

На правах рукописи



Рылов Сергей Александрович

МЕТОДЫ И АЛГОРИТМЫ СЕГМЕНТАЦИИ
МУЛЬТИСПЕКТРАЛЬНЫХ СПУТНИКОВЫХ ИЗОБРАЖЕНИЙ
ВЫСОКОГО ПРОСТРАНСТВЕННОГО РАЗРЕШЕНИЯ

05.13.18 – математическое моделирование, численные методы
и комплексы программ

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Новосибирск – 2016

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институте вычислительных технологий Сибирского отделения Российской академии наук, г. Новосибирск.

Научный руководитель: кандидат физико-математических наук,
доцент **Пестунов Игорь Алексеевич**

Официальные оппоненты: **Пяткин Валерий Павлович**,
доктор технических наук, профессор,
ИВМиМГ СО РАН, г. Новосибирск,
заведующий лабораторией

Федоров Роман Константинович,
кандидат технических наук,
ИДСТУ СО РАН, г. Иркутск,
ведущий научный сотрудник

Ведущая организация: федеральное государственное бюджетное
образовательное учреждение высшего
образования «Сибирский государственный
университет геосистем и технологий»,
г. Новосибирск

Защита состоится «27» января 2017 г. в 15:00 на заседании диссертационного совета ДМ 003.046.01 на базе Федерального государственного бюджетного учреждения науки Института вычислительных технологий Сибирского отделения Российской академии наук по адресу: 630090, г. Новосибирск, пр. Академика Лаврентьева, 6, конференц-зал ИВТ СО РАН.

С диссертацией можно ознакомиться в библиотеке и на сайте Федерального государственного бюджетного учреждения науки Института вычислительных технологий Сибирского отделения Российской академии наук:
<http://www.ict.nsc.ru/ru/structure/discouncil/rilov-sa-0>

Автореферат разослан «__» ноября 2016 г.

Ученый секретарь
диссертационного совета
к.ф.-м.н.

 Лебедев А.С.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность. В настоящее время активно развиваются средства и технологии дистанционного зондирования Земли (ДЗЗ) из космоса. С каждым годом растет число запускаемых спутников с мультиспектральными сенсорами высокого (5 м и лучше) пространственного разрешения: WorldView-2/3, GeoEye-1, Pleiades-1A/1B, KazEOSat-1, Ресурс-П (1/2/3) и др. Спутниковые данные незаменимы при решении задач, связанных с оперативным мониторингом протяженных или труднодоступных территорий. Увеличение объема и информативности получаемых данных ДЗЗ способствует расширению круга решаемых с их помощью практических задач (мониторинг состояния окружающей среды, инвентаризация сельскохозяйственных угодий, лесопользование, территориальное планирование, мониторинг и прогнозирование последствий чрезвычайных ситуаций и др.). Однако существенным фактором, ограничивающим применение мультиспектральных спутниковых изображений высокого разрешения, является отсутствие приемлемого инструментария для их автоматизированного анализа.

Одним из основных этапов анализа спутниковых изображений является сегментация, которая заключается в разбиении изображения на непересекающиеся области на основе схожести их спектральных, текстурных или других характеристик. Методы сегментации позволяют выделять на спутниковых изображениях области, соответствующие различным типам природных и антропогенных объектов.

Один из наиболее распространенных подходов к сегментации мультиспектральных изображений основан на использовании алгоритмов кластеризации в пространстве спектральных признаков. Это связано с тем, что на практике априорные сведения о вероятностных характеристиках классов, а также обучающие выборки классов, как правило, отсутствуют. Широко используемые для этих целей и включенные в состав популярных программных пакетов (ERDAS Imagine, ENVI, ArcGIS, SNAP и др.) алгоритмы кластеризации (К-средних, ISODATA) основаны на предположении о нормальном виде плотности распределения искомых классов и зачастую не обеспечивают требуемой достоверности результатов при анализе спутниковых изображений. С другой стороны, более подходящие в данном случае непараметрические алгоритмы, способные выделять кластеры произвольной формы, а также алгоритмы, позволяющие получать иерархическое представление данных, не получили распространения из-за неприемлемо высокой вычислительной трудоемкости.

Кроме того, при сегментации спутниковых изображений высокого пространственного разрешения существенно возрастает внутриклассовая неоднородность спектральных характеристик объектов и, как следствие, повышается информативность текстурных признаков. При этом применение методов, учитывающих лишь спектральные характеристики, зачастую приводит к чрезмерной раздробленности искомых классов. К настоящему времени разработано большое количество эффективных методов формирования текстурных признаков для полутоновых изображений, однако для

мультиспектральных изображений проблема спектрально-текстурной сегментации остается открытой.

Таким образом, актуальной является задача разработки вычислительно эффективных (позволяющих обрабатывать спутниковые изображения на персональном компьютере в диалоговом режиме) непараметрических алгоритмов кластеризации и новых методов спектрально-текстурной сегментации для обработки мультиспектральных спутниковых изображений высокого пространственного разрешения.

Целью диссертационной работы является разработка вычислительно эффективных методов и алгоритмов сегментации мультиспектральных спутниковых изображений высокого пространственного разрешения.

Для достижения данной цели требовалось решить следующие задачи.

1. Исследовать возможность применения существующих алгоритмов кластеризации и методов спектрально-текстурной сегментации для обработки мультиспектральных спутниковых изображений высокого пространственного разрешения.
2. Разработать, теоретически обосновать и программно реализовать вычислительно эффективные непараметрические алгоритмы кластеризации, позволяющие работать в условиях малой априорной информации о вероятностных характеристиках классов.
3. Разработать и программно реализовать вычислительно эффективный метод спектрально-текстурной сегментации мультиспектральных спутниковых изображений высокого пространственного разрешения.
4. Провести экспериментальное исследование разработанных методов и алгоритмов на модельных и реальных данных.

На защиту выносятся следующие положения, соответствующие трем пунктам (3, 4, 5) паспорта специальности 05.13.18 – «Математическое моделирование, численные методы и комплексы программ» (технические науки).

1. Разработанный в рамках сеточного подхода алгоритм кластеризации ССА и основанный на нем ансамблевый алгоритм ECCA для сегментации мультиспектральных изображений.
2. Иерархический сеточный алгоритм кластеризации НСА и основанный на нем ансамблевый алгоритм НЕСА для сегментации мультиспектральных изображений.
3. Метод спектрально-текстурной сегментации мультиспектральных спутниковых изображений высокого пространственного разрешения ESEG.
4. Комплекс программ «ECCA-Pack» для обработки и анализа мультиспектральных спутниковых данных, в котором реализованы разработанные автором методы и алгоритмы сегментации изображений.
5. Методы решения задач оперативного мониторинга паводковой ситуации и картографирования типов растительности по данным спутниковой съемки высокого пространственного разрешения с помощью разработанных методов и алгоритмов сегментации изображений.

Научная новизна диссертационной работы состоит в следующем.

1. Предложены новые вычислительно эффективные непараметрические алгоритмы кластеризации данных, разработанные в рамках сеточного и ансамблевого подходов и позволяющие выделять кластеры разной формы, размера и плотности. Ансамблевый подход впервые применен для повышения устойчивости и качества результатов сеточного алгоритма кластеризации.
2. Предложены новые вычислительно эффективные иерархические алгоритмы кластеризации данных на основе сеточного и ансамблевого подходов, обеспечивающие разделение пересекающихся кластеров. Впервые предложен вычислительно эффективный метод построения ансамбля иерархических разбиений в рамках сеточного подхода.
3. Предложен новый метод описания мультиспектральной текстуры, не требующий введения единой метрики в пространстве разнородных спектрально-текстурных признаков. На его основе разработан вычислительно эффективный алгоритм спектрально-текстурной сегментации для мультиспектральных изображений.
4. Предложен новый метод автоматического выделения водных объектов на мультиспектральных спутниковых изображениях высокого пространственного разрешения, разработанный на основе предложенных алгоритмов кластеризации.

Практическая значимость полученных результатов обусловлена следующим. Разработанные методы и алгоритмы сегментации мультиспектральных спутниковых изображений превосходят ранее известные по качеству и скорости работы, что позволяет повысить эффективность автоматизированной обработки спутниковых данных высокого пространственного разрешения при решении прикладных задач.

Основные результаты работы были использованы при выполнении проектов РФФИ (№№ 14-07-31320-мол_а, 11-07-12083-офи_м, 11-07-00202-а, 13-07-12202-офи_м, 13-04-90446-Укр_ф_а), партнерского интеграционного проекта СО РАН № 74 и проекта РНФ № 14-14-00453.

Результаты диссертационной работы используются в Сибирском центре ФГБУ «НИЦ «Планета» (для оперативного мониторинга паводковой ситуации), а также в Центральном сибирском ботаническом саду СО РАН (для картографирования типов растительности по данным спутниковой съемки), что подтверждено актами о внедрении.

Методология исследования опирается на современные информационно-вычислительные технологии. При решении поставленных задач использовались методы математической статистики, кластерного анализа, распознавания образов, обработки цифровых изображений, компьютерного моделирования, а также современные ГИС-технологии.

Экспериментальные исследования проводились с использованием программного комплекса, разработанного автором.

Достоверность результатов обеспечивается корректным применением разработанных методов и алгоритмов и подтверждается проведенными экспериментальными исследованиями на модельных и прикладных задачах.

Представление работы. Результаты работы были представлены на следующих научных мероприятиях: Всероссийской конференции молодых ученых по математическому моделированию и информационным технологиям (Новосибирск, 2011, 2012; Томск, 2013; Тюмень, 2014; Красноярск, 2015); Международном научном конгрессе «ГЕО-Сибирь» (Новосибирск, 2012, 2013, 2015, 2016); Всероссийской открытой конференции «Современные проблемы дистанционного зондирования Земли из космоса» (Москва, 2013-2015); Всероссийской конференции «Обработка пространственных данных и дистанционный мониторинг природной среды и масштабных антропогенных процессов» (Барнаул, 2013); Международной научной конференции «Региональные проблемы дистанционного зондирования Земли» (Красноярск, 2014-2015); Российской конференции с международным участием «Распределенные информационно-вычислительные ресурсы» (Новосибирск, 2014); Open German-Russian Worokshop on Pattern Recognition and Image Understanding (Germany, Koblenz, 2014); Всероссийской конференции «Обработка пространственных данных в задачах мониторинга природных и антропогенных процессов» (с. Усть-Сема, Республика Алтай, 2015); Международной научно-практической конференции «Информационные технологии, системы и приборы в АПК – АГРОИНФО-2015» (НСО, п. Краснообск, 2015); научно-методическом семинаре «Информационно-вычислительные технологии в задачах поддержки принятия решений» в ИВТ СО РАН (Новосибирск, 2012-2016); объединенном семинаре (СЦ ФГБУ «НИЦ «Планета», ИВМиМГ СО РАН) «Дистанционное зондирование и цифровая обработка изображений» (Новосибирск, 2015); объединенном семинаре «Информационно-вычислительные технологии» в ИВТ СО РАН (Новосибирск, 2016).

Публикации. По теме диссертации опубликовано 18 работ, в том числе 4 статьи в изданиях, рекомендованных ВАК, 6 – в других рецензируемых изданиях, 8 – в трудах международных и всероссийских конференций. Получено 3 свидетельства о государственной регистрации программ для ЭВМ.

Личный вклад автора. Автор принимал активное участие в постановке задач и интерпретации результатов. Сеточные алгоритмы кластеризации ССА и ЕССА разработаны автором совместно с Куликовой Е.А. и Бериковым В.Б. Все остальные представленные в диссертации методы и алгоритмы разработаны автором лично. Программная реализация, выбор алгоритмических решений, проведение численных экспериментов, работы по апробации и тестированию разработанных алгоритмов выполнены автором лично.

Структура и объем диссертации. Работа состоит из введения, пяти глав, заключения, списка литературы из 151 наименования и пяти приложений. Полный объем диссертации составляет 135 страниц, включая 40 рисунков и 4 таблицы.

СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Во введении обоснована актуальность работы, сформулирована цель и поставлены задачи исследования. Показаны научная новизна и практическая значимость полученных результатов, представлены основные положения, выносимые на защиту.

Глава 1 посвящена анализу современного состояния проблемы сегментации мультиспектральных спутниковых изображений высокого пространственного разрешения. На основе проведенного анализа делается вывод об актуальности создания вычислительно эффективных непараметрических алгоритмов кластеризации и новых методов спектрально-текстурной сегментации, обеспечивающих высокое качество обработки спутниковых изображений высокого пространственного разрешения.

В главе 2 предложены вычислительно эффективные непараметрические алгоритмы кластеризации ССА и ЕССА для сегментации мультиспектральных спутниковых изображений, разработанные в рамках комбинации плотностного и сеточного подходов.

Задача кластеризации рассматривается в рамках вероятностно-статистического подхода: предполагается, что множество векторов признаков кластеризуемых объектов X является реализацией некоторой многомерной случайной величины, плотность распределения которой неизвестна, и отсутствует какая-либо априорная информация о ее параметрическом виде. Для непараметрической оценки плотности использовался сеточный подход.

Для описания формальной постановки задачи кластеризации в рамках сеточного подхода в разделе 2.1 введены необходимые определения.

Пусть множество X состоит из векторов, лежащих в d -мерном пространстве признаков: $X = \{x_i = (x_i^{(1)}, \dots, x_i^{(d)}) \in \mathbb{R}^d, i = \overline{1, N}\}$. Векторы x_i лежат в прямоугольном гиперпараллелепипеде $\Omega = [l^{(1)}, r^{(1)}] \times \dots \times [l^{(d)}, r^{(d)}]$, где $l^{(j)} = \min_{x_i \in X} x_i^{(j)}, r^{(j)} = \max_{x_i \in X} x_i^{(j)}$.

Сеточная структура определяется как разбиение пространства признаков гиперплоскостями: $x^j = (r^j - l^j) \cdot i/m + l^j, i = 0, \dots, m$, где m – число разбиений Ω по каждой размерности. Минимальным элементом этой структуры является *клетка*. Множество клеток, смежных с B , обозначается через A_B . *Плотность* D_B клетки B определяется как число элементов множества X , попавших в клетку B . Непустая клетка B_i *непосредственно связна* с непустой клеткой B_j ($B_i \rightarrow B_j$), если B_j – максимальная по номеру клетка, удовлетворяющая условиям: $B_j = \arg \max_{B_k \in A_{B_i}} D_{B_k}$ и $D_{B_j} \geq D_{B_i}$. Непустые клетки B_i и B_j *непосредственно связны* ($B_i \leftrightarrow B_j$), если $B_i \rightarrow B_j$ или $B_j \rightarrow B_i$. Непустые клетки B_i и B_j *связны* ($B_i \sim B_j$), если существуют k_1, \dots, k_l такие, что $k_1 = i, k_l = j$ и для всех $p = 1, \dots, l - 1$ выполнено $B_{k_p} \leftrightarrow B_{k_{p+1}}$. Введение отношения связности является отношением эквивалентности и порождает разбиение множества непустых клеток на *компоненты связности* (классы эквивалентности) $\{G_1, \dots, G_S\}$. Представителем компоненты связности G называется максимальная по номеру клетка $Y(G)$, удовлетворяющая условию: $Y(G) = \arg \max_{B \in G} D_B$.

Выделенные компоненты связности соответствуют одномодовым кластерам, а их представители – модам плотности этих кластеров.

В разделе 2.2 предложен сеточный алгоритм кластеризации ССА, основанный на использовании введенных компонент связности, который объединяет их в кластеры в соответствии с изложенным ниже критерием.

Смежные компоненты связности G_i и G_j *непосредственно связны* ($G_i \leftrightarrow G_j$), если существует набор клеток $P_{ij} = \{Y(G_i) = B_{k_1}, \dots, B_{k_t}, B_{k_{t+1}}, \dots, B_{k_l} = Y(G_j)\}$ такой, что: 1) для всех $t = 1, \dots, l-1$ клетка $B_{k_t} \in G_i \cup G_j$ и $B_{k_t}, B_{k_{t+1}}$ – смежные клетки; 2) $\min_{B_{k_t} \in P_{ij}} D_{B_{k_t}} / \min(D_{Y(G_i)}, D_{Y(G_j)}) > T$, где $T \in [0; 1]$ – порог объединения. Компоненты связности G_i и G_j *связны* ($G_i \sim G_j$), если существуют k_1, \dots, k_l такие, что $k_1 = i$, $k_l = j$ и для всех $p = 1, \dots, l-1$ выполнено $G_{k_p} \leftrightarrow G_{k_{p+1}}$. Кластер определяется как максимальное множество попарно связных компонент связности.

Предложенный алгоритм ССА позволяет выделять кластеры сложной формы, разного размера и плотности (рисунок 1) и обладает линейной вычислительной сложностью, что подтверждено экспериментальными исследованиями. Однако результаты его работы демонстрируют сильную зависимость от параметра t , определяющего размер элементов сеточной структуры.

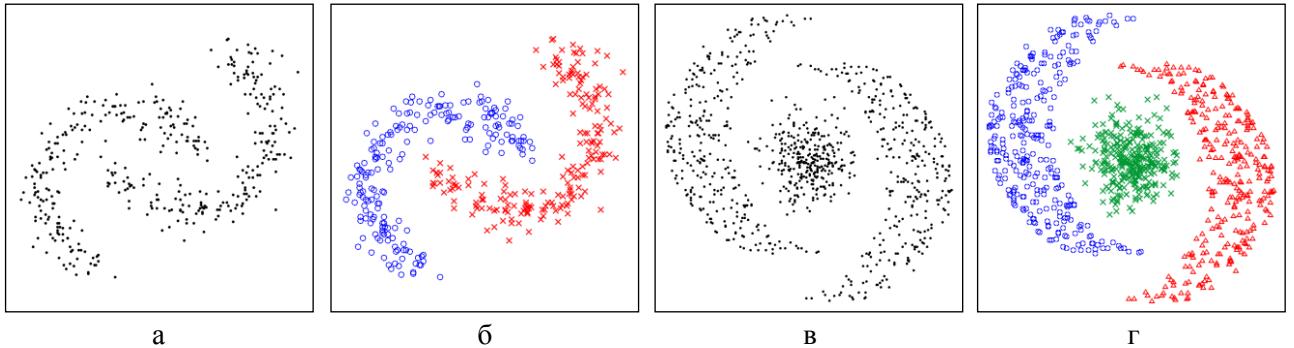


Рисунок 1 – Модельные данные (а, в); результаты кластеризации алгоритмом ССА (б, г)

Известно, что повысить качество результатов кластеризации и их устойчивость к изменению параметров возможно с помощью применения ансамблевого подхода. Он заключается в создании нескольких вариантов разбиения данных и дальнейшем формировании согласованного решения. Однако большинство методов построения ансамблевого решения являются вычислительно трудоемкими.

В разделе 2.4 представлен подход к построению ансамбля сеточных алгоритмов кластеризации с помощью согласованной матрицы различий H , которая определяется следующим образом.

Пусть с помощью некоторого алгоритма кластеризации $\mu(\theta)$, зависящего от вектора параметров $\theta \in \Theta$ (где Θ – множество допустимых значений параметров), получен набор частных решений $\mathcal{Q}_L = \{\mathcal{Q}^{(1)}, \dots, \mathcal{Q}^{(l)}, \dots, \mathcal{Q}^{(L)}\}$, где $\mathcal{Q}^{(l)}$ – l -й вариант разбиения. Обозначим через $B(\mathcal{Q}^{(l)})$ бинарную матрицу размера $N \times N$, которая для l -й группировки определяется по формуле

$$B_{ij}(\mathcal{Q}^{(l)}) = \begin{cases} 0, & \text{если объекты } i \text{ и } j \text{ отнесены в один кластер;} \\ 1, & \text{иначе.} \end{cases}$$

Тогда, согласованная матрица различий $H = \{H_{ij}\}$, $i, j = 1, \dots, N$ для L частных решений $\{\mathcal{Q}^{(1)}, \dots, \mathcal{Q}^{(l)}, \dots, \mathcal{Q}^{(L)}\}$ вычисляется по формуле

$$H_{ij} = \frac{1}{L} \sum_{l=1}^L B_{ij}(\mathcal{Q}^{(l)}). \quad (1)$$

В разделе 2.5 предложен ансамблевый алгоритм кластеризации ECCA, основанный на использовании матрицы H и формирующий согласованное решение для набора разбиений, полученных сеточным алгоритмом ССА при различных значениях параметра сетки. При этом объектами кластеризации являются представители компонент связности $\{Y(G_1), \dots, Y(G_s)\}$, что позволяет установить соответствие между кластерами из разных разбиений и избежать проблемы высокой трудоемкости. Для нахождения коллективного решения к построенной согласованной матрице различий применяется иерархический алгоритм кластеризации невзвешенной средней связи. В результате формируется иерархическая структура (дендrogramma), базовыми элементами которой являются компоненты связности.

Представленные в разделах 2.3 и 2.6 результаты экспериментальных исследований на модельных данных и реальных изображениях показали способность алгоритмов ССА и ECCA выделять многомодовые кластеры сложной структуры даже в присутствии «шума». Показано, что использование ансамблевого подхода повышает устойчивость получаемых результатов к изменению параметров алгоритма, что значительно облегчает их настройку. Вычислительная эффективность разработанных алгоритмов позволяет применять их для обработки мультиспектральных спутниковых изображений большого размера (до 100 миллионов пикселей) в диалоговом режиме.

На рисунке 2 представлен результат сегментации цветного изображения размера 640×480 пикселей алгоритмом ECCA. Время обработки¹ составило 0.04 с. Также на рисунке представлен результат сегментации этого изображения алгоритмом кластеризации ISODATA, который не позволил отделить представленный на изображении самолет от фона даже при увеличении числа выделяемых кластеров.

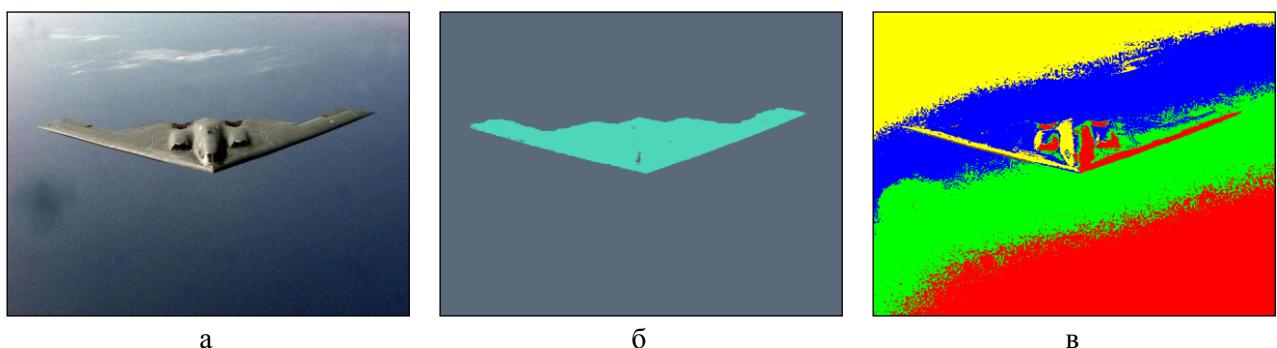


Рисунок 2 – Исходное изображение (а); результаты сегментации алгоритмами ECCA (б) и ISODATA (в)

¹ Все вычисления выполнялись на ПЭВМ с четырехядерным процессором Intel Core i7, 3.2 ГГц

Глава 3 посвящена разработке и исследованию вычислительно эффективных сеточных иерархических алгоритмов кластеризации НСА и НЕСА.

Иерархическое представление удобно при интерпретации результатов кластеризации. Однако существующие иерархические алгоритмы кластеризации не позволяют разделять пересекающиеся кластеры и характеризуются высокой вычислительной трудоемкостью, поэтому они не применяются для обработки спутниковых изображений.

Для построения иерархии в разделе 3.1 введена специальная метрика между элементами сеточной структуры, основанная на непараметрической оценке плотности распределения данных. При этом базовыми элементами иерархии выступают введенные во второй главе компоненты связности, число которых мало относительно объема исходных данных.

Расстояние h_{ij} между парой смежных компонент связности G_i и G_j определяется по формуле

$$h_{ij} = \min_{P_{ij} \in \mathfrak{R}_{ij}} \left[1 - \min_{B_{k_t} \in P_{ij}} D_{B_{k_t}} / \min(D_{Y(G_i)}, D_{Y(G_j)}) \right], \quad (2)$$

где $\mathfrak{R}_{ij} = \{P_{ij}\}$ – множество всех цепочек, связывающих представителей компонент связности $P_{ij} = \langle Y(G_i) = B_{k_1}, \dots, B_{k_t}, B_{k_{t+1}}, \dots, B_{k_l} = Y(G_j) \rangle$ таких, что для всех $t = 1, \dots, l-1$: 1) $B_{k_t} \in G_i \cup G_j$; 2) $B_{k_t}, B_{k_{t+1}}$ – смежные клетки.

Матрица расстояний между компонентами связности $\{\tilde{h}_{ij}\}$ строится на основе матрицы расстояний между смежными компонентами связности $\{h_{ij}\}$ следующим образом. Пусть $\Theta_{ij} = \{\mathcal{Q}_{ij}\}$ – множество цепочек, элементами которых являются компоненты связности, $\mathcal{Q}_{ij} = \langle G_i = G_{k_1}, \dots, G_{k_t}, G_{k_{t+1}}, \dots, G_{k_l} = G_j \rangle$ таких, что для всех $t = 1, \dots, l-1$ компоненты $G_{k_t}, G_{k_{t+1}}$ – смежные. Тогда расстояние между компонентами G_i и G_j вычисляется по формуле

$$\tilde{h}_{ij} = \min_{\mathcal{Q}_{ij} \in \Theta_{ij}} [\max_t h_{k_t, k_{t+1}}]. \quad (3)$$

Утверждение 3.1. Введенное расстояние \tilde{h}_{ij} обладает свойством ультраметрики на множестве компонент связности.

Это означает, что \tilde{h}_{ij} является метрикой, для которой выполняется усиленное неравенство треугольника: $\tilde{h}_{ij} \leq \max(\tilde{h}_{ik}, \tilde{h}_{kj}), \forall i, j, k$. Известно, что существует однозначное соответствие между матрицами расстояний со свойством ультраметрики и дендрограммами, то есть такие матрицы описывают иерархические разбиения.

Процедура получения ультраметрики $\{\tilde{h}_{ij}\}$ из матрицы расстояний смежных объектов $\{h_{ij}\}$ в литературе известна как операция минимального транзитивного замыкания² (min-transitive closure). Для ее реализации, как правило, применяются вычислительно трудоемкие алгоритмы со сложностью $O(n^3)$ (для матрицы размера $n \times n$). Из следующего утверждения следует, что операцию минимального транзитивного замыкания возможно выполнить с помощью алгоритма построения дендрограммы методом ближайшего соседа, который можно реализовать с вычислительной сложностью $\mathcal{O}(n^2)$.

² Mirzaei A., Rahmati M. A novel hierarchical-clustering-combination scheme based on fuzzy-similarity relations // IEEE Transactions on Fuzzy Systems. – 2010. – Vol. 18. – No. 1. – P. 27-39.

Утверждение 3.2. Дендрограмма, соответствующая ультраметрике $\{\tilde{h}_{ij}\}$, совпадает с дендрограммой, полученной с помощью применения метода ближайшего соседа к матрице расстояний $\{h_{ij}\}$.

В разделе 3.3 предложен сеточный алгоритм кластеризации НСА, который позволяет выделять иерархическую структуру данных на основе введенной метрики. Алгоритм НСА состоит из следующих основных этапов.

- 1) Формирование сеточной структуры. Для каждого вектора данных $x_i \in X$ определяется содержащая его клетка и вычисляются плотности клеток.
- 2) Выделение компонент связности $\{G_1, \dots, G_s\}$ и их клеток представителей $\{Y(G_1), \dots, Y(G_s)\}$.
- 3) Формирование матрицы расстояний $\{h_{ij}\}$ между смежными компонентами связности в соответствии с определением.
- 4) Применение алгоритма построения дендрограммы методом ближайшего соседа к матрице расстояний $\{h_{ij}\}$.

Предложенный алгоритм НСА обладает линейной вычислительной сложностью и позволяет выделять иерархическую структуру данных, при этом он способен эффективно разделять пересекающиеся в пространстве признаков кластеры. Однако результаты его работы, как и в случае с алгоритмом ССА, значительно зависят от значения параметра сетки m .

В разделе 3.5 представлен способ построения ансамбля иерархических разбиений, который предлагается использовать для повышения устойчивости результатов кластеризации алгоритма НСА.

Для комбинации иерархических разбиений удобно использовать представление дендрограмм в виде матриц расстояний (которые обладают свойством ультраметрики). Пусть для анализируемых данных имеется набор иерархических разбиений, представленных в виде матриц расстояний $\{M^{(1)}, \dots, M^{(L)}\}$. Определим *согласованную матрицу расстояний* $R = \{R_{ij}\}$ по формуле

$$R_{ij} = \frac{1}{L} \sum_{l=1}^L M_{ij}^{(l)}. \quad (4)$$

Согласованная матрица расстояний R может не обладать свойством ультраметрики. Однако задача нахождения ближайшей к ней ультраметрики является NP-трудной оптимизационной задачей. Поэтому для получения согласованного иерархического результата к матрице R применяется один из агломеративных методов построения дендрограммы.

В разделе 3.6 предложен ансамблевый алгоритм кластеризации НЕСА, позволяющий согласовать иерархические результаты кластеризации, полученные сеточным алгоритмом НСА. В результате применения алгоритма НСА к исходным данным с различными значениями параметра сетки $m \in \{m_1, \dots, m_L\}$ получается L матриц расстояний между компонентами связности $\{\tilde{h}^{(1)}, \dots, \{\tilde{h}^{(L)}\}\}$. На их основе строится согласованная матрица расстояний $\{R_{ij}\}$:

$$R_{ij} = \frac{1}{L} \sum_{k=1}^L \tilde{h}^{(k)} (G_i^{(k)}, G_j^{(k)}), \quad (5)$$

где $G_i^{(k)}$ является такой компонентой связности на сетке с номером k , в которой находится клетка-представитель компоненты $G_i^{(L)}$, полученной на сетке с наибольшим номером (L). Для построения окончательной дендрограммы к полученной согласованной матрице расстояний применяется метод невзвешенной средней связи.

Представленные в разделах 3.4 и 3.7 результаты экспериментальных исследований алгоритмов НСА и НЕСА на модельных и реальных данных показали способность предложенных алгоритмов выделять кластеры сложной формы, разного размера и плотности, а также разделять пересекающиеся классы. Иерархическое представление значительно упрощает интерпретацию получаемых результатов кластеризации. Результаты моделирования подтвердили, что использование ансамблевого подхода позволяет существенно повысить качество результатов и их устойчивость к изменению параметра сетки. График на рисунке 3 показывает увеличение точности кластеризации с ростом числа используемых сеток (элементов ансамбля) для линейно неразделимой модели «Бананы» (рисунок 1,а).

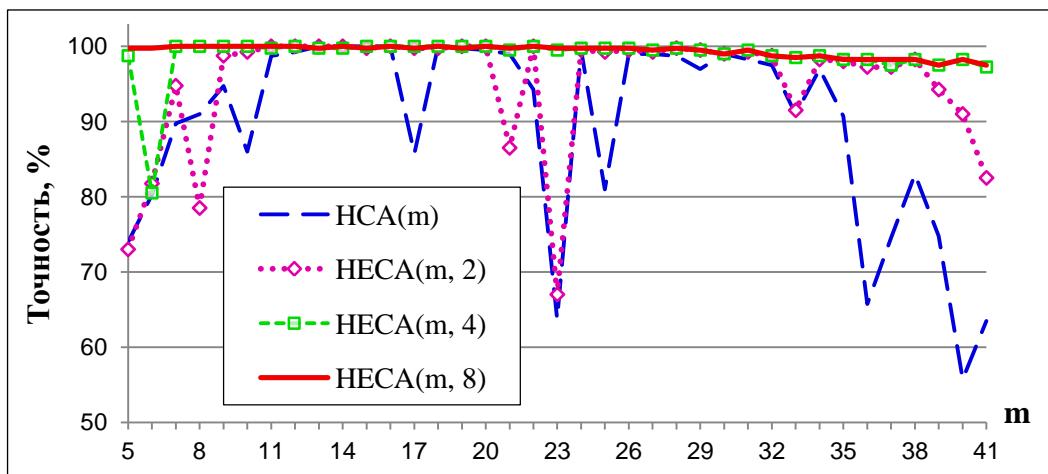


Рисунок 3 – Зависимость точности кластеризации от значения параметра сетки (m) для алгоритмов НСА и НЕСА при различном числе используемых сеток (2, 4, 8)

В диссертации приведено сравнение алгоритмов НСА и НЕСА с известными алгоритмами кластеризации (K-means, EM, DBSCAN, OPTICS, DeLiClu, SLINK), реализованными в программном пакете ELKI³. Показано значительное преимущество алгоритмов НСА и НЕСА в скорости работы.

На рисунке 4 представлены результаты кластеризации двумерной модели⁴, состоящей из восьми кластеров, различающихся по форме, размеру и плотности. Ансамблевый алгоритм НЕСА позволил успешно выделить все представленные кластеры с точностью 99.3%. С другой стороны, ни один алгоритм из пакета ELKI не смог правильно выделить все кластеры. Наилучшие результаты показал плотностной иерархический алгоритм DeLiClu с точностью 94.6%. Иерархический алгоритм ближайшего соседа (SLINK) продемонстрировал точность кластеризации 90.3%.

³ URL: <http://elki.dbs.ifi.lmu.de/>

⁴ URL: <https://drive.google.com/open?id=0ByK9GtU5ExExRnZwdFNmRHRWdFk>

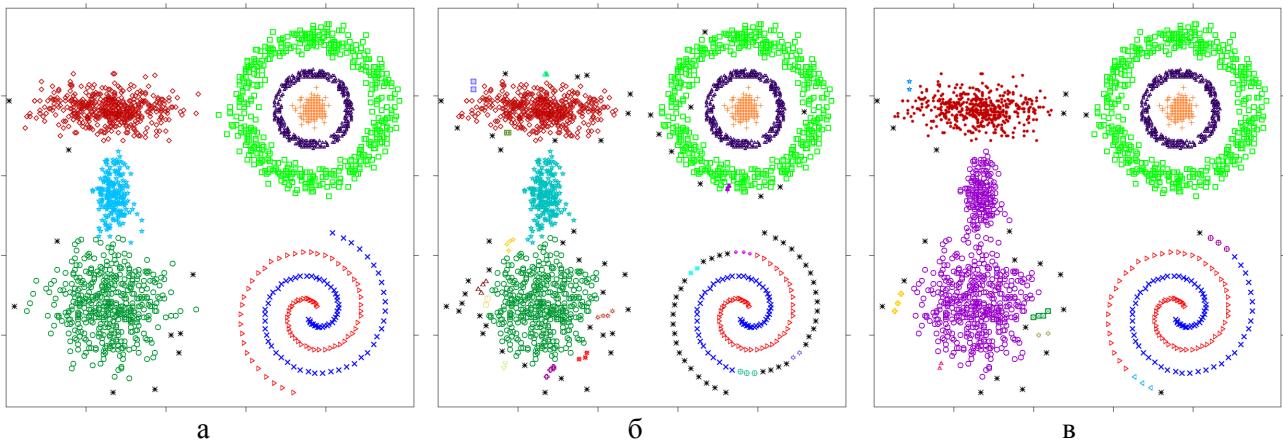


Рисунок 4 – Результаты кластеризации модельных данных, состоящих из восьми классов, алгоритмами HECA (а), DeLiClu (б) и SLINK (в)

На рисунке 5 представлен результат сегментации мультиспектрального изображения, полученного со спутника WorldView-2, алгоритмом НСА. Размер изображения составляет 2048×2048 пикселей. При сегментации использовалось четыре канала: 1, 2, 4, 7. Время обработки составило 0.28 с.

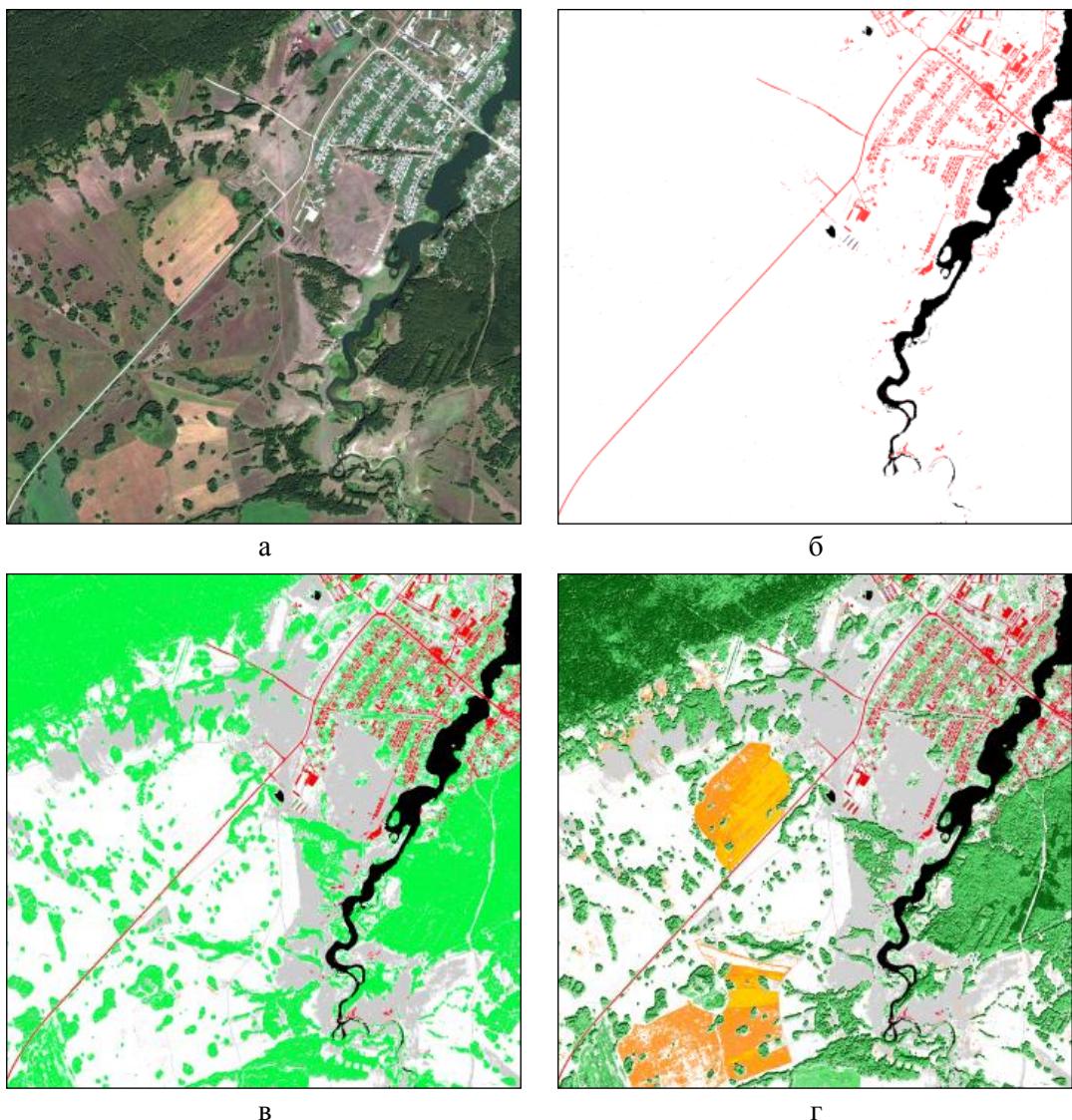


Рисунок 5 – Изображение со спутника WorldView-2 (RGB-композит, каналы 5, 3, 2) (а); результат кластеризации алгоритмом НСА на различных уровнях иерархии (б-г)

В главе 4 предложен новый метод описания мультиспектральной текстуры, в рамках которого разработан вычислительно эффективный алгоритм спектрально-текстурной сегментации изображений ESEG.

Применительно к мультиспектральным изображениям текстуру можно интерпретировать как характер распределения векторов спектральных яркостей по полю изображения объекта, обусловленный закономерностями взаимного расположения составляющих объект элементов. Для описания спектрально-текстурных характеристик областей изображения предлагается использовать векторы частот встречаемости кластеров, полученных с помощью кластеризации изображения по спектральным признакам одним из предложенных в диссертации алгоритмов. Основная идея метода заключается в том, что в локальных областях изображения для мультиспектральных текстур одного и того же класса процентное содержание пикселей из разных «спектральных» кластеров будет примерно одинаково, а для текстур разных классов будет заметно отличаться. Это наблюдение справедливо для большинства текстур, соответствующих объектам природного происхождения (лес, болото, тундра и т.п.), для которых структурные особенности, такие как пространственная ориентация, не являются важными.

В разделе 4.1 предложен метод описания мультиспектральной текстуры с помощью *векторов спектральных частот*. Пусть картосхема, полученная в результате кластеризации по спектральным признакам, состоит из K кластеров. Тогда для произвольного фрагмента изображения f вектор спектральных частот $z(f) = (z^{(1)}, \dots, z^{(K)})$ определяется следующим образом: $\{z^{(i)} = N_f^{(i)} / N_f, i = \overline{1, K}\}$, где N_f – число всех пикселей фрагмента f , $N_f^{(i)}$ – число пикселей фрагмента f , относящихся к кластеру с номером i . Расстояние между векторами спектральных частот $x = (x^{(1)}, \dots, x^{(K)})$ и $y = (y^{(1)}, \dots, y^{(K)})$ определяется с помощью манхэттенской метрики: $\mu(x, y) = 1 - \sum_{i=1}^K \min(x^{(i)}, y^{(i)})$. Предложенный подход позволяет избежать проблемы введения единой метрики в пространстве разнородных спектрально-текстурных признаков.

В разделе 4.2 представлен алгоритм спектрально-текстурной сегментации ESEG, который состоит из трех основных шагов.

- 1) Формирование векторов спектральных частот. Исходная картосхема разбивается на множество непересекающихся фрагментов размера $h \times h$ пикселей. Для фрагментов вычисляются векторы спектральных частот.
- 2) Кластеризация полученных векторов спектральных частот с помощью алгоритма CLUSTER⁵ с параметром R (максимальный радиус кластера). Пиксели фрагментов изображения относятся к тем кластерам, к которым были отнесены соответствующие им векторы спектральных частот.
- 3) Попиксельная коррекция границ кластеров, основанная на переклассификации всех граничных пикселей с помощью вычисления векторов спектральных частот фрагментов с центром в этих пикселях.

⁵ Jain A.K., Murty M.N., Flynn P.J. Data clustering: a review // ACM computing surveys. – 1999. – Vol. 31. – No. 3. – P. 264-323.

Приведенные в разделе 4.3 результаты экспериментальных исследований на модельных и реальных изображениях показали, что алгоритм ESEG является вычислительно эффективным и обеспечивает выделение текстурных классов на мультиспектральных изображениях.

На рисунке 6 представлен результат сегментации алгоритмом ESEG модельного изображения размера 622×622 пикселя. Изображение составлено из пяти цветных текстур, при этом две из них (верхняя левая и верхняя правая) имеют одинаковый средний цвет. В результате были выделены все классы с точностью 99.5%. Общее время обработки изображения составило 0.1 с.

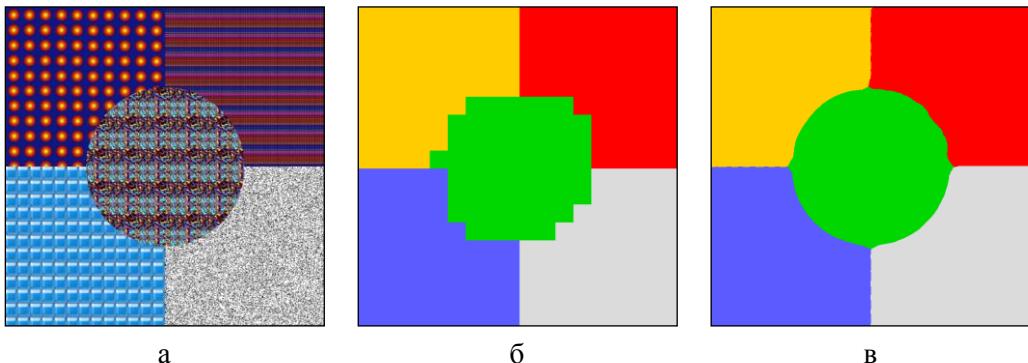


Рисунок 6 – Модельное изображение (а); результат сегментации алгоритмом ESEG до (б) и после (в) коррекции границ

Экспериментальные исследования демонстрируют также эффективность разработанного алгоритма при сегментации мультиспектральных спутниковых изображений высокого пространственного разрешения и способность выделять текстурные объекты, которые оказываются раздробленными при кластеризации только по спектральным признакам. Для тестового изображения, полученного со спутника WorldView-2, методом визуально-инструментального дешифрирования была построена картосхема для информационного класса «лес». Точность сегментации алгоритмом ESEG для данного класса составила 97%. По результатам сегментации спутникового изображения (WorldView-2) Карадагского природного заповедника и данным полевых исследований были выявлены лесные массивы различного породного состава и степени сомкнутости и составлена карта его растительного покрова.

В разделе 4.4 на основе алгоритма ESEG предложена многоэтапная схема сегментации спутниковых изображений высокого пространственного разрешения, позволяющая учитывать объекты, выделенные с помощью различных специализированных методов. В процессе обработки для всех выделенных объектов формируется пространственная маска, пиксели которой исключаются из рассмотрения на этапе выполнения спектрально-текстурной сегментации. Разработанная схема позволяет получать качественные результаты сегментации для объектов различной природы.

На рисунке 7 представлен пример сегментации мультиспектрального изображения размера 2048×2048 пикселей, полученного со спутника WorldView-2. На снимке были предварительно выделены следующие объекты: вода, песок и искусственные объекты (здания, сооружения, дороги). Сегментация по спектральным признакам выполнялась алгоритмом ECCA.

В результате была получена картосхема, содержащая 52 кластера, для большинства из которых нет соответствия среди реальных объектов. На рисунке 7,в. представлен результат сегментации изображения алгоритмом ESEG с учетом выделенных объектов. Время работы алгоритма ECCA составило 0.4 с, а алгоритма ESEG – 0.3 с.

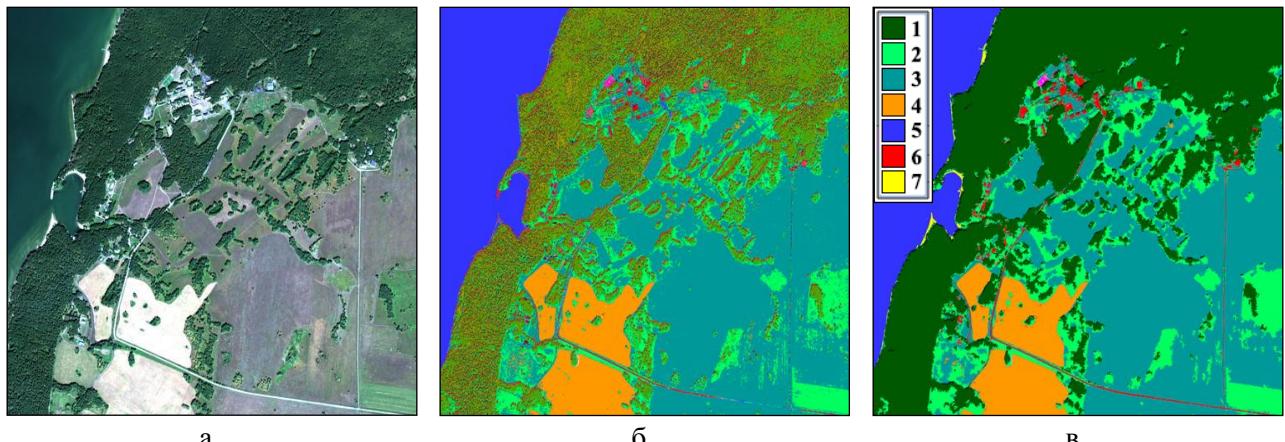


Рисунок 7 – Изображение со спутника WorldView-2 (RGB-композит, каналы 5, 3, 2) (а); результат сегментации алгоритмом ECCA (б); результат спектрально-текстурной сегментации с учетом выделенных объектов (в) (1 – лесные территории, 2 – трава, 3 – незасеянные поля и пустоши, 4 – с/х посевы, 5 – вода, 6 – здания и дороги, 7 – песок)

Глава 5 посвящена описанию разработанного программного комплекса и решению практических задач.

В разделе 5.1 представлена структура и основные функциональные блоки созданного программного приложения «ECCA-Pack», которое предназначено для сегментации мультиспектральных спутниковых изображений. Приложение реализовано на языке программирования Java, обладает пользовательским графическим интерфейсом и включает в себя алгоритмы кластеризации и методы сегментации, разработанные автором.

В разделе 5.2 рассмотрена задача оперативного мониторинга паводковой ситуации. Применение известных методов автоматического выделения водных объектов к снимкам высокого разрешения, полученных с отечественных спутников Ресурс-П и Канопус-В, в оперативной работе оказалось неудовлетворительным: к водным объектам часто относились сильно увлажненные почвы, тени от облаков и некоторые антропогенные объекты. На основе алгоритма кластеризации ECCA был разработан новый метод автоматического выделения водных объектов на мультиспектральных спутниковых изображениях. В отличие от известных методов, предложенный подход позволяет выделять водные объекты на снимках не только низкого и среднего, но и высокого пространственного разрешения. При этом время обработки спутниковых изображений размером десятки миллионов пикселей составляет всего несколько секунд. Разработанный метод используется в Сибирском центре ФГБУ «НИЦ «Планета» при создании оперативных карт паводковой обстановки по данным с отечественных космических аппаратов Ресурс-П и Канопус-В для потребителей Росгидромета и региональных служб МЧС.

В разделе 5.3 описан метод решения задачи крупномасштабного моделирования структуры степной растительности по данным спутниковой съемки высокого пространственного разрешения с помощью разработанных алгоритмов сегментации. В результате были выявлены закономерности формирования растительного покрова горностепного пояса на уровне отдельных фитоценозов, что имеет большое значение для охраны уникальных реликтовых сообществ и произрастающих в них редких и уникальных видов растений.

В заключении сформулированы основные результаты работы.

1. Разработан и исследован новый сеточный алгоритм кластеризации ССА, позволяющий выделять многомодовые кластеры разной формы, размера и плотности. Показана линейная зависимость времени работы алгоритма от числа элементов обрабатываемых данных. Вычислительная эффективность алгоритма ССА позволяет обрабатывать мультиспектральные спутниковые изображения большого размера (до 100 миллионов пикселей) в диалоговом режиме.
2. Впервые предложен подход к построению ансамбля сеточных алгоритмов кластеризации с помощью согласованной матрицы различий. В рамках этого подхода создан ансамблевый алгоритм кластеризации ЕССА, позволяющий значительно повысить устойчивость результатов работы алгоритма ССА к изменению значений его параметров. Проведенные экспериментальные исследования показали, что алгоритм ЕССА является вычислительно эффективным и обеспечивает высокое качество получаемых результатов.
3. В рамках сеточного подхода предложена метрика, основанная на непараметрической оценке плотности распределения данных. На ее основе разработан и исследован иерархический сеточный алгоритм кластеризации НСА. В отличие от известных иерархических алгоритмов, НСА позволяет разделять пересекающиеся кластеры и обладает линейной вычислительной сложностью, что обеспечивает возможность его применения для сегментации спутниковых изображений большого размера. Получаемое иерархическое представление результатов существенно упрощает их интерпретацию.
4. Предложен оригинальный способ построения ансамбля иерархических разбиений в рамках сеточного подхода. По сравнению с известными методами формирования ансамбля иерархических разбиений, предложенный подход является вычислительно эффективным и применим для обработки мультиспектральных спутниковых изображений. На основе этого подхода создан ансамблевый алгоритм кластеризации НЕСА, позволяющий значительно повысить устойчивость получаемых результатов работы алгоритма НСА. Проведенные экспериментальные исследования показали существенное превосходство НЕСА в скорости работы и качестве получаемых результатов по сравнению с известными алгоритмами кластеризации.
5. Предложен новый метод описания мультиспектральной текстуры, не требующий введения единой метрики в пространстве разнородных спектрально-текстурных признаков. С его помощью разработан и исследо-

дован вычислительно эффективный алгоритм спектрально-текстурной сегментации мультиспектральных изображений (ESEG). На основе созданного алгоритма разработана многоэтапная схема сегментации спутниковых изображений высокого пространственного разрешения, предназначенная для исследования природных и антропогенных объектов.

6. Создан комплекс программ «ECCA-Pack» для обработки и анализа мультиспектральных спутниковых изображений, в котором реализованы разработанные автором методы и алгоритмы.
7. Разработан новый метод автоматического выделения водных объектов на мультиспектральных спутниковых изображениях высокого пространственного разрешения, основанный на использовании ансамблевого алгоритма кластеризации ECCA. Разработанный метод внедрен в СЦ ФГБУ «НИЦ «Планета», где используется при создании оперативных карт паводковой обстановки по данным с отечественных космических аппаратов Ресурс-П и Канопус-В для потребителей Росгидромета и региональных служб МЧС.
8. Программный комплекс «ECCA-Pack» внедрен в Центральном сибирском ботаническом саду СО РАН, где используется для решения задачи картографирования типов растительности по данным спутниковой съемки высокого пространственного разрешения.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

В рецензируемых журналах, рекомендованных ВАК:

1. Пестунов, И.А. Ансамблевый алгоритм кластеризации больших массивов данных / И.А. Пестунов, В.Б. Бериков, Е.А. Куликова, С.А. Рылов // Автометрия. – 2011. – Т. 47. – № 3. – С. 49-58.
[Англ. версия] Pestunov, I.A. Ensemble of clustering algorithms for large datasets / I.A. Pestunov, V.B. Berikov, E.A. Kulikova, S.A. Rylov // Optoelectronics, instrumentation and data processing. – 2011. – Vol. 47. – No. 3. – P. 245-252. (Scopus).
2. Пестунов, И.А. Алгоритмы спектрально-текстурной сегментации спутниковых изображений высокого пространственного разрешения / И.А. Пестунов, С.А. Рылов // Вестник Кемеровского государственного университета. – 2012. – Т. 52. – № 4/2 (52). – С. 104-110.
3. Пестунов, И.А. Иерархические алгоритмы кластеризации для сегментации мультиспектральных изображений / И.А. Пестунов, С.А. Рылов, В.Б. Бериков // Автометрия. – 2015. – Т. 51. – № 4. – С. 12-22.
[Англ. версия] Pestunov, I.A. Hierarchical clustering algorithms for segmentation of multispectral images / I.A. Pestunov, S.A. Rylov, V.B. Berikov // Optoelectronics, instrumentation and data processing. – 2015. – Vol. 51. – No. 4. – P. 329-338. (Scopus).
4. Синявский, Ю.Н. Методы и технология сегментации мультиспектральных изображений высокого разрешения для исследования природных и антропогенных объектов / Ю.Н. Синявский, И.А. Пестунов, С.А. Рылов и др. // Вычислительные технологии. – 2016. – Т. 21. – № 1. – С. 127-140.

В других рецензируемых изданиях:

5. Пестунов, И.А. Сегментация изображений на основе кластеризации в пространстве спектральных и текстурных признаков / И.А. Пестунов, С.А. Рылов // Известия Кыргызского государственного технического университета им. И. Рazzакова. – 2011. – № 24. – С. 266-270.
6. Пестунов, И.А. Сегментация спутниковых изображений высокого разрешения по спектральным и текстурным признакам / И.А. Пестунов, С.А. Рылов // Интерэкспо ГЕО-Сибирь. – 2012. – Т. 1. – № 4. – С. 86-91.
7. Пестунов, И.А. Технология и программный инструментарий для сегментации спутниковых изображений высокого пространственного разрешения / И.А. Пестунов, С.А. Рылов, П.В. Мельников, Ю.Н. Синявский // Интерэкспо ГЕО-Сибирь. – 2013. – Т. 4. – № 1. – С. 202-208.
8. Рылов, С.А. Использование графических процессоров NVIDIA при кластеризации мультиспектральных данных сеточным алгоритмом ССА / С.А. Рылов, И.А. Пестунов // Интерэкспо ГЕО-Сибирь. – 2015. – Т. 4. – № 2. – С. 51-56.
9. Синявский, Ю.Н. Совместная обработка разнородных данных при сегментации спутниковых изображений высокого разрешения / Ю.Н. Синявский, И.А. Пестунов, С.А. Рылов, П.В. Мельников // Интерэкспо ГЕО-Сибирь. – 2015. – Т. 4. – № 2. – С. 57-61.
10. Рылов, С.А. Спектрально-текстурная классификация гиперспектральных изображений высокого пространственного разрешения / С.А. Рылов, П.В. Мельников, И.А. Пестунов // Интерэкспо ГЕО-Сибирь. – 2016. – Т. 4. – № 1. – С. 78-84.

В трудах международных и всероссийских конференций:

11. Пестунов, И.А. Спектрально-текстурная сегментация спутниковых изображений высокого пространственного разрешения / И.А. Пестунов, С.А. Рылов // Сборник трудов XI Всероссийской конференции с участием иностранных ученых «Проблемы мониторинга окружающей среды» (ЕМ-2011). – Кемерово: КемГУ, 2011. – С. 357-361.
12. Харук, В.И. Обнаружение и классификация усыхающих кедровых древостоев по спутниковым данным высокого пространственного разрешения / В.И. Харук, С.А. Рылов и др. // Сборник материалов Международной конференции и школы молодых ученых по измерениям, моделированию и информационным системам для изучения окружающей среды «ENVIROMIS-2014». – Томск: Томский ЦНТИ, 2014. – С. 178-180.
13. Ермаков, Н.Б. Крупномасштабное картографирование структуры степной растительности и выявление сообществ с редкими и уникальными видами растений на территории Южной Сибири с использованием снимков высокого разрешения / Н.Б. Ермаков, С.А. Рылов и др. // Материалы межд. научной конф. «Региональные проблемы дистанционного зондирования Земли». – Красноярск: СФУ, 2014. – С. 224-229.
14. Пестунов, И.А. Метод построения ансамбля сеточных иерархических алгоритмов кластеризации для сегментации спутниковых изображений /

- И.А. Пестунов, С.А. Рылов // Материалы межд. научной конф. «Региональные проблемы дистанционного зондирования Земли». – Красноярск: СФУ, 2014. – С. 215-223.
15. Рылов, С.А. Технология обработки данных с космических аппаратов «Канопус-В», «Песурс-П» и «Метеор-М» для мониторинга и картографирования паводковой ситуации / С.А. Рылов, О.Г. Новгородцева и др. // Материалы II межд. научной конф. «Региональные проблемы дистанционного зондирования Земли». – Красноярск: СФУ, 2015. – С. 207-212.
16. Rylov, S. Hierarchical Ensemble Clustering Algorithm for Multispectral Image Segmentation / S. Rylov, I. Pestunov, V. Berikov // Proceedings 9th Open German-Russian Workshop on Pattern Recognition and Image Understanding (OGRW-2014). – Koblenz: University of Koblenz-Landau, 2015. – P. 123-127. [Электронный ресурс]. – URL: http://kola.opus.hbz-nrw.de/volltexte/2015/1136/pdf/OGRW_2014_Proceedings.pdf
17. Рылов, С.А. Иерархические алгоритмы кластеризации мультиспектральных изображений и их использование при создании тематических карт паводковой обстановки / С.А. Рылов, О.Г. Новгородцева, О.А. Дубровская, И.А. Пестунов // Сборник трудов Всероссийской конференции «Обработка пространственных данных в задачах мониторинга природных и антропогенных процессов». – Новосибирск, 2015. – С. 165-171. [Электронный ресурс]. – URL: <http://conf.nsc.ru/files/conferences/SDM-2015/294652/SDM-2015%20Thesis.pdf>
18. Рылов, С.А. Мониторинг паводковой ситуации по спутниковым изображениям высокого разрешения с учетом наличия теней / С.А. Рылов, О.Г. Новгородцева, И.А. Пестунов // Материалы 6-й межд. научно-практической конф. «АГРОИНФО-2015». Ч. 1: Информационные технологии, системы и приборы в АПК. – Новосибирск, 2015. – С. 434-439.
- Зарегистрированные программы для ЭВМ:**
1. Пестунов И.А., Куликова Е.А., Рылов С.А. Программа для кластеризации больших массивов данных ECCA/2010 // Свид. о гос. регистрации программы для ЭВМ № 2011610996 от 27.01.2011 г.
 2. Пестунов И.А., Рылов С.А. Программа ESEG для сегментации мультиспектральных спутниковых изображений высокого пространственного разрешения по спектральным и текстурным признакам // Свид. о гос. регистрации программы для ЭВМ № 2013611866 от 07.02.2013 г.
 3. Пестунов И.А., Рылов С.А. Программа ShadowExtractor для автоматического выделения теней на мультиспектральных аэрокосмических изображениях высокого разрешения // Свид. о гос. регистрации программы для ЭВМ № 2016610256 от 11.01.2016 г.