искусственный интеллект, машинное обучение, статистика и системы управления базами данных. В принципе, общая проблема состоит в том, чтобы вытащить информацию из баз данных и трансформировать ее в удобный для пользователя вид. Взрыв интереса к географическим исследованиям ничем не отличается от похожих революций в маркетинге, биологии и астрономии. Сегодня, проблемы, связанные с изучением географической среды, привлекают большое количество исследователей к изучению анализа данных и других компьютерных наук. В последнее время разрабатываются новые методы обработки и анализа данных, которые направлены на активное вовлечение экспертов из разных предметных областей.

Предметом исследования этой работы является группировка данных морской среды с учетом знания предметной области на основе методов кластеризации. С помощью нового метода кластеризации Carticlus, мы выделим различные морские территории с разными характеристиками и сравним результаты с реальной ситуацией.

Для моей цели я использую данные о Баренцевом Море, полученные в ходе экспедиций Государственного научного центра "Арктический и антарктический научно-исследовательский институт" с использованием различных океанографических станций. Эти данные представляют собой параметры морской среды:

* Широта
* Долгота
* Температура
* соленость

В рамках написания ВКР были реализованы новый метод кластеризации Carticlus и новый метод преобразования баз данных "картификация". Также были получены результаты этого алгоритма на данных Баренцева моря. Полученные результаты были анализированы и сравнены с реальной ситуацией на Баренцевом море.

**Ключевые слова:** Кластеризация, картификация, Баренцево море.

**Abstract**

Our world consists of different areas which contain a massive amount of data and interact with each other. Scientists apply computer-oriented operations of detecting patterns in voluminous assemblies of data comprehending techniques at the intercrossing of artificial intelligence, machine learning, statistics, mathematic and database systems. Basically, the overall problem is to pick out information from databases and transfashion it up to user-friendly organization. The geographic study outburst is not quite different from resembling revolutions in marketing, biology, and astronomy. Today the problems associated with the study of the geographic environment attract a strong interest in modern Data Analysis and Computer Science. Recently, newly developed methods of data processing and analysis are focused on the active participation of experts of the data domain.

Therefore, the principal goal of this paper is to group data of the marine environment. With the help of the new method of the cluster analysis "Carticlus", I will reveal sea areas with similar trends (parameters) and draw an analogy between obtained information and reality.

For our study purposes we use State Research Center "Arctic and Antarctic Reserch Institute" database of the Barents Sea parameters observations, which was composed of seven main parameters:

* temperature
* total salt content
* latitude
* longitude

As a result I have realization of Carticlus algorithm and results of clasterization of our data.

**Оглавление**

Аннотация................................................................................................................2

Abstract......................................................................................................................4

Введение...................................................................................................................6

Глава 1. Географическая информационная система (ГИС).................................7

1.1 Определение ГИС............................................................................................7

1.2 Области использования ГИС..........................................................................9

1.3 Геоинформационное картографирование.....................................................9

1.4 Примеры ГИС в России................................................................................11

1.4.1 GIS-LAB....................................................................................................12

1.4.2 Интеллектуальная ГИС «Данные наук о Земле по территории России»...................................................................................................................13

1.4.3 Проект АРГО............................................................................................14

1.5 Выводы и результаты по главе.....................................................................19

Глава 2. Картификация.........................................................................................20

2.1 Базовые обозначения....................................................................................20

2.2 Формальное преобразование........................................................................21

2.3 Со-встречаемость объектов..........................................................................22

2.4 Алгоритм картификации ..............................................................................24

2.5 Пример............................................................................................................25

2.6 Метод Carticlus...............................................................................................27

2.7 Результаты......................................................................................................29

2.8 Выводы и результаты по главе.....................................................................40

Заключение.............................................................................................................41

Список литературы................................................................................................42

Приложения...........................................................................................................45

**Введение**

Объектом исследования являются данные о Баренцевом Море. Эти данные представляют собой параметры морской среды:

* Широта
* Долгота
* Температура
* соленость

Эти данные не являются общедоступными и получены в ходе экспедиций Государственного научного центра "Арктический и антарктический научно-исследовательский институт" с использованием различных океанографических станций.

Предметом исследования является группировка данных морской среды с учетом знания предметной области на основе нового метода кластеризации Cartuclus. Так как этот метод относительно новый, то любые эксперименты с его использованием имеют значительный научный и исследовательский интерес.

Для наших целей я буду использовать новый метод преобразования данных, картификация, представленный Bart Goethals в его работе "Cartification: A Neighborhood Preserving Transformation for Mining High Dimensional Data." [15], и основанный на нем метод кластеризации Carticlus [15].

В первой главе моей работы я расскажу о географических информационных системах (ГИС), их областях применения и приведу примеры некоторых ГИС. В третьей главе будет описан метод картификации и результаты его применения к географическим данным Баренцева моря, а также я сравню результаты Carticlus и метода k-means.

**Глава 1. Географическая информационная система (ГИС).**

В настоящее время почти все стороны нашей жизни стали подверглись информатизации, и крайне трудно найти какую-либо область человеческой деятельности - от дошкольного образования до высоких игр государственной политики, - где не чувствовалось бы ее исполинское влияние. Информатика следует по пятам всем остальным наукам, при этом увлекая их за собой, модифицируя, а иногда и закрепощая их в стремлении к недостижимому информационному совершенству.

**1.1 Определение ГИС**

В географии компьютерные технологии произвели геоинформатику и географические информационные системы (ГИС). Геоинформационные системы - эффективный инструмент исследования в географии для сбора, хранения, анализа и визуализации географических данных и прикрепленных к ним информации о необходимых объектах. При этом термин "географические" значит в этом случае не столь "территориальность", а более относится к комплексности и структурировании научного подхода.

В первые ГИС была применены в Канаде и США, примерно, в середине 60-х годов, но на данный момент в промышленно развитых государствах функционируют тысячи ГИС, используемые в таких областях как, политика, экономика, методы оптимизации, экология и образование. Эти системы обхватывают се пространственные уровни: локальный, национальный, региональный, муниципальный, при этом объединяя всевозможную информацию о нашей планете: данные дистанционного зондирования, картографическую, статистику, различные списочные сведения, информацию полевых наблюдений, гидрометеорологические данные, итоги бурения и глубоководного зондирования.

Создание ГИС не простое дело. В этом деле участвуют различные международные организации (ООН, Продовольственная программа), государственные учреждения, всевозможные ведомства и министерства, службы картографии, геологии, управления статистикой, частные фирмы , а также научно-исследовательские институты и университеты. На разработку и поддержание ГИС выделяются значительные финансовые средства, в дело вступают целые отрасли промышленности, проектируется разветвленная геоинформационная структура, с прикрепленными к ней телекоммуникационными сетями.

Что такое ГИС на практике? Ее сущность состоит в том, что с помощью нее можно различными методами собирать информацию, проектировать базы данных, обрабатывать, хранить, преобразовывать и выводить в виде картографической формы, таблицы, графики или текстов. Вообще, с исследовательской точки зрения ГИС - это инструмент моделирования и изучения природных и социально-экономических систем. Эти системы применяются для изучения тех природно-общественных объектов и явлений, которые являются предметом изучения наук о Земле, а также дистанционного зондирования и картографии. Если рассматривать технологический аспект, ГИС является инструментом сбора, хранения, отображения, преобразования и распространения пространственно-координированной географической информации. Рассматривая производственную область, можно сказать, что ГИС - это комплекс устройств и программного обеспечения, используемых для обеспечения управления и оптимизации управления, и самый важный элемент этой системы - автономные картографические системы. В итоге, ГИС можно одновременно понимать как средство научного исследования, метод и результат ГИС-отрасли.

**1.2 Области использования ГИС**

Гис разрабатываются для решения всевозможных задач, которые можно сгруппировать в таком виде:

* Рациональное использование и нахождение природных ресурсов
* Обеспечение кадастра
* Отраслевое и территориальное планирование и оптимальное управление промышленностью, сельским хозяйством, транспортом, финансами и энергетикой
* Наблюдение и контроль условий человеческой жизни , рекреаций, безработицы и др.
* Наблюдение опасных природных явлений и экологических ситуаций, оценивание воздействия техногенных процессов на окружающую среду и их последствия
* Образование
* Поддержка деятельности силовых структур и правоохранительных органов
* Картографирование: производство тематических карт и атласов , поддержка существующих карт, а также оперативное картографирование

Таким образом, множественность использования ГИС для различных сфер порождает большую группу их типов, различающихся по тематике, назначению и территориальному охвату.

**1.3 Геоинформационное картографирование**

Соединение картографии и географической информатики обернулось появлением на научном горизонте нового направления – геоинформационного картографирования. Его суть состоит в автоматизации картографически-информационного моделирования социально-экономических и природных геосистем, основывающихся на ГИС и базах знаний.

В настоящее время картография испытывает перестройку, похожую на изменения, которые происходили во времена перехода от рукописных карт к печатным оттискам. Если смотреть глобально, то в некоторых областях геоинформационное картографирование почти полностью вытеснило традиционные методы составления карт и их издания.

Геоинформационное картографирование отличается, от других методов картографирования, четкой целевой установкой и прикладным характером. На данный момент, около 80% карт, спроектированных с помощью ГИС, имеют прогнозный или оценочный характер, моделируют то или иное районирование территории.

ГИС позволяет соединить две основных ветви картографии: создание и использование карт. Теперь почти все трудоемкие ранее операции, которые были связаны с расчетом длин и площадей, модифицированием изображений или их совмещением, стали простыми процедурами. Появилась динамическая электронная картометрия. Получается, что использование и создание карт стали почти единым процессом, т.к. в процессе информационного анализа происходит непрерывное модифицирование и взаимное сравнение изображений.

Оперативное картографирование – еще одно порождение ГИС-технологии. Оперативное картографирование подразумевает использование и создание карт в реальном времени для своевременного предупреждения неблагоприятных процессов, инвентаризации комплексов, оптимального управления в различных сферах деятельности – от политических событий до военных действий. Кадастровая информация, результаты референдумов, наблюдения и замеры, являются исходными данными для данного вида картографирования.

Отдельно стоит выделить возможности картографической анимации. Различные анимационные программы производят смену кадров карт, изменение положения отдельных участков на экране, динамику трехмерных диаграмм, дискретизации районов местности и т.п.

В ближайшем будущем все возможные перспективы модернизации Земляных наук связаны с геоинформационном картографированием. Так как они позволяют избежать подготовки различных печатных карт.

Новые атласы и карты не захламляют шкафы и не пахнут краской. Они смотрят на нас с экрана компьютера и изменяют свою цветовую гамму и вид, подстраиваясь под наше настроение. Весь процесс движется к тому, что можно будет создавать высоко детализированные голограммы местности, а модели пейзажей заменят живописные полотна.

**1.4 Примеры ГИС**

В настоящее время ГИС-технологиям в Росси уделяется должно внимание, так как возросла потребность в геоданных. Повышение конкуренции среди производителей геоинформационных технологий дает некоторые преимущества пользователям ГИС. Открытость и доступность программных средств позволяет использовать и модифицировать программы. Появляются различные пользовательские клубы, телеконференции, которые территориально разобщены, но связаны единой тематикой. Начинает формироваться геоинформационная мировая инфраструктура. Рассмотрим некоторые примеры ГИС-проектов в России.

**1.4.1 GIS-LAB**

GIS-Lab [3] («ГИС Лаборатория») создан 18 ноября 2001 г. и является независимым информационным ресурсом посвященным Географическим информационным системам (ГИС) и Дистанционному зондированию Земли (ДЗЗ) и, одновременно, сообществом людей занимающихся и интересующихся этими областями знаний.

GIS-Lab - неформальное сообщество, размещенное в виртуальном пространстве сети Интернет. Сайт GIS-Lab не принадлежит и не является частью какой-либо определенной организации и не является домашней страницей конкретного человека. Работой GIS-Lab управляет группа активных участников, вложивших значительное время в развитие сайта и сообщества, создавая статьи, данные, участвуя в проектах.

Они не делают разницы между специалистами и обычными людьми интересующимися геотехнологиями. Участником их сообщества может быть любой человек, независимо от его образования, возраста и квалификации.

Цели:

* Всеобщий, не дискриминирующий доступ к геоданным и геотехнологиям;
* Создание и поддержка открытого сообщества специалистов в области геотехнологий, активно помогающих друг другу.

Задачи:

* Создание обучающих материалов различного уровня сложности;
* Создание новых и организация существующих открытых данных;
* Исследование новых технологий и разработка программного обеспечения;
* Организация и поддержка коллективных проектов важных для достижения целей;
* Разносторонняя поддержка начинающих пользователей и специалистов.

На данный момент завершенные проекты [4]:

* Коллективный проект по уточнению географического положения участковых избирательных комиссий по всей стране
* Коллективный проект по уточнению географического положения учреждений и созданию геоданных
* Коллективный проект по переводу материалов OSGeo Live DVD
* Коллективный проект по созданию открытого слоя данных по расположению УИКов г. Москвы
* Открытый коллективный проект по созданию открытого слоя границ субъектов РФ, на базе данных Росреестра
* Открытый коллективный проект по переводу базы данных VMap0, на данный момент - названий населенных пунктов, с уточнением топонимики

**1.4.2 Интеллектуальная ГИС «Данные наук о Земле по территории России»**

Назначение проекта [10]:

Сегодня в мире функционирует значительное число систем геофизических наблюдений и непрерывного мониторинга, что приводит к постоянному росту объемов геолого-геофизических данных, имеющих географическую привязку. Обработка, анализ и представление геолого-геофизических данных, а также вновь получаемых результатов, являются важными задачами современных наук о Земле.

В связи с этим, возникает задача создания современных информационных систем, предоставляющих доступ к данным различной тематики и позволяющих выполнять как тривиальные операции по работе с массивами пространственных данных, так и сложный многоуровневый интеллектуальный анализ. Географические информационные системы (ГИС) — это информационные системы, обеспечивающие сбор, хранение, обработку, отображение и распространение пространственных данных, а также получение на их основе новой информации и знаний о пространственно-координированных явлениях. В настоящее время ГИС имеют очень широкую сферу применения и представляют собой эффективный инструмент для обобщения и полноценного анализа пространственной информации, полученной различными методами.

Для решения разнообразных геофизических задач необходимо проводить анализ и обработку геолого-геофизических данных. В течение последних лет в Геофизическом центре Российской академии наук (ГЦ РАН) ведутся работы по созданию многодисциплинарной аналитической ГИС для поддержки исследований в области наук о Земле.

Основными задачами в разработке ГИС являются интеграция в единой геоинформационной среде баз данных по наукам о Земле и реализация системы представления пространственных данных, которая обеспечивает взаимодействие между ГИС и пользователями посредством тонкого клиента. Главная особенность разработанной ГИС состоит в наличии пространственных данных различных тематик и интегрированных в ней методов дискретного математического анализа, что позволяет решать широкий спектр задач в области наук о Земле и делает возможным принятие управленческих решений в различных областях научной и практической деятельности.

Цели:

* Разработка интеллектуальной ГИС для обеспечения совместных междисциплинарных исследований в области наук о Земле;
* Разработка методов распознавания, классификации и кластеризации и их интеграция с ГИС и базами геоданных.

Направление исследований:

* Анализ и интерпретация геофизических данных с использованием интеллектуальной геоинформационной системы;
* Комплексная оценка ресурсов стратегического минерального сырья;
* Исследование влияния изменений климата на состояние здоровья населения России.

Результат [11]: ГИС «Россия» спроектирована в соответствии с современными представлениями о системах обработки пространственных данных. Представление архитектуры ГИС характеризуется разделением на несколько независимых уровней и основывается на сервис-ориентированной концепции и открытых стандартах. Центральным компонентом современной архитектуры является ГИС-сервер.

**1.4.3 Проект АРГО**

Суть этого проекта [1] - создание, на основе дрейфующих измерительных буев, глобальной и долговечной сети константных океанографических станций.

Эта сеть ежедневно поставляет данные в очень большом количестве. Измерения буев производятся с дискретностью около 10 суток, а нижняя глубина измерений - 2000 м. Таким образом, каждый буй в течение 10 суток плавает на заданной глубине, потом погружается на глубину 2000 м. С глубины 2000 м он всплывает на поверхность, производя измерения температуры и солености. После этого, за 6 последующих часов данные отправляются на некоторые спутники АРГОС, и они уже в последствии пересылают их береговые центры АРГОС. Затем буй вновь опускается на заданную глубину плавания и весь цикл продолжается, пока не сядут батареи( одного заряда хватает, примерно, на 4 года).

Не исключается вероятность того, что буй может попасть в рыбацкие сети или будет выброшен на берег, то есть закончит свой цикл преждевременно. Также, из-за того что буи дрейфуют на заданных глубинах, некоторые участки Мирового океана могут быть неизмеренными на данном промежутке времени. Для устранения этой проблемы производится пополнение буев, а также их повторное использование. В будущем будет возможно самостоятельное движение буев под конец цикла, и изменение параметров вычислений ( глубина и т.д.) с помощью функций обратной связи.

В дополнение к перечисленным параметрам измерений буев, также добавляются определение характеристик подводных ( на глубине дрейфа) течений и поверхностных течений ( за время пребывания на поверхности).

Совокупность измерений температуры, течений, солености, плюс ,определенная из этих параметров, плотность, а также данных со специальных спутников о возвышении морской поверхности предоставляет ученым полную картину состояния океана.

Полученная сеть океанографических станций имеет ценность как для наблюдения за изменениями состояния Мирового океана, так и для метеорологических служб. Молодая океанографическая дисциплина - оперативная океанография, появилась на основе сети поверхностных буев и комплекса прибрежных метеостанций. На рисунке 1 [14] изображена карта буев измерителей по состоянию на февраль 2014 года.

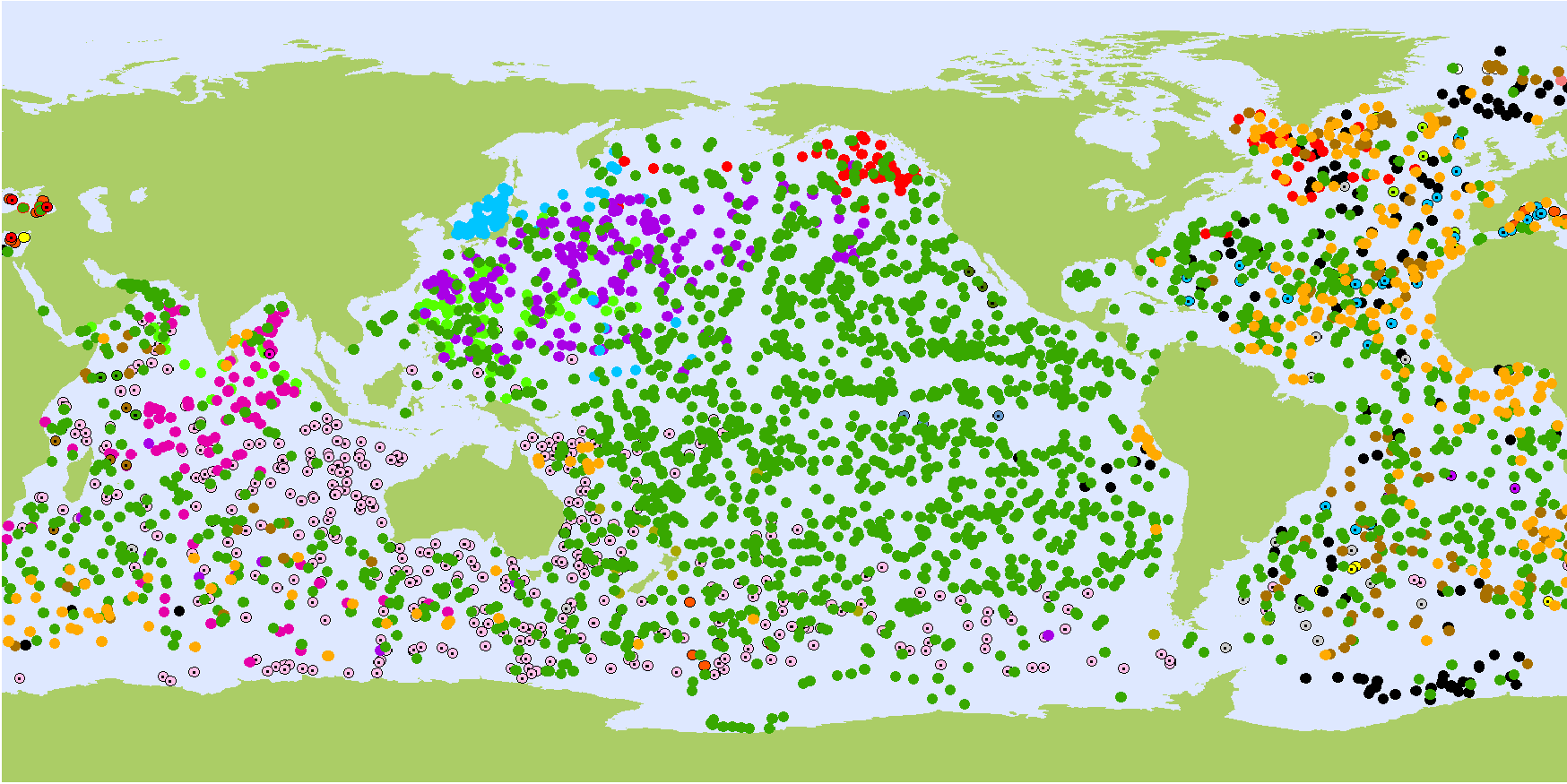


Рисунок 1.Карта распределения буев измерителей (февраль 2014).

Технология производства буев была создана во время проекта по изучению циркуляции вод Мирового океана (WOCE). В настоящее время буи производятся в корпорациях ВЕББ (Фалмут, США), трех организациях США и во Франции.

Все полученные данные измерений с буев отправляются через спутниковые станции в центры данных АРГО , а также в национальные центры данных АРГО.

На данный момент функционируют два всемирных Центра данных АРГО: Монтера (США) и Тулуза (Франция). Национальные центры имеются во всех странах-участницах проекта АРГО (Франция, США, Канада, Великобритания, Германия, Австралия, Южная Корея, Япония)(Карты солености и температуры Мирового океана можно посмотреть в приложении 1).

Все полученные данные проекта АРГО имеют статус свободно доступных для мирового сообщества . А комплексные наблюдения, которые прошли контроль, могут быть получены через национальные центры данных АРГО ( примерная задержка 5 месяцев).

В течение 10 ближайших лет глобальная сеть буев АРГО позволит улучшить понимание процессов, происходящих в Мировом океане, и его влияние на процессы в атмосфере, а именно:

* определить структуру вод Мирового океана и ее изменчивость;
* уточнить характер глобальной циркуляции вод в Мирвом океане;
* оценить меридиональный перенос тепла в океане;
* определить влияние долгопериодных аномалий температуры поверхности океана на изменение циркуляции атмосферы;
* изучить причинно-следственные связи таких явлений, как Эль-Ниньо и др.;
* дать оценку роли Мирового океана в изменении климата.

Спектр задач может быть увеличен в зависимости от обилия данных Мирового океана как по времени, так и по пространству.

Для пополнения данных Мирового океана следует усовершенствовать или разработать:

* методики восстановления параметров поверхности океана на основе спутниковых данных и данных, полученных с буев-измерителей;
* расчетные методы для картирования параметров, позволяющих оценить состояние океана (вертикальное и горизонтальное распределение, динамические высоты, карты течений на поверхности и на глубине 2000м и др.);
* новые численные модели циркуляции океана и усовершенствовать существующие для прогнозов гидрометеорологических параметров;
* процедуры четырехмерного объективного анализа океанических параметров.

**1.5 Выводы и результаты по главе**

В главе было рассказано о Географических информационных системах, областях их применения, решаемых ими задач и их актуальности. Также было рассказано об одной из ветвей геоинформационных систем - Геоинформационном картографировании, которое является соединением картографии и географической информатики. И о ее важной роли в науках о Земле. Был приведен обзор примеров различных ГИС (GIS-LAB и Интеллектуальная ГИС «Данные наук о Земле по территории России» и проект АРГО).

**Глава 2. Картификация**

Bart Goethals в своей работе "Cartification: A Neighborhood Preserving Transformation for Mining High Dimensional Data." [15] рассказывает о новом методе кластеризации для многомерных данных. Этот метод трансформирует каждый объект данных в тележку, в которой находятся k ближайших соседей этого объекта. После этого применяются частые множества [16] [17] для выявления кластерной структуры данных.

Теперь покажем как происходит переход от многомерного набора данных к картифиционной базе данных.

**2.1 Базовые обозначения**

Этот метод переводит информацию близких объектов многомерного массива данных в базу данных транзакций.

*Вход: многомерный массив данных.*

Пусть - набор d измерений в наших данных. Объект данных о - это кортеж значений над . d-размерная база данных, D, набор из n объектов данных, такие что каждый объект из D представим как d-размерный вектор .

Этот метод пытается сохранить в кластерах локально-близкие объекты, которые обладают наибольшей близостью. Поэтому будем создавать многократные "взгляды" на данные, используя различные метрики для каждого "взгляда"(о "взглядах" я расскажу немного ниже).

Будем обозначать множество различных метрик, для проведения всех "взглядов". Индивидуальная мера m - это просто функция, которая показывает взаимную близость объектных пар. К примеру, используем как метрику для множества произвольных признаков . Эти метрики могут быть как меры сходства, так и меры различия, до тех пор пока объекты могут быть упорядочены в соответствии их попарной схожести.

*Выход: транзакционное представление данных.*

Для множества элементов , набор элементов Х определяется как . Транзакция это пара, состоящая из уникального номера транзакции и набора элементов. База данных транзакций - это набор транзакций над . *Помощь* набора элементов Х в базе данных - это число транзакций в , в которых участвует Х, т.е.

Набор элементов называется *частым*, если его помощь больше, чем заданный порог , называемый *минимальная помощь* или *minsup*.

**2.2 Формальное преобразование**

Картификация переводит оригинальный набор данных в объединение различных взглядов на данные, такие что каждый из них вносит информацию о близости на основе заданной на нем метрики.

*Определение 1* (Взгляд объекта) [15] : Пусть - колличество объектов в D, и m мера различия. Тележка объекта о - это кортеж близких к нему элементов, расположенные по убыванию, в соответствии их близости к о. Формально,

также,

Информация близости объекта р, который находится далеко от о, имеет малый вклад в различимость локальных кластерных структур, поэтому можем удалять самых дальних соседей из тележек. Таким образом, тележки могут быть урезаны, чтобы включать только топ-к Ближайших Соседей (кБС) объекта о, относительно метрики m.

*Определение 2* (взгляд меры) [15] : Локальный взгляд отражает информацию близости определенной меры сходства m:

Использование всего лишь одной меры накладывает определённые ограничения, поэтому нужно использовать множественные взгляды , чтобы собрать больше информации о близких объектах.

*Определение 3* (Преобразованная база данных) [15]: Пусть - множество метик, и - локальный взгляд метрики сходства m, тогда картифицированная база данных С будет иметь вид:

После преобразования в тележку каждого объекта и для каждой меры попадают соответствующие близкие объекты. Объединение информации близости из выборки метрик даст нам кластерную структуру.

**2.3 Со-встречаемость объектов**

Частая встречаемость набора объектов Х в С указывает на то, что эти объекты находятся в непосредственно близости в изначальном наборе данных. Таким образом, скорее всего они будут связаны между собой. Каждой метрике соответствует определенный набор отношений в наших данных и картификация извлекает информацию близости отдельно для каждой метрики. Таким образом, дополнительные метрики привносят дополнительную информацию, и точность индивидуальных метрик не ухудшается из-за нерелевантных взглядов.

*Определение 4* (со-встречающийся набор объектов) [15] : *Встречаемость* набора объектов Х – это количество тележек С, которые являются суперсетом Х. Формально:

Кластер Х затем определяется как набор со-встречающихся объектов относительно параметра minsup:

Получается, что мы находим метрику близости набора объектов Х подсчетом со-встречаемости этих объектов в какой-либо тележке в нашей картифицированной базе данных . Также, кластеризованные объекты должны иметь большое количество одинаковых объектов в соответствующей близости, т.е. они демонстрируют высокую взаимную близость к друг другу.

Формально, кластеризированные объекты будут выделены из остаточных объектов для какой-нибудь меры из [15] :

Таким образом, существует набор тележек, которые отражают близость всех кластеризованных объектов Х для каких-то взглядов . При этом близость кластера сохраняется в этих взглядах, т.е. объекты кластеризованны вместе в близости центрального объекта о, а другие объекты q полностью отделились от этого центрального объекта.

**2.4 Алгоритм картификации**

Покажем как работает алгоритм картификации для преобразования исходной базы данных.

Пусть имееться исходная база данных D имеющая атрибутов и набор метик . Получаем такой псевдокод [15] :

Ввод : База данных D, набор метрик

Вывод: Картифицированная база данных , полученная из D

1:

2: For each do

3:

4: For each

5:

6:

7:

8: Return C

Теоретически, необходимо примерно памяти для этого алгоритма, где n - количество объектов, - количество метрик. НО на практике, как было сказано ранее, мы можем обрубать количество объектов в тележке, тем самым уменьшая необходимую память. Худший случай вычислительной сложности - .

**2.5 Пример**

Мной реализован алгоритм картификации в среде MATLAB (Смотри приложение 2).

Для простоты разберем пример для двумерной базы данных( таблица 1):

Табилца 1. Данные

|  |  |
| --- | --- |
| 1 | 2 |
| 2 | 1 |
| 3 | 3 |
| 4 | 4 |
| 5 | 6 |
| 10 | 10 |
| 19 | 19 |
| 20 | 20 |
| 21 | 21 |
| 22 | 23 |

Изобразим на плоскости (Рисунок 1):

Рисунок 1 Плоскоть с изображенными данных

Для каждого объекта была произведена картификация с числом ближайших соседей 3, для евклидовой метрикии для каждого атрибута. Вот какие тележки получаем:

По х :

1 {1 2 3}

2 {2 1 3}

3 { 3 2 4}

4 { 4 3 5}

5 { 5 4 3}

6 { 6 5 4}

7 {7 8 9}

8 { 8 7 9}

9 { 9 8 10}

10 { 10 9 8}

По у :

1 { 1 2 3}

2 { 2 1 3}

3 { 3 1 4}

4 { 4 3 1}

5 { 5 4 3}

6 { 6 5 4}

7 { 7 8 9}

8 { 8 7 9}

9 { 9 8 7}

10 { 10 8 9}

Теперь посмотрим на частоту встречаемости объектов в полученных тележках (Рисунок 2):

Рисунок 2. Частота встречаемости объектов

Видно, что объекты 3 и 9 имеют наибольшую частоту, и как следствие будут центрами кластеров {1 2 3 4 5} и {7 8 9 10} соответственно.

Далее перейдем к описанию метода кластеризации.

**2.6 Метод Carticlus**

Перед описанием метода стоит сказать, что для создания картифицированной базы данных я использовал Евклидову метрику по всем атрибутам нашей базы данных.

Я создал картифицированную базу данных и теперь буду проводить кластеризацию на основе метода, предложенного Бартом [15]. Этот метод кластеризации называется Carticlus и основан на поиске частых множеств [16] [17].

Этот алгоритм получает на вход картифицированную базу данных, минимальную поддержку minsup, количество желаемых кластеров n и минимальную длину кластера minlen. Имеем такой псевдокод [15]:

Ввод: Картифицированная база данных , минимальная помощь minsup, минимальная длина кластера minlen, колличество кластеров n

Вывод: Набор кластеров

3: Repeat

4: X// случайно выбираем частый объект

5: P

6: Repeat

7: // случайно выбираем объект из

8: P

9: Until P =

10: If

11:

12: Until или максимальное число итераций достигнуто

13: // итоговые кластеры

14: for each do

15: for each do

16: if

17:

18: continue with the next X

19:

20: return

Подробнее распишем шаги алгоритма. Создаем множества и и P(линии 1-2). Затем выбираем случайно частый элемент из и помещаем его в X (), затем помещаем в P объекты из , объединение которых с X встречается в нашей картифицированной базе данных больше чем minsup раз (линии 4-5). После переходим во внутренний цикл, в котором в X мы помещаем объединение X и случайного объекта из P, а P видоизменяем и заполняем объектами из , объединение которых с X встречается в нашей картифицированной базе больше чем minsup раз (линии 7-8). Все это мы повторяем до тех пор, пока P не станет пустым. В итоге получаем кластер и если его мощность больше или равно minlen, то добавляем его к множеству кластеров. Таким образом, мы выполняем 4-11 пока мощность не будет больше либо равна n или не будет достигнуто максимальное число итераций. Дальше (линии 14 -20) мы проверяем полученные кластеры на схожесть (линия 16) и похожие кластеры «склеиваем». В итоге получаем набор итоговых кластеров.

**2.7 Результаты**

Теперь рассмотрим некоторые результаты работы метода Carticlus на наших данных (реализацию данного алгоритма в среде Matlab смотри в приложении 3).

Имеем такую карту измерений (2169 объектов измерения) (рисунок 3):

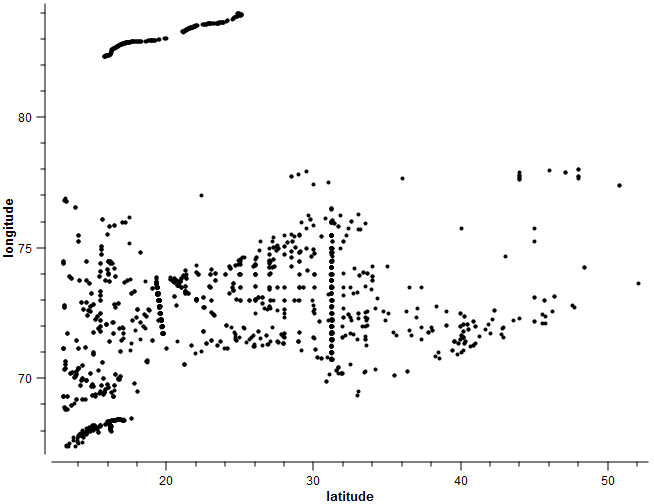


Рисунок 3. Карта измерений за январь 2011.

Метод Carticlus сильно зависит от входных параметров. По этой причине, чтобы получить нормальный вид кластеризации были проделаны множество экспериментов с различными параметрами.

Для начала была изучена структура водных масс Баренцева моря. Их всего 5 видов: Атлантические воды, Северные воды, прибрежные воды, Воды норвежского моря и материковые воды. По этому начальное количество кластеров было выбрано 5. После этого был проделан ряд экспериментов с вариациями минимальной поддержки minsup и длинны тележек k. На рисунках 4 и 5 представлены результаты неудачных экспериментов.

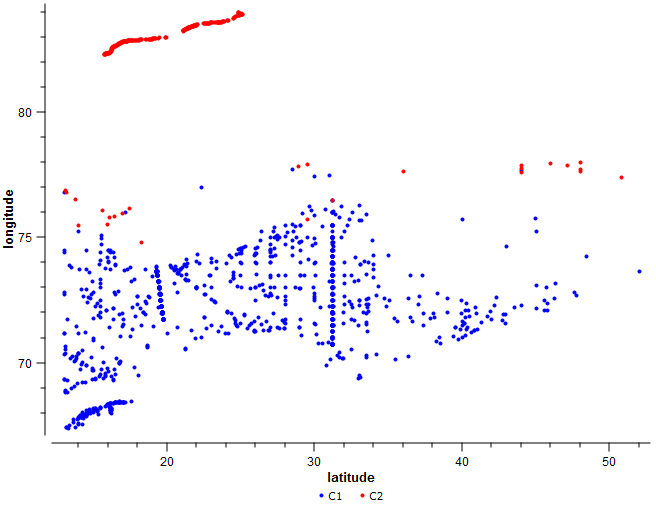


Рисунок 4. Пример неудачной кластеризации. Параметры: k=1700, minsup=5, n=5.

В этом неудачном примере кластеризации слишком мал параметр minsup и в итоге получилось много малых кластеров, которые сливались, так как имели много общих частых объектов. В итоге получился один Большой кластер.

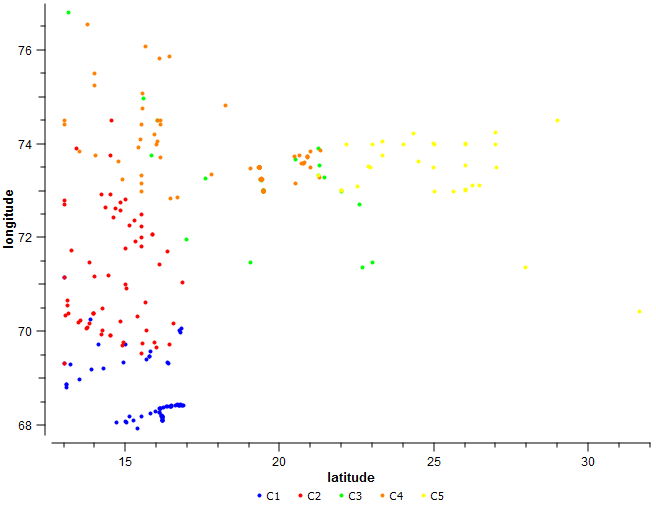


Рисунок 5. Пример неудачно кластеризации. Параметры: k=20, minsup=700, n=5.

В этом примере завышен параметр минимальной поддержки, поэтому алгоритм потерял много объектов.

В итоге после 144 экспериментов я подобрал нормальные параметры для взаимодействия наших данных и метода Carticlus. На рисунке 6 изображен результат кластеризации при нормальных параметрах (k=435, minsup=150, n=5).

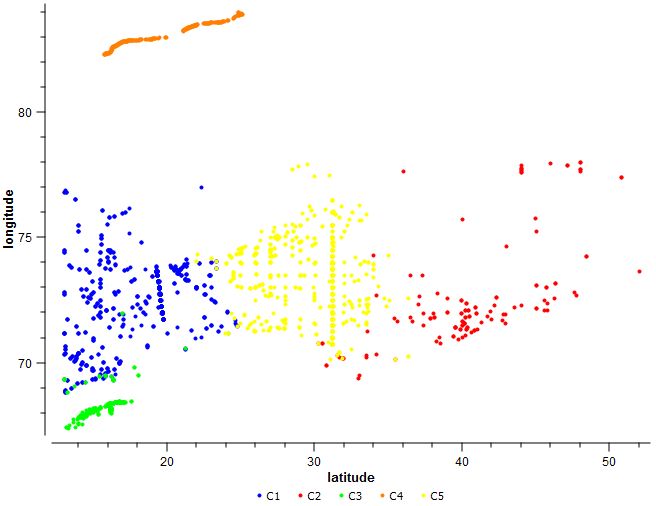


Рисунок 6. Результат кластеризации при k=435, minsup=150, n=5.

Попытаемся соотнести полученные кластеры с реальными данными о Баренцевом море. Известно, что море находится под сильным влиянием теплых вод течения Гольфстрим, поэтому южная и западная его части не замерзают (отметим, что южная и западная части у нас отнесены в отдельные кластеры: желтый и зеленый). Выделяются следующие водные массы: атлантические воды с повышенными температурой и соленостью; арктические воды с пониженными температурой и соленостью; прибрежные воды, поступающие из Белого моря, Норвежского моря и с материковым стоком. Последние характеризуются летом высокой температурой и низкой соленостью, а зимой низкими и температурой, и соленостью. Таким образом, на характеристики Баренцева моря влияет Атлантический океан (синяя и

оранжевая часть графика), Северный ледовитый океан (красная часть графика) и прибрежные воды (зеленая и желтая часть графика). Таким образом при данных параметрах мы получили хорошо интерпретируемые кластеры, которые отражают реальное состояние водных масс Баренцева моря.

Теперь давайте сравним результаты Carticlus с результатами давно известного метода k-means. Для нормальной работы k-means была произведена линейная нормировка данных:

Также нужно определить оптимальное количество кластеров. Критерий, который оптимизирует метод k-средних, записывается как

,

где - это обозначение -го кластера, - центроид -го кластера, - данное наблюдение (объект), а – расстояние между объектами (векторами).

В моей работе количество кластеров я определял с помощью «метода локтя» (Elbow method): для каждого натурального k из некоторого диапазона строится значение целевой функции алгоритма J(R). Количество классов определяется, как значение k, начиная с которой величина J(R) падает не так резко.

Отметим, что встроенный в Orange (программное обеспечение для анализа данных) [18] метод kmeans рандомизированный, поэтому берётся среднее значение J(R) по N запускам.

Результат метода локтядля кластеризации по географическим координатам, солености, температуре, глубине можно увидеть на рисунке 7 (реализовано в Matlab. см. приложение 4) :

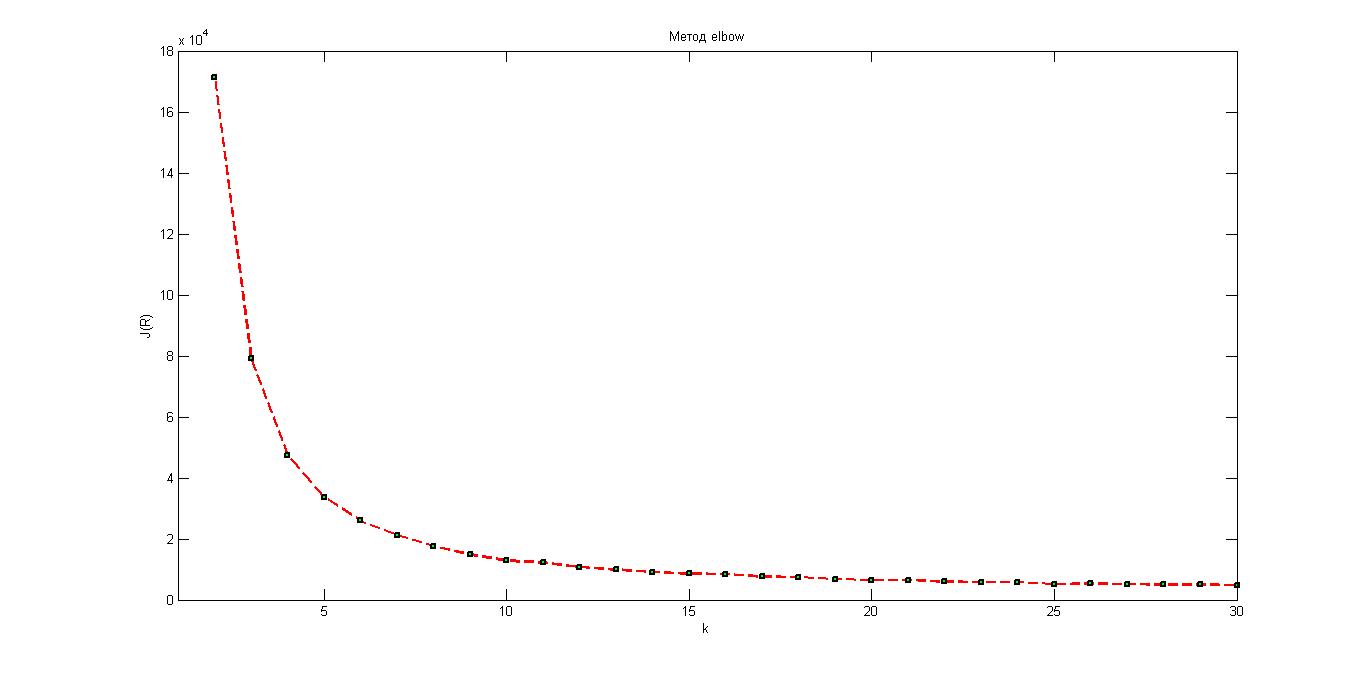


Рисунок 7. Использование метода локтя для определения количества кластеров для k-means.

Судя по графику, начиная с , величина J(R) падает не так резко, поэтому далее будем разбивать данные на 5 кластеров.

Для применения метода k-means я пользовался специальным программным обеспечением для анализа данных Orange [18]. На рисунке 8 изображен результат применения метода k-means к нашим данным.

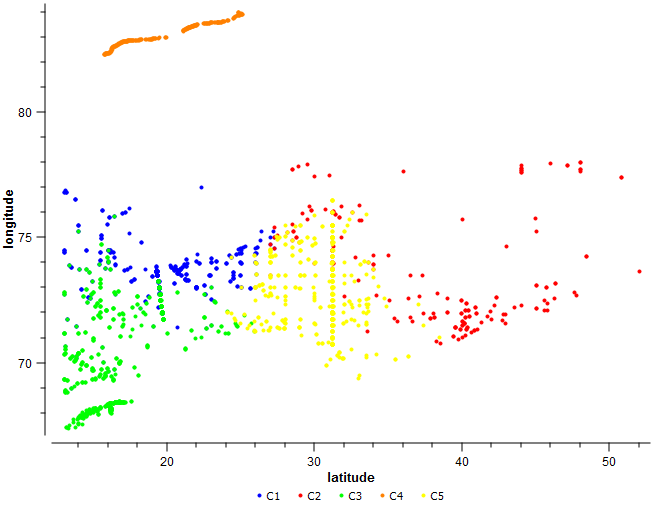


Рисунок 8. Кластеризация методом k-means для k=5.

Из рисунка видно, что k-means тоже выдал результат, который примерно отражает распределение водных масс в Баренцевом море. Сравним распределения температур в кластерах, так как распределение солености для всего моря в январе примерно одинаково (рисунок 9).

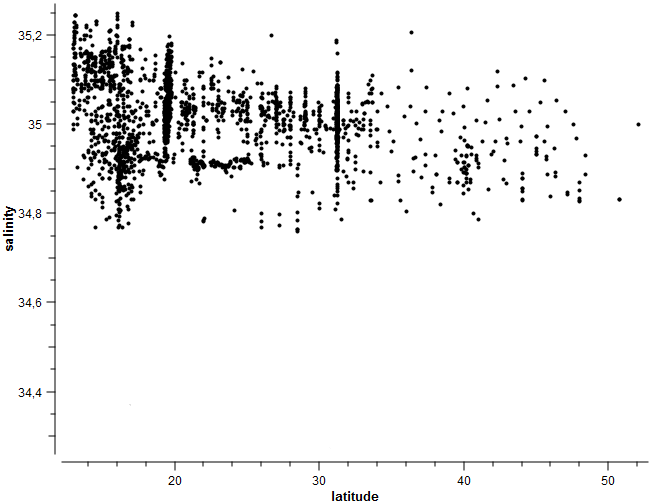


Рисунок 9. Распределение солености среди всех измерений за январь 2011 года.

Из рисунка видно, что по всем измерениям соленость находиться в интервале от 34.7% до 35.5%.

Распределение показаний температур среди итоговых кластеров метода Carticlus и k-means для данных по Баренцевому морю за январь 2011 представлены соответственно на рисунках 10 и 11.

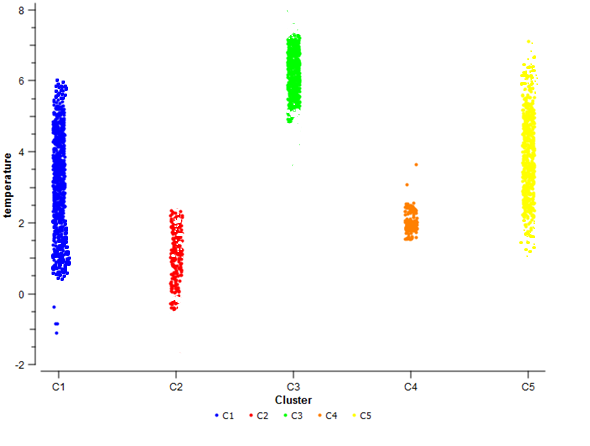


Рисунок 10. Распределение температур среди итоговых клстеров метода Carticlus.

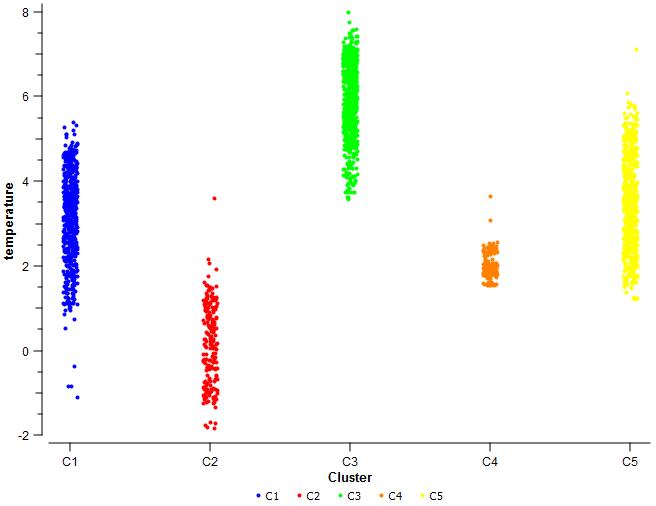


Рисунок 11. Распределение температур среди итоговых кластеров метода k-means.

Если сравнить эти результаты между собой, то можно заметить, что количество объектов отнесенных к красному и зеленому кластерам увеличилось для результатов k-means, а число объектов отнесенных к желтому и синему кластерам уменьшилось. Эта же тенденция наблюдается и для рисунков 6 и 8, а именно :

1) Curticlus (количество объектов в кластерах):

|  |  |
| --- | --- |
| синий | 922 |
| красный | 139 |
| зеленый | 256 |
| оранжевый | 221 |
| желтый | 657 |

2) k-means (количество объектов в клстерах):

|  |  |
| --- | --- |
| синий | 604 |
| красный | 294 |
| зеленый | 612 |
| оранжевый | 221 |
| желтый | 465 |

В итоге получаем, что два метода примерно одинаково отражают реальную ситуацию на Баренцевом море и отличаются структурами кластеров в силу особенностей их алгоритмов.

В конце этой главы я оценю качество кластеров, полученные алгоритмом Carticlus, методом символов.

Пусть a(i) - средняя непохожесть объекта i с другими объектами внутри кластера. a(i) можно интерпретировать как показатель, насколько хорошо объект i принадлежит кластеру.

Пусть b(i) - наименьшая средняя непохожесть объекта i для других кластеров, членом которых он не является.

Тогда можем определить символ объекта:

,где . Т.е. среднее по кластеру показывает насколько тесно сгруппированы объекты.

Найдем символы для каждого кластера алгоритма Carticlus используя Matlab. Результат приведен на рисунке 12.

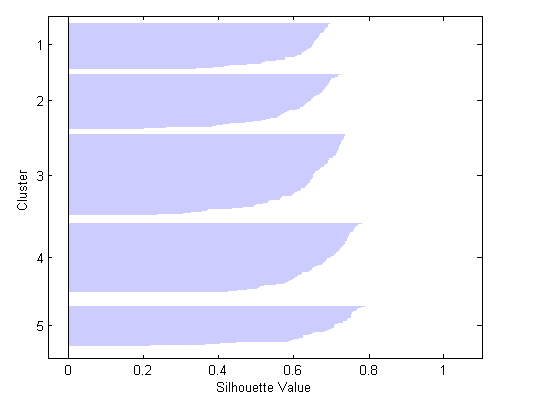


Рисунок 12. Силуэты для кластеров метода Carticlus. Номера кластеров и их цвета в Cartuclus кластеризации: 1- оранжевый, 2- желтый, 3- синий, 4-зеленый, 5 – красный.

Из результата можно заключить, что кластеры полученные методом Carticlus хорошо сформированы.

В заключении нужно сказать, что метод Carticlus крайне сильно зависит от входных параметров, поэтому перед его использованием нужно уделить время для изучения предметной области данных, которые вы собираетесь использовать, а также для подбора нормальных параметров.

**2.8 Выводы и результаты по главе**

В данной главе мы описали метод преобразования баз данных "картификация" [15] и основанный на этом преобразовании метод кластеризации Carticlus [15]. Были представлены примеры использования этих методов и результаты применения их к данным Баренцева моря. Также было проведено сравнение результатов с методом k-means.

**Заключение**

Предметом исследования является группировка данных морской среды с учетом знания предметной области на основе нового метода кластеризации Cartuclus. Так как этот метод относительно новый, то любые эксперименты с его использованием имеют значительный научный и исследовательский интерес.

В рамках написания ВКР были реализованы новый метод кластеризации Carticlus и новый метод преобразования баз данных "картификация". Также были получены результаты этого алгоритма на данных Баренцева моря. Полученные результаты были анализированы и сравнены с реальной ситуацией на Баренцевом море. А также были сравнены результаты Carticlus с методом k-means для наших данных, и на основе этого сравнения были сделаны соответствующие выводы.

В дальнейшем планируется реализация методов автоматического определения входных параметров алгоритма для любых баз данных, дальнейшее усовершенствование алгоритма и разработка новых методов использующих картифицированные базы данных.

**Список литературы**

1. Проект "АРГО" [Электронный ресурс] - Режим доступа: http://rus.ferhri.ru/argo/index\_r.htm
2. Проект "АРГО" [Электронный ресурс]: свободная электронная библиотека - Режим доступа: http://ru.wikipedia.org/wiki/Арго\_(океанография)
3. Gis-lab [Электронный ресурс]: официальный сайт проекта - Режим доступа: http://gis-lab.info/about.html
4. Gis-lab, общественные проекты [Электронный ресурс]: официальный сайт проекта - Режим доступа: http://gis-lab.info/projects.html
5. ГИС "Возобновляемые источники энергии России", Солнечная энергетика [Электронный ресурс]: официальный сайт проекта - Режим доступа: http://www.gis-vie.ru/index.php?option=com\_content&view=article&id=50&Itemid=27
6. ГИС "Возобновляемые источники энергии России", Ветровая энергетика [Электронный ресурс]: официальный сайт проекта - Режим доступа: http://www.gis-vie.ru/index.php?option=com\_content&view=article&id=51&Itemid=53
7. ГИС "Возобновляемые источники энергии России", Гидроэнергетика [Электронный ресурс]: официальный сайт проекта - Режим доступа: http://www.gis-vie.ru/index.php?option=com\_content&view=article&id=52&Itemid=54
8. ГИС "Возобновляемые источники энергии России", Биоэнергетика [Электронный ресурс]: официальный сайт проекта - Режим доступа: http://www.gis-vie.ru/index.php?option=com\_content&view=article&id=53&Itemid=55
9. ГИС "Возобновляемые источники энергии России", Геотермальная энергетика [Электронный ресурс]: официальный сайт проекта - Режим доступа: http://www.gis-vie.ru/index.php?option=com\_content&view=article&id=54&Itemid=56
10. Интеллектуальная ГИС "Данные наук о Земле по территории России" [Электронный ресурс] : официальный сайт проекта - Режим доступа: http://gis.gcras.ru/index.html
11. Интеллектуальная ГИС "Данные наук о Земле по территории России", сервисы [Электронный ресурс] : официальный сайт проекта - Режим доступа: http://gis.gcras.ru/services.html
12. France operational center, project "Coriolis", Browse Temperature & Salinity gridded maps [Электронный ресурс] : официальный сайт проекта - Режим доступа: http://www.coriolis.eu.org/Data-Services-Products/View-Download/Browse-T-S-maps
13. Красноперов Р.И., Лебедев А.Ю., Пятыгина О.О., Рыбкина А.И., Шибаева А.А. Многодисциплинарная аналитическая ГИС для обработки и представления данных дистанционного зондирования. // Современные проблемы дистанционного зондирования Земли из космоса. 2012. Т. 9. № 3.
14. project ARGO information centre [Электронный ресурс]: официальный сайт проекта - Режим доступа: http://wo.jcommops.org/cgi-bin/WebObjects/Argo
15. Cartification: A Neighborhood Preserving Transformation for Mining High Dimensional Data. E. Aksehirli, B. Goethals, E. Müller, and J. Vreeken. In Data Mining, 2013. ICDM 2013. Thirteenth IEEE International Conference on. IEEE 2013.
16. Игнатов Д.И., Кузнецов С.О. О поиске сходства Интернет-документов с помощью частых замкнутых множеств признаков // Труды 10-й национальной конференции по искусственному интеллекту с международным участием (КИИ’06). – М.:Физматлит, 2006, Т.2, стр.249-258.
17. Игнатов Д.И., Кузнецов С.О. Бикластеризация объектно-признаковых данных на основе решеток замкнутых множеств// Труды 12-й национальной конференции по искусственному интеллекту, М., Физматлит, Т. 1., С.175-182, 2010.
18. Программное обеспечение Orange [Электронный ресурс]: официальный сайт - Режим доступа : http://orange.biolab.si

**Приложения**

**Приложение 1. Карты солености и температуры Мирового океана для разных глубин.**

Карты взять из результатов Французского центра сбора данных(проект CORIOLIS[12]) на 3 мая.

1) Глубина 10 метров:

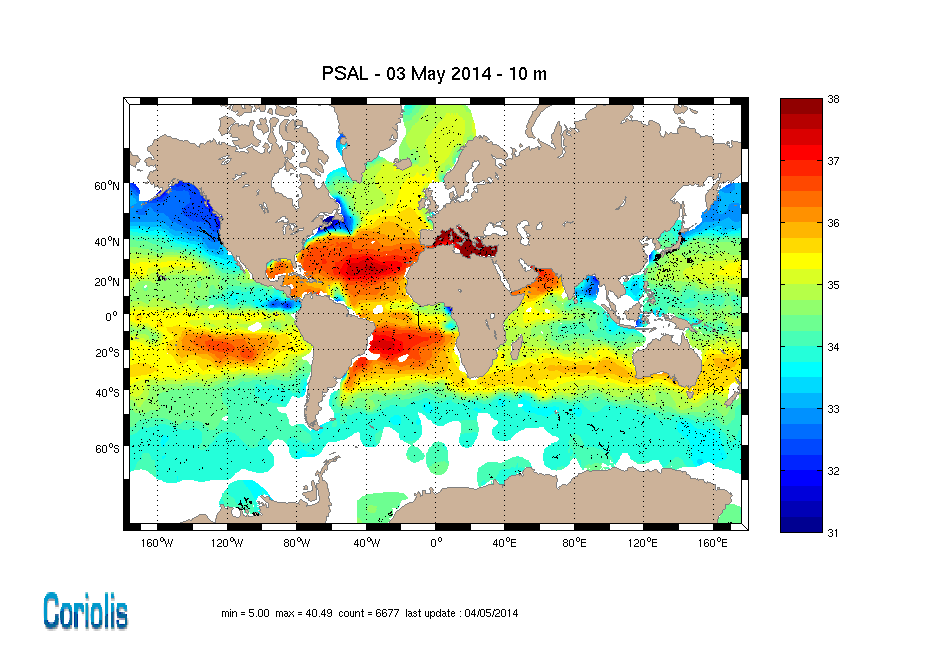


Рисунок 13 Соленость

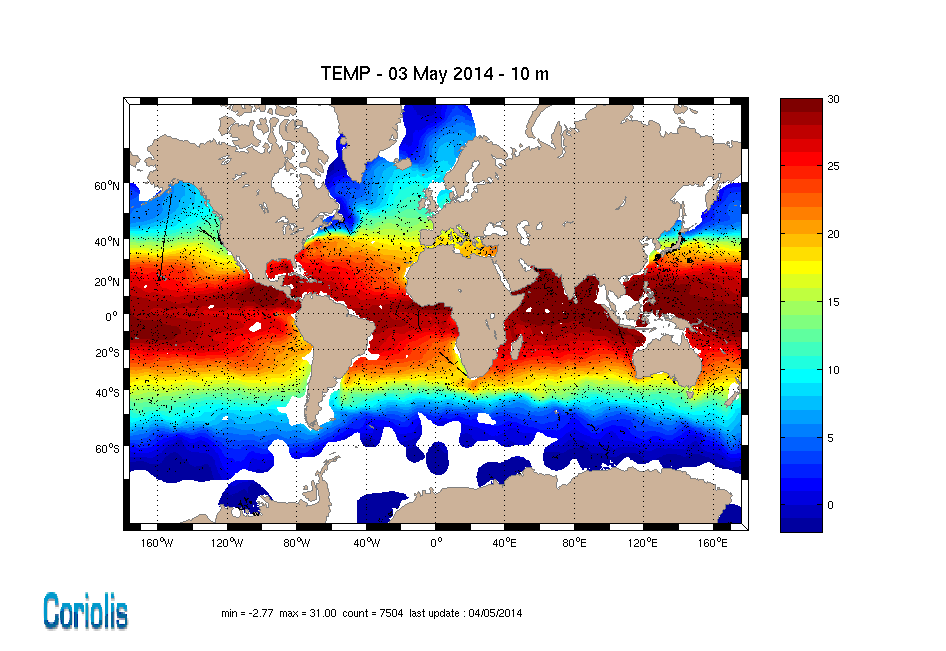


Рисунок 14 Температура

2) Глубина 100 метров:

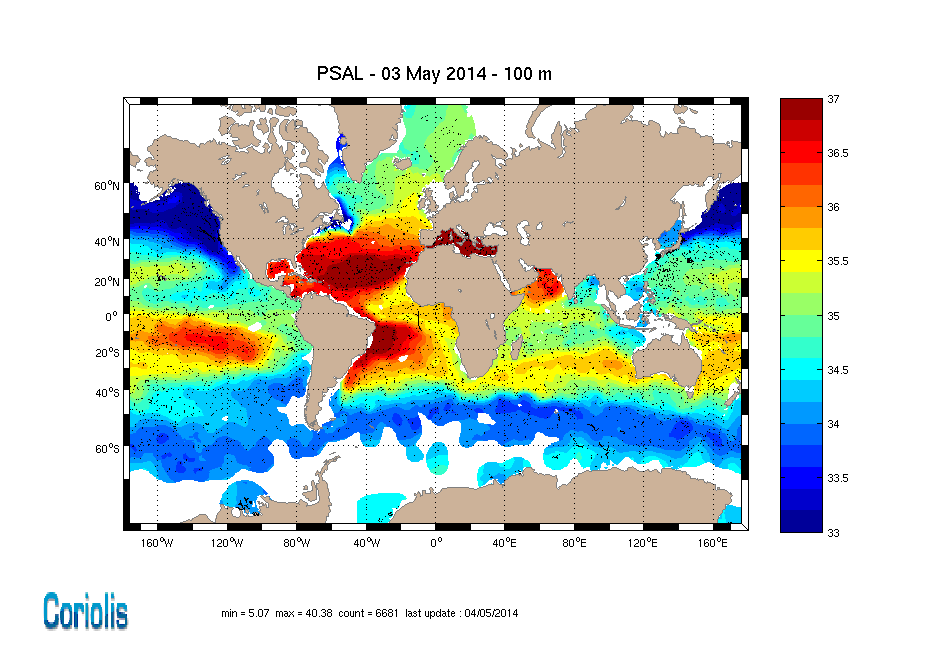


Рисунок 3 Соленость

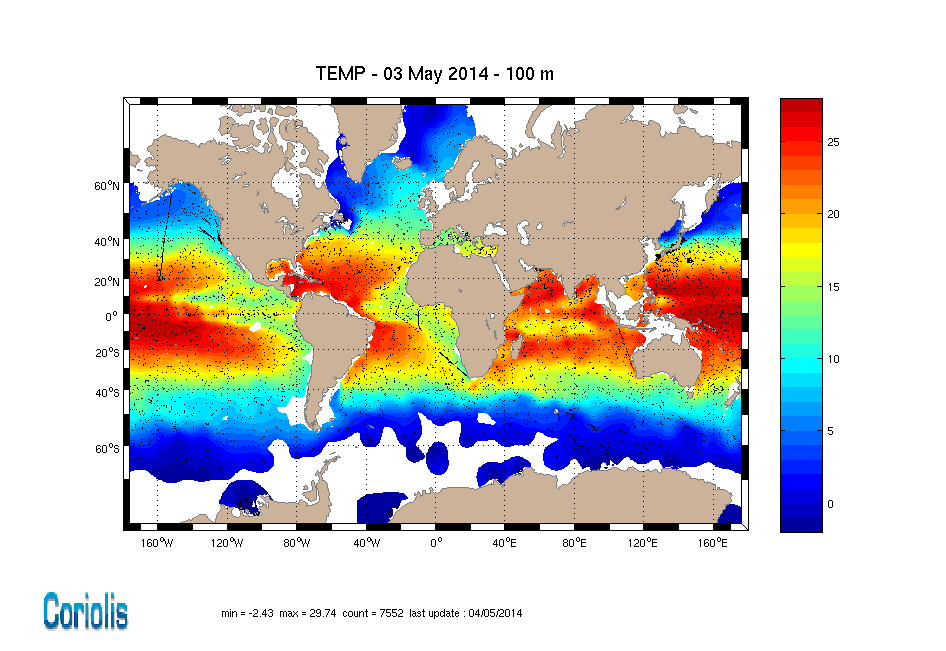


Рисунок 4 Температура

3) Глубина 300 метров:

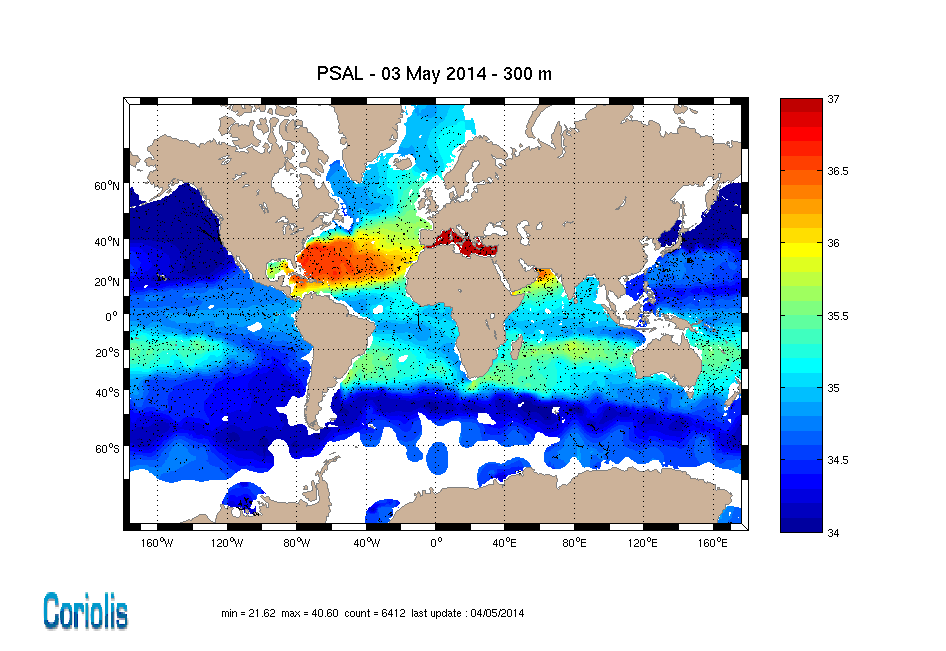


Рисунок 5 Соленость

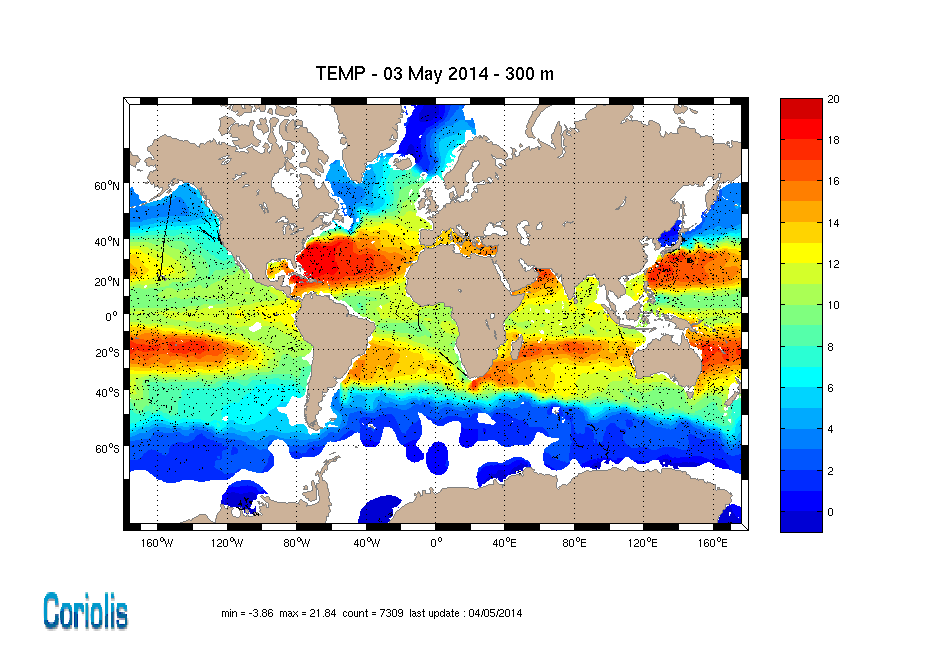


Рисунок 6 Температура

4) Глубина 1000 метров:

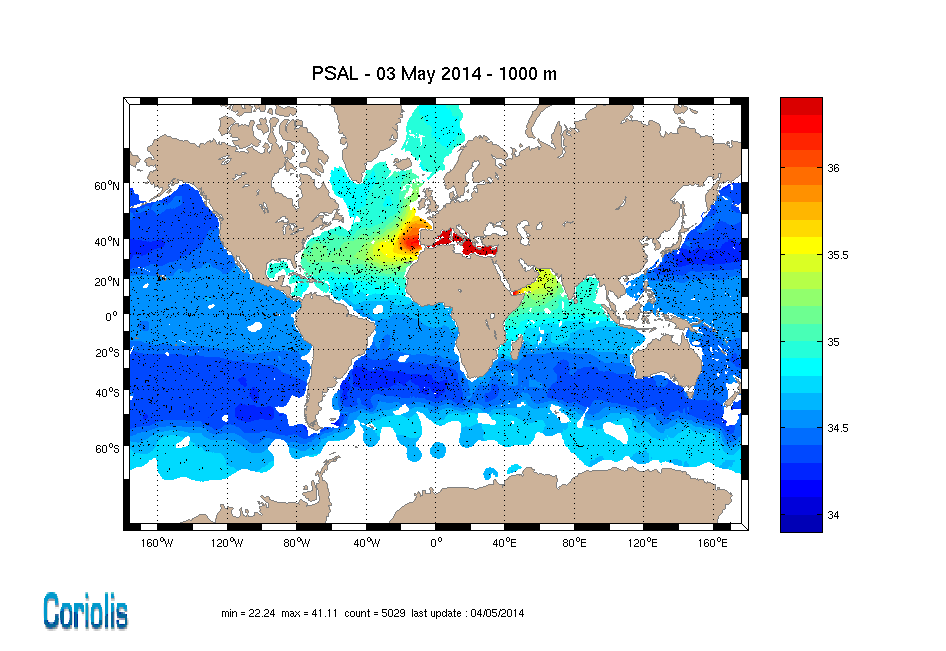


Рисунок 7 Соленость

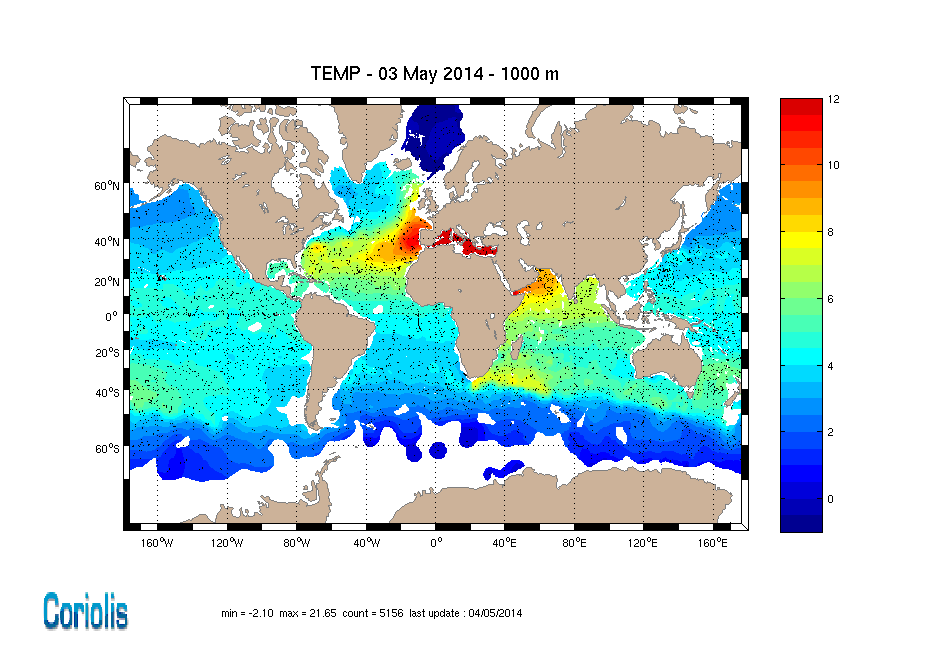


Рисунок 8 Температура

5) Глубина 1600 метров:

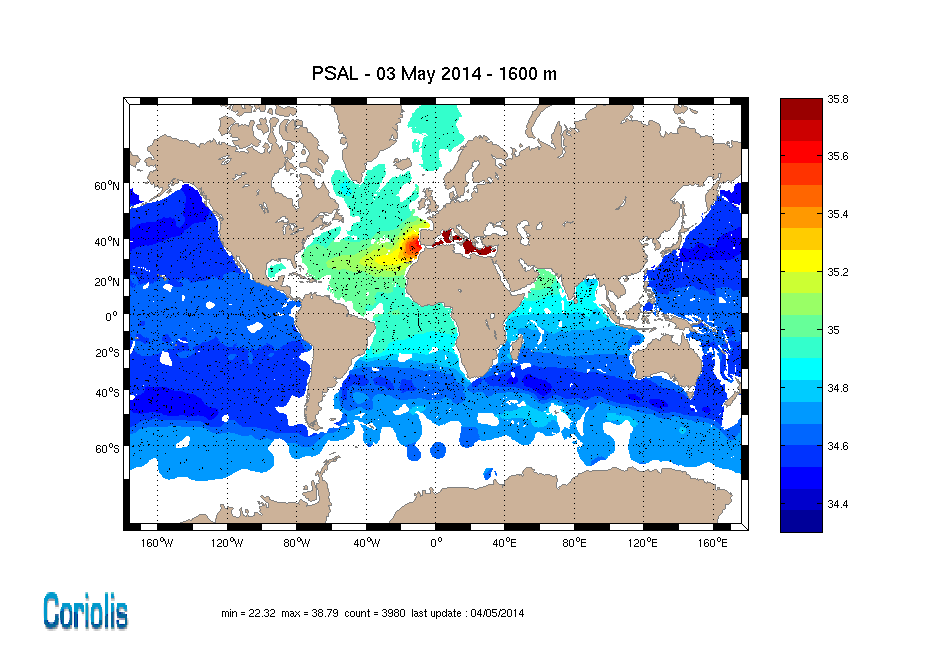


Рисунок 9 Соленость

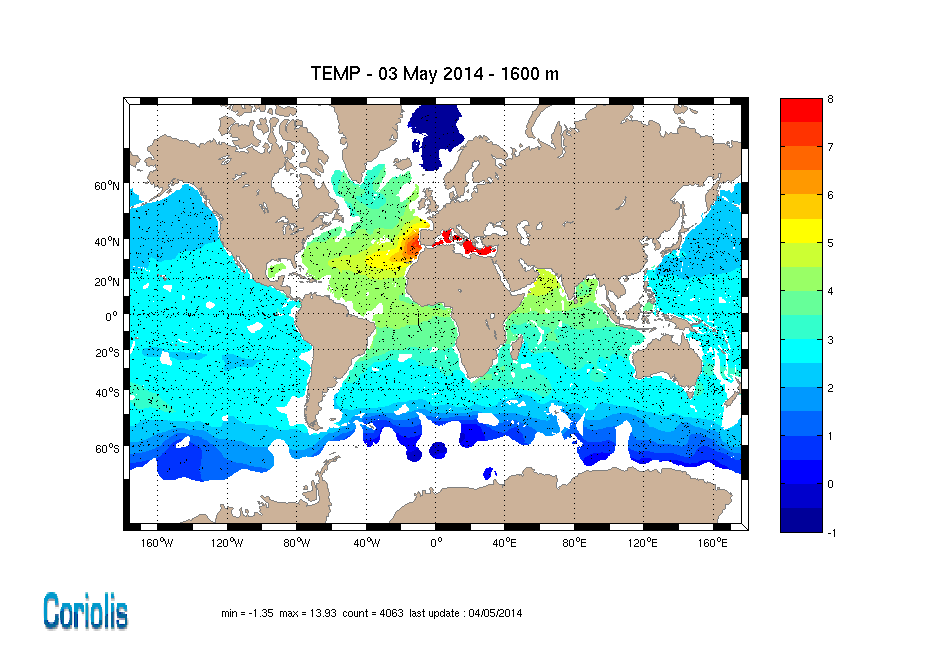


Рисунок 8 Температура

**Приложение 2. Код программы, которая картифицирует базу данных.**

Первый код собирает все объекты находящиеся в определенной папке в одну сруктуру:

%% Считывание всех данных из папки и запись в файл

function data\_all()

clear all

clc

fileList = getAllFiles('D:\data'); % указать путь к папке, где лежат только(!) данные

[num\_files, ~] = size(fileList);

%набор признаков: широта, долгота, соленость, температура, глубина

attributes = {'latitude' 'longitude' 'salinity' 'temperature' 'soundings'};

[~, d] = size(attributes); % размерность пространства признаков

%запись данных из всех файлов в один массив

data = [];

%начинаем со второго, т.к. первый в списке - путь к файлам

for i = 2:num\_files

data\_part = load(fileList{i});

[n1, ~] = size(data\_part);

for j = 1:n1

data = [data; data\_part(j,:)];

end

clear data\_part

end

%запись в файл

save('data.txt', 'data', '-ascii');

end

%% Получение списка всех файлов в папке

function fileList = getAllFiles(dirName)

dirData = dir(dirName);

dirIndex = [dirData.isdir];

fileList = {dirData(~dirIndex).name}';

if ~isempty(fileList)

fileList = cellfun(@(x) fullfile(dirName,x),...

fileList,'UniformOutput',false);

end

subDirs = {dirData(dirIndex).name};

validIndex = ~ismember(subDirs,{'.','..'});

for iDir = find(validIndex)

nextDir = fullfile(dirName,subDirs{iDir});

fileList = [fileList; getAllFiles(nextDir)];

end

end

Второй код получает на вход единую базу данных из первого кода. Этот код создает тележки для всех объектов при всех метриках:

%% Пример вызова функции

function test\_of\_Cartification

clear all

clc

%считываем заранее записанные в один файл данные

data = load('data.txt');

%набор признаков: широта, долгота, соленость, температура, глубина( можно задать как хотите)

attributes = {'latitude' 'longitude' 'salinity' 'temperature' 'soundings'};

[~, d] = size(attributes); % размерность пространства признаков

%пример вызова функции

%Картификация по подмножеству признаков S из 1 и 3 признака

C = Cartification(data, [1 3]);

C{2}{900}

end

%% Алгоритм преобразования исходной базы данных

function C = Cartification(data, Attr)

Metrics = {'euclidean' 'seuclidean' 'mahalanobis' 'cityblock' 'minkowski' 'cosine' 'correlation' 'hamming' 'jaccard'};

X = data(:, Attr);

[n,~] = size(X);

[~, n\_m] = size(Metrics);

k = 200; % top-k Nearest Neighbors, настраиваемый параметр

%если нужно, чтобы использовались все объекты, то поставить k = n;

C = {};

for i = 1:n\_m

clear C\_m Y S Ind

C\_m = {};

Y = pdist(X, Metrics{i}); % получаем матрицу попарных расстояний

S = squareform(Y);% делаем из нее квадратную

[~, Ind] = sort(S,2); % сортируем по возрастанию внутри каждой строки

%Для каждого объекта берем только k ближайших соседей

for j = 1:n

C\_m{j} = Ind(j, 1:k);

end

C{i} = C\_m;

end

end

**Инструкция по использованию кода:**

Сначала запускается первый код в Matlab. Там в 7 строке меняем путь к папке, где лежат все ваши данные, желательно избегать русскоязычных названий, разные версии Matlab по-разному на это реагируют. Тут сканируется папка, получается список всех файлов в ней, после чего из каждого из файлов грузятся данные и записываются в массив data. Данные следующего файла начинают записываться сразу после данных предыдущего файла, никаких разграничений нет. После массив сохраняется в data.txt

Теперь открываем второй код в Matlab.

Здесь сначала грузятся исходные данные, потом задаются другие входные параметры алгоритма преобразования в базу транзакций.

На выходе мы получаем структуру C, где всего 1 строка и столбцов ровно столько, сколько метрик всего(их строенных 9). Каждый элемент структуры С – тоже структура, размер каждой из которых – число объектов.

Что и где хранится:

Например, C{2} – это структура, которая содержит структуру с индексами k-ближайших объектов во 2 мере близости

C{2}{3} – массив индексов объектов, ближайших к 3 объекту во 2 метрике.

C{2}{3}(1) - первый ближайший объект к 3 объекты во 2 метрике.

Обратите внимание на скобки, к элементам структур обращение идет через фигурные, тогд как к обычным массивам через обычные.

**Приложение 3. Код программы реализации метода кластеризации Carticlus.**

%% Пример вызова функции

function test\_of\_Clust\_with\_print\_proc

clear all

clc

close all

%предобработка данных, работа 1 алгоритма.

data = load('19002010.0808.10019.txt');

k = 12; % размер тележек

[n2, ~] = size(data);

%набор признаков: широта, долгота, соленость, температура, глубина

attributes = {'latitude' 'longitude' 'salinity' 'temperature' 'soundings'};

[~, d] = size(attributes); % размерность пространства признаков

%Вызов функции и запись для каждого признака в общую базу

for i = 1: d

C{i} = Cartification(data, i, k);

end

%Задание параметров для 2 алгоритма

minsup = 15;

min\_intersect = 7;%минимальное число пересечений в кластерах

minlen = 9;% минимальное число объектов в кластере

n = 20;% желаемое число кластеров

max\_iter = 1000;

%кластеризация

[Hi\_result, num\_clust] = CartiClus\_Algoritm(C, minsup,minlen,n, n2, max\_iter, min\_intersect);

% Вывод кластеров на экран

disp(['Число кластеров = ', num2str(num\_clust)]);

index = zeros(n2,1);

index(:,1)= num\_clust + 1;

for i = 1:num\_clust

disp(['Кластер №', num2str(i)]);

ind = Hi\_result{i};

index(Hi\_result{i})= i;

disp(ind)

%графики

% %широта-долгота

h = figure();

plot(data(ind,2), data(ind,1),'o');

xlabel(attributes{2});

ylabel(attributes{1});

title(['Кластер №', num2str(i)]);

print(h, '-djpeg', ['D:\Graphs\lat\_log', num2str(i),'.jpeg']);

%соленость-температура

h = figure();

plot(data(ind,3), data(ind,4),'o');

xlabel(attributes{3});

ylabel(attributes{4});

title(['Кластер №', num2str(i)]);

print(h, '-djpeg', ['D:\Graphs\sal\_tem', num2str(i),'.jpeg']);

%трехмерные проекции

% долгота-широта-температура

h = figure();

scatter3(data(ind,2), data(ind,1), data(ind,3),55,'o','LineWidth',2);

xlabel(attributes{2});

ylabel(attributes{1});

zlabel(attributes{3});

title(['Кластер №', num2str(i)]);

print(h, '-djpeg', ['D:\Graphs\lat\_log\_tem', num2str(i),'.jpeg']);

% долгота-широта-соленость

h = figure();

scatter3(data(ind,2), data(ind,1), data(ind,4),55,'o','LineWidth',2);

xlabel(attributes{2});

ylabel(attributes{1});

zlabel(attributes{4});

title(['Кластер №', num2str(i)]);

print(h, '-djpeg', ['D:\Graphs\lat\_log\_sal', num2str(i),'.jpeg']);

% температура-соленость-долгота

h = figure();

scatter3(data(ind,3), data(ind,4), data(ind,2),55,'o','LineWidth',2);

xlabel(attributes{3});

ylabel(attributes{4});

zlabel(attributes{2});

title(['Кластер №', num2str(i)]);

print(h, '-djpeg', ['D:\Graphs\log\_sal\_tem', num2str(i),'.jpeg']);

% температура-соленость-широта

h = figure();

scatter3(data(ind,3), data(ind,4), data(ind,1),55,'o','LineWidth',2);

xlabel(attributes{3});

ylabel(attributes{4});

zlabel(attributes{1});

title(['Кластер №', num2str(i)]);

print(h, '-djpeg', ['D:\Graphs\lat\_sal\_tem', num2str(i),'.jpeg']);

clear ind

end

close all

%Общий график

%

k = find(index == num\_clust+1);

data(k,:)=[];

index(k,:) = [];

% создаем палитру,чтобы у каждого кластера был свой цвет

palitra = hsv(num\_clust);

colors = palitra(index,:);

scatter(data(:,2), data(:,1),55,colors,'o','LineWidth',1);

xlabel('Долгота');

ylabel('Широта');

figure()

scatter(data(:,3), data(:,4),55,colors,'o','LineWidth',1);

xlabel('Соленость');

ylabel('Температура');

end

%% Алгоритм преобразования исходной базы данных

function C = Cartification(data, Attr, k)

% top-k Nearest Neighbors

X = data(:, Attr);

[n,~] = size(X);

C = {};

Y = pdist(X);

S = squareform(Y);% делаем из нее квадратную

[~, Ind] = sort(S,2); % сортируем по возрастанию внутри каждой строки

%Для каждого объекта берем только k ближайших соседей

for j = 1:n

C{j} = Ind(j, 1:k);

end

end

%%Алгоритм 2. Кластеризация

function [Hi\_result, num\_clust] = CartiClus\_Algoritm(C, minsup,minlen,n, num\_obj, max\_iter, min\_intersect)

% возвращаю кластеры и их число

Hi={};

%Заполним матрицу, отвечающую за support 2 объектов

Supp = zeros(num\_obj, num\_obj);

for i = 1:num\_obj

for j = i:num\_obj

Supp(i,j) = support(C, i,j);

Supp(j,i)= Supp(i,j);

end

end

iter = 0;%число итераций

num\_X = 0;%число кластеров в Hi

while (num\_X < n && iter < max\_iter)

clear X P P\_without\_X

iter = iter +1

o = randi(num\_obj, 1); %генерируем случайно целочисленное значение-номер объектра в исходной базе

X = o

P = find(Supp(o,:) > minsup)

[~,size\_P] = size(P);

k = find(P == o);

P\_without\_X = P(:,[1:k-1 k+1:size\_P])

clear k

while size\_P > 0

p = P(randi(size\_P, 1)); %берем случайно объект из P

X = [X p]; %добавляем p к Х

X = unique(X)

%создаем P без X

k = find(P\_without\_X == p);

[~,size\_P\_without\_X] = size(P\_without\_X);

if k

P\_without\_X = P\_without\_X(:,[1:k-1 k+1:size\_P\_without\_X])

size\_P\_without\_X = size\_P\_without\_X - 1;

end

%считаем support для подмножества объектов из P\Х и Х

Supp\_X\_o = zeros(1,size\_P\_without\_X); %вектор support

for i = 1: size\_P\_without\_X

Supp\_X\_o(i) = support\_set(C, [X P\_without\_X(i)]);

end

%ищем среди них превышающие порог

k = find(Supp\_X\_o > minsup);

%если такие есть, то пишем их в P, если нет, то P - пусто

if k

P = P\_without\_X(1, k)

else

P = []

end

[~,size\_P] = size(P);

end

[~,size\_X] = size(X);

if size\_X >= minlen && no\_in\_Hi(X, Hi, num\_X)

num\_X = num\_X + 1;

Hi{num\_X} = X;

X

end

end

%Конец двух вложенных циклов

Hi\_result={};%это Hi'

num\_clust = 0;% конечное число кластеров

%как и писалось X' берется из Hi, а не Hi'

for i = 1:num\_X-1

flag = 0; %отвечает за то, склеился ли кластер с каким-то еще или нет, и попал ли при этом в HI'.

X = Hi{i};

for j = i+1:num\_X

X\_help = Hi{j};

X\_= [];

ins = intersect(X, X\_help);%перечесение X и X'

[~,k] = size(ins);

%если пересечение больше минлен, то склеили, проверили suppб если

%норм, добавили в хи'

if k > min\_intersect

X\_ = [X X\_help];%склеили кластеры

X\_ = unique(X\_);

if no\_in\_Hi(X\_, Hi\_result, num\_clust)

flag = 1;

num\_clust = num\_clust + 1;

Hi\_result{num\_clust} = X\_;

end

end

end

%если кластер ни с чем не склеился,то мы его все равно пишем в Hi\_res

if( flag ==0)

num\_clust = num\_clust + 1;

Hi\_result{num\_clust} = X;

end

end

end

%функция подсчета supp для 2 объектов

function supp = support(C, obj1, obj2)

supp = 0;

[~, n1] = size(C);

[~, n2] = size(C{1});

%в базе тележек ищем есть ли 1 и второй объект в тележке

%если есть и тот, и другой, то увеличивает support на 1

for i = 1:n1

for j = 1:n2

[~,~,v1] = find(C{i}{j} == obj1);

[~,~,v2] = find(C{i}{j} == obj2);

if(v1==1 & v2==1)

supp = supp +1;

end

end

end

end

%функция подсчета supp для множества объектов

function supp\_X = support\_set(C, X)

supp\_X = 0;

[~, num\_obj] = size(X);%число объектов в множестве

[~, n3] = size(C{1}{1});

%если чсило объектов в множестве больше, чем в любой тележке, то support=0

if num\_obj > n3

return

end

[~, n1] = size(C);

[~, n2] = size(C{1});

%в базе тележек ищем есть ли все объекты, если какого-то объекта нет, то

%идем дальше

%если есть все, то supp+=1

for i = 1:n1

for j = 1:n2

flag = 0; %счетчик

for k = 1:num\_obj

clear v

[~,~,v] = find(C{i}{j} == k);

if v~=1

break

else flag = flag+1;

end

end

%если в тележку вошли все объекты множества, то увеличиваем на 1

%supp

if flag == num\_obj

supp\_X = supp\_X +1;

end

end

end

end

%функция проверки есть ли кластер X в Hi

function flag = no\_in\_Hi(X, Hi, num\_X)

[~, n] = size(X);

flag = 1;

for i = 1:num\_X

[~,n\_hi] = size(Hi{i});

if (n == n\_hi)

if X == Hi{i}

flag = 0;

return

end

end

end

end

**Приложение 4. Метод локтя.**

%% Считываем информацию (путь к файлу стоит заменить на свой)

Elbow\_txt = '0101\_m.txt';

Elbow\_matrix = load(Elbow\_txt);

%% Меняем число кластеров, считаем и рисуем критерий

Elbow\_number = size(Elbow\_matrix);

k\_max = 30; %% для определенности так возьмем

J = zeros(k\_max, 1);%% целевая функция/критерий

N = 50;

k = 2:k\_max;

Z = zeros(N,1);

for i = 1:k\_max

for j= 1:N

% [idx, C, d] = kmeans(Elbow\_matrix, i,'emptyaction', 'singleton');

[idx, C, d] = kmeans(Elbow\_matrix(:,3:4), i,'emptyaction', 'singleton');

% [idx, C, d] = kmeans(Elbow\_matrix(:,3:5), i,'emptyaction', 'singleton');

Z(j) = sum(d);

end

J(i)=mean(Z(:,1));

end

plot(k, J, '--rs','LineWidth',2,...

'MarkerEdgeColor','k',...

'MarkerFaceColor','g',...

'MarkerSize',3);

xlabel('k');

ylabel('J(R)');

title('Метод elbow');

xlim([1 k\_max])