

Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное научное учреждение
«ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР «КРАСНОЯРСКИЙ
НАУЧНЫЙ ЦЕНТР СИБИРСКОГО ОТДЕЛЕНИЯ
РОССИЙСКОЙ АКАДЕМИИ НАУК»
(ФИЦ КНЦ СО РАН)

ГРНТИ 658.512

Рег. № НИОКТР АААА-А18-
118011890021-4

Рег. № ИКРБС

УТВЕРЖДАЮ
Врио директора ФИЦ КНЦ СО РАН,
д.с.-х.н.

_____ А.А. Шпедт

«27» января 2021 г.

ОТЧЕТ

О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ МЕТОДЫ И ТЕХНОЛОГИИ АНАЛИТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ И ПОСТРОЕНИЯ ПРОГРАММНО-ТЕХНИЧЕСКИХ КОМПЛЕКСОВ И ИНТЕГРИРОВАННЫХ СИСТЕМ (заключительный)

Направление фундаментальных исследований
IV.35 Когнитивные системы и технологии, нейроинформатика и
биоинформатика, системный анализ, искусственный интеллект, системы
распознавания образов, принятие решений при многих критериях
(№ 0356-2019-0016)

Руководитель проекта
заведующий отделом
д.т.н., профессор

_____ Л.Ф. Ноженкова

СПИСОК ИСПОЛНИТЕЛЕЙ

Руководитель темы д.т.н., профессор		Л.Ф. Ноженкова (Введение, Раздел 1, Заключение)
Исполнители: зам. дир., к.т.н.		С.В. Исаев (Раздел 1)
г.н.с., д.т.н.		А.В. Лапко (Раздел 2)
в.н.с., д.т.н.		В.А. Лапко (Раздел 2)
в.н.с., д.ф.-м.н.		М.Г. Садовский (Раздел 3)
с.н.с., к.т.н.		О.С. Исаева (Раздел 1)
с.н.с., к.т.н.		А.В. Коробко (Раздел 1)
с.н.с., к.т.н.		В.В. Ничепорчук (Раздел 1)
с.н.с., к.т.н.		Т.Г. Пенькова (Раздел 1)
с.н.с., к.ф.-м.н.		М.Ю. Сенашова (Раздел 3)
н.с., к.т.н.		Д.В. Жучков (Раздел 1)
н.с., к.т.н.		Е.В. Ковязина (Раздел 1)
н.с., к.т.н.		В.В. Морозов (Раздел 1)
н.с., к.т.н.		А.И. Ноженков (Раздел 1)
н.с.		Д.Д. Кононов (Раздел 1)

М.Н.С.	_____ (подпись, дата)	А.А. Коробко (Раздел 1)
М.Н.С.	_____ (подпись, дата)	А.М. Метус (Раздел 1)
прогр. 1-ой кат.	_____ (подпись, дата)	С.Н. Кочетков (Раздел 1)
прогр. 1-ой кат.	_____ (подпись, дата)	И.А. Ларионова (Раздел 1)
прогр. 1-ой кат.	_____ (подпись, дата)	А.В. Малышев (Раздел 1)
прогр. 1-ой кат.	_____ (подпись, дата)	А.С. Михалев (Раздел 1)
прогр. 1-ой кат.	_____ (подпись, дата)	А.С. Черниговский (Раздел 1)
программист	_____ (подпись, дата)	Н.В. Кулясов (Раздел 1)
инженер	_____ (подпись, дата)	А.А. Сиротинин (Раздел 1)
Нормоконтролер	_____ (подпись, дата)	А.В. Вяткин

РЕФЕРАТ

Отчёт 103 с., 47 рис., 1 прил.

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ, БОЛЬШИЕ ДАННЫЕ, МОДЕЛЬНО-ОРИЕНТИРОВАННЫЕ СИСТЕМЫ, ПРОГРАММНО-ТЕХНИЧЕСКИЕ КОМПЛЕКСЫ, ИНТЕГРИРОВАННЫЕ СИСТЕМЫ.

Объектом исследования являются методы и технологии информационно-аналитической поддержки принятия решений.

Цель работы – создание новых методов интеллектуальной аналитической обработки больших объемов данных, развитие и реализация новых технологических подходов к созданию сложных программно-технических комплексов и интегрированных систем разного назначения.

Использованы оригинальные методы консолидации и оперативной аналитической обработки данных, методы интеллектуальной поддержки конструкторских задач, информационно-графического моделирования, методы анализа больших и сверхбольших данных.

Основные результаты работы:

- Модели, методы и алгоритмы обнаружения технологических нарушений и снижения производительности в цикле производства алюминия по данным мониторинга, разработаны средства визуализации для оценивания текущего состояния технологии и динамики изменения производительности отдельных единиц комплекса производства алюминия.

- Метод построения многомерной аналитической модели состояния сложных объектов и систем на основе выявления и формального описания закономерностей. Выполнено исследование и аналитическое моделирование состояния гидроэнергетической системы на примере одного из гидроагрегатов Красноярской ГЭС. Сравнение аналитических моделей за разные периоды времени позволяет отслеживать динамику состояния системы и отдельных элементов.

- Новые элементы технологии интеллектуального имитационного моделирования в жизненном цикле бортовой аппаратуры космических аппаратов. Создан метод, позволяющий сопоставлять результаты имитационных экспериментов с результатами испытаний реального оборудования. Метод позволяет автоматически выявлять причинно-следственные связи между получаемыми данными испытаний и параметрами работы бортовых систем.

– Методические и программные средства многоэтапной консолидации гетерогенных данных с данными распределенной отраслевой информационной системы. Методы управления консолидированными данными в интегрированных системах территориального и корпоративного управления.

– Новые методы и средства информационно-аналитической поддержки управления территориальной безопасностью территорий, в том числе методика формирования территориально-ориентированного стандарта, представляющего нормативную модель безопасности с учетом индивидуальных особенностей территорий, метод расчета интегральных оценок природно-техногенной безопасности территорий, программные приложения, обеспечивающие безопасность туризма.

– Веб-платформа для автоматизации процесса создания прикладных модельно-ориентированных систем сбора и анализа данных. Платформа обеспечивает возможность динамического расширения тематического наполнения построенных систем на основе моделей, описывающих предметную область. Новые функциональные возможности реализованы за счет интеграции методов интеллектуального анализа данных и средств динамической инфографики и динамического картографирования.

– Модернизированная структура институционального репозитория ИВМ СО РАН, расширяющая круг применяемых web-технологий, схем и форматов метаданных цифровых объектов и системы их хранения. Институциональный репозиторий, формируемый в технологиях архивов открытого доступа, позволяет проводить качественный учет и продвижение научных публикаций, обеспечить открытость результатов научных исследований для мирового сообщества, повысить индивидуальные показатели цитирования.

– Новые модели, методы и программное обеспечение для выявления, оценки и предотвращения угроз кибербезопасности информационно-телекоммуникационных систем.

– Непараметрические методы распознавания образов и модели стохастических зависимостей, которые обеспечивают эффективную обработку информации в условиях малых выборок; быстрые алгоритмы оптимизации непараметрических решающих правил ядерного типа в условиях статистических данных большого объема;

– Методика проверки гипотез о распределениях случайных величин на основе непараметрических алгоритмов распознавания образов, развитая на многомерные законы распределения в условиях статистических данных большого объема.

– Новая методика проверки гипотезы о независимости случайных величин, основанная на использовании непараметрического алгоритма распознавания образов, соответствующего критерию максимального правдоподобия.

- Информационная система для оценивания состояний природных объектов по данным дистанционного зондирования, основанная на предложенных непараметрических системах принятия решений. Система использована для исследования состояний лесных массивов.

- Список новых структурных элементов в символьных последовательностях, их комбинаторные и статистические свойства, связь между структурами и семантикой генома.

- Быстрый (вычислительно эффективный) метод выявления длинных и сверхдлинных общих подстрок в двух или нескольких символьных последовательностях.

- Методы выявления особенностей структурирования и связей выделенных структурных единиц с биологическими свойствами носителей генетической информации (таксономия, функция и т.п.). Проведение вычислительных экспериментов.

- Методы анализа медицинских данных и выявления клинически значимых характеристик в больших массивах данных.

Область применения полученных результатов – интеллектуальный анализ больших объемов данных мониторинга производственных и природных процессов, создание сложных программно-технических комплексов и интегрированных систем для информационно-аналитической поддержки принятия решений в разных прикладных областях.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ.....	9
1. МЕТОДЫ И ТЕХНОЛОГИИ АНАЛИТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ И ПОСТРОЕНИЯ ПРОГРАММНО-ТЕХНИЧЕСКИХ КОМПЛЕКСОВ И ИНТЕГРИРОВАННЫХ СИСТЕМ.....	13
1.1. Методы аналитической обработки данных в интегрированных системах территориального и корпоративного управления	13
1.1.1. Анализ данных на основе многомерной аналитической модели состояния гидроэнергетической системы	13
1.1.2. Модели, методы и алгоритмы обнаружения технологических нарушений и снижения производительности в цикле производства алюминия	16
1.1.3. Методы консолидации и анализа гетерогенных данных с данными распределенной отраслевой информационной системы	18
1.1.4. Метод и алгоритмы формирования унифицированного представления структуры межсистемного обмена данными с учетом вариативности форматов	24
1.1.5. Методы и технологии информационно-аналитической поддержки управления безопасностью территорий	27
1.2. Разработка новых методов интеллектуального имитационного моделирования и испытаний бортовых систем космических аппаратов	33
1.2.1. Интеллектуальное имитационное моделирование командно-программного управления командно-измерительной системой	34
1.2.2. Поддержка конструирования программно-аппаратных комплексов бортовой аппаратуры космического аппарата	35
1.2.3. Метод анализа результатов испытаний бортовой аппаратуры по прецедентам имитационной модели.....	37
1.2.4. Метод анализа имитационной модели по результатам испытаний бортовой аппаратуры космического аппарата	39
1.3. Веб-платформа для автоматизации процесса создания прикладных модельно-ориентированных систем сбора и анализа данных	42
1.4. Развитие инфраструктуры научных коммуникаций	45
1.5. Разработка и апробация средств анализа и оценки угроз кибербезопасности в корпоративной сети.....	47
1.5.1. Выявление сетевых аномалий для противодействия киберугрозам.....	47
1.5.2. Мониторинг кибербезопасности корпоративной сети	51
1.5.3. Разработка в расширенной ролевой модели безопасности на основе RBAC ..	54
2. НЕПАРАМЕТРИЧЕСКИЕ СИСТЕМЫ ПРИНЯТИЯ РЕШЕНИЙ В УСЛОВИЯХ НЕОДНОРОДНЫХ ДАННЫХ	59

2.1. Разработка и исследование быстрых алгоритмов оптимизации непараметрических решающих правил ядерного типа в условиях статистических данных большого объёма...	59
2.2. Разработка статистических методов распознавания образов и восстановления многомерных зависимостей в условиях неоднородных данных	60
2.3. Разработка непараметрических алгоритмов распознавания образов и моделей стохастических зависимостей в условиях малых выборок	64
2.4. Разработка и исследование быстрых алгоритмов оптимизации непараметрических решающих правил ядерного типа в условиях статистических данных большого объёма...	66
2.5. Разработка непараметрических систем проверки гипотез о распределениях многомерных случайных величин и их независимости, основанных на использовании методов распознавания образов	71
3. РАЗРАБОТКА МЕТОДОВ ПОИСКА, КЛАССИФИКАЦИИ И АНАЛИЗА РАЗЛИЧНЫХ СТРУКТУР И СВЯЗЕЙ МЕЖДУ НИМИ В НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЯХ	76
3.1. Список новых структурных элементов в символьных последовательностях, их комбинаторные и статистические свойства, связь между структурами и семантикой генома	76
3.2. Быстрый (вычислительно эффективный) метод выявления длинных и сверхдлинных общих подстрок в двух или нескольких символьных последовательностях	86
3.3. Методы выявления особенностей структурирования и связей выделенных структурных единиц с биологическими свойствами носителей генетической информации	88
3.4. Анализ медицинских данных, выявление клинически значимых характеристик в больших массивах данных.....	90
ЗАКЛЮЧЕНИЕ	96
ПРИЛОЖЕНИЕ. Научные публикации	101

ВВЕДЕНИЕ

Развитие перспективных цифровых технологий в России и во всем мире повышает значимость исследований, направленных на создание новых методов интеллектуальной аналитической обработки больших объемов данных и развитие технологий построения инструментальных цифровых платформ, позволяющих на новый уровень перевести создание прикладных систем во всех сферах производства и жизнеобеспечения. Взрывной рост объёмов данных сопровождается дефицитом новых методов и технологий их обработки и анализа, требующихся для принятия обоснованных и эффективных решений. Примерами могут служить огромные объемы данных мониторинга производственных процессов в энергетике, алюминиевой промышленности и других отраслях экономики, растущие объемы данных экологического мониторинга, спутниковых данных дистанционного зондирования Земли, нарастающие объемы данных как результат активизации применения цифровых технологий в здравоохранении и социальной сфере.

«Большие данные» несут в себе потенциал повышения эффективности всех процессов, но для его реализации необходимо создание новых методов интеллектуальной аналитической обработки больших объемов данных, развития и реализации новых технологических подходов к созданию прикладных систем разного назначения для комплексной информационной, аналитической и интеллектуальной поддержки принятия решений. К числу перспективных относятся новые методы интеллектуализации аналитической обработки данных, основанные на построении многомерных аналитических моделей состояния сложных объектов и систем путем выявления и формального описания закономерностей. Применение методов и технологий интеллектуального анализа к данным мониторинга с последующей визуализацией результатов анализа и выявленных закономерностей в многомерном пространстве данных дает возможность анализа динамики изменения аналитической модели, что в свою очередь, позволяет отслеживать динамику производственных процессов и систем и открывает перспективу создания технологий повышения показателей их технологической эффективности.

Развитие отечественного производства космических аппаратов требует создания методов и технологий поддержки конструирования программно-аппаратных комплексов, обеспечивающих сквозную автоматизацию всех этапов жизненного цикла бортовой аппаратуры. Для повышения экономической эффективности производственных процессов необходимо создание новых технологий, включая технологии имитационного моделирования, поддержки подготовки и проведения испытаний, технологий анализа функционирования бортовой аппаратуры. В основе должны лежать интеллектуальные

методы, способствующие интеграции знаний специалистов предметной области для эффективного решения задач комплексной поддержки высокотехнологичного производства и формирования информационной памяти предприятий.

Развитие современных информационных технологий приводит к повышению уровня «цифровизации» и переводу в киберпространство научных и производственных процессов. Глобальные компьютерные сети глубоко интегрированы в деятельность различных предприятий. Оперативность доступа к информации способствует сокращению издержек производства, повышает эффективность и рентабельность компаний. С другой стороны, такая интеграция может привести к возникновению проблем и потере конфиденциальности данных при несанкционированных внешних воздействиях. Актуальной задачей является организация такой системы безопасности, которая не мешала бы выполнять основные задачи без увеличения рисков киберугроз. Это возможно только при периодической настройке и обновлении системы безопасности, с учетом современных трендов и новых угроз.

Нестационарность объектов исследования, сложность получения необходимой для принятия решений информации и большая размерность данных обуславливают актуальность разработки непараметрических алгоритмов распознавания образов и моделей стохастических зависимостей. Вычислительная эффективность непараметрических статистик значительно снижается при увеличении объёма исходных данных, что особенно характерно в условиях исследования нестационарных объектов, требующих частую корректировку алгоритмов обработки исходной информации. Поэтому возникает необходимость разработки новых методик «быстрой» оптимизации непараметрических оценок плотностей вероятностей, обеспечивающих сокращение временных затрат при определении коэффициентов размытости ядерных функций. Особую актуальность эти исследования приобретают для обработки и анализа больших объемов данных дистанционного зондирования Земли.

Изучение особенностей и деталей структуры нуклеотидных последовательностей в настоящее время является важнейшей задачей биологии. Выявление связи между структурными компонентами и соответствующими им функциями представляет собой классическую проблему молекулярной и системной биологии и, несмотря на обширный поток публикаций и исследований в этом направлении, она всё ещё далека от завершения. В настоящее время все больший интерес вызывают не кодирующие области геномов, поскольку обнаруживаются ранее неизвестные функции этих областей. Анализ внутренней структурированности не кодирующих областей геномов хлоропластов

наземных растений на предмет выявления характерных типов структур является важным направлением исследований в современной биоинформатике.

Цель работы – создание новых методов интеллектуальной аналитической обработки больших объемов данных, развитие и реализация новых технологических подходов к созданию сложных программно-технических комплексов и интегрированных систем разного назначения.

Основные задачи проекта и запланированные результаты:

- Создание новых методов консолидации и анализа больших объемов разнородных данных, обеспечивающих оперативную совместную аналитическую обработку и адаптивное манипулирование данными из разрозненных источников.

- Оригинальная реализация модельно-ориентированного подхода к разработке программного обеспечения, включающая новые технологические элементы, позволяющие создавать адаптивные информационные системы с возможностью оперативного управления тематическим наполнением.

- Создание технологии межсистемного информационного взаимодействия, обеспечивающей стандартизированный обмен данными между гетерогенными ресурсами и гибкую адаптацию к изменениям условий функционирования систем, технология может применяться при создании корпоративных и распределенных систем, например, в сфере государственных услуг, в том числе в сфере размещения заказов.

- Разработка новых технологических подходов к имитационному моделированию, информационно-графическому моделированию и визуализации результатов аналитической обработки данных, как в интегрированных системах, так и в программно-аппаратных комплексах. Новые методы и технологии могут использоваться в задачах проектирования программных систем для имитационного моделирования бортовой аппаратуры космических аппаратов, при создании программно-аппаратных комплексов для поддержки испытаний бортовой аппаратуры.

- Адаптация существующих и создание новых моделей и методов обеспечения безопасности функционирования интегрированных информационно-телекоммуникационных систем. Созданные методы и модели могут применяться как при разработке новых, так и для защиты существующих систем в научно-образовательной сфере.

- Разработка методических, алгоритмических и программных средств построения эффективных статистических систем восстановления стохастических зависимостей и распознавания образов в условиях неоднородных данных малого объёма, основанных на

принципах коллективного оценивания и гибридного моделирования, позволяющих обойти проблему дискретизации области определения плотностей вероятности. Результаты могут быть использованы при исследовании социально-экономических, медико-биологических и сложных технических систем, при обработке данных дистанционного зондирования территорий Севера и Сибири.

– Разработка методов выявления структур и способов их анализа (классификации), результатами их применения станут перечень новых структурных элементов и их классификация по информационным, комбинаторным и статистическим свойствам. Эти результаты могут применяться при секвенировании, сборке и аннотировании геномов различных организмов, в том числе хвойных растений. Исследование будет способствовать построению исчерпывающей и непротиворечивой теории, объясняющей фундаментальные связи структур биологических макромолекул и их свойств.

Все перечисленные задачи решены, запланированные результаты получены в полном объеме и представлены в настоящем отчете. Представлены также примеры применения полученных результатов для мониторинга производственных и природных процессов, создания сложных программно-технических комплексов и интегрированных систем для информационно-аналитической поддержки принятия решений в разных прикладных областях.

Настоящий заключительный отчет представляет результаты работ за период 2017 – 2020 гг. по теме «Методы и технологии аналитической обработки данных и построения программно-технических комплексов и интегрированных систем».

1. МЕТОДЫ И ТЕХНОЛОГИИ АНАЛИТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ И ПОСТРОЕНИЯ ПРОГРАММНО-ТЕХНИЧЕСКИХ КОМПЛЕКСОВ И ИНТЕГРИРОВАННЫХ СИСТЕМ

Ответственный исполнитель блока д.т.н. Ноженкова Л.Ф.

1.1. Методы аналитической обработки данных в интегрированных системах территориального и корпоративного управления

1.1.1. Анализ данных на основе многомерной аналитической модели состояния гидроэнергетической системы

Предложен метод построения многомерной аналитической модели состояния сложных объектов и систем на основе выявления и формального описания закономерностей. Построение многомерной аналитической модели заключается в последовательном применении совокупности методов и технологий интеллектуального анализа к данным мониторинга системы с последующей визуализацией результатов анализа и выявленных закономерностей в многомерном пространстве данных. Аналитическая модель описывает особенности функционирования системы и ее элементов в различных режимах и условиях работы в отдельные моменты времени.

Выполнено построение многомерной аналитической модели состояния гидроэнергетической системы (гидроагрегата). Комплексный интеллектуальный анализ данных мониторинга позволил определить особенности в работе гидроагрегата, обнаружить структуру и закономерности в данных, выявить признаки взаимного влияния его конструктивных узлов, определить соотношения диапазонов значений ключевых технических параметров в различных режимах функционирования оборудования. На рисунке 1.1 представлен результат кластеризации данных на плоскости двух первых главных компонент, где объекты – это моменты времени работы одного из гидроагрегатов Красноярской ГЭС за три месяца, агрегированные по часам, атрибуты – контролируемые параметры: A_POWER – активная мощность, МВт; R_POWER – реактивная мощность, МВт; RV_GB_LB – относительная вибрация генераторного подшипника, мкм; RV_HB_LB – относительная вибрация пяты подпятника, мкм; RV_TB_LB – относительная вибрация турбинного подшипника, мкм; AV_GB_LB – абсолютная вибрация генераторного подшипника; AV_FH_LB – абсолютная вибрация опоры подпятника, мкм; AV_TB_LB – абсолютная вибрация турбинного подшипника, мкм.

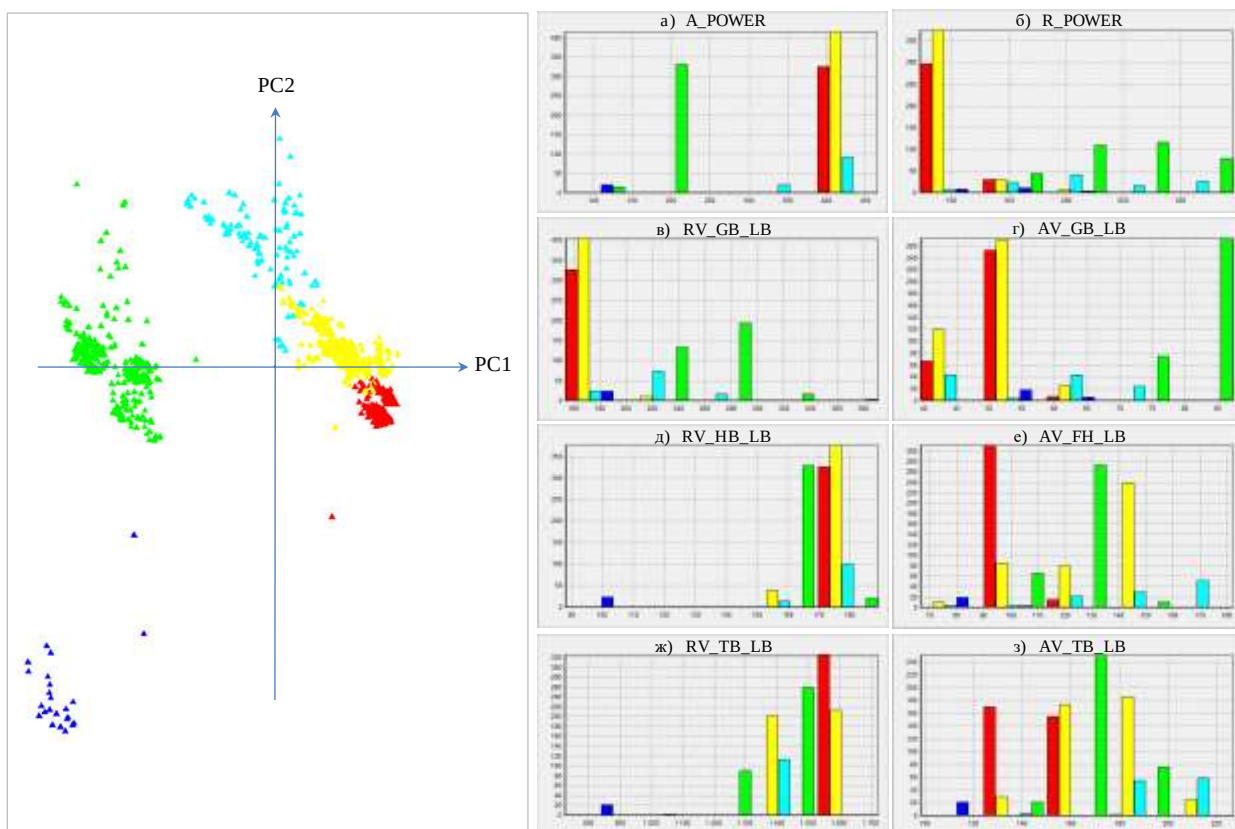


Рис. 1.1 – Пятикластерная структура данных мониторинга гидроэнергетической системы

Для выявления типового поведения системы в определенных режимах работы гидроагрегата выполнена кластеризация по диапазону значений контролируемых параметров. Исследование взаимного влияния параметров выполнено путем детализация рассматриваемых моментов времени до минут, построены диаграммы сопоставления временных рядов контролируемых параметров для характерных отрезков времени. На рисунке 1.2 представлена диаграмма, описывающая состояние гидроагрегата в моменты времени с высокими значениями относительной вибрации турбинного подшипника. Из рисунка видно, что незначительное уменьшение активной мощности (в диапазоне ее высоких значений, $\sim 450\text{ МВт}$) ведет к небольшому снижению относительной вибрации турбинного подшипника (RV_TB_LB), но при этом к существенному росту значений абсолютной вибрации опоры подшипника (AV_FH_LB) ($\approx 50\%$) и абсолютной вибрации генераторного подшипника (AV_GB_LB) ($\approx 30\%$). В результате анализа установлено, что при среднем уровне активной мощности вибрация генераторного подшипника достигает максимальных значений, в то время как при высоком уровне активной мощности она минимальна, но при этом относительная вибрация пяты подшипника и турбинного подшипника имеет достаточно высокие значения. Кроме того, высокому уровню активной мощности соответствуют низкие значения реактивной мощности. Также удалось выявить признаки взаимного влияния функциональных узлов гидроагрегата в различных режимах

работы. Замечено, что турбинный подшипник и подпятник имеют схожий характер поведения, их относительная вибрация повышается при повышении уровня активной мощности, а абсолютная вибрация, наоборот, уменьшается. При этом характер поведения генераторного подшипника отличается, его абсолютная и относительная вибрации уменьшаются при повышении уровня активной мощности.

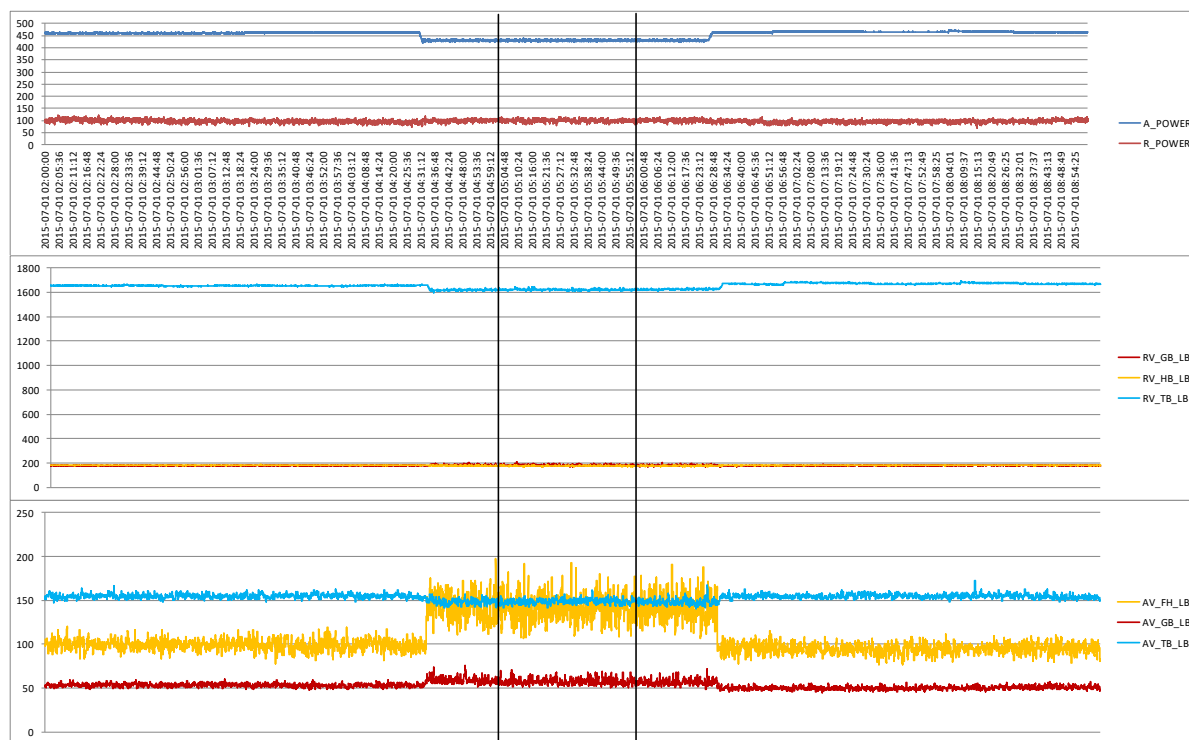


Рис. 1.2 – Сопоставление временных рядов контролируемых параметров в моменты времени с высокими значениями относительной вибрации турбинного подшипника

Выполненный детальный анализ характерных моментов времени позволил выявить определенные зависимости значений для контролируемых параметров. Например, незначительное изменение амплитуды активной мощности на верхней границе средних значений (~ 200 МВт) ведет к существенному увеличению относительной вибрации генераторного подшипника ($\approx 60\%$), абсолютной вибрации опоры подпятника ($\approx 45\%$) и абсолютной вибрации турбинного подшипника ($\approx 17\%$); небольшое уменьшение активной мощности в диапазоне своих высоких значений (~ 450 МВт) ведет к небольшому снижению относительной вибрации турбинного подшипника (в диапазоне своих максимальных значений, ~ 1600 мкм) и к существенному увеличению абсолютной вибрации опоры подпятника ($\approx 50\%$) и генераторного подшипника ($\approx 30\%$). Выявленные зависимости между параметрами, диапазоны и соотношения их значений в различных режимах эксплуатации являются уникальными характеристиками и отличительными свойствами конкретного объекта управления – гидроагрегата. Изменение технического состояния гидроагрегата с

течением времени проявляется в изменении структуры и закономерностей в данных, представленных многомерной аналитической моделью. Сравнение аналитических моделей за разные периоды времени позволяет отслеживать динамику износа оборудования и изменения технического состояния системы в целом.

1.1.2. Модели, методы и алгоритмы обнаружения технологических нарушений и снижения производительности в цикле производства алюминия

В рамках создания информационно-аналитической системы оперативного контроля технологии производства алюминия на основе применения совокупности методов математической статистики, интеллектуального анализа данных и машинного обучения разработаны модели и алгоритмы обнаружения технологических нарушений и снижения производительности в цикле производства алюминия по данным мониторинга. В ходе выполнения комплексного структурного многомерного анализа данных среднесуточного мониторинга определены закономерности функционирования комплекса производства алюминия и характерные условия возникновения отклонений в технологическом процессе (рис. 1.3).

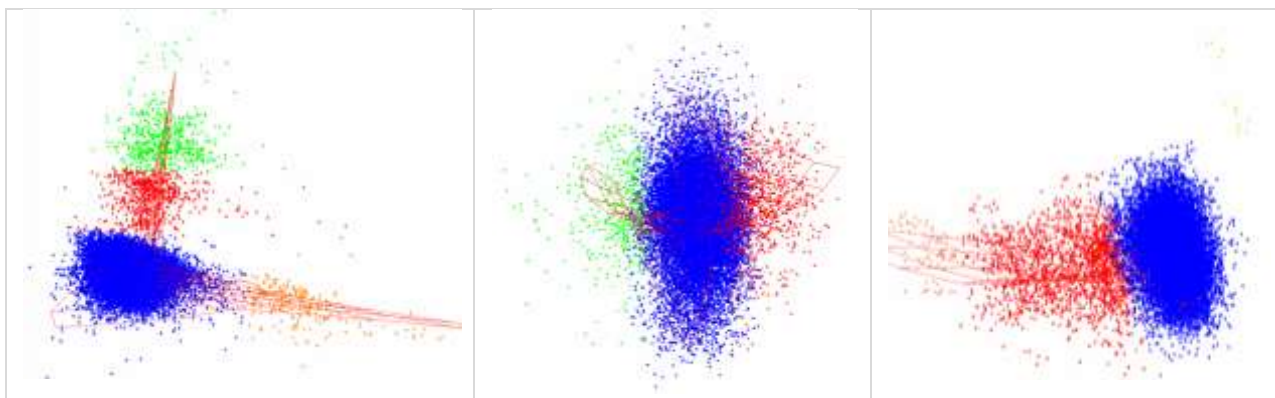


Рис. 1.3 – Результаты кластеризации данных в пространстве главных компонент:

а) ХАЗ; б) БоАЗ; в) БрАЗ

Например, результаты кластеризации объектов ХАЗа (Хакасский алюминиевый завод) (рис. 3а) показали, что синий кластер характеризует основной режим работы; красный кластер объединяет события возникновения «анодного эффекта», зеленый кластер связан с событиями планового или аварийного отключения электролизеров; желтый кластер характеризует возникновение технологического нарушения типа «конус». Результаты кластеризации объектов БоАЗа (Богучанский алюминиевый завод) (рис. 3б) показали, что синий кластер характеризует основной режим работы, зеленый кластер объединяет события возникновения «анодного эффекта» и «конуса», красный кластер характеризует возникновение технологического нарушения типа «отставание». Детальный

анализ событий с технологическими нарушениями показал, что условия и природа возникновения нарушений типа «конус» и «отставание» отличаются. При нарушении «отставание» значения параметра Температура электролита уменьшаются, а значения параметра Потребление глинозема увеличиваются, в то время как, при нарушении типа «конус» наоборот значения параметра Температура электролита увеличиваются, а значения параметра Потребление глинозема уменьшаются. Результаты кластеризации объектов БрАЗа (Братский алюминиевый завод) (рис. 3в) позволили выявить технологические особенности, связанные с возрастом электролизеров и их производительностью: синий кластер характеризует основной режим работы, красный кластер характеризуется большим возрастом электролизеров, умеренным уровнем потребления глинозема и высокими значениями уровня металла; желтый кластер характеризуется большим возрастом электролизеров, очень низким уровнем потребления глинозема и высокими значениями уровня металла; зеленый кластер объединяет события с особыми условиями технологического процесса. Полученные результаты позволили подтвердить многие гипотезы и сформировать «аналитические портреты» работы отдельных единиц комплекса производства алюминия.

На основе выявленных зависимостей и результатов анализа прецедентов событий определены диапазоны допустимых значений контролируемых параметров, сформированы правила оценивания текущего состояния технологии и возникновения нарушений для отдельных единиц комплекса производства алюминия. По данным мгновенного мониторинга разработаны модели прогнозирования ключевых параметров, позволяющие своевременно обнаруживать отклонения в технологии (рисунок 1.4).

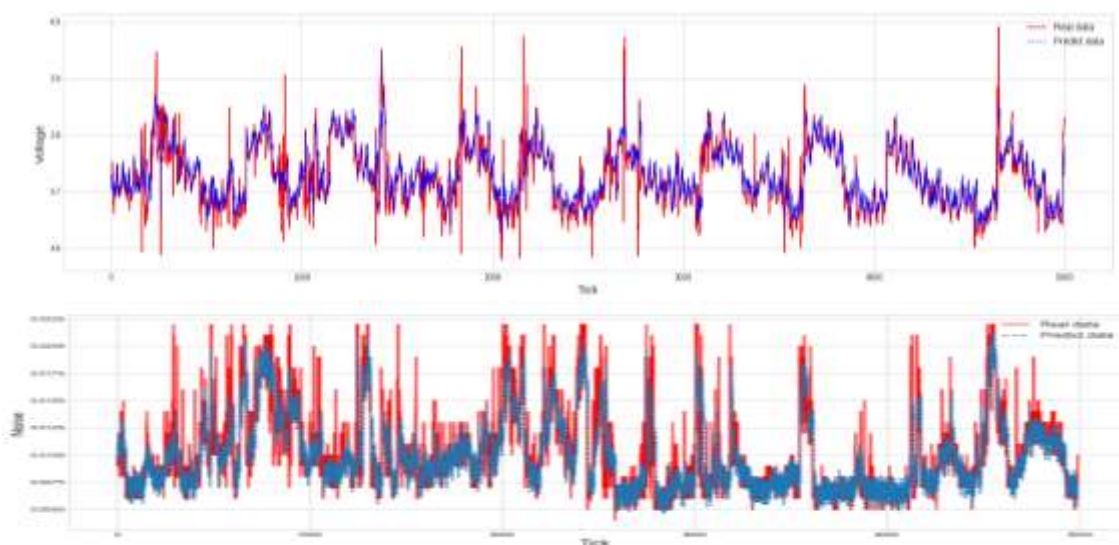


Рис. 1.4 – Результаты работы прогнозной модели для значений напряжения и «шума»

Разработаны алгоритмы идентификации изменений характера временных рядов параметров, свидетельствующих о возникновении технологических нарушений, созданы средства визуализации для оценивания текущего состояния технологии и динамики изменения производительности отдельных объектов по комплексу параметров производства алюминия (рис. 1.5).



Рис. 1.5 – Средства оперативного контроля технологических параметров в цикле производства алюминия с индикацией обнаружения отклонений

1.1.3. Методы консолидации и анализа гетерогенных данных с данными распределенной отраслевой информационной системы

Разработаны методические и программные средства многоэтапной консолидации гетерогенных данных, в ходе которой сведения, получаемые от удаленных веб-ресурсов, согласуются с данными распределенной отраслевой информационной системы.

На первом этапе консолидации осуществляется получение данных из внешнего источника с предварительной фильтрацией входящего потока сведений – из обработки исключаются сведения, не отвечающие установленным требованиям. Затем выполняется загрузка сведений в принимающую базу и постобработка полученных данных – заполнение дополнительных служебных атрибутов, частичное изменение форматов и форм представления информации. Обработанные данные поступают на этап адресной рассылки конечным получателям – выполняется формирование пакетов с учётом категории получателя, а также его ведомственной и территориальной принадлежности. В систему введены две основных категории получателей данных – первая категория

получает только собственные данные, вторая категория получает данные подчиненных структур. Таким образом, достигается согласованность данных на различных уровнях организационной иерархии.

На следующем этапе в локальных экземплярах баз данных выполняется загрузка сведений из поступающих пакетов с использованием буферных структур, допускающих дополнительную обработку и согласование данных. Далее следует двухэтапная сверка разноструктурных гетерогенных данных. Первый этап проверки выполняется для обнаружения факта расхождения данных, второй этап служит для формирования детального отчёта об обнаруженных нарушениях. Если в ходе проверки данных обнаруживаются несоответствия, то выполняется согласование локальных сведений со сведениями, поступившими из внешнего источника. Разработанные средства контроля данных также используются для дополнительных проверок согласованности данных при последующих этапах их обработки. Повторное использование процедур сверки данных позволяет поддерживать согласованность данных в случае их изменений. Заключительным этапом процесса консолидации данных является дублирующий контроль в центральной базе данных, который позволяет удостовериться в корректности сведений, передаваемых в основную базу системы. Апробация разработанных средств консолидации гетерогенных данных выполнена для организации муниципальных закупок г. Красноярска при обмене данными между Автоматизированной системой поддержки муниципальных закупок и Общероссийским официальным сайтом zakupki.gov.ru.

Выполнен комплекс работ по созданию новых и развитию существующих методов и алгоритмов аналитической обработки и интеграции разнородных данных. Проведен анализ вычислительных затрат расчета аналитических показателей при реализации многошаговых методик. На основе формализации базовых операций над OLAP-кубом проанализирована производительность формируемых запросов при различных сочетаниях аналитических операций с точки зрения мощности обрабатываемого многомерного куба данных. Результаты исследования позволили определить оптимальную последовательность операций для повышения эффективности алгоритмов расчета аналитических показателей. На примере данных гидрологического мониторинга территорий Красноярского края определен наиболее эффективный путь извлечения оперативных данных за счет сокращения мощности OLAP-куба.

Предложена метамодель формирования многомерных оценок аналитических показателей на основе спецификации Common Warehouse Metamodel (CWM), которая представляет открытый технологический стандарт, обеспечивающий информационное взаимодействие инструментов обработки данных за счет формирования унифицированной

структуры метаданных. На основе CWM разработана расширенная метамодель OLAP, определяющая логическую структуру аналитических моделей. Разработана метамодель аналитических операций над OLAP-кубом, позволяющая реализовать алгоритм расчета многомерных оценок показателей. Расширение стандартной метамодели для задач мониторинга природно-техногенной безопасности территорий выполнено путем включения класса «*Standard*», который задает многомерные нормативные значения показателя, атрибута «*isTerritory*» класса «*Dimension*», который обеспечивает привязку данных к измерению территории, а также аналитической операции над OLAP-кубом «*Estimation*», которая реализует алгоритм расчета многомерных оценок. В соответствии с предложенной метамоделью, на основе расширяемого языка разметки предложены DTD-описания, определяющие структуру XML-файлов для хранения и передачи аналитических моделей. Предложенные метаописания обеспечивают оперативную актуализацию многомерных моделей и представляют единый формат работы аналитических средств с OLAP-моделями.

Исследованы особенности представления данных в слабоструктурированных форматах и сопоставлены со свойствами реляционной схемы. Предложены алгоритмы трансформации разнородных схем данных в универсальную объектную модель, построенную на основе теории многомерного аналитического моделирования. Предложен формальный подход к виртуальной интеграции гетерогенных источников на основе динамической шины данных. Проведены исследования быстродействия интегральной аналитической модели как алгебраической решетки кубов-концептов, сделаны выводы о состоятельности предлагаемого подхода.

Разработана формальная основа объединения гетерогенных данных путем формирования референтного множества аналитических измерений, содержащего все возможные аспекты анализа данных разнородных источников, и обнаружения общих аспектов анализа с применением методов семантико-синтаксического анализа.

Предложена метамодель построения унифицированных объектных аналитических моделей источников различных форматов, позволяющая преодолеть гетерогенность (разнородность) объединяемых для совместного анализа данных. Разработаны и реализованы формальные правила построения унифицированных моделей для XML-документов и реляционных баз данных на основе анализа исходных схем хранения данных. Создание формального фундамента в виде метамодели позволило разработать универсальный форматонезависимый алгоритм формирования многомерных представлений гетерогенных источников данных на основе анализа их объектных аналитических моделей. Предложена и апробирована уникальная распределенная

архитектура системы аналитической интеграции гетерогенных данных, обеспечивающая доступ к гетерогенным источникам данных, их объединение и совместный анализ.

В рамках развития авторского подхода к совместной OLAP-обработке гетерогенных данных на основе виртуальной интегральной аналитической модели была поставлена и решена задача разработки теоретических основ объединения данных из разрозненных источников с учетом вариативности форматов. Прикладной целью настоящих исследований является построение интегральной аналитической платформы для совместного эксплоративного анализа разнородных данных, находящихся в открытом доступе, и данных корпоративных информационных систем.

Предложен способ построения интегральной аналитической модели, объединяющей разнородные схемы хранения, путем формирования референтного множества измерений, содержащего все возможные аспекты анализа данных. Основная трудность при формировании референтного множества измерений заключается в определении и слиянии общих аспектов анализа для гетерогенных источников данных.

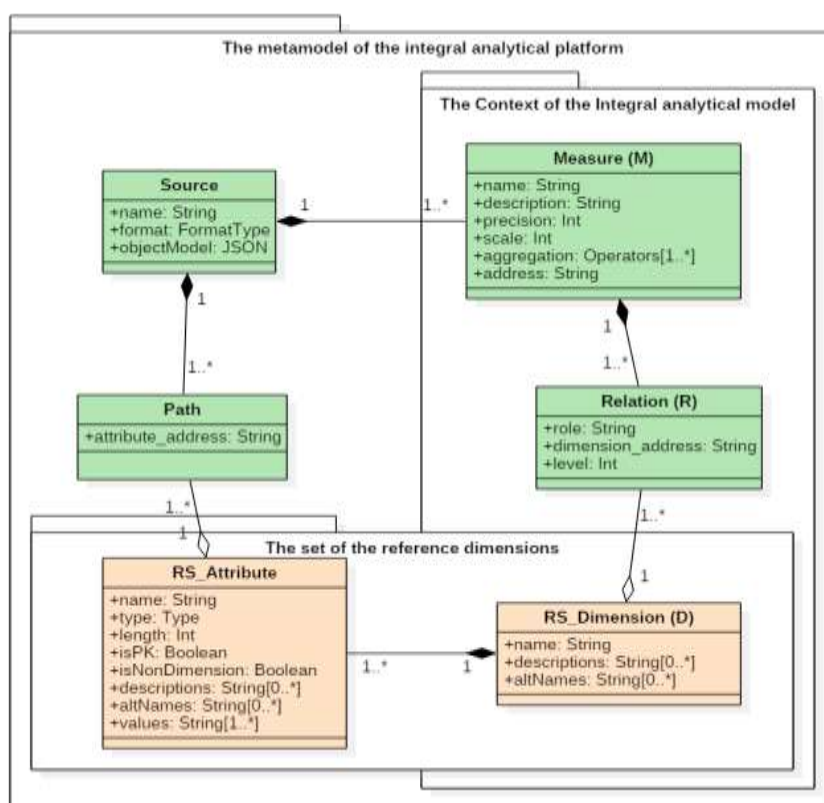


Рис. 1.6 – Метамодель интегральной аналитической платформы

Для решения проблемы представления объединённого множества измерений предложено накапливать варианты наименования и описания измерений, общих для нескольких источников, а также перечень возможных атрибутов и их свойств. Предложена концептуальная модель («The metamodel of the integral analytical platform»)

интегральной аналитической платформы (рис. 1.6). Модель описывает необходимые сущности, обеспечивающие:

- формирование и хранение всех возможных свойств и атрибутов измерений референтного множества («The set of reference dimensions») – «RS_Dimension (D)» и «RS_Attribute»;
- построение и выполнение запросов на получение данных непосредственно из гетерогенных источников – «Path» и «Source»;
- построения интегральной аналитической модели как решетки кубов-концептов на основе бинарной матрицы аналитической сопоставимости обобщенных измерений и показателей («The Context of the integral Analytical model») – «Measure (M)», «Relation (R)» и «RS_Dimensiond (D)».

Разработанная концептуальная модель позволяет учитывать как особенности представления исходных гетерогенных данных, так и их аналитические свойства. Для решения проблемы автоматического сопоставления измерений предложен алгоритм определения степени сходства разнородных измерений путем семантико-синтаксического анализа текстовых свойств и атрибутов аналитических измерений.

Изучены свойства интегральной аналитической модели (ИАМ), быстроедействие алгоритмов построения и использования модели в виде решетки OLAP-кубов. Исследование выполнено в виде вычислительного эксперимента. Объектом исследования выбрана модель размещения муниципального заказа, объединяющая два источника: реляционный (Региональная система формирования потребностей) и база XML-документов (Единая информационная система в сфере закупок). В качестве контрольного образца ИАМ для сравнения параметров быстрогодействия используется модель аналогичной размерности и плотности заполнения, сформированная с помощью генератора случайных чисел. Выполнен сравнительный анализ следующих аспектов: структуры моделей, скорости генерации решетки по контексту, количество концептов (вершин) решетки и свойств концептов (объема и содержания). Сравнительная оценка скорости (сложности) выполнения процедуры поддержки исследования ИАМ, заключающейся в сужении поля поиска информации по ходу формирования (уточнения) пользовательского аналитического запроса, выполнена для исходного бинарного контекста модели и построенной решетки OLAP-кубов. Вычислительный эксперимент реализован в среде «JupyterLab» на языке python 3.7. На рисунке 1.7 представлен пример визуализации свойств кубов-концептов.

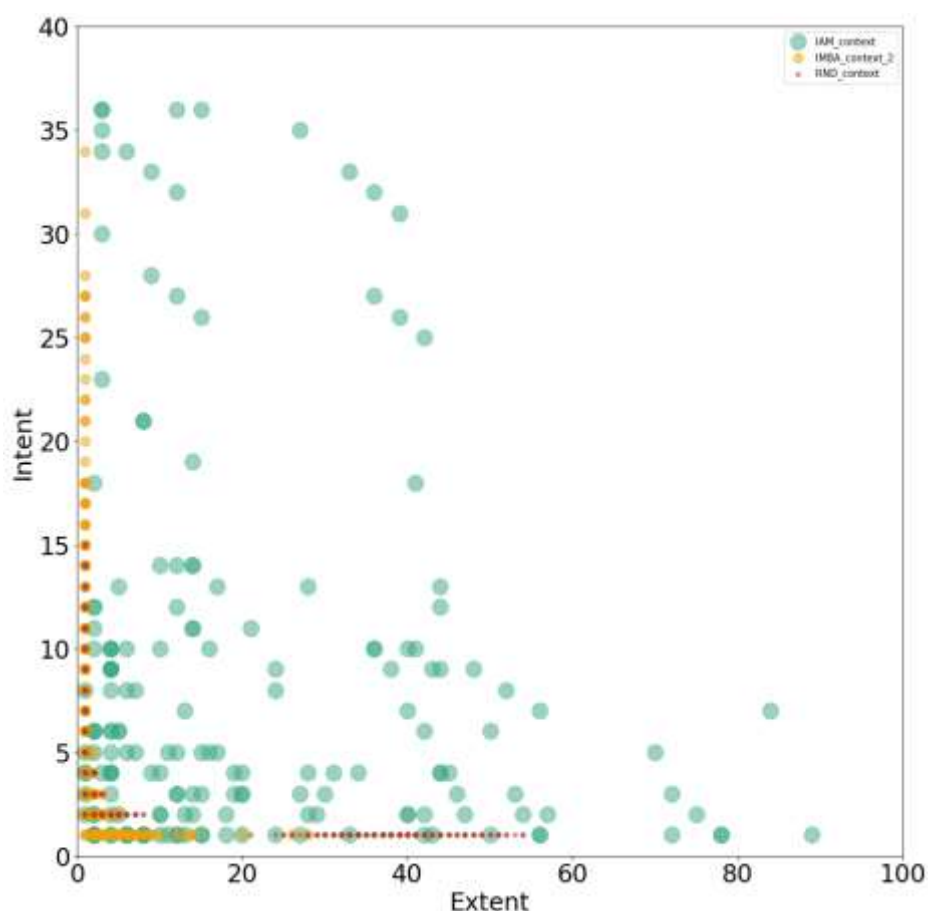


Рис. 1.7 – Диаграмма рассеяния для объема и содержания кубов-концептов ИАМ

В ходе исследования выявлены значимые структурные особенности реального контекста ИАМ. Функциональные зависимости между атрибутами исходных схем хранения находят отражение в иерархических зависимостях между измерениями многомерной модели, а в контексте ИАМ они принимают вид отношений «обусловленности» в соответствии с современной теорией априорного формирования системы измеряемых свойств в методологии FCA. Наличие отношений обусловленности (и несовместимости) между измерениями ИАМ существенно снижает скорость вычисления концептов и уменьшает их количество по сравнению с контрольной моделью. Кроме того, анализ свойств концептов позволил выявить границы мощности объема и содержания концептов в силу естественных ограничений числа аналитических связей в реальных базах данных.

Программная реализация предложенной метамодели интегральной аналитической платформы и алгоритма сопоставления разнородных аналитических измерений служит апробацией теоретических результатов и обеспечивает формирование интегральной аналитической модели гетерогенных данных. Результаты проведенного вычислительного эксперимента подтверждают обоснованность применения предложенной модели для решения современных задач оперативного анализа данных, требующих быстрогодействия в

условиях большого объема и разнородности обрабатываемой информации. Развитие предложенного подхода сопряжено с систематизацией ранее полученных результатов и созданием технологии интегрального аналитического моделирования. Совершенствование теоретической базы предложенного подхода заключается в поиске способов интеллектуализации процесса формирования ИАМ в части построения многомерной модели и учета вариативности аналитических связей.

1.1.4. Метод и алгоритмы формирования унифицированного представления структуры межсистемного обмена данными с учетом вариативности форматов

Предложен метод, разработаны модели и алгоритмы формирования унифицированного представления структуры межсистемного обмена данными с учетом вариативности форматов. Согласно предложенному методу обмен данными между системами можно представить в виде пары: $M = (\langle O_{In}, I_{In} \rangle, \langle O_{Out}, I_{Out} \rangle)$, где I – системы информационного обмена, O – объекты информационного обмена. Каждая система может участвовать в информационном обмене в качестве «отправителя» (I_{In}) или «получателя» (I_{Out}). Объект информационного обмена можно представить как тройку: $O = \langle D, F, S \rangle$, где D – передаваемые или получаемые системой данные, F – формат хранения данных, S – структура, в которой представлены данные. Для системы «отправителя» и системы «получателя» объект информационного обмена определяется как $O_{In} = \langle D_{In}, F_{In}, S_{In} \rangle$ и $O_{Out} = \langle D_{Out}, F_{Out}, S_{Out} \rangle$ соответственно. Процесс унифицированного обмена данными заключается в преобразовании информационного объекта «отправителя» в информационный объект «получателя» на основе формирования унифицированного представления структур данных и может быть представлен цепочкой:

$$O_{In} \langle D_{In}, F_{In}, S_{In} \rangle \xrightarrow{Q \langle D_{In}, S_{In} \rangle} U_{In} \langle D_{In}, S_{In} \rangle \xrightarrow{G \langle U_{In}, S_{Out} \rangle} U_{Out} \langle D_{Out}, S_{Out} \rangle \\ \xrightarrow{\bar{Q} \langle D_{Out}, S_{Out}, F_{Out} \rangle} O_{Out} \langle D_{Out}, F_{Out}, S_{Out} \rangle,$$

где Q – оператор унификации, выполняющий преобразование данных «отправителя» в унифицированный вид: $Q: O_{In} \rightarrow U_{In}$, U_{In} – данные «отправителя» в терминах унифицированной модели; $\bar{Q}$ – оператор деунификации, выполняющий преобразование данных в терминах унифицированной модели в формат данных «получателя»: $\bar{Q}: U_{Out}, F_{Out} \rightarrow O_{Out}$, где U_{Out} – данные «получателя» в терминах унифицированной модели, F_{Out} – это формат хранения данных «получателя». При этом унифицированная модель U представляет собой универсальную иерархическую структуру, независимую от формата данных, в терминах которой может быть представлена любая структура данных. G – оператор, выполняющий преобразование унифицированной структуры данных

«отправителя» в унифицированную структуру данных «получателя»: $G: U_{In}, S_{Out} \rightarrow U_{Out}$. Оператор G работает с определенной парой структур (структурой данных «отправителя» и структурой данных «получателя») и не зависит от содержания данных, что позволяет использовать оператор многократно для обмена данными заданных структур. На рисунке 1.8 представлена метамодель унифицированного информационного обмена в виде диаграммы классов, описывающая логику процесса, основные сущности и отношения.

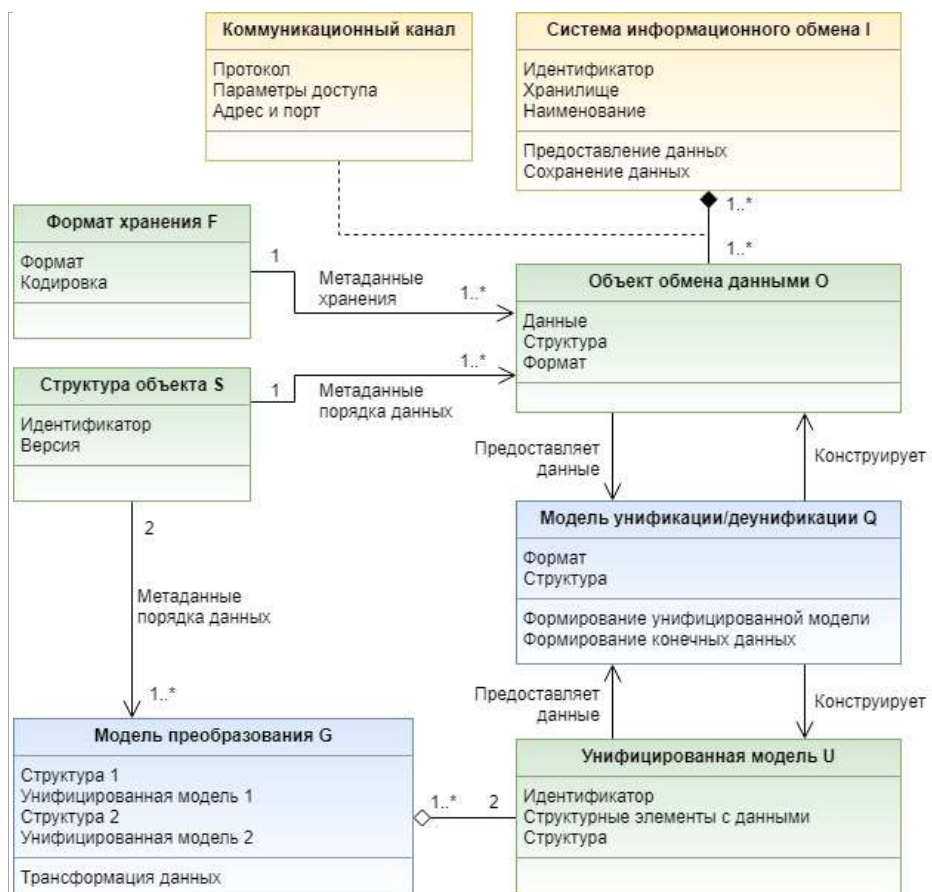


Рис. 1.8 – Метамодель унифицированного информационного обмена данными

В соответствии с предложенной моделью разработана архитектура программных средств, обеспечивающая организацию унифицированного обмена данными. На рисунке 1.9 представлена диаграмма, описывающая основные модули архитектуры и их взаимодействие. Процесс унифицированного обмена данными между гетерогенными системами можно разделить на два этапа: подготовка и применение. Этап подготовки включает такие процессы, как описание структуры данных и описание взаимного соответствия элементов структур данных, которые выполняются «отправителем» и «получателем» данных с помощью визуальных компонент: *Structure Builder* и *Link Builder*. *Structure Builder* – конструктор, предназначенный для описания структуры данных (S_{In} и S_{Out}) в терминах унифицированной модели. *Link Builder* – конструктор, предназначенный

для установления связей между элементами двух разных унифицированных структур, а также для описания операций над отдельными элементами (*Links Set*). *Cross Operand Builder* – конструктор, который на основании описаний структур данных и взаимосвязей между элементами формирует сценарий преобразования данных $A(S_{In}, S_{Out})$ в виде последовательности команд выполнения операций над элементами унифицированных структур данных (перестановка, удаление, агрегация, перерасчет и др.). Построенный сценарий преобразования данных сохраняется в базу данных *Central DB* для дальнейшего применения.

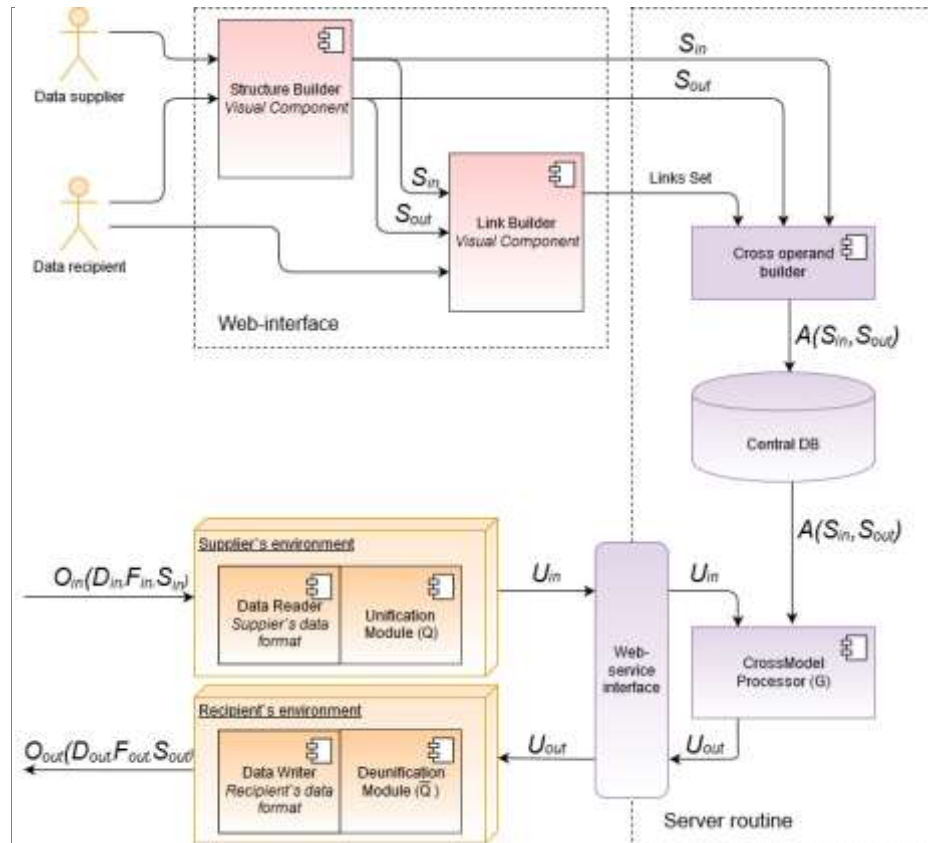


Рис. 1.9 – Архитектура программных средств унифицированного обмена данными

Этап применения направлен на преобразование данных. На вход цепочки преобразования «отправителем» передается объект информационного обмена $O_{In} = \langle D_{In}, F_{In}, S_{In} \rangle$. Формирование унифицированного представления выполняется с помощью двух модулей: *Data Reader* и *Unification Module*. *Data Reader* – модуль, выполняющий считывание данных из источника в формате «отправителя» (F_{In}). *Unification Module* – модуль унификации (Q), выполняющий перевод данных в термины унифицированной модели (U_{In}) и освобождающий данные от их формата хранения (F_{In}). Модули работают на стороне «отправителя» (*Supplier's environment*), а на сервер обработки передаются готовые к преобразованию данные в унифицированном виде (U_{In}). *Cross Model Processor* (G) – модуль преобразования, выполняющий преобразование данных отправителя в

унифицированном виде (U_{In}) в требуемые получателем данные в унифицированном виде (U_{Out}) на основе сценария $A(S_{In}, S_{Out})$. Сценарий преобразования выбирается из базы данных *Central DB* по паре структур (S_{In}, S_{Out}) , где S_{In} – структура, в которой представлены унифицированные данные «отправителя» (U_{In}), S_{Out} – структура, в которой представлены унифицированные данные «получателя» (U_{Out}). В результате обработки – выполнения команд сценария – модулем преобразования *Cross Model Processor* формируются данные «получателя» в терминах унифицированной модели (U_{Out}). Формирование объекта информационного обмена «получателя» выполняется на стороне «получателя» (*Recipient's environment*) с помощью двух модулей: *Deunification Module* и *Data Writer*. *Deunification Module* – модуль деунификации ($\bar{Q}$), выполняющий перевод данных в терминах унифицированной модели (U_{Out}) в данные в формате хранения «получателя» (F_{Out}). *Data Writer* – модуль, выполняющий запись объекта информационного объекта «получателя» $O_{Out} = \langle D_{Out}, F_{Out}, S_{Out} \rangle$ в хранилище. Возможность хранения сценария преобразования $A(S_{In}, S_{Out})$ позволяет при однократной подготовке структур данных многократно выполнять преобразование данных из структуры S_{In} в структуру S_{Out} .

Разработанные в рамках предложенного метода алгоритмические и программные средства позволяют автоматизировать обмен данными между информационными системами с возможностью адаптации к изменениям условий и форматов передаваемых данных. Метод и алгоритмы апробированы для задачи организации закупок, где осуществляется обмен данными между корпоративной системой размещения заказа, единой информационной системой и электронными торговыми площадками.

1.1.5. Методы и технологии информационно-аналитической поддержки управления безопасностью территорий

Выполнены работы по развитию методов аналитической обработки мониторинговых данных и формированию комплексных оценок состояния объекта контроля в системах информационно-аналитической поддержки территориального управления. Для обеспечения информационной поддержки стратегического управления природно-техногенной безопасностью территорий разработана иерархическая система аналитических показателей, характеризующих природные и техногенные факторы риска возникновения чрезвычайных ситуаций. На основе консолидации многолетних данных с применением технологий хранилищ данных и оперативной аналитической обработки рассчитаны многомерные показатели состояния безопасности территорий Красноярского края в динамике по годам и различным видам обстановок.

Предложен метод интегрального аналитического оценивания природно-техногенной безопасности территорий, обеспечивающий формирование комплексного показателя на основе результатов многомерного аналитического моделирования состояния окружающей среды и объектов техносферы. Согласно предложенному методу, интегральное оценивание выполняется с применением территориально-ориентированного стандарта, который представляет собой нормативную модель безопасности, учитывающую индивидуальные особенности территорий. Разработана методика формирования стандарта, включающая определение коэффициентов значимости аналитических показателей, нормативных значений показателей, функций агрегирования и коэффициентов чувствительности оценок для различных видов обстановок. На основе анализа накопленных статистических и отчётных данных ведомственных систем мониторинга разработан стандарт природно-техногенной безопасности Красноярского края.

Для формирования комплексных оценок разработан алгоритм формирования интегральной оценки природно-техногенной безопасности территорий. Интегральная оценка рассчитывается на основе оценок базовых показателей с учетом их значимости в показателе верхнего уровня иерархии. Оценки базовых показателей представляют собой агрегированные по территории оценки показателей, рассчитанные по каждому отдельному пункту наблюдения. Многомерные оценки показателей формируются в виде OLAP-моделей, характеризуют соответствие показателя нормативу и учитывают скорость изменения оценок при отклонении фактического значения показателя от установленного норматива. Для интерпретации количественных оценок предложен алгоритм формирования оценочной шкалы с применением аппарата нечетких множеств. Каждому значению оценки, на основе функций принадлежности, ставится в соответствие значение лингвистической переменной «Уровень природно-техногенной безопасности»: «Низкий», «Пониженный», «Удовлетворительный», «Приемлемый», «Улучшенный». Функция принадлежности определяется с помощью метода нечеткой кластеризации по распределению значений интегральных оценок путём сопоставления нечеткого кластера нечеткому множеству. С использованием предложенного алгоритма сформированы оценочные шкалы для интерпретации оценок природно-техногенной безопасности территорий Красноярского края. На рисунке 1.10 представлен пример результата оценивания уровня природно-техногенной безопасности Красноярского края в виде картограммы.

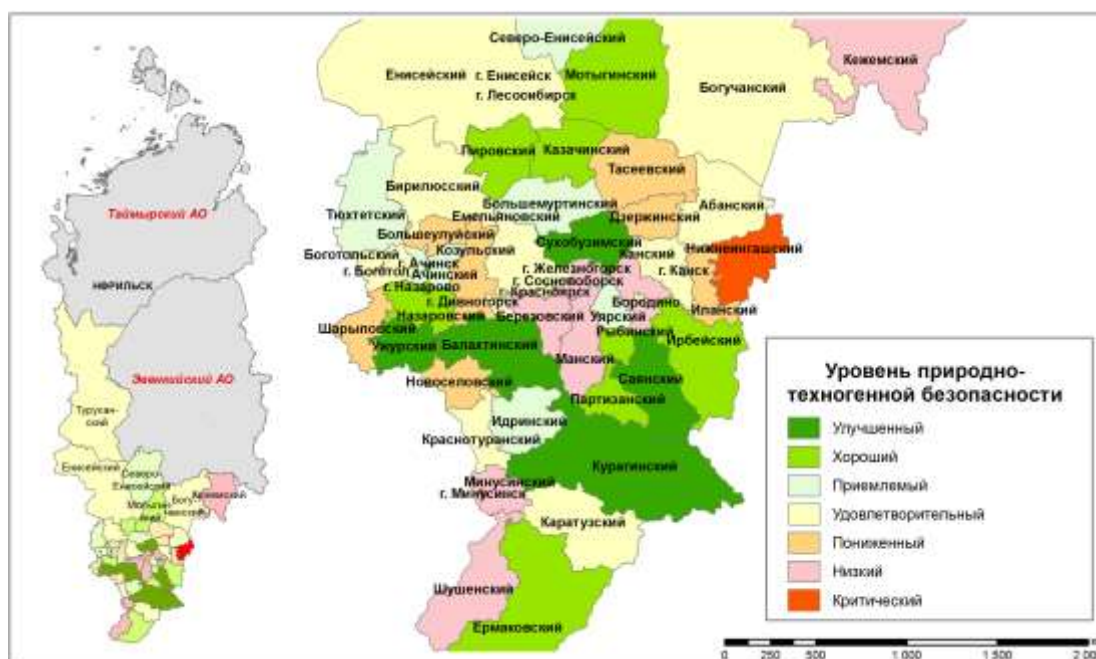


Рис. 1.10 – Результат оценивания природно-техногенной безопасности территорий Красноярского края в виде картограммы

Предложенный метод позволяет получать комплексные оценки текущего состояния, выполнять сравнительный анализ территорий и, в случае необходимости, детализировать оценки до отдельных сфер мониторинга и показателей, что дает возможность определить первопричины текущего состояния и сформировать целевые управляющие рекомендации. Апробация предложенного метода выполнена для комплексного оценивания природно-техногенной безопасности территорий Красноярского края.

Для информационно-аналитической поддержки оперативного и стратегического управления территориальной безопасностью разработаны средства динамической визуализации результатов многомерного аналитического моделирования состояния объектов техносферы и окружающей среды, а также результатов комплексного оценивания природно-техногенной безопасности территорий. На основе метода интегрального аналитического оценивания природно-техногенной безопасности территорий, с применением web-технологий, разработаны средства, обеспечивающие расчет статистики событий по данным оперативного мониторинга и формирование аналитических оценок безопасности территорий с учетом нормативных значений показателей. Результаты обработки представляются в виде динамических кросс-таблиц, диаграмм и картограмм с индикацией уровня безопасности в разрезе аналитических показателей, территорий и временного периода. Элементы визуализации настраиваются с

учетом многомерного характера данных: иерархия показателей, иерархия территорий и иерархия оценок. На рисунках 1.11 и 1.12 представлены результаты оценивания состояния территорий Красноярского края (населенных пунктов и районов) по показателю «Количество аварий на производственных объектах» в виде картограммы (за один отчетный период) и диаграммы (в динамике по годам) соответственно. Индикация уровня безопасности выполняется в соответствии с алгоритмом формирования оценочной шкалы на основе модели нечеткой кластеризации.

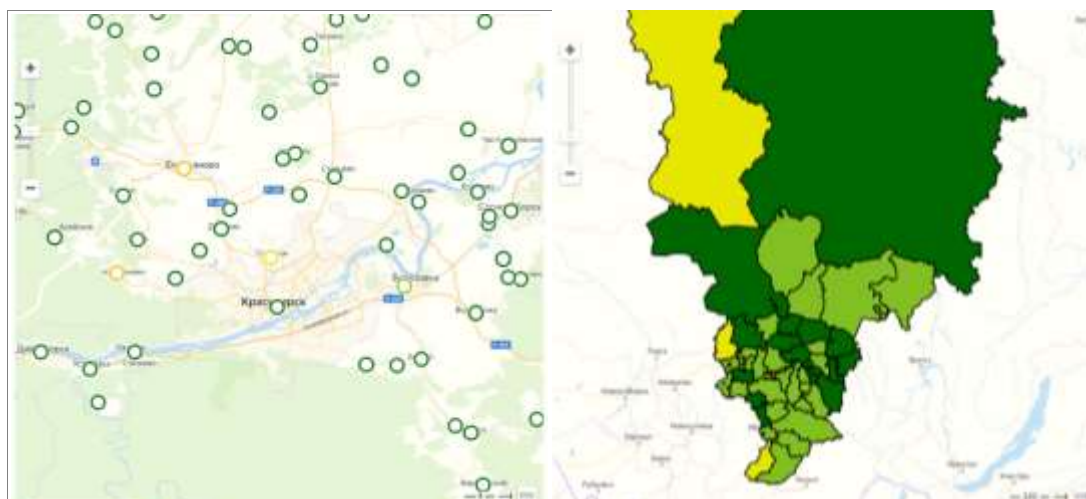


Рис. 1.11 – Пример визуализации результатов оценивания состояния безопасности в разрезе территорий Красноярского края за отчетный период в виде картограммы

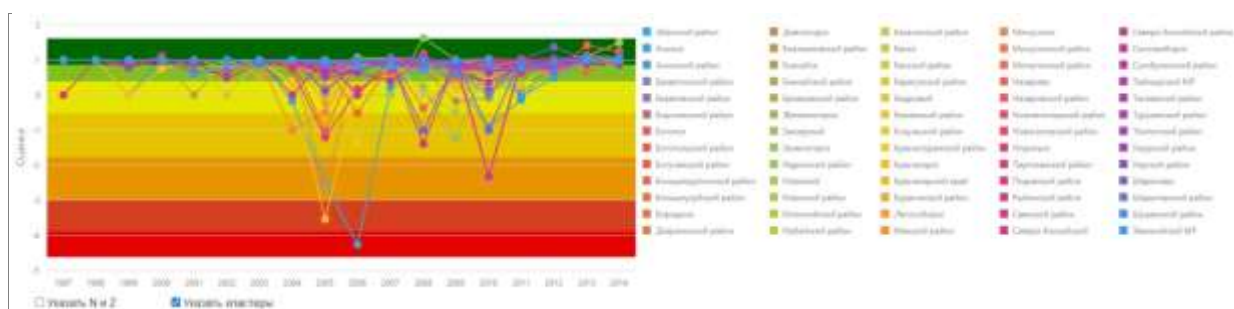


Рис. 1.12 – Пример визуализации результатов оценивания состояния безопасности в разрезе территорий Красноярского края в динамике по годам в виде диаграммы

Разработаны мобильные приложения для туристов и веб-система мониторинга туристической деятельности в режиме онлайн. Мобильные приложения предназначены для регистрации туристических групп, создания и редактирования туристических маршрутов, их согласования с МЧС, а веб-система – для мониторинга передвижения туристических групп по заявленным маршрутам. Вся информация автоматически передается в Территориальный центр мониторинга и прогнозирования чрезвычайных

ситуаций, а также непосредственно спасательным и аварийным службам. Приложения разработаны на основе современных информационных технологий, использующих стандартный функционал смартфонов, планшетов и других устройств с выходом в сеть Интернет – это делает доступным их использование большинством современных туристов.

Для разработки приложения была выбрана одна из популярнейших платформ, позволяющих разрабатывать как веб, так и гибридные приложения – Angular. Angular – это открытый фреймворк, написанный на языке TypeScript, разрабатывается компанией Google и открытым сообществом. Для интеграции приложения с оборудованием устройств используется технология Ionic. Ionic, кроме инструментов, обеспечивающих доступ к оборудованию устройств, представляет библиотеку стандартных визуальных компонент, имеющих характерные для каждой из платформ черты. Angular и Ionic позволяют реализовать клиентскую часть приложения. Серверная часть реализована с помощью Cloud Firestore, объединяющего механизм хранения данных и доступа к ним. Основная часть Cloud Firestore – это документо-ориентированная NoSQL база данных, обеспечивающая хранение и синхронизацию между сервером и несколькими пользователями в реальном времени за счет специальных слушателей событий. Кроме этого, данный сервис предоставляет возможность управления пользователями (Firebase Authentication) и обеспечивает разделение доступа к различным частям данных, основываясь на ролях. Интеграция мобильного приложения и веб-системы с сервисами Cloud Firestore, а также Firebase Authentication реализована с помощью Google-библиотек Angular. Результат сборки клиентского приложения для туриста – исполняемый файл для целевой платформы. Результат сборки веб-системы – JavaScript-файл, который размещен в интернете для обеспечения доступа к нему всем пользователям. Для этого использован хостинг Firebase от компании Google. Хостинг Firebase обеспечивает безопасность приложения и позволяет легко, с помощью одной команды, развернуть приложение Angular.

Мобильное приложение безопасности туриста позволяет регистрировать туристические группы с указанием состава участников, времени начала и окончания маршрута, интервалов прохождения контрольных точек маршрута (Рис. 1.13). Туристический маршрут может быть выбран из справочника существующих маршрутов, или заново сформирован с использованием цифровой мультимасштабной карты.

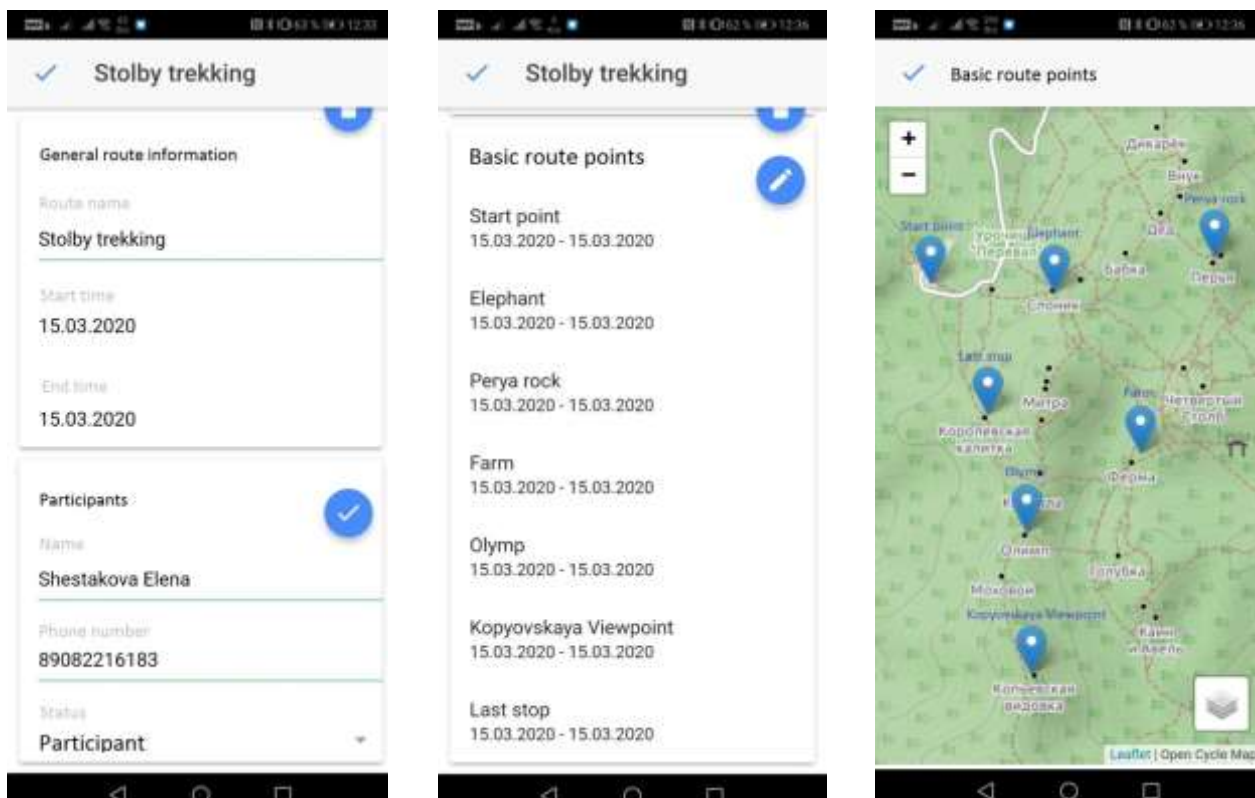


Рис. 1.13 – Мобильное приложение безопасности туриста

В мобильных приложениях предусмотрено получение экстренных оповещений, прогнозов и информации о запретах, действующих на территориях.

Для работы с приложением требуется выполнить процедуры регистрации и подтверждения с использованием электронной почты или альтернативными способами. Это позволяет вести реестр пользователей, а в экстренных случаях использовать голосовую связь. Входные данные приложения включают оповещения, прогнозы обстановки, запреты, характеристики маршрутов, описание опасностей, образовательную информацию для туристов. Выходные данные представляют данные о маршруте, составе туристической группы, отчётные данные о прохождении маршрута.

Возможность получения мобильными приложениями оповещений, прогнозов, а также данных оперативного мониторинга природных и техногенных опасностей реализована путем интеграции с веб-системой, функционирующей в Территориальном центре мониторинга и прогнозирования чрезвычайных ситуаций Красноярского края. Оперативные службы согласуют маршруты, контролируют наличие туристических групп в зоне ответственности аварийно-спасательных формирований и проводит обработку поступающих сообщений. Работа операторов и администратора реализована в виде веб-ресурса, доступного для просмотра через веб-браузеры. Пользователи получают

информацию об опасных погодных и гидрологических явлениях, фрагменты суточной оперативной сводки, а также имеют возможность просмотра тематических ресурсов.

Мобильные приложения и веб-система мониторинга туристической деятельности разработаны на основе современных информационных технологий и адаптированы к законодательству России, что позволяет качественно по-новому обеспечивать безопасность людей за счет прямого взаимодействия с оперативными спасательными службами.

1.2. Разработка новых методов интеллектуального имитационного моделирования и испытаний бортовых систем космических аппаратов

Комплекс методов интеллектуального имитационного моделирования и анализа функционирования бортовых систем предназначен для поддержки решения задач конструирования бортовой аппаратуры космических аппаратов (БА КА). К таким задачам относятся: моделирование командно-программного управления бортовыми системами космического аппарата, проведение имитационных экспериментов, проведение испытаний БА КА, анализ результатов испытаний по прецедентам имитационного моделирования, анализ имитационной модели по результатам испытаний (рис. 1.14).

Для решения указанных задач предложены методы имитационного моделирования командно-программного управления космическим аппаратом, поддержки конструирования программно-аппаратных комплексов бортовой аппаратуры, анализа результатов испытаний командно-измерительных систем по прецедентам имитационного моделирования и анализа имитационной модели по результатам испытаний бортовой аппаратуры.

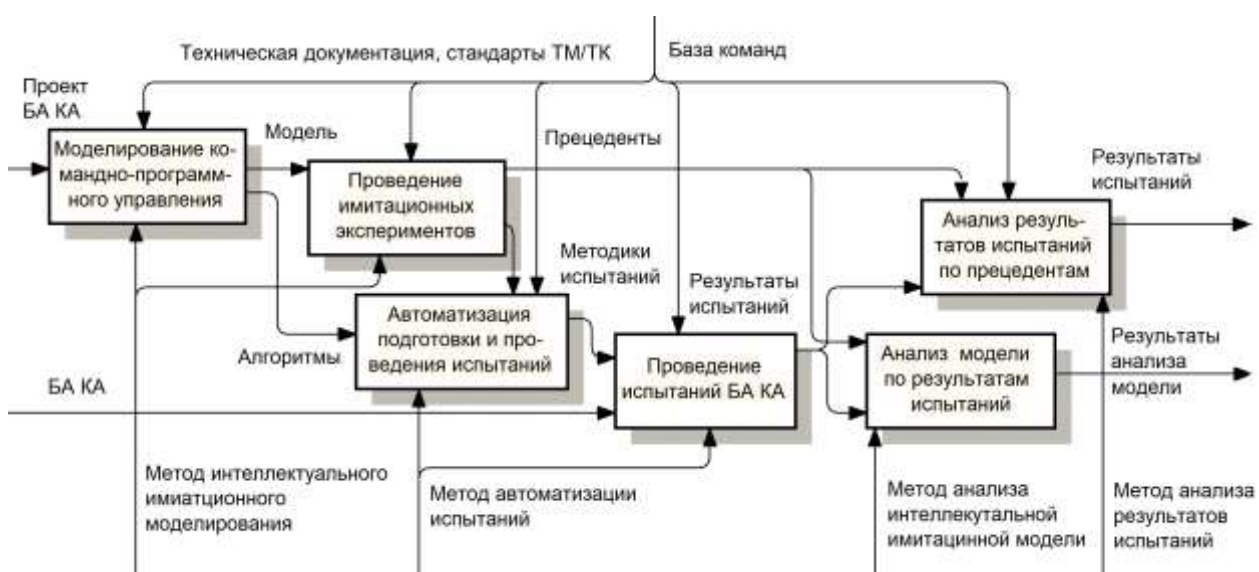


Рис. 1.14 – Задачи поддержки конструирования БА КА и методы их решения

В основе всех методов лежит интеллектуальная имитационная модель, которая содержит базу знаний об особенностях работы бортовой аппаратуры при приеме, обработке и передаче команд управления. Проведены имитационные эксперименты, которые позволили описать алгоритмы функционирования объекта испытаний и выделить контрольные точки для анализа прохождения команд. На основе выбранных критериев анализа разработано программное обеспечение, предназначенное для подготовки испытательных процедур и анализа их выполнения. Программное обеспечение встроено в программно-аппаратный комплекс контрольно-проверочной аппаратуры. Предложенный комплекс методов обеспечивает решение широкого круга задач поддержки проектирования, разработки и испытаний бортовых систем космических аппаратов.

1.2.1. Интеллектуальное имитационное моделирование командно-программного управления командно-измерительной системой

Инструментами интеллектуального имитационного моделирования построена схема потоков данных, которая позволила определить типы данных, направления информационного взаимодействия и контрольные точки для анализа прохождения и отработки команд (рис. 1.15).

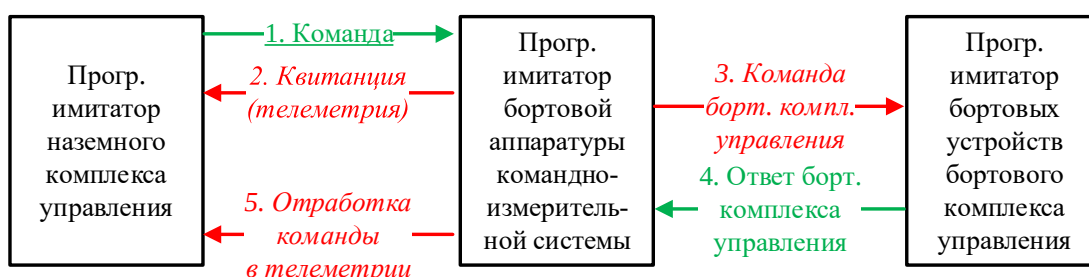


Рис. 15 – Потоки данных для имитации командно-программного управления

Разработаны алгоритмы информационного взаимодействия имитаторов бортовых систем (рис. 1.16). В алгоритмах реализованы функции программных имитаторов: имитатор наземного комплекса управления (НКУ), имитатор командно-измерительной системы (КИС), имитатор бортовых устройств бортового комплекса управления (БУ БКУ) и входящего в его состав бортового цифрового вычислительного комплекса (БЦВК). Предложенные алгоритмы моделируют функции бортовых систем при приёме-передаче команд управления (на рисунке – «Обработка команд») и функции анализа прохождения команд по контрольным значениям параметров телеметрической информации (на рисунке – «Контроль ТМ»). Создание алгоритмов позволило исследовать функции бортовой аппаратуры, определить критерии контроля функционирования и информационного

взаимодействия бортовых систем. На основе построенных имитационных моделей выполняется поддержка конструирования бортовых систем.

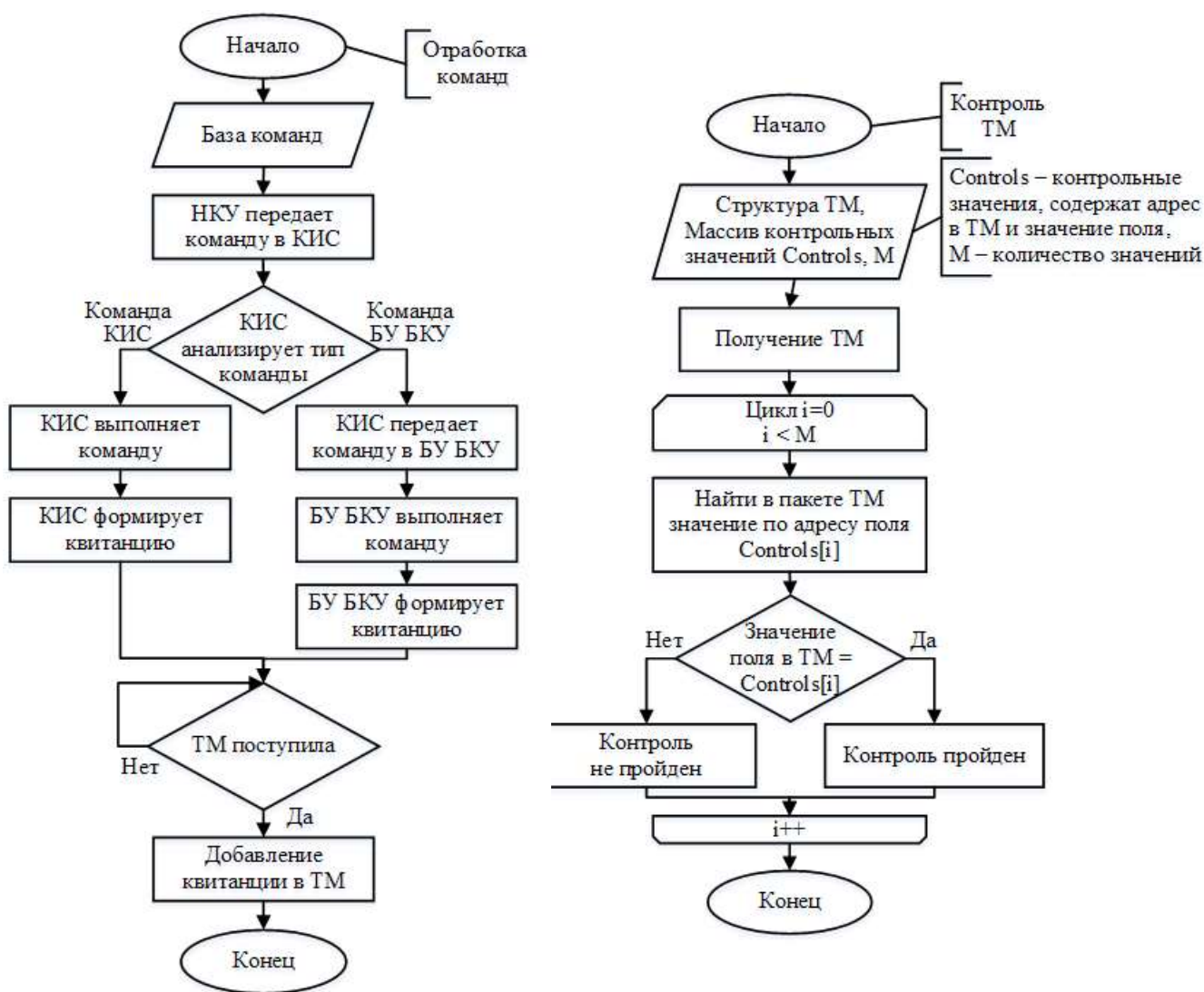


Рис. 1.16 – Алгоритмы информационного взаимодействия имитаторов

1.2.2. Поддержка конструирования программно-аппаратных комплексов бортовой аппаратуры космического аппарата

Для решения задач поддержки конструирования выполнена автоматизация испытаний командно-программного управления бортовой аппаратурой космического аппарата. Предложен информационный состав методик испытаний, который определяет воздействия на объект контроля, действия имитаторов окружения и критерии анализа результатов испытаний. Диаграмма формирования методики испытаний командно-программного управления бортовой аппаратурой показана на рисунке 1.17. Формирование методик испытаний сводится к заданию настроек: «Пакет данных команды», «Интерфейс КПА-КИС» и «Порядок отправки», которые определяют поведение имитаторов в

соответствии с выбранным вариантом командно-программного управления. Критерии анализа функционирования бортовой аппаратуры задаются в контрольных точках: «Контроль квитанции по ТМ», «Контроль МИ КИС БКУ», «Контроль реакции на команду».

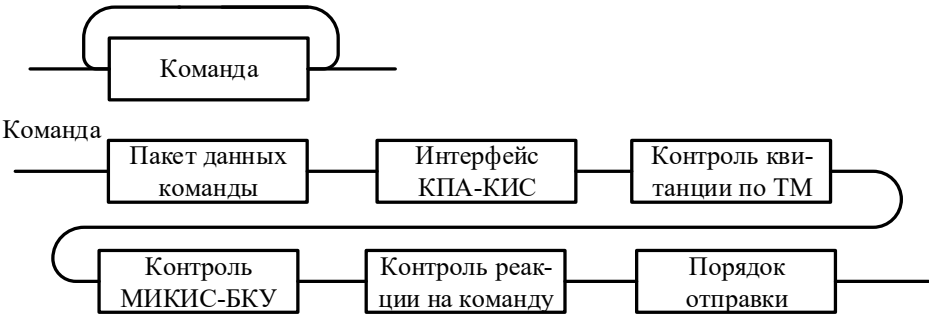


Рис. 1.17 – Диаграмма формирования методики испытаний

Реализация метода автоматизации испытаний выполнена в программном обеспечении контрольно-проверочной аппаратуры ПО КПА. На рисунке 1.18 приведён пример визуализации результатов испытаний командно-программного управления бортовой аппаратурой.

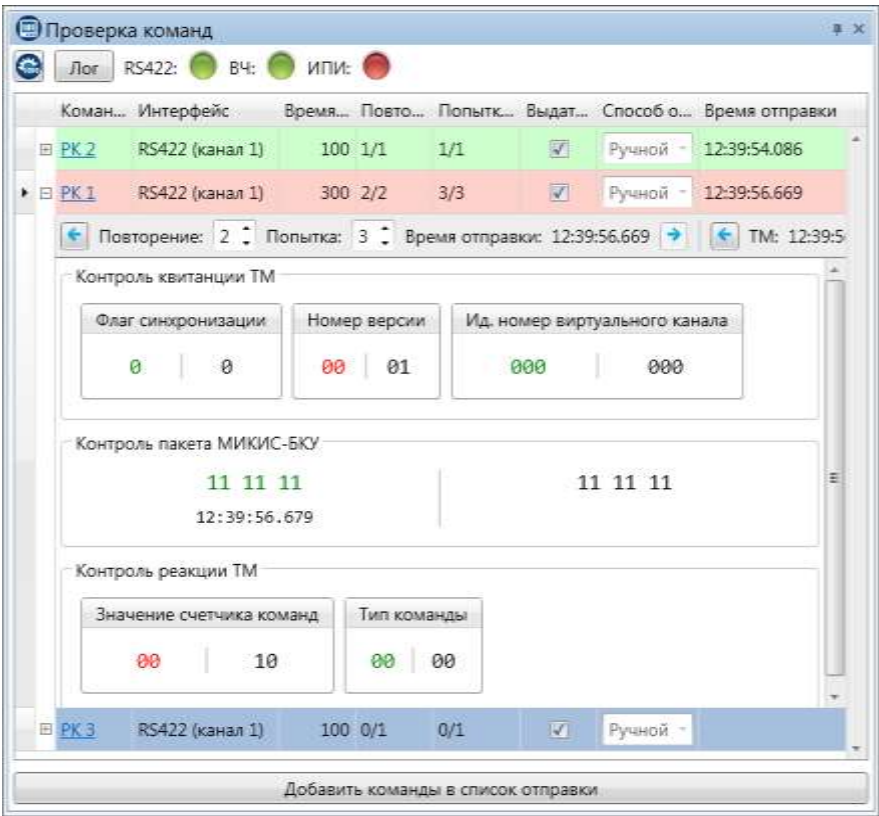


Рис. 1.18 – Подсистема испытаний командно-программного управления

Программное обеспечение выполняет заданные в методике испытаний последовательности команд и отображает результаты анализа контрольных точек. Ведется

журнал выполняемых действий, передаваемых и получаемых данных с возможностью формирования отчета по всем проводимым испытаниям.

1.2.3. Метод анализа результатов испытаний бортовой аппаратуры по прецедентам имитационной модели

Для анализа результатов испытаний бортовой аппаратуры разработана база прецедентов имитационного моделирования. Прецедент состоит из последовательности действий, выполняемых для достижения определенного результата и параметров, характеризующих этот результат. Структура базы прецедентов приведена на рисунке 1.19.



Рис. 1.19 – Структура базы прецедентов

Созданы структуры данных для записи правил, их зависимостей и последовательностей выполнения, параметров телеметрии, состояний элементов модели, а также структуры данных для хранения массивов байт, содержащих пакеты команд и телеметрии. Разработано программное обеспечение, которое позволяет сохранять в базу прецедентов результаты проведения имитационных экспериментов.

Предложен метод анализа испытаний по базе прецедентов, который позволяет исследовать логику поведения бортовых систем на имитационных моделях и применять результаты моделирования для анализа результатов испытаний. Метод рассматривает правила выполнения отправки команд управления, их получение, квитирование, обработку, изменение параметров приемо-передающего тракта, а также состояния бортовых систем, управление работой оборудования (основные / резервные комплекты) и пр. Результаты таких действий отображаются в отдельных параметрах телеметрической информации. Задача сводится к следующей последовательности действий: настройка испытательных процедур передаются в объект контроля и имитационную модель;

выполнение имитационного моделирования; результаты моделирования выбираются в качестве эталонных значений для сопоставления с результатами испытаний и сохраняются в базе прецедентов; проводятся испытания технических устройств; выполняется анализ результатов испытаний по базе прецедентов. Алгоритм анализа приведён на рисунке 1.20.

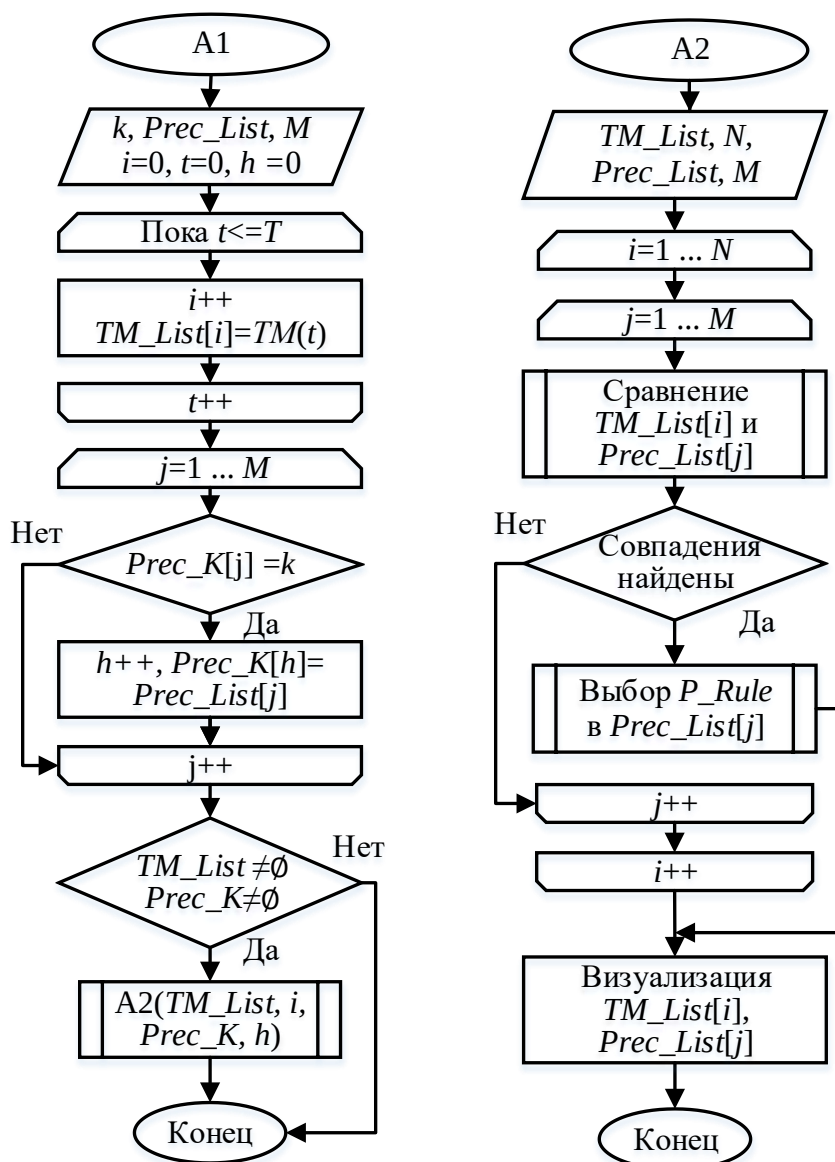


Рис. 1.20 – Алгоритм анализа результатов испытаний по базе прецедентов

На начальном этапе работы алгоритма выбираются данные для сравнения и формируются: TM_List – множество пакетов телеметрической информации, полученных при испытаниях бортовой аппаратуры, $Prec_List$ – множество прецедентов имитационной модели. $Prec_ListK = \langle P_TM, P_Rule, P_K \rangle$, где P_TM – телеметрия, формируемая имитационной моделью, P_Rule – правила базы знаний, P_K – отправленная команда, T – период ожидания ответа на команду. Если построенные множества TM_List и $Prec_ListK$ не пустые, то выполняется «A1 – Алгоритм анализа испытаний по базе прецедентов», который выполняет поиск параметров телеметрии для прецедентов в P_TM и их сравнение

с телеметрией объекта контроля в *TM_List*. Совпадение значений параметров свидетельствует о том, что произошли одинаковые события, как в имитационной модели, так и во время испытаний реального оборудования. Изменение состояния модели описывается правилами *P_Rule*. Результатом работы алгоритма являются прецеденты модели, соответствующие телеметрии бортовых систем, и правила базы знаний.

Выполнена реализация метода анализа результатов испытаний в программном обеспечении контрольно-проверочной аппаратуры. Результаты работы программного обеспечения приведены на рисунке 1.21.

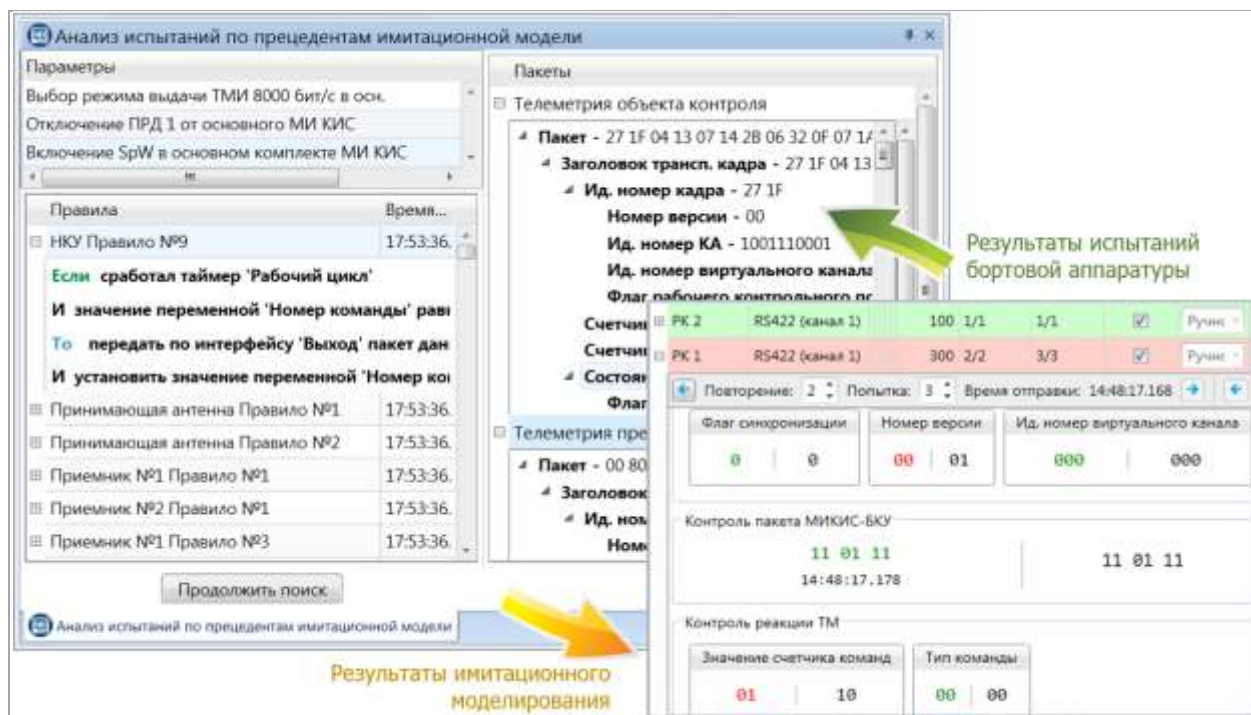


Рис. 1.21 – Анализ результатов испытаний по прецедентам моделирования

Сравнение телеметрии бортовых систем с базой прецедентов дополняет методики анализа и проведения испытаний. Прецедент содержит наборы правил базы знаний, которые выполнялись для его получения, что позволяет выявлять особенности работы, не наблюдаемые при анализе отдельных параметров. Предложенный подход рассматривает все возможные изменения параметров, отраженные в имитационной модели, и применяется для различных видов испытаний. В свою очередь, результаты испытаний бортовых систем применяются для анализа имитационной модели.

1.2.4. Метод анализа имитационной модели по результатам испытаний бортовой аппаратуры космического аппарата

Для анализа интеллектуальной имитационной модели и оценки соответствия результатов её работы реальному оборудованию предложен метод, обеспечивающий

сопоставление результатов имитационных экспериментов с данными, получаемыми при испытаниях бортовых систем. Алгоритм работы следующий:

1. Выполнить имитационное моделирование. Определить команды, которые поступают на вход имитаторов бортовых систем по коммутационным интерфейсам.
2. Выбрать правила в базе знаний, в которых заданы методы отработки команд.
3. Для каждого правила построить цепочки правил, выполняемых в процессе логического вывода для имитаторов бортовых систем.
4. Для каждой цепочки правил объединить зависимые цепочки, логический вывод в которых выполняется при выполнении исходных цепочек правил. Построенные цепочки правил описывают действия модели при различных способах командно-программного управления бортовой аппаратурой.
5. Найти коммутационные соединения, по которым передаются данные при различных взаимодействиях элементов модели. Коммутационные соединения определяют точки контроля для анализа работы имитационной модели.
6. Выполнить сравнение в выбранных контрольных точках телеметрии, построенной имитационной моделью с результатами испытаний бортовых систем, формируется список ошибок. При отсутствии в базе знаний правил, либо несовпадении результатов моделирования и данных, полученных при испытаниях от объекта контроля, требуется доработка имитационной модели путём формирования новых правил базы знаний.

Критерии интерпретации результатов анализа:

- Если в выбранной цепочке правил результаты отработки команд в имитационной модели и объекте контроля совпадают, то выполняется переход к анализу следующей цепочки правил.
- Если для команды, заданных в имитационной модели нет данных испытаний готового оборудования, то программа испытаний дополняется.
- Если для команды проведены испытания, но в имитационной модели такая команда отсутствует, то такая команда помещается в список ошибок модели.
- Если в модели не находится значений, соответствующих телеметрии объекта контроля, то формируется список ошибок, в который помещаются выбранные цепочки правил.

Выполнена программная реализация метода анализа имитационной модели. Созданы графические инструменты, которые позволяют визуально в интерактивном режиме строить цепочки зависимых правил. Пример графического представления цепочек правил, моделирующих отработку команд, показан на рисунке 1.22.

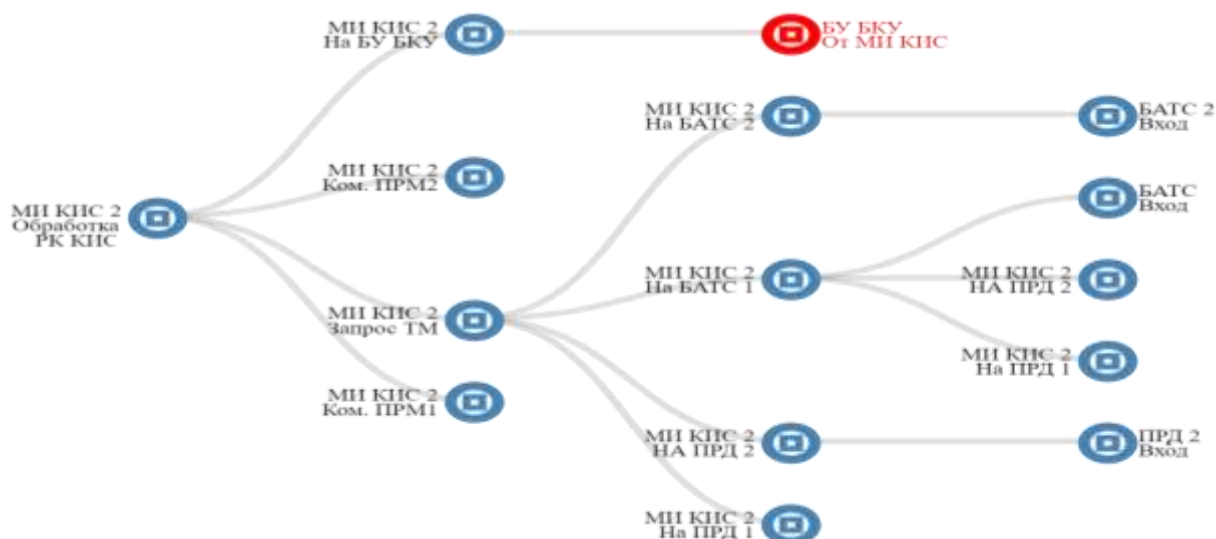


Рис. 1.22 – Графическое представление цепочек правил

Программное обеспечение позволяет выбирать связанные элементы, для которых построены цепочки и отображать полученные зависимости. На рисунке следующие обозначения узлов графа: внешний круг – определяет элемент модели, внутренний квадрат – тип логического элемента (передающий /принимающий интерфейс / таймер). Для каждого узла отображается расширенный список правил.

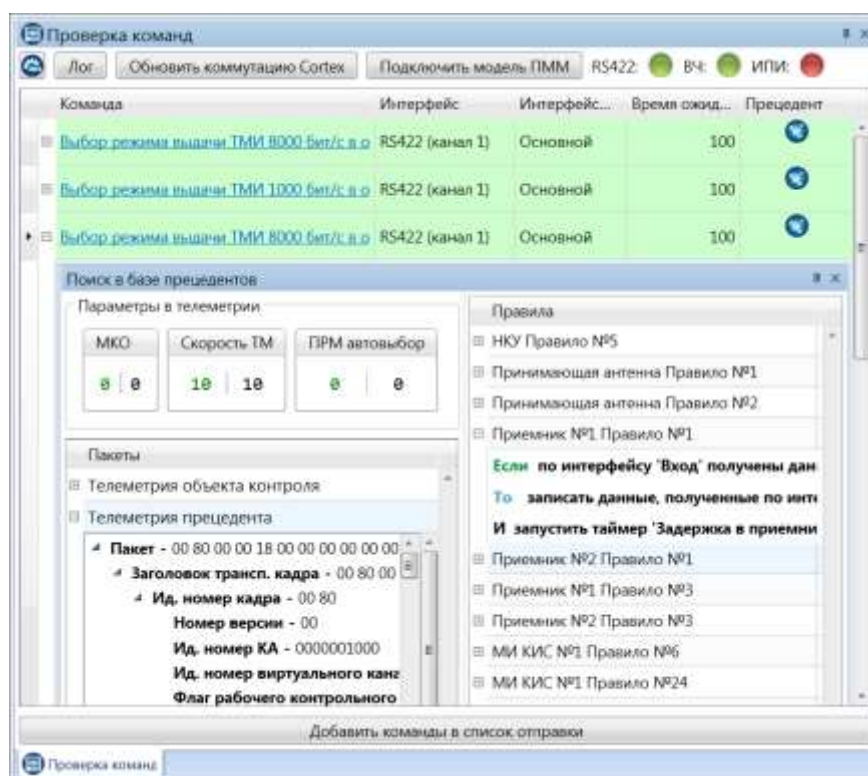


Рис. 1.23 – Результаты анализа базы знаний

Выполняется сравнение множества команд имитационной модели с множеством команд, для которых проводились испытания и сопоставление данных, получаемых в

результате проведения имитационных экспериментов с испытаниями бортовой аппаратуры. Пример работы программного обеспечения показан на рисунке 1.23.

В программном обеспечении отображаются команды, результаты испытаний, параметры телеметрии и прецеденты имитационного моделирования. Несовпадение результатов моделирования и данных, полученных от объекта контроля, требует доработки базы знаний имитационной модели.

После проведения анализа и доработки модель позволяет наглядно исследовать работу бортовой аппаратуры, имитирует широкий спектр команд, правил и значений параметров телеметрии, описывающих различные задачи функционирования и имитации командно-программного управления космическим аппаратом.

1.3. Веб-платформа для автоматизации процесса создания прикладных модельно-ориентированных систем сбора и анализа данных

Получен комплекс теоретических и прикладных результатов в ходе развития оригинального подхода к созданию прикладных систем сбора данных. Разработана и реализована программная платформа построения модельно-ориентированных систем, включая модель взаимодействия клиентской и серверной части, алгоритм автоматической генерации пользовательского интерфейса на основе управляющей модели и алгоритм построения аналитических отчетов. Разработаны модули платформы, обеспечивающие аналитическую обработку собранных данных, инфографическое представление результатов анализа с помощью встроенных инструментов визуализации и передачу данных отчетов сторонним системам анализа данных через REST-интерфейс. На базе платформы созданы системы мониторинга функционирования Красноярского водохранилища, анализа результативности научной деятельности и сбора и анализа данных диагностики загрязнения почв Красноярского края.

В рамках формализации авторского подхода к созданию модельно-ориентированных систем описано построение управляющих моделей систем сбора данных на основе анализа информационных потоков. Описаны шаги по выделению и формализации основных описательных характеристик объектов исследования, созданию форм ввода первичных данных об экспериментах и исследованиях. Подход проиллюстрирован на примере задачи оценки степени деградации почв, антропогенной и техногенной нагрузки на почвы методом биOLUMИнесцентного ферментативного анализа (рис. 1.24).

Ecological monitoring of soil cover | Data input

Korobko Anna

Movable nitrogen and phosphorus in the soil

- Minino. Meadow, virgin land. P. 4-6. A. 5
- Minino. Meadow, virgin land. P. 4-6. AB. 6th
- Minino. Meadow, virgin land. P. 4-6. B1k. 7th**
- Minino. Meadow, virgin land. P. 4-6. B2k. 8
- Minino. Meadow, virgin land. P. 4-6. Bk. 23
- Minino. Meadow, virgin land. P. 4-6. Ck. 24
- Minino. Meadow, virgin land. P. 4-6. Ck. 9
- Минино. Пашня. P. 6-1. Ak. 22
- Минино. Пашня. P. 6-1. An. 21
- Pogorelskiy Bor. Coniferous forest. P. 4-5. A1. 1
- Pogorelskiy Bor. Coniferous forest. P. 4-5. A1. 2

Objects of research

- Horizon Index
- Research
 - Soil section
 - Movable nitrogen and phosphorus in the soil**
 - Humus condition
 - Place of sampling
 - Section horizon

Minino. Meadow, virgin land. P. 4-6. B1k. 7th

Sample: 7

Section horizon:

Minimum depth, cm: 55

Maximum depth, cm: 60

Gumus, %: 0.31

Reaction of acidic soils (pH sol):

Current soil reaction (pH H₂O): 8.3

Reaction to alkali (NaOH):

Nitrification capacity (N-NO₃), mg / kg: 2.2

Phosphorus oxide (P₂O₅), mg / 100 g: 10.7

The value of the residual glow (T), %: 96.19

Standard deviation: 13.17

Рис. 1.24 – Форма ввода данных на примере задачи экологического мониторинга почв

В качестве развития средств аналитической обработки данных, консолидируемых прикладными модельно-ориентированными системами, предложен механизм гибридизации таких методов анализа и представления данных как инфографика, интеллектуальный анализ данных и геоинформационные технологии. Реализация различных комбинаций аналитических методов обеспечивает поддержку нативного исследования консолидированных данных, позволяя анализировать отдельные объекты, выбранные на карте или сгруппированные в ходе кластеризации, представляя их в виде современных диаграмм, таблиц или картограмм.

Разработанная на основе оригинальных методических и алгоритмических решений, Веб-платформа помимо типовых функций модельно-ориентированной системы по ведению, хранению и аналитической обработке данных, обеспечивает возможность динамического расширения тематического наполнения построенных систем на основе моделей, описывающих предметную область. На базе платформы в рамках создания модельно-ориентированной системы поддержки гидробиологических исследований Красноярского водохранилища разработан комплекс аналитических моделей, объединяющих абиотические, физико-химические и биологические показатели функционирования водохранилища. На рисунке 25 представлены данные о суммарной

численности водорослей из аналитической модели "Усреднённые показатели фитопланктона и абиотических показатели" в виде инфографики на карте.

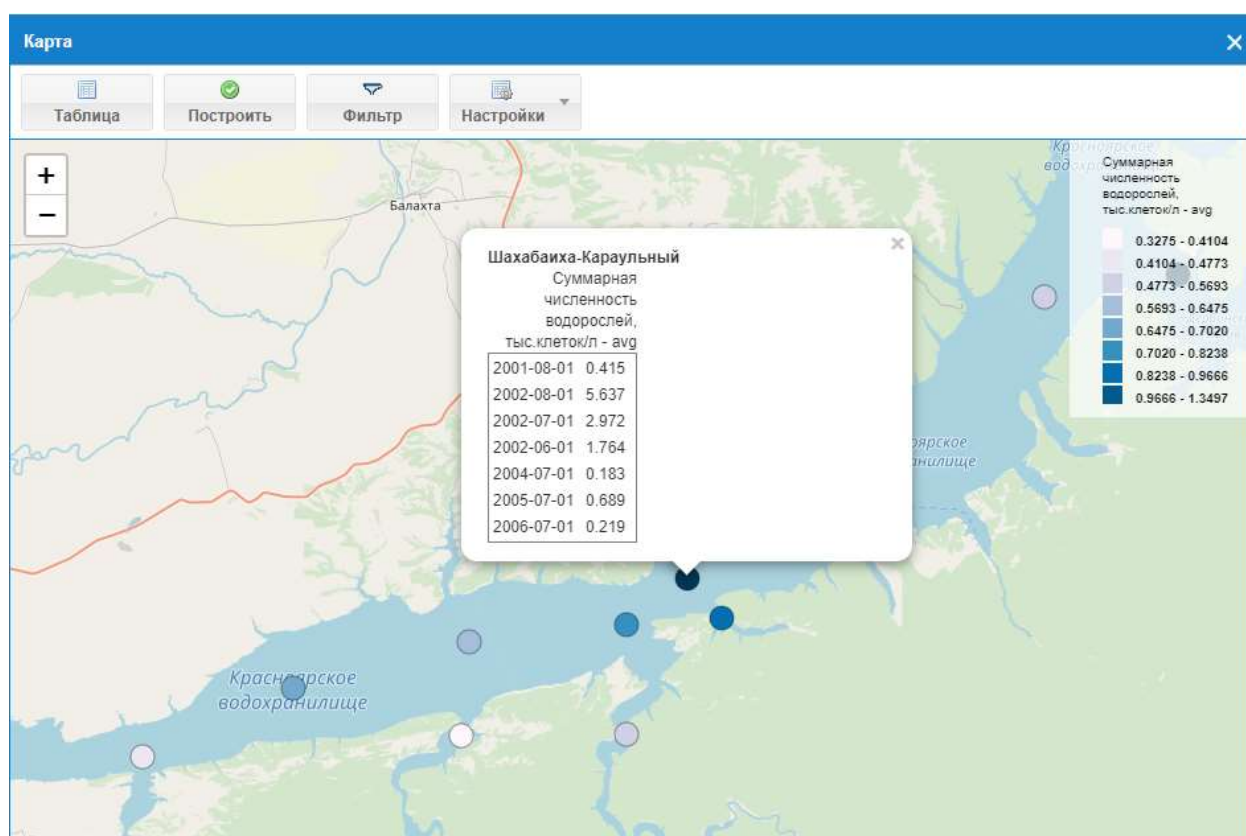


Рис. 1.25 – Отображение на карте суммарной численности водорослей из аналитической модели усреднённых показателей фитопланктона и абиотических показателей

Создаваемые на базе платформы системы получают набор инструментов для построения отчётов и начального анализа данных: таблица, линейная диаграмма, точечная диаграмма с возможностью кластеризации методом К-средних и картограмма. Для более продвинутого анализа имеется возможность использовать сторонние современные аналитические системы, такие как Power BI, Qlik Sense или Tableau. Эти системы позволяют объединять данные для анализа нескольких источников, проводить предварительную очистку и трансформацию данных, а также имеют богатый набор инструментов для построения отчётов. Для анализа данных сторонними инструментами в системах сбора данных, построенных на базе платформы, реализован REST-интерфейс передачи отчётов. Посредством данного интерфейса актуальные данные каждого отчёта в формате JSON могут быть переданы сторонним системам для анализа или другой обработки. Отчуждение данного функционала позволяет сосредоточиться на главной цели создаваемых систем – сборе данных.

Процесс построения модельно-ориентированных систем сбора данных на базе веб-платформы описан в виде методики в руководстве пользователя платформы. Методика включает следующие этапы: регистрация системы, заключающаяся в создании базы данных и под-домена; регистрация пользователей и назначение прав доступа; формирование системы путем создания и модификации прикладной и управляющей моделей с помощью интерфейса системы; наполнение системы с помощью интерфейса сбора данных; создание аналитических отчетов. Разработанная методика позволяет пользователям в кратчайшие сроки создавать и развивать модельно-ориентированные системы сбора данных на базе Веб-платформы, без привлечения программистов.

Созданные методические и программные средства обеспечили развитие функционала веб-платформы для построения модельно-ориентированных систем сбора данных с поддержкой полного цикла организации и обработки данных предметной области. Построение ряда систем иллюстрирует возможности платформы по созданию прикладных систем, формированию тематического ядра и анализу данных предметной области в режиме реального времени.

1.4. Развитие инфраструктуры научных коммуникаций

Цифровые репозитории открытого доступа являются одним из основополагающих элементов инфраструктуры открытой науки. Однако для обеспечения качественной интеграции в инфраструктуру, как репозитории, так и размещаемые в них объекты хранения должны отвечать определенным требованиям по поддерживаемым протоколам, схемам и структурам данных, а также их содержательному наполнению. Программное обеспечение, предназначенное для реализации цифровых репозиториях, построено на java- или xml-технологиях.

На основе тестовой модели выполнено развитие структуры институционального репозитория ИВМ СО РАН, развернутого на новой версии свободно распространяемой программной платформы с открытыми кодами DSpace 6.3 под ОС Linux OpenSuse, обеспечивающей хранение и доступ к электронным документам, отражающим результаты научных исследований. В соответствии с требованиями системы научной коммуникации документ в репозитории должен обладать качествами FAIR (Findable, Accessible, Interoperable and Re-usable), т.е. быть легко найден, доступен любому, у кого есть Интернет-связь, обеспечивать интероперабельность при манипуляциях с ним и давать возможность многократного использования. В качестве базовой схемы данных репозитории предлагают Дублинское ядро (DC) или Дублинское ядро с квалификаторами (QDC), оформленное как xsd-схема. Для создания специализированных коллекций,

например, тематических или музейных, а также разработки cгis-систем на основе репозитория на практике используются расширения базовой схемы данных. В международной практике существует ряд стандартизованных или рекомендованных расширений схемы, например, mods или dc-lib, в разной степени удовлетворяющих библиотечное сообщество по содержательному наполнению. Программное обеспечение репозитория должно обеспечивать преобразование данных репозитория в наиболее распространенные схемы с помощью xslt-преобразований. В ходе работы бы проведен анализ типов и видов хранимых объектов репозитория, а также схем преобразования данных на примере тестовых массивов библиотек ФИЦ КНЦ СО РАН. Создан дополнительный массив метаописаний «Наборы данных» для тестирования возможности их связей с объектами репозитория. В результате синтеза проведенных работ была сформирована модернизированная схема данных, совместимая по содержательному наполнению как с DC, так и с наиболее распространенными стандартизованными схемами, принятыми в практике библиотечной работы, а также xslt-преобразования для таких схем. Содержательная совместимость схем метаданных объектов репозитория позволила использовать для кооперации с внешними информационными ресурсами обычные запросы протоколов SRU/SRW и OAI/PMH без потерь информативности. Использование стандартизованных международных форматов позволило обеспечить интероперабельность на уровне данных при построении интегрированных распределенных систем, а оптимизация поисковой системы репозитория углубить и расширить интеграцию с поисковыми системами Интернет.

Интеграция открытого архива с Интернет-сервисами выполнена в следующих направлениях: оптимизация индексирования поисковыми машинами Интернет с отслеживанием ботов и выборкой индексируемых страниц; подключение сервисов лицензирования публикаций Creative Common; подключение служб проекта Sherpa/Romeo, позволяющей отслеживать соответствие доступа к публикациям издательским ограничениям, что особенно важно для публикаций в зарубежных журналах, отличающихся большим разнообразием правил доступа.

На основе тестовой модели модернизирована структура институционального репозитория ИВМ СО РАН, расширен круг применяемых web-технологий, схем и форматов метаданных цифровых объектов и системы их хранения. Кооперация с внешними ресурсами организована на основе системы запросов протоколов SRU/SRW и OAI/PMH. Исследованы особенности форматов MODS, METS и QDC, используемых для преобразования и хранения данных в репозитории без потерь информативности. Оптимизирована поисковая система репозитория. Исследование возможностей

существующих web-технологий открытых архивов, их адаптация, развитие и внедрение позволяет создавать и продвигать собственные ресурсы и результаты исследований, эффективно и согласованно влияя на формирование информационной среды Интернет. Институциональный репозиторий, формируемый в технологиях архивов открытого доступа, позволяет произвести качественный учет и продвижение научных публикаций, обеспечить открытость результатов научных исследований для мирового сообщества, повысить индивидуальные показатели цитирования.

1.5. Разработка и апробация средств анализа и оценки угроз кибербезопасности в корпоративной сети

1.5.1. Выявление сетевых аномалий для противодействия киберугрозам

Выполнены исследования, направленные на снижение риска киберугроз за счет привлечения методов анализа больших данных. Источником данных для анализа являлись журналы различных интернет-сервисов. Цель работы – исследование возможности группировки потребителей Интернет-ресурсов и выявления проблемных групп, характеризующихся аномальной активностью. Выявление аномалий может свидетельствовать об активности компьютерных вирусов, утечках информации и прочих угрозах кибербезопасности организаций. С целью повышения эффективности функционирования систем информационной безопасности в настоящее время активно развиваются системы управления событиями информационной безопасности (Security information and event management). Основная их функция – координация действий и отражение актуального состояния всех систем, что полностью согласуется с концепцией кибербезопасности. Разработана функциональная схема универсальной сигнатурной SIEM системы (рис. 1.26), которая может настраиваться с учетом технических, программных и правовых особенностей организации. Конфигурирование производится наполнением СУБД, содержащей общие параметры системы, сигнатуры угроз и сценарии реагирования на выявленные угрозы, а также может содержать базу знаний, содержащую прецеденты угроз для дальнейшего предотвращения. Переход к использованию системы управления событиями информационной безопасности позволяет в полной мере осуществлять контроль за состоянием различных систем безопасности, которые в комплексе позволяют наиболее качественно выявлять и противодействовать различным киберугрозам организации.

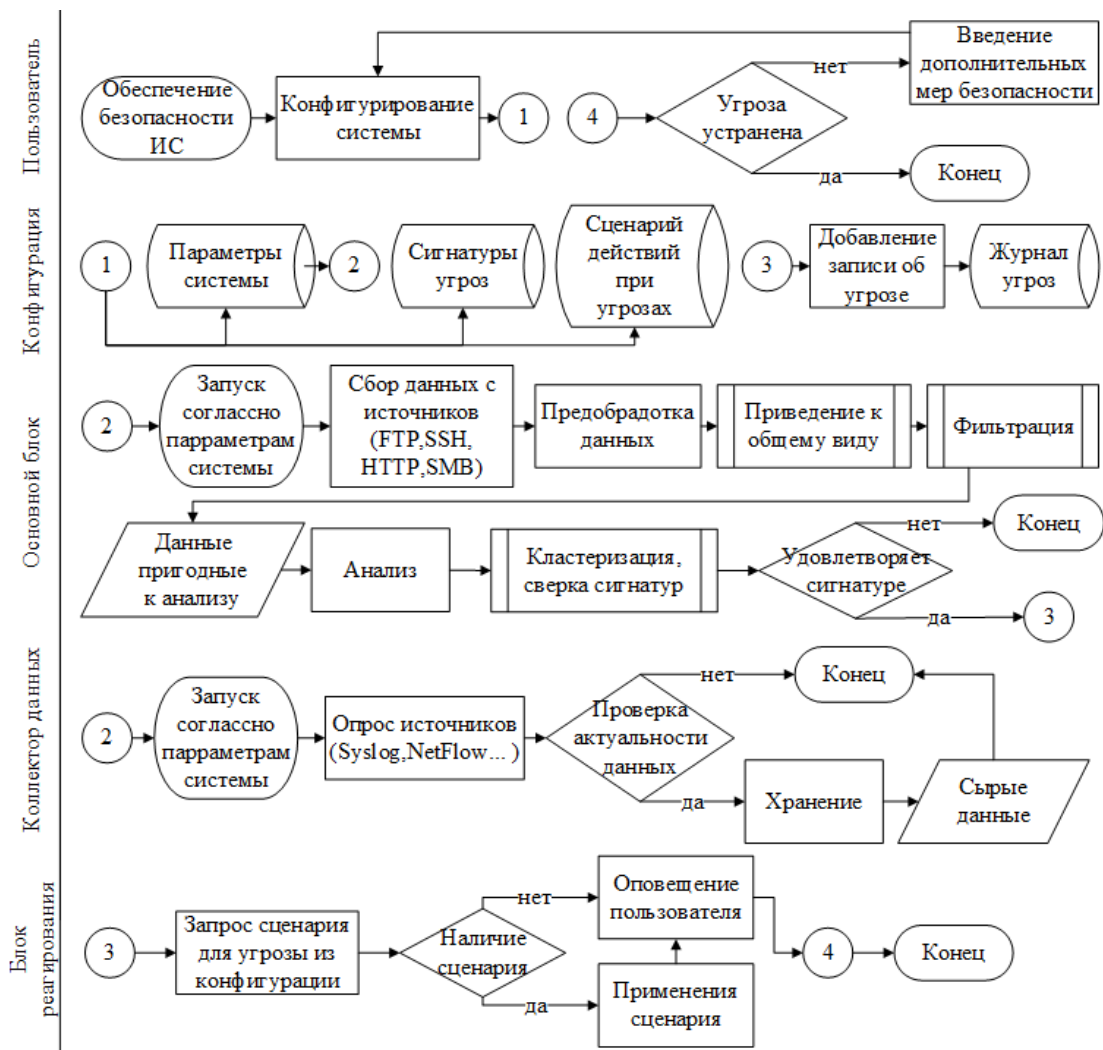


Рис. 1.26 – Модель универсальной SIEM системы

Проанализированы наиболее эффективные подходы к контролю и управлению сетевой инфраструктурой, представленные в «библиотеке информационно-технической инфраструктуры» (ITIL – IT Infrastructure Library), объединившей в себе доказавшие свою эффективность практики, применяемые в управлении ИТ-компаниями. Описаны базовые элементы инфраструктуры сети ФИЦ КНЦ СО РАН. В качестве основных информационных систем приняты система электронной почты и прокси-сервер для контроля доступа пользователей к сети интернет. В свою очередь, пользователи на рабочих местах могут использовать персональные компьютеры, ноутбуки, принтеры, сетевые хранилища и различные персональные переносные устройства. Была разработана информационная модель необходимых инфраструктурных объектов и перечень их атрибутов. Согласно ITIL, одним из хороших решений для контроля за подобной инфраструктурой является база данных управления конфигурациями (CMDB). Для рассмотренной инфраструктуры и набора данных предложено реализовать CMDB в виде веб-сервиса (рис. 1.27), учитывающего необходимые особенности. Помимо хранения

информации данное решение должно в автоматическом режиме изменять конфигурацию сетевого оборудования и манипулировать учётными данными пользователей на информационных системах. Например, при добавлении в сервис новых учётных записей пользователя прокси-сервера они автоматически добавляются в его конфигурацию, исключая необходимость дополнительных манипуляций со стороны администратора.

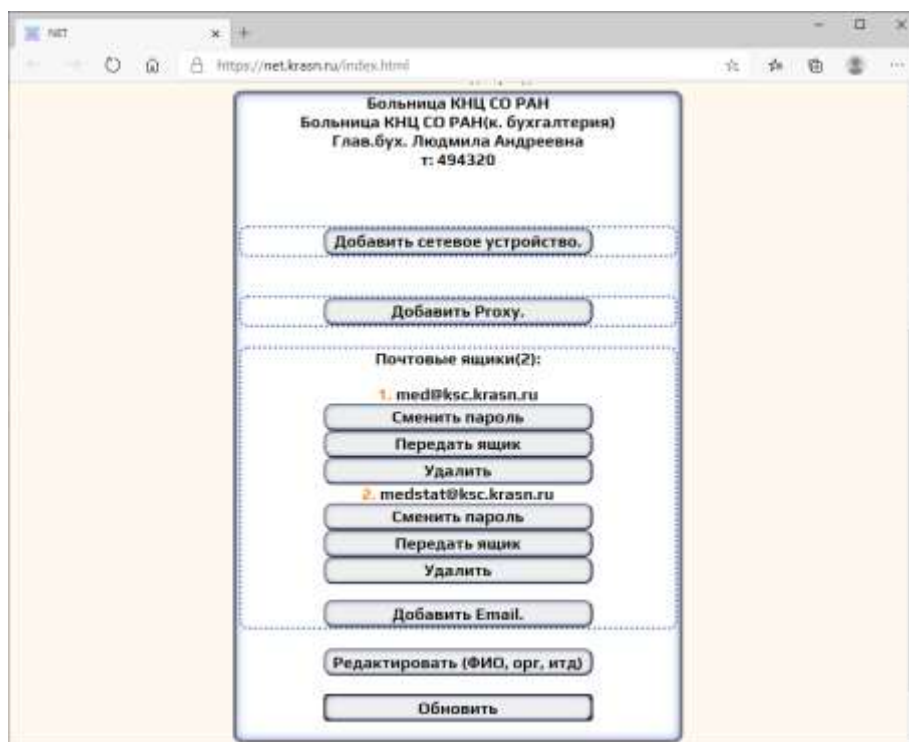


Рис. 1.27 – Пример интерфейса *CMDB*

Реализован предложенный вариант CMDB системы в виде web-сервиса, позволяющего вести учёт ресурсов информационно-технической инфраструктуры и контролировать её актуальную конфигурацию, что уменьшает вероятность ошибок администрирования, которые могут привести к уязвимостям в информационной безопасности предприятия. Данный сервис при необходимости можно адаптировать для использования с существующими ИТ сервисами такими, как службы поддержки для ускорения обработки запросов пользователя, так и в системы управления информационной безопасностью (SIEM), предоставив тем самым информацию о пользователях и устройствах

Проведены исследования по разработке методов оценки различных аспектов кибербезопасности информационно-телекоммуникационных систем. При использовании интернет-ресурсов накапливается большое количество данных, как на стороне пользователей, так и на стороне интернет-провайдеров. Обработка таких данных с привлечением различных методов Big-data, позволяет получить новые знания о поведении

интернет-пользователей. Выявление аномалий потребления ресурсов основано на анализе результатов кластеризации пользователей Интернет-ресурсов за разные промежутки времени и может свидетельствовать об активности компьютерных вирусов, утечках информации и прочих угрозах кибербезопасности организаций.

В качестве исходных данных использованы журналы сервера доступа в интернет за месяц. Общий объем данных составляет около 8 Гигабайт, а количество обрабатываемых запросов около 100 миллионов. Полученные данные были нормализованы с помощью введения справочников доменов и пользователей. Затем выполнено понижение размерности пространства характеристик, с помощью перехода от адресов ресурсов (около 4000) к тематикам сайтов (около 160). Полученные данные были разбиты на кластеры, с помощью метода иерархической кластеризации. При расчете расстояний использовался алгоритм минимизации дисперсии Уорда. Наиболее показательными оказались разбиения на 2 и 3 кластера. Для визуализации полученных результатов использованы дендрограммы, метод главных компонент (PCA) и линейный дискриминантный анализ (LDA). Наиболее наглядный результат для трех и большего количества кластеров показал метод LDA (рис. 1.28).

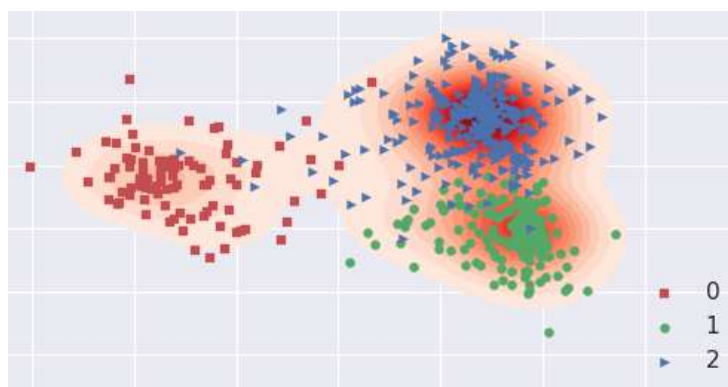


Рис. 1.28 – Результат применения LDA для визуализации 3-х кластерной структуры пользователей Интернет-ресурсов

Предложенная методика анализа может быть применена для обнаружения аномалий использования интернет ресурсов, как визуально, так и с помощью выделения одиночных наиболее удаленных элементов. Обнаруженные аномалии могут рассматриваться как один из аспектов снижения кибербезопасности. Дополнительно может быть реализован динамический контроль распределения пользователей по кластерам. Признаком опасного изменения является переход пользователя в отдельный кластер. Перспективными представляются дальнейшие исследования в данном направлении.

1.5.2. Мониторинг кибербезопасности корпоративной сети

Выполнены работы по мониторингу кибербезопасности корпоративной сети Федерального исследовательского центра «Красноярский научный центр СО РАН» и развитию моделей безопасности интернет-сервисов. Источниками данных для анализа являлись:

1. Данные по IP-трафику Netflow: объем более 400 Гб, более 2 млрд. элементов;
2. Данные журналов почтового сервера за год: более 1.5 млн. элементов;
3. Данные журнала системы блокирования угроз за год: около 200 тыс. элементов;
4. Данные журналов www-сервисов за 1 год: объем около 12 Гб, более 44 млн. элементов.

Анализ показал наличие постоянного потока попыток несанкционированного доступа. При посуточном агрегировании наблюдается тренд на изменение количества попыток доступа более чем в 2 раза за год. Выявленные пики обусловлены продолжительностью атак, при средней их интенсивности.

Анализ данных частоты сканирования отдельных сервисов (рис. 1.29) позволил выделить популярные сервисы, на которые производится наибольшее количество атак и под которые маскируются угрозы: Telnet/23, MS SQL/1433, HTTP/80, Personal Agent/5555, SSH/22, HTTP Alternate/8080, RDP/3389. Эти сервисы могут служить дополнительными индикаторами атак. По результатам анализа, протоколы Telnet и MS SQL были добавлены к сервисам функционирующей системы блокирования (SSH, RDP, SMTP), что повысило ее эффективность.

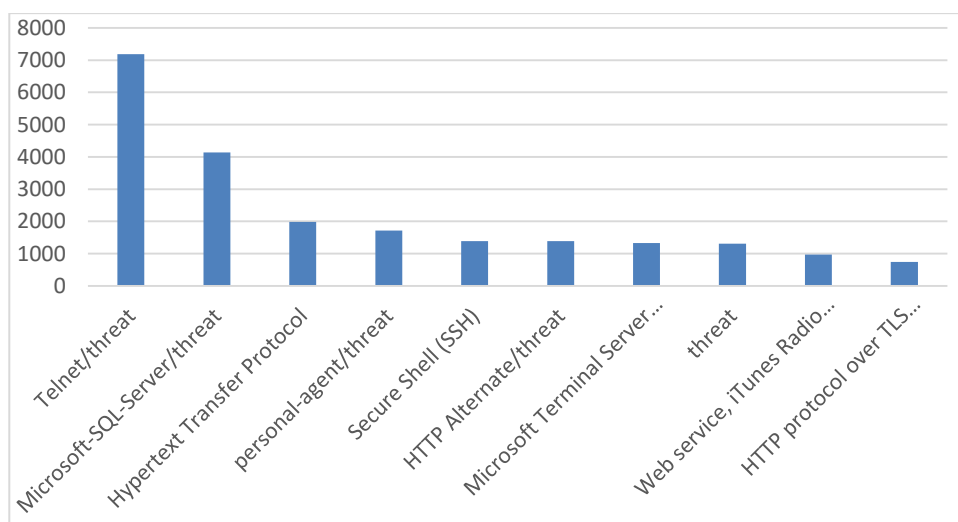


Рис. 1.29 – Частоты сканирования по сервисам

Проведен анализ данных системы электронной почты, который позволил выявить географическое распределение источников несанкционированной почты (рис. 1.30), а также показал различие в источниках вирусов и массовых рассылок.

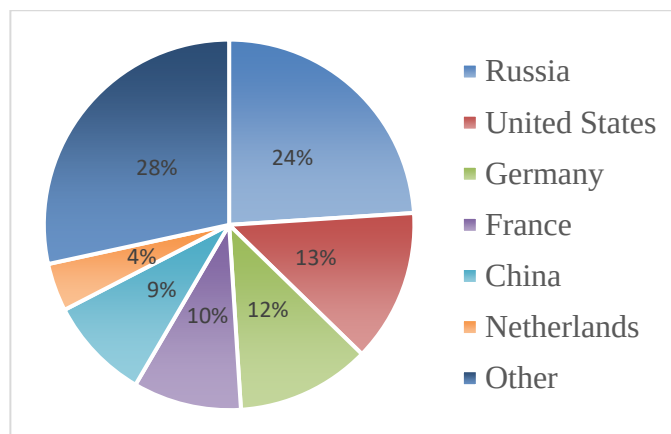


Рис. 1.30 – Географическое распределение источников несанкционированной почты

Апробированы и применены новые методы противодействия массовым рассылкам, основанные на анализе корреспонденции публичных адресов, за счет которых удалось в несколько раз уменьшить процент спама (рис. 1.31).

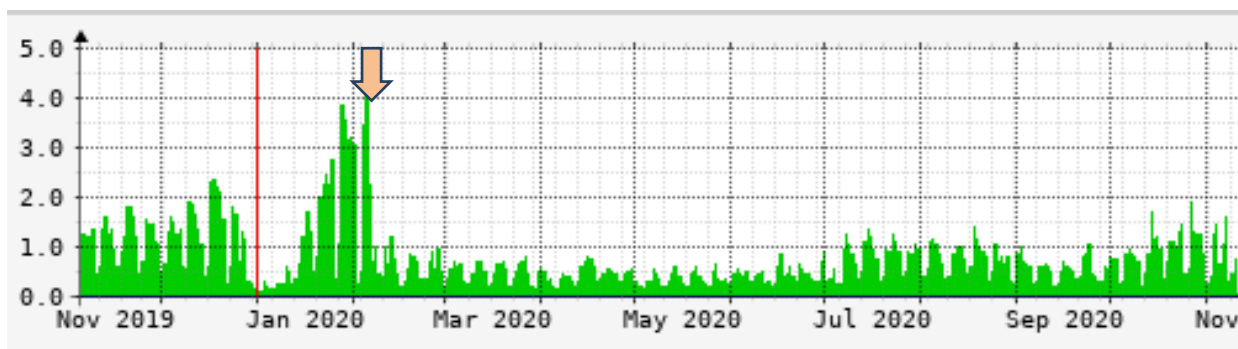


Рис. 1.31 – Скорость поступления писем, идентифицированных как спам

Анализ данных сервиса блокирования вторжений позволил сделать вывод о существовании большой сети источников угроз, запускаемых в 00 часов GMT. Анализ географического положения источников угроз показывает подавляющее лидерство Китая (39%) и США (12%). Таким образом, можно сделать вывод о целенаправленных попытках вторжения со стороны Китая, т.к. по остальным данным он не имеет лидирующего положения.

Проведенный анализ данных журналов веб-сервисов показывает их периодичность и тенденцию увеличению количества запросов, связанную с развитием сервиса и увеличением числа посетителей. Географическое распределение запросов (Россия – около 80% всех посещений, США – 7%, Германия – 2%) и ошибок (Россия – 54%, США – 27%, Китай – 3%) демонстрирует повышенную активность попыток доступа к несуществующим или закрытым разделам со стороны США и Китая. Это обуславливает существенную вероятность повышенного количества источников угроз в этих странах.

Была посчитана корреляция и произведена нормировка запросов и ошибок по времени сервера и клиента (рис. 1.32). Выявлено, что высокий коэффициент корреляции ошибок вызван неправильной работой веб-сайтов и веб-сервисов. Однако оставшаяся часть ошибок показывает наличие сканирований и атак со стороны веб-пауков.

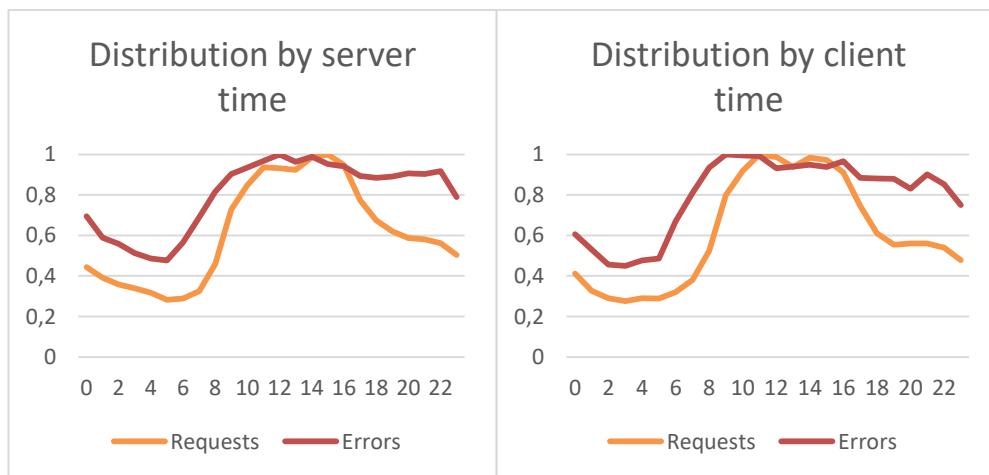


Рис. 1.32 – Распределение запросов и ошибок по серверному и клиентскому времени

Преобладающая доля ошибок приходится на выявление уязвимостей популярных систем управления сайтами (CMS). Кроме того, анализ доли различных браузеров в журналах запросов и ошибок показывает повышенный процент ошибок со стороны веб-пауков, что может являться признаком угрозы.

Апробирован и применен новый метод противодействия попыткам подбора паролей к интернет-сервисам, что позволило уменьшить количество попыток подбора паролей с нескольких десятков тысяч до нескольких сотен (рис. 1.33).

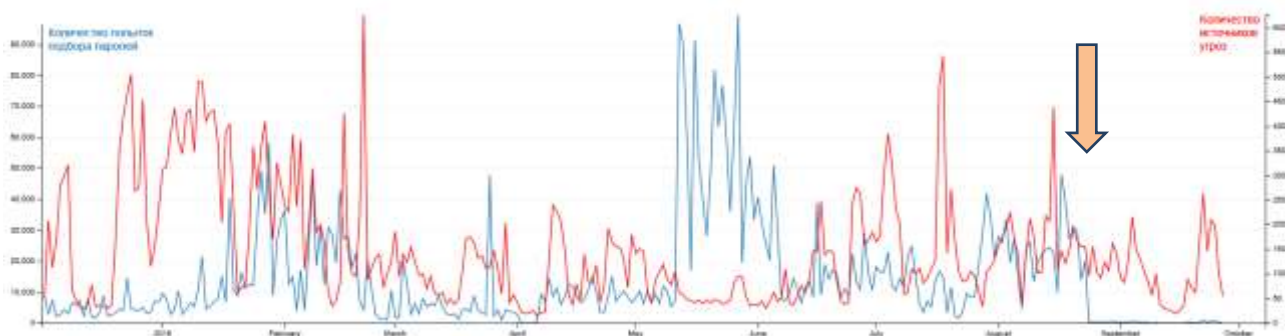


Рис. 1.33 – Анализ попыток подбора паролей и источников угроз с помощью системы обнаружения аномалий

По результатам проведенного анализа было выполнено следующее: 1) увеличен пороговый временной интервал для более уверенного обнаружения сканирования сети; 2) в систему мониторинга добавлены новые порты для отслеживания вредоносной активности; 3) оптимизированы настройки межсетевых экранов для более эффективного блокирования нежелательных хостов; 4) обновлены настройки веб-сервера для предотвращения атак на уязвимости известных CMS; 5) выполнена оптимизация сетевых настроек внутренних коммутаторов для блокирования нежелательного трафика между различными подразделениями.

1.5.3. Разработка в расширенной ролевой модели безопасности на основе RBAC

В рамках развития методики разработки защищенных веб-приложений предложена новая модель разграничения доступа на основе параметрической информации. Данная модель основана на базовой ролевой модели RBAC (Role Based Access Control), описанной в работах Sandhu, и является дальнейшим продолжением работы по созданию моделей безопасности, адаптированных для функционирования в веб-приложениях и веб-сервисах.

Разграничение доступа с использованием классической ролевой модели описано в работе Sandhu. Базовая модель ролевого разграничения доступа RBAC определяется следующим образом:

$$\langle U, R, P, S, UA(U), PA(R), user(S), roles(S) \rangle,$$

где U — множество пользователей, R — множество ролей, P — множество прав доступа, S — множество сессий пользователей. Также задаются следующие функции:

- $UA: U \rightarrow 2^R$ — функция, определяющая для каждого пользователя множество ролей, на которые он может быть авторизован;
- $PA: R \rightarrow 2^P$ — функция, определяющая для каждой роли множество прав доступа, при этом $\forall p \in P, \exists r \in R$ такой, что $p \in PA(r)$;
- $user: S \rightarrow U$ — функция, определяющая для каждой сессии пользователя, от имени которого она авторизована;
- $roles: S \rightarrow 2^R$ — функция, определяющая для пользователя множество ролей, на которые он авторизован в данной сессии; при этом в каждый момент времени $\forall s \in S$ выполняется условие $roles(s) \subseteq UA(user(s))$.

В предыдущих работах создана расширенная модель безопасности на основе RBAC, которая учитывает специфику работы веб-приложений. Расширенная ролевая модель безопасности дополняет базовую модель RBAC новыми элементами: Tk (токен),

Rq (запрос). Токен (Tk) представляет собой набор атрибутов пользователя, позволяющих осуществить его аутентификацию в системе. Запрос (Rq) — набор информации, которая пересылается между клиентом и сервером по протоколу HTTP(S). На множестве запросов вводится бинарное отношение включения, которое задаёт нестрогий частичный порядок на множестве запросов Rq и позволяет упорядочить запросы согласно иерархии. Также задается функция $RqA()$, отображающая права доступа на множество запросов: $RqA: P \rightarrow 2^{Rq}$. Для всех запросов из множества Rq вводится иерархия запросов RqH. Данная модель учитывает специфику работы веб-приложений и позволяет разграничивать доступ на основе пути запроса (URI). Однако при разработке и внедрении на практике различных политик безопасности возникла необходимость дополнительного разграничения доступа, в результате было создано три новые модели безопасности: параметрическая, сервис-ориентированная и мультидоменная.

Параметрическая модель безопасности создана на базе расширенной ролевой модели. В модель были добавлены новые элементы: RqP (параметр запроса), функции $RqPA()$, $params()$. Полученная новая модель описывается следующим образом:

$\langle U, R, P, S, Tk, Rq, RqP, UA(U), PA(R), RqA(P), RqPA(P), user(S), roles(S), token(Tk), requests(S), params(Rq) \rangle$.

Введено новое определение: параметр запроса (RqP) — наборов пар (ключ, значение) запроса HTTP(S), который принадлежит запросу Rq и используется для передачи дополнительных данных в запросе. Задана новая функция $RqPA()$, отображающая разрешения на множество параметров: $RqPA: P \rightarrow 2^{RqP}$. Также задана новая функция $params()$, отображающая запросы на множество параметров: $params: Rq \rightarrow 2^{RqP}$.

Полученная модель безопасности позволяет разграничивать доступ с использованием не только пути запроса (URI), но и параметров запроса. Дополнительное разграничение доступа может быть задано явно с использованием функции отображения $RqPA()$ или с помощью специального метаязыка описания фильтрации. Созданная модель не ограничивается фиксированным набором параметров, а предусматривает динамическое изменение данного набора в зависимости от выполнения заданных условий.

Сервис-ориентированная модель безопасности создана на основе параметрической модели безопасности для обеспечения разграничение доступа в веб-сервисах. В параметрическую модель безопасности были добавлены новые элементы: Rq

(метод запроса) и дополнительные функции RqMA(), method(). Полученная новая модель описывается следующим набором элементов:

$\langle U, R, P, S, Tk, Rq, RqM, RqP, UA(U), PA(R), RqA(P), RqMA(P), RqPA(P), user(S), roles(S), token(Tk), requests(S), method(Rq), params(Rq) \rangle$.

Задано новое определение: метод запроса (RqM) — элемент запроса Rq, который используется для передачи действия при обработке запроса по протоколу HTTP. Метод запроса RqM принадлежит запросу Rq. Задана новая функция RqMA(), отображающая разрешения на множество методов: $RqMA: P \rightarrow 2^{RqM}$. Также задана новая функция method(), отображающая запросы на разрешённые методы: $method: Rq \rightarrow RqM$.

Полученная модель безопасности позволяет разграничивать доступ с использованием не только пути и параметров запроса, но и метода запроса. Данный подход позволяет гибко разграничивать доступ в веб-сервисах, использующих архитектуру REST. Также подход применим для веб-сервисов, использующих подмножество технологии RPC (Remote Procedure Call), например, gRPC. В этом случае множество методов RqM будет содержать набор методов, определённых в программном интерфейсе (API). Разработанная модель безопасности позволяет повысить защищённость веб-сервисов за счёт применения более гибкого разграничения доступа на основе методов запросов и может быть использована для решения широкого круга задач с применением технологий REST и RPC.

Мультидоменная модель безопасности создана на основе сервис-ориентированной модели безопасности. Разработка данной модели вызвана необходимостью дополнительного разграничения доступа в сложных системах, функции которых декомпозированы в отдельные подсистемы, имеющие собственные пространства имен (например, домены). Чтобы учесть требования таких систем, в существующую расширенную ролевую модель безопасности была добавлена возможность мультидоменного разграничения доступа. Для этого в модель добавлены новые элементы: RqD (домен запроса) и функции RqDA(), domains(), domain(). Мультидоменная модель безопасности описывается следующим образом:

$\langle U, R, P, S, Tk, Rq, RqD, RqM, RqP, UA(U), PA(R), RqA(P), RqDA(P), RqMA(P), RqPA(P), user(S), roles(S), token(Tk), requests(S), domains(S), domain(Rq), method(Rq), params(Rq) \rangle$.

Введено новое определение: домен запроса (RqD) — элемент запроса Rq, который используется для передачи доменного имени узла, к которому производится обращение при обработке запроса по протоколу HTTP(S). Домен запроса RqD принадлежит запросу,

один запрос может выполняться на нескольких доменах. Задана функцию $RqDA()$, отображающая разрешения на множество доменов: $RqDA: P \rightarrow 2^{RqD}$. Также задана функция $domains()$, отображающая множество доменов для данной сессии: $domains: S \rightarrow 2^{RqD}$. Для ограничения запросов задана функция $domain()$, отображающая запросы на домены: $domain: Rq \rightarrow RqD$. Схема элементов модели представлена на рисунке 1.34.

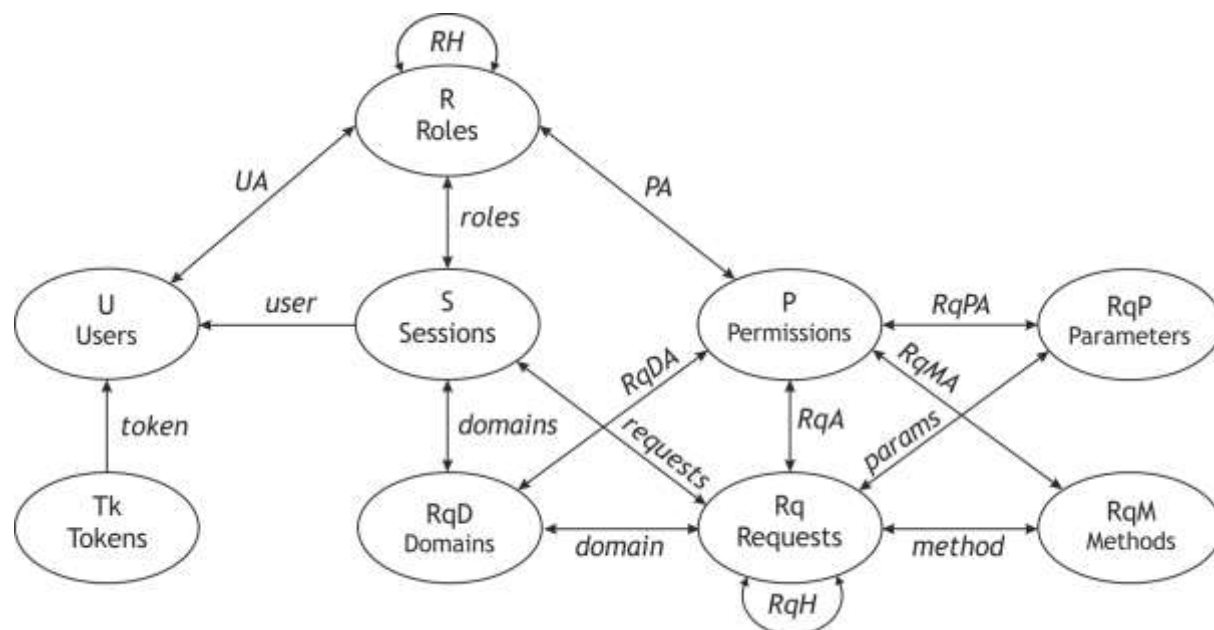


Рис. 1.34 – Схема элементов мультидоменной расширенной ролевой модели безопасности

Полученная модель безопасности позволяет разграничивать доступ с использованием следующих элементов: путь запроса (URI), параметр запроса, метод запроса, домен. С помощью функции $RqDA()$ можно ограничить набор доменов, которые могут выполняться для данного разрешения. Функция $domain()$ позволяет разграничить использование доменов для каждого отдельного запроса. Данное расширение модели позволяет осуществлять дополнительное разграничение доступа для сложных систем, имеющих несколько пространств имён для запросов.

Критерии оценки безопасности при разработке веб-приложений

В рамках развития методики разработки защищенных веб-приложений предложен метод интегральной оценки защищенности веб-приложений. Метод основывается на международном стандарте ASVS (Advanced Security Verification Standard), который содержит требования безопасности, применяемые при проектировании, разработке и тестировании современных веб-приложений. Данный стандарт определяет три уровня верификации в зависимости от типа системы. Каждый уровень верификации определяет набор критериев, которые необходимо проверить для оценки защищенности системы.

Критерии оценки стандарта ASVS сгруппированы по категориям, например, аутентификация, управление сессиями, безопасность HTTP и т.д. В рамках развития метода оценки защищенности веб-приложений в данный список была добавлена группа критериев для проверки модели безопасности, включая: 1) проверку выбранной модели безопасности на соответствие целям и задачам; 2) проверку элементов модели на полноту охвата предметной области; 3) проверку элементов модели на непротиворечивость; 4) проверку алгоритмов контроля доступа на покрытие тестовыми данными.

Разработан метод интегральной оценки защищенности веб-приложений на основе векторов групповых оценок критериев. Каждый критерий нормализуется в интервале $[0,1]$, где 0 – полное невыполнение требований критерия, а 1 – полное соответствие требованиям критерия. Для каждой группы критериев строится вектор из нормализованных значений критериев $\mathbf{K}_i = (k_{i1}, k_{i2}, \dots, k_{im_i})$, где m_i – количество критериев внутри i -й группы, n – число групп критериев. Задается вектор групповых оценок критериев $\mathbf{G} = (g_1, g_2, \dots, g_n)$. Для подсчета групповых оценок используется квадрат Евклидова расстояния между нулевым вектором 0 и $\mathbf{K}_i$ в пространстве размерностью m_i с нормализацией в интервале $[0,1]$: $g_i = \sum_{j=1}^{m_i} k_{ij}^2 / m_i$.

Чем ближе полученное значение g_i к 1, тем больше критериев внутри группы удовлетворяют заданным требованиям. На основе вектора групповых оценок безопасности построена интегральная оценка безопасности с использованием средневзвешенного значения: $I = \sum_{i=1}^n \omega_i g_i$, где $\sum_{i=1}^n \omega_i = 1$.

Коэффициенты ω_i назначаются экспертами для каждой группы критериев, что позволяет увеличить или уменьшить вес определенной группы в итоговой оценке. Данный подход обеспечивает гибкость при оценке информационных систем, в которых компоненты вносят разные веса в итоговую оценку.

Применение разработанных методов и программных продуктов позволяет эффективно выявлять и противодействовать киберугрозам на уровне информационно-телекоммуникационных систем организации и проводить комплексную оценку кибербезопасности организации.

Основные результаты исследований реализованы, апробированы и внедрены в корпоративной сети ФИЦ КНЦ СО РАН, а также в ряде информационных систем муниципального управления города Красноярск.

2 НЕПАРАМЕТРИЧЕСКИЕ СИСТЕМЫ ПРИНЯТИЯ РЕШЕНИЙ В УСЛОВИЯХ НЕОДНОРОДНЫХ ДАННЫХ

Ответственный исполнитель д.т.н. Лапко А.В.

2.1. Быстрые алгоритмы оптимизации непараметрических решающих правил ядерного типа в условиях статистических данных большого объёма

Разработаны статистические методы распознавания образов и восстановления многомерных зависимостей в условиях выборок неоднородных данных большого объёма, основанные на дискретизации области их определения, принципах коллективного оценивания и гибридного моделирования. Установлены условия асимптотической сходимости соответствующих им непараметрических решающих функций. На этой основе определены аналитические зависимости показателей эффективности оценивания решающих функций от особенностей исходных данных и используемых методов их обработки.

Разработаны непараметрические алгоритмы распознавания образов и модели стохастических зависимостей, которые обеспечивают эффективную обработку информации в условиях малых выборок. Идея синтеза предложенных непараметрических систем основана на декомпозиции исходных обучающих выборок по размерности, построении частных решающих функций и последующем их обобщении в виде линейных либо нелинейных коллективов решающих функций. Исследованы свойства показателей эффективности подобных систем по сравнению с традиционными алгоритмами и моделями ядерного типа.

Предложены и исследованы быстрые алгоритмы оптимизации непараметрических решающих правил ядерного типа в условиях статистических данных большого объёма. Обоснован и разработан быстрый выбор оптимальных коэффициентов размытости ядерных функций в непараметрических оценках многомерной плотности вероятности типа Розенблатта-Парзена, уравнения разделяющей поверхности в двухальтернативной задаче распознавания образов. Их применение позволяет на несколько порядков сократить временные затраты при синтезе непараметрических решающих правил.

На основе анализа асимптотических свойств регрессионной оценки плотности вероятности предложен метод оптимальной дискретизации области значений двумерной случайной величины, что является одной из актуальных задач математической статистики. Разработаны процедуры оценивания нелинейного функционала от плотности

в формуле расчёта оптимального количества интервалов дискретизации. Полученные результаты обобщены на многомерный случай на основе использования метода аналогий.

Методика проверки гипотез о распределениях случайных величин на основе непараметрических алгоритмов распознавания образов развита на многомерные законы распределения в условиях статистических данных большого объёма. Обоснована возможность замены задачи проверки гипотезы о тождественности двух законов распределений многомерных случайных величин на проверку гипотезы о равенстве ошибки распознавания образов пороговому значению 0,5. Обучающая выборка формируется по статистическим данным сравниваемых законов распределения. При больших объёмах статистических данных синтез непараметрического алгоритма распознавания образов осуществляется на основе регрессионных оценок плотностей вероятности распределения случайных величин в классах. Предлагаемые алгоритмы распознавания образов позволяют уменьшить объём обучающей выборки за счёт декомпозиции области значений случайных величин.

Предложена и исследована новая методика проверки гипотезы о независимости случайных величин, которая основывается на использовании непараметрического алгоритма распознавания образов, соответствующего критерию максимального правдоподобия. В отличие от традиционной постановки задачи распознавания образов априори отсутствует обучающая выборка. Законы распределения случайных величин в классах оцениваются по исходным статистическим данным для условий их зависимости и независимости. По минимальному значению оценки вероятности ошибки распознавания образов в классах принимается решение о зависимости либо независимости случайных величин.

На основе предложенных непараметрических систем принятия решений разработана информационная система для оценивания состояний природных объектов по данным дистанционного зондирования, которая использована при исследовании состояний лесных массивов.

2.2. Статистические методы распознавания образов и восстановления многомерных зависимостей в условиях неоднородных данных

Разработана методика синтеза и анализа новой непараметрической оценки многомерной решающей функции для двухальтернативной задачи распознавания образов в условиях обучающих выборок большого объёма и неоднородных данных. Её синтез осуществляется путём декомпозиции исходной информации и анализа количественных характеристик получаемых множеств случайных величин на основе кривой регрессии.

Определены условия асимптотической несмещённости и состоятельности подобной регрессионной оценки многомерной решающей функции, соответствующей оптимальному байесовскому уравнению разделяющей поверхности между классами. На этой основе определена аналитическая зависимость показателей качества аппроксимации исследуемой решающей функции от методов дискретизации пространства признаков классифицируемых объектов и их размерности, что позволило сформировать эффективную методику оптимизации непараметрических алгоритмов распознавания образов.

Исследован альтернативный подход построения непараметрических уравнений разделяющих поверхностей между классами в условиях больших выборок, основанный на декомпозиции исходных данных по их объёму, использовании принципов коллективного оценивания и технологии параллельных вычислений. Установлены условия асимптотической сходимости предложенного линейного коллектива решающих функций в двухальтернативной задаче распознавания образов и определена их зависимость от количества его составляющих. По сравнению с традиционной непараметрической решающей функции исследуемый коллектив имеет меньшую дисперсию, но несколько большее значение асимптотического выражения среднеквадратического отклонения. Это объясняется увеличением смещения коллектива непараметрических решающих функций, так как снижаются объёмы обучающих выборок при оценивании его составляющих. Неравномерность распределения элементов обучающей выборки между классами снижает аппроксимационные свойства коллектива непараметрических решающих функций. Данная тенденция особенно характерна для малых размерностей признаков классифицируемых объектов.

Методика синтеза гибридных систем распознавания образов, сочетающих преимущества параметрических и локальных аппроксимаций решающих функций, обобщена на условия неоднородных данных. Обучающая выборка в подобных условиях формируется при расширении возможности системы контроля, наличии заполненных пропусков данных, нестационарности классифицируемых объектов и разнотипности их признаков. Идея предлагаемого подхода, например, при наличии заполненных пропусков данных состоит в выполнении следующих действий:

- осуществить декомпозицию неоднородной обучающей выборки V на множества их строк с заполненными пропусками данных V_1 и без таковых V_2 ;
- по выборке V_1 построить непараметрическую оценку уравнения разделяющей поверхности $\bar{f}_{12}(x)$ и соответствующий алгоритм распознавания образов $\bar{m}_{12}(x)$ в

двухальтернативной задаче классификации;

- с использованием выборки V_2 и алгоритма $\bar{m}_{12}(x)$ организовать вычислительный эксперимент и сформировать данные V'_2 , включающие значения признаков X классифицируемых объектов и корректирующей функции $q(x)$, характеризующей расхождение между «решениями» $\bar{\sigma}(x)$ алгоритма $\bar{m}_{12}(x)$ и «указаниями учителя» $\sigma(x)$. При наличии ошибки $\sigma(x) \neq \bar{\sigma}(x)$ корректирующая функция принимает значение обратное по знаку уравнения разделяющей поверхности $\bar{f}_{12}(x)$ и превышает его на величину Δ ;

- восстановить по выборке V'_2 корректирующую функцию $q(x)$ на основе непараметрической регрессии $\tilde{q}(x)$;

- построить гибридный алгоритм распознавания образов $\bar{\bar{m}}_{12}(x)$ с использованием решающей функции вида $\bar{\bar{f}}_{12}(x) = \bar{f}_{12}(x) + \tilde{q}(x)$.

Если принимаются несколько градаций по степени заполнения пропусков данных в обучающей выборке, то в гибридной системе распознавания образов формируется многоуровневая процедура корректировки уравнения разделяющей поверхности между классами. Разработаны и исследованы модификации гибридных алгоритмов распознавания образов в условиях неоднородных данных, отличающиеся условиями классификации и видом корректирующих решающих функций.

Исследованы аппроксимационные свойства линейных коллективов многомерных непараметрических регрессий в условиях равномерного и неравномерного распределения исходных статистических данных между их составляющими. Установлены свойства их асимптотической несмещённости и сходимости в среднеквадратическом. Полученные результаты позволили получить аналитические зависимости показателей качества аппроксимации коллектива непараметрических регрессий от объёма исходных данных, их размерности и количества составляющих коллектива. Неравномерное распределение исходных данных между составляющими коллектива по сравнению с равномерным распределением способствует снижению значений среднеквадратического отклонения и увеличению дисперсии. По сравнению с традиционной непараметрической регрессией, их коллективы характеризуются меньшей дисперсией. Полученные результаты имеют большое значение при проектировании непараметрических моделей стохастических зависимостей в условиях больших объёмов статистических данных с использованием технологии параллельных вычислений.

Аппроксимационные свойства непараметрической оценки плотности вероятности Розенблатта-Парзена $\bar{p}(x)$ зависят от выбора коэффициентов размытости ядерных

функций. Проведён анализ методов оптимизации $\bar{p}(x)$ путём выбора значений коэффициентов размытости из условий минимума среднеквадратического отклонения непараметрической оценки плотности вероятности (метод 1) и максимума функции правдоподобия (метод 2). Эффективность сравниваемых методов оптимизации непараметрической оценки плотности вероятности при восстановлении равномерного закона распределения случайной величины сопоставима. Они характеризуются близкими значениями среднеквадратического отклонения оценки плотности вероятности. Выбор коэффициента размытости ядерной функции из условия минимума оценки среднеквадратического критерия позволяет получить меньшее значение дисперсии $\bar{p}(x)$. При этом смещение непараметрической оценки плотности вероятности значительно больше, если при её оптимизации использовать критерий максимума функции правдоподобия. В этих условиях более предпочтительно использовать метод 1 оптимизации непараметрической оценки плотности вероятности. При восстановлении нормального закона распределения в условиях больших объёмов n статистических данных значения смещения, дисперсии и среднеквадратического отклонения $\bar{p}(x)$ с использованием сравниваемых методов оптимизации отличаются несущественно. При $n < 200$ и ступенчатой ядерной функции целесообразно использовать метод 1, который обеспечивает значительно меньшее значение дисперсии $\bar{p}(x)$. Если основу синтеза $\bar{p}(x)$ составляет ядерная функция Епанечникова, то в условиях малых выборок более предпочтительным является метод 2. С уменьшением области определения плотности вероятности смещение, дисперсия и среднеквадратическое отклонение $\bar{p}(x)$ увеличиваются, но их соотношения для сравниваемых методов оптимизации и приведённые выше рекомендации сохраняются. Полученные результаты открывают возможность сравнения традиционных методов с рандомизированным подходом оптимизации непараметрической оценки плотности вероятности. Их отличие состоит в признании случайной природы коэффициентов размытости ядерных функций. Тогда задача оптимизации непараметрической оценки плотности вероятности может быть сведена к выбору оптимальных параметров закона распределения коэффициентов размытости ядерных функций.

Программы «Восстановление стохастической зависимости в условиях неопределённости», «Программа для реализации непараметрических алгоритмов распознавания образов парзеновского типа в многомерном пространстве признаков классифицируемых объектов» получили свидетельства о государственной регистрации программ для ЭВМ.

2.3. Непараметрические алгоритмы распознавания образов и моделей стохастических зависимостей в условиях малых выборок

Разработана методика построения линейных и нелинейных непараметрических коллективов решающих функций в задачах распознавания образов и восстановления стохастических зависимостей в условиях малых выборок. Подобные условия встречаются при исследовании медико-биологических, экологических и социально-экономических объектов, что обуславливается нестационарностью объекта исследования, высокой стоимостью и сложностью получения необходимой информации. Предлагаемые системы обеспечивают эффективную обработку информации большой размерности, когда отношение «объем/размерность» обучающих выборок мало.

Идея предлагаемого подхода при синтезе непараметрических систем принятия решений в условиях малых выборок состоит в выполнении следующих действий:

- в соответствии со спецификой решаемой задачи осуществить декомпозицию исходного пространства признаков $x = (x_v, v = \overline{1, k})$ на их наборы $x(j), j = \overline{1, m}$;
- в пространствах набора признаков $x(j), j = \overline{1, m}$ по имеющимся обучающим выборкам построить частные непараметрические решающие функции $\bar{\varphi}_j(x(j)), j = \overline{1, m}$ и осуществить их оптимизацию по коэффициентам размытости ядерных функций в режиме «скользящего экзамена»;
- осуществить интеграцию частных решающих функций в форме линейного либо нелинейного коллектива $\Psi(\bar{\varphi}_j(x(j)), j = \overline{1, m})$.

В частности, линейный коллектив непараметрических регрессий $\bar{y}_j = \bar{\varphi}_j(x(j)), j = \overline{1, m}$ запишется в виде

$$\bar{y} = \sum_{j=1}^m \alpha_j \bar{\varphi}_j(x(t)),$$

где

$$\sum_{j=1}^m \alpha_j = 1, \alpha_j = W_j^{-1} / \sum_{t=1}^T W_t^{-1};$$

W_j – оценка среднеквадратической ошибки аппроксимации искомой зависимости с использованием частной модели $\bar{\varphi}_j(x(j)), j = \overline{1, m}$.

В нелинейном коллективе обобщение частных моделей $\bar{\varphi}_j(x(j)), j = \overline{1, m}$ на основе обучающей выборки $\bar{V} = (\bar{\varphi}_j(x^i(j)), j = \overline{1, m}, y^i, i = \overline{1, n})$ объема n осуществляется с помощью непараметрической статистики

$$\bar{y} = \sum_{i=1}^n y^i \prod_{j=1}^m \beta(\bar{\varphi}_j(x(j)), \bar{\varphi}_j(x^i(j))) / \sum_{i=1}^n \prod_{j=1}^m \beta(\bar{\varphi}_j(x(j)), \bar{\varphi}_j(x^i(j))).$$

В приведённом нелинейном коллективе преобразование

$$\beta(\bar{\varphi}_j(x(j)), \bar{\varphi}_j(x^i(j))) = \Phi\left(\frac{\bar{\varphi}_j(x(j)) - \bar{\varphi}_j(x^i(j))}{c_j}\right),$$

где ядерные функции $\Phi(u)$ удовлетворяют условиям положительности, симметричности и нормированности, а их коэффициенты размытости $c_j \rightarrow 0$ при $n \rightarrow \infty$.

В линейном коллективе весовые коэффициенты α_j , $j = \overline{1, m}$ частных решающих функций определяются путём анализа их показателей эффективности. При синтезе нелинейных коллективов задача обнаружения скрытых закономерностей переносится в пространство значений частных решающих функций.

Предложенный подход развит на решение задачи распознавания образов в условиях обучающих выборок малого объёма. Ближайшими аналогами рассматриваемых систем распознавания образов являются алгебраический подход Ю.И. Журавлева и метод комитетов Вл.Д. Мазурова.

Установлено достоверное преимущество нелинейной непараметрической системы классификации над традиционным алгоритмом распознавания образов, что соблюдается при малых объёмах обучающих выборок. Обнаружен экстремальный характер зависимости оценки вероятности ошибки распознавания образов от количества частных решающих правил. По сравнению с традиционным непараметрическим классификатором парзеновского типа применение предлагаемой непараметрической системы позволяет в 1,5 – 3 раза снизить ошибку распознавания образов в условиях малых выборок. Непараметрический нелинейный коллектив моделей стохастических зависимостей в условиях малых выборок при значениях отношения $n/k < 10$ имеет более высокие аппроксимационные свойства по сравнению с традиционной непараметрической регрессией. С ростом объёма обучающих выборок показатели эффективности сравниваемых моделей приближаются к сопоставимым.

Преимущество разработанных систем принятия решений по сравнению с традиционными непараметрическими моделями и алгоритмами состоит в возможности учёта частичных априорных сведений о виде взаимосвязи между переменными и обходе проблем малых выборок за счёт снижения размерности задачи. Их эффективная реализация обеспечивается возможностью использования технологий параллельных вычислений.

2.4. Быстрые алгоритмы оптимизации непараметрических решающих правил ядерного типа в условиях статистических данных большого объёма

Непараметрические оценки плотности вероятности ядерного типа широко используются при синтезе решающих правил в условиях априорной неопределённости. Вычислительная эффективность традиционных методов их оптимизации значительно снижается с ростом объёма статистических данных. Для решения этой проблемы предложен быстрый алгоритм выбора коэффициентов размытости ядерных функций в непараметрической оценке плотности вероятности многомерной случайной величины, основанный на принципах коллективного оценивания с использованием тестового семейства плотностей вероятностей.

Реализация предложенной методики основана на результатах исследования асимптотических свойств непараметрической оценки плотности вероятности $\bar{p}(x_1, K, x_k)$ в условиях представления коэффициентов размытости ядерных функций в виде $c_v = c \sigma_v$, где σ_v – среднеквадратическое отклонение случайной величины x_v , $v = \overline{1, k}$. Из условия минимума среднеквадратического отклонения $\bar{p}(x_1, K, x_k)$ от восстанавливаемой плотности вероятности $p(x_1, K, x_k)$ определён оптимальный параметр:

$$c^* = \left(\frac{k \left(\|\Phi(u)\|^2 \right)^k}{n B \prod_{v=1}^k \sigma_v} \right)^{\frac{1}{k+4}},$$

где $B = \int_{-\infty}^{\infty} \int_K \left(\sum_{v=1}^k \sigma_v^2 p_v^{(2)}(x_1, K, x_k) \right)^2 dx_1 K dx_k$; $\|\Phi(u_v)\|^2 = \int_{-\infty}^{\infty} \Phi^2(u_v) du_v$; $p_v^{(2)}(x_1, K, x_k)$ – вторая производная $p(x_1, K, x_k)$ по компонентам x_v , $v = \overline{1, k}$.

Существует семейство многомерных плотностей вероятности независимых случайных величин, для которых значение $B \prod_{v=1}^k \sigma_v$ определяется только видом плотности вероятности и не зависят от её параметров.

На этой основе разработана методика быстрого выбора коэффициентов размытости ядерных функций для непараметрической оценки многомерной плотности вероятности. Идея предлагаемого подхода состоит в вычислении оптимальных значений параметров коэффициентов размытости ядерных функций для непараметрических оценок, составляющих тестовое семейство многомерных плотностей вероятности. Они обобщаются при определении оценки параметра c^* в процедуре коллективного типа. Для

найденного значения $\bar{c}^*$ можно определить оценку среднеквадратического отклонения статистики $\bar{p}(x_1, K, x_k)$ от искомой плотности вероятности $p(x_1, K, x_k)$.

Полученные результаты создают основу широкого применения непараметрических оценок плотностей вероятности ядерного типа в прикладных задачах, характеризующихся наличием статистических данных большого объема.

Традиционный подход оптимизации непараметрических алгоритмов распознавания образов предполагает выбор коэффициентов размытости ядерных функций из условия минимума оценки вероятности ошибки классификации. Его вычислительная эффективность значительно снижается с увеличением объема обучающей выборки. На основе анализа асимптотических свойств непараметрической оценки уравнения разделяющей поверхности в двухальтернативной задаче распознавания образов установлена возможность выбора коэффициентов размытости ядерных функций по результатам оптимизации непараметрических оценок плотностей вероятностей случайных величин в классах. Этот вывод позволяет изложенную выше методику быстрого выбора коэффициентов размытости для ядерных оценок плотностей вероятностей развить на решение задачи оптимизации непараметрической оценки уравнения разделяющей поверхности между классами и на несколько порядков сократить временные затраты её решения.

Для непараметрического решающего правила, соответствующего критерию максимального правдоподобия, получены асимптотические выражения среднеквадратических отклонений $\bar{W}(c)$, $\bar{W}(c_1, c_2)$ ядерной оценки уравнения разделяющей поверхности $\bar{f}_{12}(x_1, K, x_k)$ от байесовской решающей функции $f_{12}(x_1, K, x_k)$ при условии $c_v = c \sigma_v$ и $c_{1v} = c_1 \sigma_{1v}$, $c_{2v} = c_2 \sigma_{2v}$, $v = \overline{1, k}$. Здесь C параметр коэффициента размытости ядерных функций в статистике $\bar{f}_{12}(x_1, K, x_k)$ соответствует традиционному подходу оптимизации непараметрического алгоритма распознавания образов, а параметры c_1 и c_2 – быстрым процедурам выбора коэффициентов размытости из условия минимума среднеквадратических отклонений ядерных оценок $\bar{p}_1(x_1, K, x_k)$, $\bar{p}_2(x_1, K, x_k)$ от плотностей вероятностей распределения случайных величин x_v , $v = \overline{1, k}$ в классах. Показано, что отношение $\bar{W}(c^*)/\bar{W}(c_1^*, c_2^*)$ при оптимальных значениях $c = c^*$ и $c_1 = c_1^*$, $c_2 = c_2^*$ является постоянным и не зависит от объема обучающей выборки. Данный результат сопровождается значительным снижением дисперсии $\bar{f}_{12}(x_1, K, x_k)$ при

использовании коэффициентов размытости (c_1^* , c_2^*) и относительно меньшим увеличением её смещения. Если дисперсии законов распределения случайных величин в классах отличаются, то преимущество предлагаемого подхода увеличивается.

Развитие методов дискретизации области значений случайных величин является перспективным направлением обработки статистических данных большого объёма. Они используются при построении доверительных границ плотности вероятности, проверки гипотез о распределениях случайных величин, решении задач автоматической классификации и распознавания образов. Известен ряд методов дискретизации интервала значений одномерной случайной величины, которые представлены в работах Sturges H.A., Wang M.P., Shimazaki H., Scott D.W. и др.

Предложена новая регрессионная оценка плотности вероятности

$$\bar{p}(x_1, K, x_k) = \frac{1}{\prod_{v=1}^k c_v} \sum_{j=1}^N \bar{P}^j \prod_{v=1}^k \Phi\left(\frac{x_v - z_v^j}{c_v}\right),$$

восстанавливаемая по данным $V' = (z_1^j, K, z_k^j, \bar{P}^j, j = \overline{1, N})$. Массив данных V' формируется по исходной выборке $V = (x_v^i, v = \overline{1, k}, i = \overline{1, n})$ путём дискретизации области значений случайной величины $x = (x_1, K, x_k)$ на многомерные интервалы. Составляющие массива «сжатой информации» V' определяются координатами центров элементов вводимой вычислительной сетки и частотой $\bar{P}^j$ попадания значений случайных величин из выборки V в эти элементы. В статистике $\bar{p}(x_1, K, x_k)$ ядерные функции $\Phi(uv)$ удовлетворяют условиям положительности, симметричности и нормированности; c_v , $v = \overline{1, k}$ – коэффициенты размытости ядерных функций $\Phi(uv)$, которые удовлетворяют условиям.

Задача дискретизации области значений случайных величин является одной из актуальных в математической статистике. Результаты её решения используются при оценивании доверительных границ плотности вероятности и проверки гипотез о распределениях случайных величин с помощью критерия Пирсона. Известен ряд обоснованных методов дискретизации интервала значений одномерной случайной величины, которые представлены в работах Sturges H.A., Wang M.P., Shimazaki H., Scott D.W. и др.

Впервые выбор оптимального количества интервала дискретизации двумерной случайной величины из условия минимума среднеквадратического отклонения регрессионной оценки плотности вероятности ядерного типа обоснован в работе Лапко

А.В., Лапко В.А. Определено оптимальное количество двумерных интервалов дискретизации

$$N^* = \sqrt{\frac{3}{4} \Delta_1 \Delta_2 \|p(x_1, x_2)\|^2 n},$$

где Δ_1, Δ_2 – интервалы значений случайных величин x_1, x_2 ; n – объём статистической выборки $(x_1^i, x_2^i, i = \overline{1, n})$; $p(x_1, x_2)$ – плотность вероятности случайных величин; значение $\|p(x_1, x_2)\|^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p^2(x_1, x_2) dx_1 dx_2$.

Оптимальное количество интервалов дискретизации области значений двумерной случайной величины зависит от функционала $b = \|p(x_1, x_2)\|^2$. С использованием статистики Розенблатта-Парзена $\bar{p}(x)$ предлагаются две её оценки, установлены условия их компетентности. Представление функционала b в виде оценки $\bar{b}_2(c)$ математического ожидания $\bar{p}(x_1, x_2)$ обладает преимуществом по сравнению с $b_1(c) = \|\bar{p}(x_1, x_2)\|^2$. Данное утверждение подтверждается выводами асимптотических исследований и результатами вычислительных экспериментов восстановления плотности вероятности зависимых двумерных случайных величин с нормальными законами распределения. Аппроксимационные свойства статистик $b_1(c)$ и $\bar{b}_2(c)$ снижаются с ростом коэффициента корреляции между случайными величинами. Количество интервалов дискретизации N^* области значений двумерной случайной величины и её оценок, определяемых с использованием статистик $b_1(c)$, $\bar{b}_2(c)$, совпадают. Этот вывод обосновывается тем, что преимущество аппроксимационных свойств $\bar{b}_2(c)$ над $b_1(c)$ компенсируется в процессе определения целых значений количества интервалов дискретизации.

Значительные трудности аналитических преобразований при получении формул оптимальной дискретизации определяют их дальнейшее развитие на многомерный случай с использованием метода аналогий.

На этой основе получена формула определения количества интервалов для многомерного случая

$$N_k^* = \left(\alpha(k) n \prod_{v=1}^k \Delta_v \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p^2(x_1, K, x_k) dx_1 K dx_k \right)^{1/2} = \alpha_k \sqrt{n},$$

где $\alpha(k) = (2k-1)/k^2$; Δ_v – длина интервала значений случайной величины x_v , $v = \overline{1, k}$.

Количество интервалов дискретизации в многомерном случае пропорционально коэффициенту α_k и квадратному корню от объёма n исходных статистических данных. Значения коэффициента α_k определяются размерностью k случайной величины, видом

плотности вероятности и не зависят от её параметров. Увеличение объёма статистических данных сопровождается ростом количества интервалов дискретизации области определения плотности вероятности и пропорционально значению $n^{1/2}$. Отношение количества многомерных интервалов дискретизации области определения различных плотностей вероятности не зависит от объёма статистических данных, а определяется только видом плотности вероятности. Использование правила Скотта для многомерных независимых случайных величин с нормальным законом распределения приводит к завышенному количеству интервалов дискретизации, которые не согласуются с предлагаемыми рекомендациями и эвристическими методами дискретизации.

Развитие предлагаемой методики дискретизации многомерной случайной величины предполагает оценивание функционала $\prod_{v=1}^k \Delta_v \int_{-\infty}^{+\infty} K \int_{-\infty}^{+\infty} p^2(x_1, K, x_k) dx_1 K dx_k$. Его значения определяются только видом плотности вероятности и не зависят от её параметров.

На основе непараметрического алгоритма автоматической классификации и предлагаемого метода дискретизации области значений многомерной случайной величины разработан программный комплекс НАС v.1.0 обработки многомерных статистических данных большого объёма. Обнаруженные классы характеризуются одномодальными фрагментами плотности вероятности в пространстве спектральных признаков анализируемых объектов. Количество классов априори не задаётся. Основу алгоритма автоматической классификации составляют процедуры «сжатия» исходной информации и методика целенаправленного анализа вероятностных характеристик интервалов дискретизации области значений многомерных случайных величин. Программный комплекс реализован в среде Visual Studio Community 2017, получил государственную регистрацию и имеет дополнительные функциональные возможности: статистический анализ количественных характеристик законов распределения случайных величин в обнаруженных классах, отображение результатов автоматической классификации в пространстве признаков анализируемых объектов и пространственных координат.

Программный комплекс применен для обработки спектральных данных дистанционного зондирования территории горной лесотундры западной части Алтае-Саянского региона, полученных с аппарата Worldview-2 (рис. 2.1). Каждый элемент земной поверхности характеризовался четырьмя спектральными каналами: синий, зелёный, красный, ближний инфракрасный.

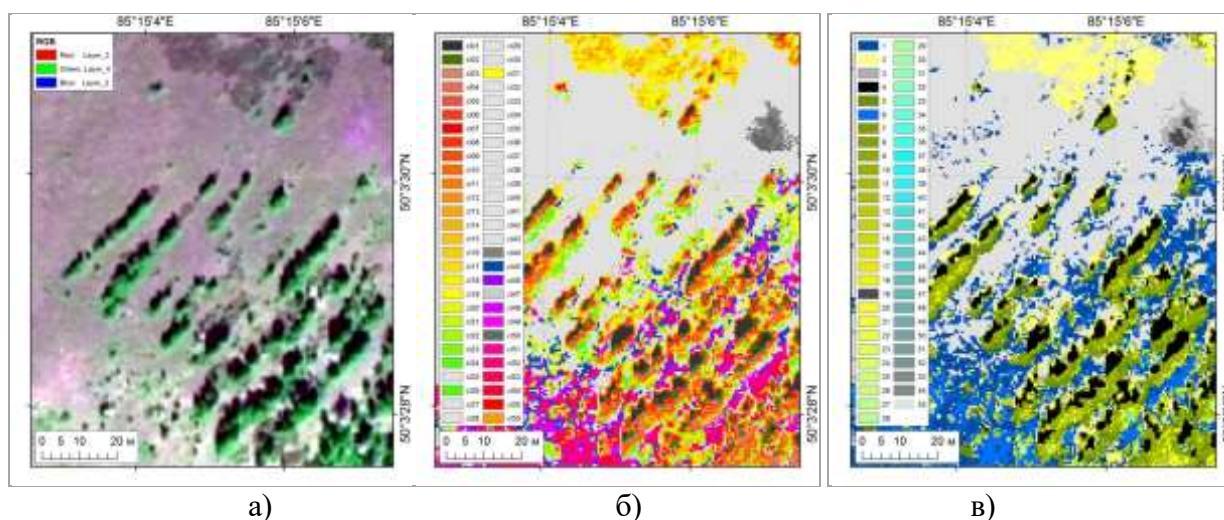


Рис. 2.1 (а) – Фрагмент спутниковой съемки Worldview-2. Отображение результатов автоматической классификации, полученных Erdas Imagine (б) и разработанного непараметрического алгоритма (в) при количестве классов 55

Визуально пространственное распределение результатов классификации предложенного алгоритма классификации и используемого программного продукта Erdas Imagine сопоставимы. Выделяются полосы деревьев и тени от них, участки тундры, кустарников и каменистые поверхности. В общем случае фиксация количества классов в программном продукте Erdas Imagine сокращает возможности проверки гипотез о неоднородности спектральных данных исследуемой территории по сравнению с результатами непараметрического алгоритма автоматической классификации. Причём интерпретация полученных результатов с использованием визуального дешифрирования является ограниченной. Обнаруженные с помощью предлагаемого метода дополнительные закономерности неоднородности спектральных данных создают основу планирования полевых обследований для проверки и обоснования полученных результатов классификации.

Разработанный программный комплекс NAC v. 1.0 получил государственную регистрацию (Свидетельство о государственной регистрации программы для ЭВМ № 2019660994 от 16.08.2019) и используется при обработке больших массивов данных дистанционного зондирования природных объектов.

2.5. Непараметрические системы проверки гипотез о распределениях многомерных случайных величин и их независимости

Задачи проверки гипотез о распределениях случайных величин являются актуальными при оценивании состояний объектов различной природы и выборе их информативных признаков. При проверке гипотез о распределения случайных величин

используется критерий согласия Пирсона, который требует использования трудно формализуемого этапа разбиения области значений случайных величин на многомерные интервалы. Обоснована возможность использования непараметрических алгоритмов распознавания образов, соответствующих критерию максимального правдоподобия, в задачах проверки статистических гипотез о распределениях случайных величин.

Пусть X_1 и X_2 – две генеральные совокупности с произвольными законами распределения. Необходимо по независимым выборкам $V_1 = (x^i, i = \overline{1, n_1})$ и $V_2 = (x^i, i = \overline{1, n_2})$ многомерных данных $x = (x_v, v = \overline{1, k})$, извлечённым из данных генеральных совокупностей, проверить гипотезу

$$H_0: F_1(x) \equiv F_2(x)$$

о тождественности их функций распределения $F_1(x)$, $F_2(x)$ в условиях больших объёмов n_1, n_2 статистических данных.

Если при решении двухальтернативной задачи распознавания образов в соответствии с критерием максимального правдоподобия вероятность ошибки классификации равна значению 0,5, то законы распределения случайных величин в области определения классов очевидно совпадают. Поэтому появляется возможность перехода от задачи сравнения законов распределения многомерных случайных величин к проверке гипотезы $\bar{H}_0$ о равенстве вероятности ошибки распознавания образов значению 0,5. Предлагаемая методика проверки гипотезы о тождественности законов распределения многомерных случайных величин в условиях статистических данных большого объёма состоит в выполнении следующих действий:

1. Осуществить «сжатие» исходной статистической информации V_1 и V_2 в массивы данных $\bar{V}_t = (z_1^j, \Lambda, z_k^j, \bar{P}^j, j = \overline{1, N_t})$, $t = 1, 2$ путём декомпозиции области значений случайных величин $x = (x_1, K, x_k)$ на многомерные интервалы. Здесь N_t – количество многомерных интервалов в области определения классов, $z_k^j = (z_1^j, \Lambda, z_k^j)$ – центры интервалов дискретизации, а $\bar{P}^j$ – частоты принадлежности случайных величин из массивов данных V_t , $t = 1, 2$ этим интервалам. Методика дискретизации области значений случайных величин представлена в тексте и публикациях отчёта.

2. По данным $\bar{V}_t$, $t = 1, 2$ сформировать обучающую выборку для синтеза непараметрического алгоритма распознавания образов, соответствующего критерию

максимального правдоподобия с использованием регрессионных оценок плотностей вероятностей случайных величин в классах. Регрессионная оценка плотности вероятности приведена в тексте отчёта.

3. Определить оценку вероятности ошибки распознавания образов $\bar{\rho}$ в режиме «скользящего экзамена».

4. Проверить гипотезу $H'_0: \bar{\rho}=0,5$; используя методику доверительного оценивания вероятности ошибки распознавания образов либо критерий Колмогорова с заданным риском отвергнуть поставленную гипотезу. Если гипотеза H'_0 подтверждается, то законы распределения $p_1(x)$ и $p_2(x)$ тождественны.

Непараметрические алгоритмы распознавания образов были использованы при проверке гипотезы о независимости двухмерных случайных величин.

Пусть имеется выборка $V = (x^i, i = \overline{1, n})$ объёма n , составленная из независимых наблюдений двухмерной случайной величины $x = (x_1, x_2)$. Выборка V извлекается из генеральных совокупностей, характеризующихся плотностями вероятности $p(x_1)p(x_2)$ или $p(x_1, x_2)$. Необходимо по статистическим данным V проверить гипотезу $\bar{H}_0: p(x_1, x_2) = p(x_1)p(x_2)$ о независимости случайных величин x_1, x_2 .

Для проверки гипотезы $\bar{H}_0$ будем решать двухальтернативную задачу распознавания образов. Под классами Ω_1, Ω_2 понимаются области определения плотностей вероятностей $p(x_1)p(x_2), p(x_1, x_2)$. В этих условиях байесовское решающее правило, соответствующее критерию максимального правдоподобия имеет вид,

$$m(x): \begin{cases} x \in \Omega_1, \text{ если } p(x_1, x_2) < p(x_1)p(x_2) \\ x \in \Omega_2, \text{ если } p(x_1, x_2) > p(x_1)p(x_2). \end{cases}$$

В отличие от традиционной постановки задачи распознавания образов при синтезе решающего правила $m(x)$ априори отсутствует обучающая выборка, содержащая сведения о принадлежности элементов выборки V к тому или иному классу. Эти сведения должны обнаруживаться в процессе реализации методики проверки гипотезы $\bar{H}_0$, которая основана на выполнении следующих действий. По выборке V восстановить плотности вероятности $p(x_1, x_2), p(x_1)p(x_2)$, используя их непараметрические оценки типа Розенблатта – Парзена.

Осуществить оптимизацию непараметрических оценок плотностей вероятностей

$\bar{p}(x_1, x_2)$ и $\bar{p}(x_1)\bar{p}(x_2)$ по коэффициентам размытости ядерных функций, например из условия максимума функции правдоподобия. Определить оценки вероятностей ошибки распознавания образов, предложенным непараметрическим алгоритмом ситуаций из обучающей выборки для классов Ω_1, Ω_2 , соответствующих независимым и зависимым признакам x_1, x_2 . По минимальному значению оценок вероятностей ошибок распознавания образов принимается решение о зависимости либо независимости случайных величин. Эффективность предложенной методики подтверждается результатами вычислительных экспериментов.

Выводы. Непараметрические алгоритмы распознавания образов и модели многомерных стохастических зависимостей в условиях больших выборок неоднородных данных, основанные на дискретизации области их определения, принципах коллективного оценивания и гибридного моделирования, обладают свойствами асимптотической сходимости. Существуют аналитические зависимости их свойств от особенностей исходных данных и методов обработки. Полученные выводы допускают обобщение при синтезе линейных и нелинейных коллективов непараметрических регрессий. На этой основе разработана методика синтеза эффективной структуры непараметрических и гибридных систем распознавания образов в условиях выборок неоднородных данных большого объёма.

Непараметрические алгоритмы распознавания образов и модели стохастических зависимостей, обеспечивают эффективную обработку информации в условиях малых выборок. Методы оптимальной дискретизации области значений двумерной случайной величины, основанные на анализе асимптотических свойств регрессионной оценки плотности вероятности, что является одной из актуальных задач математической статистики. Их эффективная реализация использует процедуру оценивания нелинейного функционала от плотности в формуле расчёта оптимального количества интервалов дискретизации. Полученные результаты допускают обобщение на многомерный случай на основе использования метода аналогий.

Быстрые алгоритмы выбора оптимальных коэффициентов размытости ядерных функций в непараметрической оценке многомерной плотности вероятности типа Розенблатта-Парзена позволяют на несколько порядков сократить время их синтеза по сравнению с традиционным подходом, что является актуальным при решении задач принятия решений в условиях статистических данных большого объёма. Полученные результаты имеют важное значение при доверительном оценивании плотности вероятности, проверке гипотез о распределениях случайных величин с использованием

критерия Пирсона и решении задач автоматической классификации.

Проверка гипотезы о тождественности двух многомерных законов распределения может быть заменена на проверку гипотезы о равенстве ошибки распознавания образов значению 0,5. Для этого необходимо сформировать обучающую выборку для решения задачи классификации по исходным статистическим данным. Для сокращения временных затрат на проверку гипотезы о тождественности многомерных случайных величин в условиях статистических данных большого объёма целесообразно использовать регрессионные оценки плотности вероятности распределения случайных величин в классах. Их синтез основан на декомпозиции области значений случайных величин, что позволяет сократить объём исходных статистических данных при незначительном снижении эффективности принимаемых решений.

Методика проверки гипотезы о независимости случайных величин эффективно реализуется на основе использования непараметрического алгоритма распознавания образов, соответствующего критерию максимального правдоподобия. При этом законы распределения случайных величин в классах формируются на основе исходных статистических данных в предположении зависимости и независимости анализируемых переменных. Гипотеза подтверждается либо отвергается на основе результатов анализа оценок вероятностей ошибок распознавания образов в классах.

Основные результаты исследований реализованы в программных комплексах NSARSD v. 1.0, NAC v. 1.0, имеют государственную регистрацию и используются при обработке данных дистанционного зондирования природных объектов.

3 МЕТОДЫ ПОИСКА, КЛАССИФИКАЦИИ И АНАЛИЗА СТРУКТУР И СВЯЗЕЙ МЕЖДУ НИМИ В НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЯХ

Ответственный исполнитель раздела д.ф.-м.н. Садовский М.Г.

3.1 Новые структурные элементы в символьных последовательностях, их комбинаторные и статистические свойства, связь с семантикой генома

Аннотация. Исследованы структуры полных геномов хлоропластов и некодирующих областей этих геномов. Был проанализирован 391 геном хлоропластов наземных растений, размещенных в EMBL-банке. Под структурой понимается расположение в пространстве частот триплетов точек, соответствующих фрагментам генома.

Изучены 15 бактериальных геномов для анализа распределения начал считывания по генетической последовательности. Ключевой вопрос заключался в том, распределяются ли начала равномерно и однородно вдоль генетической последовательности, или существуют некоторые области их повышенной локальной плотности.

Изучена структурированность суммарного транскриптома сибирской лиственницы. Для этого контиги из общего транскриптома были помечены ридами, содержащими тканеспецифичные транскрипты. Распределение контигов от общего транскриптома было разработано с учетом взаимной энтропии частот встречаемости ридов из тканеспецифичных транскриптов. Изучен митохондриальный геном лиственницы сибирской (*Larix sibirica* Ledeb.) - одной из краеугольных пород бореальных лесов Сибири.

Комплексное профилирование экспрессии генов членов семейства генов гомеобокса позволяет определить их роль в росте и развитии *Pinus sylvestris*. Гены гомеобокса кодируют транскрипционные факторы, играющие важную роль в развитии организма. Гомеодомены распространены у огромного количества видов. Следовательно, их можно идентифицировать даже у немодельных организмов. Понимание функций генов гомеобокса поддерживает исследования развития тканей и дает способы его регулирования.

Выводы. Для геномов хлоропластов наземных растений обнаружена характерная восьмикластерная структура. Эта структура образована центральным кластером состоящим из точек, относящихся к некодирующим участкам, вокруг него тремя лучами собраны шесть кластеров, относящихся к кодирующим участкам и отдельно отстоит

восьмой кластер, состоящий как из кодирующих, так и из не кодирующих участков. Анализ внутренней структурированности не кодирующих областей геномов хлоропластов наземных растений позволил обнаружить пять типов характерных структур. На основании одновременной кластеризации наборов не кодирующих областей хлоропластов разных видов можно предположить, что исходным типом структуры была «линза с двумя хвостами», из которой в процессе эволюции редуцировались остальные типы структур.

Было установлено, что некоторые геномы демонстрируют чрезвычайно далекий от однородности характер распределения стартов считывания, в то время как другие показывают более низкий уровень неоднородности. Уровень неоднородности был определен через расстояние Кульбака-Лейблера между реальным распределением строк заданной длины и наиболее вероятным продолжением из более коротких строк.

Рассмотрены транскриптомы лиственницы сибирской (*Larix sibirica* Ledeb.) и сосны сибирской (*Pinus sibirica* Du Tour). Транскриптомы демонстрируют необычную симметрию в распределении частотных словарей триплетов. Была обнаружена октаэдрическая структура распределения контигов транскриптомов, причем для *L. sibirica* она имела зеркальную симметрию, а для *P. sibirica* – вращательную. При рассмотрении митогенома лиственницы сибирской было обнаружено, что его размер намного больше, чем у других изученных к настоящему времени голосеменных. Подобно другим видам, митогеном лиственницы сибирской содержит относительно мало генов, и, несмотря на его огромный размер, повторяющиеся области низкой сложности покрывают только 14,46% последовательности митогенома.

Гены гомеобокса для сосны обыкновенной недостаточно изучены. Собран *de novo* транскриптом *Pinus sylvestris*, полученный из пяти тканей. Транскриптом включает 775 502 транскрипта. Было обнаружено 243 гомеобокс-содержащих транскрипта, и с использованием этих последовательностей был проведен DE-анализ. Мы получили 5 кластеров генов гомеобокса DE (визуализированные как тепловая карта экспрессии генов). Аннотированы гены DE. Полученные результаты дают представление о развитии почек и зрелых тканей *Pinus sylvestris*.

Описание результатов. Рассматривались генетические последовательности длины L , состоящие из символов алфавита $\mathfrak{M} = \{A, C, G, T\}$. Для этой последовательности символов составлялся частотный словарь толщины 3. Под частотным словарем W_3 толщины 3 символьной последовательности, соответствующей ДНК, понимается список всех триплетов $v_1 v_2 v_3$ идущих подряд символов с указанием частот

этих триплетов. Таких комбинаций может быть 64. В качестве частоты f_ω рассматривается отношение количества копий n_ω выбранного триплета к общему количеству всех триплетов N , где N — сумма всех n_ω : $f_\omega = \frac{n_\omega}{N}$. Любой частотный словарь W_3 ставит в соответствие геному множество точек в 64-мерном метрическом пространстве.

Для обнаружения структуры в генетической последовательности проводилась предварительная обработка, которая ставила в соответствие данной последовательности множество точек в 64-мерном пространстве триплетов. Делалось это следующим образом: последовательность сканировалась окном длины Δ с шагом t . Для каждого положения i рамки определялся участок генетической последовательности, совпадающий с рамкой считывания, для которого вычислялся частотный словарь $W_3^{(i)}$, соотносящийся с i -ой точкой 64-мерного пространства. Кроме того, с каждой точкой 64-мерного пространства связывались следующие параметры: номер центрального символа рассматриваемого участка и относительная фаза.

Номер центрального символа участка совпадает с номером этого символа в последовательности. Относительная фаза определяется тем, попал рассматриваемый участок в кодирующую или не кодирующую область последовательности. Участок относится к кодирующим, если он целиком попадал в кодирующую область последовательности. Для не кодирующего участка соответствующая ему точка помечается символом J . Если участок относится к кодирующим, для него возможны 6 вариантов маркировки: $B_0, B_1, B_2, F_0, F_1, F_2$. Для кодирующего участка, аннотированного в последовательности как считывающийся в прямом направлении, вычисляется остаток от деления на 3 разности номеров центрального символа участка и первого символа кодирующей области, к которой он относится. В соответствии с величиной остатка от деления точка помечалась символом B_0, B_1 или B_2 . Для участка, аннотированного как считывающийся в обратном направлении, вычисляется остаток от деления на 3 разности номеров последнего символа кодирующей области, к которой относится участок, и центрального символа участка. В зависимости от значения остатка от деления точка помечалась символами F_0, F_1 или F_2 . Для всех генетических последовательностей длина рамки считывания $\Delta = 603$, шаг $t = 11$.

Для того, чтобы визуализировать множество точек 64-мерного пространства, с

помощью программы *VidaExpert* строилась проекция из 64-мерного пространства частот троек в пространство первых трёх главных компонент, построенных по этому множеству точек. Чтобы получить двумерные рисунки трехмерного пространства рассматривались проекции на плоскость пространства первых двух главных компонент и второй и третьей главной компоненты.

Было обнаружено, что все геномы хлоропластов наземных растений в плоскости первой и второй главных компонент имеют выраженную трех лучевую структуру. Для подавляющего большинства геномов типичен вид в плоскости первых двух главных компонент и второй и третьей главной компоненты, который показан на рис. 3.1.

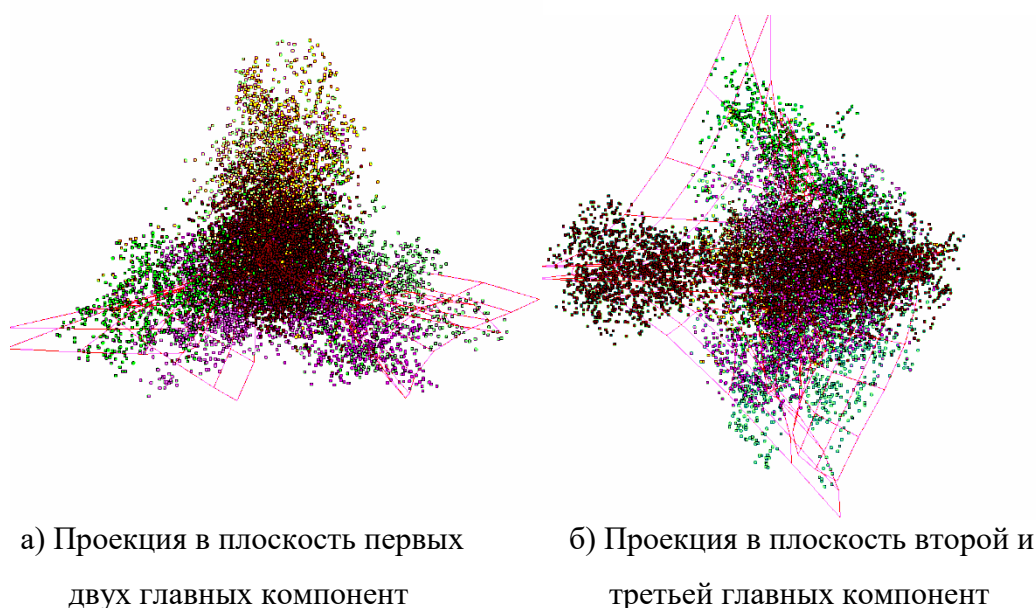


Рис. 3.1 – Типичный вид структуры данных геномов наземных растений в проекциях пространства первых трех главных компонент (структура генома *Nimphaea alba*)

Как видно на рис. 3.1а, структура генома хорошо кластеризована относительно точек из кодирующих и не кодирующих участков генома. Центральную часть рисунка занимают точки из некодирующих участков, они окрашены в коричневый цвет. Один луч состоит из точек, относящихся к прямой и обратной относительной фазе 2 (точки, отмеченные светло желтым и темно желтым цветом), второй луч содержит точки, относящиеся к прямой относительной фазе 0 и обратной относительной фазе 1 (точки темно сиреневого и светло зеленого цвета), и третий состоит из точек, относящихся к обратной относительной фазе 0 и прямой относительной фазе 1 (точки светло сиреневого и темно зеленого цвета). На рис. 3.1б видно, что точки, занимающие центральную часть распределения в проекции рис. 1а, разделены на два кластера. Первый находится в центре трёхлучевой структуры, а второй обособлен от остальных точек.

Исследована внутренняя структурированность некодирующих областей геномов хлоропластов. Кластеризация проводилась методом упругих карт для проекции 64-мерного пространства в пространство первых трех главных компонент (рис. 3.2).

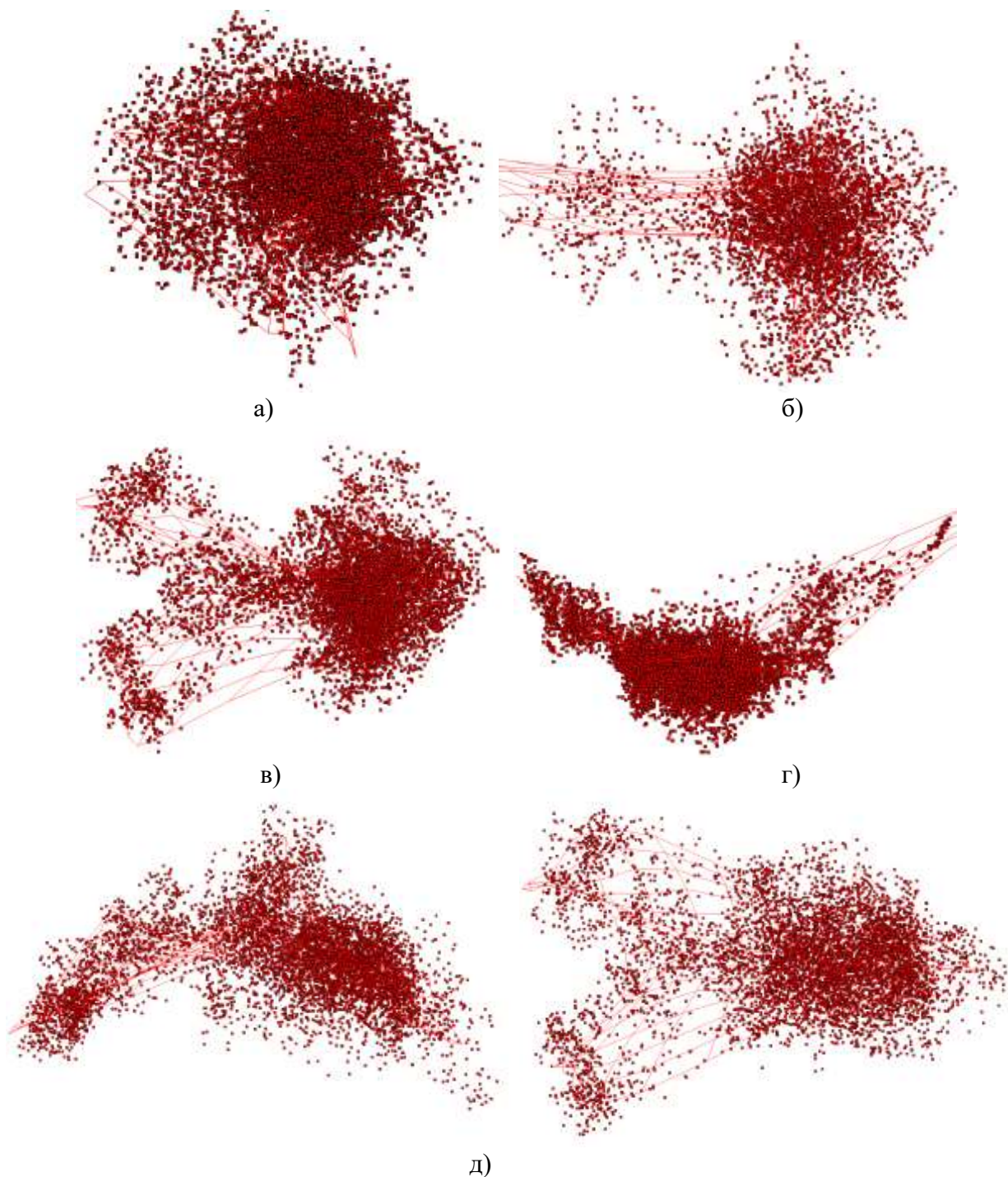


Рис. 3.2 – Структуры некодирующих областей геномов хлоропластов: а) структура типа «Шар» на примере генома *Erodium chrysanthum*, б) «Шар с одним хвостом» на примере генома *Pseudotsuga sinensis*, в) «Шар с двумя хвостами» на примере генома *Liriodendron tulipifera*, г) «Линза с одним хвостом» на примере генома *Ricinus communis*, д) «Линза с двумя хвостами» на примере генома *Lupinus luteus*, в двух проекциях

Кроме того, были рассмотрены структуры, которые можно выделить в ансамблях формально выделяемых фрагментов, принадлежащих разным видам, в случае одновременной кластеризации таких геномов. Для того, чтобы показать, что целостность структуры проявляется не только в каждом отдельно взятом геноме, но и в их множестве, были построены общие структуры для наборов из 10 геномов. Множество точек, принадлежащее каждому из геномов, окрашивалась в отдельный цвет (см. рис. 3). В большинстве случаев объединение некодирующих участков 10 произвольно взятых геномов хлоропластов, разделённых на формально выделяемые фрагменты, как описано выше, порождали структуру практически идентичную между собой; типичный пример такой структурированности показан на рис. 3.3.

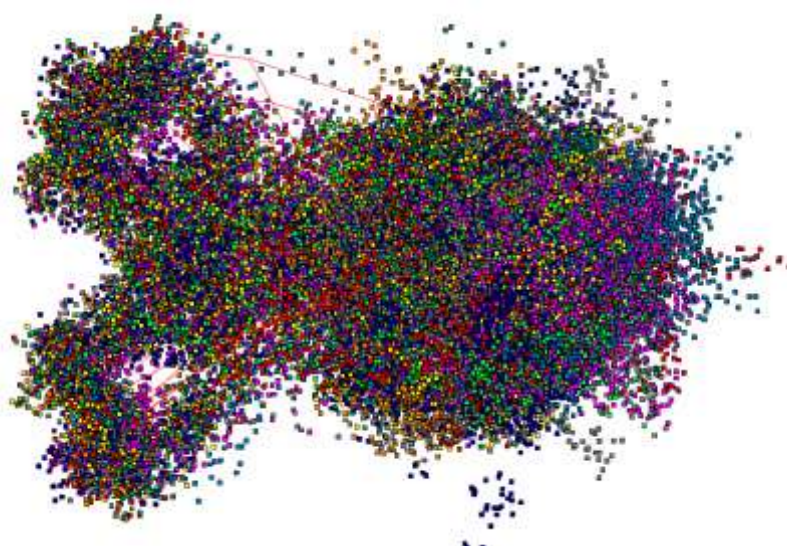


Рис. 3.3 – Объединённая структура геномов *Genlisea margaretae*, *Schwalbea americana*, *Pinguicula ehlersiae*, *Orobanchе gracilis*, *Ptilidium pulcherrimum*, *Magnolia kwangsiensis*, *Jacobaea vulgaris*, *Mankyua chejuensis*, *Sesamum indicum* и *Vaccinium macrocarpon*

Транскриптом представляет собой последовательности экспрессируемых генов и соответствует молекуле мРНК, выделенной из биологической клетки или ткани. Систематическое сравнение (довольно коротких) фрагментов постоянной длины формально идентифицированных в геноме обнаруживает симметрию в распределении частотных словарей триплетов, полученных по этим фрагментам. Общая схема распределения выглядит как суперпозиция двух треугольников, где вершинам соответствуют фрагменты, относящиеся к одной и той же относительной фазе. Проще говоря, это соответствует сдвигу рамки считывания в случае последовательной обработки генетического текста. Сам транскриптом может рассматриваться как набор этих фрагментов, с небольшими исключениями. Во-первых, длина транскриптов различна

и может влиять на ожидаемую картину. Во-вторых, в транскриптоме нет фрагментов, соответствующих некодирующим областям генома. Этот факт приводит к наиболее вероятной конфигурации кластеров, соответствующих транскриптам с одиноковым индексом относительной фазы, то есть октаэдру. Все эти структуры можно увидеть в пространстве первых трех главных компонент. Транскриптом *L. sibirica* дает почти идеальную октаэдрическую структуру, в то время как транскриптом *P. sibirica* отличается довольно значительно, у него структура представляет собой два параллельных треугольника с плоскостями, включающими кластеры из одного и того же стренда. Кроме того, было обнаружено, что октаэдрическая структура распределения контигов транскриптомов для *L. sibirica* имеет зеркальную симметрию, а для *P. sibirica* – вращательную (рис. 3.4). Такая деформация структуры может возникнуть в силу биологических особенностей исследуемого материала: мы изучали транскриптом *P. Sibirica* полученный не из обычной древесины, а из так называемой «ведьминой метлы». «Ведьминая метла» – это фрагменты кроны растения с аномальным морфогенезом, для которого неизбежны серьезные генетические изменения в его геноме.

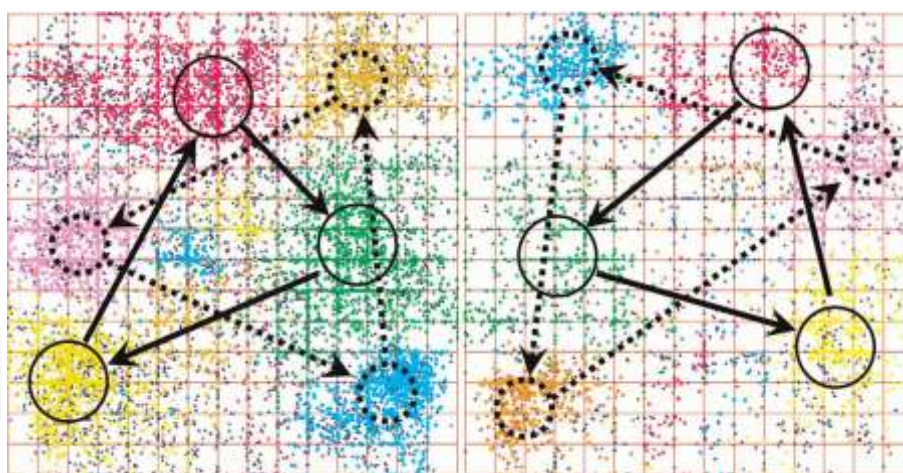


Рис. 3.4 – Зеркальная (слева) симметрия в транскриптоме *L. sibirica* против вращательной симметрии (справа) в транскриптоме *P. sibirica*. Непрерывные стрелки и окружности соответствуют прямым относительным фазам, а пунктирные – обратным

Для поиска последовательностей митохондриальной ДНК (мтДНК) использовались два источника последовательностей ДНК: образцы, обогащенные мтДНК, и считывания нуклеотидов, полученные в проекте по секвенированию полного генома *de novo*, соответственно. Сборка митогенома лиственницы сибирской содержала девять контигов, самый короткий и самый большой контиг составляли 24 767 п.н. и 4 008 762 п.н. соответственно. Общий размер генома оценивается в 11,7 Мбит. Всего в этом митогеноме аннотировано 40 кодирующих белок, 34 тРНК и 3 гена рРНК и

многочисленные повторяющиеся элементы (RE). Всего было найдено 864 сайта редактирования РНК C-to-U для 38 из 40 генов, кодирующих белок. Огромный размер этого генома, о котором в настоящее время сообщается больше всего, можно частично объяснить переменным количеством мобильных генетических элементов и интронов, но маловероятно последовательностями, связанными с плазмидами. Мы обнаружили несколько плазмидоподобных вставок, представляющих лишь 0,11% всего митогенома лиственницы сибирской. Удалось произвести высококачественную сборку митогенома сибирской лиственницы с использованием коротких и длинных считываний нуклеотидов, генерируемых Illumina HiSeq 2000 и MinION. Финальная сборка митогенома лиственницы сибирской состоит из 9 контигов общей длиной 11 662 539 п.н. (N50 = 3 031 766 п.н.). Самый длинный контиг – 4 008 762 б.п., самый короткий – 24 767 б.п. Наконец, были аннотированы 40 генов, кодирующих белок, 34 гена тРНК и 3 гена рРНК. Этот митогеном в настоящее время является самым крупным из известных. Собранный геном имеет достаточное качество для дальнейших подробных исследований и сравнительных анализов с митогеномами других растений.

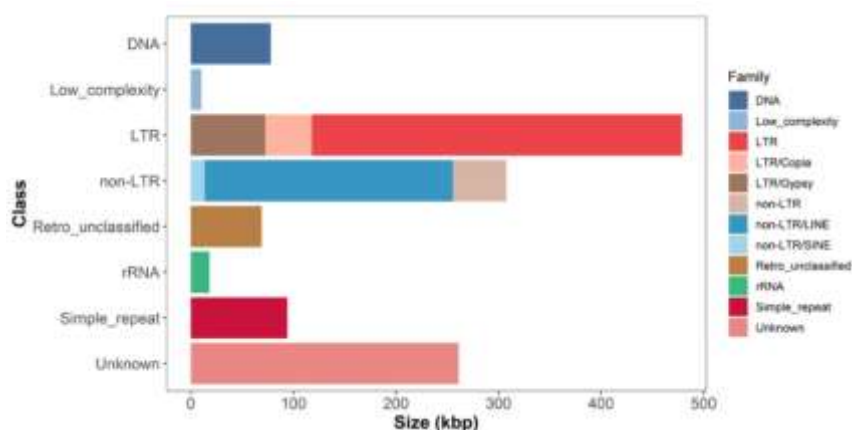


Рис. 3.5 – Список повторяющихся элементов, идентифицированных в митогеноме лиственницы сибирской

Вопрос распределения стартов считывания по нуклеотидной последовательности является довольно острым, так как многочисленные неоднородности в этом распределении могут принести проблемы в сборке, аннотации и дальнейшем анализе генетических объектов. Для выявления неоднородности в распределении был использован обобщенный подход. Мы рассматриваем геном как символьную последовательность и воздерживаемся от внесения каких-либо биологических знаний до окончания анализа. Другими словами, мы ищем очень неожиданные участки в символьной последовательности. Как только такие участки найдены, изучается их биологическая роль. Было установлено, что участки распределены по геному очень неслучайно, с явным

предпочтением некоторых биологически заряженных локусов (рис. 3.6). Чтобы выявить структурированность в строках, содержащих нуклеотиды с различным количеством стартов считывания, мы использовали идею информационной емкости как усредненной меры, указывающей на характер распределения в целом. Чтобы улучшить анализ, была реализована идея информационно ценных слов. Для каждого слова длины ω подсчитываются его частота f_ω и ожидаемая частота $f_\omega^{\%}$. Те слова, для которых разность f_ω и $f_\omega^{\%}$ превышает некоторый порог α считаются информационно ценными. Далее эти слова локализуются в геноме и определяется соответствующая им аннотация. Есть предположение, что такие слова будут соответствовать некоторым специфическим участкам генома.

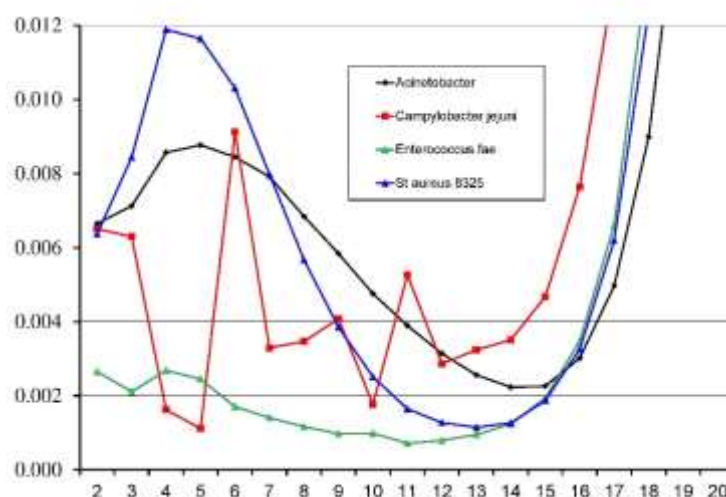


Рис. 3.6 – Информационная емкость для двоичной символьной последовательности, представляющей распределение стартов

Чтобы получить представление о предполагаемых функциях генов гомеобокса у *P. sylvestris*, его образцы экспрессии были проанализированы в различных тканях. Было идентифицировано 46 статистически значимо дифференциально экспрессируемых троичных «генов» и отображено в виде тепловой карты. Были проанализированы пять полученных кластеров транскриптов DE (рис. 3.7). В первом кластере были обнаружены два транскрипта, кодирующие белки гомеобокса 9, связанные с WUSCHEL (WOX9), и богатый поздним эмбриогенезом белок D-34, участвующий в раннем развитии, росте меристемы и позднем развитии эмбриона соответственно, которые были активированы в эмбрионе и мегагаметофите. стадиях и понижается на других стадиях. Второй кластер включает транскрипты, кодирующие факторы транскрипции (TF), а именно белок гомеодомена 1 цинкового пальца, белок гомеобокс-лейциновой молнии ROC (1,2,8), белок гомеобокс-лейциновой молнии PROTODERMAL FACTOR 2, участвующий в

развитии семядолей, прорастании семян и дифференцировка эпидермальных клеток, подавленная в ткани флоэмы. В третьем кластере транскрипты, аннотированные как белки WUSCHEL-относящиеся к гомеобоксу (1,4 и 8), ответственные за развитие органов растений, деление клеток, развитие многоклеточных организмов. Четвертый кластер включает белки, такие как BEL1-подобный гомеодомен 4, который определяет форму листа путем подавления роста в определенных субдоменах листа, белок гомеобокса лейциновой молнии, гомеобоксный белок узловатый-1-подобный 3 (KNAT3), участвующие в обнаружении и клеточном ответе на цитокиновый раздражитель, ответ на световой раздражитель. Транскрипты в пятом кластере определяют белки, включая белок гомеобокс-лейциновой молнии, участвующий в морфогенезе семядолей, позитивную регуляцию транскрипции и белок WUSCHEL. участвует в дифференцировке клеток. Было обнаружено, что активация транскриптов происходит в зародыше.

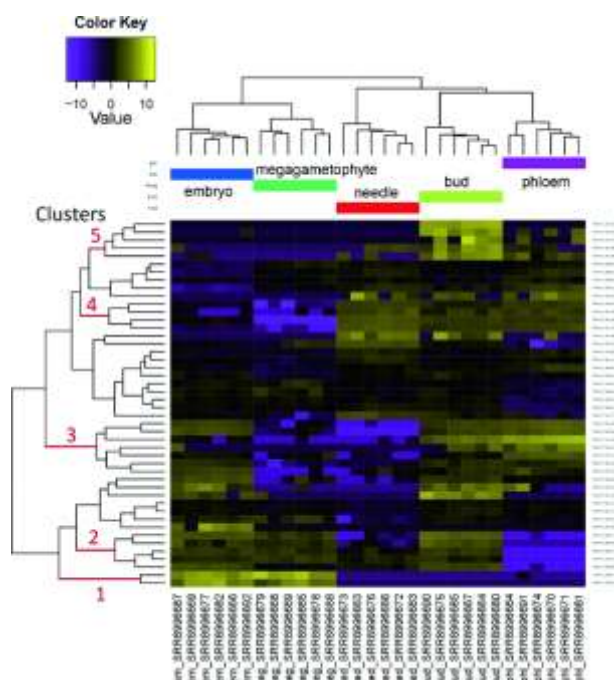


Рис. 3.7 – Тепловая карта экспрессии генов пяти тканей *P. sylvestris*.

Горизонтальная ось представляет ткани, правая вертикальная ось представляет «гены», идентифицированные Тринити

После анализа профилей дифференциальной экспрессии генов гомеобокса можно оценить, какие из генов гомеобокса или набор генов имеют решающее значение для развития различных тканей. Было обнаружено, что транскрипты кластера 5-го значительно активируются в ткани зачатка, что предполагает важность роли этих генов в развитии этой ткани. Транскрипты этого класса в основном включают члены генов класса HD-ZIP I, а также один транскрипт, аннотированный как член класса WOX. Кроме того, гены класса BEL предположительно участвуют в развитии зрелых тканей *P. sylvestris*.

3.2 Быстрый метод выявления длинных и сверхдлинных общих подстрок в двух или нескольких символьных последовательностях

Аннотация. Разработан алгоритм быстрого поиска повторяющихся подпоследовательностей нуклеотидов в одной или нескольких генетических последовательностях, основанный на сравнении сильно прореженных частотных словарей. Найденные подпоследовательности совпадают с точностью до некоторой заданной плотности локальной погрешности.

Предложен новый, устойчивый к ошибкам метод сравнения и анализа символьных последовательностей. Метод основан на вычислении функции свертки, где функция определяется над двоичными числовыми последовательностями, полученными путем конкретного преобразования исходной последовательности символов. Метод допускает высокопараллельное выполнение и имеет большое значение для поиска мутаций вставки / удаления. Для вычисления функции свертки в реализации метода используется быстрое преобразование Фурье.

Выводы. Разработанный алгоритм быстрого поиска повторяющихся подпоследовательностей нуклеотидов в одной или нескольких генетических последовательностях на основе сравнения сильно прореженных частотных словарей позволяет находить существенно больше почти совпадающих подпоследовательностей по сравнению с BLAST-подобными алгоритмами.

Важным преимуществом метода, основанного на вычислении свёртки двух полиномов, полученных специальным преобразованием исходных сравниваемых последовательностей, является его высокая чувствительность и простота применения для поиска таких малых ошибок, которые доставляют наибольшие трудности для всех остальных методов поиска с ошибками — поиск с (малыми) выпадениями и вставками.

Описание результатов. Разработан алгоритм быстрого поиска повторяющихся подпоследовательностей нуклеотидов в одной или нескольких генетических последовательностях. Алгоритм позволяет искать не только строго совпадающие подпоследовательности, но и подпоследовательности с заданной плотностью локальной погрешности. Под погрешностями понимаются несовпадения подпоследовательностей, возникающие в следствии удаления символа, вставки символа, либо замены одного символа на другой.

Метод основан на сравнении специально построенных очень сильно прореженных частотных словарей, что позволяет уменьшить число сравнений на $4\div 6$ порядков; это позволяет ускорить поиск на несколько порядков. Метод построения таких

специализированных частотных словарей состоит в следующем: если ожидаемая длина общей подпоследовательности в двух сравниваемых последовательностях $\mathfrak{S}_1$ и $\mathfrak{S}_2$ равна N , тогда построим частотный словарь, содержащий слова длины $l = \lfloor \sqrt{N} \rfloor$, выделяемые в последовательности $\mathfrak{S}_1$ с шагом l . В последовательности $\mathfrak{S}_2$ построим частотный словарь, содержащий слова длины l , но выделяемые с шагом длины $l-1$. Если существует общая подпоследовательность длины N , то вне зависимости от того, где именно в $\mathfrak{S}_1$ и $\mathfrak{S}_2$ располагается эта общая подпоследовательность, всегда с необходимостью найдётся хотя бы одно общее слово в указанных словарях.

Были построены две модификации данного алгоритма, позволяющие осуществлять приблизительный поиск. Первая модификация осуществляет поиск точных соответствий, но допускает малые ошибки при построении расширения от зерна поиска (длины l) до всей подпоследовательности длины N , при котором расширение вправо и влево от зерна поиска также проводится по словам длины l и с шагом длины l , но при этом возможны несовпадения в отдельных символах. Критерием остановки является либо превышение локальной плотности ошибок структурных элементов в нуклеотидных последовательностях и их анализом.

Пусть имеется символьная последовательность из алфавита $\aleph$ мощности $K = |\aleph|$ длины $N \approx 10^{10}$ символов и образец (также символьная последовательность) из того же алфавита длины $L \leq N$. Требуется найти все включения образца в рассматриваемую последовательность, при том условии, что каждое такое включение содержит не более δ ошибок (несовпадений).

На первом шаге исходная символьная последовательность из алфавита мощности $K = |\aleph|$ преобразуется в набор из K последовательностей каждая. В результате исходная последовательность заменяется на K бинарных (т.е. из алфавита $\mathfrak{S} = \{0,1\}$) последовательностей. Это преобразование является ключевым для работы предлагаемого метода и обеспечивает алгоритмическое распараллеливание в K раз. Аналогично, образец, для которого требуется найти все его (быть может, неточные) включения в анализируемую последовательность, также преобразуется в набор из K последовательностей алфавита $\mathfrak{S} = \{0,1\}$.

Опишем это преобразование на примере нуклеотидных последовательностей, имеющих алфавит $\aleph = \{A, C, G, T\}$ мощности 4. Исходная символьная последовательность

преобразуется в четыре числовые из алфавита $\mathfrak{Z}$ по следующему правилу. В первой последовательности заменим все символы $\mathbb{A}$ на 1, а все остальные — на 0. Аналогично, во второй, третьей и четвёртой последовательностях заменим на 1 символы $\mathbb{C}$, $\mathbb{G}$ и $\mathbb{T}$, соответственно, а остальные — на 0. Аналогично преобразуем образец в четыре бинарных образца.

Остановимся на определении дискретной функции совпадений по каждому символу в отдельности. Пусть это будет последовательности $u_0, u_1, \dots, u_{N-1}, u_N$ и $n_0, n_1, \dots, n_{L-1}, n_L$, соответствующие символу $\mathbb{A}$. На следующем шаге инвертируем одну из последовательностей — пусть это будет тот образец, для которого ведётся поиск. Это означает, что все $K \in \{0,1\}$ последовательностей для образца вида $n_0, n_1, \dots, n_{L-1}, n_L$ преобразуется в обратную последовательность $z_1, z_2, \dots, z_{L-1}, z_L = n_L, n_{L-1}, \dots, n_1, n_0$.

Будем рассматривать получившиеся бинарные последовательности как коэффициенты полиномов (степени $N-1$ и $L-1$). Легко проверить, что коэффициенты $w_0, w_1, \dots, w_{N+L-3}, w_{N+L-2}$ их произведения равны скалярным произведениям

$$\begin{aligned} w_j &= \sum_{s=0, K, j} u_s z_{j-s} = \sum_{s=0, K, j} u_s n_{L-1-j+s}, \text{ если } j \leq L-1 \text{ и} \\ w_j &= \sum_{s=0, K, j} u_{s+j-L+1} z_{j-s} = \sum_{s=0, K, j} u_{s+j-L+1} n_{L-1-j+s}, \text{ если } j > L-1 \end{aligned} \quad (1)$$

Тогда наличие одинаковых подпоследовательностей из символа $\mathbb{A}$ при некотором j_0 будет характеризоваться уверенным скачком значения (1) на фоне возможных случайных совпадений при других $j \neq j_0$. Для быстрого перемножения двух многочленов будем использовать БПФ [8]. Поскольку оно даёт наибольший выигрыш для числа элементов со степенями двойки, то обе бинарные последовательности дополняются нулями в позициях до ближайшей степени двойки, большей $N+L-2$.

3.3 Методы выявления особенностей структурирования и связей структурных единиц с биологическими свойствами носителей генетической информации

Аннотация. Исследована связь между кластеризацией частотных словарей как нуклеотидов, так и аминокислот семейства митохондриальных генов *atp6*, *atp8* и *atp9*, их функцией и таксономией носителей.

Для 223 митохондриальных геномов были вычислены частотные словари как нуклеотидов так и аминокислот. Этим частотным словарям были поставлены в соответствие точки в соответствующем метрическом пространстве. Было обнаружено, что

точки группируются в три кластера, каждый из которых соответствует одному из генов *atp6*, *atp8* и *atp9* независимо от выбора алфавита.

Описание результатов. Рассмотрено взаимное влияние трех основных генетических объектов – структуры, функции и таксономии. Чтобы сделать это, создана база данных, включающая гены АТФ-синтазы грибных митохондрий, а именно гены *atp6*, *atp8* и *atp9*. Затем гены были преобразованы в частотные словари, вычисленные как для нуклеотидов, так и для аминокислот, так, что каждый ген был представлен в виде точки в соответствующем Евклидовом пространстве, где полученные частоты являются координатами. При анализе пространственного расположения точек была обнаружена внутренняя структурированность данных. Методом упругих карт были выявлены три кластера, кроме того, структурированность другого типа была найдена при линейной классификации (метод K-means). На следующем этапе был проанализирован состав кластеров, образованных 343 АТФ-синтазными генами митохондрий грибов с точки зрения видового и генного состава. Были проверены кластеры, полученные обоими упомянутыми выше методами (рис. 3.8, 3.9).

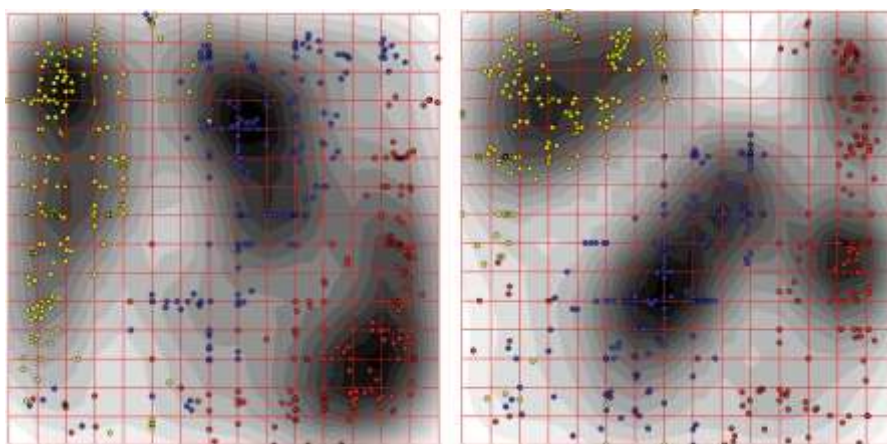


Рис. 3.8 – Распределение генов семейства *atp* во внутренних координатах упругой карты. Частотные словари для триплетов нуклеотидов. Гены *atp6* обозначены красным цветом, *atp8* синим и *atp9* желтым; слева частотный словарь для шага $t = 1$, справа для $t = 3$

Обнаружено, что в формировании кластеров большое влияние оказывают именно гены. Такое преобладание заранее не очевидно. Преобладание влияния генов доказывает превосходство функции над таксономией в формировании структурированности в триаде структура – функция – таксономия.

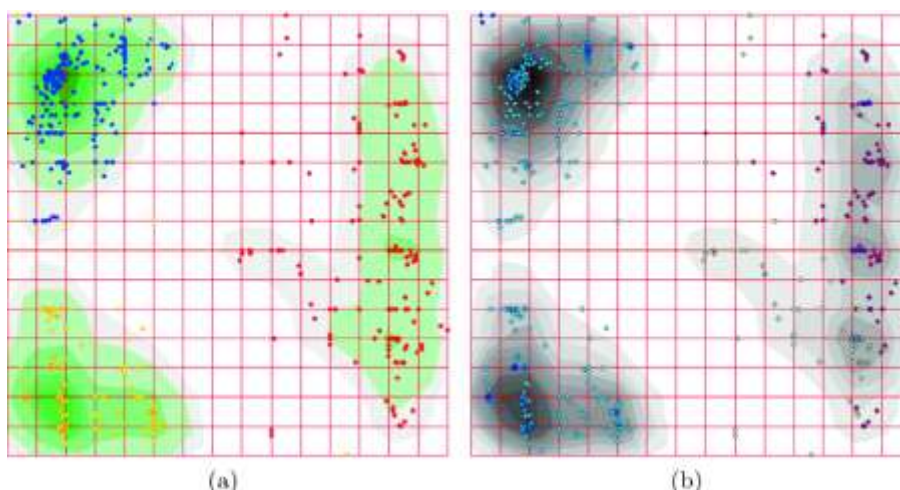


Рис. 3.9 – Распределение генов *atr* на упругой карте: *atr6* – синий, *atr8* – красный, *atr9* – желтый (слева). Справа такое же распределение с классификацией К-средних, 4 класса

3.4 Анализ медицинских данных, выявление клинически значимых характеристик в больших массивах данных

Аннотация. Анализ данных по острой недостаточности мозгового кровообращения (ОНМК). Данное заболевание является очень грозным, трудно поддающимся лечению и требующего продолжительной реабилитации больного. Для этого типа заболеваний особую остроту представляет поиск и выделение предикторов течения заболевания и прогноза его отдалённых последствий. Медики выделяют три генерализованных типа больных такими заболеваниями: лёгкий, средний и тяжелый.

Болезни экстрапирамидной системы, в частности болезнь Паркинсона (БП) и эссенциальный тремор (ЭТ), являются одними из самых распространенных нозологий в неврологии. Они имеют очень похожую симптоматику, что затрудняет их дифференциальную диагностику, значительно снижая успешность лечения.

Злокачественные новообразования являются второй причиной смертности в мире. Ряд научных работ посвящен взаимосвязи между отклонениями показателей общего анализа крови (ОАК) и клиническими проявлениями рака, оценкой риска прогрессирования опухоли или появлением отдалённых метастазов.

Анализ выявленных кластеров по такому показателю, как отдалённые последствия, показал, что один из кластеров, который содержал больных с динамикой содержания BDNF и VEGF, которую традиционно рассматривают как проявление выздоровления, на самом деле включал в себя исключительно больных, скончавшихся в течение месяца после начала заболевания. Иными словами, такая специфическая динамика содержания BDNF и VEGF может служить индикатором развития осложнений и предиктором отдалённых последствий.

Подтверждена гипотеза об обратной зависимости выраженности тремора и степени снижения обоняния. Найдены параметры, по которым можно дифференцировать больных с БП и ЭТ.

Показана эффективность оценки общего анализа крови (ОАК) для прогнозирования и коррекции лечения пациентов с раком молочной железы и раком толстой кишки. Результаты общего анализа крови использовались для отслеживания динамики пациентов, пролеченных курсом интраоперационной химиотерапии.

Описание результатов. Был проведен анализ данных по острой недостаточности мозгового кровообращения (ОНМК). Для выделения предикторов течения заболевания было проведено исследование, включавшее 56 пациентов с разными типами заболевания ОНМК. У этих больных бралась венозная кровь на первый, седьмой и 21 дни после госпитализации, в крови определялось содержание двух важнейших гормонов, влияющих на восстановление таких больных:

- нейротрофического фактора, выделяемого мозгом (BDNF), и
- васкулярного эндотелиального фактора роста (VEGF).

Кроме того, для таких больных определялось их состояние по двум шкалам, введенным в практику лечения и применяющимся во всём мире: Рэнкин и Ривермид. Именно эти шкалы определяют тип (легкий, средний и тяжёлый) течения заболевания.

Стандартные методы статистики не выявляли различий между такими группами больных, с точки зрения содержания BDNF и VEGF в их крови. Напротив, анализ методом упругих карт выявил три кластера, которые объединяли больных по данным шкал Рэнкина и Ривермид. Выяснилось два важнейших для практической медицины обстоятельства: 1) выделяемые кластеры, строго говоря, включают больных с различной тяжестью течения заболевания, у больных в выделенных кластерах различная динамика содержания BDNF и VEGF в крови.

Было обнаружено, что кластер, содержащий пациентов, по существующей методике считающихся выздоравливающими, на самом деле содержал больных, скончавшихся в течение месяца после начала заболевания. Таким образом, такая специфическая динамика содержания BDNF и VEGF может служить индикатором развития осложнений и предиктором отдалённых последствий.

Дифференциальная диагностика больных с болезнью Паркинсона и с эссенциальным тремором очень важна, поскольку болезни экстрапирамидной системы являются одними из самых распространенных нозологий в неврологии и имеют очень похожую симптоматику. Это затрудняет их диагностику, значительно снижая успешность лечения.

БП и ЭТ объединяет такой симптом, как дрожание. Тремор является одной из основных жалоб, по поводу которой пациенты обращаются за помощью к врачу, так как качество повседневной жизни заметно ухудшается уже на ранних стадиях развития указанных заболеваний. Отличительной особенностью БП является тремор покоя, который значительно снижается или исчезает во время движения или других активных действий пациента. Для ЭТ, напротив, характерен постурально-кинетический тремор. За несколько лет до возникновения двигательных расстройств при БП развиваются и обонятельные нарушения. Это обусловлено тем, что нейродегенерация начинается именно в обонятельных луковицах головного мозга. Для ЭТ характерен другой механизм развития болезни и, соответственно, данного симптома у больных с ЭТ нет. Следовательно, ольфакторная дисфункция может быть ранним и значимым дифференциальным признаком БП.

Во всем мире в последние десятилетия активно используют обонятельные тесты. В нашей работе был выбран Sniffin's sticks test. Сниффин Стикс тест состоит из трёх субтестов, каждый из которых описывает больного 16-ю переменными. Кластеризация обследуемых по результатам 3-х субтестов показала, что 1 субтест не способен различить исследуемые группы между собой, доказательство этому служит рисунок 3.10. На рисунке 3.10а видно, что все группы слились в единые точки (кластер), это наглядно доказывает, что и больные, и здоровые люди набрали одинаково малые баллы в 1 субтесте, в ситуации со 2 (рис. 3.10б) и 3 субтестом (рис. 3.10в) группы больных и здоровых различимы.

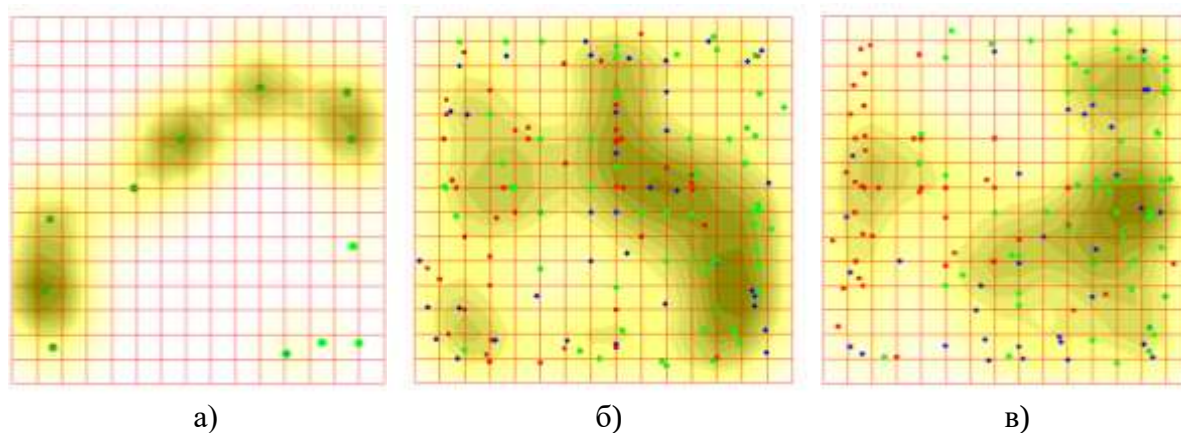


Рис. 3.10 – Упругая карта распределения обследуемых по результатам 1 субтеста (а), 2 субтеста (б), 3 субтеста (в). Зелёным цветом отмечены здоровые люди, красным – люди с БП, синим – люди с ЭТ

Дрожание было проанализировано у 90 человек: 30 условно здоровых людей и по 30 пациентов с диагнозами БП и ЭТ. Были исследованы различия в значениях частоты и амплитуды электромиограммы (ЭМГ) правой и левой руки в каждой из представленных

групп людей. Кроме того, был проведён анализ на наличие различий по данным показателям между исследуемыми группами. По показателю частоты колебаний потенциала ЭМГ были найдены достоверные различия с датчиков на левой кисти: между людьми с БП и людьми с диагнозом ЭТ — во всех пяти тестах; между здоровыми людьми и людьми с диагнозом ЭТ — в первом, третьем и четвёртом тесте; между здоровыми людьми и людьми, страдающими БП — в пятом тесте.

Так как показатели дрожания количественно значительно преобладают над показателями обонятельной функции (280 и 52 соответственно), было принято решение оставить только те показатели тремора, которые являются наиболее информативными. Упругая карта, включающая только выбранные показатели тремора (рис. 3.11а), а также карта, включающая данные обонятельных тестов и выборочные данные исследования тремора рук пациентов, показана на рис. 3.11б.

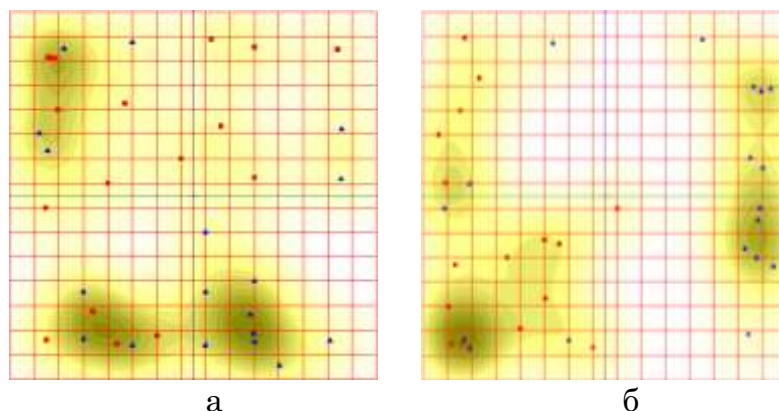


Рис. 3.11 – Распределение людей с БП и ЭТ по выборочным результатам исследования тремора (а) и сочетания показателей тремора и результатов обонятельных тестов (б).

Красным квадратом обозначены люди с БП, синим треугольником – пациенты с ЭТ

В пространстве объединённых данных тремора и обонятельной функции пациенты с различными заболеваниями хорошо разделились друг от друга, образовав 3 отдельных кластера (рис. 3.11).

Проведен программный анализ базы данных, которая содержит лабораторные результаты развернутого анализа крови пациентов с разными нозологическими формами рака. Образцы крови были взяты у пациентов, наблюдающиеся в Красноярском краевом онкологическом диспансере им. А.И. Крыжановского. База данных содержит 867 записей и включает следующие параметры развернутого анализа крови: BAS (абсолютное содержание базофилов в периферической крови), EOS (абсолютное содержание эозинофилов в периферической крови), НСТ (показатель гематокрита), НGB (концентрация гемоглобина в цельной крови), IG (содержание иммуноглобулинов в

периферической крови), LYM (абсолютное количество лимфоцитов в периферической крови), MCH (среднее содержание гемоглобина на эритроцит), MCHC (уровень насыщения эритроцитов гемоглобином), MCV (средний объем эритроцитов), MON (абсолютное количество моноцитов в периферической крови), MPV (средний объем тромбоцита), NEUT (абсолютное количество нейтрофилов в периферической крови), P-LCR (часть больших тромбоцитов), PCT (процент тромбоцитарной массы в объеме крови), PDW (относительная ширина распределения тромбоцитов по объему), PLT (абсолютное количество тромбоцитов в периферической крови), RBC (абсолютное количество эритроцитов в периферической крови), RDW-CV (изменение относительной ширины распределения тромбоцитов по объему), RDW-SD (стандартное отклонение относительной ширины распределения тромбоцитов по объему), WBC (абсолютное количество лейкоцитов в периферической крови) и скорость оседания эритроцитов (СОЭ).

Предварительная обработка данных включала стандартный статистический анализ, а именно:

- для каждой переменной были определены среднее значение и стандартное отклонение;
- каждая переменная проверена на соответствие нормальности распределения;
- для каждой переменной была разработана гистограмма для оценки модели распределения.

На следующем этапе был реализован анализ главных компонент [19]. МГК был применен для определения эффективного (линейного) измерения данных, а корреляционная матрица была рассчитана для определения пар переменных с высоким линейным ограничением. Последний использовался для выбора переменных из записей РАК, которые необходимо исключить из дальнейшего анализа. Таким образом, каждый пациент представлен точкой в 21-мерном метрическом пространстве. Везде ниже мы использовали евклидовы метрики для измерения расстояния в пространстве. Выявлялись неоднородности в распределении пациентов, включенных в базу данных, только в 21-мерном евклидовом пространстве по данным РАК. Если такие кластеры были обнаружены, то проводился их предметный анализ с точки зрения нозологических форм, особенностей пациентов или характеристик заболевания. В свою очередь, был рассмотрен и двойкий вопрос. Изучалась структура и уровень однородности распределения пациентов в разных кластерах. Другими словами, было очень важно узнать какой показатель РАК играет ключевую роль в формировании кластера.

На рис. 3.12 показана кластеризация, полученная с помощью мягкой (16×16) упругой карты, с параметрами эластичности, заданными по умолчанию. На карте выделяется три или четыре кластера, в зависимости от параметра контраста.

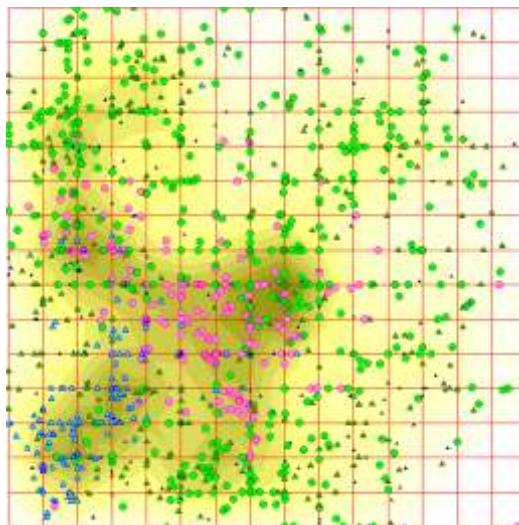


Рис. 3.12 – Распределение пациентов по эластичной карте 16×16 . Мужчины показаны треугольниками, женщины – кружками. Все больные онкологическими заболеваниями показаны оттенками зелёного, все здоровые – голубым (мужчины) и розовым (женщины)

Было обнаружено, что все показатели крови, в размере 21, разделены на две группы с точки зрения распределения паттернов пациентов на эластичной карте. Вариант, при котором данные больных, в виде точек, хаотично размещались на карте, в виде так называемого “звездного неба”, наблюдался для базофилов, эозинофилов, лейкоцитов и иммуноглобулинов. Для остальных показателей наблюдалось волновое распространение.

Не существует стратификации пациентов с точки зрения соответствия кластеров, идентифицированных методом эластичной карты, возрасту, полу или нозологической формы опухоли. Напротив, пациенты стратифицированы по типу ответа на раковую опухоль. Между составом кластера и характером заполнения карты прослеживается очевидная корреляция. Последнее отражает тип и способ мобилизации ресурсов организма, вызванных злокачественным новообразованием, а стратегия стратификации следует принципу Либиха. Это в основном новая стратификация, наблюдаемая у онкологических больных, и все еще необходимы дальнейшие исследования детальных взаимосвязей между типом ответа и соответствующими важными клиническими аспектами.

ЗАКЛЮЧЕНИЕ

За период выполнения проекта получены следующие результаты.

Разработаны модели и алгоритмы обнаружения технологических нарушений и снижения производительности в цикле производства алюминия, основанные на применении совокупности методов математической статистики, интеллектуального анализа данных и машинного обучения. В результате комплексного структурного многомерного анализа данных среднесуточного мониторинга определены закономерности функционирования комплекса производства алюминия и характерные условия возникновения отклонений в технологическом процессе, что позволило сформировать «аналитические портреты» работы отдельных единиц комплекса производства алюминия. Разработаны алгоритмы идентификации изменений характера временных рядов параметров, свидетельствующих о возникновении технологических нарушений, созданы средства визуализации для оценивания текущего состояния технологии и динамики изменения производительности отдельных объектов по комплексу параметров производства алюминия.

Разработан метод построения многомерной аналитической модели состояния сложных объектов и систем на основе выявления и формального описания закономерностей. Выполнено исследование и аналитическое моделирование состояния гидроэнергетической системы на примере одного из гидроагрегатов Красноярской ГЭС. Сравнение аналитических моделей за разные периоды времени позволяет отслеживать динамику состояния системы и отдельных элементов.

Разработаны новые элементы технологии интеллектуального имитационного моделирования в жизненном цикле бортовой аппаратуры космических аппаратов. Выполнена тесная интеграция интеллектуальной имитационной модели и методов автоматизации испытаний бортовых систем. Создан метод, позволяющий сопоставлять результаты имитационных экспериментов с результатами испытаний реального оборудования. Метод позволяет автоматически выявлять причинно-следственные связи между получаемыми данными испытаний и параметрами работы бортовых систем. Имеется возможность оценивания корректной работы реального оборудования по модели. С другой стороны, в случае, если полученные при испытаниях данные отражают корректное поведение бортовых систем, метод может применяться для анализа соответствия имитационной модели реальному оборудованию. Новый метод анализа модели по результатам испытаний позволяет упростить отладку имитационной модели и обладает высоким потенциалом масштабирования для поддержки больших моделей.

Разработаны методы управления консолидированными данными в

интегрированных системах территориального и корпоративного управления. Разработаны методические и программные средства многоэтапной консолидации гетерогенных данных, в ходе которой сведения, получаемые от удаленных веб-ресурсов, согласуются с данными распределенной отраслевой информационной системы. Апробация разработанных средств консолидации гетерогенных данных выполнена для организации муниципальных закупок г. Красноярск при обмене данными между Автоматизированной системой поддержки муниципальных закупок и Общероссийским официальным сайтом zakupki.gov.ru.

Разработаны новые методы и средства информационно-аналитической поддержки управления территориальной безопасностью территорий, в том числе методика формирования территориально-ориентированного стандарта, представляющего нормативную модель безопасности с учетом индивидуальных особенностей территорий, метод расчета интегральных оценок природно-техногенной безопасности территорий, программные приложения, обеспечивающие безопасность туризма.

Создана инструментальная веб-платформа для автоматизации процесса создания прикладных модельно-ориентированных систем сбора и анализа данных. Платформа позволяет создавать прикладные системы на основе моделей, описывающих предметную область, что обеспечивает возможность расширения тематического наполнения. Новые функциональные возможности реализованы за счет интеграции методов интеллектуального анализа данных и средств динамической инфографики и динамического картографирования. Реализация различных комбинаций аналитических методов обеспечивает поддержку в создаваемых на основе платформы прикладных системах возможностей нативного исследования консолидированных данных, позволяет анализировать как отдельные объекты, выбранные на карте, так и результаты анализа, представляя их в виде современных диаграмм, таблиц или картограмм.

Модернизирована структура институционального репозитория ИВМ СО РАН, расширен круг применяемых web-технологий, схем и форматов метаданных цифровых объектов и системы их хранения. Применение web-технологий открытых архивов, их адаптация, развитие и внедрение позволяют создавать и продвигать собственные ресурсы и результаты исследований, эффективно и согласованно влияя на формирование информационной среды Интернет. Институциональный репозиторий, формируемый в технологиях архивов открытого доступа, позволяет проводить качественный учет и продвижение научных публикаций, обеспечить открытость результатов научных исследований для мирового сообщества, повысить индивидуальные показатели цитирования.

Разработаны новые модели безопасности на основе расширенной ролевой модели безопасности RBAC: параметрическая, сервис-ориентированная, мультидоменная. Они позволяют реализовывать сложные политики разграничения доступа в веб-приложениях, учитывать особенности функционирования веб-сервисов и обеспечивать разграничение доступа с использованием методов запроса в системах, использующих технологии REST и RPC, а также обеспечивать разграничение доступа в системах, использующих сложные политики безопасности в нескольких пространствах доменных имён. Созданные модели безопасности могут применяться для решения широкого круга задач. Разработан метод интегральной оценки безопасности веб-приложений, который использует комплексный набор критериев, состоящий из классических критериев ASVS, а также дополнительных критериев, учитывающих свойства модели безопасности и корректность алгоритмов разграничения доступа. Созданный метод позволяет повысить уровень защищённости веб-приложений и веб-сервисов за счёт комплексной оценки всех компонентов системы.

Разработаны статистические методы распознавания образов и восстановления многомерных зависимостей в условиях выборок неоднородных данных большого объёма, основанные на дискретизации области их определения, принципах коллективного оценивания и гибридного моделирования. Установлены условия асимптотической сходимости соответствующих им непараметрических решающих функций. На этой основе определены аналитические зависимости показателей эффективности оценивания решающих функций от особенностей исходных данных и используемых методов их обработки.

Разработаны непараметрические алгоритмы распознавания образов и модели стохастических зависимостей, которые обеспечивают эффективную обработку информации в условиях малых выборок. Идея синтеза предложенных непараметрических систем основана на декомпозиции исходных обучающих выборок по размерности, построении частных решающих функций и последующем их обобщении в виде линейных либо нелинейных коллективов решающих функций. Исследованы свойства показателей эффективности подобных систем по сравнению с традиционными алгоритмами и моделями ядерного типа.

Предложены и исследованы быстрые алгоритмы оптимизации непараметрических решающих правил ядерного типа в условиях статистических данных большого объёма. Обоснован и разработан быстрый выбор оптимальных коэффициентов размытости ядерных функций в непараметрических оценках многомерной плотности вероятности типа Розенблатта-Парзена, уравнения разделяющей поверхности в двухальтернативной задаче распознавания образов. Их применение позволяет на несколько порядков сократить

временные затраты при синтезе непараметрических решающих правил.

На основе анализа асимптотических свойств регрессионной оценки плотности вероятности предложен метод оптимальной дискретизации области значений двумерной случайной величины, что является одной из актуальных задач математической статистики. Разработаны процедуры оценивания нелинейного функционала от плотности в формуле расчёта оптимального количества интервалов дискретизации. Полученные результаты обобщены на многомерный случай на основе использования метода аналогий.

Обоснована возможность использования непараметрических алгоритмов распознавания образов, соответствующих критерию максимального правдоподобия, в задачах проверки статистических гипотез о распределениях случайных величин. Задача проверки гипотезы о тождественности двух законов распределений многомерных случайных величин может быть заменена на проверку гипотезы о равенстве ошибки распознавания образов пороговому значению 0,5. При больших объёмах статистических данных синтез непараметрического алгоритма распознавания образов осуществляется на основе регрессионных оценок плотностей вероятности распределения случайных величин в классах. Предлагаемые алгоритмы распознавания образов позволяют уменьшить объём обучающей выборки за счёт декомпозиции области значений случайных величин.

Предложена и исследована новая методика проверки гипотезы о независимости случайных величин, которая основывается на использовании непараметрического алгоритма распознавания образов, соответствующего критерию максимального правдоподобия. В отличие от традиционной постановки задачи распознавания образов априори отсутствует обучающая выборка. Законы распределения случайных величин в классах оцениваются по исходным статистическим данным для условий их зависимости и независимости. По минимальному значению оценки вероятности ошибки распознавания образов в классах принимается решение о зависимости либо независимости случайных величин.

Создана информационная система для оценивания состояний природных объектов по данным дистанционного зондирования земли, основанная на предложенных непараметрических методах принятия решений. Система используется для исследования состояний лесных массивов.

Разработан алгоритм быстрого поиска повторяющихся подпоследовательностей нуклеотидов в одной или нескольких генетических последовательностях на основе сравнения сильно прореженных частотных словарей, который позволяет находить существенно больше почти совпадающих подпоследовательностей по сравнению с BLAST-подобными алгоритмами. Важным преимуществом метода, основанного на

вычислении свёртки двух полиномов, полученных специальным преобразованием исходных сравниваемых последовательностей, является его высокая чувствительность и простота применения для поиска таких малых ошибок, которые доставляют наибольшие трудности для всех остальных методов поиска с ошибками – поиск с (малыми) выпадениями и вставками.

Исследована связь между кластеризацией частотных словарей как нуклеотидов, так и аминокислот семейства митохондриальных генов *atp6*, *atp8* и *atp9*, их функцией и таксономией носителей. Для 223 митохондриальных геномов были вычислены частотные словари как нуклеотидов так и аминокислот. Этим частотным словарям были поставлены в соответствие точки в соответствующем метрическом пространстве. Было обнаружено, что точки группируются в три кластера, каждый из которых соответствует одному из генов *atp6*, *atp8* и *atp9* независимо от выбора алфавита.

Предложены методы анализа медицинских данных, основанные на выявлении клинически значимых характеристик в больших массивах данных. Полученные результаты имеют высокую значимость для решения задач поиска предикторов течения заболевания и прогноза отдаленных последствий острой недостаточности мозгового кровообращения, дифференциальной диагностики болезней экстрапирамидной системы (болезнь Паркинсона и эссенциальный тремор), оценки риска прогрессирования опухоли или появления отдалённых метастазов злокачественных новообразований.

ПРИЛОЖЕНИЕ. Научные публикации

в изданиях, индексируемых в российских и международных
информационно-аналитических системах научного цитирования

1. Mikhalev A.S. Lugovaya N.M. Penkova T.G. Prediction of technological process parameters in aluminum production based on artificial neural networks // J. Phys.: Conf. Ser. – 2020. – Vol. 1679. – 032092. doi:10.1088/1742-6596/1679/3/032092.
2. Nicheporchuk V.V., Nozhenkov A.I. Emergencies situational modeling technology for territorial management support // Procedia Structural Integrity. – 2020. – Vol. 20. – P. 248-253 doi: 10.1016/j.prostr.2019.12.147.
3. Nozhenkov A., Korobko A. Technologies and applications for the support of tourism in Krasnoyarsk region // IOP Conference Series: Materials Science and Engineering. – 2020. – Vol. 862, Issue 5. – 052066. doi: 10.1088/1757-899X/862/5/052066.
4. Исаева О.С., Кулясов Н.В., Исаев С.В. Метод структурно-графического анализа и верификации интеллектуальной имитационной модели // Вестн. Том. гос. ун-та. Управление, вычислительная техника и информатика. – 2020. – № 50. – С. 79-88. doi: 10.17223/19988605/50/10.
5. Исаева О.С., Ноженкова Л.Ф., Мишуров А.В., Камышников А.Н., Евстратько В.В., Черниговский А.В. Создание гетерогенной имитационной модели командно-измерительной системы космического аппарата // Космические аппараты и технологии. – 2020. – № 4(32). – С. 107-115. doi: 10.26732/j.st.2020.2.05.
6. Korobko A.V., Korobko A.A. A software platform for constructing model-driven systems for primary data consolidation // J. Phys.: Conf. Ser. – 2020. – Vol. 1679. – 032056. doi: <https://doi.org/10.1088/1742-6596/1679/3/032056>.
7. Kovyazina E.V. The Digital Repository Integration with External Information Services // CEUR Workshop Proceedings. DICR 2019 - Selected Proceedings of the 17th International Conference Distributed Information-Computational Resources. – 2019. – Vol.2569. – P. 55-58. doi: 10.25743/ICT.2019.97.64.033.
8. Ковязина Е.В. Особенности работы с цифровыми документами: проблемы и решения // КНИГА. КУЛЬТУРА. ОБРАЗОВАНИЕ. ИННОВАЦИИ.: сборник докладов Пятого Международного профессионального форума «Крым-2019». Государственная публичная научно-техническая библиотека России, 2020. – С. 117-121. doi: 10.33186/978-5-85638-223-4-2020-117-121.

9. Lapko A.V., Lapko V.A. Dependencies Between Histogram Parameters and the Kernel Estimate of the Probability Density of a Multidimensional Random Variable // *Measurement Techniques*. – 2020. – Vol. 62. No. 11. – P. 945-952. doi: 10.1007/s11018-020-01717-z.
10. Lapko A.V., Lapko V.A. Estimating the integral of the square of derivatives of symmetric probability densities of one-dimensional random variables // *Measurement Techniques*. – 2020. – Vol. 63, No. 3. – P. 171-176. doi: 10.1007/s11018-020-01768-2.
11. Lapko A.V., Lapko V.A. Integral Estimate from the Square of the Probability Density for a One-Dimensional Random Variable // *Measurement Techniques*. – 2020. – Vol. 63, No. 7. – P. 534-542. doi: 10.1007/s11018-020-01820-1.
12. Лапко А.В., Лапко В.А., Бахтина А.В. Сравнение линейных коллективов многомерных непараметрических регрессий // *Информатика и системы управления*. – 2020. – №1(63). – С. 84-94. doi: 10.22250/isu.2020.63.84-94.
13. Лапко А.В., Лапко В.А. Непараметрический алгоритм распознавания образов в задаче проверки гипотезы о тождественности законов распределения многомерных случайных величин в условиях статистических данных большого объема // *Информатика и системы управления*. – 2020. – №2(64). – С. 103-112. doi: 10.22250/isu.2020.64.103-112.
14. Sadovsky M., Senashova M., Malyshev A. Amazing symmetrical clustering in chloroplast genomes // *BMC Bioinformatics*. – 2020. – Vol. 21 (Suppl. 2). – P.1-14. doi: 10.1186/s12859-020-3350-z.
15. Gorban I. K., Senashova M. Yu., Sadovsky M. G. Towards the structuredness of non-coding regions of chloroplast genomes // *J. Phys.: Conf. Ser.* – 2020. – Vol. 1679. – 042004. doi:10.1088/1742-6596/1679/4/042004.
16. Putintseva Yu., Bondar E. , Simonov E., Sharov V., Oreshkova N., Kuzmin D., Konstantinov Yu., Shmakov V., Belkov V., Keech O., Sadovsky M. Siberian larch (*Larix sibirica* Ledeb.) mitochondrial genome assembled using both short and long nucleotide sequence reads is currently the largest known mitogenome // *BMC genomics*. – 2020. – 21(1):1 – P.12. doi:10.1186/s12864-020-07061-4.
17. Guseva T., Biriukov V., Sadovsky M. Role of Homeobox Genes in the Development of *Pinus sylvestris* // *LNBI*. 2020. – Vol.12108. – P. 429-437. doi: 10.1007/978-3-030-45385-5_38.
18. Molyavko A., Shaidurov V., Karepova E., Sadovsky M. Highly Parallel Convolution Method to Compare DNA Sequences with Enforced In/Del and Mutation Tolerance. *Bioinformatics and Biomedical Engineering (IWBBIO 2020)* // *Lecture Notes in Computer Science*. – Vol. 12108. – P. 472-481. doi/10.1007/978-3-030-45385-5_42.

19. Fedotovskaya V., Putintseva Yu., Shpagina T., Kolesnikova A., Sadovsky M. Function vs. Taxonomy: Further Reading from Fungal Mitochondrial ATP Synthases // LNBI. – 2020. – Vol. 12108. – P. 438-444. doi/10.1007/978-3-030-45385-5_39.
20. Roslavytceva V., Bushmelev E., Astanin P., Zabrodskaya T., Salmina A., Prokopenko S., Laptchenkova V., Sadovsky M. Blood Plasma Trophic Growth Factors Predict the Outcome in Patients with Acute Ischemic Stroke // LNBI. – 2020. – Vol. 12108. – P. 27-39. doi:10.1007/978-3-030-45385-5_3.
21. Арутюнян А.Г., Туник М.Е., Абрамов В.Г., Туценко К.О., Хорошавина А.А., Похабов Д.В., Садовский М.Г. Новые подходы к лечению и профилактике спастичности у больных в остром периоде очаговых поражений головного мозга с использованием ботулинического токсина типа А // Сибирское медицинское обозрение. – 2020. – № 4. – С. 84-91. doi: 10.20333/2500136-2020-4-84-91.
22. Похабов Д.Д., Туник М.Е., Абрамов В.Г., Туценко К.О., Хорошавина А.А., Похабов Д.В., Садовский М.Г. Определение пороговой чувствительности обонятельного восприятия у больных экстрапирамидными заболеваниями // Сибирское медицинское обозрение. – 2020. – № 2. – С. 58-66. doi: 10.20333/2500136-2020-2-58-66.
23. Хорошавина А.А., Туценко К.О., Абрамов В.Г., Туник М.Е., Похабов Д.В., Садовский М.Г. Novel smart data analysis technologies to differentiate Parkinson's disease from essential tremor // IOP Conference series. – 2020, – Vol. 1679. – 042005. doi:10.1088/1742-6596/1679/4/042005.
24. Туценко К.О., Хорошавина А.А., Абрамов В.Г., Карлова Е.А., Садовский М.Г. Diagnostics of Parkinson's disease by positron-emission tomography with 18F-DOPA: new approach based on smart clustering // IOP Conference series. – 2020. – Vol. 1679. – 042006. doi:10.1088/1742-6596/1679/4/042006.
25. Феллер А.А., Мартынова Е.А., Садовский М.Г., Черняев Д. В., Семёнов Э. В., Слепов Е.В., Зуков Р.А. CBC effectively stratifies the patients with different types of malignant tumors // IOP Conference series. – 2020, – Vol. 1679. – 042019. doi:10.1088/1742-6596/1679/4/042019.
26. Горбунова Е.А., Медведева Н.Н., Зуков Р.А., Садовский М.Г. Advanced nonlinear statistics reveals the relations between anthropometry and bioimpedance of physique in post-surgery complications of the patients with gastric cancer // IOP Conference series. – 2020. – Vol. 1679. – 042018. doi:10.1088/1742-6596/1679/4/042018.