Министерство науки и высшего образования Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ НАУЧНОЕ УЧРЕЖДЕНИЕ «ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР «КРАСНОЯРСКИЙ НАУЧНЫЙ ЦЕНТР СИБИРСКОГО ОТДЕЛЕНИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК»

Институт вычислительного моделирования
СО РАН – обособленное подразделение ФИЦ КНЦ СО РАН

ГРНТИ 658.512

№ АААА-А18-118011890021-4

УТВЕРЖДАЮ
Врио директора ФИЦ КНЦ СО РАН

_____ А.А. Шпедт
«__» _____ 2020 г.

ОТЧЕТ
О ВЫПОЛНЕНИИ ПРОЕКТА
«МЕТОДЫ И ТЕХНОЛОГИИ АНАЛИТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ И ПОСТРОЕНИЯ ПРОГРАММНО-ТЕХНИЧЕСКИХ КОМПЛЕКСОВ И ИНТЕГРИРОВАННЫХ СИСТЕМ»

(промежуточный)

Номер проекта в плане НИР: 0356-2019-0016

Приоритетное направление: Информационно-телекоммуникационные системы
Программа ФНИ (номер и наименование): IV.35 Когнитивные системы и технологии, нейроинформатика и биоинформатика, системный анализ, искусственный интеллект, системы распознавания образов, принятие решений при многих критериях

Протокол Ученого совета _____

№______ от «__» _____ 2020 г.

Руководитель проекта _____ Л.Ф. Ноженкова
д.т.н., профессор "___" _____ 2020 г.

Красноярск, 2020

СПИСОК ИСПОЛНИТЕЛЕЙ

Руководитель темы д.т.н., профессор	_____ (*подпись, дата*)	Л.Ф. Ноженкова (Введение, Раздел 1, Заключение)
Исполнители: зам. дир., к.т.н.	_____ (*подпись, дата*)	С.В. Исаев (Раздел 1
г.н.с., д.т.н.	_____ (*подпись, дата*)	А.В. Лапко (Раздел 2)
в.н.с., д.т.н.	_____ (*подпись, дата*)	В.А. Лапко (Раздел 2)
в.н.с., д.ф.-м.н.	_____ (*подпись, дата*)	М.Г. Садовский (Раздел 3)
с.н.с., к.т.н.	_____ (*подпись, дата*)	О.С. Исаева (Раздел 1)
с.н.с., к.т.н.	_____ (*подпись, дата*)	А.В. Коробко (Раздел 1)
с.н.с., к.т.н.	_____ (*подпись, дата*)	В.В. Ничепорчук (Раздел 1)
с.н.с., к.т.н.	_____ (*подпись, дата*)	Т.Г. Пенькова (Раздел 1)
с.н.с., к.ф.-м.н.	_____ (*подпись, дата*)	М.Ю. Сенашова (Раздел 3)
н.с., к.т.н.	_____ (*подпись, дата*)	Д.В. Жучков (Раздел 1)
н.с., к.т.н.	_____ (*подпись, дата*)	Е.В. Ковязина (Раздел 1)
н.с., к.т.н.	_____ (*подпись, дата*)	В.В. Морозов (Раздел 1)
н.с., к.т.н.	_____ (*подпись, дата*)	А.И. Ноженков (Раздел 1)
н.с.	_____ (*подпись, дата*)	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 54 с., 27 рис., 2 прил.

ИНТЕЛЛЕКТУАЛЬНАЯ АНАЛИТИЧЕСКАЯ ОБРАБОТКА МНОГОМЕРНЫХ ДАННЫХ, 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 интеграции и интеллектуальной аналитической обработки больших объемов данных, развитие и реализация новых технологических подходов к созданию сложных программно-технических комплексов и интегрированных систем разного назначения.

Разработаны и программно реализованы новые методы интеллектуализации аналитической обработки данных в интегрированных системах территориального и корпоративного управления. Предложен метод построения многомерной аналитической модели состояния сложных объектов и систем на основе выявления и формального описания закономерностей. Разработаны алгоритмические и программные средства построения модельно-ориентированных систем на базе интерактивной веб-платформы. Разработаны алгоритмы формирования унифицированного представления структуры межсистемного обмена данными с учетом вариативности форматов. Выполнено развитие веб-технологии функционирования архивов открытого доступа. Проведен анализ актуальных угроз кибербезопасности корпоративной сети, разработаны новые формальные модели безопасности и методы защиты.

Разработаны новые технологические принципы поддержки конструирования программно-аппаратных комплексов бортовой аппаратуры космических аппаратов. Построена гетерогенная модель, объединяющая базы знаний, описывающие программные имитаторы функционирования бортовых систем и виртуальные приборы, имитирующие устройства приёма-передачи данных. Для поддержки проведения наземных испытаний модель реализует функции бортовой аппаратуры при выполнении командно-программного управления космическим аппаратом, представляя методы приема-передачи широкого спектра команд и формирования значений телеметрического кадра. Разработан метод анализа результатов испытаний бортовой аппаратуры космического аппарата по прецедентам гетерогенной имитационной модели. Метод позволяет выполнять

исследование функционирования бортовой аппаратуры на модели, что значительно повышает эффективность процессов изготовления оборудования.

Разработаны и исследованы быстрые алгоритмы оптимизации непараметрических решающих правил ядерного типа в условиях статистических данных большого объёма. Обоснован и разработан быстрый алгоритм выбора оптимальных коэффициентов размытости ядерных функций в непараметрической оценке многомерной плотности вероятности типа Розенблатта-Парзена, разработан быстрый алгоритм оптимизации непараметрической оценки уравнения разделяющей поверхности ядерного типа в двухальтернативной задаче распознавания образов, предложена методика оптимальной дискретизации области определения многомерной плотности вероятности. Полученные результаты реализованы в программном комплексе автоматической классификации данных дистанционного зондирования природных объектов.

Выполнено исследование структурированности генетических данных больших и сверх-больших объёмов. Изучены статистические свойства некодирующих областей хлоропластных геномов растений с использованием кластеризации частотных словарей триплетов отдельных фрагментов генома, определяемых регулярным порядком. На основании одновременной кластеризации наборов некодирующих областей хлоропластов выявлен исходный тип структуры, из которой в процессе эволюции редуцировались остальные типы структур.

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

Курс на развитие перспективных «сквозных» цифровых технологий в России ставит новые цели и 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. В России, как и во всем мире, переход от этапа «Индустрия 4» к этапу «Индустрия 5» и «Интеллектуальному обществу 5.0» сопровождается взрывным ростом объёмов данных и дефицитом 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. В ИВМ СО РАН СО РАН на протяжении многих лет ведется мониторинг и анализ киберугроз корпоративной сети Красноярского научного центра, на основе которых предлагаются новые методы повышения эффективности кибербезопасности и модели доступа к информации в интернет-системах.

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 соответствующим 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 некодирующие области геномов, поскольку обнаруживаются ранее неизвестные функции этих областей. Анализ внутренней структурированности некодирующих 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:

1. Разработка и реализация методов интеллектуализации аналитической обработки данных в интегрированных системах территориального и корпоративного управления.
2. Разработка методов поддержки конструирования программно-аппаратных комплексов бортовой аппаратуры космических аппаратов.
3. Разработка и апробация средств анализа и оценки угроз кибербезопасности в корпоративной сети.
4. Разработка и исследование быстрых алгоритмов оптимизации непараметрических решающих правил ядерного типа в условиях статистических данных большого объёма.
5. Изучение структурированности генетических данных больших и сверх-больших объёмов.

Настоящий отчет представляет результаты работ за 2019 год 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 д.т.н. Ноженкова Л.Ф.

1.1 Разработка и реализация методов интеллектуализации аналитической обработки данных в интегрированных системах территориального и корпоративного управления

1.1.1 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 отдельных элементов в различных режимах и условиях работы в отдельные моменты времени. Выполнено построение аналитической модели состояния гидроэнергетической системы (гидроагрегата) на основе применения метода анализа главных компонент и кластерного анализа к данным системы вибрационного контроля. Комплексный интеллектуальный анализ мониторинговых данных позволил определить особенности в работе гидроагрегата, обнаружить структуру и закономерности в данных, выявить признаки взаимного влияния его конструкционных узлов, определить соотношения диапазонов значений ключевых технических параметров в различных режимах функционирования оборудования. На рис. 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 (с 07:00 25.06.2015 по 23:00 15.09.2015), 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.

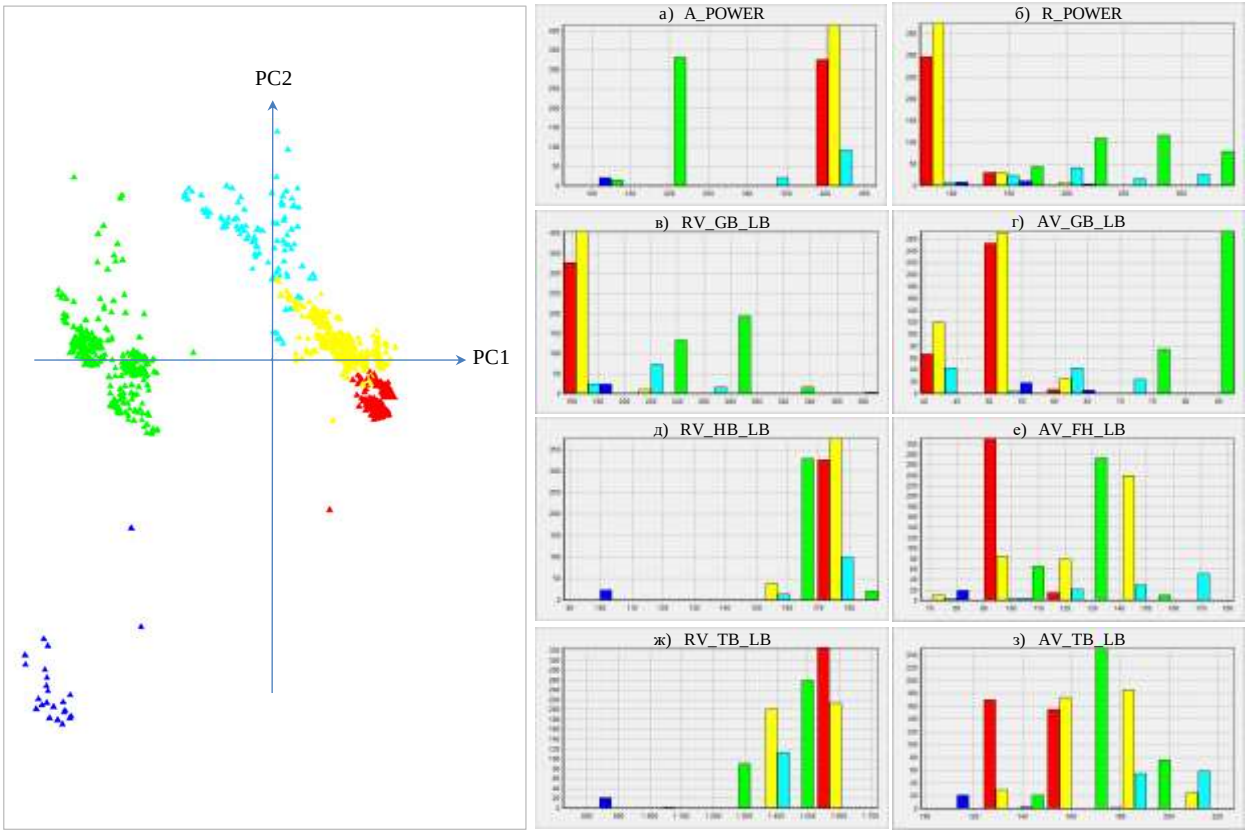


Рисунок 1.1 – Пятикластерная структура данных

С целью выявления характерного поведения системы в определенных режимах функционирования гидроагрегата выполнена кластеризация по диапазону значений контролируемых параметров. С целью исследования взаимного влияния параметров выполнена детализация рассматриваемых моментов времени до минут и построены диаграммы сопоставления временных рядов контролируемых параметров для характерных отрезков времени. На рис. 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 своих высоких значений, ~450МВт) ведет к небольшому снижению относительной вибрации турбинного подшипника (RV_TB_LB), но при этом к существенному росту значений абсолютной вибрации опоры подпятника (AV_FH_LB) (≈50%) и абсолютной вибрации генераторного подшипника (AV_GB_LB) (≈30%). Такое поведение системы наблюдалось преимущественно в дневные часы, непрерывно до 16 часов подряд, в конце июня – начале июля.

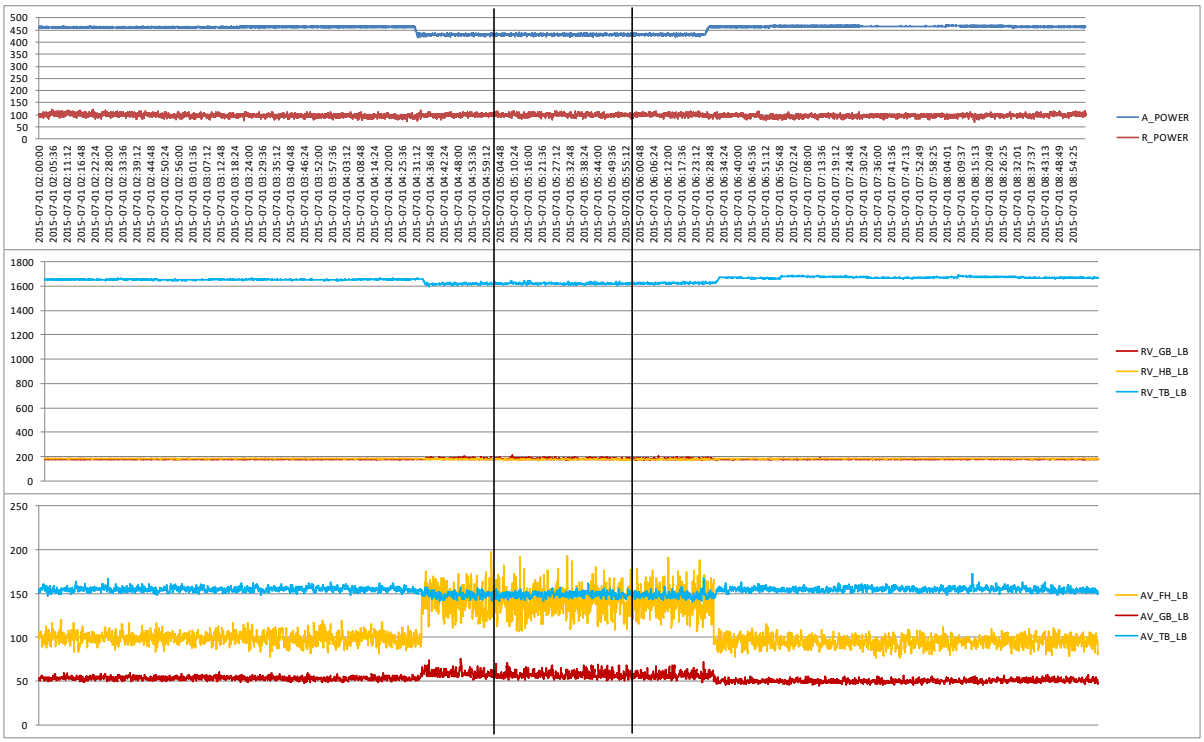


Рисунок 1.2 – Сопоставление временных рядов контролируемых параметров в моменты времени с высокими значениями относительной вибрации турбинного подшипника

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 подпятника 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. Выполненный детальный анализ характерных моментов времени позволил выявить определенные зависимости значений для ключевых контролируемых параметров. Например, незначительное изменение амплитуды активной мощности на верхней границе средних значений (~200 МВт) ведет к существенному увеличению относительной вибрации генераторного подшипника (≈60%), абсолютной вибрации опоры подпятника (≈45%) и абсолютной вибрации турбинного подшипника (≈17%); небольшое уменьшение активной мощности в диапазоне своих высоких значений

(~450МВт) ведет к небольшому снижению относительной вибрации турбинного подшипника (в диапазоне своих максимальных значений, ~1600 мкм) и к существенному увеличению абсолютной вибрации опоры подпятника (≈50%) и генераторного подшипника (≈30%). Выявленные зависимости между параметрами, диапазоны и соотношения их значений в различных режимах эксплуатации являются уникальными характеристиками и отличительными свойствами конкретного гидроагрегата. Изменение технического состояния гидроагрегата с течением времени проявляется в изменении структуры и закономерностей в данных, представленных аналитической моделью. Сравнение аналитических моделей за разные периоды времени позволяет отслеживать динамику изменения износа оборудования и состояния системы в целом.

1.1.2 Предложен метод, разработаны алгоритмы формирования унифицированного представления структуры межсистемного обмена данными с учетом вариативности форматов на основе ранее предложенной унифицированной модели. Согласно предложенному методу процесс обмена данными между системами можно представить в виде: $M = (\langle O_{In}, I_{In}\rangle, \langle O_{Out}, I_{Out}\rangle)$, где I – системы информационного обмена, O – объекты информационного обмена. Каждая система может участвовать в информационном обмене в качестве «отправителя» (индекс *In*) или «получателя» (индекс *Out*). Одна и та же система может выступать одновременно в качестве «отправителя» и «получателя», например, когда необходимо перевести данные из одного формата в другой в рамках одной информационной системы. Система информационного обмена характеризуется рядом параметров, описывающих ее особенности: адрес, канал связи, способы хранения данных, параметры доступа к данным и т.д. Объект информационного обмена O можно представить как тройку: $\mathrm{O} = \langle \mathrm{D}, \mathrm{F}, \mathrm{S}\rangle$, где D – передаваемые данные, F – формат хранения данных, S – структура, в которой представлены данные. Процесс информационного обмена на основе унифицированного представления структуры данных можно представить цепочкой: $O_{In}\langle D_{In}, F_{In}, S_{In}\rangle \xrightarrow{Q\langle F_{In}, S_{In}\rangle} P_{In}\langle U_{In}, S_{In}\rangle \xrightarrow{G\langle S_{In}, S_{Out}\rangle} P_{Out}\langle U_{Out}, S_{Out}\rangle \xrightarrow{\bar{Q}\langle F_{Out}, S_{Out}\rangle} O_{Out}\langle D_{Out}, F_{Out}, S_{Out}\rangle$, где Q – оператор унификации, выполняющий преобразование данных «отправителя» в информационный пакет P_{In}, содержащий исходные данные в унифицированном виде, G – оператор, выполняющий преобразование входной структуры данных в выходную с получением унифицированного пакета P_{Out}, имеющего структуру данных «получателя», $\bar{Q}$ – оператор деунификации, выполняющий преобразование унифицированных данных в формат данных «получателя».

На рис. 1.3 представлена метамодель организации унифицированного информационного обмена в виде диаграммы классов, описывающая логику, основные сущности и отношения между ними.

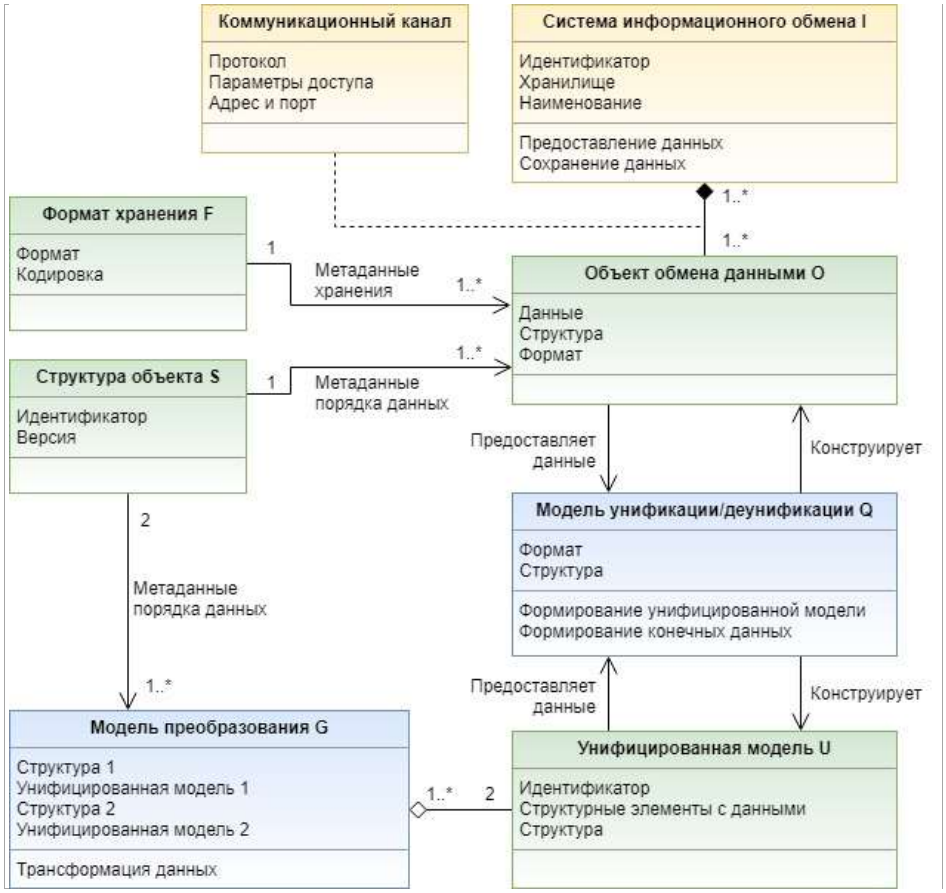


Рисунок 1.3 – Метамодель унифицированного информационного обмена данными

На рисунке объект информационного обмена *O* представлен классом «*Объект информационного обмена*». Это входные данные, поступающие в межсистемный информационный обмен от систем «отправителей» и выходные данные, поступающие к системам «получателям». Основными атрибутами класса являются данные, формат хранения (*F*) и структура (*S*). Формат хранения и структура данных являются ключевыми метаданными о способе обработки входящей или исходящей информации. Формат данных (XML, JSON, RDB, CSV, DOCX и др.) определяет способ чтения или записи данных. Структура данных определяет состав и порядок расположения отдельных информационных элементов. Унифицированная модель *U* представлена классом «*Унифицированная модель*». В данной модели информация представлена в виде отдельных элементов, упорядоченных по иерархическому принципу. Унифицированная модель не имеет формата хранения, атомарность элементов обеспечивает возможность преобразования одной структуры данных в другую. Модель унификации *Q* представлена

классом «Модель унификации /деунификации». Основная функция модели унификации – преобразование информационного объекта заданного формата хранения в унифицированную модель и обратно. Здесь выполняются операции чтения исходных данных, построение и наполнение структуры в терминах унифицированной модели, а также обратные операции чтения и компоновки преобразованных данных в конкретный формат хранения для передачи их системе «получателю». Модель преобразования *G* представлена классом «Модель преобразования». Модель преобразования обеспечивает преобразование одной унифицированной модели в другую на основе взаимного сопоставления элементов входных и выходных данных.

В рамках предложенного метода разработаны алгоритмы унификации и деунификации данных и алгоритмы преобразования унифицированных моделей. Разработанные 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 задач организации закупок, где осуществляется обмен данными между корпоративной системой размещения заказа, единой информационной системой и электронными торговыми площадками.

1.1.3 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.

Гетерогенность информации, доступной для анализа, в первую очередь заключается в разнообразии форматов и схем хранения данных. Форматы данных определяют основную логику организации информации, допустимые виды объектов и отношения между ними на концептуальном уровне. И даже в рамках одного формата информация может быть представлена разными схемами хранения, правилами наименования объектов и уровнями агрегации.

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. 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. 1.4).

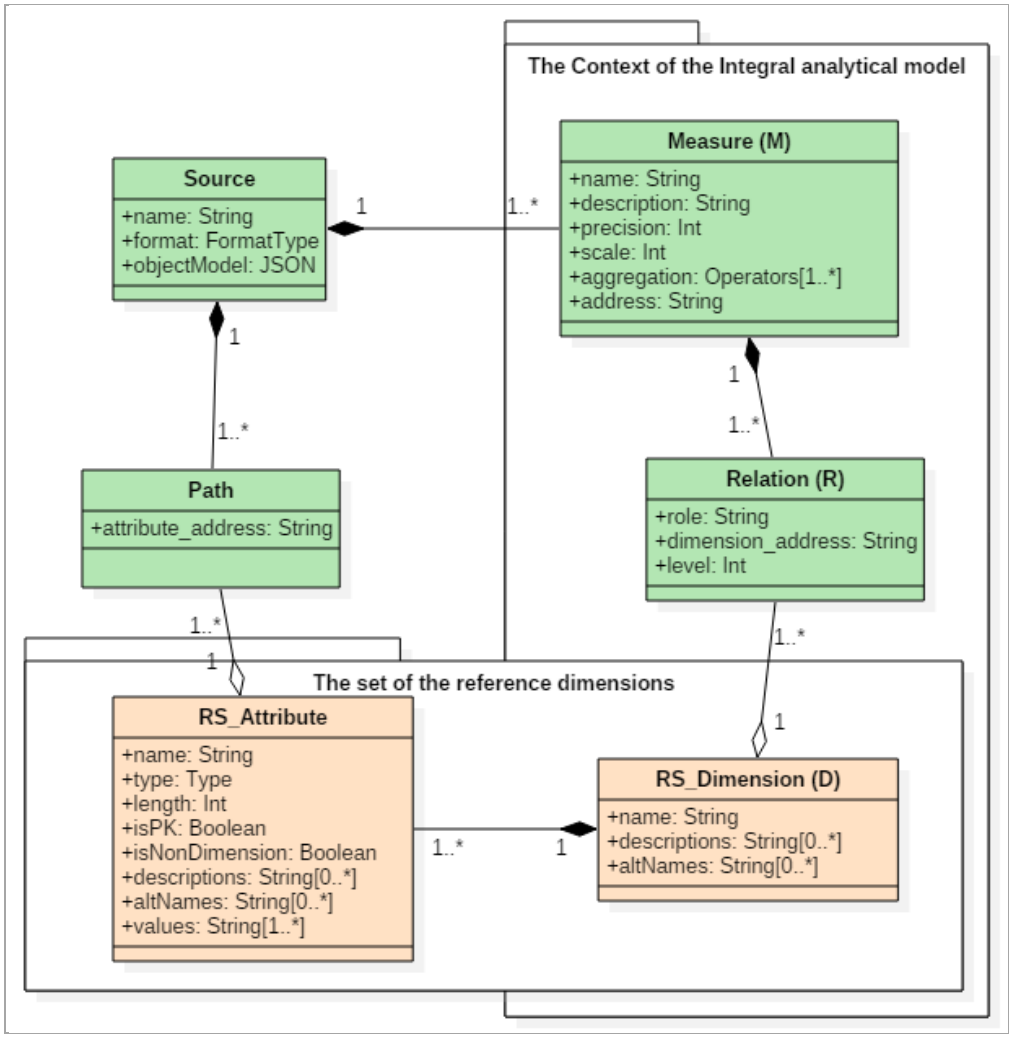


Рисунок 1.4 – Метамодель интегральной аналитической платформы

Концептуальная 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 (рис. 1.5).

```
import pandas as pd
import psycopg2 as pg
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity


""" Calculate the cosine distance between two text arrays as a matrix"""
def cosine_distance(text1, text2):...


if __name__ == '__main__':
...

    connection = pg.connect("host='...' dbname=omdb user=omdb password='omdb1209'")
    sql1 = """..."""
    sql2 = """..."""
    """Get the set of dimensions with their descriptions"""
    df_descr = pd.read_sql_query(sql1, con=connection)
    """Get the set of dimensions with their alternative names"""
    df_alt_n = pd.read_sql_query(sql2, con=connection)

    """Divide the sets according to data sources"""
    df_descr_rel = df_descr[df_descr.source == 'Client:8081']
    df_descr_xml = df_descr[df_descr.source == 'XSD client']
    df_names_rel = df_descr_rel['dim'].drop_duplicates()
    df_names_xml = df_descr_xml['dim'].drop_duplicates()
    df_alt_n_rel = df_alt_n[df_alt_n.sour == 'Client:8081']
    df_alt_n_xml = df_alt_n[df_alt_n.sour == 'XSD client']

    """Compare names of dimensions"""
    new_db = cosine_distance(df_descr_rel['dim'], df_descr_xml['dim'])
    for i in range(0, len(df_names_rel)):
        for j in range(0, len(df_names_xml)):
            """Compare arrays of alternative names for each pair of dimensions"""
            text1 = df_alt_n_rel[df_alt_n_rel.dim == df_names_rel.iat[i]]['alt_name']
            text2 = df_alt_n_xml[df_alt_n_xml.dim == df_names_xml.iat[j]]['alt_name']
            if text1.count() > 0 and text2.count() > 0:
                alt_n_val = cosine_distance(text1, text2).max().max()
                """Add the maximal similarity score of alternative names for pair of dimensions i, j"""
                new_db.iat[i, j] = max(alt_n_val, new_db.iat[i, j])
            text1 = df_descr_rel[df_descr_rel.dim == df_names_rel.iat[i]]['description']
            text2 = df_descr_xml[df_descr_xml.dim == df_names_xml.iat[j]]['description']
            if text1.count() > 0 and text2.count() > 0:
                discr_val = cosine_distance(text1, text2).max().max()
                """Add the maximal similarity score of descriptions for pair of dimensions i, j"""
                new_db.iat[i, j] = max(discr_val, new_db.iat[i, j])
```

Рисунок 1.5 – Алгоритм сопоставления разнородных измерений на языке Python

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.

1.1.4 Разработаны алгоритмические и программные средства построения прикладных систем на базе интерактивной веб-среды. Создана веб-платформа, предназначенная для автоматизации процесса создания модельно-ориентированных систем сбора данных. Помимо типовых функций модельно-ориентированной системы по ведению, хранению и аналитической обработке данных, платформа обеспечивает возможность динамического расширения тематического наполнения построенных систем на основе моделей, описывающих предметную область (прикладная и управляющая модели). Для каждой построенной системы платформа реализует разделение прав доступа пользователей на уровне ролей: оператора, аналитика и модератора. Роль оператора открывает доступ к интерфейсу ввода данных. Роль аналитика позволяет выполнять анализ собранных данных, строить различные графики, выводить информацию на карту, а также экспортировать данные в Excel. Роль модератора позволяет изменять модель предметной области, тем самым изменяя интерфейс и поведение системы.

Создание прикладной системы на базе платформы включает следующие основные этапы: регистрация системы, заключающаяся в создании базы данных и поддомена; регистрация пользователей и назначение прав доступа; формирование структурных элементов системы путем построения и модификации прикладной и управляющей моделей с помощью интерфейса системы (рис. 1.6); наполнение системы с помощью интерфейса сбора данных (рис. 1.7); формирование аналитических отчётов.

Создаваемые на базе платформы прикладные системы имеют необходимый набор инструментов для анализа данных и построения отчётов. Для возможности анализа данных с помощью сторонних инструментов реализован REST-интерфейс передачи отчётов. Посредством данного интерфейса актуальные данные каждого отчёта в формате JSON могут быть переданы сторонним аналитическим средствам для обработки.

Выполнена апробация разработанных алгоритмических и программных средств веб-платформы для задач оценки экологического состояния почв населенных пунктов Красноярского края и анализа проб техногенно-поверхностных образований вблизи промышленных предприятий города Красноярска.

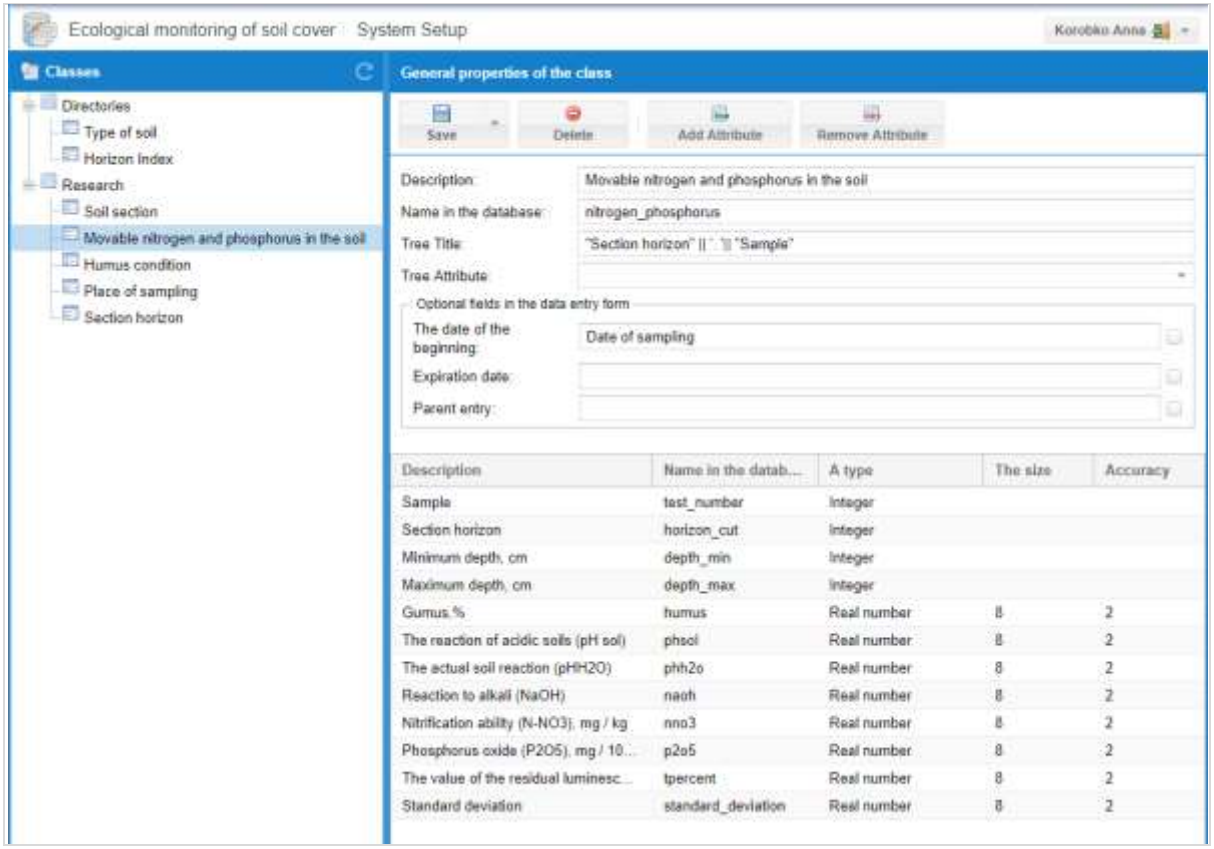


Рисунок 1.6 – Редактор классов на примере задачи экологического мониторинга почв

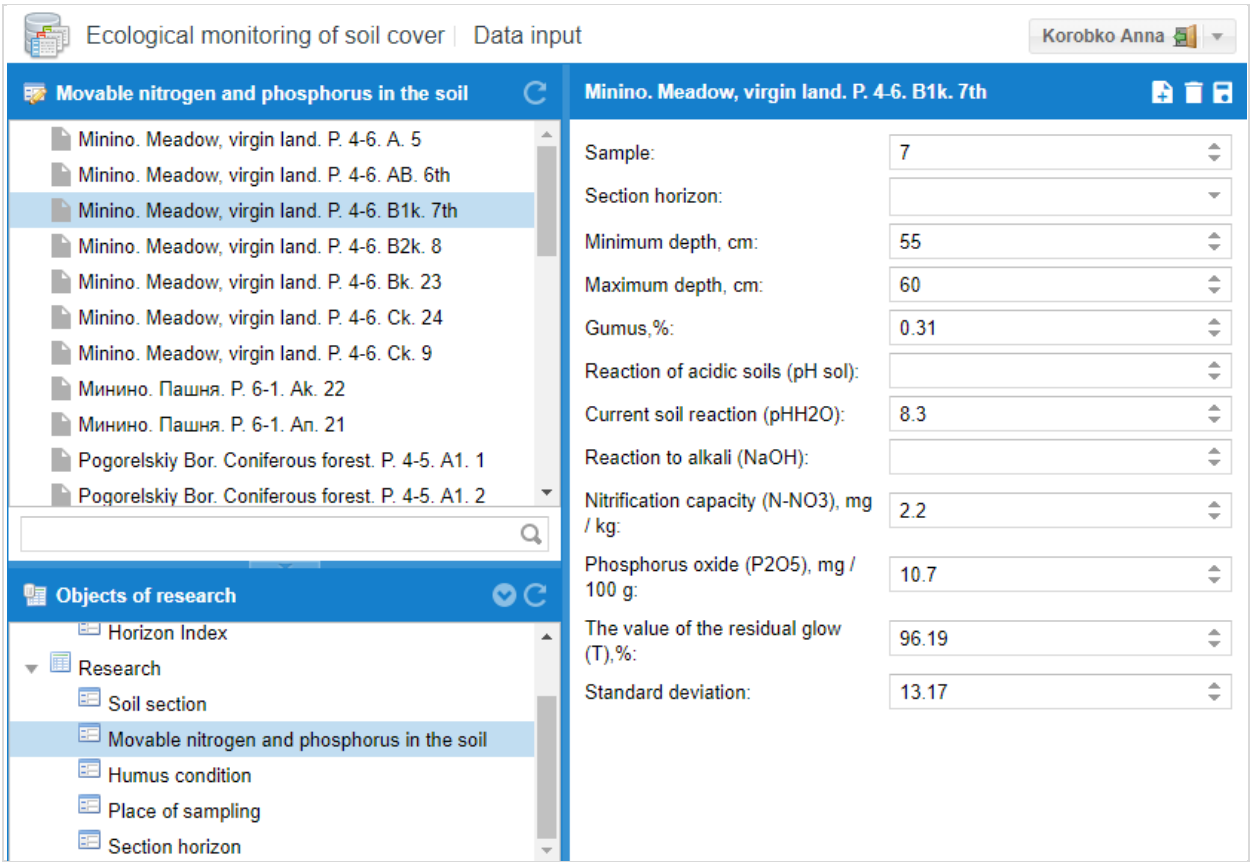


Рисунок 1.7 – Форма ввода данных на примере задачи экологического мониторинга почв

1.1.5 Разработана Технология ситуационного моделирования опасных событий, обеспечивающая информационную поддержку управления безопасностью территорий и позволяющая решать задачи предупреждения и ликвидации различных видов опасных событий. Технология основана на ранее предложенной системной модели построения информационно-аналитических систем поддержки управления природно-техногенной безопасностью, предусматривающей совместное использование расчётных методик оценки последствий опасных ситуаций различной сложности, метода динамического картографирования, экспертных систем и веб-технологий. Процесс ситуационного моделирования представлен на рис. 1.8.

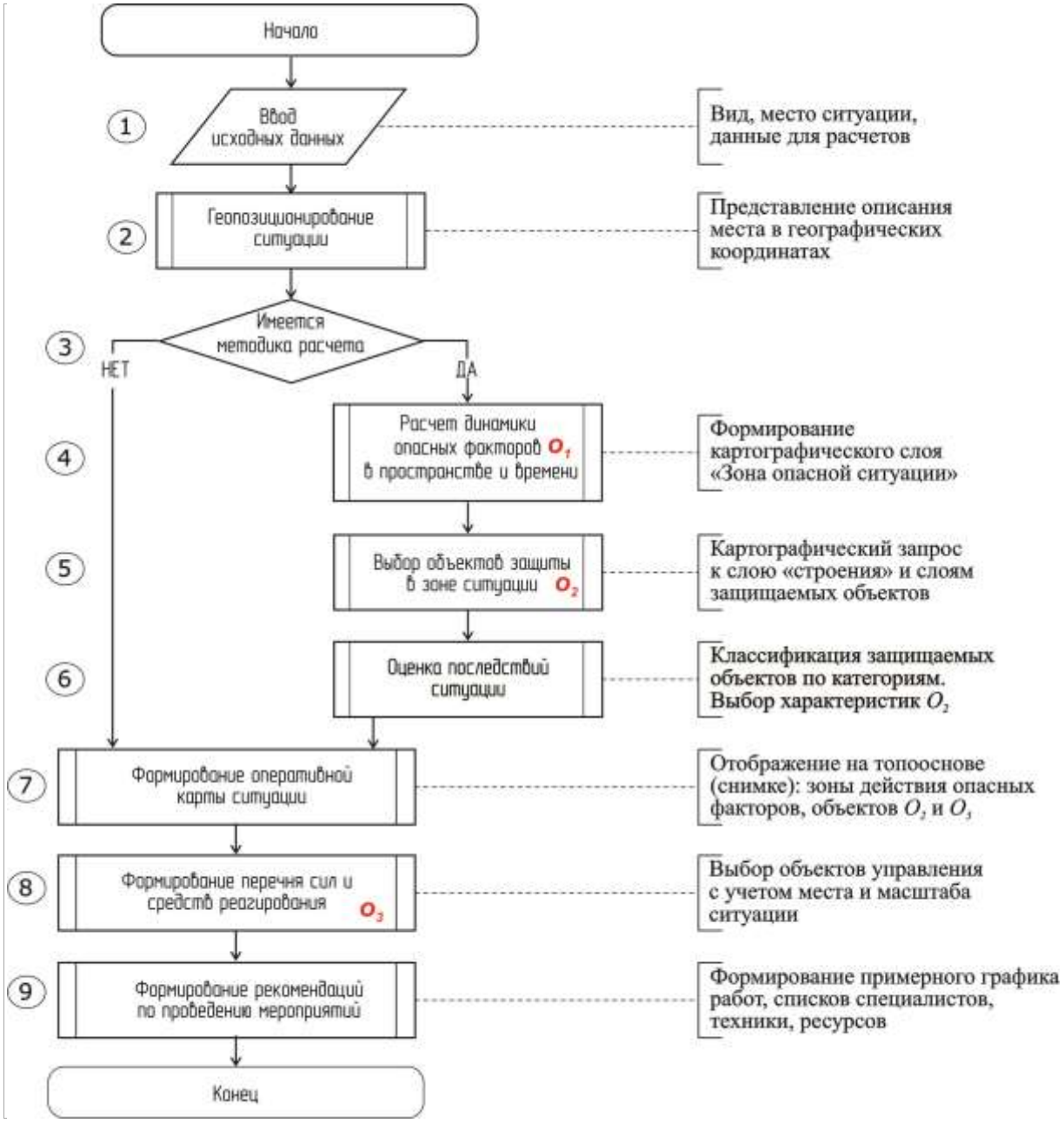


Рисунок 1.8 – Схема ситуационного моделирования

Процесс моделирования инициируется вызовом сценария (1), соответствующего виду опасного события. Геопозиционирование (2) позволяет оценить удаленность формирований, время их реагирования и вероятность эскалации ситуации. Проверка наличия методик расчёта (3) определяет три основных направления моделирования (4-6): фоновая оценка масштабов опасных событий на основе упрощённых методик; реконструкция и создание новых сценариев опасных событий на основе уточнённых методик; пространственный анализ опасных событий на протяженных объектах. Формирование оперативных карт (7) происходит с использованием ГИС, позволяющей интегрировать топографические основы или космические снимки из веб-ресурсов. По результатам анализа сложившейся ситуации определяются необходимые силы и средства (8-9) с детализацией определенных специалистов, техники и ресурсов. Применение динамически настраиваемых таблиц, карт и графиков на основе веб-технологий, в отличие от традиционных отчетных форм, позволяет адаптировать формируемые решения под конкретную ситуацию и предпочтения лиц, принимающих решения, исключить избыточность информации, замедляющую выработку системы неотложных мероприятий.

На основе гибридного подхода, сочетающего преимущества технологий разработки нативных и веб- приложений, реализован комплекс многофункциональных сервисов безопасности жизнедеятельности: сервис безопасности туристической деятельности, предназначенный для регистрации туристических групп и контроля прохождения ими заданных маршрутов с возможностью получения экстренных оповещений; информационная среда для старост населённых пунктов, предназначенная для оперативного мониторинга обстановки и оповещения об опасных явлениях; сервис «Общественный контроль», предназначенный для оперативной фиксации нарушений и угроз опасных событий с автоматической передачей сведений оперативным службам экстренного реагирования. На рис. 1.9 представлен пример отображения данных оперативного мониторинга в информационной среде старост населённых пунктов.

Возможность получения мобильными приложениями оповещений, прогнозов, а также данных оперативного мониторинга природных и техногенных опасностей реализована за счет интеграции с автоматизированной системой мониторинга и прогнозирования чрезвычайных ситуаций, функционирующей в Территориальном центре мониторинга и прогнозирования чрезвычайных ситуаций Красноярского края.

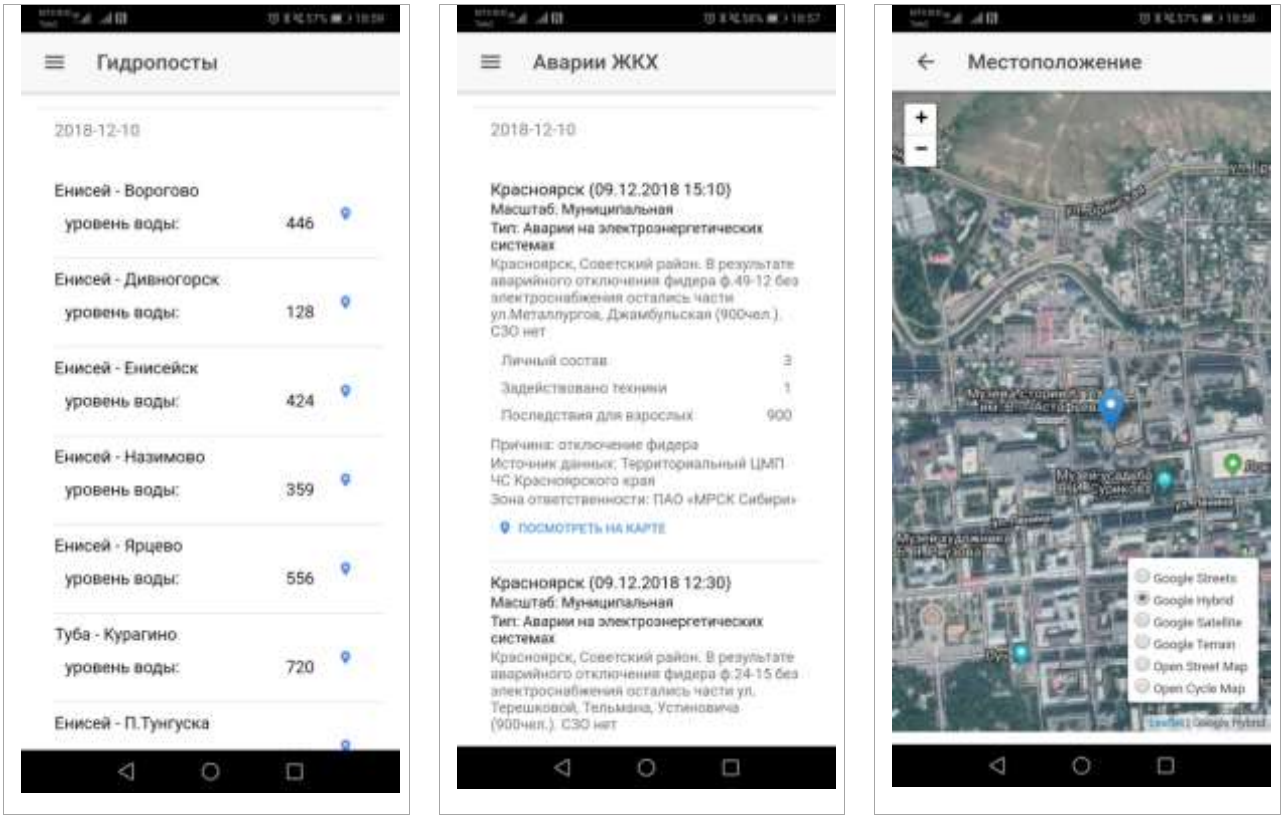


Рисунок 1.9 – Отображение данных оперативного мониторинга
в информационной среде старост населённых пунктов

1.1.6 Выполнено развитие ранее предложенного метода интегрального аналитического оценивания природно-техногенной безопасности территорий, обеспечивающего формирование комплексного показателя на основе результатов многомерного аналитического моделирования состояния окружающей среды и объектов техносферы. Для интерпретации количественных значений оценок предложен подход к формированию оценочной шкалы, которая обеспечивает отображение количественных значений в качественное выражение с использованием аппарата нечетких множеств. Каждому значению оценки ставится в соответствие значение лингвистической переменной «Уровень природно-техногенной безопасности», на основе функций принадлежности $\mu_j(x)$, где x – значение количественной оценки показателя, j – номер, соответствующий уровням безопасности: «Низкий», «Пониженный», «Удовлетворительный», «Приемлемый», «Улучшенный» и т.д. Согласно предложенному подходу, функция принадлежности определяется с помощью метода нечеткой кластеризации (fuzzy c-means clustering), который позволяет определить функции принадлежности по распределению значений интегральных оценок, проводя сопоставление нечеткого кластера нечеткому множеству. При этом функция

принадлежности нечеткого кластера соответствует функции принадлежности нечеткого множества.

С использованием предложенного подхода сформированы оценочные шкалы для интерпретации оценок состояния природно-техногенной безопасности территорий Красноярского края. В таблице 1.1 представлен пример оценочной шкалы для комплексного показателя «Природно-техногенная безопасность» для г. Красноярска.

Таблица 1.1 – Пример оценочной шкалы комплексного показателя «Природно-техногенная безопасность»

	Уровень природно-техногенной безопасности	Центр нечеткого кластера	Функция принадлежности (μ_j)
1	Улучшенный	0,9093267666288	0,986412206
2	Хороший	0,8136647929115	0,008926737
3	Приемлемый	0,7275432132811	0,002325299
4	Удовлетворительный	0,6561627709198	0,00114636
5	Пониженный	0,6003893396118	0,0
6	Низкий	0,4804927665189	0,0
7	Критический	0,1016071897181	0,0

На рис. 1.10 представлен пример результата оценивания уровня природно-техногенной безопасности Красноярского края в виде картограммы.

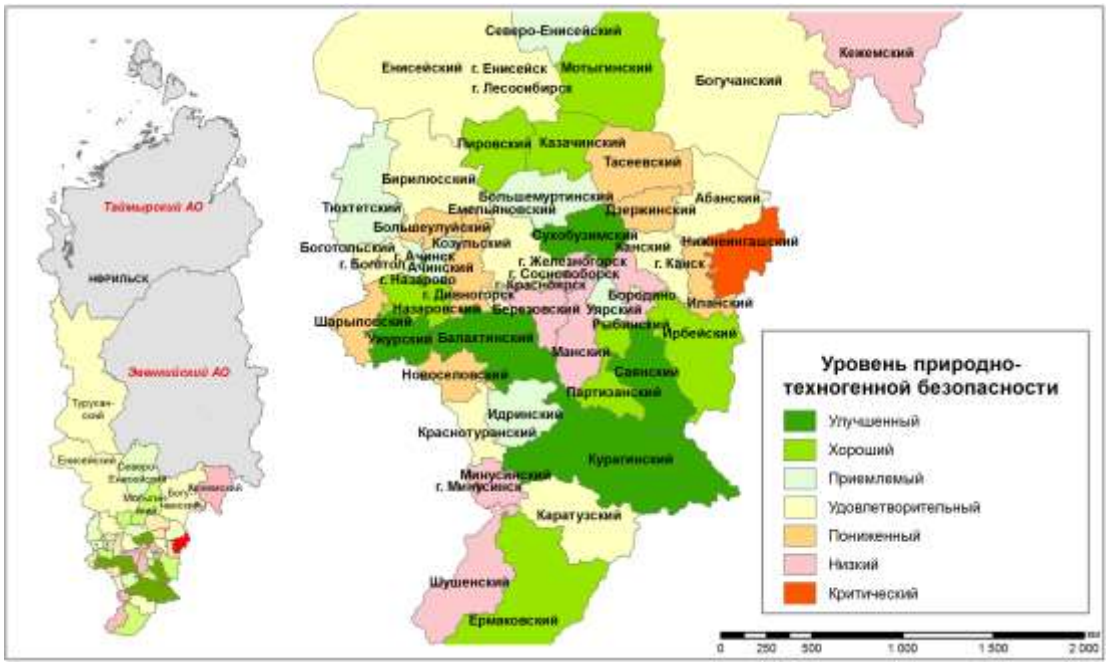


Рисунок 1.10 – Результат оценивания природно-техногенной безопасности территорий Красноярского края в виде картограммы

В рамках развития веб-технологии функционирования архивов открытого доступа и инфраструктуры открытой науки, обеспечения взаимопроникновения технологий научных цифровых библиотек (RDL), работающих на основе институциональных репозиториев (IR), информационных систем текущих научных исследований (CRIS) и центров обработки данных (DC), выполнены работы по совершенствованию схем метаданных публикаций, обеспечивающие минимальные потери информации при обмене данными открытого архива в системе автоматизации библиотек. Выполнены работы по оптимизации индексирования открытого архива поисковыми системами Интернет, онлайновой поддержке открытого лицензирования публикаций Creative Common (CC) и интеграции с системой Sherpa/Romeo, обеспечивающей соответствие доступа к ресурсам издательским требованиям. Проведены работы по миграции тестовой версии открытого архива ИВМ СО РАН в операционную систему Linux OpenSuse, что позволило обеспечить большую стабильность и устойчивость к сбоям программной платформы DSpace 6.3. Выгрузка метаданных системы автоматизации библиотек ИРБИС64+ в xml-формате схемы RUSMARC дополнена выгрузкой в базовой для DSpace схеме QDC. Такой способ миграции данных позволяет сформировать пакет объекта DSpace, включающий как метаданные, так и полный текст публикации. 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 службы проекта Sherpa/Romeo, позволяющей отслеживать соответствие доступа к публикациям издательским ограничениям, что особенно важно для публикаций в зарубежных журналах, отличающихся большим разнообразием правил доступа.

1.2 Разработка методов поддержки конструирования программно-аппаратных комплексов бортовой аппаратуры космических аппаратов. Методы анализа результатов испытаний бортовой аппаратуры на основе прецедентов имитационной модели

Разработаны новые технологические принципы поддержки конструирования программно-аппаратных комплексов бортовой аппаратуры космических аппаратов. Построена гетерогенная модель, объединяющая базы знаний, описывающие программные имитаторы функционирования бортовых систем и виртуальные приборы, имитирующие устройства приёма-передачи данных. Для поддержки проведения наземных испытаний модель реализует функции бортовой аппаратуры при выполнении командно-программного управления космическим аппаратом, представляя методы приема-передачи широкого спектра команд и формирования значений телеметрического кадра. Предложен метод анализа результатов испытаний бортовой аппаратуры космического аппарата, основанный на применении прецедентов имитационной модели.

Важным этапом подготовки и проведения испытаний является построение критериев оценки получаемых результатов. Такие критерии, как правило, задаются в виде наборов граничных условий или эталонных значений. Широкий спектр команд управления не позволяет в полной мере сформировать критерии для оценки всех особенностей работы объекта контроля. В настоящее время это трудоемкий, ручной и подверженный ошибкам процесс. Анализ соответствия испытываемых устройств техническим заданиям требует от конструктора бортовой аппаратуры высокой квалификации и широких знаний о различных особенностях их функционирования. В общем случае для решения этой задачи требуется исследовать функционал $Y=F^o(X)$, устанавливающий соответствие между входными переменными – X и выходными – Y. Как правило, его аналитическое представление построить затруднительно. Для анализа логики функционирования объекта контроля предложено заменить исследование данного функционала анализом результатов имитационного моделирования. Отправка команд, их получение, квитирование, отработка, изменение параметров приемо-передающего тракта, а также состояния бортовых систем, работающие комплекты оборудования (основные / резервные) и пр. отображаются в отдельных параметрах телеметрической информации. Задача сводится к следующему: X^i передаются в объект контроля и имитационную модель. Результаты моделирования Y^m рассматриваются как эталонные значения для анализа результатов испытаний Y^j. Для проведения анализа предварительно должны быть выполнены

имитационные эксперименты и построена база прецедентов, определены сценарии передачи команд управления и соответствующие им варианты параметров телеметрии.

Алгоритм анализа испытаний по базе прецедентов показан на рис. 1.11.

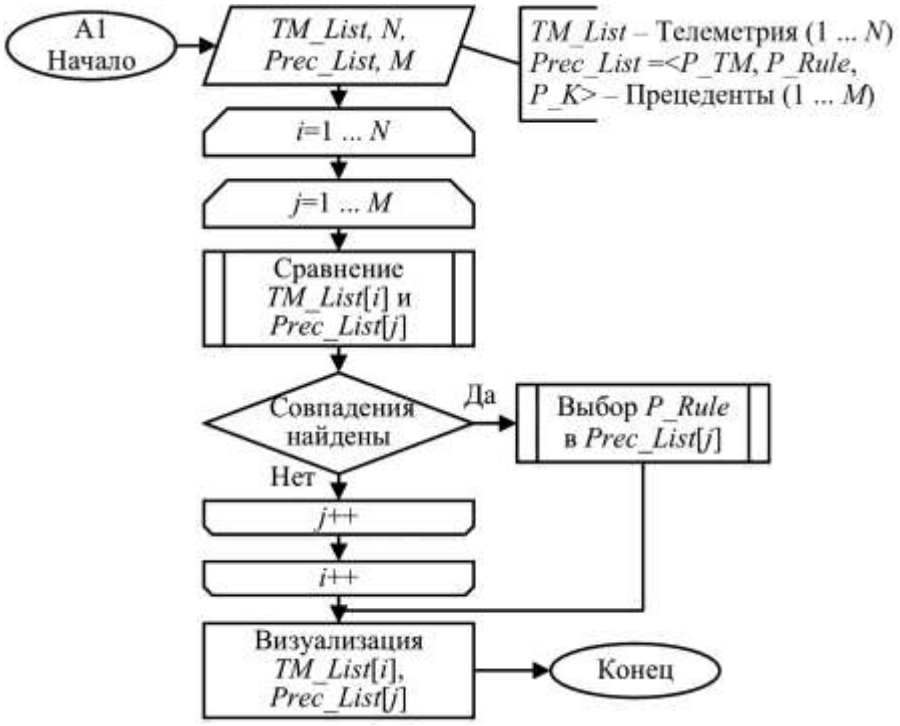


Рисунок 1.11 – Алгоритм анализа испытаний по базе прецедентов

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_List* =<*P_TM*, *P_Rule*, *P_K*>, где *P_TM* – телеметрия, формируемая имитационной моделью, *P_Rule* – правила базы знаний, *P_K* – команда, в случае, если прецедент описывает моделирование приема, передачи или обработки команды. Алгоритм 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. Применение правил базы знаний для формирования прецедентов обеспечивает детальное описание действий, которые привели к получению тех или иных значений параметров телеметрической информации. Когда выполняется приём и передача команд, поиск по

базе прецедентов может быть ограничен результатами моделирования, полученными при отработке заданной команды. Алгоритм анализа отработки команды по базе прецедентов приведён на рис. 1.12.

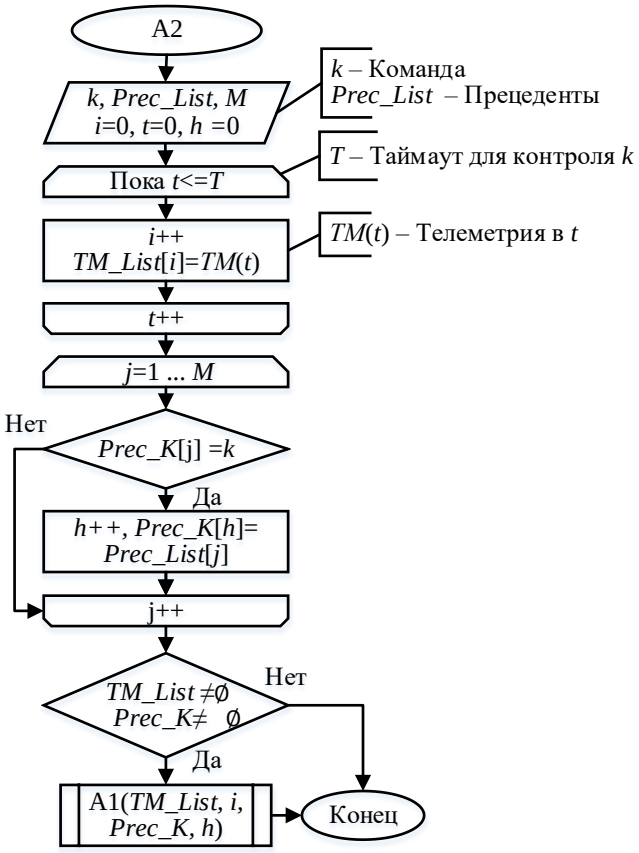


Рисунок 1.12 – Алгоритм анализа испытаний командно-программного управления

На начальном этапе работы алгоритма выбираются данные для сравнения, формируется список *TM_List*, содержащий пакеты телеметрической информации, полученных от объекта контроля после отправки команды *K* за период времени *T*. Из множества прецедентов имитационной модели *Prec_List* выбираются прецеденты, полученные при моделировании приема, передачи и отработки команды *K*. Из них формируется множество *Prec_ListK*. Если построенные множества *TM_List* и *Prec_ListK* не пустые, то выполняется «А1 – Алгоритм анализа испытаний по базе прецедентов».

Имитационная модель строится на основе конструкторских документов и протоколов информационного взаимодействия, что не только позволяет наглядно исследовать работу бортовой аппаратуры на этапе проектирования, но и служит основой для анализа уже изготовленного оборудования. Сравнение телеметрии бортовых систем с базой прецедентов дополняет методики анализа и проведения испытаний. Предложенный подход рассматривает все возможные изменения параметров, отраженные в

имитационной модели, и применяется для различных видов испытаний, вне зависимости от отправляемых в объект контроля команд. Прецедент содержит наборы правил базы знаний, которые выполнялись для его получения, что позволяет выявлять особенности работы, не наблюдаемые при анализе отдельных параметров.

Для оценки качества построенной модели предложены интерактивные графические инструменты, выполняющие интерпретацию формального описания модели в объекты инфографики. Инструменты позволяют строить графы зависимости между элементами модели (рис. 1.13), выявлять недостающие или избыточные структуры (рис. 1.14), ошибки

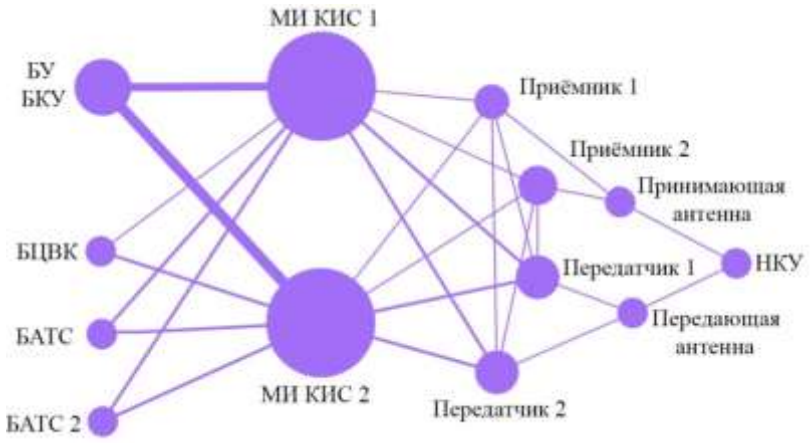


Рисунок 1.13 – Граф нагрузки на элементы модели

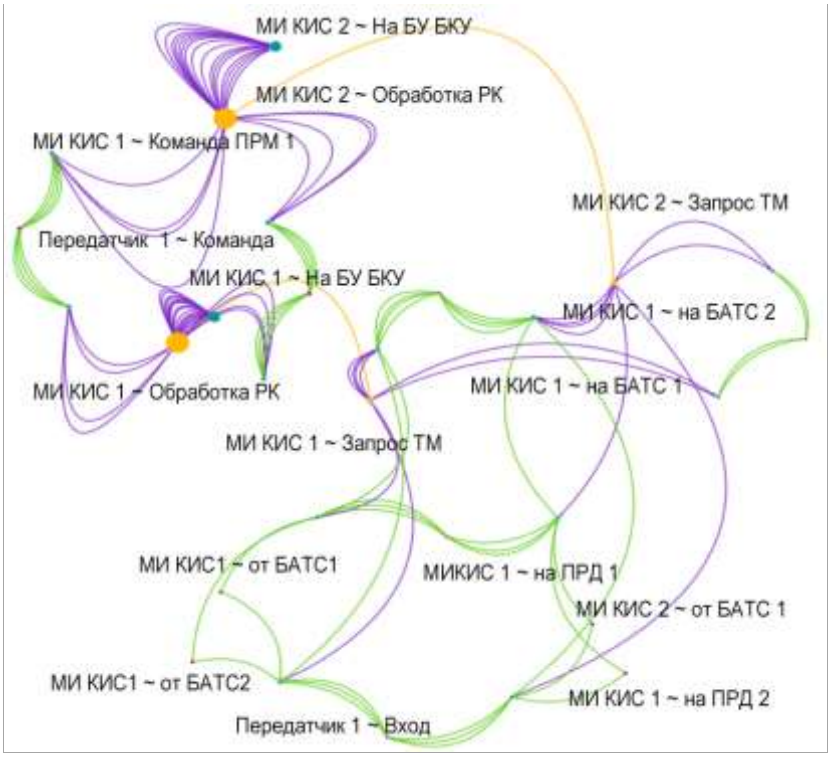


Рисунок 1.14 – Граф функциональных зависимостей

базы знаний (рис. 1.15) и обеспечивают анализ полноты функционального представления. Помимо автоматического контроля интерактивные графические инструменты могут быть использованы для проверки соответствия моделей техническим описаниям, заданным в конструкторской документации.

Рисунок 1.15 – Таблица ошибок функциональных зависимостей

Прошедшая процедуру верификации модель применяется для анализа результатов автономных испытаний бортовой аппаратуры космического аппарата. Выполняется сравнение прецедентов, полученных при проведении имитационных экспериментов с результатами испытаний готового оборудования (рис. 1.16).

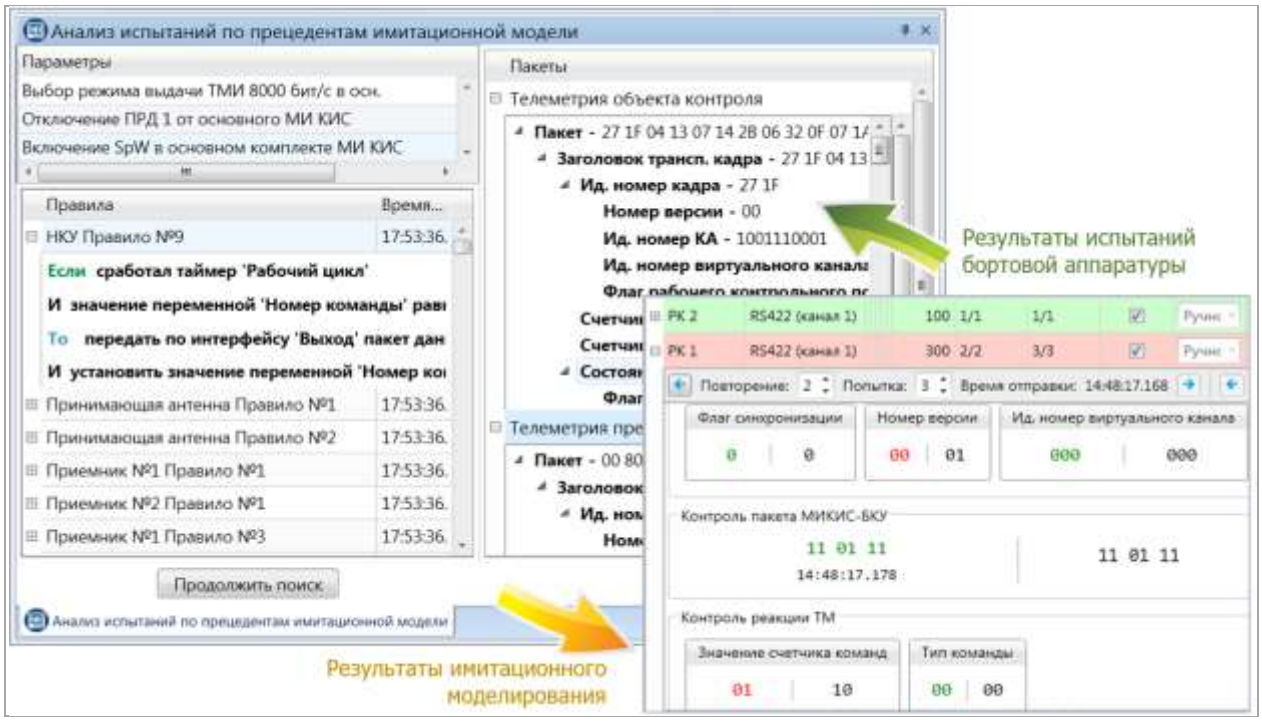


Рисунок 1.16 – Сравнение прецедентов моделирования
и результатов испытаний бортовой аппаратуры

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. Такой подход отражает особенности работы бортовых систем, которые могли остаться незамеченными при использовании других методов подготовки и проведения испытаний.

1.3 Разработка и апробация средств анализа и оценки угроз кибербезопасности в корпоративной сети

В целях изучения современных трендов кибербезопасности проведена актуализация значимости киберугроз для корпоративной сети. Было зафиксировано существенное повышение уровня угроз по каналу электронной почты (рис. 1.17), что подтверждается исследованиями компании Positive Technologies. Ведущим мотивом атак становится получение данных (54%), а основным способом распространения вредоносного программного обеспечения является электронная почта (36%) с использованием методов социальной инженерии. Для повышения безопасности электронной почты был проведен анализ журналов авторизаций пользователей, в том числе с использованием протоколов с шифрованием данных. Были выявлены и заблокированы потенциально опасные источники угроз. Недостаточность имеющихся данных для выявления активов с наибольшими рисками обусловила необходимость получения новых источников данных, что в свою очередь потребовало модификации программного обеспечения модуля авторизации пользователей электронной почты.

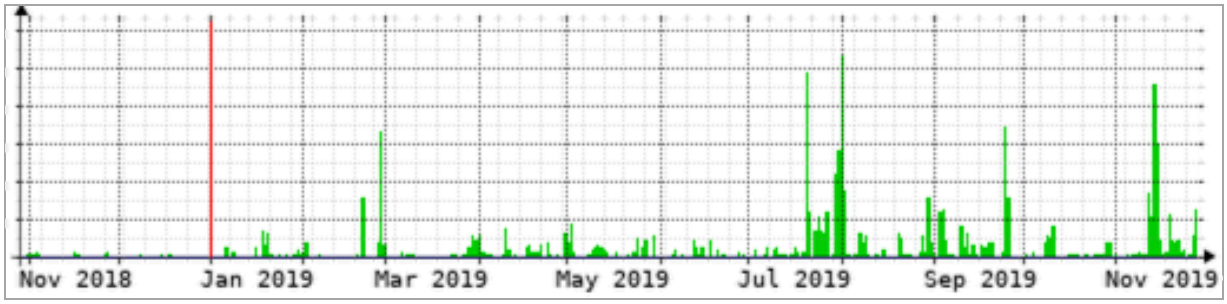


Рисунок 1.17 – Изменение количества обнаруженных вирусов в электронной почте

Продолжены работы по совершенствованию формальной расширенной ролевой модели безопасности (RBAC), адаптированной для веб-приложений и веб-сервисов, являющихся третьим по риску распространения вредоносного ПО (12%). Авторская расширенная ролевая модель безопасности дополняет классическую RBAC следующими элементами: «токен» (token), «запрос» (request), «параметр» (parameter). 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. На множестве запросов вводится бинарное отношение включения, которое задаёт нестрогий частичный порядок на множестве запросов Rq. Также задается функция RqA(), отображающая права доступа на множество запросов. Для всех запросов из множества Rq вводится иерархия запросов

RqH. Параметр (RqP) – наборов пар (ключ, значение), который передаётся при обработке запроса HTTP. Задаётся функция RqPA(), отображающая разрешения на множество параметров. Также задана функция params(), отображающая запросы на множество параметров. Данная модель безопасности учитывает специфику работы веб-приложений, однако возможности модели недостаточны для разграничения доступа в веб-сервисах. Например, при разработке веб-сервисов с использованием архитектуры REST (REpresentational State Transfer) необходимо разграничивать доступ на основе дополнительной информации – методов HTTP (GET, POST, PUT и др.). Для решения данной задачи в расширенную ролевую модель были добавлены новые элементы: «метод запроса» (RqM) и дополнительные функции: RqMA(), отображающая разрешения на множество методов ($P \to 2^{RqM}$) и methods(), отображающая запросы на множество методов ($Rq \to 2^{RqM}$). Полученная новая формальная модель описывается следующим набором элементов: <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), methods(Rq), params(Rq)>. Схема элементов модели приведена на рис. 1.18.

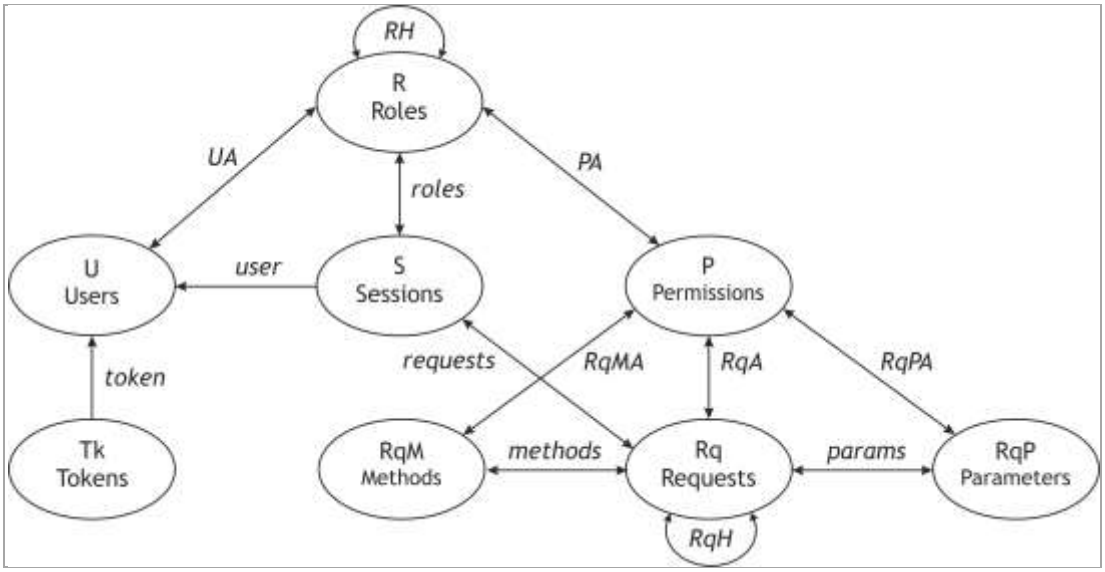


Рисунок 1.18 – Схема элементов новой модели безопасности

Построенная модель безопасности позволяет разграничивать доступ с использованием не только пути и параметров запроса, но и метода запроса. Предложенный 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).

Выполненный анализ актуальности киберугроз корпоративной сети позволил выстроить приоритеты и возможные методы защиты. Были выявлены и заблокированы потенциально опасные источники угроз. Для дальнейшего развития методов защиты определены источники данных и способы их получения. Предложенная модель безопасности позволяет повысить защищенность веб-сервисов за счет применения более гибкого разграничения доступа на основе методов запросов. Созданная модель может быть использована для решения широкого круга задач с применением технологий REST и RPC.

2 Непараметрические системы принятия решений в условиях неоднородных данных

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

2.1 Разработка и исследование быстрых алгоритмов оптимизации непараметрических решающих правил ядерного типа в условиях статистических данных большого объёма

2.1.1 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 $\bar{p}(x_1,\text{K}\,,x_k)$ 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 $\bar{p}(x_1,\text{K}\,,x_k)$ от восстанавливаемой плотности вероятности $p(x_1,\text{K}\,,x_k)$ определён оптимальный параметр:

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

где $B = \int\limits_{-\infty}^{\infty} \text{K} \int\limits_{-\infty}^{\infty} \left(\sum\limits_{v=1}^{k} \sigma_v^2 \; p_v^{(2)}(x_1,\text{K}\,,x_k) \right)^2 dx_1 \text{K}\; dx_k$; $\left\| \Phi(u_v) \right\|^2 = \int\limits_{-\infty}^{\infty} \Phi^2(u_v)\, du_v$;

$p_v^{(2)}(x_1,\text{K}\,,x_k)$ – вторая производная $p(x_1,\text{K}\,,x_k)$ 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 $\bar{p}(x_1,\text{K}\,,x_k)$ от $p(x_1,\text{K}\,,x_k)$.

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

2.1.2 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 объёма 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 порядки сократить временные затраты её решение.

Для непараметрического решающего правила, соответствующего критерию максимального правдоподобия, получены асимптотические выражения среднеквадратических отклонений $\overline{W}(c)$, $\overline{W}(c_1,c_2)$ ядерной оценки уравнения разделяющей поверхности $\bar{f}_{12}(x_1,\text{K}\,,x_k)$ от байесовской решающей функции $f_{12}(x_1,\text{K}\,,x_k)$ 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 $\bar{f}_{12}(x_1,\text{K},x_k)$ 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 $\bar{p}_1(x_1,\text{K},x_k)$, $\bar{p}_2(x_1,\text{K},x_k)$ от плотностей вероятностей распределения случайных величин x_v, $v=\overline{1,k}$ в классах. Показано, что отношение $\overline{W}(c^*)/\overline{W}(c_1^*,c_2^*)$ при оптимальных значениях $c=c^*$ и $c_1=c_1^*$, $c_2=c_2^*$ является постоянным и не зависит от объёма обучающей выборки. Данный результат сопровождается значительным снижением дисперсии $\bar{f}_{12}(x_1,\text{K},x_k)$ 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.

Результаты исследований являются перспективными при обработке данных дистанционного зондирования большого объёма.

2.1.3 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, восстанавливаемая по выборке $V=\left(x_v^i,\ v=\overline{1,k},\ i=\overline{1,n}\right)$

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

где c_v – коэффициент размытости ядерных функций $\Phi(u_v)$, которые удовлетворяют условиям положительности, симметричности и нормированности. Составляющие массива «сжатой информации» $V'=\left(z_1^j,\text{K},z_k^j,\bar{P}^j,\ j=\overline{1,N}\right)$ определяются координатами центров

элементов вводимой вычислительной сетки и частотой $\overline{P}^{j}$ попадания значений случайных величин из выборки V в эти элементы.

Ранее (2017, 2018 гг.) авторами из условия минимума асимптотического выражения среднеквадратического отклонения $\overline{p}(x)$ от восстанавливаемой плотности вероятности $p(x)$ получены формулы определения оптимального количества интервалов дискретизации N_k^* при k=1, 2. Используя метод аналогий, полученные результаты обобщены на многомерный случай:

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

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 $\prod_{v=1}^{k} \Delta_v \int_{-\infty}^{+\infty} \text{K} \int_{-\infty}^{+\infty} p^2(x_1, \text{K}, x_k)\, dx_1 \text{K}\, dx_k$. Его значения определяются только видом плотности вероятности и не зависит от её параметров.

2.1.4 На основе непараметрического алгоритма автоматической классификации разработан программный комплекс NAC v. 1.0 обработки многомерных статистических

данных большого объёма. Обнаруженные классы характеризуются одномодальными фрагментами плотности вероятности в пространстве 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. Результаты автоматической классификации представлены на рис. 2.2.

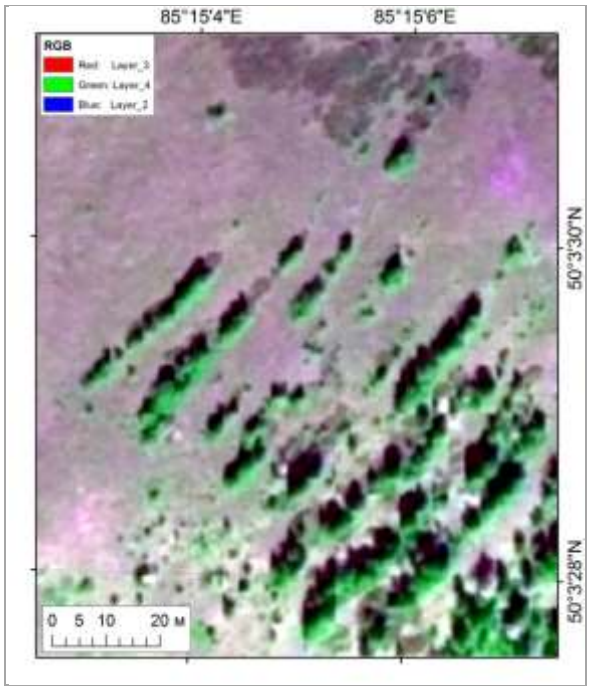
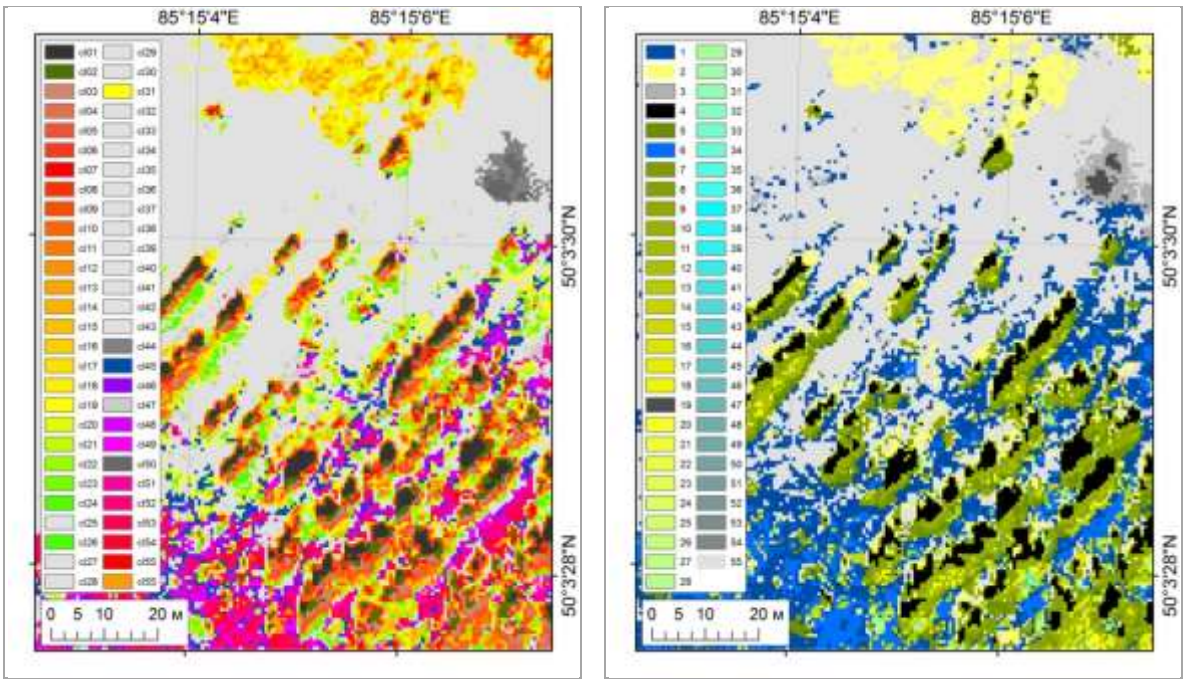


Рисунок 2.1 – Фрагмент спутниковой съемки Worldview-2



а) б)

Рисунок 2.2 – Пространственное отображение результатов автоматической классификации, полученных с использование программного продукта Erdas Imagine (а) и разработанного непараметрического алгоритма (б) при количестве классов M =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.

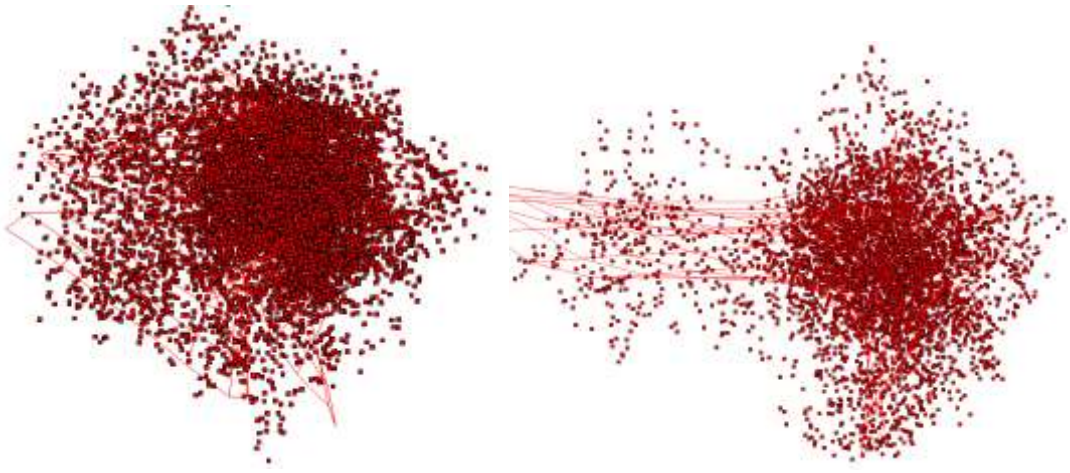
Разработанный программный комплекс NAC v. 1.0 получил государственную регистрацию (Свидетельство о государственной регистрации программы для ЭВМ № 2019660994 от 16.08.2019) и используется при обработке больших массивов данных дистанционного зондирования природных объектов.

3 Разработка методов поиска, классификации и анализа различных структур и связей между ними в нуклеотидных последовательностях

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

3.1 Изучение структурированности генетических данных больших и сверх-больших объёмов

3.1.1 В настоящее время вызывают все больший интерес, некодирующие области геномов поскольку обнаруживаются ранее неизвестные функции этих областей. За отчетный период исследована внутренняя структурированность некодирующих областей. Под структурированностью понимается кластеризация частотных словарей триплетов отдельных фрагментов генома, определяемых регулярным порядком, вне зависимости от функциональной роли того или иного участка в 64-мерном пространстве частот триплетов. Каждая некодирующая область была разбита на блоки с набором перекрывающихся фрагментов одинаковой длины, и эти фрагменты были преобразованы в словари триплетных частот. Словари были сгруппированы в 64-мерном евклидовом пространстве. Кластеризация проводилась методом упругих карт для проекции 64-мерного пространства в пространство первых трех главных компонент. Было выделено пять типов распределений: шар, шар с хвостом, шар с двумя хвостами, линза с хвостом и линза с двумя хвостами (рис. 3.1).



а) б)

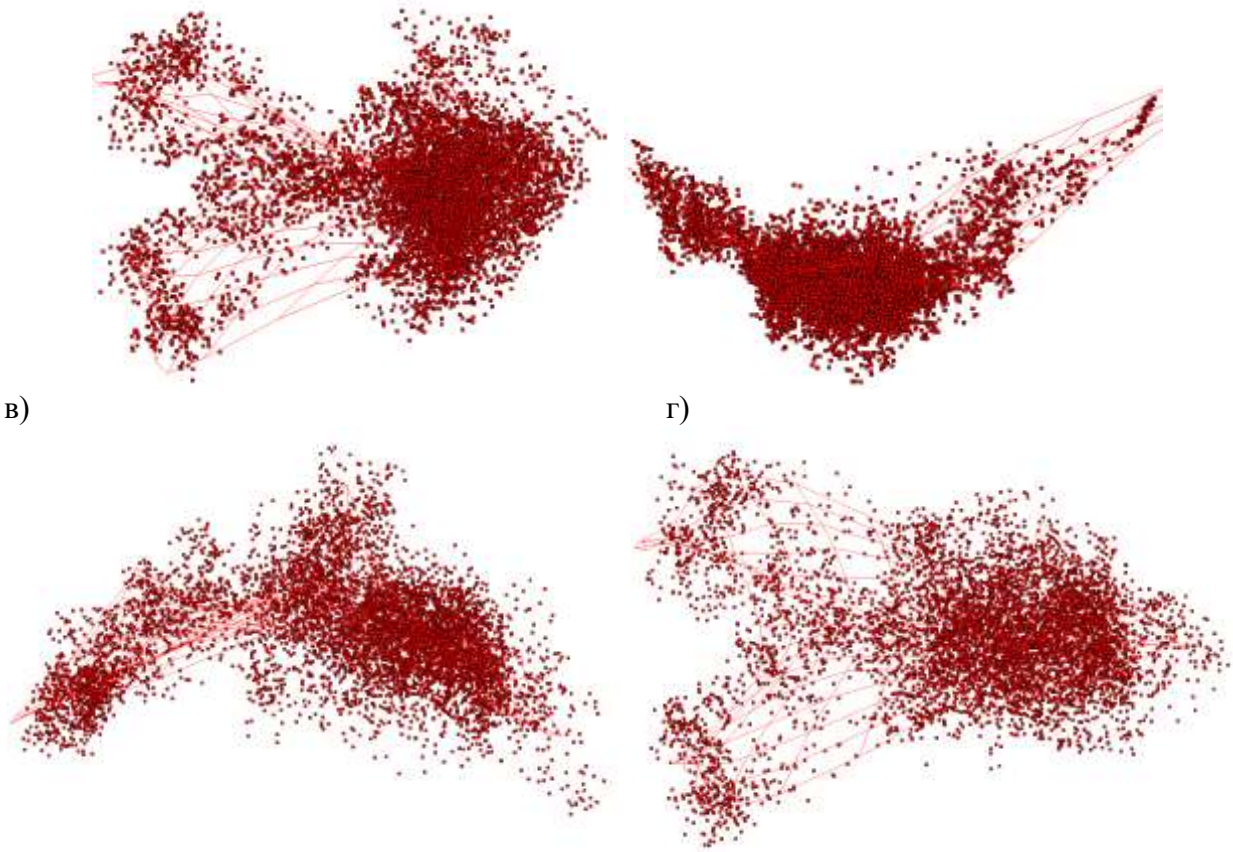


Рисунок 3.1 – Структуры некодирующих областей геномов хлоропластов: а) структура типа «Шар» на примере генома *Erodium chrysanthum,* б) «Шар с одним хвостом» на примере генома *Pseudotsuga sinensis,* в) «Шар с двумя хвостами» на примере генома *Liriodendron tulipifera,* г) «Линза с одним хвостом» на 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, окрашивалось в отдельный цвет (рис. 3.2). 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. 3.2.

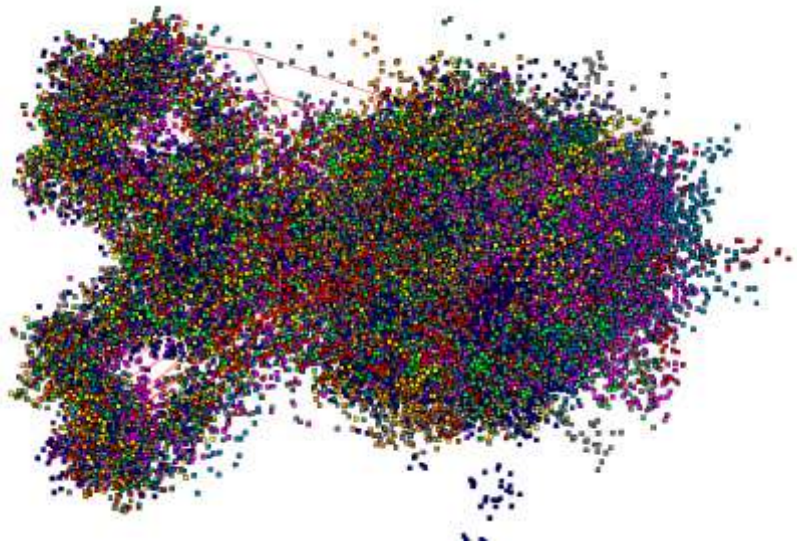


Рисунок 3.2 – Объединённая структура геномов Genlisea margaretae, Schwalbea americana, Pinguicula ehlersiae, Orobanche gracilis, Ptilidium pulcherrimum, Magnolia kwangsiensis, Jacobaea vulgaris, Mankyua chejuensis, Sesamum indicum и Vaccinium macrocarpon

3.1.2 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 частототных 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 последова тельной обработки генетического текста. Сам транскриптом может рассматриваться как набор этих фрагментов, с небольшими исключениями. Во-первых, длина транскриптомов 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 транскриптомам 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 же стренда. Кроме того, было обнаружено, что октаэдрическая структура распределения контигов транскритпомов для L. sibirica имеет зеркальную симметрию, а для P. sibirica – вращательную (рис. 3.3). 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». «Ведьмина метла» – это фрагменты кроны растения с аномальным морфогенезом, для которого неизбежны серьезные генетические изменения в его геноме.

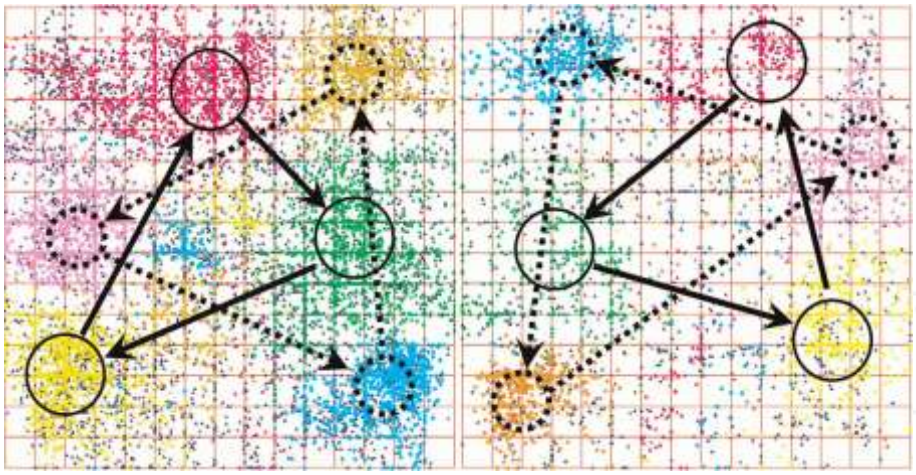


Рисунок 3.3 – Зеркальная (слева) симметрия в транскриптоме *L. sibirica* против вращательной симметрии (справа) в транскриптоме *P. sibirica*.
Непрерывные стрелки и окружности соответствуют прямым относительным фазам, а пунктирные – обратным

3.1.3 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 участи 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. 3.4). 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 $\tilde{f}_\omega$. Те слова, для которых разность f_ω и $\tilde{f}_\omega$ 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.

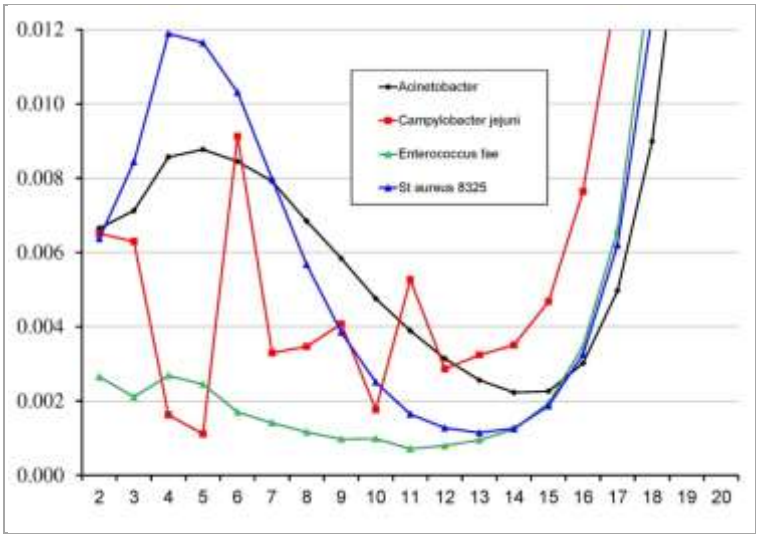


Рисунок 3.4 – Информационная емкость для двоичной символьной последовательности, представляющей распределение стартов

3.1.4 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. Затем гены были преобразованы в триплетные частотные словари так, что каждый ген был представлен в виде точки в 64-мерном 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. Обнаружено, что в формировании кластеров большое влияние оказывают имено гены (рис. 3.5). Такое преобладание заранее не очевидно. Преобладание влияния генов доказывает превосходство функции над таксономией в формировании структурированности в триаде структура – функция – таксономия.

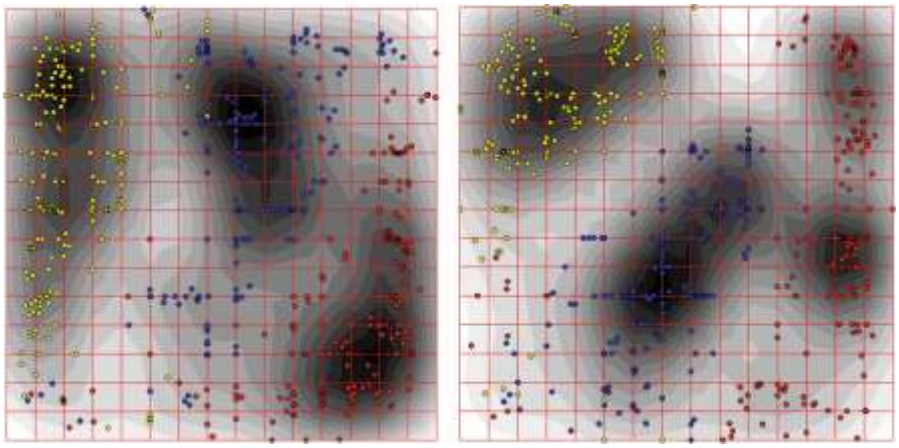


Рисунок 3.5 – Распределение генов семейства *atp*, представленных во внутренних координатах упругой карты. Гены *atp6* обозначены красным цветом, *atp8* синим и *atp9* - желтым; слева для слева частотный словарь вычислен для шага $t = 1$, справа для $t = 3$

3.1.5 Визуализация общего транскриптома лиственницы сибирской (посредством трансформации длинных контигов в словари триплетных частот $W_{(3,1)}$ и $W_{(3,3)}$) показывает его структурированность. На рис. 3.6 показано распределение контигов. Кластеры, показанные на рис. 3.6 (a) и (b) получены с помощью упругой карты. Кластеризация для K-means при K=2 и K=3 показана на рис. 3.6 (c) и (d). Следует сказать, что эти два разбиения на классы очень устойчивы: более 85% прогонов K-means дают одинаковое распределение точек. При K≥4 распределение становится неустойчивым. Рис. 3.6 отвечает на ключевой вопрос работы – обеспечивает ли общий транскриптом лучшую сборку тканеспецифичных контигов или нет. Проанализировано распределение контигов с повышенным содержанием тканеспецифических прочтений. Для этого сначала определили контиги с высоким уровнем взаимной энтропии, затем проверили, какие ткани преобладают в контиге, и пометили его в соответствии с распространенностью ткани. На рис. 3.7 показано полученное распределение; розовыми кругами обозначен камбий, зелеными треугольниками обозначены иглы и коричневыми пятиугольниками обозначены побеги. Очевидно, что тканеспецифические обогащенные контиги не образуют выделенные кластеры. Кроме того, поведение невязок Чаргаффа выглядит весьма примечательным: для K-means при классификации с K=2 внутриклассовые невязки составляют $\xi_1 = 5,45 \times 10^{-4}$ и $\xi_1 = 5,90 \times 10^{-4}$ соответственно с межклассовой невязкой $\mu_{(1,2)} = 8,20 \times 10^{-4}$. Невязка между двумя классами больше, чем невязки внутри классов. Для K=3 ситуация иная. Здесь внутриклассовые невязки отличаются для трех классов: $\xi_1 = 6,29 \times 10^{-4}$, $\xi_2 = 2,64 \times 10^{-4}$ и $\xi_3 = 5,91 \times 10^{-4}$ соответственно.

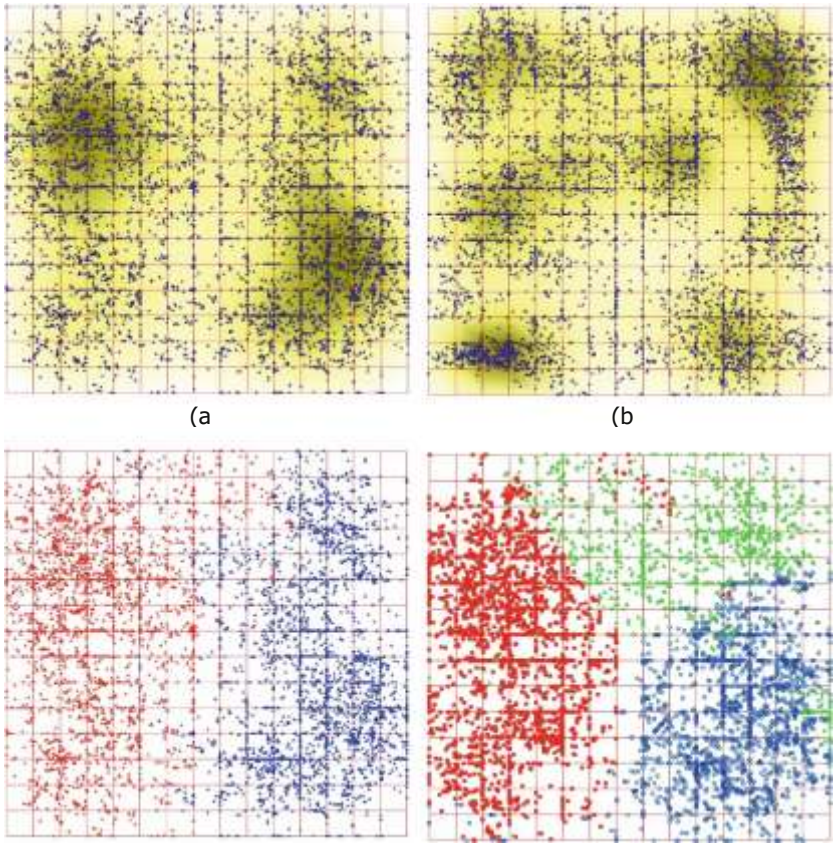


Рисунок 3.6 – Распределение транскриптов с большей взаимной энтропией, чем у транскриптома в целом; (a) случай $W_{(3,1)}$, (б) случай $W_{(3,3)}$.

K-means показан в (c) *K=2* и в (d) *K=3*; оба случая посчитаны для словаря $W_{(3,1)}$

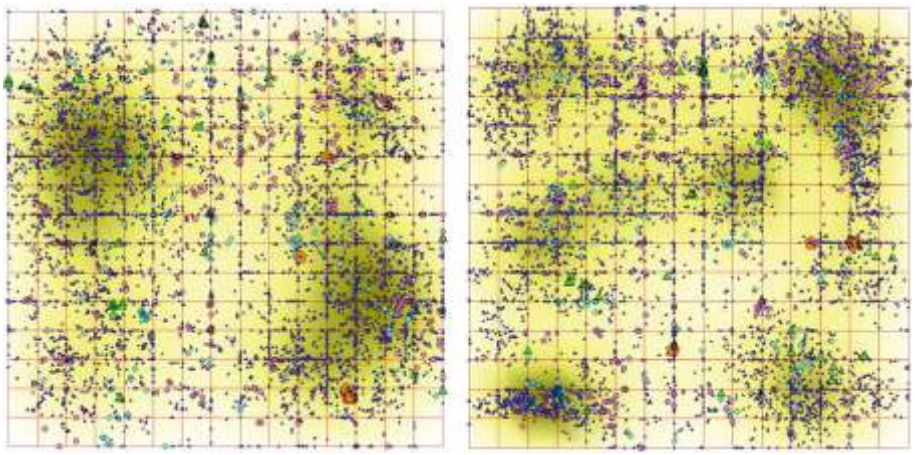


Рисунок 3.7 – Распределение контигов с более высоким предпочтением встречаемости тканевых специфических чтений; случай $W_{(3,1)}$ слева и случай $W_{(3,3)}$ справа

Очевидно, что второй класс выпадает из общей картины невязок Чаргаффа. Этот факт говорит о том, что второй класс включает в себя контиги двух видов, в отличие от первого и третьего класса. Это предположение подтверждается цифрами межклассовых

невязок: $\mu_{(1,2)} = 3,32 \times 10^{-4}$, $\mu_{(2,3)} = 4,27 \times 10^{-4}$, но $\mu_{(1,3)} = 3,71 \times 10^{-5}$. Это обычная практика, когда исследователь не гарантирован от необходимости изучения общего транскриптома, а не (тканевого) специфического. Такие ситуации могут иметь место, когда новый (или редкий) образец находится в стадии анализа. Следовательно, необходимо иметь инструмент для оценки пределов знаний, которые могут извлекаться из общего транскриптома. Значительное количество химерных транскриптов может создать проблему в анализе общего транскриптома, скажем, в дифференциальной оценке выражения. Если тканевая специфичность различных считываний известна априори, то можно исключить химерные контиги из ансамбля, используя оценку энтропии.

ЗАКЛЮЧЕНИЕ

За отчетный период выполнения проекта получены следующие результаты.

Предложен 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.

Предложен метод формирования унифицированного представления структуры межсистемного обмена данными с учетом вариативности форматов. Разработанные на его основе алгоритмы позволяют автоматизировать обмен данными между информационными системами с возможностью адаптации к изменениям условий и форматов передаваемых данных. Метод и алгоритмы апробированы в задачах организации закупок, где осуществляется обмен данными между корпоративной системой размещения заказов, единой информационной системой и электронными торговыми площадками.

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. Предложен алгоритм определения степени сходства разнородных измерений путем семантико-синтаксического анализа текстовых атрибутов аналитических измерений. Программная реализация метамодели и алгоритма обеспечивает формирование интегральной аналитической модели гетерогенных данных.

Разработаны алгоритмические и программные средства построения прикладных систем на базе интерактивной веб-платформы. Выполнено развитие веб-платформы, обеспечивающей автоматизацию создания модельно-ориентированных систем сбора данных. Для обеспечения анализа данных с помощью сторонних инструментов реализован REST-интерфейс передачи отчётов. Выполнена апробация разработанных средств в задачах оценки экологического состояния почв населенных пунктов Красноярского края.

Разработана технология ситуационного моделирования опасных событий, обеспечивающая информационную поддержку управления безопасностью территорий и позволяющая решать задачи предупреждения и ликвидации всех видов опасных событий. Технология основана на ранее предложенной системной модели построения

информационно-аналитических систем поддержки управления природно-техногенной безопасностью, предусматривающей совместное использование расчётных методик оценки последствий опасных ситуаций различной сложности, метода динамического картографирования, экспертных систем и веб-технологий. На основе гибридного подхода, сочетающего преимущества технологий разработки нативных и веб-приложений, реализован комплекс многофункциональных сервисов безопасности жизнедеятельности, интегрированных с информационными системами оперативного комплексного мониторинга чрезвычайных ситуаций.

Выполнено развитие метода интегрального аналитического оценивания природно-техногенной безопасности территорий, обеспечивающего формирование комплексного показателя на основе результатов многомерного аналитического моделирования состояния окружающей среды и объектов техносферы. Для интерпретации количественных значений оценок предложен подход к формированию оценочной шкалы, который обеспечивает отображение количественных значений в качественное выражение с использованием аппарата нечетких множеств. Согласно предложенному подходу, функция принадлежности определяется с помощью метода нечеткой кластеризации, который позволяет определить функции принадлежности по распределению значений интегральных оценок.

Усовершенствованы схемы метаданных публикаций, обеспечивающие минимальные потери информации при обмене данными открытого архива и системы автоматизации библиотек. Особое внимание уделено оптимизации индексирования открытого архива поисковыми системами Интернет, онлайновой поддержке открытого лицензирования публикаций Creative Common (CC) и интеграции с системой Sherpa/Romeo, обеспечивающей соответствие доступа к ресурсам издательским требованиям.

Предложен метод анализа результатов испытаний бортовой аппаратуры космического аппарата, основанный на применении прецедентов имитационной модели. Построение гетерогенной имитационной модели и её применение для автоматизации анализа результатов испытаний повышает эффективность принятия решений при формировании технических спецификаций на оборудование и его последующем тестировании и служит основой подготовки производства бортовой аппаратуры космического аппарата.

Выполнен анализ актуальности киберугроз корпоративной сети, который позволил выстроить приоритеты и возможные методы защиты. Выявлены и заблокированы потенциально опасные источники угроз. Для дальнейшего развития методов защиты

определены новые источники данных и способы их получения. Выполнено развитие формальной ролевой модели безопасности (RBAC), путем добавления элементов, характерных для веб-приложений и веб-сервисов. Предложенная модель позволяет повысить защищенность за счет применения более гибкого разграничения доступа на основе методов запросов. Созданная модель может быть использована для решения широкого круга задач с применением технологий REST и RPC.

Разработан быстрый алгоритм выбора оптимальных коэффициентов размытости ядерных функций в непараметрической оценке многомерной плотности вероятности типа Розенблатта-Парзена. Предложена методика оптимальной дискретизации области определения многомерной плотности вероятности, которая основана на обобщении формул дискретизации для одномерного и двухмерного случаев. Методика быстрого выбора коэффициентов размытости ядерных функций в непараметрической оценке многомерной плотности вероятности позволяет на порядки сократить время их синтеза по сравнению с традиционным подходом, что является актуальным при решении задач принятия решений в условиях статистических данных большого объёма. Формула оптимальной дискретизации области значений многомерной случайной величины получена на основе обобщения формул дискретизации для одномерного и двухмерного случаев и использования метода аналогий. 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. Полученные результаты реализованы в программном комплексе и используются для решения задач автоматической классификации данных дистанционного зондирования природных объектов.

Изучены статистические свойства некодирующих областей хлоропластных геномов 391 растения с использованием кластеризации частотных словарей триплетов отдельных фрагментов генома, определяемых регулярным порядком. Анализ внутренней структурированности некодирующих областей геномов хлоропластов наземных растений позволил обнаружить пять типов характерных структур. На основании одновременной кластеризации наборов некодирующих областей хлоропластов разных видов можно предположить, что исходным типом структуры была «линза с двумя хвостами», из которой в процессе эволюции редуцировались остальные типы структур.

Применена новая методика для изучения комбинаторных и статистических свойств транскриптомных последовательностей на основе распределения нуклеотидных триплетных частотных словарей, полученных при конверсии транскриптомных последовательностей. Рассмотрены транскриптомы лиственницы сибирской (Larix sibirica

Ledeb.) и сосны сибирской (Pinus sibirica Du Tour). Транскриптомы демонстритуют необычную симметрию в распределении частотных словарей триплетов. Была обнаружена октаэдрическая структура распределения контигов транскритпомов, причем для L. sibirica она имела зеркальную симметрию, а для P. sibirica – вращательную. Проведенные исследования позволяют предположить, что октаэдрическая структура возможно является универсальной для растений.

План работ по проекту выполнен полностью на высоком научном уровне, что подтверждается публикациями коллектива исполнителей проекта в ведущих рецензируемых отечественных и зарубежных научных изданиях.

ПРИЛОЖЕНИЕ А Научные публикации

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

1. Penkova T., Korobko A.V. Investigation of hydropower equipment functioning features using data mining techniques // Lecture Notes in Computer Science. – Part I, Vol. 11619. – P. 434-446. DOI: 10.1007/978-3-030-24289-3_32.
2. Пенькова Т.Г., Коробко А.В., Валов Ю.Н. Исследование особенностей функционирования гидроагрегата на основе комплексного анализа данных вибрационного контроля // Приборы и системы. Управление, контроль, диагностика. – 2018. – № 12. – С. 36-45. DOI: 10.25791/pribor.12.2018.000.
3. Ничепорчук В.В., Ноженков А.И., Коробко А.А. Мобильные приложения мониторинга безопасности жизнедеятельности // Образовательные ресурсы и технологии. – 2018. – № 4(25). – С. 60-65. DOI: 10.21777/2500-2112-2018-4-60-65.
4. Penkova T., Metus A.M. Method of Comprehensive Estimation of Natural and Anthropogenic Territory Safety in the Case of Krasnoyarsk Region // Lecture Notes in Computer Science. – 2019. – Part I, Vol. 11619. – P. 421-433. DOI: 10.1007/978-3-030-24289-3_31.
5. Исаева О.С., Ноженкова Л.Ф., Колдырев А.Ю. Интеллектуальный анализ испытаний бортовой аппаратуры космического аппарата // Вычислительные технологии. – 2019. – Т. 24, № 3. – С. 59–74. DOI: 10.25743/ICT.2019.24.3.005.
6. Исаева О.С., Колдырев А.Ю. Алгоритмы анализа испытаний командно-программного управления бортовой аппаратурой космического аппарата // Информатика и системы управления. – 2019. – № 1(59). – С. 46-57. DOI: 10.22250/isu.2019.59.46-57.
7. Isaeva O., Nozhenkova L. Spacecraft onboard equipment testing automation technology on the basis of simulation model // IOP Conference Series: Materials Science and Engineering. – 2019. – P. 1-6. DOI:10.1088/1757-899X/537/3/032067.
8. Kulyasov N., Isaeva O., Isaev S. Method of creation and verification of the spacecraft onboard equipment operation model // IOP Conference Series: Materials Science and Engineering. – 2019. – P. 1-6. DOI:10.1088/1757-899X/537/2/022042.
9. Isaeva O.S., Nozhenkova L., Koldyrev A.Yu. Methods of test procedures' generation on the basis of simulation model's knowledge base // Advances in Intelligent Systems Research. – 2019 – Vol. 164. – P. 32-35. DOI: 10.2991/mmssa-18.2019.8.

10. Isaeva O.S., Koldyrev A.Yu., Chernigovskiy A.S., Mishurov A.V., Kamyshnikov A. N., Evstratko V. V. Automated support for spacecraft onboard equipment design on the basis of a heterogeneous model // Journal of Physics: Conference Series. –2019. – V. 1353, N. 1. – P. 012011. DOI:10.1088/1742-6596/1353/1/012011.
11. Lapko A.V., Lapko V.A. Fast algorithm for choosing kernel function blur coefficients in a nonparametric probability density estimate // Measurement Techniques. – 2018. – Vol. 61. – No. 6. – P. 540-545. DOI: 10.1007/s11018-018-1463-9.
12. Lapko A.V., Lapko V.A. Fast algorithm for choosing blur coefficients in multidimensional kernel probability density estimates // Measurement Techniques. – 2019. – Vol. 61, No. 10. – P. 979-986. DOI: 10.1007/s11018-019-01536-x
13. Lapko A.V., Lapko V.A. Methods for rapid selection of kernel function blur coefficients in a nonparametric pattern recognition algorithm // Measurement Techniques. – 2019. – Vol. 62, No. 4. – P. 300-306. DOI: 10.1007/s11018-019-01621-1
14. Lapko A.V., Lapko V.A. Discretization method for the range of values of a multi-dimensional random variable // Measurement Techniques. – 2019. – Vol. 62, No. 1. – P. 16-22. DOI: 10.1007/s11018-019-01579-0
15. Lapko A.V., Lapko V.A., Im S.T., Tuboltsev V.P., Avdeenok V.A. Nonparametric algorithm of identification of classes corresponding to single-mode fragments of the probability density of multidimensional random variables // Optoelectronics, Instrumentation and Data Processing. – 2019. – Vol. 55, No. 3. – P. 230-236. DOI: 10.3103/S8756699019030038.
16. Michael Sadovsky, Maria Senashova, Inna Gorban, Vladimir Gustov. Non-Coding Regions of Chloroplast Genomes Exhibit a Structuredness of Five Types // LNBI. – 2019. – Vol. 11465. – P. 346-355. DOI:10.1007/978-3-030-17938-0_31.
17. Sadovsky M., Kobets V., Khodos G., Kuzmin D., Sharov V. Reads in NGS Are Distributed over a Sequence Very Inhomogeneously // LNBI. – 2019. – Vol. 11465. – P. 271-282. DOI: 10.1007/978-3-030-17938-0_25.
18. Sadovsky M., Fedotovskaya V., Kolesnikova A., Shpagina T., Putintseva Yu. Function vs. Taxonomy: The Case of Fungi Mitochondria ATP Synthase Genes // LNBI. – 2019. – Vol. 11465. – P. 335-345. DOI: 10.1007/978-3-030-17938-0_30.
19. Sadovsky M., Guseva T., Biriukov V. Triplet Frequencies Implementation in Total Transcriptome Analysis // LNBI. – 2019. – Vol. 11465. – P. 370-378. DOI: 10.1007/978-3-030-17938-0_33.

ПРИЛОЖЕНИЕ Б

Выписка из плана научно-исследовательских работ на 2020 год

Содержание работы	Планируемый результат выполнения работы
1. Разработка гибридных технологий обработки и визуализации многомерных данных в интегрированных информационно-аналитических системах. Разработка методов анализа имитационной модели по результатам испытаний бортовой аппаратуры космического аппарата.	1. Методы когнитивного моделирования и инфографического представления результатов оперативной аналитической обработки многомерных данных. Алгоритмы расчета аналитических показателей, описывающих метрические свойства и динамику изменения многомерных аналитических моделей состояния сложных объектов и систем. Методы связывания данных в модели открытого доступа на примере цифрового архива ФИЦ КНЦ СО РАН. Реализация моделей и методов защиты интегрированных информационно-телекоммуникационных систем. Методы анализа имитационной модели по результатам испытаний бортовой аппаратуры космического аппарата.
2. Разработка непараметрических систем проверки гипотез о распределениях многомерных случайных величин и их независимости, основанных на использовании методов распознавания образов.	2. Новые непараметрические критерии проверки гипотез о распределениях многомерных случайных величин и их независимости.
3. Анализ многомерных данных медицинских клинических исследований, содержащих пробелы. Разработка методов анализа таких данных, их визуализации и обработки, с помощью методов кластеризации, основанных на теории графов.	3. Анализ медицинских данных, выявление клинически значимых характеристик в больших массивах данных.