

Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт космических и информационных технологий
Кафедра систем искусственного интеллекта

УТВЕРЖДАЮ
Заведующий кафедрой
_____ Г. М. Цибульский

подпись

« ____ » _____ 2018 г.

БАКАЛАВРСКАЯ РАБОТА

09.03.02 — «Информационные системы и технологии»
Разработка сервиса мониторинга неоднородной структуры
сельскохозяйственных земель

Руководитель _____ доцент, канд. техн. наук Р. В. Брежнев
подпись, дата

Выпускник _____ А. В. Непомнящих
подпись, дата

Красноярск 2018

Продолжение титульного листа бакалаврской работы по теме
«Разработка сервиса мониторинга неоднородной структуры
сельскохозяйственных земель»

Нормоконтролер

подпись, дата

Р. В. Брежнев

Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт космических и информационных технологий
Кафедра систем искусственного интеллекта

УТВЕРЖДАЮ

Заведующий кафедрой

_____ Г. М. Цибульский

подпись

« ____ » _____ 2018 г.

ЗАДАНИЕ
НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ
В форме бакалаврской работы

Студенту Непомнящих Алексею Владимировичу

Группа КИ14-11Б, направление 09.03.02 «Информационные системы и технологии», профиль 09.03.02.04 «Информационные системы и технологии в медиаиндустрии»

Тема выпускной квалификационной работы «Разработка сервиса мониторинга неоднородной структуры сельскохозяйственных земель».

Утверждена приказом по университету № 4533/с от 29.03.2018 г.

Руководитель ВКР Брежнев Р. В. Старший преподаватель кафедры систем искусственного интеллекта ИКИТ СФУ.

Исходные данные для ВКР: введение; глава 1. Теоритическая часть; выводы по главе 1; глава 2. Программная часть; выводы по главе 2; глава 3. Экспериментальная апробация модуля; заключение; список использованных источников; приложение (листинг программы); приложение (плакаты презентации).

Руководитель

подпись

Р. В. Брежнев

Выпускник

подпись

А. В. Непомнящих

«___» _____ 2018 г.

График

выполнения выпускной квалификационной работы студентом направления 09.03.02 «Информационные системы и технологии», профиля 09.03.02.04 «Информационные системы и технологии в медиаиндустрии».

Таблица 1 — График выполнения этапов ВКР

Наименование этапа	Срок выполнения этапа	Результат выполнения этапа	Примечание руководителя (отметка о выполнении этапа)
Ознакомление с целью и задачами работы	09.03 — 15.03	Краткий обзор по теме ВКР	Выполнено
Сбор и анализ литературных источников	15.04 — 25.04	Список использованных источников	Выполнено
Решение задач ВКР	26.04 — 14.05	Доклад с презентацией по решенным задачам	Выполнено
Компановка отчета и презентации по результатам решения задач ВКР	15.05 — 25.05	Отчет по результатам решения задач ВКР	Выполнено
Предварительная защита результатов ВКР	07.06	Доклад и презентация по проделанной работе	Выполнено
Нормоконтроль	04.06 — 16.06	Пояснительная записка и презентация к ВКР	Выполнено
Защита ВКР	22.06	Доклад и презентация по результатам бакалаврской работы	

Руководитель

подпись

Р. В. Брежнев

Выпускник

подпись

А. В. Непомнящих

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	4
1 Обзор кластеризации методом <i>k</i> -means	5
1.1 Кластеризация методом <i>k</i> -means	7
1.2 Определение качества кластеризации	8
1.3 Вывод по главе 1	13
2 Проектирование программного модуля кластеризации	13
2.1 Диаграмма прецедентов	14
2.2 Входные и выходные данные	14
2.2.1 Индекс NDVI	14
2.2.2 Landsat-8	16
2.3 Язык программирования	17
2.4 Диаграмма компонентов	18
2.5 Вывод по главе 2	19
3 Экспериментальная апробация модуля	20
ЗАКЛЮЧЕНИЕ	22
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	23
ПРИЛОЖЕНИЕ А Листинг программы	25
ПРИЛОЖЕНИЕ Б Плакаты презентации	27
ПРИЛОЖЕНИЕ В Акт об использовании результатов проектирования в рамках бакалаврской работы	34
ПРИЛОЖЕНИЕ Г Результат проверки в системе «Антиплагиат»	35

ВВЕДЕНИЕ

Классификация – основополагающий процесс интеллектуальной деятельности человека. При встрече с новым явлением, мы стараемся найти ему аналог в знакомой нам области. При рассмотрении группы каких-либо объектов, мы непроизвольно разделяем их на подгруппы близких друг другу элементов. Классификация присутствует при упорядочении известных нам фактов, явлений, предметов. Так же классификация играет значимую роль в науке: примерами служат теории Менделеева и Дарвина.

Можно сказать, что классификация одно из основополагающих понятий науки. Но поскольку классификация – это упорядочивание объектов по их схожести, а объектом можно назвать всё, что можно описать вектором дескрипторов, включая действия и процессы, то можно прийти к выводу, что классификация - это характерная способность всех живых организмов.

Если бы они не были способны собирать схожие внешние раздражители в группы для определения классов раздражителей, для которых нужны соответствующие положительные или отрицательные реакции, они были бы недостаточно приспособлены для дальнейшего выживания. Поэтому процедуры используемые в кластер-анализе для выявления групп похожих объектов просто систематизируют и оценивают количественно один из фундаментальных процессов присущих не только людям, но и абсолютно всем живым существам.

В процессе создания математических моделей описывающих естественный процесс классификации наблюдаемых явлений и объектов было получено множество алгоритмов и их модификаций с той или иной эффективностью решающих свою задачу. К одним из самых популярных методов кластеризации относится кластеризация методом k -means, реализованная в виде программного модуля в ходе выполнения работы.

1 Обзор кластеризации методом *k*-means

Основная цель кластеризации – выделить в исходных многомерных данных такие однородные подмножества, чтобы объекты внутри групп были похожи в известном смысле друг на друга, а объекты из разных групп – не похожи. Под «похожестью» понимается близость объектов в многомерном пространстве признаков, и тогда задача сводится к выделению в этом пространстве естественных скоплений объектов, которые и считаются однородными группами.

Кластер по-английски означает пучок, скопление, группа элементов, характеризующихся каким-либо общим свойством. Строго говоря, кластером называется такая группа объектов из рассматриваемого множества, для которой средний квадрат внутригруппового расстояния до центра группы меньше среднего квадрата расстояния до общего центра в исходной совокупности.

Если данные представлены в виде матрицы объект - признак, то анализируемые объекты удобно интерпретировать геометрически как точки в многомерном пространстве признаков. Если признаков всего три, то исследуемые объекты представляются в виде точек в трёхмерном евклидовом пространстве. Следует считать, что геометрическая близость двух или нескольких точек в этом пространстве обозначает близость физических состояний этих объектов и их однородность. Тогда проблема кластеризации состоит в разбиении рассматриваемой совокупности точек на сравнительно небольшое число кластеров, таких, что точки, принадлежащие к одному кластеру, максимально «близки» друг к другу, а точки из разных кластеров максимально «далеки» друг от друга.

Пусть X – конечное множество входных векторов (X),
 C – конечное множество кластеров (C),
 μ – конечное множество центров масс соответствующих кластеров.

$$\sum_{i \in C} \sum_{x \in X} \|x - \mu_i\|^2 \quad (1)$$

, если для любых

При использовании критерия минимума суммы квадратов отклонений показатель качества кластеризации имеет вид:

$$\sum_{i \in C} \sum_{x \in X} \|x - \mu_i\|^2 \quad (2)$$

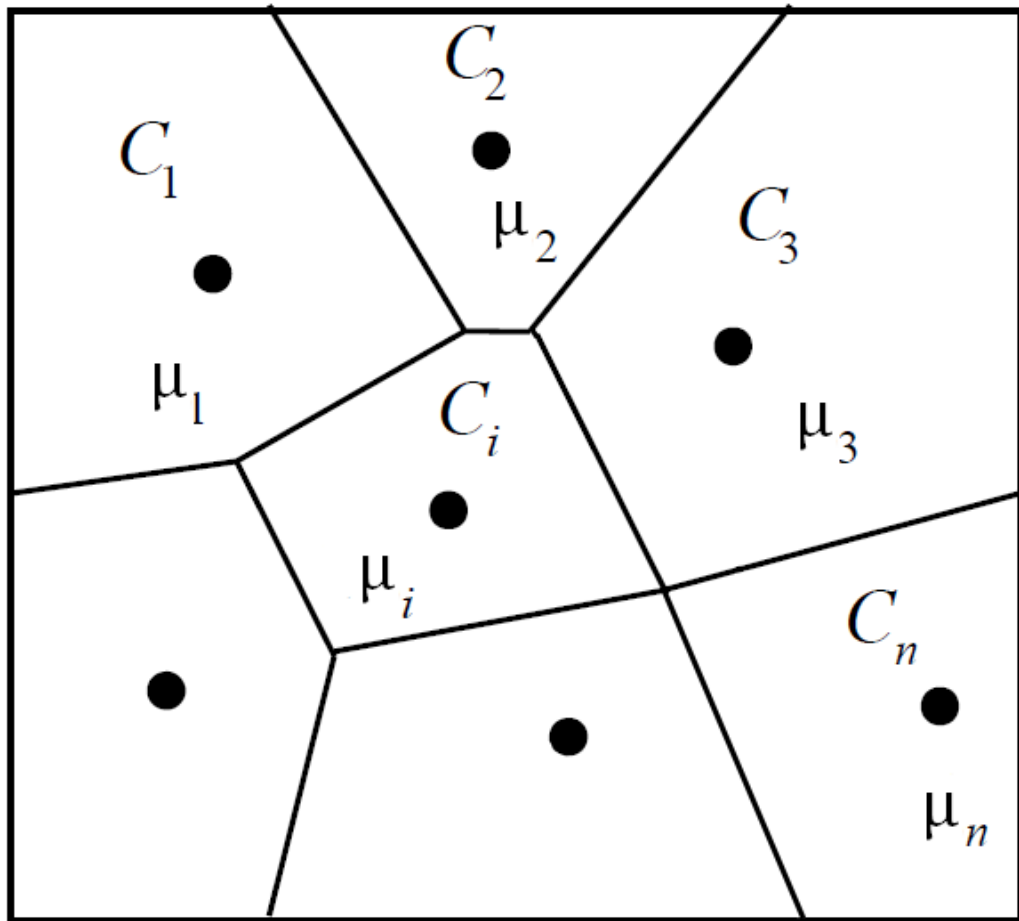


Рисунок 1 — Форма кластеров при кластеризации двумерных данных по критерию минимума суммы квадратов отклонений

1.1 Кластеризация методом k -means

В текущей работе применяется метод кластеризации k -means, принадлежащий к группе итеративных методов. В методе k -means за центр кластера принимается центроид, который вычисляется как среднее арифметическое всех объектов класса. Расстояние между объектом и кластером понимается как евклидово расстояние между объектом и центроидом кластера. Например, если в кластере состоящем из двух объектов, провести прямую между точками описывающими эти объекты, то центроидом этого кластера будет середина этого отрезка. Объект относится к тому кластеру, расстояние до центроида которого минимально.

Алгоритм метода k -means выглядит следующим образом:

- 1) для начала процедуры кластеризации должны быть заданы k случайно выбранных точек, которые будут служить центроидами кластеров (число которых априорно задается пользователем);
- 2) затем происходит определение принадлежности точек кластерам: каждая точка принадлежит ближайшему к ней кластеру;
- 3) вычисляются центроиды новых кластеров, как среднее арифметическое всех принадлежащих им точек;
- 4) шаги 2-3 повторяются, пока не будет найдено такое разбиение объектов на заранее заданное количество кластеров k , которое минимизирует сумму квадратов отклонений объектов кластера до центроида соответствующего кластера.

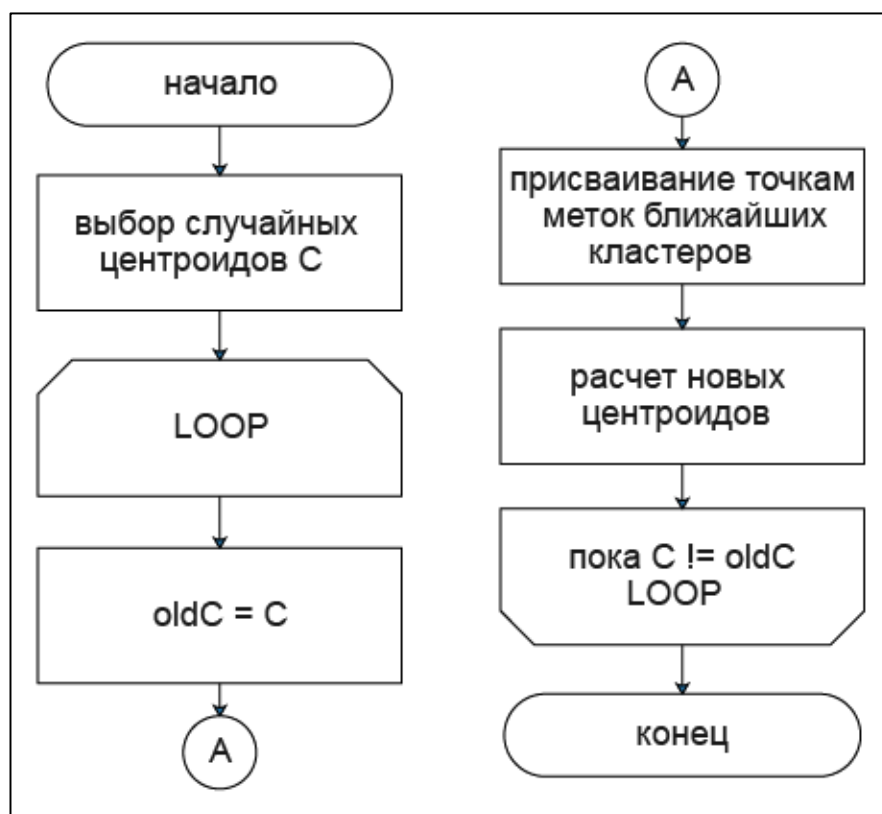


Рисунок 2 — Блок-схема кластеризации методом k-means

1.2 Определение качества кластеризации

На сегодняшний день нет чётких разработок по проверке оптимальности результатов кластеризации, но имеются способы определения количественных критериев, следуя которым можно предпочесть одно разбиение другому. С этой целью в кластер-анализе вводится понятие критерия качества разбиения, определённого на множестве всех разбиений. Критерий зависит от объёмов кластеров и расстояний между объектами, вошедшими в отдельные кластеры. Наилучшим разбиением считается то разбиение S^* из всех S , на котором достигается экстремум (минимум или максимум) выбранного критерия качества.

Ниже приведены примеры критериев качества кластеризации рассмотренные в ходе выполнения работы:

1) Сумма квадратов расстояний до центров кластеров.

Данный критерий качества кластеризации рассчитывается по формуле:

$$\sum_j \sum_i d_{ij}^2 \quad (3)$$

где j — номер кластера,

\bar{x}_j — центр j -го кластера,

x_i — вектор значений переменных для i -го объекта, входящего в j -ый кластер,

d_{ij} — расстояние между i -м объектом и центроидом j -го кластера.

При увеличении количества кластеров значение дисперсии будет снижаться до нуля. По графику зависимости дисперсии от количества кластеров видно, что снижение дисперсии замедляется, на графике это происходит в точке называемой «локтем».

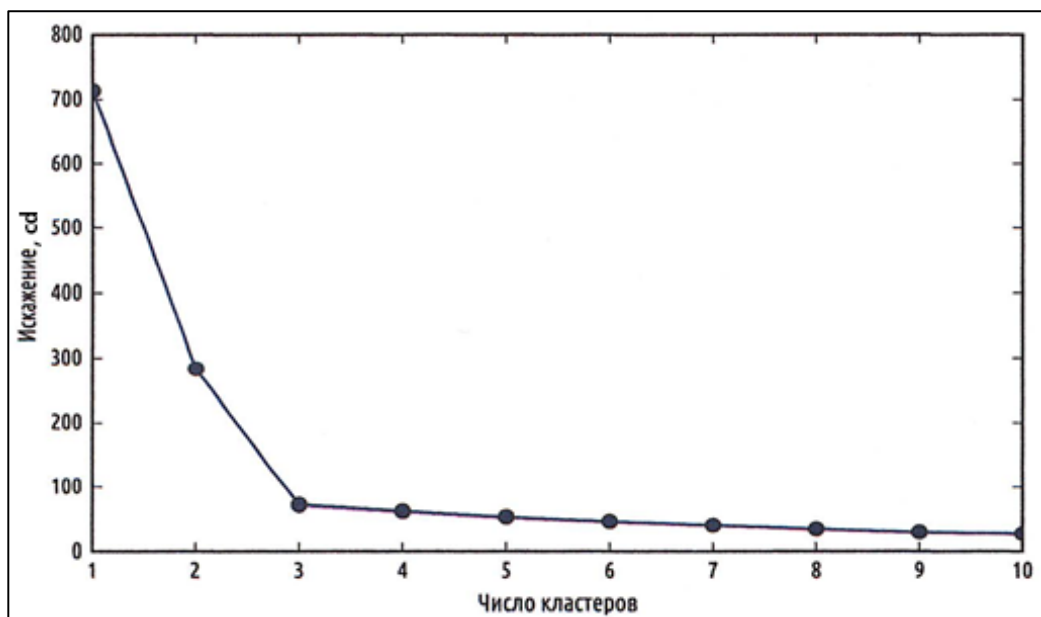


Рисунок 3 — График зависимости дисперсии разбиения от количества кластеров

Исходя из данных представленных на рисунке 3 можно предположить, что для наилучшего качества кластеризации будет достаточно трех кластеров, а дальнейшее увеличение количества кластеров не приведет к существенному уменьшению дисперсии.

Метод «локтя» прост в применении, но требует визуального анализа графика, что не подходит для автоматического определения оптимального количества кластеров, также при расчете дисперсии не учитывается расположение относительно других кластеров, поэтому для определения оптимального количества кластеров решено применить следующий критерий качества разбиения.

2) Коэффициент силуэта.

Данный критерий качества кластеризации предложен бельгийским статистиком Питером Россью в работе «Silhouettes: a graphical aid to the interpretation and validation of cluster analysis» 1987 года.

«Силуэт» каждого кластера определяется следующим образом: допустим элемент принадлежит кластеру . Обозначим среднее расстояние от этого объекта до других объектов из того же кластера через . Теперь обозначим среднее расстояние от до объектов из другого кластера , через . Положим . Смысл этой величины можно определить как меру несхожести отдельного элемента с элементами ближайшего кластера. Таким образом, «силуэт» каждого отдельного элемента определяется как

(4)

Оценка для всей кластерной структуры достигается усреднением показателя по элементам:

$$-\sum \quad (5)$$

где n — количество кластеров в разбиении.

Силуэтный коэффициент показывает, на сколько среднее расстояние до объектов текущего кластера отличается от среднего расстояния до объектов ближайшего кластера. Данная величина лежит в диапазоне от минус одного до единицы. Значения, близкие к минус одному, соответствуют плохой (разрозненной) кластеризации, значения, близкие к нулю, говорят о том, что кластеры пересекаются и накладываются друг на друга, значения, близкие к единице, соответствуют «плотным» четко выделенным кластерам. Таким образом, чем больше коэффициенты силуэтов, тем более четко выделены кластеры.

График содержащий графическое представление силуэтных коэффициентов разбиения представленный на рисунке 4 свидетельствует о качественной кластеризации так как все из полученных силуэтных коэффициентов больше нуля, а средний коэффициент, изображенный в виде пунктирной линии, достаточно отдалён от нуля.

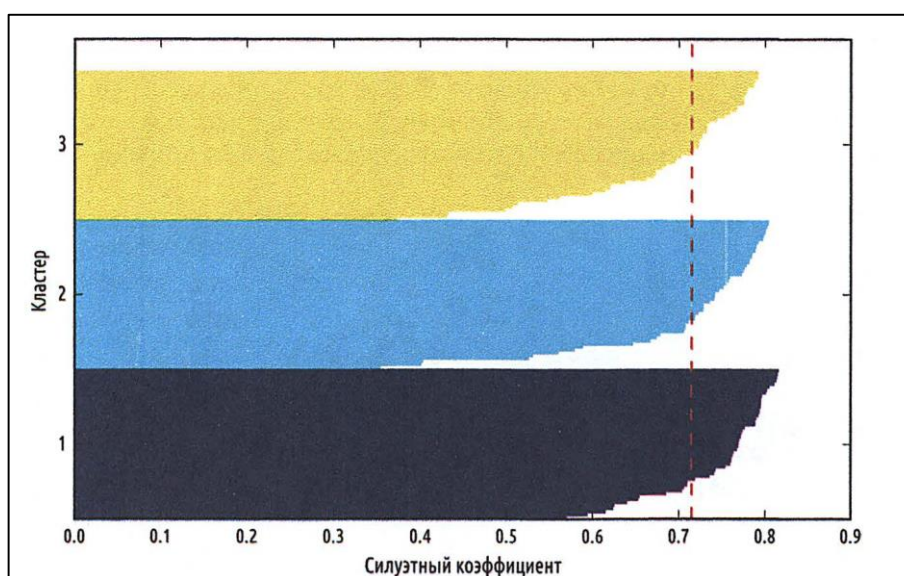


Рисунок 4 — График силуэтных коэффициентов

График силуэтных коэффициентов так же как и график зависимости дисперсии разбиения от количества кластеров наглядно показывает качество разбиения, а средний силуэтный коэффициент дает оценку разбиению в виде конкретного числа. Благодаря этому появляется возможность провести серию кластеризаций с разным числом кластеров как показано на рисунке 5 и определить оптимальное количество кластеров путем поиска максимального значения среднего силуэтного коэффициента. Именно поэтому этот функционал качества кластеризации выбран для автоматического определения оптимального количества кластеров в программном модуле кластеризации.

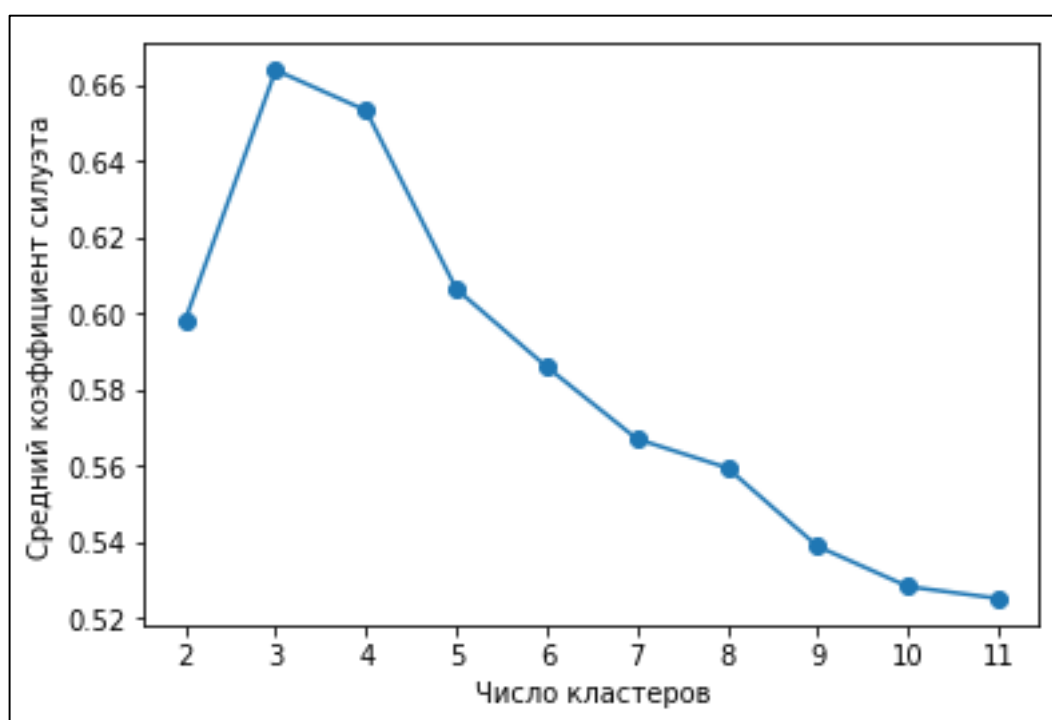


Рисунок 5 — График зависимости среднего коэффициента силуэта от количества кластеров

На графике представленном на рисунке 5 максимальное значение среднего силуэтного коэффициента соответствует разбиению с тремя кластерами, что говорит о приемлемом качестве кластеризации при разбиении на три кластера.

1.3 Вывод по главе 1

В ходе обзора метода кластеризации k-means были определены преимущества к которым можно отнести простоту алгоритма и его способность автоматически определять пороги для дальнейшей разметки изображения, а так-же наличие возможности автоматического определения оптимального количества кластеров с помощью среднего силуэтного коэффициента.

2 Проектирование программного модуля кластеризации

В ходе обзора алгоритма локализации неоднородных областей методом k-means определена последовательность действий, которую необходимо реализовать в виде программного модуля.

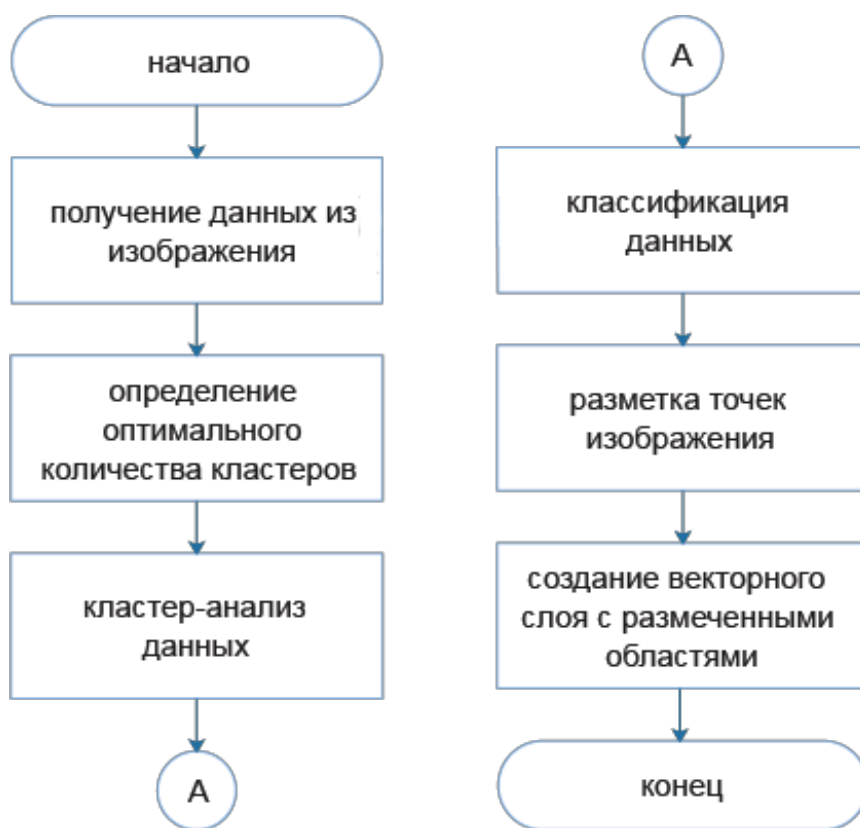


Рисунок 6 – Блок-схема работы программного модуля

2.1 Диаграмма прецедентов

Выделяется два основных прецедента реализуемых в программном модуле кластеризации изображения методом k-means.

Анализ и выбор оптимального количества кластеров, которое определяется с помощью среднего коэффициента силуэтов кластеров.

Непосредственно процесс кластеризации изображения.

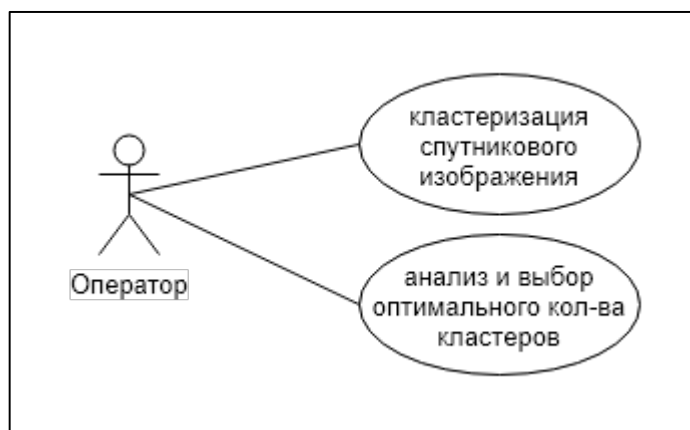


Рисунок 7 — Диаграмма прецедентов реализуемого программного модуля

2.2 Входные и выходные данные

В качестве входных данных программный модуль будет принимать изображение в формате GeoTIFF с одноканальным растровым слоем с рассчитанным индексом NDVI на основе данных полученных со спутника программы Landsat-8.

В качестве выходных данных выступает векторный слой в формате ESRI SHP содержащий размеченные области исходного изображения.

2.2.1 Индекс NDVI

Характерным признаком растительности и ее состояния является спектральная отражательная способность, характеризующаяся большими различиями в отражении излучения разных длин волн. Знания о связи структуры и состояния растительности с ее спектрально отражательными способностями позволяют использовать аэрокосмические снимки для

картографирования и идентификации типов растительности и их стрессового состояния.

Расчет большей части вегетационных индексов базируется на двух наиболее стабильных (не зависящих от прочих факторов) участках кривой спектральной отражательной способности растений (рис. 1). На красную зону спектра приходится максимум поглощения солнечной радиации хлорофиллом, а на ближнюю инфракрасную зону максимальное отражение энергии клеточной структурой листа, то есть высокая фотосинтетическая активность (связанная, как правило, с большой фитомассой растительности) ведет к более низким значениям коэффициентов отражения в красной зоне спектра и большим значениям в ближней инфракрасной. Как это хорошо известно, отношение этих показателей друг к другу позволяет четко отделять растительность от прочих природных объектов. Для работы со спектральной информацией часто прибегают к созданию так называемых «индексных» изображений. На основе комбинации значений яркости в определенных каналах, информативных для выделения исследуемого объекта, и расчета по этим значениям «спектрального индекса» объекта строится изображение, соответствующее значению индекса в каждом пикселе, что и позволяет выделить исследуемый объект или оценить его состояние. Спектральные индексы, используемые для изучения и оценки состояния растительности, получили общепринятое название вегетационных индексов. NDVI – нормализованный разностный индекс растительности был впервые описан Rouse B.J. в 1973 г. – простой количественный показатель количества фитомассы. Говоря вегетационный индекс, часто подразумевают именно его.

Индекс вычисляется по следующей формуле:

где RED — коэффициент отражения в красной спектральной зоне,
NIR — коэффициент отражения в ближней инфракрасной спектральной зоне.

Для растительности индекс NDVI принимает положительные значения, и чем больше зеленая фитомасса, тем они выше. На значения индекса влияет также видовой состав растительности, ее сомкнутость, состояние, экспозиция и угол наклона поверхности, цвет почвы под разреженной растительностью. Индекс умеренно чувствителен к изменениям почвенного фона, кроме случаев, когда густота растительного покрова ниже 30%. Индекс может принимать значения от минус одного до единицы. Для зеленой растительности индекс обычно принимает значения от 0,2 до 0,8.

Основным преимуществом вегетационных индексов является легкость их получения и широкий спектр решаемых с их помощью задач. Так, NDVI часто используется как один из основных инструментов при проведении более сложных типов анализа, результатом которых могут являться карты продуктивности лесов и сельскохозяйственных земель, карты ландшафтов и природных зон, почвенные карты. Также на его основе возможно получение численных данных для использования в расчетах оценки и прогнозирования урожайности и продуктивности, биологического разнообразия, степени нарушенности и ущерба от различных стихийных бедствий, техногенных аварий и т. д.

2.2.2 Landsat-8

Landsat — наиболее продолжительный проект съемки Земли из космоса. Первый из спутников в рамках программы был запущен в 1972 г.; последний на данный момент (Landsat-8) — 11 февраля 2013 г.

Данные со спутника Landsat-8 доступны для всех пользователей. Ежедневно спутник снимает порядка 400 сцен которые после обработки в соответствии с текущим стандартом продуктов Landsat, хранятся в Центре хранения данных Геологической службы США. Большая часть данных предоставляется пользователям менее чем через 24 часа после приема.

На борту космического аппарата установлены многоканальный сканирующий радиометр OLI (Operational Land Imager) и сканирующий

двухканальный ИК-радиометр TIRS (Thermal Infrared Sensor). Радиометр OLI позволяет получать изображения земной поверхности с максимальным разрешением 15 м/с использованием усовершенствованных технологий космической съемки. ИК-радиометр TIRS предназначен для получения «теплового» изображения земной поверхности с разрешением 100 м.

Таблица 1 — Каналы космоснимков Landsat-8

№ канала	Спектральный канал	Длины волн, мкм	Разрешение, м
1	Побережья и аэрозоли (Coastal/Aerosol, New Deep Blue)	0.433 – 0.453	30
2	Синий (Blue)	0.450 – 0.515	30
3	Зеленый (Green)	0.525 – 0.600	30
4	Красный (Red)	0.630 – 0.680	30
5	Ближний ИК (NIR)	0.845 – 0.885	30
6	Ближний ИК (SWIR 2)	1.560 – 1.660	30
7	Ближний ИК (SWIR 3)	2.100 – 2.300	30
8	Панхроматический (PAN)	0.500 – 0.680	15
9	Пористые облака (SWIR)	1.360 – 1.390	30
10	Дальний ИК (TIR1)	10.30 – 11.30	100
11	Дальний ИК (TIR2)	11.50 – 12.50	100

2.3 Язык программирования

Для реализации алгоритма локализации неоднородностей на изображении методом *k-means* был выбран язык программирования высокого уровня Python.

2.4 Диаграмма компонентов

Соответственно списку задач были определены библиотеки языка программирования Python необходимые для реализации алгоритма, была составлена диаграмма компонентов представленная на рисунке 8.

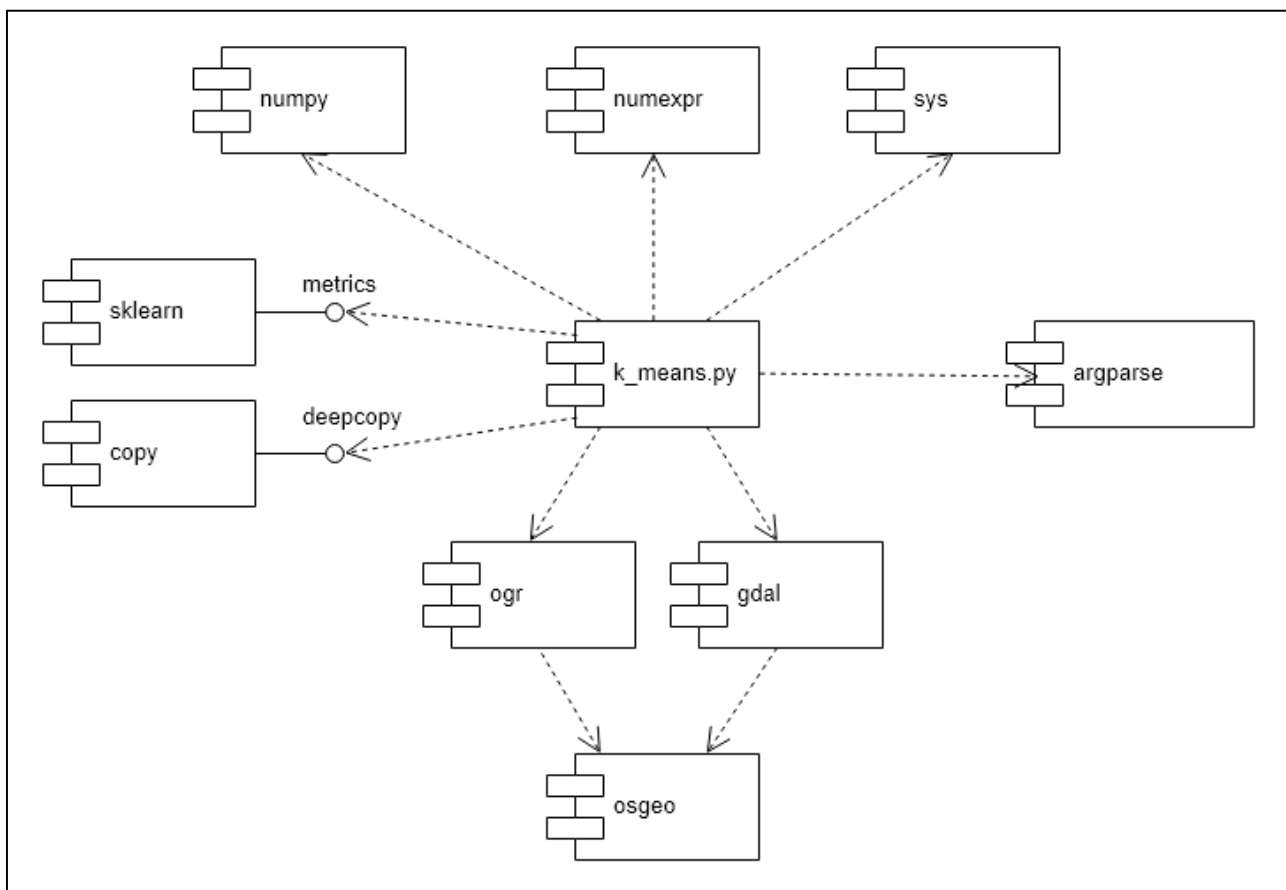


Рисунок 8 — Диаграмма компонентов реализуемого программного модуля

Библиотека OGR необходима для создания векторного изображения со слоем содержащим размеченные области.

Библиотека GDAL необходима для открытия изображения и получения данных из растрового слоя.

Библиотека Numpy необходима для выполнения математических вычислений производимых над данными в ходе выполнения алгоритма кластеризации k-means.

Библиотека Numexpr необходима для выполнения некоторых математических вычислений, которые не достаточно оптимизированы в библиотеке Numpy.

Библиотека Sys позволяет реализовать программный модуль в виде сценария командной строки который может принимать аргументы такие как название обрабатываемого изображения и название для выходных данных.

Функция deersору библиотеки сору необходима на этапе сравнения который происходит на каждой итерации кластеризации ввиду особенностей реализации языка программирования Python.

Модуль metrics библиотеки sklearn позволяет получить средний силуэтный коэффициент кластеризованного изображения. Предпочтение было дано готовому модулю по причине большей скорости выполнения функции по сравнению с реализацией посредством библиотек Numpy и Numexpr.

Библиотека Arparse необходима для удобной работы с аргументами командной строки.

2.5 Вывод по главе 2

В ходе выполнения проектной части работы выполнено следующее:

1. Разработан алгоритм, который необходимо реализовать в виде программного модуля;
2. Выбран язык программирования для написания программного кода модуля.
3. Определены библиотеки языка программирования необходимые для реализации программного модуля.
4. Написан программный код на языке программирования Python реализующий алгоритм определенный на первом шаге. Программный код представлен в приложении А.

3 Экспериментальная апробация модуля

В ходе выполнения работы создан программный модуль который принимает в качестве входных аргументов название растрового файла в формате GeoTIF и название для выходных файлов.

В качестве входных данных выступает фрагмент снимка со спутника Landsat-8 с рассчитанным NDVI по маске полей представленный на рисунке 9.

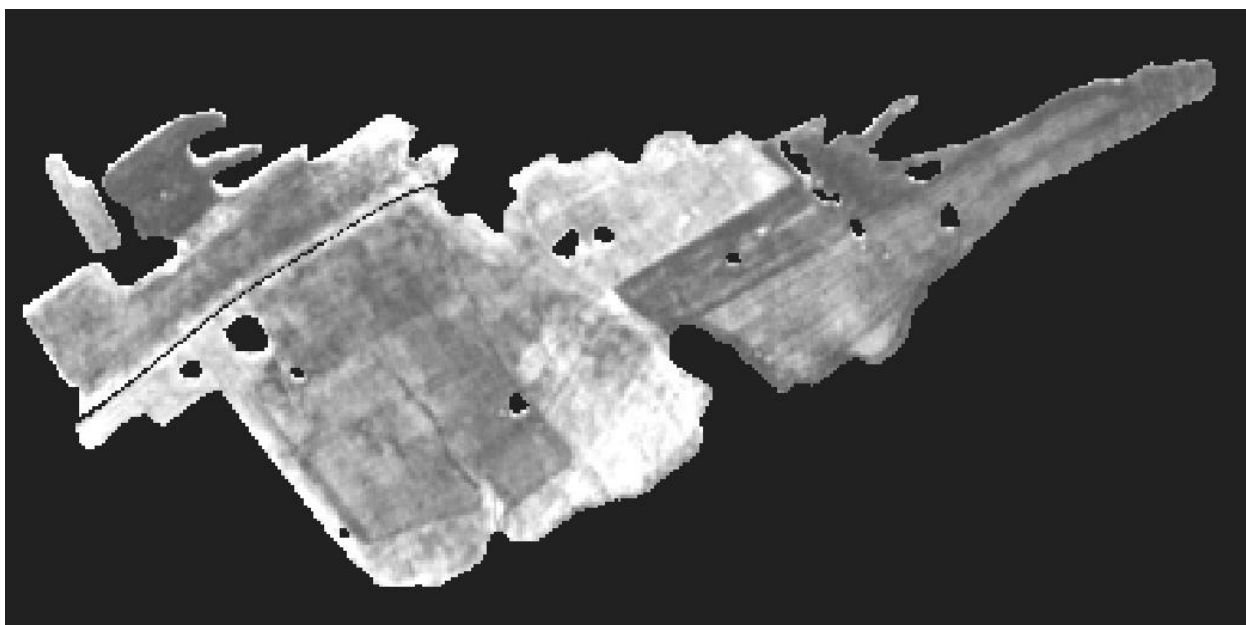


Рисунок 9 – Входное изображение

В качестве выходных данных выступает файл векторной графики в формате SHP ESRI содержащий размеченные области. Результат представлен на рисунке 10.

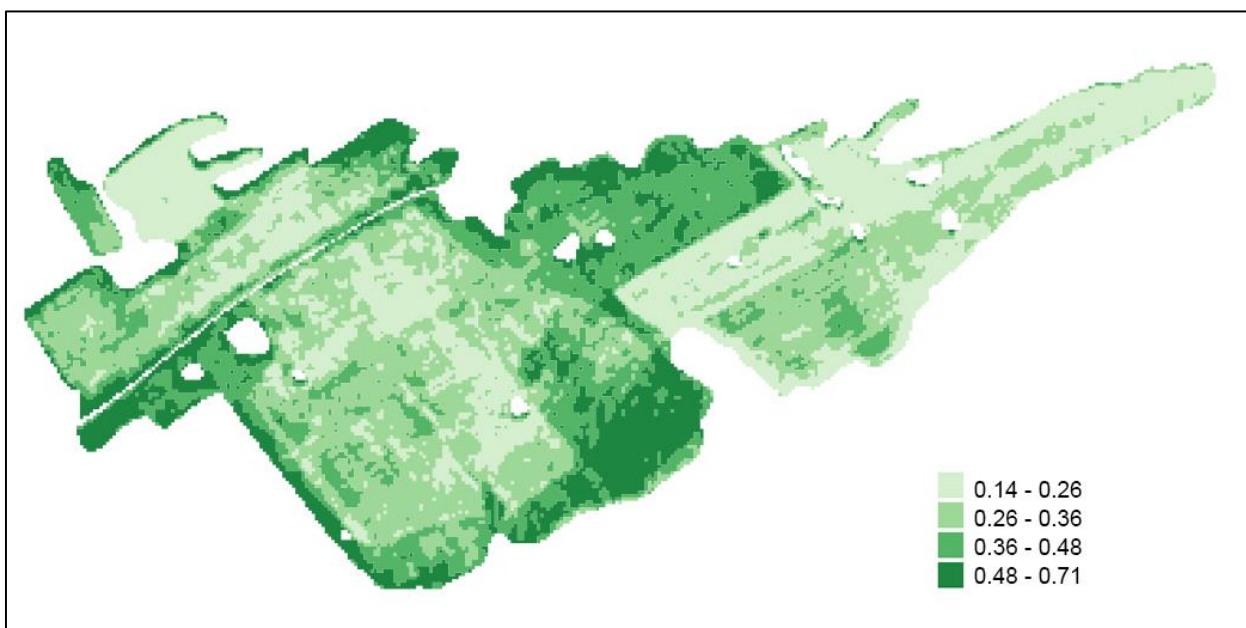


Рисунок 10 – Выходное изображение

После расчета среднего силуэтного коэффициента для каждого разбиения из проведенной серии кластеризаций программный модуль определил, что максимальное значение силуэтного коэффициента достигается при разбиении входного изображения на четыре кластера. После чего сохранил результат разбиения с максимальным средним силуэтным коэффициентом в виде файла векторной графики в формате SHP ESRI.

ЗАКЛЮЧЕНИЕ

В ходе выполнения выпускной квалификационной работы были выполнены все поставленные задачи. Произведен обзор кластеризации методом k -means, найден способ автоматического определения оптимального количества кластеров, спроектирован и реализован программный модуль локализации неоднородностей методом k -means с предварительным расчетом оптимального количества кластеров с помощью подсчета среднего силуэтного коэффициента разбиения. Произведена экспериментальная апробация программного модуля на данных спутника Landsat-8.

Разработанный сервис встроен в программно аппаратный комплекс ГИС ИКИТ СФУ, о чем свидетельствует составленный акт об использовании модуля в структуре работы комплекса.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Цифровая обработка аэрокосмических изображений. Версия 1.0 [Электронный ресурс] : электрон. учеб. пособие / В. Б. Кашкин, А. И. Сухинин.– Красноярск : ИПК СФУ, 2008. – (Цифровая обработка аэрокосмических изображений : УМКД № 54-2007 / рук. творч. коллектива В. Б. Кашкин.
2. Разработка геоприложений на языке Python / пер. с англ. А. В. Логунова. - М: ДМК Пресс, 2017
3. GDAL/OGR Python API [Электронный ресурс]: Документация GDAL Python API. – Режим доступа: <http://gdal.org/python>
4. Numpy and Scipy Documentation Numpy and Scipy documentation [Электронный ресурс]: Документация библиотеки NumPy. – Режим доступа: <https://docs.scipy.org/doc>
5. Факторный, дискриминантный и кластерный анализ: Пер. с англ./Дж. О. Ким, Ч. У. Мьюллер, У. Р. Клекка и др.; Под ред. И. С. Енюкова. — М.: Финансы и статистика, 1989
6. Кластерный анализ. И. Д. Мандель М.: Финансы и статистика, 1988
7. Обзор методов фильтрации и сегментации цифровых изображений / В. В. Стругайло. Науч. издание МГТУ им. Н. Э. Баумана «Наука и образование» №5, Май 2012
8. Сравнение алгоритмов кластерного анализа на случайном наборе данных / Е. С. Подвальный, А. В. Плотников, А. М. Белянин Воронежский государственный технический университет
9. Python и машинное обучение /С. Рашка пер. с англ. А. В. Логунова. - М.: ДМК Пресс, 2017.
10. Кластерный анализ. / Б. Дюран, П. Одрел пер. с англ. Е. З. Демиденко. Под ред. А. Я. Боярского. М., «Статистика», 1977.

11. Компьютерная обработка естественно-научных данных методом многомерной прикладной статистики / Л. И. Дубровская, Г. Б. Князев Учебное пособие. – Томск: ТМЛ-Пресс, 2011.
12. Статистическая классификация и кластерный анализ. / Л. Х. Гитис М.: Издательство Московского государственного горного университета, 2003.
13. Определение физического смысла комбинации каналов снимков Landsat для мониторинга состояния наземных и водных экосистем / С. И. Евдокимов, С. Г. Михалап. Серия «Естественные и физико-математические науки» 7/2015.
14. Классификация и кластер / Дж. Вэн Райзин пер. с англ. П. П. Кольцова под ред. Ю. И. Журавлева, Издательство Мир Москва, 1980.
15. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis / Peter J. Rousseeuw, Journal of Computational and Applied Mathematics №20, 1987.
16. Компьютерная обработка и распознавание изображений / В. Т. Тарасенко, Т. Ю. Фисенко, СПб: СПбГУ ИТМО, 2008.
17. Сравнение алгоритмов кластерного анализа на случайном наборе данных / Е. С. Подвальный, А. В. Плотников, А. М. Белянин.
18. Анализ и обработка изображений: принципы и алгоритмы / В. В. Яншин, М.: Машиностроение, 1995.
19. Компьютерные методы анализа видеоинформации / Д. А. Денисов, Издательство Красноярского университета, 1993.
20. ГОСТ 19.701-90. Схемы алгоритмов, программ, данных и систем. Условные обозначения и правила выполнения
21. СТО 4.2–07–2014. Стандарт организации. Система менеджмента качества. Общие требования к построению, изложению и оформлению документов учебной деятельности.

ПРИЛОЖЕНИЕ А

Листинг программы

```
#!/bin/python2
import numpy as np
import numexpr as ne
import argparse

from sklearn import metrics
from osgeo import gdal, ogr
from copy import deepcopy

parser = argparse.ArgumentParser(prog='k_means.py')
parser.add_argument('-i', '--input', type=str)
parser.add_argument('-o', '--out', type=str, default='')
parser.add_argument('-c', '--clusters', type=int, default=0)
args = parser.parse_args()

file = args.input
out = args.out
cls = args.clusters

if not out:
    out = file[file.rfind('/') + 1: file.rfind('.')] + '_result'

res = gdal.Open(file)
band = res.GetRasterBand(1)
data = band.ReadAsArray()
arr = data[data > 0]
arr.sort()

distance = lambda c: ne.evaluate('(arr - c) ** 2')
vdistance = np.vectorize(distance)

def getNearC(centroids, arr):
    tmp = vdistance(centroids[0])

    for g in centroids[1:]:
        tmp = np.append(tmp, vdistance(g))

    tmp = tmp.reshape(-1, arr.shape[0]).T

    nearc = tmp.argmax(axis=1)
    dist = tmp.min(axis=1).sum()

    return nearc, dist

def kmeans(nc, arr, v = False):
    centroids = np.random.uniform(arr.min(), arr.max(), nc)
    centroids.sort()
    tmp = np.zeros(centroids.shape, dtype=float)
    steps = np.array([])

    while not np.all(np.isclose(tmp, centroids, rtol=1e-04)):
        nearc, dist = getNearC(centroids, arr)
        tmp = deepcopy(centroids)

        for i in range(centroids.shape[0]):
```

```

        centroids[i] = arr[np.where(nearc == i)].mean()

    if v:
        steps = np.append(steps, deepcopy(centroids))

    return centroids, dist, steps.reshape(-1, nc)

def kmeansN(arg):
    nc, arr, n, v = arg
    result = [kmeans(nc, arr, v) for i in range(n)]
    return sorted(result, key = lambda x: x[1])[0]

if cls:
    optimumC = cls
else:
    results = []
    arg = [(i, arr, 3, False) for i in range(1, 10)]

    for i in range(1, 10):
        results.append(kmeansN((i, arr, 3, False)))

    results = sorted(results, key=lambda x: x[1], reverse=True)
    silhouettes = np.array([])
    for cl in results[1:]:
        nearc = getNearC(cl[0], arr)[0]
        m = metrics.silhouette_score(arr.reshape(-1, 1), nearc)
        silhouettes = np.append(silhouettes, m)

    optimumC = silhouettes.argmax() + 2

c, d, _ = kmeansN((optimumC, arr, 4, False))
nearc = getNearC(c, arr)
intervals = []
for i in range(c.size):
    intervals.append((arr[np.where(nearc[0] == i)].min(),
arr[np.where(nearc[0] == i)].max()))

clusters = np.zeros(data.shape)
for n, i in enumerate(intervals):
    clusters[np.where((i[0] <= data) & (data <= i[1]))] = n + 1

rasterFile = out + '.tif'
driver = gdal.GetDriverByName("GTiff")
cols, rows = data.shape
outdata = driver.Create(rasterFile, rows, cols, 1, gdal.GDT_UInt16)
outdata.SetGeoTransform(res.GetGeoTransform())
outdata.GetRasterBand(1).WriteArray(clusters)

layerName = out
drv = ogr.GetDriverByName("ESRI Shapefile")
dstDs = drv.CreateDataSource(layerName + '.shp')
dstLayer = dstDs.CreateLayer(layerName, srs=None)
fd = ogr.FieldDefn('DN', ogr.OFTInteger)
dstLayer.CreateField(fd)

gdal.Polygonize(outdata.GetRasterBand(1), None, dstLayer, 0)

```

ПРИЛОЖЕНИЕ Б

Плакаты презентации

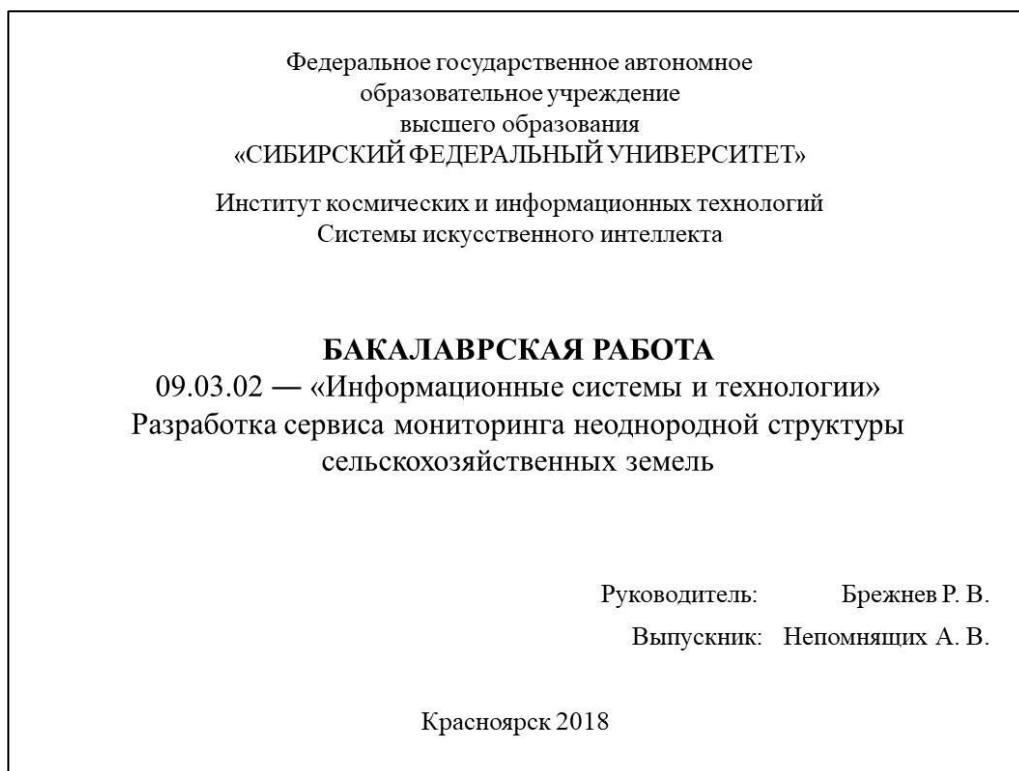


Рисунок Б.1 – Плакат презентации №1

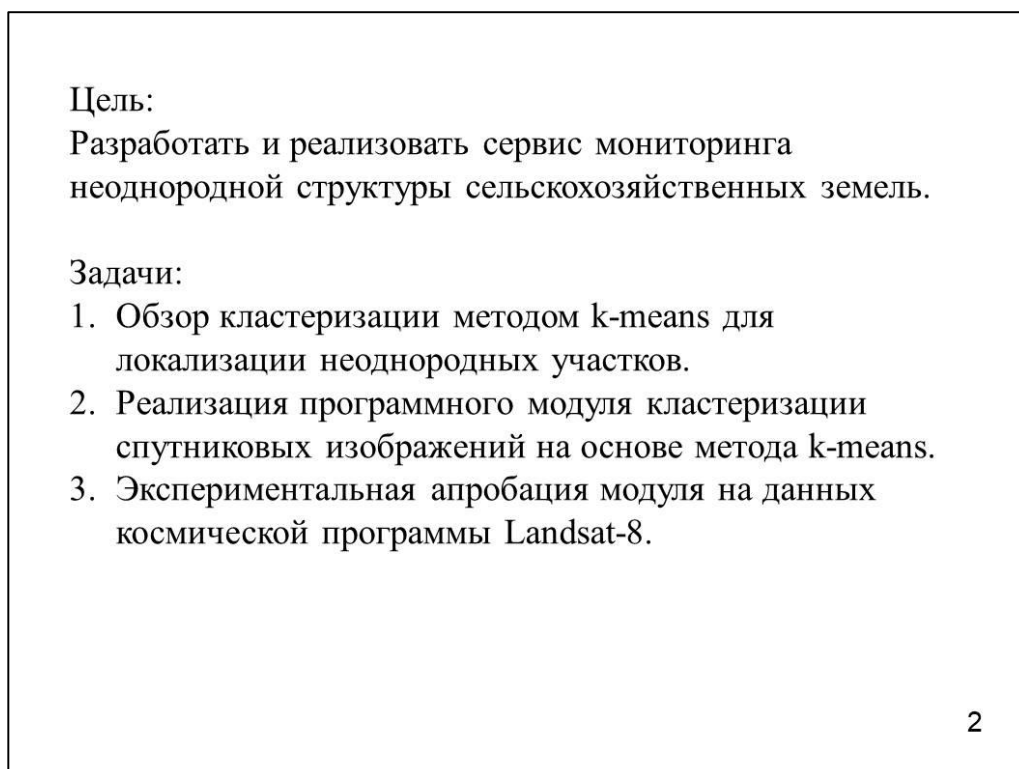


Рисунок Б.2 – Плакат презентации №2

Алгоритм работы с сервисом

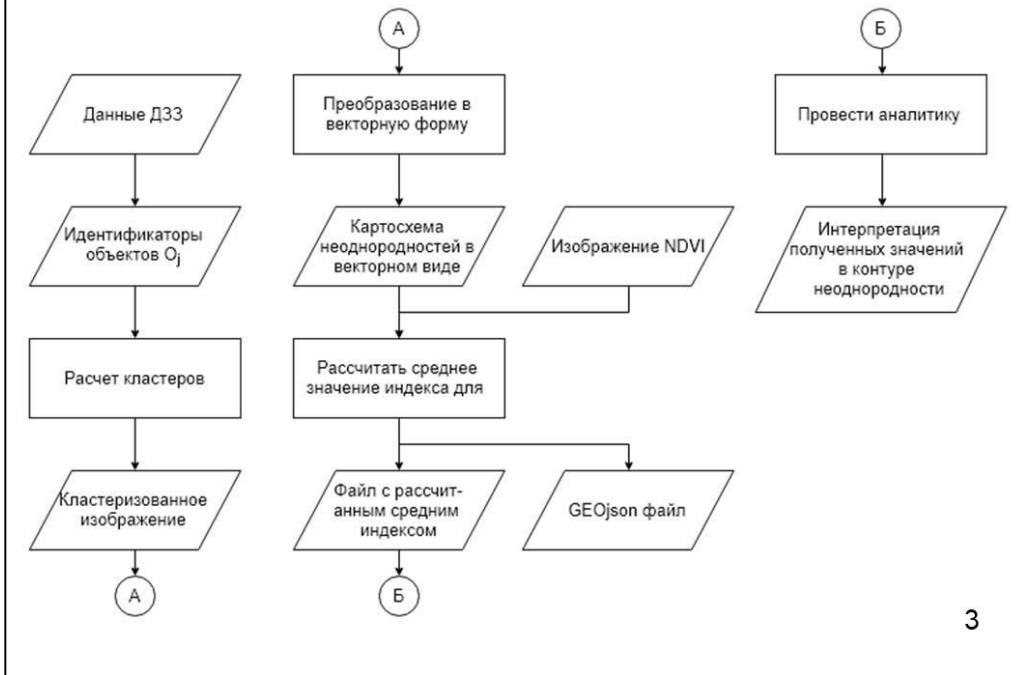
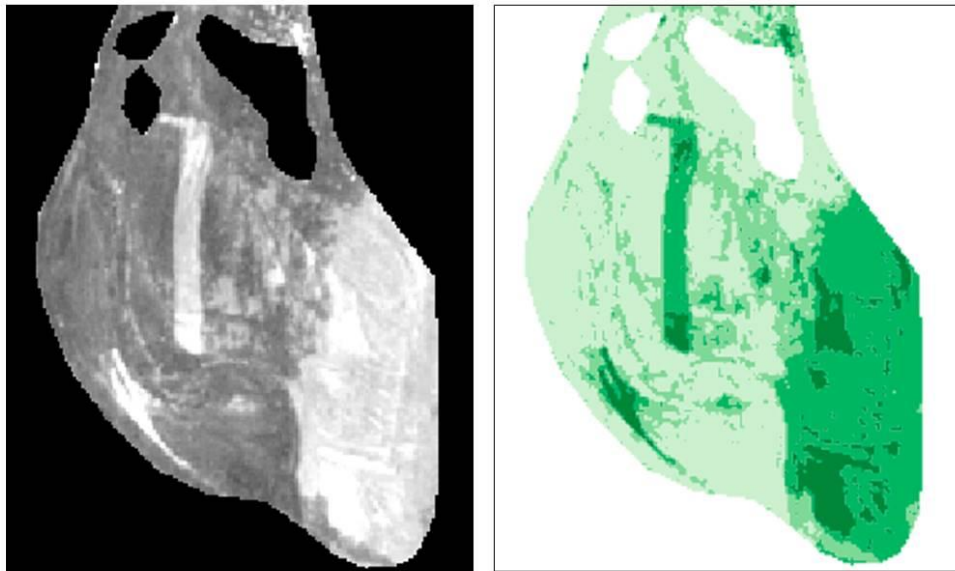


Рисунок Б.3 – Плакат презентации №3

Пример снимка с локализованными неоднородностями



4

Рисунок Б.4 – Плакат презентации №4

Локализация неоднородных участков

Кластеризация методом k-means

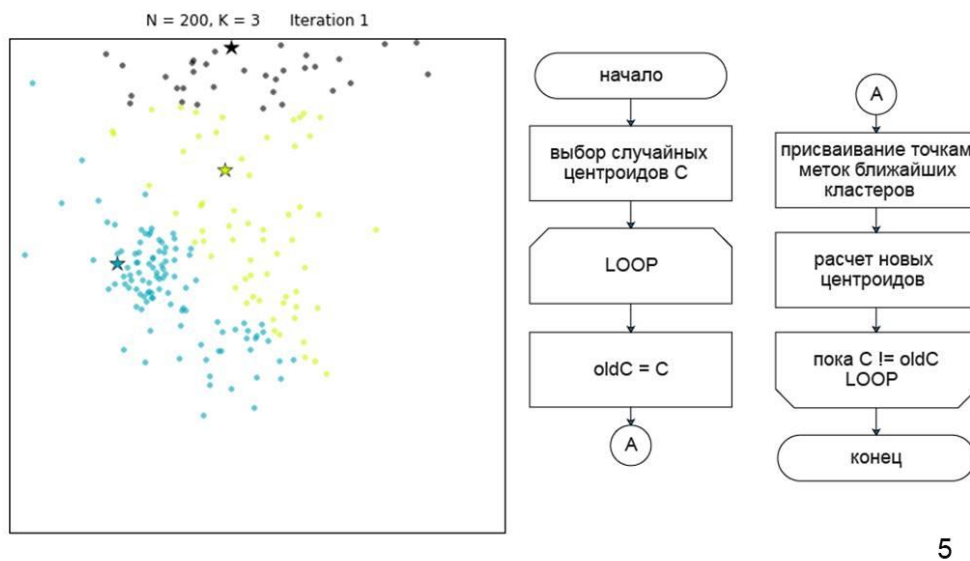


Рисунок Б.5 – Плакат презентации №5

Локализация неоднородных участков

Кластеризация методом k-means

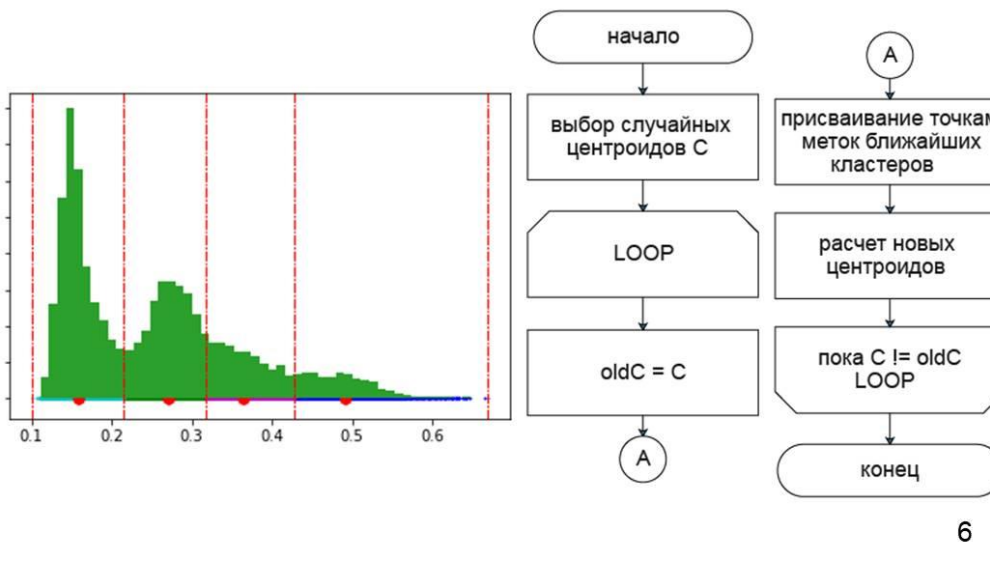
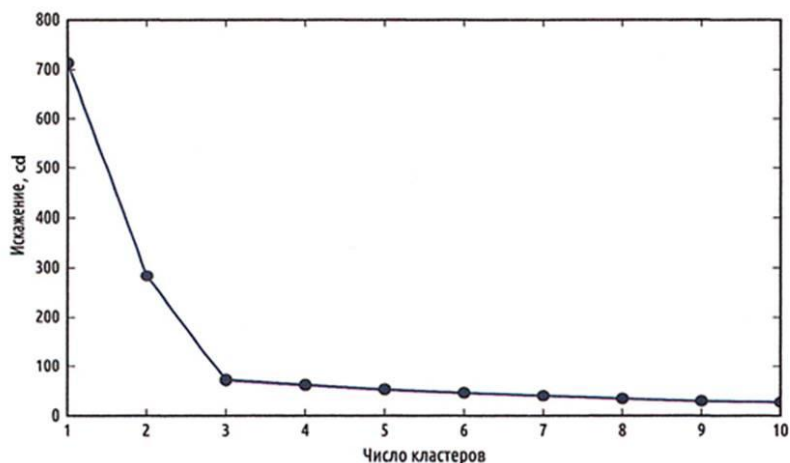


Рисунок Б.6 – Плакат презентации №6

Поиск оптимального количества кластеров
Сумма квадратов отклонений

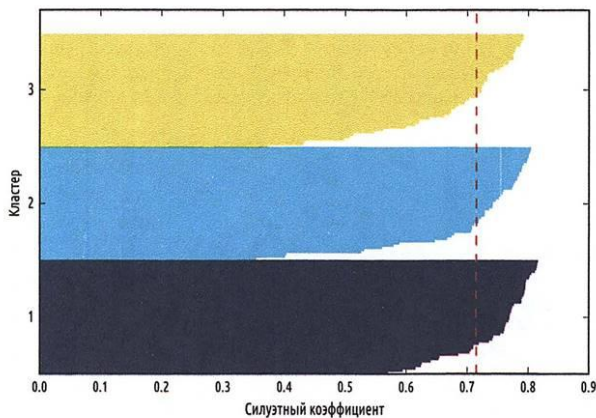


$$Q(S) = \sum_{l=1}^k \sum_{i \in S_l} d^2(X_i, \bar{X}_l)$$

7

Рисунок Б.7 – Плакат презентации №7

Поиск оптимального количества кластеров
Коэффициент силуэта



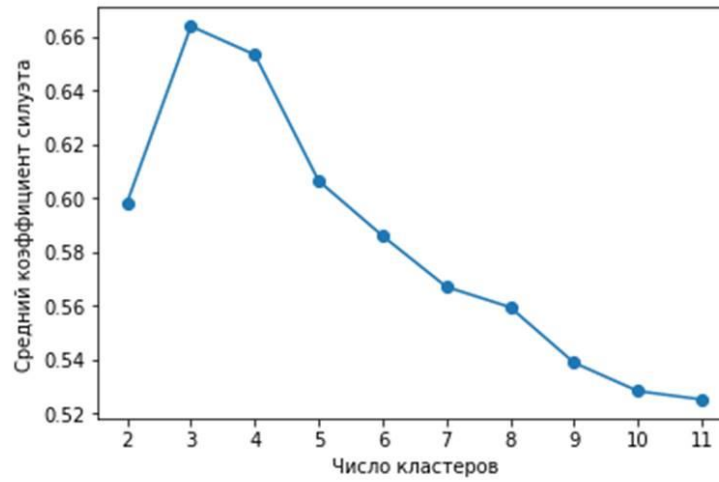
$$Q(S) = \frac{1}{n} \sum_{i=1}^n \frac{b(i) - a(i)}{\max[a(i), b(i)]}$$

где n – общее количество объектов;
 $b(i)$ – среднее расстояние до объектов ближайшего кластера;
 $a(i)$ – среднее расстояние до объектов текущего кластера.

8

Рисунок Б.8 – Плакат презентации №8

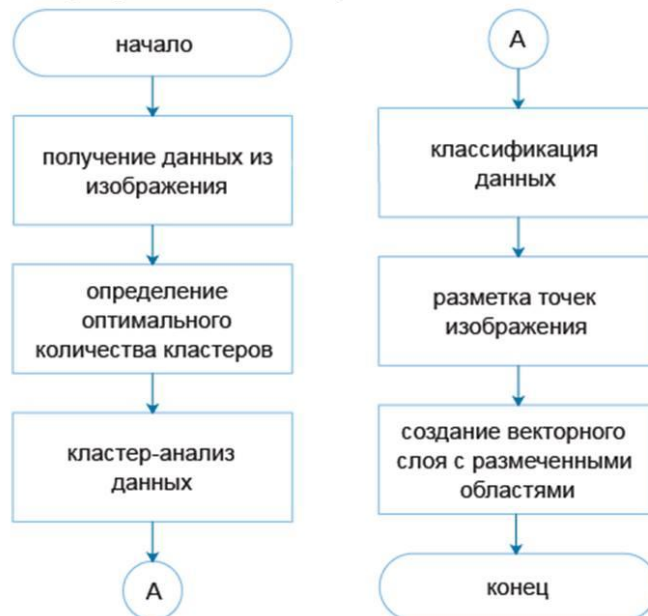
Поиск оптимального количества кластеров Коэффициент силуэта



9

Рисунок Б.9 – Плакат презентации №9

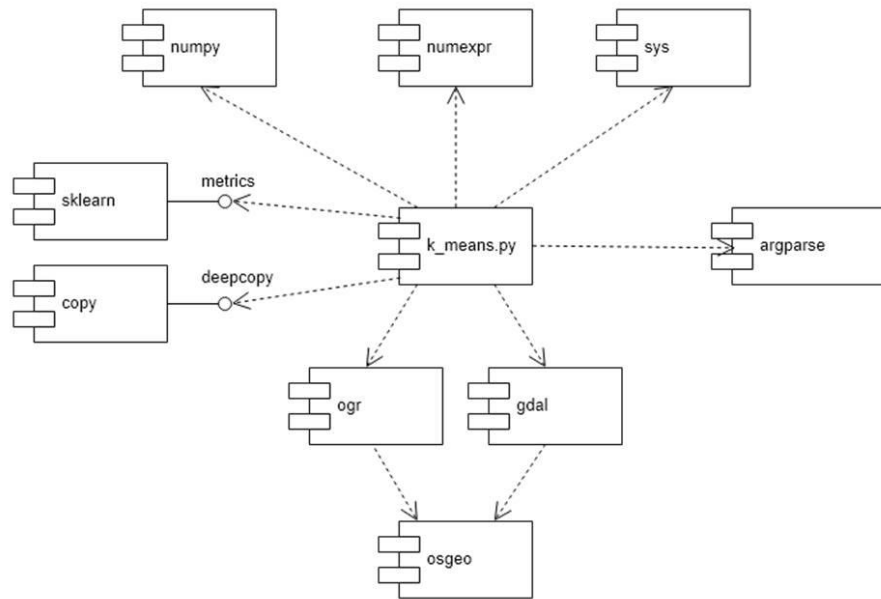
Разработка программного модуля Алгоритм программного модуля



10

Рисунок Б.10 – Плакат презентации №10

Разработка программного модуля Диаграмма компонентов

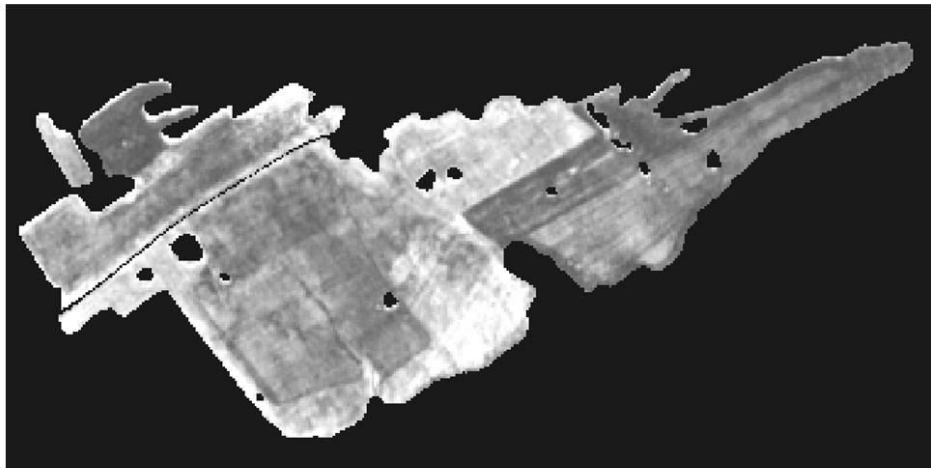


11

Рисунок Б.11 – Плакат презентации №11

Экспериментальная апробация Входные данные

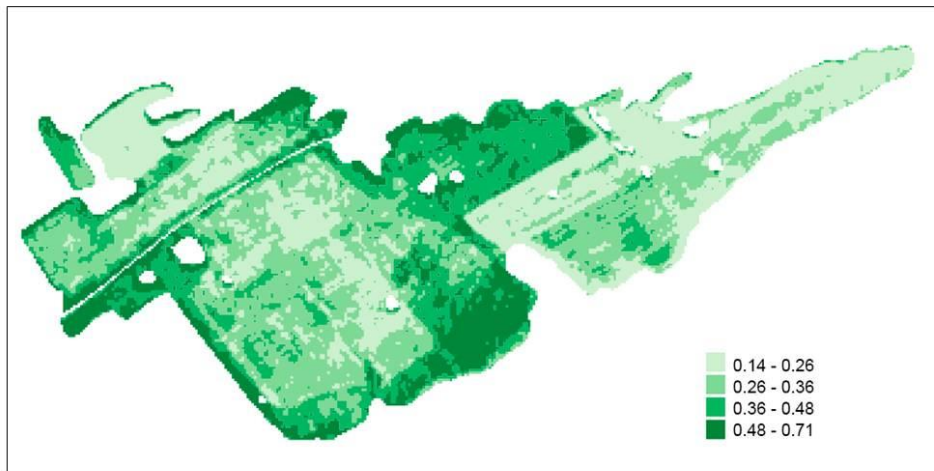
`k_means.py -i input.tif [-o output] [-c 0...n]`



12

Рисунок Б.12 – Плакат презентации №12

Экспериментальная апробация Результат



13

Рисунок Б.11 – Плакат презентации №13

Заключение

В ходе выполнения выпускной квалификационной работы были выполнены все поставленные задачи:

1. Обзор кластеризации методом k-means для локализации неоднородных участков.
2. Реализация программного модуля кластеризации спутниковых изображений на основе метода k-means.
3. Экспериментальная апробация модуля на данных космической программы Landsat-8.

14

Рисунок Б.11 – Плакат презентации №14

ПРИЛОЖЕНИЕ В

Акт об использовании результатов проектирования в рамках бакалаврской работы




<p>МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РФ Федеральное государственное автономное образовательное учреждение высшего образования «Сибирский федеральный университет» (СФУ) ИНСТИТУТ КОСМИЧЕСКИХ И ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ 660074, Красноярск, ул.Киренского, 26 Телефон: (3912) 912-575 Факс: (3912) 912-575 E-mail: адрес GTsybulsky@sfu-kras.ru</p>	<p>УТВЕРЖДАЮ</p> <p>зам. директора ИКИТ</p> <p>О.И. Киселев</p> 
<p><u>15.06.2018</u> № <u>47</u> на № _____ от _____</p>	
<h3>А К Т О Б И С П О Л Ь З О В А Н И И</h3> <p>результатов проектирования в рамках бакалаврской работы</p> <p>«15» июня 2018 г. г. Красноярск</p> <p>Комиссия в составе: руководитель НУЛ ИПКМ кафедры СИИ ИКИТ Маглинец Юрий Анатольевич, доцент кафедры СИИ ИКИТ Брежнев Руслан Владимирович, осуществила приемо-сдаточные испытания программного модуля кластеризации спутниковых изображений, встроенного программно-аппаратный комплекс ГИС ИКИТ СФУ и используемого в сервисе мониторинга неоднородной структуры земель сельскохозяйственного назначения.</p> <p>Модуль разработан студентом гр. КИ14-11Б Непомнящих Алексеем Владимировичем под руководством доцента кафедры «Системы искусственного интеллекта» ИКИТ СФУ Брежнева Руслана Владимировича в рамках выполнения бакалаврской работы.</p> <p>В настоящее время программный модуль внедрен в опытную эксплуатацию. Использование данного модуля позволяет в автоматическом режиме осуществлять кластеризацию спутниковых изображений для пространственной локализации и последующей интерпретации неоднородных участков исследуемых пространственных объектов.</p>	
<p>Доцент кафедры СИИ ИКИТ</p> <p>Руководитель НУЛ ИПКМ</p>	<p> </p> <p>Р. В. Брежнев</p> <p>Ю.А. Маглинец</p>

Рисунок В.1 — Акт об использовании результатов проектирования в рамках бакалаврской работы

ПРИЛОЖЕНИЕ Г

Результат проверки в системе «АНТИПЛАГИАТ»

Министерство образования и науки Российской Федерации
Федеральное государственное автономное образовательное
учреждение высшего образования
«Сибирский федеральный университет»

НАУЧНАЯ БИБЛИОТЕКА

660049, Красноярск, пр. Свободный, 79/10, тел. (3912) 2-912-820, факс (3912) 2-912-773
E-mail: bik@sfu-kras.ru

ОТЧЕТ о результатах проверки в системе «АНТИПЛАГИАТ»

Автор: Непомнящих Алексей Владимирович

Заглавие: Разработка сервиса мониторинга неоднородной структуры сельскохозяйственных земель

Вид документа: Выпускная квалификационная работа бакалавра

По результатам проверки оригинальный текст составляет 72,68%

Рисунок В.1 — Результат проверки в системе «АНТИПЛАГИАТ»