Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное образовательное учреждение высшего образования
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Школа – Инженерная школа информационных технологий и робототехники
Направление подготовки – 09.04.01 «Информатика и вычислительная техника»
Отделение школы (НОЦ) – Отделение информационных технологий

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Тема работы
Разработка алгоритма кластеризации климатических данных

УДК 004.421.2: 519.237:551.58

Студент

Группа	**ФИО**	**Подпись**	**Дата**
8ВМ03	Кавешников А.В.		

Руководитель

Должность	**ФИО**	**Учёная степень, звание**	**Подпись**	**Дата**
Доцент	Иванова Ю.А.	к.т.н.		

КОНСУЛЬТАНТЫ:

По разделу «Финансовый менеджмент»

Должность	**ФИО**	**Учёная степень, звание**	**Подпись**	**Дата**
Доцент ОСГН ШБИП	Былкова Т.В.	к.э.н.		

По разделу «Социальная ответственность»

Должность	**ФИО**	**Учёная степень, звание**	**Подпись**	**Дата**
Профессор ОКД ШБИП	Федоренко О.Ю	д.м.н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	**ФИО**	**Ученая степень, звание**	**Подпись**	**Дата**
Профессор	Спицын В.Г.	д.т.н.		

Томск – 2022

ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ

Код	Результаты обучения	Требования ФГОС ВО, СУОС, критерии АИОР, требования профессиональных стандартов
1	**2**	**3**
Общепрофессиональные компетенции		
Р1	Самостоятельно приобретать, и применять математические, естественнонаучные, социально-экономические и профессиональные знания в области современных информационно- коммуникационных технологий для решения междисциплинарных инженерных задач.	Требования ФГОС ВО (3++), СУОС ТПУ (УК-1,4; ОПК-1,4), критерий 5 АИОР (п. 1.1), соответствующий международным стандартам EUR-ACE и FEANI. Запросы студентов, отечественных и зарубежных работодателей.
Р2	Разрабатывать оригинальные алгоритмы и программные средства, в том числе с использованием современных интеллектуальных технологий, для решения профессиональных задач.	Требования ФГОС ВО (3++), СУОС ТПУ (УК-1,2; ОПК-2,5,6), критерий 5 АИОР (п. 1.1, 1.2), соответствующий международным стандартам EUR-ACE и FEANI. Запросы студентов, отечественных и зарубежных работодателей. Требования профессиональных стандартов 06.016, 06.017 (ПК-2, ПК-3).
Р3	Применять на практике новые научные принципы и методы исследований. Демонстрировать способность анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями.	Требования ФГОС ВО (3++), СУОС ТПУ (УК-1,2; ОПК-3,4), критерий 5 АИОР (п. 1.2), соответствующий международным стандартам EUR-ACE и FEANI. Запросы студентов, отечественных и зарубежных работодателей.
Р4	Разрабатывать и модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем.	Требования ФГОС ВО (3++), СУОС ТПУ (УК-1,4; ОПК-5,6,7), критерий 5 АИОР (п. 1.6, п. 2.2,2.6.), соответствующий международным стандартам EUR-ACE и FEANI. Запросы студентов, отечественных и зарубежных работодателей. Требования профессиональных стандартов: 06.028, 06.016, 06.017 (ПК-1, ПК-3).
Р5	Анализировать и оценивать уровни своих компетенций в сочетании со способностью и готовностью к дальнейшему образованию и профессиональной мобильности. Активно владеть одним из иностранных языков на уровне социального и профессионального общения, применять специальную лексику и	Требования ФГОС ВО (3++), СУОС ТПУ (УК-5,6; ОПК-7,8), критерий 5 АИОР (п. 2.1, п. 2.3, п. 1.5), соответствующий международным стандартам EUR-ACE и FEANI. Запросы студентов, отечественных и зарубежных работодателей.

1	2	3
	профессиональную терминологию языка.	
Р6	Осуществлять эффективное управление разработкой программных средств и проектов. Эффективно работать, как член и руководитель группы, демонстрировать ответственность за результаты работы и готовность следовать корпоративной культуре.	Требования ФГОС ВО (3++), СУОС ТПУ (УК-3,4,5; ОПК-3,8), критерий 5 АИОР (п. 2.4, п. 2.5), соответствующий международным стандартам EUR-ACE и FEANI. Запросы студентов, отечественных и зарубежных работодателей Требования профессиональных стандартов: 06.028, 06.016, 06.017 (ПК-2, ПК-3).
Профессиональные компетенции		
Р7	Разрабатывать стратегии проектирования, критерии эффективности и ограничения применимости сверточных нейронных сетей и методов вычислительного интеллекта для разработки программно-алгоритмических систем анализа больших объёмов данных.	Требования ФГОС ВО (3++) (3++), СУОС ТПУ (УК-3; ОПК-5,6), критерий 5 АИОР (п.1.3), соответствующий международным стандартам EUR-ACE и FEANI. Запросы студентов, отечественных и зарубежных работодателей. Требования профессиональных стандартов 06.001, 06.015, 40.057, 06.003, 06.017, 06.035.
Р8	Планировать и проводить теоретические исследования и компьютерные эксперименты в области создания программных систем интеллектуального анализа больших объёмов данных.	Требования ФГОС ВО (3++) (3++), СУОС ТПУ (УК-2,5; ОПК-6,8), критерий 5 АИОР (п. 1.5), соответствующий международным стандартам EUR-ACE и FEANI. Запросы студентов, отечественных и зарубежных работодателей. Требования профессиональных стандартов 06.001, 06.015, 40.057, 06.003, 06.017, 06.035.

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное образовательное учреждение высшего образования
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Школа – Инженерная школа информационных технологий и робототехники
Направление подготовки – 09.04.01 «Информатика и вычислительная техника»
Отделение школы (НОЦ) – Отделение информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП

_____ _____ _____
(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

Магистерской диссертации

(бакалаврской работы, дипломного проекта/работы, магистерской диссертации)

Студенту:

Группа	ФИО
8ВМ03	Кавешникову Артему Владимировичу

Тема работы:

Утверждена приказом директора (дата, номер)	

Срок сдачи студентом выполненной работы:	

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе *(наименование объекта исследования или проектирования; производительност ь или нагрузка; режим работы (непрерывный, периодический, циклический и т. д .) ; вид сырья или материал изделия; т ребования к продукту, изделию или процессу; особые требования к особенностям функционирования (эксплуатации) объекта или из делия в плане безопасности эксплуатации, влияния на окружающую среду, энергозатратам; экономический анализ и т. д.).*	Предметом исследования являются алгоритмы и методы, используемые в задачах кластеризации климатических данных

Перечень подлежащих исследованию, проектированию и разработке вопросов *(аналитический обзор по литературным источникам с целью выяснения достижений мировой науки техники в рассматриваемой области; постановка задачи исследования, проектирования, конструирования;*	– Обзор литературных источников. – Поиск и формирование выборок данных. – Сравнительный анализ результатов работы применяемых алгоритмов кластеризации – Социальная ответственность.
содержание процедуры исследования, проектирования, конструирования; обсуждение результатов выполненной работы; наименование дополнительных разделов, подлежащих разработке; заключениепо работе).	– Финансовый менеджмент, ресурсоэффективность и ресурсосбережение. – Заключение.
Перечень графического материала *(с точным указанием обязательных чертежей)*	

Консультанты по разделам выпускной квалификационной работы

с указанием разделов)

Раздел	Консультант
Основная часть	Доцент ОИТ ИШИТР, к .т.н., Иванова Ю.А.
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Доцент ОСГН ШБИП к.э.н., Былкова Т.В.
Социальная ответственность	Профессор ОКД ШБИП д.м.н., Федоренко О.Ю
Английский язык	Старший преподаватель Розанова Я.В.

Названия разделов, которые должны быть написаны на русском и иностранном языках:

Аналитический обзор
Материалы и методы
Результаты экспериментов
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	

Социальная ответственность

Задание выдал руководитель / консультант (при наличии):

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Иванова Ю.А.	к.т.н.		01.03.22

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ВМ03	Кавешников Артем Владимирович		01.03.22

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	ФИО
8ВМ03	Кавешников Артем Владимирович

Школа	ИШИТР	**Отделение школы (НОЦ)**	ОИТ
Уровень образования	магистратура	**Направление/специальность**	09.04.01 «Информатика и вычислительная техника»

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:	
1. Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих	Стоимость ресурсов определялась по средней рыночной стоимости, и в соответствии с окладами сотрудников организации.
2. Нормы и нормативы расходования ресурсов	Районный коэффициент 30%; Коэффициент дополнительной заработной платы 12%; Накладные расходы 16%.
3. Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования	Коэффициент отчислений на уплату во внебюджетные фонды 30%
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Оценка коммерческого и инновационного потенциала НТИ	Провести предпроектный анализ
2. Разработка устава научно-технического проекта	Представить Устав научного проекта магистерской работы
3. Планирование процесса управления НТИ: структура и график поведения, бюджет, риски и организация закупок	Разработать план управления НТИ
4. Определение ресурсной, финансовой, экономической эффективности	Рассчитать сравнительную эффективность исследования
Перечень графического материала *(с точным указанием обязательных чертежей)***:**	
1. «Портрет» потребителя результатов НТИ *2. Сегментирование рынка* *3. Оценочная карта конкурентных технических решений* *4. Диаграмма FAST* *5. Матрица SWOT* *6. График проведения и бюджет НТИ* *7. Оценка ресурсной, финансовой и экономической эффективности НТИ* *8. Потенциальные риски*	

Дата выдачи задания для раздела по линейному графику	01.03.2022

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОСГН ШБИП	Былкова Т.В.	канд.экон.наук		01.03.2022

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ВМ03	Кавешников Артем Владимирович		01.03.2022

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	**ФИО**
8ВМ03	Кавешников Артем Владимирович

Школа	Инженерная школа информационных технологий и робототехники	**Отделение (НОЦ)**	Информационных технологий
Уровень образования	Магистратура	**Направление/специальность**	09.04.01 Информатика и вычислительная техника

Тема ВКР:

Разработка алгоритма кластеризации климатических данных	
Исходные данные к разделу «Социальная ответственность»:	
1. Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика, рабочая зона) и области его применения	Целью работы является: разработка нового алгоритма кластеризации временных рядов климатических характеристик, обеспечивающий выделение уникальных климатических классов в разных пространственно-временных масштабах. Основная работа с системой и работа по её созданию производится с использованием персонального компьютера в жилом помещении. Характеристика помещения, где проводились работы по ВКР: ширина комнаты составляет b= 4.5 м, длина a=6 м, высота H = 2,8 м. Площадь помещения будет составлять $S=ab$=27 м^2, объем $V=abh$=81.4 м^3; присутствует окно, через которое может производится вентиляция помещения, принудительная вентиляция отсутствует; в зимнее время помещение отапливается; в помещении используется комбинированное освещение.
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности при разработке проектного решения): – специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства; – организационные мероприятия при компоновке рабочей зоны.	Работа над системой регулируется следующими нормативно-правовыми актами: Трудовой Кодекс Российской Федерации от 30.12.2001 N 197-ФЗ (ред. от 25.02.2022); СанПиН 1.2.3685-21. Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания ГОСТ 12.2.032-78. Межгосударственный стандарт. Система стандартов безопасности труда. Рабочее место при выполнении работ сидя. Общие эргономические требования. ГОСТ 12.0.003-2015. Межгосударственный стандарт. Система стандартов безопасности труда. Опасные и вредные производственные факторы. Классификация. СП 52.13330.2016. Свод правил. Естественное и искусственное освещение. Актуализированная редакция СНиП 23-05-95*.

	ГОСТ 22269-76. Государственный стандарт Союза ССР. Система "человек-машина". Рабочее место оператора. Взаимное расположение элементов рабочего места. Общие эргономические требования" ГОСТ Р 50923-96. Государственный стандарт Российской Федерации. Дисплеи. Рабочее место оператора. Общие эргономические требования и требования к производственной среде. Методы измерения ГОСТ 12.1.030-81. Государственный стандарт Союза ССР. Система стандартов безопасности труда. Электробезопасность. Защитное заземление. Зануление ГОСТ 12.1.038-82. Система стандартов безопасности труда. Электробезопасность. Предельно допустимые значения напряжений прикосновения и токов ГОСТ 17.4.3.04-85. Государственный стандарт Союза ССР. Охрана природы. Почвы. Общие требования к контролю и охране от загрязнения ГОСТ Р 53692-2009. Национальный стандарт Российской Федерации. Ресурсосбережение. Обращение с отходами. Этапы технологического цикла отходов ГОСТ 12.1.004-91. Межгосударственный стандарт. Система стандартов безопасности труда. Пожарная безопасность. Общие требования
2. Производственная безопасность при разработке проектного решения: 2.1. Анализ выявленных вредных и опасных факторов 2.2. Обоснование мероприятий по снижению воздействия	**Вредные факторы:** 1. Электромагнитные поля; 2. Электростатические поля; 3. Шум и вибрации; 4. Отклонения показателей микроклимата от нормы в помещении; 5. Недостаточная освещенность рабочей зоны; 6. Психофизиологические факторы (монотонность труда, нервно-психические перегрузки, перенапряжение зрительных анализаторов). **Опасные факторы:** 1. Поражение электрическим током; 2. Короткое замыкание; ~~3.~~ Статическое электричество. **Требуемые средства коллективной и индивидуальной защиты от выявленных факторов:** системы вентиляции, источники света, защитные покрытия, защитные заземления, виброизолирующие средства, изолирующие устройства и покрытия. **Расчет:** расчет системы искусственного освещения
3. Экологическая безопасность при разработке проектного решения:	**Воздействие на селитебную зону:** отсутствует **Воздействие на литосферу:** засорение территории при утилизации компьютера и периферийных устройств (принтеры, МФУ, веб-камеры, наушники, колонки, телефоны), акумуляторных батарей, люминесцентных ламп, макулатуры. **Воздействие на гидросферу:** отсутствует **Воздействие на атмосферу:** отсутствует
4. Безопасность в чрезвычайных ситуациях при разработке проектного решения:	Возможными чрезвычайными ситуациями при разработке устройства являются пожары, грозы, ураганы, оползни. Вероятные ЧС, инициируемые объектом исследования: пожары.

Дата выдачи задания для раздела по линейному графику	01.03.2022

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Профессор ОКД ШБИП	Федоренко Ольга Юрьевна	Доктор медицинских наук		01.03.2022

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ВМ03	Кавешников Артем Владимирович		01.03.2022

РЕФЕРАТ

Выпускная квалификационная работа содержит пояснительную записку на 84 листах, включает 28 рисунков, 19 таблиц, 24 источник литературы, 1 приложение.

Объектом исследования являются: алгоритмы кластеризации климатических данных

Цель работы: разработать новый алгоритм кластеризации временных рядов климатических характеристик, обеспечивающий выделение уникальных климатических классов в разных пространственно-временных масштабах.

В данной работе рассмотрены наиболее простые и фундаментальные алгоритмы кластеризации и сферы их применения. Проанализированы алгоритмы кластеризации выявлены их достоинства и недостатки.

Предложена и обоснована метрика для кластеризации температурных временных рядов

Разработан нейросетевой алгоритм кластеризации климатических данных.

Ключевые слова: кластеризация, методы группировки, разработка метода, алгоритма , мат модель.

Оглавление

ВВЕДЕНИЕ

Одной из самых существенных и масштабных проблем современности на текущий момент можно назвать непрерывно растущий объем информации, который требует определенной систематизации, упрощения и вычленения ее существенной части. С развитием технических средств и интернет-технологий объем цифровых данных растет в огромных масштабах и исчисляется терабайтами. Осуществлять обработку таких данных вручную трудоемко, а существующие методы могут оказаться неэффективными. Поэтому для решения задач такого рода требуются все более и более новые методы обработки данных. Современные методы должны с достаточно высокой точностью осуществлять анализ, систематизацию и сбор полученной информации.

Методы, позволяющие анализировать большие объемы данных, имеют широкий спектр применения. Так, в медицине по совокупности кластерных симптомов можно с достаточно высокой точностью установить диагноз и назначить последующее лечение; в экономике набор параметров кластера может использоваться для выделения групп потребителей, их поведения и их потребительской корзины; в метеорологии кластерный анализ позволяет выделять климатические зоны и прогнозировать их изменение. С помощью алгоритмов кластеризации можно реализовать задачу распознавания образов, а также существует достаточно высокая потребность в обработке больших объемов данных в научных исследованиях. На основании вышеизложенного можно сделать вывод, что востребованность алгоритмов кластеризации и их исследования достаточно высока.

1. АНАЛИТИЧЕСКИЙ ОБЗОР

1.1 Задача кластеризации

Задачу кластеризации (обучения без учителя) можно поставить так. Имеется обучающий набор объектов - выборку $X\ell = \{x1, \ldots, x\ell\} \subset X$ и функцию расстояния между ними $\rho(x, x')$. Задача кластеризации заключается в осуществлении процесса разбиения выборки на непересекающиеся подмножества, таким образом, чтобы каждое подмножество состояло только из объектов сходных по выбранной метрике, а объекты разных подмножеств имели существенные различия. Данные подмножества именуются кластерами. В результате кластеризации каждому объекту будет присвоена метка соответствующего кластера.

В целом алгоритм кластеризации можно охарактеризовать как функцию, которая приводит все объекты выборки X в соответствие с некоторой меткой кластера Y. $X \rightarrow Y$

Цели кластеризации:

- разбиение множества объектов по некоторому признаку и упрощение дальнейшей обработки данных
- выявление структуры объектов
- сокращение объема данных;
- выделение нетипичных объектов;

1.2 Графовые методы кластеризации

Существует широкий ряд методов кластеризации, на графах. Вершины графа — объекты выборки. Ребра — пары объектов с расстоянием:

$\rho_{i,j} = \rho(x_i, x_j)$. Графовые алгоритмы являются относительно простыми в реализации, наглядными и достаточно простыми для модификации.

Алгоритм выделения связных компонент. В данном алгоритме задается входной параметр R и в графе удаляются все ребра (i, j), для которых расстояние $\rho_{i,j} > R$. Соединенными остаются только наиболее близкие пары объектов. Основная концепция алгоритма заключается в подборе, значения R

при котором граф разделяется на несколько связных компонент. Найденные связные компоненты — и есть искомые кластеры.

Компонент связности графа — это подмножество его вершин, в котором любые две вершины могут быть соединены путем, целиком лежащим в этом подмножестве. Для подбора оптимального значения параметра R как правило строят гистограмму распределения попарных расстояний $\rho(i,j)$ [1].

Алгоритм кратчайшего незамкнутого пути

Суть алгоритма в следующем: строится граф из *n-1* ребер, так, чтобы все n точки соединялись и длинна между ними была сведена к минимуму. Граф такого типа называется кратчайшим незамкнутым путем или каркасом.

Алгоритм состоит в поиске пары точек с наименьшим расстоянием и соединения их ребром.

После этого выполняется проверка на наличие изолированных узлов. Пока условие выполняется:

осуществляется поиск изолированной точки, наиболее близкой к некой неизолированной.

Далее эти две точки соединяются.

После этого p-1 самых длинных ребер удаляются.

1.3 Алгоритм FOREL

(ФОРмальный ЭЛемент). Пусть задана некоторая точка $x0 \in X$ и параметр R. Выделяются все точки выборки $xi \in X\ell$, попадающие внутрь сферы $\rho(xi , x0)$, и точка x0 переносится в центр тяжести выделенных точек. Эта процедура повторяется до тех пор, пока набор выделенных точек, а значит и местоположение центра, не перестанет изменяться. Доказано, что эта процедура сходится за конечное количество итераций. При этом сфера перемещается в место локального сгущения точек. Центр сферы x0 в общем случае не принадлежит к выборке, именно поэтому он и называется формальным элементом. Для вычисления центра необходимо, чтобы множество объектов X было не только метрическим, но и линейным

векторным пространством. Это требование естественным образом выполняется, когда объекты описываются числовыми признаками. [1]

1.4 Статистические алгоритмы

Статистические алгоритмы основаны на допущении, что кластеры можно описать с помощью семейства вероятностных распределений. Тогда задача кластеризации сводится к разделению смеси распределений по конечной выборке.

Метод k-средних является упрощением EM-алгоритма. Основное отличие в том, что в EM-алгоритме каждый объект x_i распределяется по всем кластерам с вероятностями $g_i(y) = P\{y_i = y\}$. В алгоритме k-средних каждый объект жестко приписывается только к одному кластеру. Второе отличие заключается в том, что форма кластеров в методе k-средних не настраиваемая. [1].

1.5 Иерархическая кластеризация

Иерархические алгоритмы кластеризации, называемые также алгоритмами таксономии, строят не одно разбиение выборки на непересекающиеся классы, а систему вложенных разбиений. Результат таксономии обычно представляется в виде таксономического дерева — дендрограммы. Среди алгоритмов иерархической кластеризации выделяют два основных типа:

Дивизимные или нисходящие алгоритмы разбивают выборку от изначально крупных кластеров на все более и более мелкие кластеры.

Более распространены агломеративные или восходящие алгоритмы, в которых объекты объединяются из изначально мелких кластеров во все более и более крупные кластеры.

Изначально каждый объект считается отдельным кластером. Для одноэлементных кластеров естественным образом определяется функция расстояния

$$R(\{x\}, \{x'\}) = \rho(x, x').$$

Затем начинается процесс слияния кластеров. На каждой итерации

вместо пары ближайших кластеров U и V образуется новый кластер $W = U \cup V$. Расстояние от нового кластера W до любого другого кластера S вычисляется по расстояниям *R(U, V), R(U, S) и R(V, S),* которые к этому моменту уже должны быть известны:

$$R(U \cup V, S) = \alpha U R(U, S) + \alpha V R(V, S) + \beta R(U, V) + \gamma |R(U, S) - R(V, S)|,$$

где αU , αU , β, γ — числовые параметры. [1]

На практике используются следующие способы вычисления расстояний *R(W, S)* между кластерами W и S. Для каждого из них доказано соответствие формуле Ланса-Вильямса при определенных сочетаниях параметров. [1]

Расстояние ближнего соседа:

$$R\text{б}(W, S) = \min_{w \in W, s \in S} \rho(w, s);$$
$$\alpha U = \alpha V = \tfrac{1}{2}\ , \beta = 0, \gamma = -\tfrac{1}{2};$$

Расстояние дальнего соседа:

$$R\text{д}(W, S) = \max_{w \in W, s \in S} \rho(w, s);$$
$$\alpha U = \alpha V = \tfrac{1}{2}\ , \beta = 0, \gamma = \tfrac{1}{2};$$

Среднее расстояние:

$$Rc(W, S) = \frac{1}{|W||S|} \sum_{w \in W} \sum_{s \in S} \rho(w, s);$$
$$\alpha U = |U|/|W| , \alpha V = |V|/|W| , \beta = \gamma = 0;$$

Расстояние между центрами:

$$R\text{ц}(W, S) = \rho^2 \left(\sum_{w \in W} \frac{w}{|W|} , \sum_{s \in S} \frac{s}{|S|} \right);$$
$$\alpha U = |U|/|W|, \alpha V = |V|/|W| , \beta = -\alpha U \alpha V , \gamma = 0;$$

Расстояние Уорда:

$$Ry(W, S) = \frac{|S||W|}{|S|+|W|} \rho^2 \left(\sum_{w \in W} \frac{w}{|W|} , \sum_{s \in S} \frac{s}{|S|} \right);$$
$$\alpha U = \frac{(|S|+|U|)}{(|S|+|W|)} , \alpha V = \frac{(|S|+|V|)}{(|S|+|W|)} , \beta = -\frac{|S|}{(|S|+|W|)} , \gamma = 0.$$

1.6 Обучение без учителя

Обучение без учителя — семейство алгоритмов обучения, в которой для подстройки весов не требуется целевая функция. Используется, когда известны только описания множества объектов и требуется обнаружить закономерности существующие между объектами.

В алгоритмах обучения без учителя не вычисляется ошибка обучающей выборки и не используется метод обратного распространения ошибки. Вместо этого используется данные о существующем состоянии системы и примеров обучающего множества.

Основное назначение нейронных сетей без учителя — осуществление

задач кластеризации.

1.7 Сеть Кохонена

Одной из наиболее популярных архитектур нейронных сетей, использующих обучение без учителя является нейронная сеть Кохонена (рисунок 1), применяемая для кластеризации, в основе построения которой лежит соревновательное обучение, в котором коррекция весов нейронов происходит через вычисления расстояния между векторами выходного слоя и векторами входных признаков. [2].

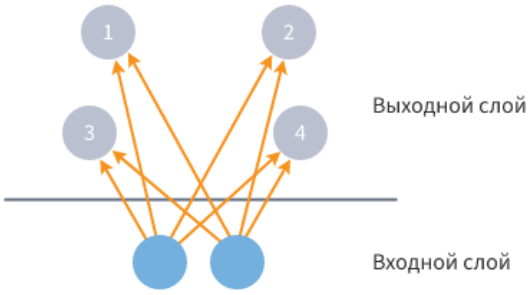

Рисунок 1. Архитектура нейронной сети Кохонена

Количество выходных нейронов сети Кохонена равно числу кластеров, которое должно быть построено сетью. Выходы нейроны избираются по принципу «победитель забирает все», т.е. нейрон с наибольшим значением выхода выдает единицу, а выходы остальных обнуляются.

Обучение сети Кохонена, как и обычной нейронной сети, заключается в подстройке весов связей между нейронами, но производится с использованием технологии соревновательного обучения. [3].

1.8 Соревновательное обучение

На этапе конкуренции на вход сети подается входной вектор признаков и производится поиск нейрона с наиболее близким к нему набором весов. Такой нейрон объявляется победителем. В процессе объединения вокруг него образуется группа (соседство) нейронов, которые будут участвовать в процессе обучения (рисунок 2). Размер группы определяется радиусом обучения. [5].

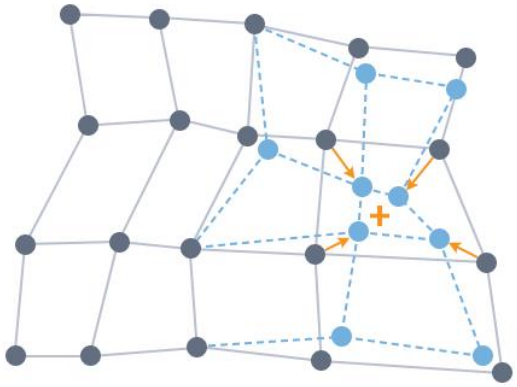

Рисунок 2. Процесс изменения структуры карты Кохонена

Наконец, на этапе подстройки происходит корректировка весов нейронов, расположенных в пределах радиуса обучения от нейрона-победителя, таким образом, чтобы их векторы стали ближе к нему. [5].

Подводя итог обзора алгоритмов кластеризации приведем сравнительный анализ алгоритмов в виде таблицы, представленной ниже.

Таблица 1. Сравнительная таблица алгоритмов кластеризации

	Итерации	Расчет полной матрицы расстояний	Хранение полной матрицы расстояний	Задание кол-ва классов	Зависимость от начальных условий	Обучение	Форма класса
Графовые методы	-	+	+	-	-	-	любая
Алгоритм Forel «Центр тяжести»	+	-	-	+	+	-	гиперсфера
K-средних	+	+	+	+	+	-	гиперсфера
Иерархические гломеративные дивизимные	-	+	+	-	+	-	любая
Нейронные сети (с учителем)	+	+	+	+	+	+	любая
Нейронные сети (без учителя)	+	+	+	-	+	+	любая

2. Алгоритм кластеризации климатических данных

2.1 Описание алгоритма

Алгоритм разработан для анализа климатических данных и использовался для работы с показателями температуры на примере данных , характеризующих изменения среднемесячной температуры.

Алгоритм использует нейросетевую архитектуру. Так как мы работаем с задачей кластеризации, то мы имеем дело с обучением без учителя. Одним из подобных инструментов является сеть Кохонена. Она представляет собой двухслойную сеть, где каждый нейрон входного слоя соединен с каждым нейроном выходного слоя. Нейроны второго – выходного слоя часто именуют кластерными элементами. Количество этих нейронов определяет, максимально возможное количество групп, на которое будут разделены данные.

Заданная нами система функционирует по принципу конкурентного обучения. Нейроны выходного слоя соревнуются между собой за право наилучшим образом сочетаться с входным вектором признаков. Победа оказывается за тем нейроном, вектор весов которого ближе всего к вектору входных признаков. Таким образом процедура кластеризации заключается в отнесении каждого входного вектора признаков к некоторому кластеру.

Обучение нейронной сети осуществляется с помощью конкурентного обучения. На каждой итерации алгоритма из входного слоя нейронной сети случайным образом берется один вектор. После этого происходит поиск нейрона выходного слоя, расстояние которого между его набором весов и набором весов входного вектора минимально. Корректировка весов для нейрона – победителя происходит по некоторому выбранному правилу. Обучение происходит с определенной заданной скоростью, задаваемой параметром $\Delta\lambda$. В процессе обучения осуществляется поиск ближайших векторов к вектору «победителю», далее происходит подстройка весов реализуемая последовательно по формуле:

$$wm_i = wm_i + la \times (x_i - wm_i),$$

где wm_i - текущий вектор выходного слоя, x_i – вектор входных данных, la- коэффициент обучения.

В качестве признаков кластеризации использованы температура и местоположение станции. В основе кластеризации лежит принцип синхронного поведения температурных временных рядов. В качестве критериев отнесения данных температуры к кластеру используется количество кластеров.

Процедура кластеризации заключается в вычислении оптимального вектора выходного слоя наиболее близко подходящего к входному вектору признаков. Процедура представляет собой итеративный процесс.

В полученные кластеры включаются климатические станции, для которых уровень расстояние между входным вектором и вектором нейрона – победителя минимально.

2.2 Разработка способа кластеризации с внедрением метрики среднегодовых температур

В ходе исследования и работы с текущим алгоритмом была поставлена задача повысить качество проводимой кластеризации. Исходный данных представлен набором среднемесячных температур. Была выдвинута гипотеза о том, что год – является оптимальным периодом, характеризующим климатическую группу. Среднегодовая температура была внедрена в качестве метрики кластеризации.

$d_{ab}=\sqrt{(t_a-t_b)_2}$, где t – среднегодовая температура.

3. МАТЕРИАЛЫ И МЕТОДЫ

Использованные технологии

3.1 NumPy

NumPy – это открытая библиотека для языка Python с открытым исходным кодом, реализующая базовые математические операции, используемая для работы с массивами и матрицами. [6].

3.2 Pandas

Pandas – это открытая библиотека для языка Python, используемая для анализа и обработки данных. Библиотека предоставляет удобные инструменты для открытия и работы с таблицами данных и временными рядами [7].

3.3 Matplotlib

Matplotlib – это открытая библиотека для языка Python, используемая для визуализации данных в виде двух и трехмерных графиков. Полученная графическая информация часто используется аналитиками в представлениях данных и находит свое использование в научных публикациях. [8].

Используемые алгоритмы кластеризации.

В нашей работе было использовано и реализовано два алгоритма: k-средних и нейросетевой алгоритм на основе принципа самоорганизующейся карты Кохонена.

3.4 K-Средних

Метод k – средних был выбран нами для реализации как наиболее известный и простой алгоритм осуществляющий качественное решение задачи кластеризации, тем не менее обладающий существенным недостатком, а именно необходимостью заранее знать количество кластеров k. Выбор пал не нейросетевые методы, ввиду их обширности, гибкости настройки, возможностей модернизации. Так как мы решаем задачу кластеризации нас интересуют нейросети с обучением без учителя. Одной из таких сетей является сеть Кохонена. [9].

3.5 Нейронная сеть Кохонена

Нейронная сеть Кохонена – алгоритм машинного обучения без учителя, позволяющий отобразить результаты в виде компактных и удобных для интерпретации двумерных карт.

Используемые наборы данных.

Входные данные представлены массивом чисел среднемесячных температур размерностью (744 x 928) (кол-во наблюдений за 62 года x 928 климатических станций). Таким образом входные данные образуют собой временные ряды

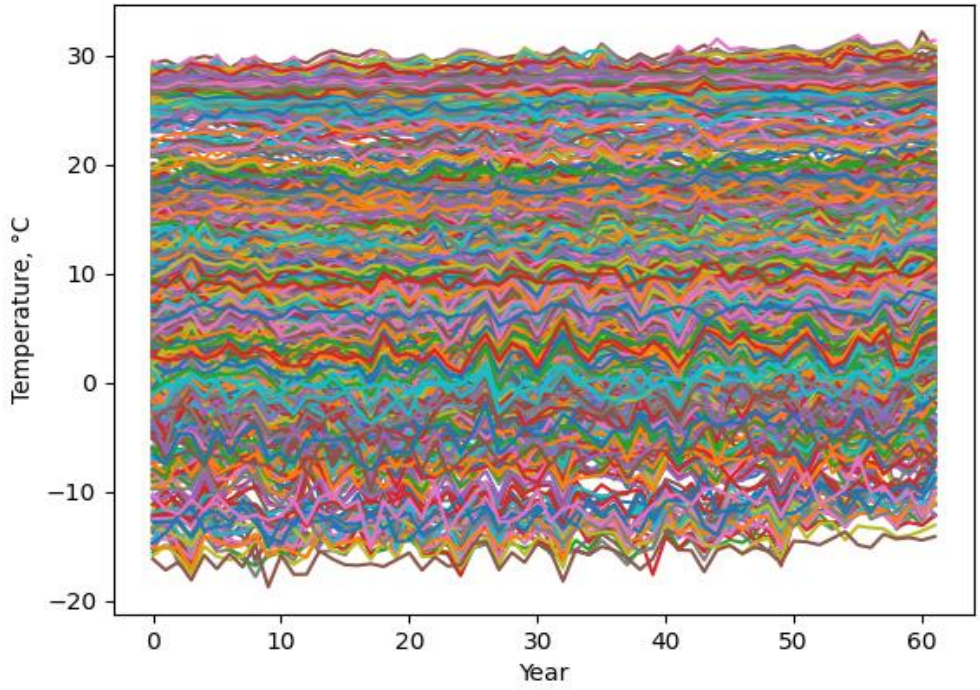

Рисунок 3. Температурные временные ряды 928 станций за 62 года

4 РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ.

4.1 Кластеризация на основе k – средних

В этом разделе мы осуществляем кластеризацию климатических данных методом k-средних. Климат характеризуется большим количеством параметров такими, как: атмосферное давление, влажность, скорость и направление ветра, температура, облачность, угол падение солнечных лучей, континентальность. Мы строим наш эксперимент на допущении, что все

климатоэкологические показатели зоны находятся в определенной корреляции и что температуру можно считать обобщающей и результирующей характеристкой. В нашем исследовании мы остановились на температуре, как одном и наиболее значимых и результирующих показателей климата.

В работе с нашими данными имеют значение и средние значения температур поведение временных рядов. На основе данных характеристик мы можем визуально оценить разделение временных рядов.

Учитывая, что большая часть наших временных рядов довольно схожа по поведению во времени мы можем уверенно применять метрику Евклидова расстояния в алгоритме k-средних.

В нашем случае мерой сходства между кластерами является средняя температура. Именно ее мы и выберем в качестве метрики.

В ходе исследования предложено произвести анализ работы алгоритма в разных временных масштабах. Минимальный отсчет температур составляет один месяц (среднемесячная температура). Таким образом было выбрано 3 интервала: 1 месяц, 1 год, весь период (62 года). Соответственным методом выбраны 3 метрики.

В ходе исследования проверялась работа алгоритма для разного количества кластеров k. Были выбраны различные значения параметра: 4, 8, 12, 16.

Среднемесячная температура

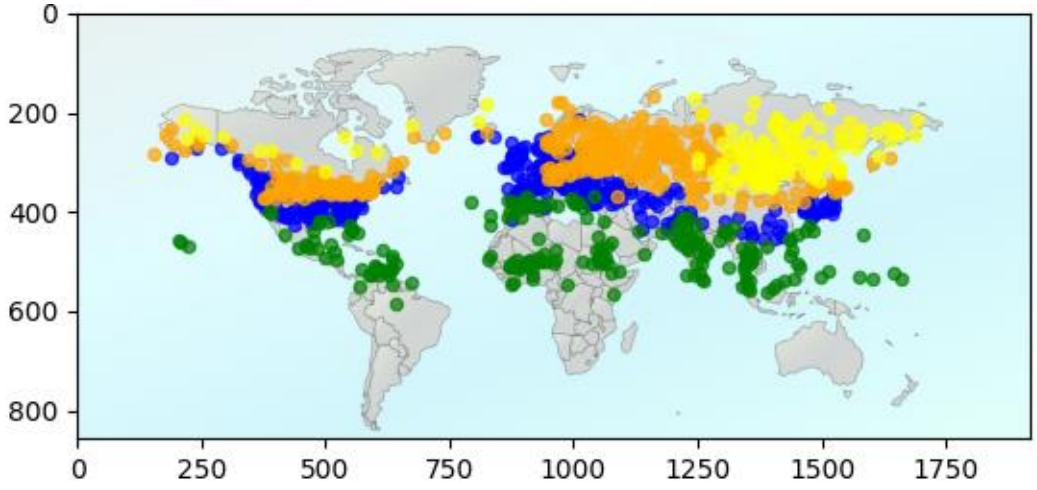

Рисунок 4. Распределение 4 кластеров по среднемесячной температуре

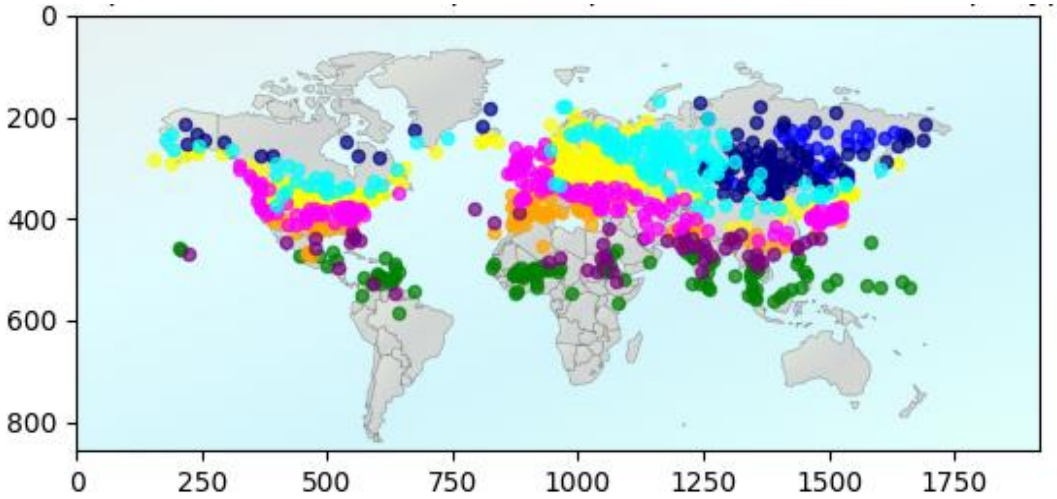

Рисунок 5. Распределение 8 кластеров по среднемесячной температуре

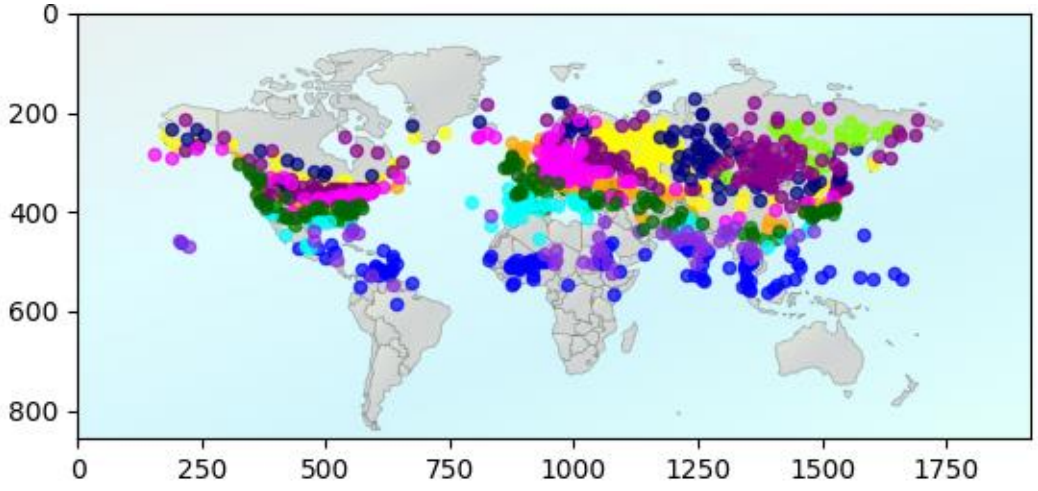

Рисунок 6. Распределение 12 кластеров по среднемесячной температуре

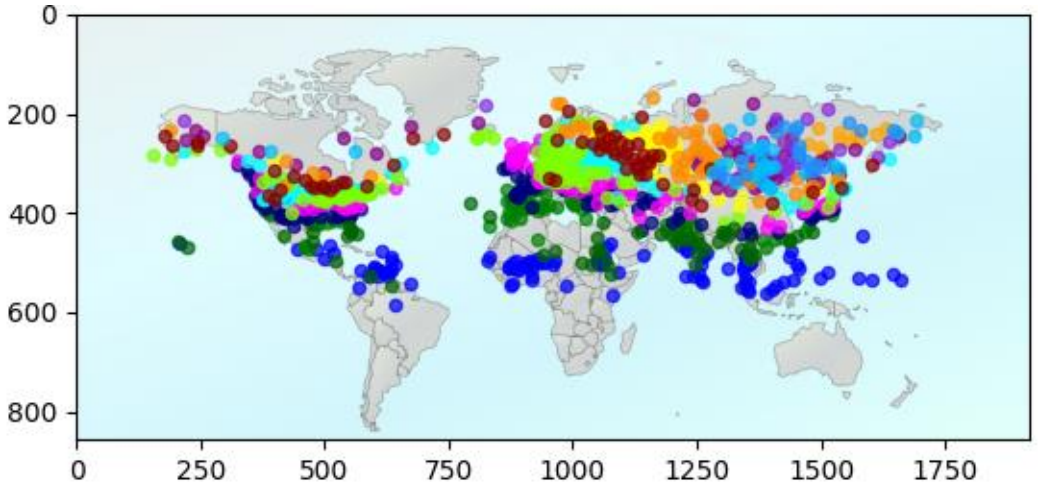

Рисунок 7. Распределение 16 кластеров по среднемесячной температуре

Вывод: как видно из изображений, распределение кластеров, полученных алгоритмом k-means по среднемесячным данным носит относительно локальный характер и характеризует локальные климатические зоны.

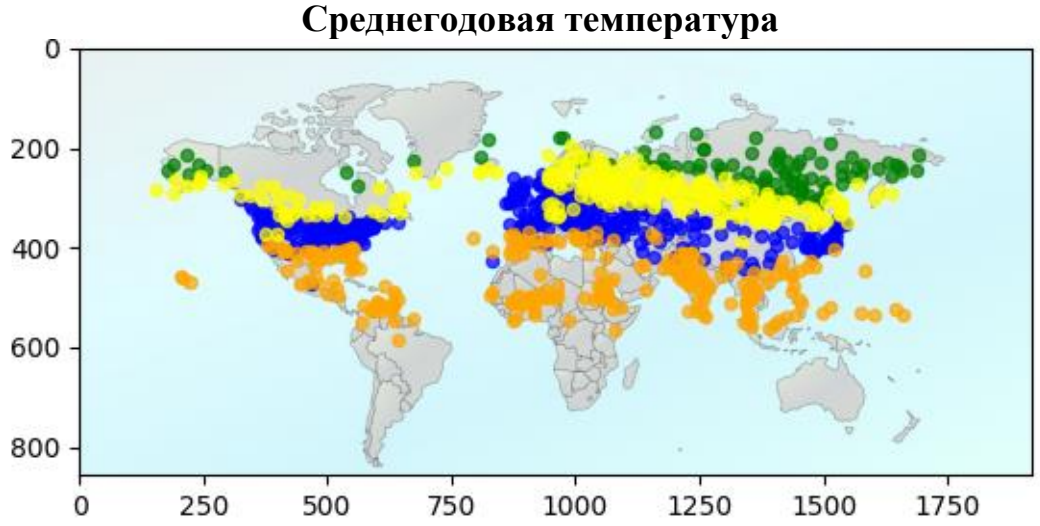

Рисунок 8. Распределение 4 кластеров по среднегодовой температуре

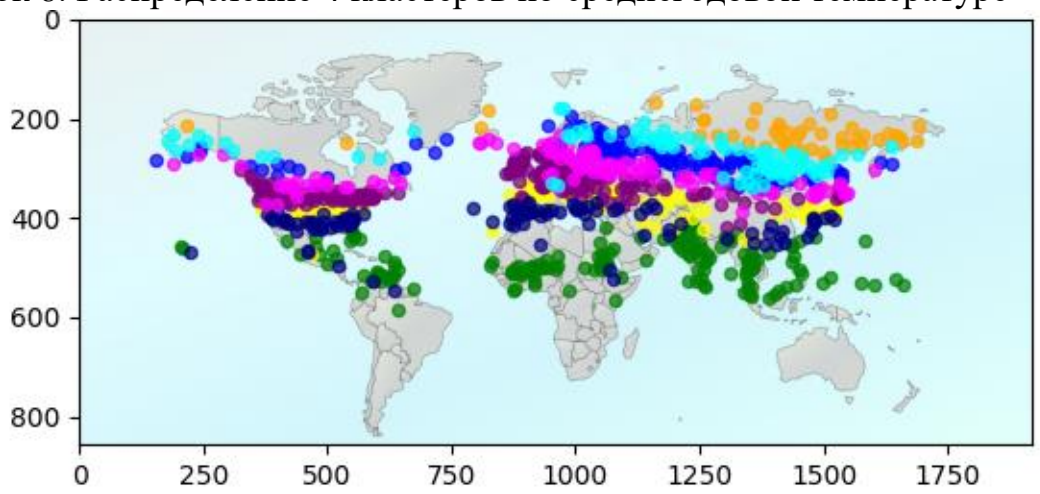

Рисунок 9. Распределение 8 кластеров по среднегодовой температуре

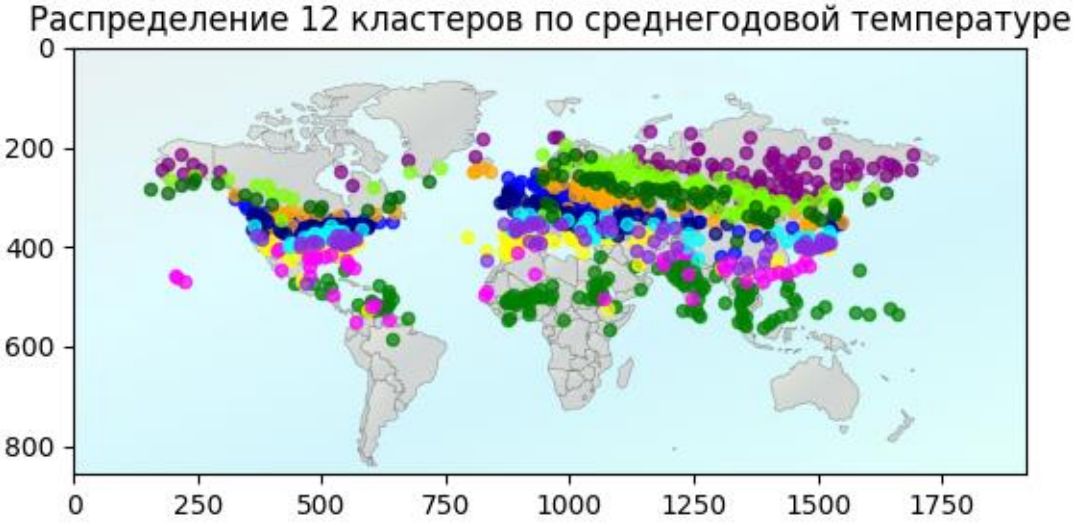

Рисунок 10. Распределение 12 кластеров по среднегодовой температуре

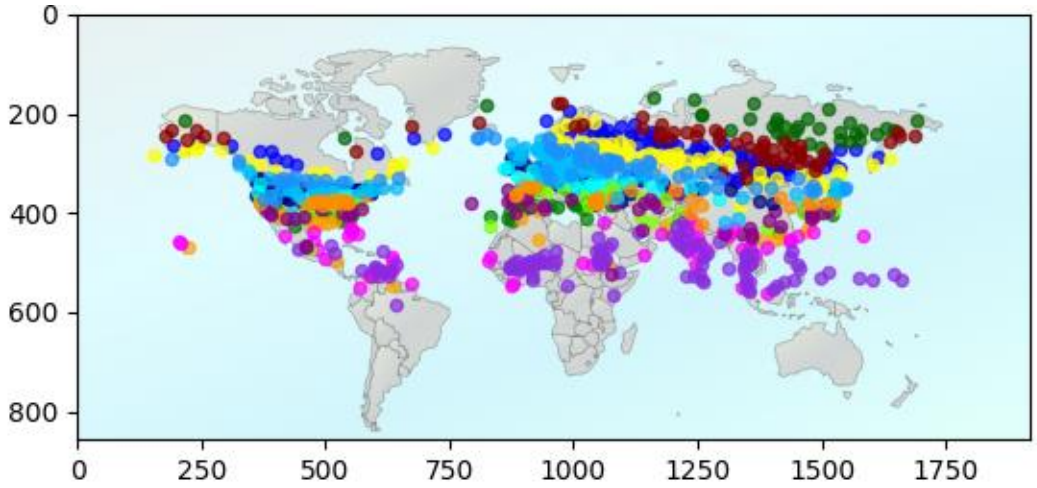

Рисунок 11. Распределение 16 кластеров по среднегодовой температуре

Вывод: распределение кластеров, полученных алгоритмом k-means по данным среднегодовых температур носит преимущественно широтный характер. И при малом количестве кластеров походит на климатические пояса Земли. При увеличение количества кластеров растет детализация климатических поясов.

Средняя температура за весь период измерений (62 года)

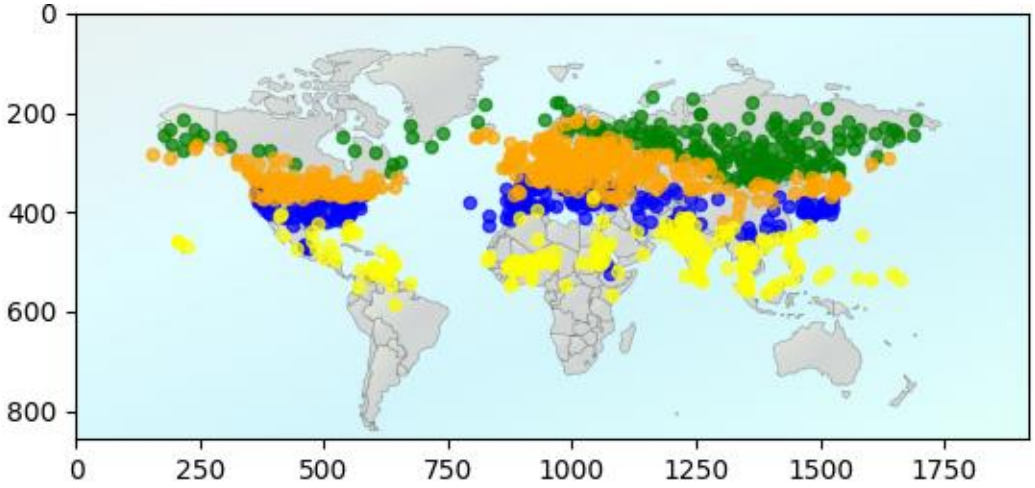

Рисунок 12. Распределение 4 кластеров по средней за 62 года температуре

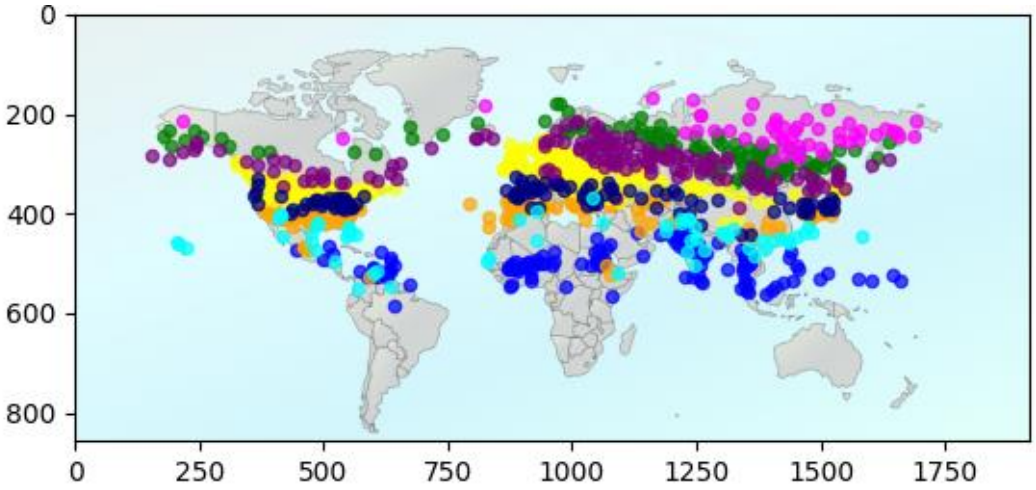

Рисунок 13. Распределение 8 кластеров по средней за 62 года температуре

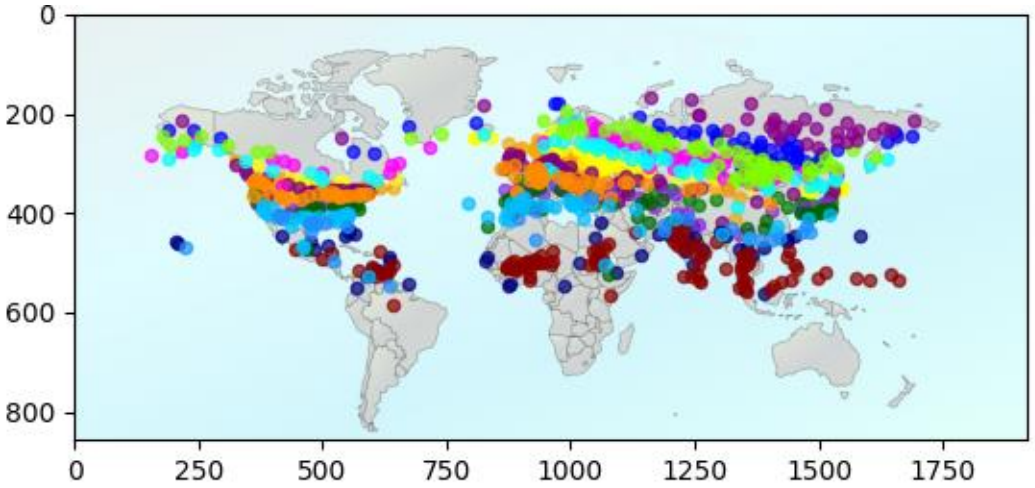

Рисунок 14. Распределение 16 кластеров по средней за 62 года температуре

Вывод: применение в качестве метрики кластеризации весь период измерений не выявил существенных изменений с метрикой периодом средней температуры в 1 год. Распределенные классы обладают схожей широтностью распределения. Можно говорить о том, что периода в 1 год вполне достаточно для выделения информативных климатических классов

4.2 Кластеризация на основе нейросетевого алгоритма Кохонена

Кластеризации на основе нейросетевого алгоритма происходила на аналогичных данных, но были различия в метрике. В данном алгоритме кластеризация происходит за счет координат и температуры метеостанции. Параметры количества кластеров аналогичны и составляют: 4, 8, 12, 16

Мера расстояния: $d_{ab}=\sqrt{(x_a-x_b)_2+(y_a-y_b)_2+(t_a-t_b)_2}$, где x – широтная координата, y – координаты долготы, y – среднемесячная/среднегодовая/средняя за 62 года температура

Среднемесячная температура

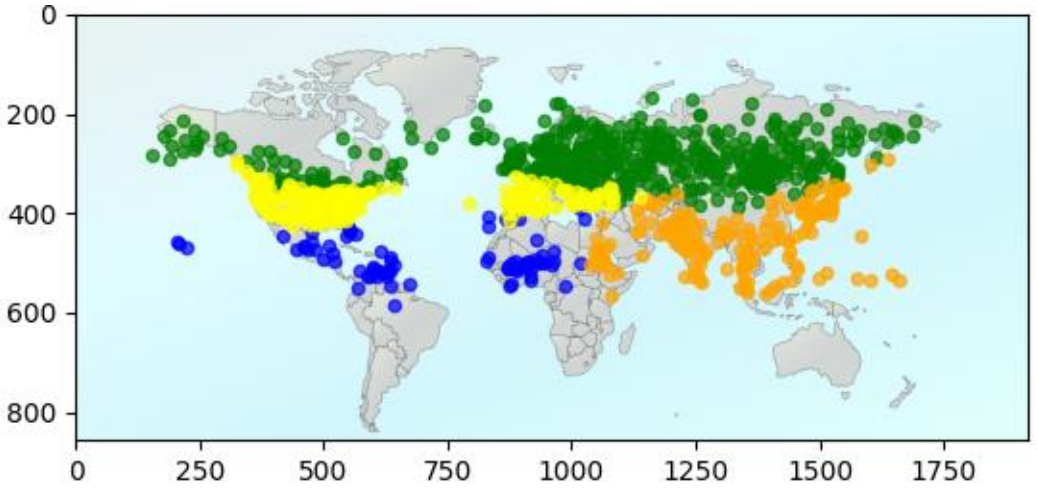

Рисунок 15. Распределение 4 кластеров по среднемесячной температуре

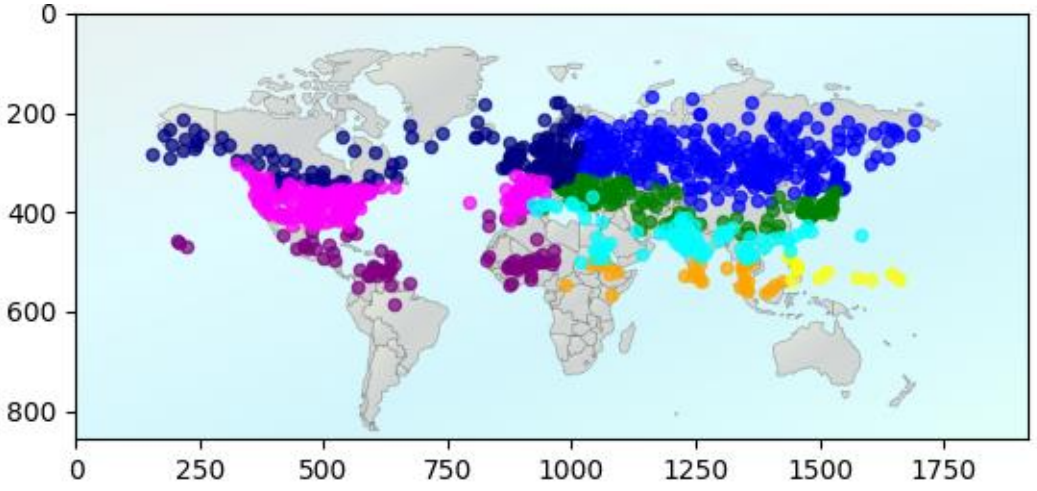

Рисунок 16. Распределение 8 кластеров по среднемесячной температуре

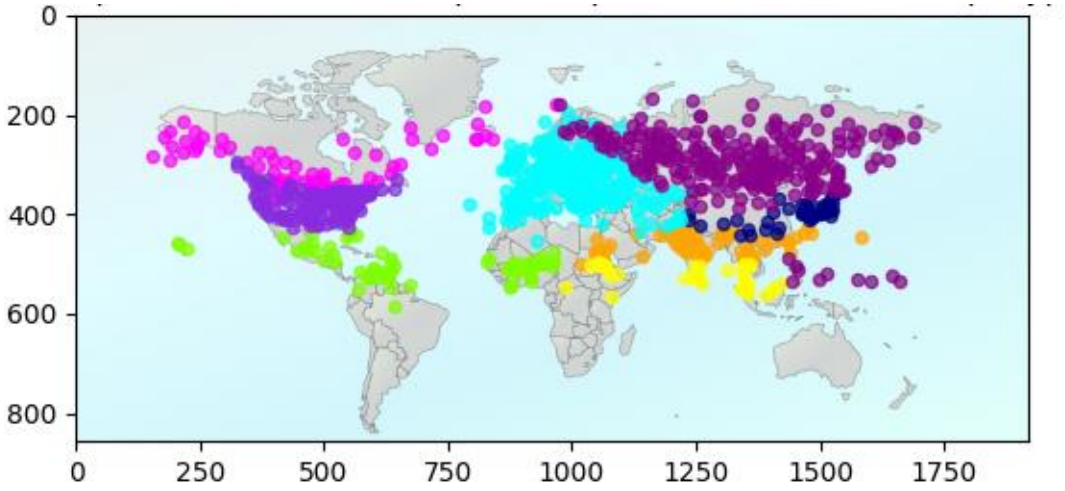

Рисунок 17. Распределение 12 кластеров по среднемесячной температуре

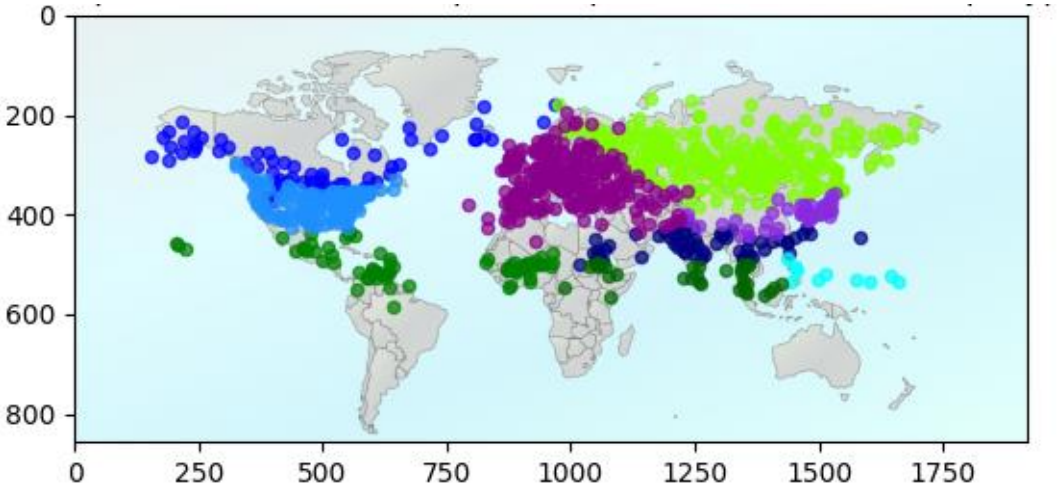

Рисунок 18. Распределение 16 кластеров по среднемесячной температуре

Вывод: кластеризация с использованием координат приводит к выделению климато-географических классов.

Среднегодовая температура

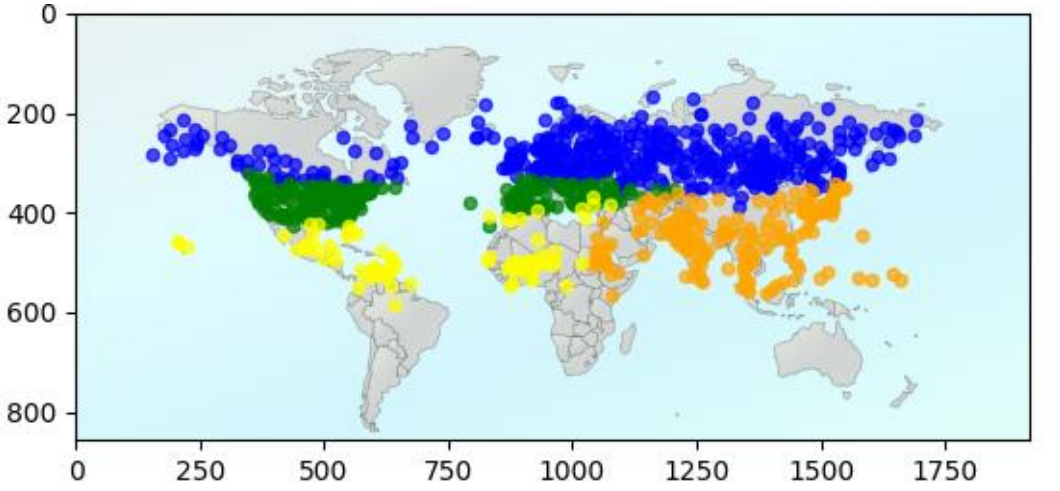

Рисунок 19. Распределение 4 кластеров по среднегодовой температуре

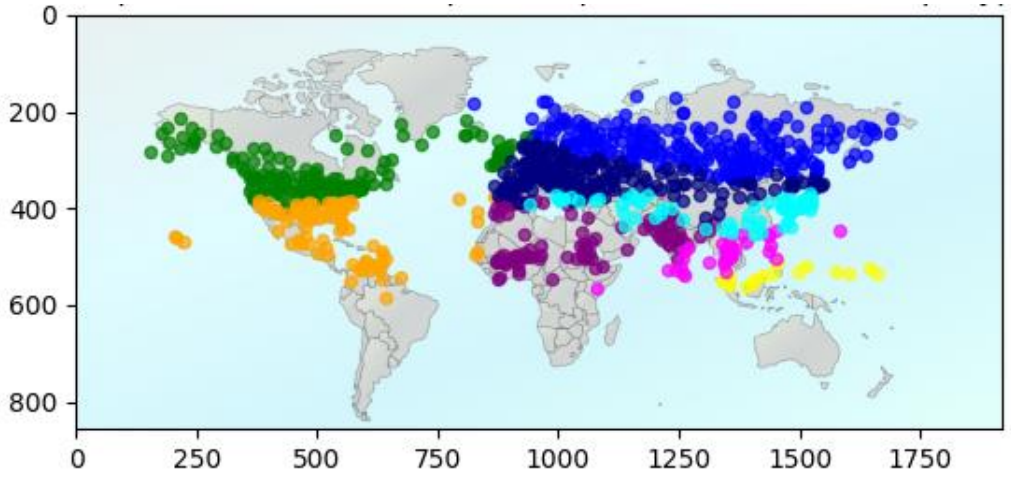

Рисунок 20. Распределение 8 кластеров по среднегодовой температуре

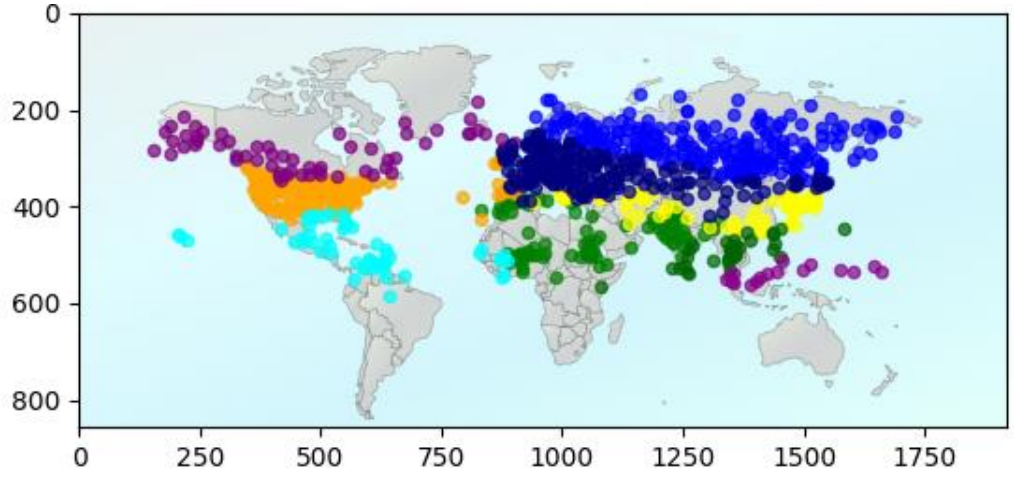

Рисунок 21. Распределение 12 кластеров по среднегодовой температуре

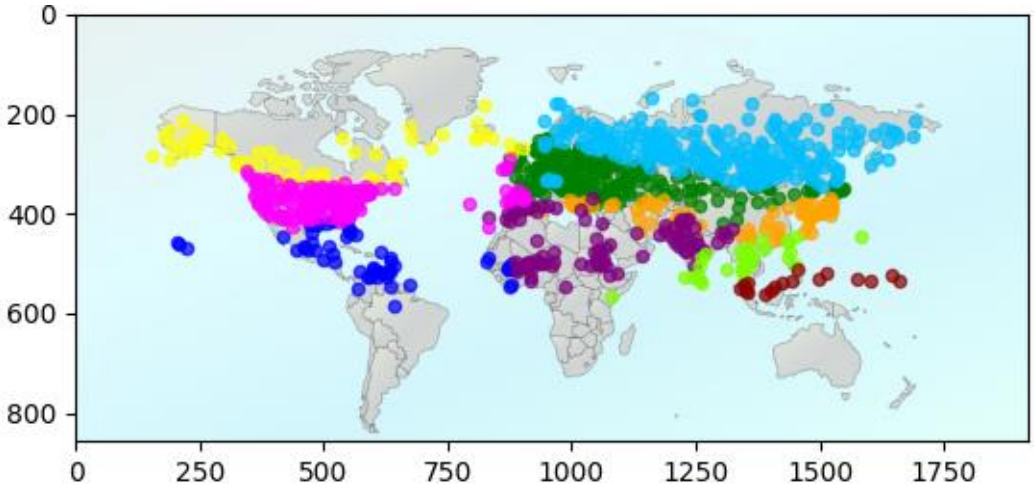

Рисунок 22. Распределение 16 кластеров по среднегодовой температуре

Вывод: распределение кластеров преимущественно широтно, как и для метода k-средних, но более локально.

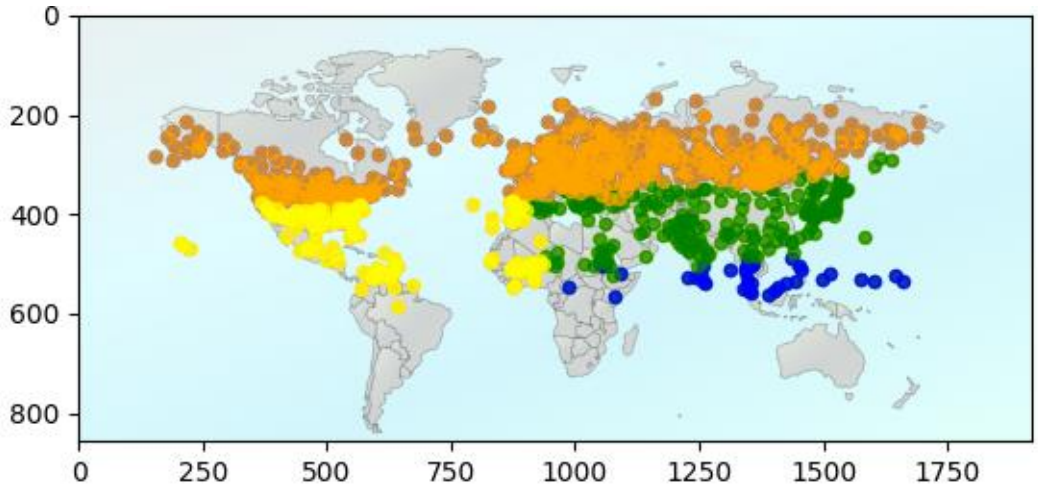

Рисунок 23. Распределение 4 кластеров по средней температуре за 62 года

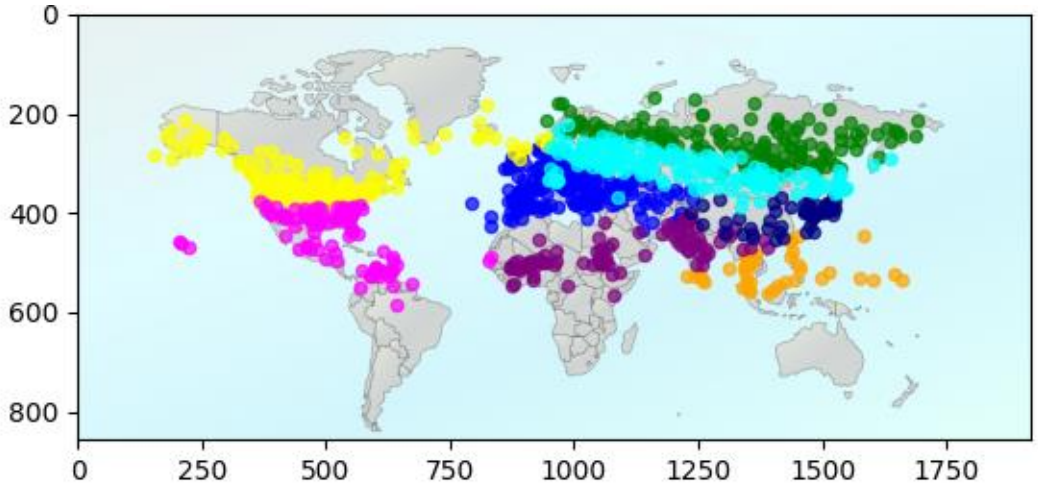

Рисунок 24. Распределение 8 кластеров по средней температуре за 62 года

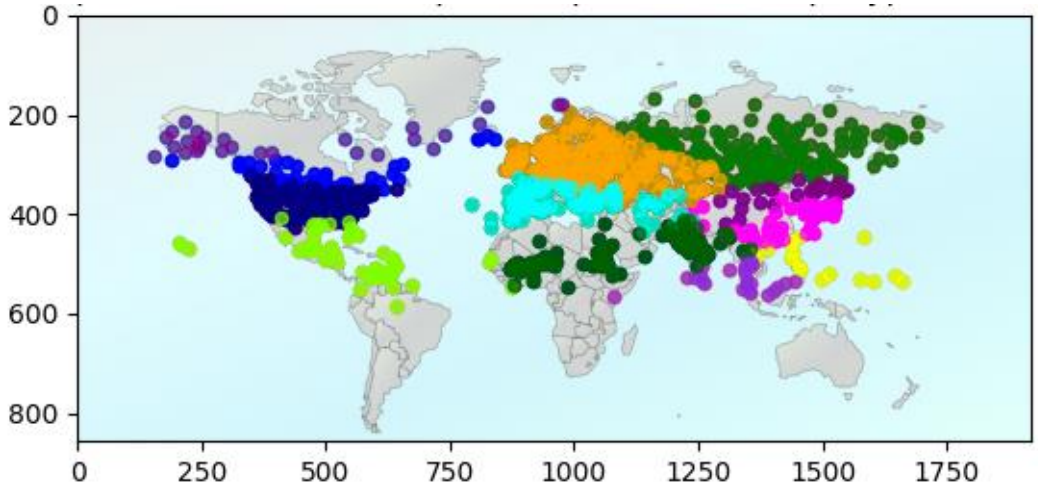

Рисунок 25. Распределение 12 кластеров по средней температуре за 62 года

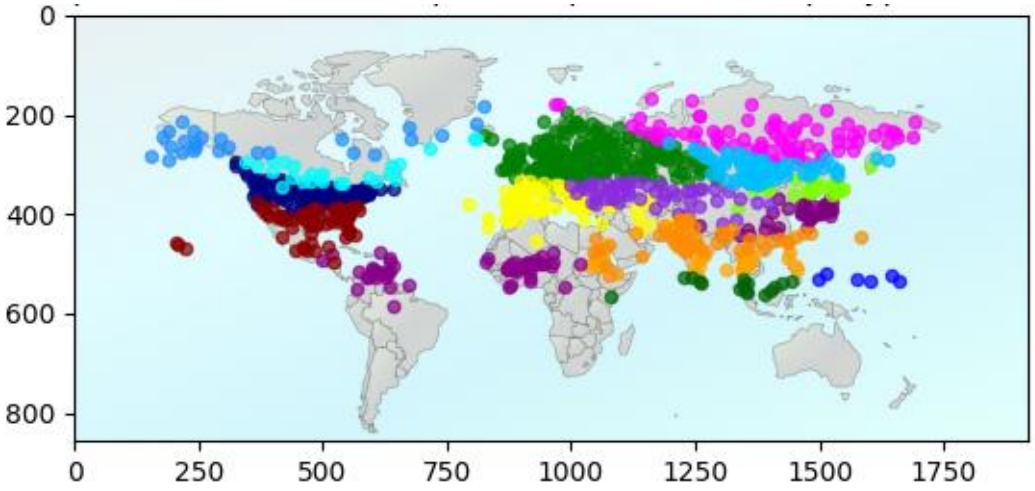

Рисунок 26. Распределение 16 кластеров по средней температуре за 62 года

Вывод: в сравнении с методом k-средних, в нейросетевом алгоритме есть существенные различия в распределении кластеров между среднегодовым расчетом и расчетом по средней температуре за 62 года. Это может быть вызвано чувствительностью алгоритма ко входным данным, либо зависимостью от координат исходных центроидов.

5. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение

5.1 Предпроектный анализ

5.1.1 Потенциальные потребители результатов исследования

Выполненная работа направлена на разработку и реализацию алгоритма, который будет решать задачи кластеризации и группировки большого объема данных, направленных на выявление групп, по некоторому признаку.

Информация о пространственно-временной изменчивости границ климатических классов востребована в сферах экономики и рационального природопользования.

В таблице приведены основные сегменты рынка по следующим критериям: размер компании-заказчика и направление деятельности. Анализ рынка проводился на основе следующих компаний: DPD (крупная компания), Томские Транспортные Линии (средняя компания), ПЭК Томск (мелкая компания).

Таблица 1. Карта сегментирования рынка

		Направление деятельности	
		.Потребности	Запасы
Размер компаний	Крупные	+	+
	Средние	+	-
	Мелкие	+	-

Согласно карте сегментирования рынка, выберем следующие сегменты: построение кластеров изображений для крупных компаний.

5.1.2 Анализ конкурентных решений

Основным методом, используемым в данной работе, является метод нечеткой кластеризации. Оценочная карта для сравнения конкурентных технических решений представлена в таблице 2

Таблица 2 – Оценочная карта для сравнения конкурентных технических решений

Критерии оценки	Вес	Баллы			Конкуренто-способность		
		$Б_{осн}$	$Б_{см}$	$Б_{пп}$	$Б_{осн}$	$Б_{см}$	$Б_{пп}$
Технические критерии оценки ресурсоэффективности							
Эффективность	0,3	4	4	4	1,2	1,2	1,2
Устойчивость	0,2	5	5	5	1	1	1
Временные затраты	0,2	5	4	4	1	0,8	0,8
Новизна метода	0,1	5	3	3	0,5	0,3	0,3
Простота реализации	0,1	5	4	5	0,5	0,4	0,5
Универсальность	0,1	4	4	4	0,5	0,4	0,4
Итого	**1**	**29**	**24**	**25**	**4,7**	**4,1**	**4,2**

По полученным результатам можно сделать вывод, что разрабатываемый алгоритм для оценки информативности является по конкурентоспособности наиболее эффективным.

5.1.3 SWOT-анализ

Матрица составляется на основе анализа рынка и конкурентных технических решений, и показывает сильные и слабые стороны проекта, возможности и угрозы для разработки. Первый этап заключается в описании сильных и слабых сторон проекта, в выявлении возможностей и угроз для реализации проекта, которые проявились или могут появиться в его внешней среде. Второй этап состоит в выявлении соответствия сильных и слабых сторон научно-исследовательского проекта внешним условиям окружающей среды. Это соответствие или несоответствие должны помочь выявить степень необходимости проведения стратегических изменений.

Соотношения параметров представлены в таблице 3.

Таблица 3 – Интерактивная матрица проекта

Сильные стороны проекта			
Возможности		Си1	Си2
	В1	+	+
	В2	+	+

Слабые стороны проекта			
Возможности		Сл1	
	В1	+	
	В2	0	
Сильные стороны проекта			
Угрозы		Си1	Си2
	У1	+	
	У2	0	
Слабые стороны проекта			
Угрозы		Сл1	
	У1	+	
	У2	+	

Матрица SWOT представлена в таблице 4.

Таблица 4 – SWOT- анализ

	Сильные стороны: С1. Алгоритм самостоятельно определяет количество кластеров. С2.Разработанный алгоритм обладает достаточно высокой точностью.	**Слабые стороны**: Сл1. При пограничных условиях возможна долгая сходимость.
Возможности: В1. Расширение функционала. В2. Автоматизация алгоритма.	При некоторых модификациях алгоритм может справляться с решением многих задач. Например, с задачей выделение климатических групп или задачей распознавания образов.	Автоматизация алгоритма позволит значительно ускорить его работу.
Угрозы: У1. Схожесть с алгоритмом k-means может повлечь за собой низкий спрос алгоритма на рынке. У2. В некоторых ситуациях возможна слабая устойчивость алгоритма.	При всей схожести алгоритма с k-means он обладает существенным отличием и преимуществом, а именно отсутствием необходимости задавать количество кластеров заранее, что делает работу с ним простой и удобной.	Неустойчивость алгоритма может спровоцировать низкий спрос на продукт.

Таким образом. самыми большими преимуществами оказывается отсутствие необходимости заранее указывать количество кластеров. Наиболее слабая сторону проекта – потенциальная неустойчивость алгоритма и долгая сходимость.

5.1.4 Оценка готовности проекта к коммерциализации

Заполним форму в таблице 5, содержащую показатели степени проработанности проекта с позиции коммерциализации и компетенциям

разработчика научного проекта.

Таблица 5 Оценка степени готовности проекта к коммерциализации

№ п/п	Наименование	Ком-мерция	Компе-тенции
1	Определен имеющийся научно-технический задел	3	3
2	Определены перспективные направления коммерциализации научно-технического задела	5	3
3	Определены отрасли и технологии для предложения	3	3
4	Определена форма для представления на рынок	2	2
5	Определены авторы и осуществлена охрана их прав	1	1
6	Оценена стоимость интеллектуальной собственности	1	1
7	Проведены маркетинговые исследования рынков сбыта	1	1
8	Разработан бизнес-план коммерциализации разработки	1	1
9	Определены пути продвижения разработки на рынок	2	2
10	Разработана стратегия (форма) реализации разработки	3	3
11	Проработаны вопросы международного сотрудничества и выхода на зарубежный рынок	1	1
12	Проработаны вопросы использования услуг инфраструктуры поддержки, получения льгот	1	1
13	Проработаны вопросы финансирования коммерциализации научной разработки	1	4
14	Имеется команда для коммерциализации разработки	1	1
15	Проработан механизм реализации научного проекта	4	5
	ИТОГО БАЛЛОВ	30	32

Итоговые значения проработанности научного проекта 30 и знания у разработчика 32, что говорит о том, что некоторые аспекты проекта практически не были проработаны. Так как, работа носит преимущественно научный характер, это приемлемо.

5.2. Инициация проекта

Устав научного проекта магистерской работы:

1. Цели и результат проекта. Информация по заинтересованным сторонам представлена в таблице 6:

Таблица 6. Заинтересованные стороны проекта

Заинтересованные стороны	**Ожидания**
Пользователь	Выделение уникальных климатических классов
Компания пользователя	Скорость работы и простота обслуживания
Разработчик	Получение прибыли
Научный руководитель, магистр	Выполненная выпускная квалификационная работа

Цели и результат проекта представлены в таблице 7.

Таблица 7 Цели и результат проекта

Цели проекта	• Изучить методы и алгоритмы кластеризации данных • Выбрать средства разработки • Реализовать и протестировать различные методы кластеризации
Ожидаемые результаты	• Реализовал алгоритм кластеризации данных без необходимости указывать количество кластеров, выделяющий уникальные климатические классы • Сдана выпускная квалификационная работа
Критерии приёмки	• Успешное тестирование функционала в соответствии с функциональным требованием
Требования к результату проекта	• Получены данные о точности работы системы

5.3 Планирование управления научно-техническим проектом

5.3.1 План проекта

Диаграмма Ганта строится в виде таблицы с разбивкой по декадам (10 дней) за период времени выполнения научного проекта. При этом работы на графике следует выделить различной штриховкой в зависимости от исполнителей, ответственных за ту или иную работу.

Таблица 8. Календарный план-график проведения НИОКР по теме

Код	Вид	Исп	Тк, кал,	Продолжительность

			дн.	Янв.			Февр.			Март			Апр.			Май		
				1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
1	Выбор направления исследования	Р, И	4/10															
2	Описание требований	И	10															
4	Составление технического задания	И	10															
5	Изучение литературы	И	30															
6	Подбор и реализация различных алгоритмов кластеризации	И	60															
7	Анализ полученных результатов	И	20															
15	Проверка работы	Р	4/10															

– Руководитель(Р) – Инженер (С)

5.3.2 Бюджет научного исследования

Таблица 9 – Бюджет затрат НТИ

Сырье	Оборудование	Осн зп	Доп зп	Накл рас	Отчисления на социальные нужды	Итого
	60 000	192224	19240	33891	63547	**368902**

Для разработки нынешней системы требуется персональный компьютер и серверы на время жизненного цикла программного продукта. Среда и средство разработки, программный софт и другие комплектующие, нужные для разработки, распространяются бесплатно и не требуют дополнительных затрат (таблица 10).

Таблица 10 – Расчет затрат на «Спецоборудование для научных работ»

№	Наименование оборудования	Кол-во	Цена, руб.	Стоимость, руб.
1.	Персональный компьютер	1	60 000	60 000
2.	Среда разработки JetBrains PyCharm	1	-	-
Итого				60 000

Должность руководителя – доцент, к.т.н. – 37700 рублей в месяц

Должность инженера – студент – 23800 рублей в месяц

Расчет основной заработной платы сводится в таблице 11.

Таблица 11 – Расчет основной заработной платы

№ п / п	Наименование этапов	Исп	Трудоемкость, чел.-дн.	Зарплата, чел.-дн., руб	Всего руб.
1		Р	8	37700	37700
2		С	150	23800	

$$C_{зп} = З_{осн} + З_{доп},$$

где $З_{осн}$ – основная заработная плата;

$З_{доп}$ – дополнительная заработная плата.

Основная заработная плата $З_{осн}$ руководителя рассчитывается по следующей формуле:

$$З_м = З_{дн} \bullet Т_{раб}$$

где $Т_{раб}$– продолжительность работ, выполняемых научно-техническим работником, рабочие дни. (таблица 14);

$З_{дн}$– среднедневная заработная плата работника, руб.

Значит, для руководителя:

$$З_м = 37700 * 1{,}3 = 49010 \text{ рублей}$$

Среднедневная заработная плата рассчитывается по формуле:

$$З_{дн}=(З_м \bullet М)/F_д$$

где $З_м$ – месячный должностной оклад работника, руб

М – количество месяцев работы без отпуска в течение года:

при отпуске в 45 раб. дней М=10,4 месяца, 6-дневная неделя;

$F_д$ – действительный годовой фонд рабочего времени научно-технического персонала (в рабочих днях) (таблица 14). Тогда,

Для руководителя:

$$З_{дн} = \frac{49010 * 10{,}4}{254} = 2006 \text{ рублей}$$

Для дипломника:

$$З_{дн} = \frac{23800 * 11,2}{217} = 1228 \text{ рублей}$$

Баланс рабочего времени представлен в таблице 12

Таблица 12 – Баланс рабочего времени

Показатели рабочего времени	**Руководитель**	**Инженер**
Календарное число дней	365	365
Количество нерабочих дней - выходные дни - праздничные дни	 52 11	 82 14
Потери рабочего времени - отпуск - невыходы по болезни	 56 –	 24 –
Действительный годовой фонд рабочего времени	254	217

Таблица 13 – Результаты расчета основной заработанной платы

Исполнители	$З_б$, руб.	$k_р$	$З_м$, руб	$З_{дн}$, руб.	$Т_р$, дн.	$З_{осн}$, руб.
Руководитель	33700	1.3	49010	2006	4	8024
Инженер	23800	1.3	30940	1228	150	184200
Итого по статье $З_{осн}$:			192224			

В таблице 14 приведен расчёт основной и дополнительной заработной платы.

Таблица 14 – Заработная плата исполнителей ВКР, руб

Заработная плата	Руководитель	Инженер
Основная зарплата	8204	184200
Дополнительная зарплата	820	18420
Зарплата исполнителя	9204	202620
Итого	211824	

Коэффициент отчислений на уплату во внебюджетные фонды (пенсионный фонд, фонд обязательного медицинского страхования и пр.) равен 0.3:

Для руководителя - 2761 руб.

Для инженера - 60786 руб.

Итого 63547 рублей

5.2.7 Накладные расходы

Накладные расходы составляют 16% от суммы основной и дополнительной заработной платы, работников, непосредственно участвующих в выполнение темы. Расчет накладных расходов:

$$С_{накл}= 33891 \text{ руб.}$$

В результате было получено, что бюджет на разработку НТИ составит **368902** руб.

5.3 Оценка сравнительной эффективности исследования

Определение эффективности происходит на основе расчета интегрального показателя эффективности научного исследования. Его нахождение связано с определением двух средневзвешенных величин: финансовой эффективности и ресурсоэффективности.

Интегральный финансовый показатель разработки определяется как:

$$I_{финр}^{испi} = \frac{Ф_{pi}}{Ф_{max}},$$

где $Ф_{pi}$ – стоимость варианта исполнения, $Ф_{max}$ – максимальная стоимость исполнения научно-исследовательского проекта.

Т.к. стоимость всех вариантов исполнения одинакова, интегральные финансовые показатели также будут одинаковы и равны 1.

Интегральный показатель ресурсоэффективности вариантов исполнения объекта исследования можно определить следующим образом:

$$I_{pi} = \sum a_i \cdot b_i,$$

где a_i – весовой коэффициент варианта исполнения разработки, b_i – балльная оценка варианта исполнения разработки.

Интегральный показатель эффективности вариантов исполнения разработки определяется на основании интегрального показателя ресурсоэффективности и интегрального финансового показателя по формулам:

$$I_{исп.1} = \frac{I_{р-исп.1}}{I_{финр}^{исп.1}},$$

$$I_{исп.2} = \frac{I_{р-исп.2}}{I_{финр}^{исп.2}}.$$

Так как интегральные финансовые показатели одинаковы и равны 1, то интегральные показатели эффективности вариантов исполнения разработки равны соответствующим интегральным показателям ресурсоэффективности.

Сравнение интегрального показателя эффективности вариантов исполнения разработки позволит определить сравнительную эффективность проекта и выбрать наиболее целесообразный вариант из предложенных.

Сравнительная эффективность проекта:

$$Э_{ср} = \frac{I_{исп.i}}{I_{исп.1}}.$$

В пункте 4.2 было рассмотрено два варианта исполнения алгоритма. На основании этого необходимо провести сравнительную характеристику вариантов исполнения (табл. 15).

Таблица 15 – Сравнительная оценка характеристик вариантов исполнения проекта

Критерии	Весовой коэффициент	Исп.1	Исп.2
1. Повышение производительности труда	0,3	5	5
2. Удобство в эксплуатации	0,2	4	3
3. Удобство в считывании исходных данных	0,2	5	2
4. Скорость работы	0,1	4	3
5. Простота эксплуатации	0,1	4	3
6. Техническая поддержка платформы	0,1	5	3
I_{pi}		**4,6**	**3,4**

На основании полученных показателей выполним сравнение интегрального показателя эффективности вариантов исполнения разработки (табл. 16).

Таблица 16 – Сравнительная эффективность разработки

Показатели	Исп.1	Исп.2
Интегральный финансовый показатель разработки	1	1

Интегральный показатель ресурсоэффективности разработки	4,6	3,4
Интегральный показатель эффективности	4,6	3,4
Сравнительная эффективность вариантов исполнения		1,35

С позиции финансовой и ресурсной эффективности на основании таблицы, первый вариант исполнения системы наиболее выгодный. Данный вариант исполнения и используется в выпускной квалификационной работе.

Итак, в ходе данной работе проведена оценка готовности проекта к коммерциализации, которая показала, что перспективность разработки низкая.

В рамках процессов инициации определены внутренние и внешние заинтересованные стороны проекта с их ожиданиями от проекта, цели и результат проекта.

План проекта представлен на диаграмме Ганта, из которого видно какой исполнитель (магистр или руководитель) какой вид работ осуществлял и в течении какого количества дней.

Рассчитан бюджет проекта, который составил 368902 рубля. С позиции финансовой и ресурсной эффективности первый вариант исполнения системы наиболее выгодный. Данный вариант исполнения представлен в выпускной квалификационной работе.

6. СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ

Целью работы является создание алгоритма кластеризации климатических данных. Основная работа с алгоритмом и работа по её созданию производится с использованием персонального компьютера в жилом помещении.

Работа проводилась в жилом помещении общежития ТПУ Усова 15б. Характеристика помещения, где проводились работы по ВКР: ширина комнаты составляет b= 4.5 м, длина a=6 м, высота H = 2,8 м. Площадь помещения будет составлять $S=ab$=27 м^2, объем $V=abh$=81.4 м^3; присутствует окно, через которое может производится вентиляция помещения, принудительная вентиляция отсутствует; в зимнее время помещение отапливается; в помещении используется комбинированное освещение. Схема помещения представлена на рисунке 5.1.

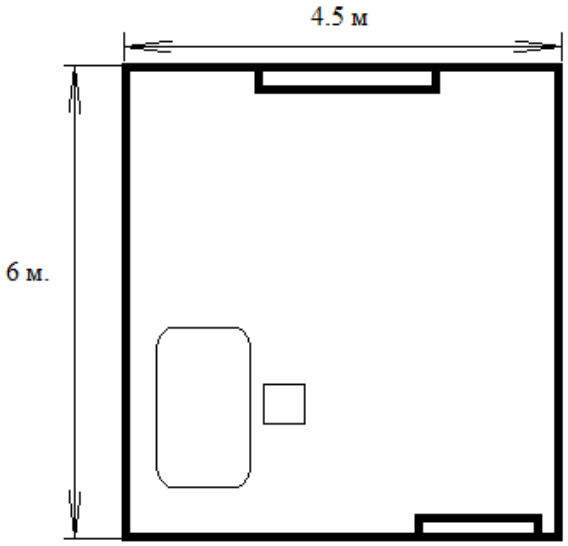

Рис. 5.1 Схема рабочего помещения.

6.1 Правовые аспекты обеспечения безопасности

Исходя из общепризнанных принципов и норм международного права и в соответствии с Конституцией Российской Федерации основными принципами правового регулирования трудовых отношений обеспечение права каждого работника на справедливые условия труда, в том числе на условия труда, отвечающие требованиям безопасности и гигиены, регламентируемых признанными в РФ нормативными правовыми документами, такими как: ГОСТ, СанПиН, СНиПТОИ [11].

Продолжительность рабочего дня не должна превышать 40 часов в неделю. В течение рабочего дня работнику должен быть предоставлен перерыв для отдыха и питания продолжительностью не более двух часов и не менее 30 минут. Всем работникам предоставляются выходные дни [11].

В целях обеспечения прав и свобод человека и гражданина работодатель и его представители должны соблюдать требования ТК РФ по получение и обработке персональных данных. В целях обеспечения защиты персональных данных, хранящихся у работодателя, работники имеют право на: полную информацию об их персональных данных и обработке этих данных; свободный бесплатный доступ к своим персональным данным; доступ к медицинской документации, отражающей состояние их здоровья; требование об исключении или исправлении неверных или неполных персональных данных, а также данных, обработанных с нарушением требований настоящего Кодекса или иного федерального закона [12].

6.2 Эргономические требования к рабочему месту

Место оператора ЭВМ регламентируется следующими нормативными актами: СанПиН 1.2.3685-21, ТК РФ от 30.12.2001 N 197-ФЗ (ред. от 25.02.2022), ГОСТ 22269-76, ГОСТ Р 50923-96, ГОСТ 12.2.032-78 [11, 12, 13, 14, 15].

В случаях, когда характер работы требует постоянного взаимодействия с компьютером (работа программиста разработчика) с напряжением внимания и сосредоточенности, при исключении возможности периодического переключения на другие виды трудовой деятельности, не связанные с ПЭВМ, рекомендуется организация перерывов на 10–15 мин. через каждые 45–60 мин. работы. Конструкция рабочей мебели (рабочий стол, кресло, подставка для ног) должна обеспечивать возможность индивидуальной регулировки соответственно росту пользователя и создавать удобную позу для работы. [1].

6.3 Производственная безопасность

Для обеспечения безопасности во время эксплуатации и разработки программы, необходимо провести анализ вредных и опасных воздействий на человека, которые могут возникать при разработке или эксплуатации проекта. Производственный фактор является вредным, в том случае если он приводит к заболеванию работника. В случае если его воздействие может привести к травме, то фактор является опасным. Выявленные вредные и опасные факторы приведены в таблице ниже.

Таблица 1 - Возможные опасные и вредные факторы

Факторы (ГОСТ 12.0.003-2015)	Нормативные документы
1.Отклонение показателей микроклимата	СанПиН 1.2.3685-21 «Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания» [36]
2. Превышение уровня шума	СП 51.13330.2011 Защита от Шума. Актуализированная редакция СНиП 23-03-2003 [40]
3.Отсутствие или недостаток естественного света	СП 52.13330.2016 Естественное и искусственное освещение. Актуализированная редакция СНиП 23-05-95*[41], [42].

4.Недостаточная освещенность рабочей зоны	СП 52.13330.2016 Естественное и искусственное освещение. Актуализированная редакция СНиП 23-05-95*[41], [42].
5.Повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека	ГОСТ 12.1.038-82 Система стандартов безопасности труда (ССБТ). Электробезопасность. Предельно допустимые значения напряжений прикосновения и токов. [44]

6.3.1 Вредные производственные факторы

6.3.1.1 Отклонения показателей микроклимата от нормы в помещении

Пониженная или же повышенная температура воздуха рабочей зоны считается вредным производственным фактором и является фактором микроклимата рабочей среды, параметры которого регулируются СанПиН 1.2.3685-21 [12].

Во время работы с ПЭВМ в производственных помещениях происходит систематическое выделение тепла вычислительной техникой, а также вспомогательными устройствами и средствами освещения. По причине того, что оператор находится поблизости от источников выделения тепла, данный фактор является одним из вредных факторов производственной среды оператора ПЭВМ, а высокая температура воздуха способствует быстрому перегреву организма и повышению утомляемости. Влажность оказывает сильное влияние на терморегуляцию организма. Высокие показатели относительной влажности (более 85 %) затрудняют терморегуляцию, показатели влажности менее 20 % вызывают пересыхание слизистых оболочек человека.

Санитарные нормы устанавливают оптимальные и допустимые значения величин показателей микроклимата рабочих мест для различных категорий работ в теплый и холодный периоды года. Для создания благоприятных условий труда и повышения производительности необходимо поддерживать оптимальные параметры микроклимата производственных помещений. Для

этого должны быть предусмотрены следующие средства: центральное отопление, вентиляция (искусственная и естественная), искусственное кондиционирование. Исходя из требований, определённой вышеприведённой нормативной документацией, в жилом помещении, в котором проводилось исследование поддерживалась температура равная 20 –21 °С, при относительной влажности в 55–58%. В зимнее время в помещении предусмотрена система водяного отопления со встроенными нагревательными элементами и терморегуляторами. Также, в некоторых случаях, целесообразно обеспечить питьевое водоснабжение. В помещениях для работы с ПЭВМ должна производиться ежедневная влажная уборка, а также систематическое проветривание после каждого часа работы [12].

6.3.1.2 Недостаточная освещенность рабочей зоны

Недостаточная освещенность рабочей зоны является вредным производственным фактором, возникающим при работе с ПЭВМ. Причиной недостаточной освещенности являются недостаточность естественного освещения, недостаточность искусственного освещения, пониженная контрастность. Работа с компьютером подразумевает постоянный зрительный контакт с дисплеем ПЭВМ и занимает от 80 % рабочего времени. Недостаточность освещения снижает производительность труда, увеличивает утомляемость и может привести к появлению профессиональных болезней зрения.

Существуют общие требования и рекомендации к организации освещения на рабочем месте. Рабочее помещение должно иметь естественное и искусственное освещение, соответствующее показателям. В качестве источников искусственного освещения должны быть использованы люминесцентные лампы, лампы накаливания – для местного освещения. Искусственное освещение в помещениях для эксплуатации ПЭВМ должно осуществляться системой общего равномерного освещения, а рабочие места следует размещать таким образом, чтобы естественный свет падал преимущественно слева, а дисплеи монитора были ориентированы боковой

стороной к световым проемам [17].

Требования к освещению согласно СП 52.13330.2016 [17]. приведены в таблице 5.1:

Таблица 5.1 – Требования к освещению на рабочих местах с ПК [17].

Вид	Требование
Освещенность на рабочем столе	200-400 лк
Освещенность на экране ПК	Не выше 200 лк
Блик на экране	Не выше 40 кд/м2
Прямая блеклость источника света	200 кд/м2
Показатель ослепленности	Не более 20
Показатель дискомфорта	Не более 15
Отношение яркости	
Между рабочими поверхностями	3:1-5:1
Между поверхностями стен и оборудования	10:1
Коэффициент пульсации	Не более 10%

Произведём расчёт этого фактора. Из схемы, которая представлена на рисунке 1 видно, что помещение имеет одно дверь и окно. Площадь помещения составляет 27 м^2.

Первой задачей размещения светильников является определение расчетной высоты подвеса H_D:

$$H_D = H - h_n - h_\partial,$$

где H – высота помещения, м;

h_n – расстояние светильников от перекрытия, как правило, принимается в пределах 0–1,5 м;

h_∂ – высота рабочей поверхности над полом, м.

Соблюдение данных мер позволит сохранить зрение работника или избежать пагубного воздействия на глаза. Так как высота потолка данного

помещения 2,8, то оптимальное значение размещения 2.1 м. Выбираем лампу дневного света ЛД-40, световой поток которой равен *ФЛД* = 2300 Лм. Выбираем светильники с люминесцентными лампами типа ОДОР-2-40. Этот светильник имеет две лампы мощностью 40 Вт каждая, длина светильника равна 1227 мм, ширина – 265 мм.

На первом этапе определим значение индекса освещенности:

$$i = S/(a+b)*h,$$

где S – площадь помещения; h–расчетная высота подвеса светильника, м; a и b–длина и ширина помещения, м.

В результате проведенных расчетов получаем значение индекса освещенности, равное $i = 1.5$.

Расстояние между соседними светильниками или рядами определяется по формуле:

$$L = \lambda \cdot h = 1{,}1 \cdot 1{,}55 = 1{,}6 \text{ м}.$$

Число рядов светильников в помещении:

$$Nb = \mathrm{b/L}$$

Число светильников в ряду:

$$Na = \mathrm{a/L}$$

Общее число светильников:

$$N = Na \cdot Nb = 4 \cdot 3 = 12.$$

Учитывая, что в каждом светильнике установлено две лампы, общеечисло ламп в помещении N=24.

Расстояние от крайних светильников или рядов до стены определяется по формуле: l = L/3 и равняется 0.53 м

Размещаем светильники в три ряда, тогда световой поток лампы определяется

по формуле:

$$Ф = \frac{E_H \cdot S \cdot K_з \cdot Z}{N \cdot \eta}$$

где $E_н$ – нормируемая минимальная освещенность по СНиП 23-05-95, лк; S– площадь освещаемого помещения, м^2; $K_з$ – коэффициент запаса, учитывающий загрязнение светильника (источника света, светотехнической арматуры, стен и пр., т.е. отражающих поверхностей), наличие в атмосфере цеха дыма, пыли; Z– rоэффициент неравномерности освещения, отношение $E_{ср}/E_{min}$. Для люминесцентных ламп при расчетах берется равным 1,1; N – число ламп в помещении; η – коэффициент использования светового потока [17].

Данное помещение относится к типу помещения со средним выделением пыли, в связи с этим $K_з$ = 1,5; состояние потолка – свежепобеленный, поэтому значение коэффициента отражения потолка $\rho_п$=70; состояние стен – побеленные бетонные стены, поэтому значение коэффициента отражения стен $\rho_с$ = 50.

Коэффициент использования светового потока, показывающий какая часть светового потока ламп попадает на рабочую поверхность, для светильников типа ОДОР с люминесцентными лампами при $\rho_п$=70%, $\rho_с$ = 50% и индексе помещения i=1,5 равен η=0,47.

Нормируемая минимальная освещенность при использовании ЭВМ и одновременной работе с документами должна быть равна 600лк.

$$Ф = \frac{E_H \cdot S \cdot K_з \cdot Z}{N \cdot \eta} = 2106 \text{ Лм}$$

Для люминесцентных ламп с мощностью 40 Вт и напряжением сети 220В, стандартный световой поток ЛД равен 2300 Лм.

Таким образом световой поток светильника не выходит за пределы требуемого диапазона, а отклонение не превышает 8%

6.3.1.3 Производственные шумы

Шум – это совокупность звуков, неблагоприятно воздействующих на организм человека и мешающих его работе и отдыху.

Допустимый уровень шума – это уровень, который не вызывает у человека беспокойства и значительных изменений показателей функционального состояния систем и анализаторов, чувствительных к шуму.

Ненормированный показатель шума на рабочих местах оказывает негативное воздействие на психологическое состояние сотрудника. У работника на поставленной ему задаче понижаются концентрация и сосредоточенность, а увеличивается утомляемость и стресс. Повышенный уровень шума приводит к нарушению слуха или являться помехой для коммуникаций между сотрудниками. Измерение уровня звука и уровней звукового давления производится на расстоянии 50 см от поверхности оборудования и на уровне расположения источника(ков) звука. Уровень шума исправного компьютера находится в пределах 35-50 дБА, что значительно ниже, чем допустимый уровень шума для данного рабочего места, определённый СанПиН 1.2.3685-21 [12].

6.3.1.4 Электромагнитные поля

Источниками электромагнитного излучения на данном рабочем месте выступают системные блоки и мониторы включённых компьютеров. Требования при работе с источниками электромагнитных излучений определяются ГОСТ 12.1.006-84 ССБТ. Допустимым считается 8-часовой рабочий день для сотрудника на своем рабочем месте, с предельно допустимым уровнем напряженности электрического поля не более 8 кА/м, уровнем магнитной индукции – 10 мТл. Соблюдение данных норм дает возможность избежать негативного воздействия электромагнитных излучений [19].

Для защиты операторов ПЭВМ от негативного воздействия электромагнитных полей в первую очередь необходимо, чтобы используемая техника удовлетворяла нормам и правилам сертификации. Кроме этого, для

уменьшения уровня электромагнитного поля от персонального компьютера рекомендуется подключать к одной розетке не более двух компьютеров, сделать защитное заземление, подключать компьютер к розетке через нейтрализатор электрического поля, использовать мониторы, уровень излучения которых понижен, установить защитные экраны и соблюдать режимы труда и отдыха. К средствам индивидуальной защиты при работе на компьютере относят спектральные компьютерные очки для улучшения качества изображения и защиты от избыточных энергетических потоков видимого света и для профилактики. Очки уменьшают утомляемость глаз на 25-30%. Их рекомендуется использовать всем операторам при работе больше 2 часов в день, а при нарушении зрения на 2 диоптрии и более – независимо от продолжительности работы [21].

6.3.2 Опасные производственные факторы

6.3.2.1 Опасность поражения электрическим током

Электробезопасность – система организационных и технических мероприятий и средств, обеспечивающих защиту людей от вредного и опасного воздействия электрического тока, электрической дуги, электромагнитного поля и статического электричества. Нормы электробезопасности на рабочем месте и вопросы требований к защите от поражения электрическим током освещены ГОСТ 12.1.038-82. и ГОСТ Р 12.1.019-2017 ССБТ [20, 21].

Помещение, где расположено рабочее место оператора ПЭВМ, относится к помещениям без повышенной опасности ввиду отсутствия следующих факторов: сырость, токопроводящая пыль, токопроводящие полы, высокая температура, возможность одновременного прикосновения человека к имеющим соединение с землей металлоконструкциям зданий, технологическим аппаратам, механизмам и металлическим корпусам электрооборудования. С целью защиты от поражения электрическим током, возникающим между корпусом приборов и инструментом при пробое

сетевого напряжения на корпус, корпуса приборов и инструментов должны быть заземлены [20, 21].

Для оператора ПЭВМ при работе с электрическим оборудованием обязательны следующие меры предосторожности: перед началом работы нужно убедиться, что выключатели и розетка закреплены и не имеют оголённых токоведущих частей. Все работы по устранению неисправностей должен производить квалифицированный персонал; При включенном сетевом напряжении работы на задней панели должны быть запрещены. При производстве монтажных работ необходимо использовать только исправный инструмент, аттестованный службой КИПиА [20, 21].

6.4 Экологическая безопасность

В данном разделе рассматривается влияние на окружающую среду деятельности по разработке проекта, а также самого продукта в результате его реализации на производстве. Нормативы экологической безопасности установлены ГОСТ 17.4.3.04-85, ГОСТ Р 53692-2009 [22,23].

Непосредственно программный продукт, разработанный в ходе выполнения магистерской диссертации, не наносит вреда окружающей среде ни на стадиях его разработки, ни на стадиях эксплуатации. В лаборатории не ведется никакого производства, однако, средства, необходимые для его разработки и эксплуатации могут наносить вред окружающей среде.

К отходам, производимым в помещении, можно отнести, в первую очередь, это бумажные отходы – макулатура, пластиковые отходы, неисправные детали персональных компьютеров и других видов ЭВМ. Бумажные отходы рекомендуется накапливать и передавать их в пункты приема макулатуры для дальнейшей переработки. Пластиковые бутылки складывать в специально предназначенные контейнеры [22, 23].

Современные ПЭВМ производят практически без использования вредных веществ, опасных для человека и окружающей среды. Исключением являются аккумуляторные батареи компьютеров и мобильных устройств. В

аккумуляторах содержатся тяжелые металлы, кислоты и щелочи, которые могут наносить ущерб окружающей среде, попадая в гидросферу и литосферу, еслиони были неправильно утилизированы.

Для утилизации аккумуляторов необходимо обращаться в специальные организации, специализировано занимающиеся приемом, утилизацией и переработкой аккумуляторных батарей. Люминесцентные лампы, применяющиеся для искусственного освещения рабочих мест, также требуют особой утилизации, т.к. в них присутствует от 10 до 70 мг ртути, которая относится к чрезвычайно -опасным химическим веществам и может стать причиной отравления живых существ, а также загрязнения атмосферы, гидросферы и литосферы. Сроки службы таких лампсоставляют около 5-ти лет, после чего их необходимо сдавать на переработку в специальных пунктах приема. Юридические лица обязаны сдавать лампы напереработку и вести паспорт для данного вида отходов [22, 23].

6.5 Безопасность в чрезвычайных ситуациях

В рабочей среде оператора ПЭВМ возможно возникновение следующих чрезвычайных ситуаций техногенного характера: пожары и взрывы в зданиях и на коммуникациях; внезапное обрушение зданий.

Среди возможных стихийных бедствий можно выделить метеорологические (ураганы, ливни, заморозки), гидрологические (наводнения,паводки, подтопления), природные пожары.

Регулирование пожаробезопасности производится ГОСТ 12.1.004-91. Наиболее характерной для объекта, где размещаются рабочие помещения, оборудованные ПЭВМ, чрезвычайной ситуацией является пожар. Помещение для работы операторов ПЭВМ по системе классификации категорий помещений по взрывопожарной и пожарной опасности относится к категории Д. В помещениях с ПЭВМ повышен риск возникновения пожара из-за присутствия множества факторов: наличие большого количества электронных схем, устройств электропитания, устройств кондиционирования воздуха [24].

Все сотрудники организации обязаны быть ознакомлены с инструкцией по пожарной безопасности, пройти инструктаж по технике безопасности и строго соблюдать его. Запрещается использовать электроприборы в условиях, не соответствующих требованиям инструкций изготовителей, или имеющие различного рода неисправности, которые в соответствии с инструкцией по эксплуатации могут привести к пожару, а также использовать электропровода икабели с поврежденной или потерявшей защитные свойства изоляцией [24].

Перед уходом из служебного помещения требуется провести его осмотр, закрыть окна, убедиться в том, что в помещении отсутствуют источники возможного возгорания, все электроприборы отключены и выключено освещение [24].

С периодичностью не менее одного раза в три года необходимо проводить замеры сопротивления изоляции токоведущих частей силового и осветительного оборудования. Увеличение устойчивости достигается за счет проведения соответствующих организационно-технических мероприятий, подготовки персонала к работе в ЧС [24].

При обнаружении пожара или признаков горения требуется [24]:

- Прекратить работу;
- Вызвать пожарную охрану;
- Сообщить непосредственному или вышестоящему начальнику;
- Отключить от сети электрооборудование;
- По возможности , принять меры по эвакуации людей и материальных ценностей;
- При общем сигнале опасности покинуть здание согласно «Плануэвакуации людей при пожаре и других ЧС» (рис 5.2);

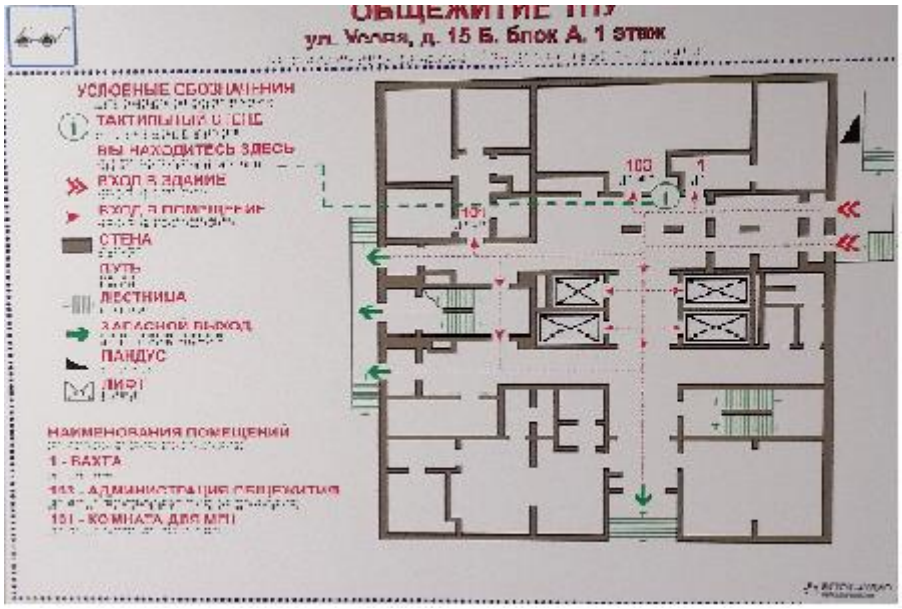

Рисунок 5.2 План эвакуации

Для тушения пожара следует применять ручные углекислотные огнетушители (типа ОУ-2, ОУ-5), находящиеся в помещениях офиса, и пожарный кран внутреннего противопожарного водопровода. Они предназначены для тушения начальных возгораний различных веществ и материалов, за исключением веществ, горение которых происходит без доступа воздуха. Огнетушители должны постоянно содержаться в исправном состоянии и быть готовыми к действию. Категорически запрещается тушить возгорания в помещениях офиса при помощи химических пенных огнетушителей [24].

Выводы по разделу

В данном разделе были рассмотрены основные вопросы соблюдения прав работника на труд, выполнения правил к безопасности труда, промышленной безопасности, экологии и ресурсосбережения.

Установлено, что рабочее место исследователя удовлетворяет требованиям безопасности и гигиены труда во время реализации проекта, а вредное воздействие объекта исследования на окружающую среду не превышает норму.

Заключение

Целью данной выпускной квалификационной работы является разработка метода и исследование алгоритма кластеризации климатических данных.

Для этого были решены следующие задачи:

– Проведен анализ существующих и актуальных на данный момент методов кластеризации. Выявлены их преимущества и недостатки;

– Было предложено использовать среднегодовое значение температур для кластеризации временных рядов;

– На основании выявленных закономерностей была подтверждена гипотеза о пригодности среднегодового и среднего значения температуры в целом (за 62 года) как метрики для кластеризации временных рядов;

– При анализе данных температурных рядов были выявлены паттерны поведения рядов, отличающихся синхронностью и как правило незначительной разницей в значениях температуры, были выявлены узловые точки, демонстрирующие сходство между климатом станций;

– С применением подготовленных данных была осуществлена кластеризация методом k-средних, взятым в качестве эталонного и выделены климатические классы;

– Реализован нейросетевой алгоритм кластеризации, опирающийся в своей архитектуре на сеть Кохонена;

– С помощью реализованного алгоритма был проведен эксперимент. Была проведена кластеризация климатических данных при установке различных параметров. Были получены уникальные климатические классы;

– Были выявлены различия результатов кластеризации с методом k-средних, взятым в качестве эталонного алгоритма;

– В отличие от k-средних у нейросетевого алгоритма в зависимости от выбранной метрики результаты могут быть различными, что говорит о чувствительности алгоритма к входным данным;

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Воронцов К.В. Алгоритмы кластеризации и многомерного шкалирования. Курс лекций. МГУ, 2007. – Режим доступа: http://www.ccas.ru/voron/download/Clustering.pdf -дата доступа: 19.05.2022.
2. Обучение без учителя. [Электронный ресурс] Режим доступа: https://wiki.loginom.ru/articles/unsupervised-learning.html - Дата доступа 19.02.2022
3. Сеть Кохонена (Kohonen Network). [Электронный ресурс] Режим доступа: https://wiki.loginom.ru/articles/kohonen-network.html - Дата доступа: 20.02.2022
4. Нейронные сети Кохонена. [Электронный ресурс] Режим доступа: https://neuronus.com/theory/nn/955-nejronnye-seti-kokhonena.html - Дата доступа: 20.02.2022
5. Конкуретное обучение (Competitive Learning). [Электронный ресурс] Режим доступа: https://wiki.loginom.ru/articles/competitive-learning.html - Дата доступа: 20.02.2022
6. NumPy [Электронный ресурс] Режим доступа:

https://ru.wikipedia.org/wiki/NumPy - Дата доступа 12.05.2022

7. Pandas [Электронный ресурс] Режим доступа: https://ru.wikipedia.org/wiki/Pandas - Дата доступа 12.05.2022
8. Matplotlib [Электронный ресурс] Режим доступа: https://ru.wikipedia.org/wiki/Matplotlib - Дата доступа 12.05.2022
9. Метод k-средних (K-Means) [Электронный ресурс] Режим доступа: https://www.helenkapatsa.ru/mietod-k-sriednikh/ - Дата доступа 14.05.2022
10. Карта Кохонена [Электронный ресурс] Режим доступа: https://basegroup.ru/deductor/function/algorithm/kohonen - Дата доступа 14.05.2022
11. Трудовой кодекс Российской Федерации от 30.12.2001 N 197-ФЗ
12. СанПиН 1.2.3685-21 "Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания"
13. ГОСТ 22269-76 «Рабочее место оператора. Взаимное расположение элементов рабочего места»
14. ГОСТ Р 50923-96 «Рабочее место оператора. Общие эргономические требования и требования к производственной среде. Методы измерения. Дисплеи»
15. ГОСТ 12.2.032-78 «Рабочее место при выполнении работ сидя»
16. СП. 51.13330.2011 «Свод правил защита от шума»
17. СП 52.13330.2016 «Свод правил естественное и искусственное

освещение»
18. СНиП 23-05-95 «Строительные нормы и правила российской федерации естественное и искусственное освещение»
19. ГОСТ 12.1.006-84 ССБТ «Электромагнитные поля радиочастот допустимые уровни на рабочих местах и требования к проведению контроля»
20. ГОСТ 12.1.038-82 Система стандартов безопасности труда (ССБТ). Электробезопасность. Предельно допустимые значения напряжений прикосновения и токов.
21. ГОСТ Р 12.1.019-2017 ССБТ «Система стандартов безопасности труда. Электробезопасность. Общие требования и номенклатура видов защиты»
22. ГОСТ 17.4.3.04-85 «Общие требования к контролю и охране от загрязнения»
23. ГОСТ Р 53692-2009. «Ресурсосбережение. обращение с отходами. этапы технологического цикла отходов»
24. ГОСТ 12.1.004-91. «Система стандартов безопасности труда. пожарная безопасность»

ПРИЛОЖЕНИЕ А

Development of climate data clustering algorithm

Студент:

Группа	ФИО	Подпись	Дата
8ВМ03	Кавешников Артем Владимирович		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Иванова Юлия Александровна	к.т.н		

Консультант – лингвист отделения иностранных языков, школы ИШИТР:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Старший преподаватель	Ануфриева Т.Н.			

Introduction

One of the most significant and large-scale problems of our time at the moment can be called a continuously growing amount of information that requires a certain systematization, simplification and isolation of its essential part. With the development of technical means and Internet technologies, the volume of digital data is growing on a huge scale and amounts to terabytes. Manual processing of such data is time-consuming, and existing methods may be inefficient. Therefore, to solve problems of this kind, more and more new methods of data processing are required. Modern methods should carry out the analysis, systematization and collection of the information received with a sufficiently high accuracy.

Methods that allow you to analyze large amounts of data have a wide range of applications. So, in medicine, based on the totality of cluster symptoms, it is possible to establish a diagnosis with a fairly high accuracy and prescribe subsequent treatment; in economics, a set of cluster parameters can be used to identify groups of consumers, their behavior and their consumer basket; In meteorology, cluster analysis makes it possible to identify climatic zones and predict their change. With the help of clustering algorithms, it is possible to implement the problem of pattern recognition, and there is also a rather high need for processing large amounts of data in scientific research. Based on the foregoing, we can conclude that the demand for clustering algorithms and their research is quite high.

Cluster analysis

The task of clustering (unsupervised) can be formulated as follows. There is a training set of objects - sample $X\ell = \{x1, \ldots, xl\} \subset X$ and the distance function between them $\rho(x, x')$. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

In general, the clustering algorithm can be characterized as a function that brings all objects of the sample X in accordance with some label of the cluster Y. $X \rightarrow Y$

Purposes of clustering:

• Splitting a set of objects according to some attribute and simplifying further data processing

• Revealing the structure of objects

• Data reduction;

• Selection of atypical objects that do not fit into any of the clusters

Graph clustering methods

A wide range of clustering methods, in which the sample is represented as a graph. The vertices of the graph are the sample objects. Edges are pairs of objects with distance $\rho(i,j) = \rho(xi , xj)$. Graph algorithms are relatively easy to implement, visual and easy enough to upgrade.

Strongly connected components. In this algorithm, the input parameter R is specified and all edges (i, j) for which the distance $\rho(i,j) > R$. are removed from the graph. Only the closest pairs of objects remain connected. The main concept of the algorithm is to select the R value at which the graph is divided into several connected components. The found connected components are the desired clusters.

A connected component of a graph is a subset of its vertices in which any two vertices can be connected by a path that lies entirely in this subset. To select the optimal value of the parameter R, as a rule, a histogram of the distribution of pairwise distances $\rho(i,j)$ is constructed [1]

Shortest path algorithm

The essence of the algorithm is as follows: a graph is built from n-1 edges, so that all n points are connected and the length between them is minimized. A graph of this type is called a shortest non-closed path or skeleton.

The algorithm consists in finding a pair of points with the smallest distance and connecting them with an edge.

After that, it checks for isolated nodes. As long as the condition is met:

a search is made for an isolated point that is closest to some non-isolated one.

Then these two points are connected.

After that, the p-1 longest edges are removed.

FOREL algorithm. Let some point x0 ∈ X and a parameter R. All sample points xi ∈ Xl falling inside the sphere ρ(xi , x0) are selected, and the point x0 is moved to the center of gravity of the selected points. This procedure is repeated until the set of selected points, and hence the location of the center, does not stop changing. It is proved that this procedure converges in a finite number of iterations. In this case, the sphere moves to the place of local concentration of points. The sphere center x0 generally does not belong to the sample, which is why it is called a formal element. To calculate the center, it is necessary that the set of objects X be not only a metric, but also a linear vector space. This requirement is naturally satisfied when objects are described by numerical features. [1]

1.2.2Statistical algorithms

Statistical algorithms are based on the assumption that clusters can be described using a family of probability distributions. Then the clustering problem is reduced to separating the mixture of distributions over a finite sample.

The k-means method is a simplification of the EM algorithm. The main difference is that in the EM algorithm each object xi is distributed over all clusters with probabilities giy = P{yi = y}. In the k-means algorithm, each object is rigidly assigned to only one cluster. The second difference is that the shape of clusters in the k-means method is not customizable. [1]

Hierarchical clustering

Hierarchical clustering algorithms, also called taxonomy algorithms, build not one partition of the sample into non-overlapping classes, but a system of nested partitions. The result of taxonomy is usually presented in the form of a taxonomic tree - a dendrogram. There are two main types of hierarchical clustering algorithms:

Divisive or top-down algorithms break the sample from initially large clusters into smaller and smaller clusters.

Agglomerative or bottom-up algorithms are more common, in which objects are combined from initially small clusters into larger and larger clusters.

Unsupervised learning

Unsupervised learning is a family of machine learning algorithms that does not require an objective function to adjust the model weights. It is used when only descriptions of a set of objects are known and it is required to detect patterns that exist between objects.

Unsupervised learning algorithms do not calculate the error of the training sample and do not use the error backpropagation method. Instead, data about the existing state of the system and examples of the training set are used.

The main purpose of unsupervised neural networks is the implementation of clustering tasks.

Kohonen network

One of the most popular neural network architectures that use unsupervised learning is the Kohonen neural network used for clustering, which is based on competitive learning, in which the neuron weights are corrected by calculating the distance between the output layer vectors and input feature vectors. [2]

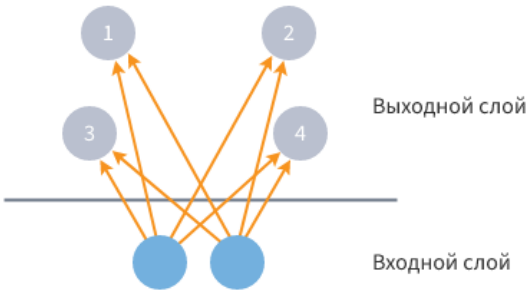

Fig 1. The structure of the Kohonen neural network

The number of output neurons of the Kohonen network is equal to the number of clusters to be built by the model, and each neuron is associated with a specific cluster. Outputs are processed on a “winner takes all” basis, i.e. the neuron with the highest output value produces one, and the outputs of the rest are set to zero.

The training of the Kohonen network, like a conventional neural network, consists in adjusting the weights of connections between neurons, but is performed using competitive learning technology. [3]

Competitive learning

At the stage of competition, the input vector of features is fed to the network input and a search is made for a neuron with the closest set of weights to it. Such a neuron is declared the winner. In the process of unification, a group (neighborhood) of neurons is formed around it, which will participate in the learning process. The group size is determined by the learning radius. [5]

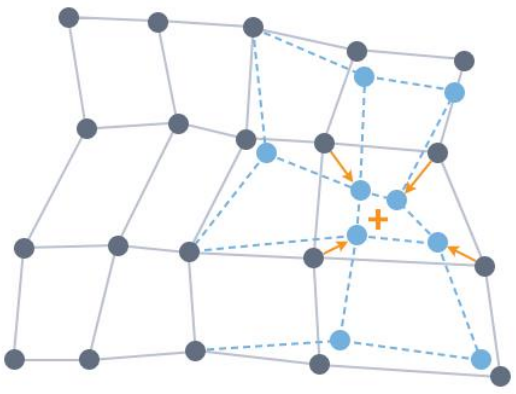

Figure 2. The process of changing the structure of the Kohonen map

Finally, at the adjustment stage, the weights of neurons located within the learning radius of the winning neuron are adjusted so that their vectors become closer to it. [5]

Summing up the review of clustering algorithms, we present a comparative analysis of the algorithms in the form of a table below.

Table 1. Comparative table of clustering algorithms

	Iterations	Calculation of the full distance matrix	Storing the full distance matrix	Setting the number of classes	Dependence on initial conditions	Learning	Class form
Graph methods	-	+	+	-	-	-	any
Forel	+	-	-	+	+	-	hypersphere
K-means	+	+	+	+	+	-	hypersphere
Hierarchical methods	-	+	+	-	+	-	any
Neural networks (supervised)	+	+	+	+	+	+	any
Neural networks (unsupervised)	+	+	+	-	+	+	any

Clustering algorithm for climate data

Description

The algorithm was developed for the analysis of climate data and was used to work with temperature data on the example of data characterizing changes in the average monthly temperature.

The algorithm uses a neural network architecture. Since we are working with a clustering problem, we are dealing with unsupervised learning. One such tool is the Kohonen network. It is a two-layer network where each neuron in the input layer is connected to each neuron in the output layer. Neurons of the second - output layer are often called cluster elements. The number of these neurons determines the maximum possible number of groups into which the data will be divided.

The system operates on the principle of competitive learning. The neurons of the output layer compete with each other for the right to best match the input feature vector. The victory goes to the neuron whose weight vector is closest to the vector of input features. Thus, the clustering procedure consists in assigning each input feature vector to a certain cluster.

Neural network training is carried out using competitive training. At each iteration of the algorithm, one vector is randomly taken from the input layer of the neural network. After that, the neuron of the output layer is searched for, the distance between its set of weights and the set of weights of the input vector is minimal. The weights for the winning neuron are adjusted according to some chosen rule. Learning takes place at a certain predetermined rate, set by the parameter $\Delta\lambda$. In the learning process, the search for the nearest vectors to the "winner" vector is carried out, then the weights are adjusted sequentially according to the formula: wm[i] = wm[i] + la * (x[i] - wm[i])

The temperature and location of the station were used as signs of clustering. Clustering is based on the principle of synchronous behavior of temperature time series. The number of clusters is used as criteria for assigning temperature data to a cluster.

The clustering procedure consists in calculating the optimal output layer vector that is closest to the input feature vector. The procedure is an iterative process.

The resulting clusters include climatic stations for which the level distance between the input vector and the winner neuron vector is minimal.

Development of a clustering method with the introduction of a metric of average annual temperatures

In the course of research and work with the current algorithm, the task was to improve the quality of the ongoing clustering. The initial data is represented by a set of average monthly temperatures. A hypothesis was put forward that the year is the optimal period characterizing the climate group. The mean annual temperature was introduced as a clustering metric.

2. MATERIALS AND METHODS

2.1 Technologies used

2.1.1 NumPy

NumPy is an open source Python library that implements basic mathematical operations used to work with arrays and matrices. [6]

2.1.2 Pandas

Pandas is an open source Python library used for data analysis and manipulation. The library provides convenient tools for opening and working with data tables and time series [8]

2.1.3 Matplotlib

Matplotlib is an open source Python library used to visualize data in 2D and 3D plots. The resulting graphical information is often used by analysts in data presentations and finds its way into scientific publications. [8]

2.2 Used clustering algorithms.

In our work, two algorithms were used and implemented: k-means and a neural network algorithm based on the Kohonen self-organizing map principle.

2.2.1 K-Means

The k-means method was chosen by us for implementation as the most well-

known and simple algorithm that provides a qualitative solution to the clustering problem, however, it has a significant drawback, namely the need to know in advance the number of clusters k. The choice fell not on neural network methods, due to their vastness, customization flexibility, and modernization opportunities. Since we are solving the clustering problem, we are interested in neural networks with unsupervised learning. One such network is the Kohonen network. **[9]**

2.2.2 Kohonen Neural Network

The Kohonen Neural Network is an unsupervised machine learning algorithm that allows you to display the results in the form of compact and easy to interpret two-dimensional maps.

2.3 Data sets used.

The input data is represented by an array of average monthly temperatures with dimensions (744 x 928) (number of observations for 62 years x 928 climatic stations). Thus, the input data form a time series.

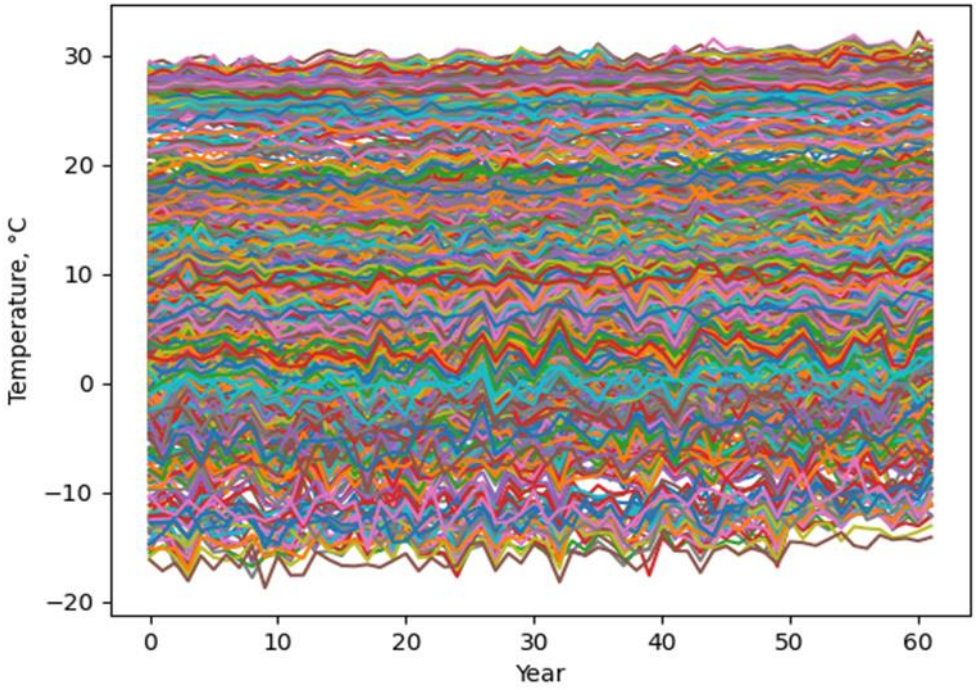

Figure 3. Temperature time series of 928 stations over 62 years

3. Results

3.1 Clustering parameters.

3.2 Clustering based on k-means

In this section, we perform k-means clustering of climate data. The climate is characterized by a large number of parameters such as: atmospheric pressure, humidity, wind speed and direction, temperature, cloudiness, angle of incidence of sunlight, continentality. We build our experiment on the assumption that all climatic and ecological indicators of the zone are in a certain correlation and that temperature can be considered a generalizing and resulting characteristic. In our study, we settled on temperature as one of the most significant and resulting climate indicators.

In working with our data, the behavior of time series also matters. Based on these characteristics, we can visually evaluate the division of the time series.

Considering that most of our time series are quite similar in their behavior over time, we can confidently apply the Euclidean distance metric in the k-means algorithm.

In our case, the measure of similarity between clusters is the average temperature. This is what we will choose as the metric.

In the course of the study, it was proposed to analyze the operation of the algorithm on different time scales. The minimum temperature reading is one month (average monthly temperature). Thus, 3 intervals were chosen: 1 month, 1 year, the entire period (62 years). Three metrics were chosen by the corresponding method.

During the study, the operation of the algorithm was tested for a different number of clusters k. Various parameter values were chosen: 4, 8, 12, 16.

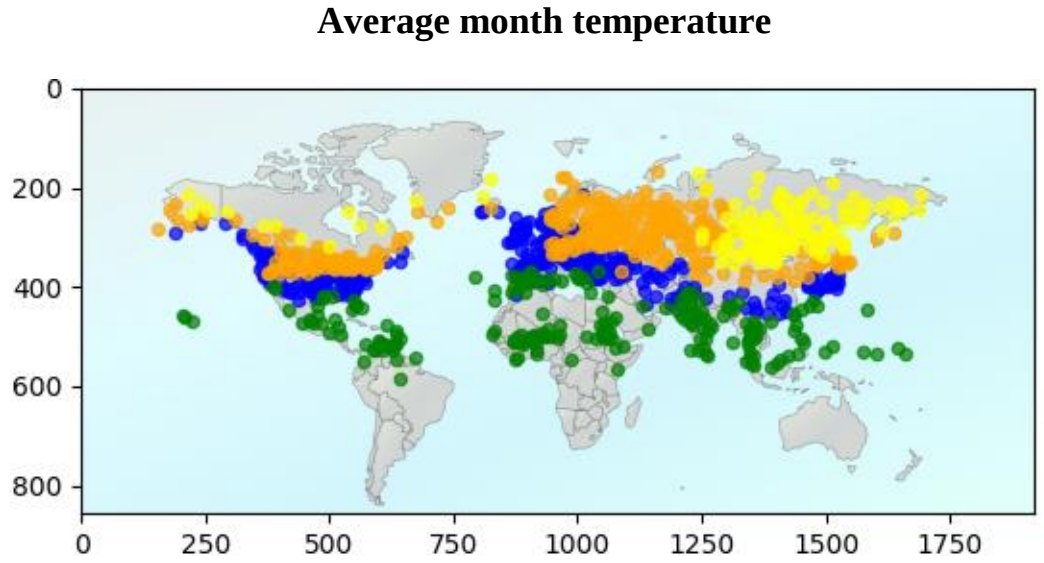

Figure 4. Distribution of 4 clusters by average monthly temperature

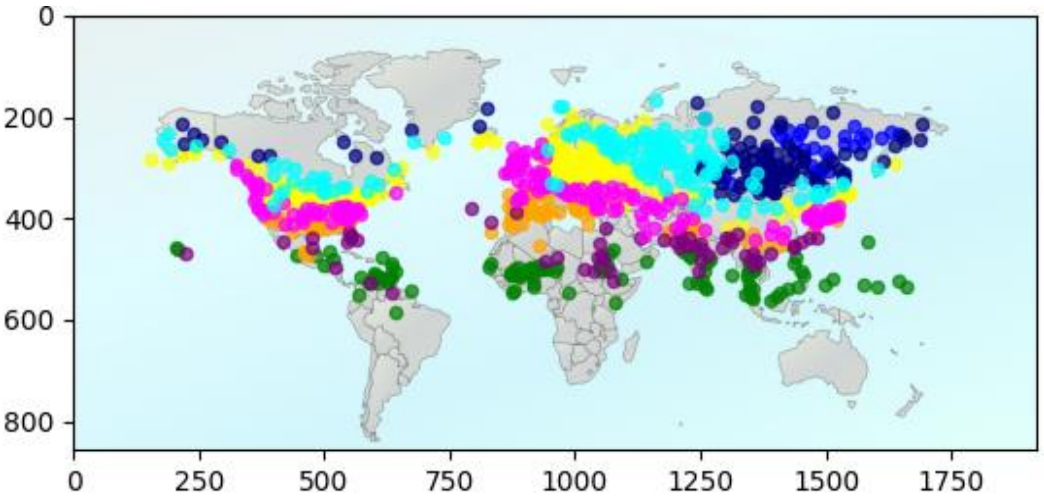

Figure 5. Distribution of 8 clusters by average monthly temperature

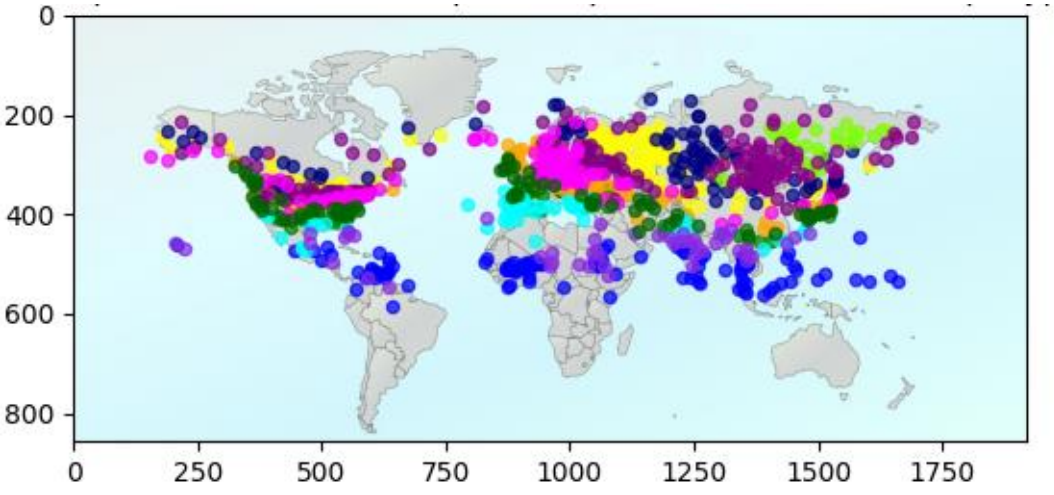

Figure 6. Distribution of 12 clusters by average monthly temperature

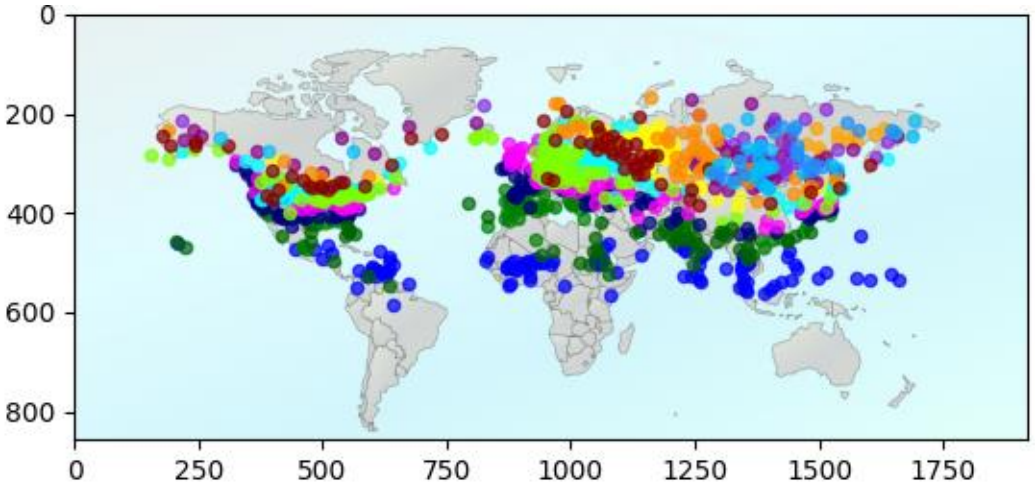

Figure 7. Distribution of 16 clusters by average monthly temperature

Conclusion: as can be seen from the images, the distribution of clusters obtained by the k-means algorithm based on monthly average data is relatively local in nature and characterizes local climatic zones.

Average annual temperature

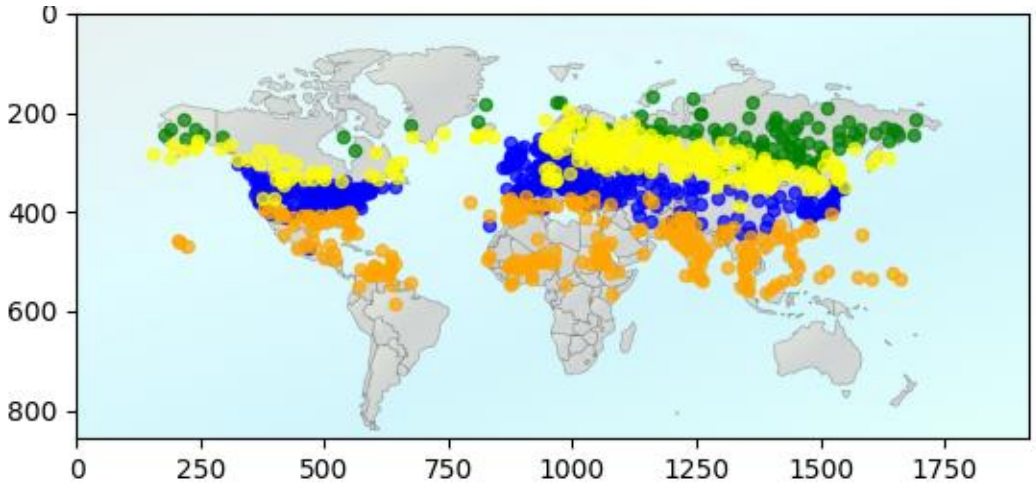

Figure 8. Distribution of 4 clusters by average annual temperature

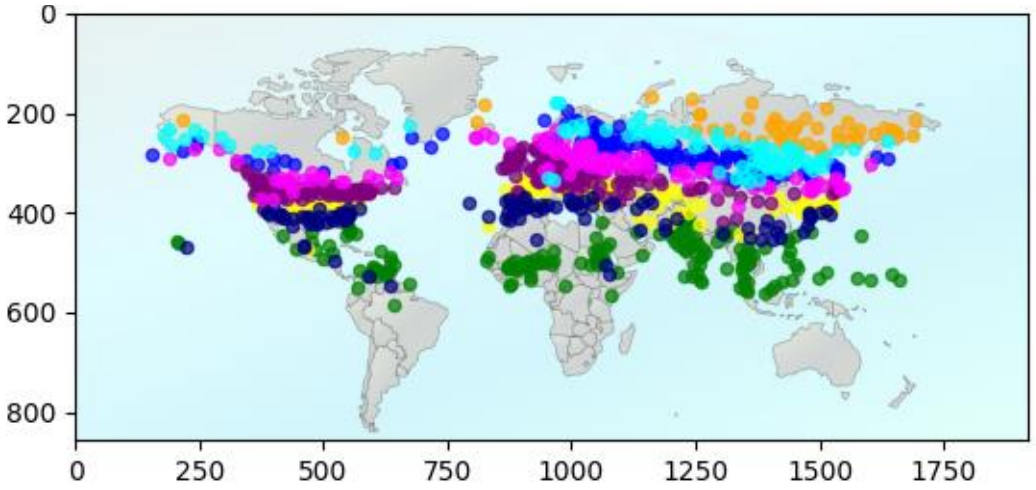

Figure 9. Distribution of 8 clusters by average annual temperature

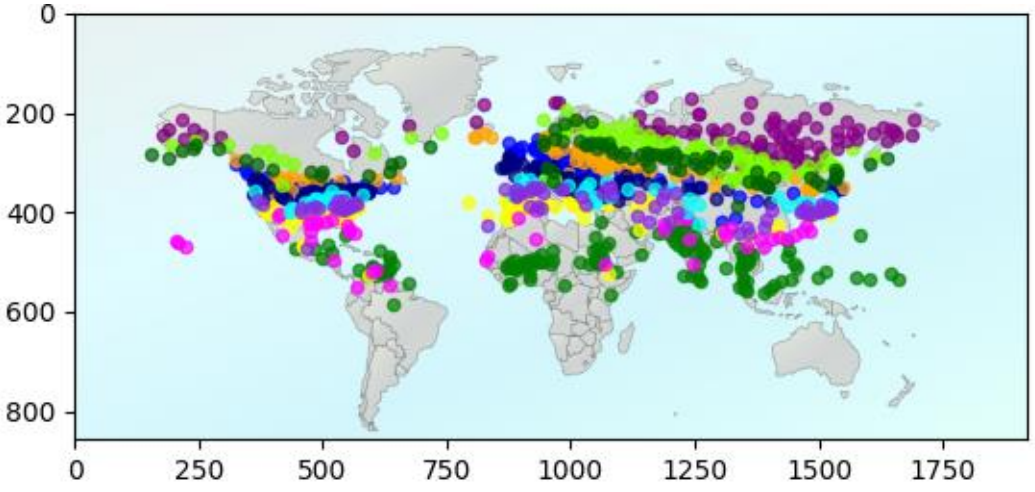

Figure 10. Distribution of 12 clusters by average annual temperature

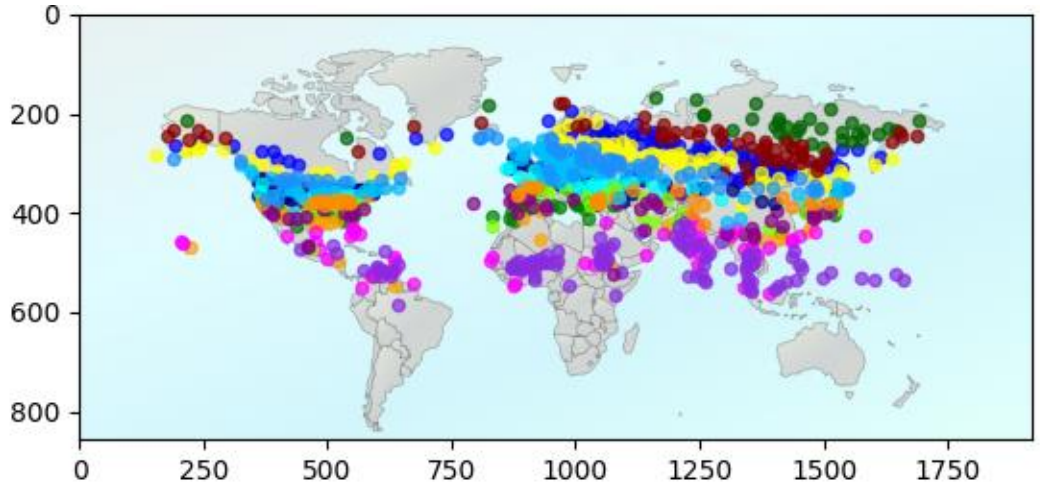

Figure 11. Distribution of 16 clusters by average annual temperature

Conclusion: the distribution of clusters obtained by the k-means algorithm from the data of average annual temperatures is predominantly latitudinal. And with a small number of clusters, it resembles the climatic zones of the Earth. With an increase in the number of clusters, the detailing of climatic zones increases.

Average temperature over the entire measurement period (62 years)

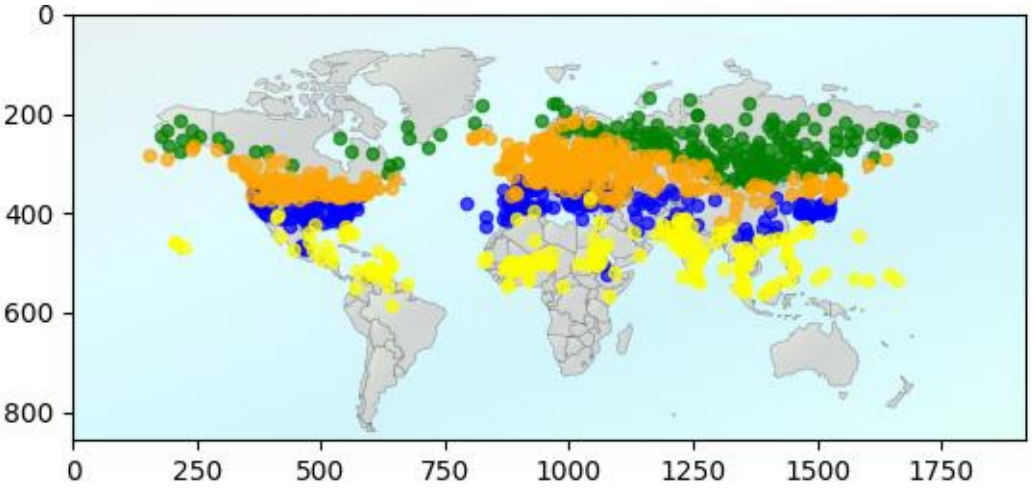

Figure 12. Distribution of 4 clusters by average temperature over 62 years

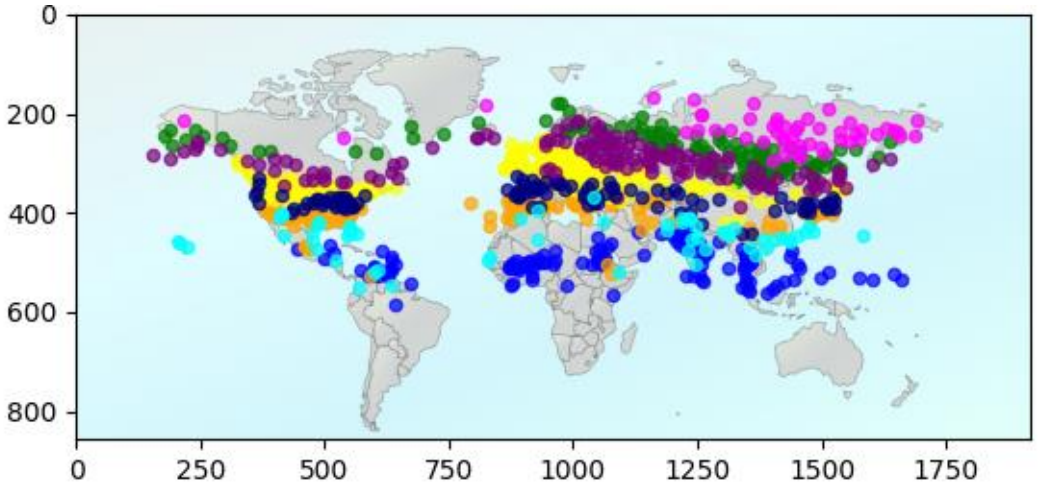

Figure 13. Distribution of 8 clusters by average temperature over 62 years

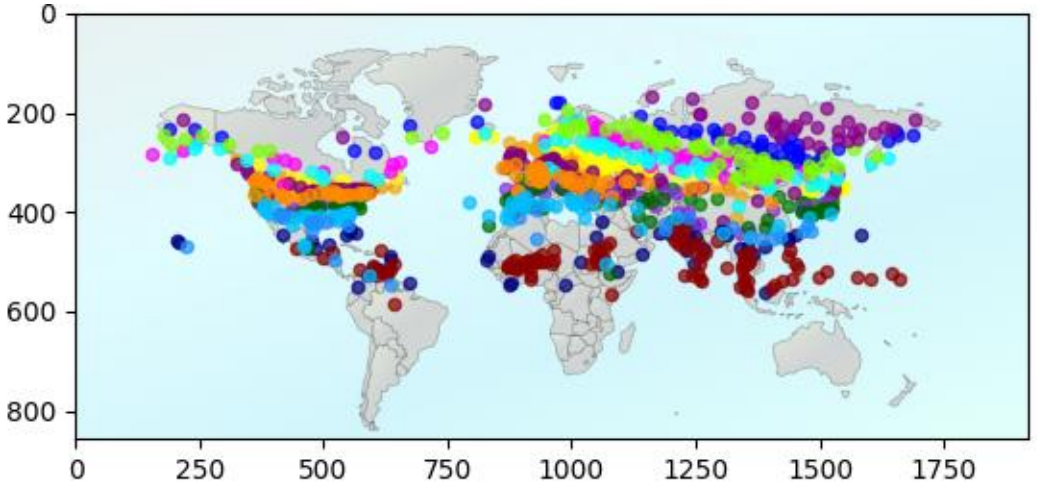

Figure 14. Distribution of 16 clusters by average temperature over 62 years

Conclusion: the use of clustering as a metric for the entire measurement period did not reveal significant changes with a metric with a period of average temperature of 1 year. Distributed classes have a similar breadth of distribution. It can be said that a period of 1 year is quite enough to identify informative climate classes.

Clustering based on neural network algorithm

Clustering based on the neural network algorithm occurred on similar data, but there were differences in the metric. In this algorithm, clustering occurs due to the coordinates and temperature of the weather station. The parameters for the number of clusters are similar and are: 4, 8, 12, 16

Average month temperature

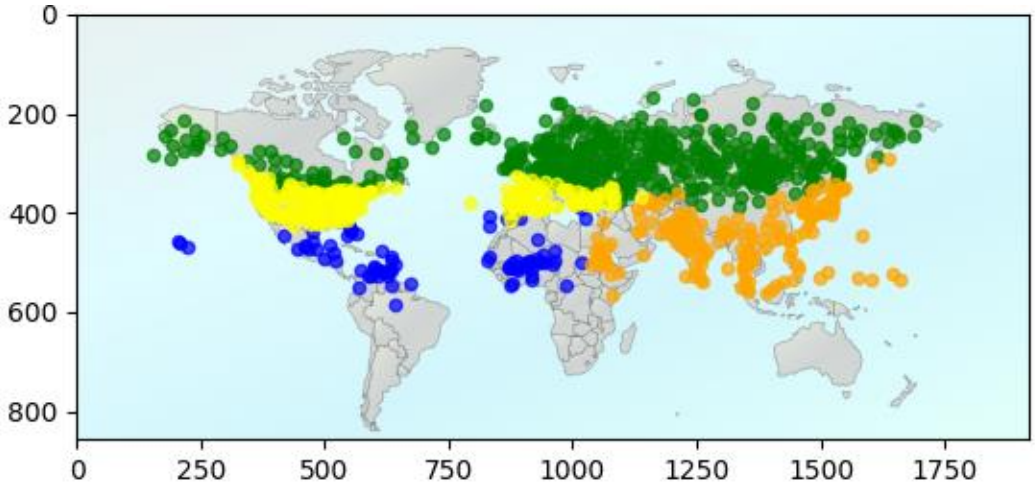

Figure 15. Distribution of 4 clusters by average monthly temperature

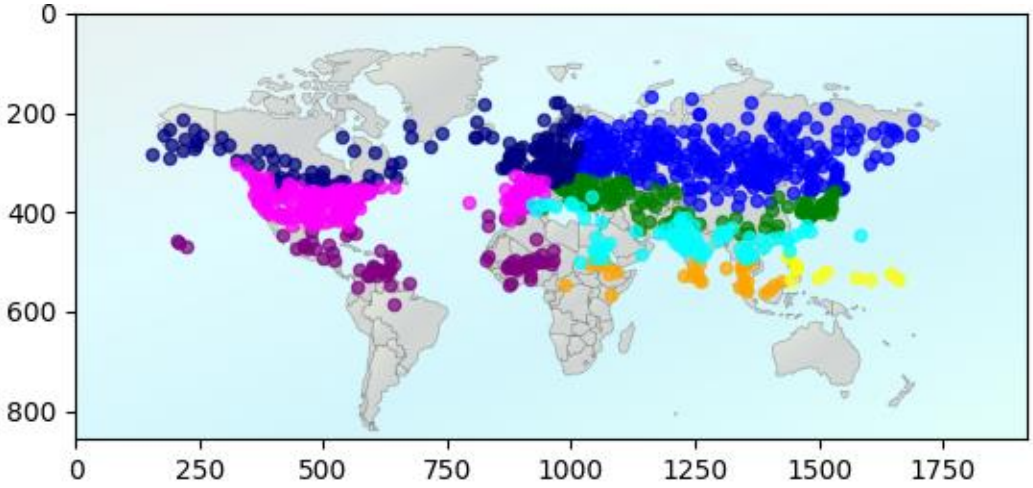

Figure 16. Distribution of 8 clusters by average monthly temperature

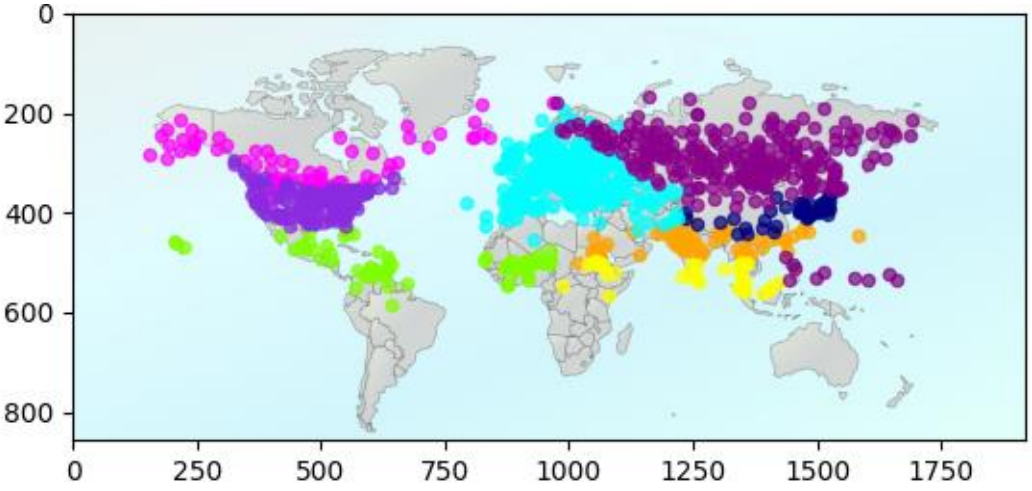

Figure 17. Distribution of 12 clusters by average monthly temperature

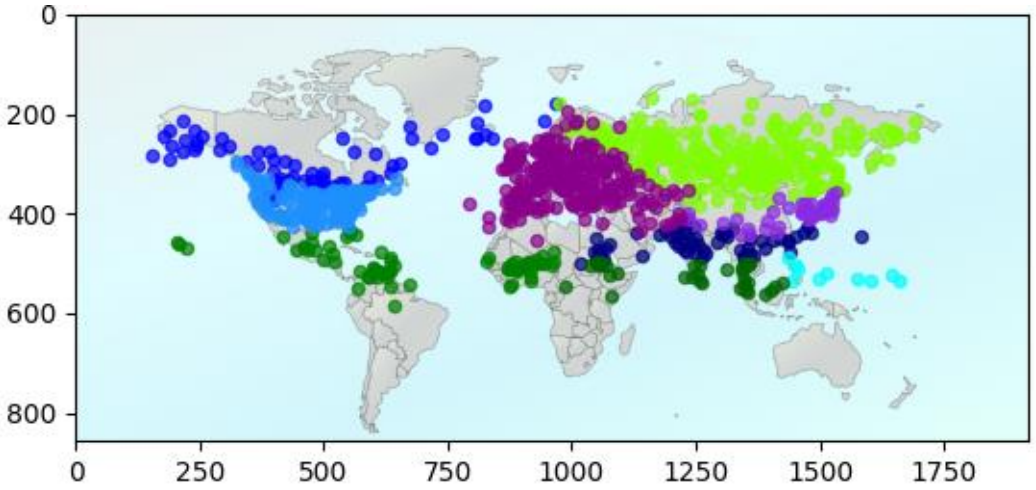

Figure 18. Distribution of 16 clusters by average monthly temperature

Conclusion: clustering using coordinates leads to the allocation of climatic and geographical classes.

Annual temperature

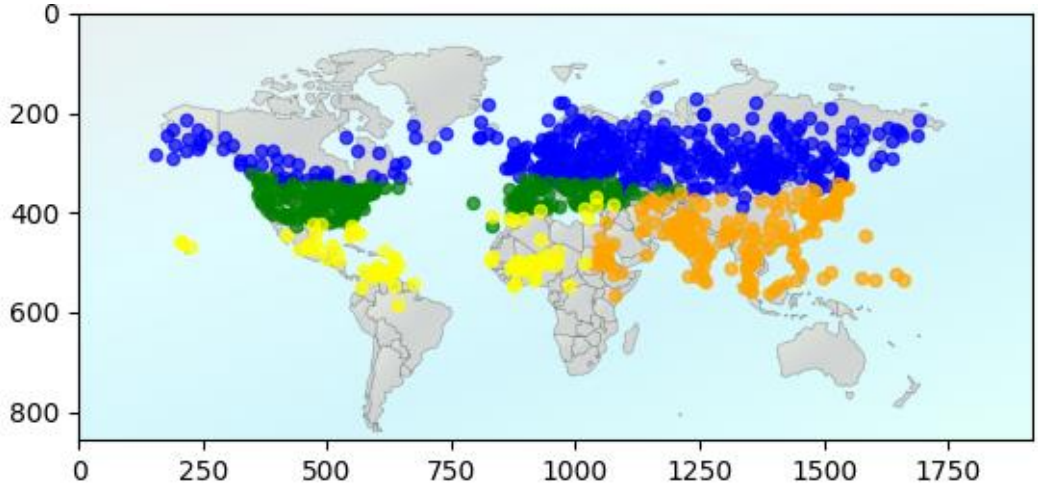

Figure 19. Distribution of 4 clusters by average annual temperature

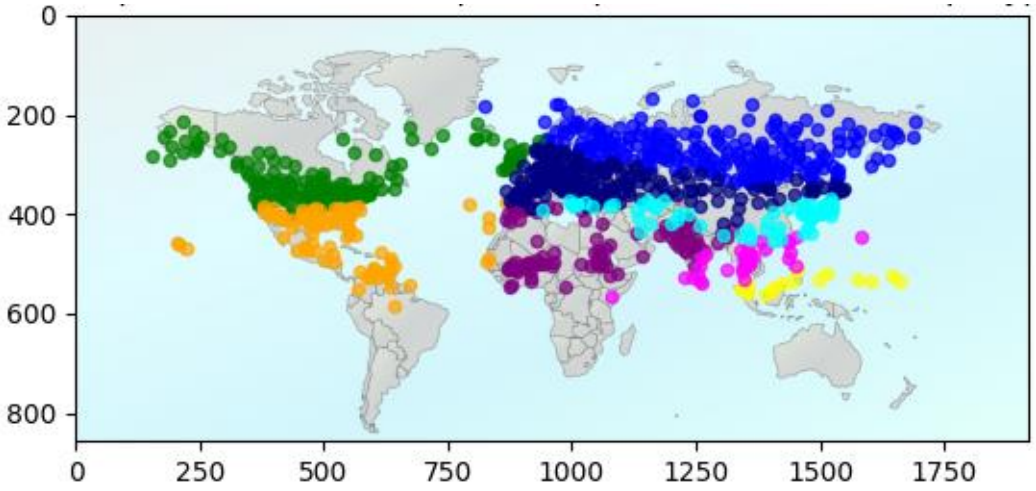

Figure 20. Distribution of 8 clusters by average annual temperature

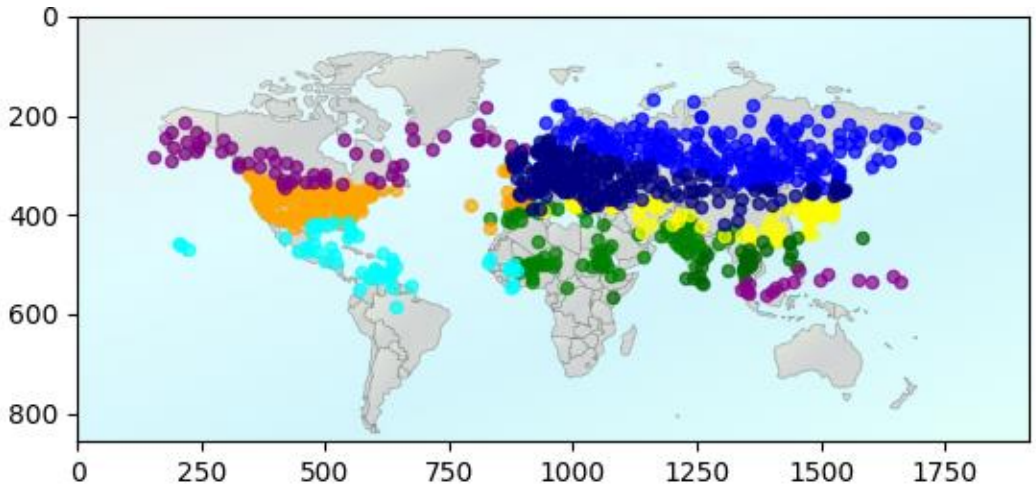

Figure 21. Distribution of 12 clusters by average annual temperature

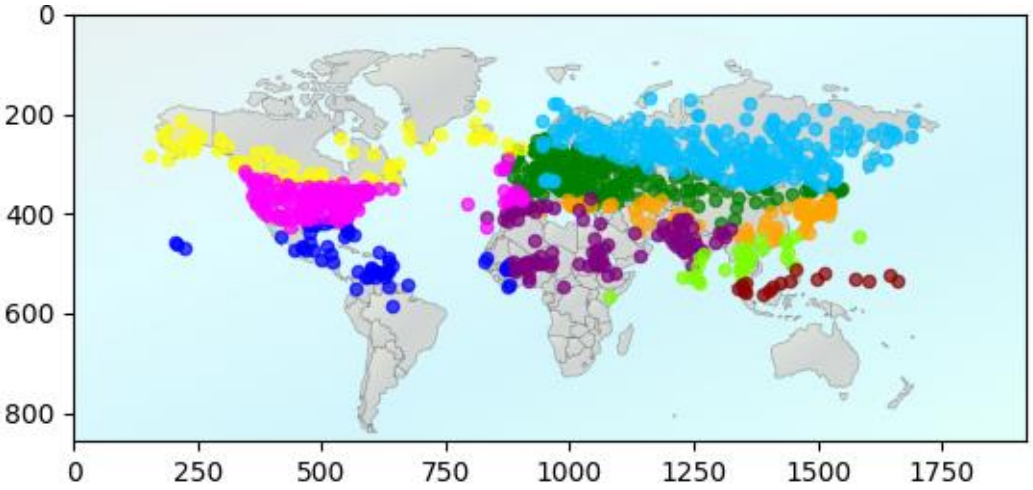

Figure 22. Distribution of 16 clusters by average annual temperature

Conclusion: the distribution of clusters is predominantly latitudinal, as for the k-means method, but more locally.

Average temperature over the entire measurement period (62 years)

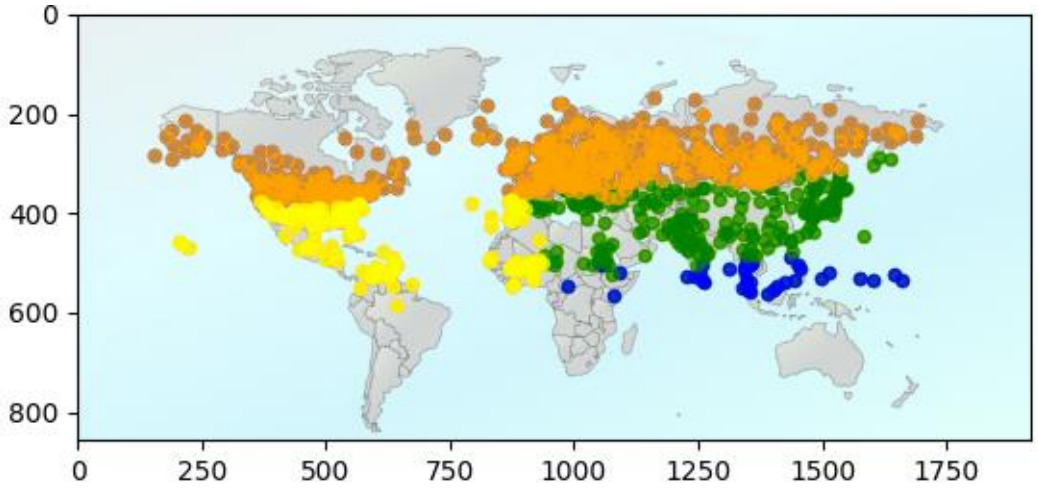

Figure 23. Distribution of 4 clusters by average temperature over 62 years

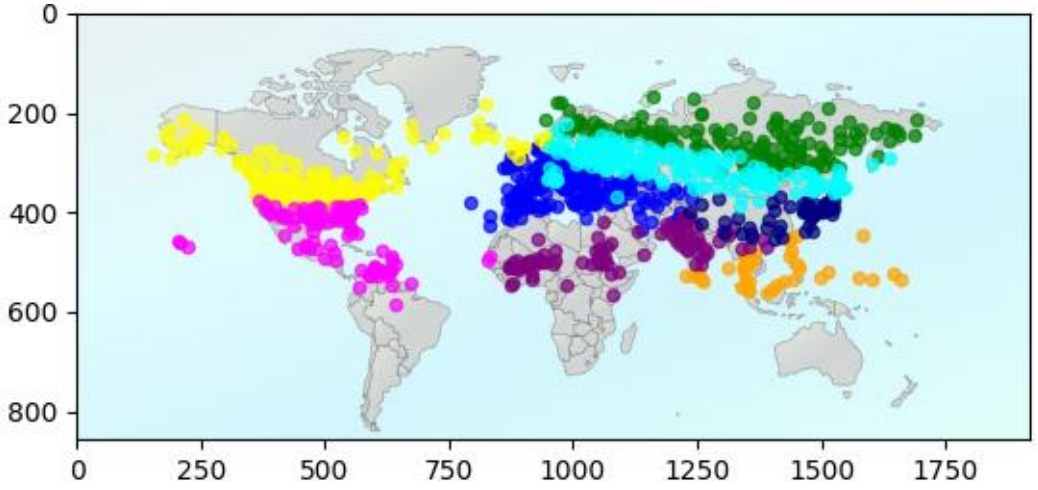

Figure 24. Distribution of 8 clusters by average temperature over 62 years

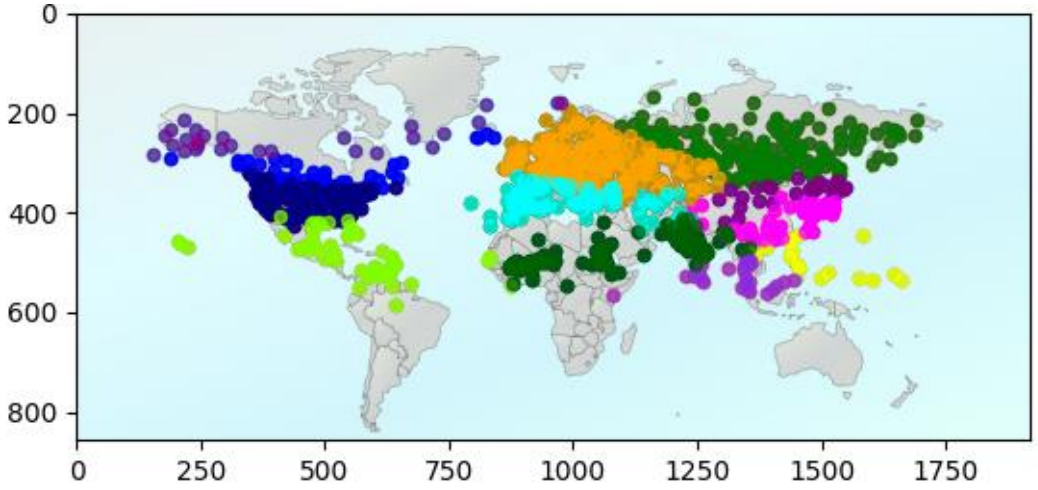

Figure 25. Distribution of 12 clusters by average temperature over 62 years

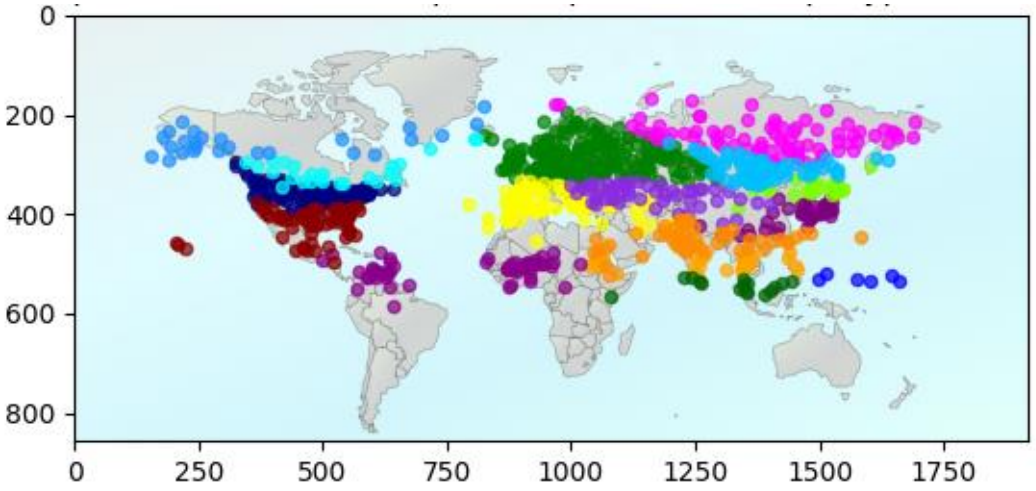

Figure 26. Distribution of 16 clusters by average temperature over 62 years

Conclusion: in comparison with the k-means method, in the neural network algorithm there are significant differences in the distribution of clusters between the average annual calculation and the calculation based on the average temperature for 62 years. This can be caused by the sensitivity of the algorithm to the input data, or by the dependence on the coordinates of the original centroids.

Conclusion

The purpose of this final qualification work is to develop a method and study the climate data clustering algorithm.

For this, the following tasks were solved:

– An analysis of existing and currently relevant clustering methods was carried out. Their advantages and disadvantages are revealed;

– An annual mean temperature metric was proposed for time series clustering;

– Based on the identified patterns, the hypothesis was confirmed about the suitability of the average annual and average temperature in general (over 62 years) as a metric for clustering time series;

– When analyzing the temperature series data, patterns of behavior of the series were identified that differ in synchronism and, as a rule, an insignificant difference in temperature values, nodal points were identified demonstrating the similarity between the climate of the stations;

- Using the prepared data, clustering was carried out using the k-means method, taken as a reference, and climate classes were identified;

– A neural network clustering algorithm has been implemented, based in its architecture on the Kohonen network;

– Using the implemented algorithm, an experiment was conducted. Clustering of climate data was carried out by setting various parameters. Unique climate classes were obtained;

– Differences in the results of clustering with the k-means method, taken as a reference algorithm, were identified;

– Unlike k-means, the results of the neural network algorithm can be different depending on the selected metric, which indicates the sensitivity of the algorithm to the input data;

– Visualization of clustering results for both algorithms was carried out;