На правах рукописи

Воробьева Гульнара Равилевна

МЕТОДОЛОГИЧЕСКИЕ ОСНОВЫ ОБРАБОТКИ НЕОДНОРОДНОЙ ПРОСТРАНСТВЕННО-ВРЕМЕННОЙ ИНФОРМАЦИИ В СИСТЕМАХ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ НА ОСНОВЕ ТЕХНОЛОГИЙ БОЛЬШИХ ДАННЫХ (НА ПРИМЕРЕ ГЕОМАГНИТНЫХ ДАННЫХ)

Специальность:
05.13.01 – Системный анализ, управление и обработка информации (информационные и технические системы)

Автореферат диссертации на соискание ученой степени доктора технических наук

Уфа – 2020

Работа выполнена в ФГБОУ ВО Уфимский государственный авиационный технический университет

Научный консультант: доктор технических наук, профессор
Юсупова Нафиса Исламовна

Официальные оппоненты:

Рогозов Юрий Иванович, доктор технических наук, профессор, ФГБОУ ВО «Южный федеральный университет», заведующий кафедрой системного анализа и коммуникаций
Андрианов Дмитрий Евгеньевич, доктор технических наук, доцент, Муромский институт (филиал) ФГБОУ ВО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых», заместитель директора по учебной работе, заведующий кафедрой информационных систем
Щербаков Максим Владимирович, доктор технических наук, доцент, ФГБОУ ВО «Волгоградский государственный технический университет», заведующий кафедрой систем автоматизированного проектирования и поискового конструирования

Ведущая организация: ФГБОУ ВО «Самарский национальный научно-исследовательский университет имени академика С.П. Королева», г. Самара

Защита диссертации состоится 1 декабря 2020 г. в 10^{00} часов на заседании диссертационного совета Д 212.288.12 на базе ФГБОУ ВО «Уфимский государственный авиационный технический университет» по адресу: 450008, г. Уфа, ул. К. Маркса, 12.

С диссертацией можно ознакомиться в библиотеке ФГБОУ ВО «Уфимский государственный авиационный технический университет» и на сайте *www.ugatu.su*.

Автореферат разослан «___» _____ 20___ года.

Ученый секретарь
диссертационного совета Сметанина Ольга Николаевна

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Эффективность систем поддержки принятия решений во многом связана с широтой спектра используемых ими данных. Поэтому принятие решений, как правило, базируется на информации, полученной из многих, зачастую разнородных, источников. Экспоненциальный рост данных приводит при этом к актуальной на сегодняшний день проблеме информационной перегруженности лиц, принимающих решения, что выражается в необходимости подбора источников данных, фильтрации, унификации и обработке предоставляемой ими информации, последующем анализу больших объемов данных доступными в системе поддержки принятия решений инструментально-программными средствами, которые, в свою очередь, не всегда эффективно справляются с возложенными на них задачами в условиях пространственной неоднородности соответствующих данных.

Решение обозначенной проблемы является критически важным в сфере техносферной безопасности, где своевременное получение информации является залогом полного предотвращения или снижения негативных последствий чрезвычайных ситуаций. Одним из факторов окружающей среды, мониторинг которого необходим для обеспечения безопасности ряда объектов техносферы, является геомагнитная обстановка, которая количественно описывается данными, регистрируемыми в режиме реального времени наземными магнитными обсерваториями и вариационными станциями. Существование множества разрозненных источников данных и отсутствие механизмов их интеграции усложняют использование результатов наблюдений за параметрами геомагнитного поля и его вариаций, поскольку в прикладных областях для принятия обоснованных решений нужен единый источник достоверных данных. Проблема интеграции геомагнитных данных усугубляется их гетерогенностью, которая на физическом уровне проявляется в использовании различных форматов данных, а на логическом – в неоднородности состава регистрируемых параметров и представлении результатов наблюдений в различных шкалах измерений.

В настоящее время проблема объединения множества разнородных источников данных решается с помощью широкого спектра программных и инструментальных средств. Так, широко распространенная технология Apache Hadoop предоставляет модули для управления большими распределенными данными в высоконагруженных веб-ориентированных приложениях. Учеными Университета Модены (Италия) предложен подход MOMIS к интеграции источников данных на основе медиатора, учитывающий масштабы и характеристики больших данных и позволяющий создавать соответствующие прикладные информационные системы. Проект SuperMag, разработанный учеными из Университета Дж.Хопкина (США) и Университета Бергена (Норвегия), обеспечивает доступ к геомагнитным данным 200 вариационных

станций и отдельных магнитометров в единой координатной системе в унифицированных единицах измерения с заданным временным разрешением.

Известные решения не учитывают особенностей пространственно-временного распределения данных и их разнородных источников, представляют временные ряды с пропусками и аномалиями, что является значимым препятствием на пути их автоматического анализа и визуализации. В проекте геомагнитных данных SuperMag, к примеру, выбросы и пропущенные во временных рядах значения удаляются из набора предоставляемых пользователю данных, что может привести к потере критически важной информации, а также негативно сказывается на качестве и информативности систем визуализации данных. Еще одним недостатком обозначенных решений является низкая вычислительная скорость обработки данных от гетерогенных источников.

Указанные ограничения перечисленных подходов не позволяют удовлетворить с их помощью возрастающие потребности потребителей пространственно-зависимой информации. При этом наблюдающийся в сфере информационных технологий бум средств и технологий визуализации данных указывает на то, что данное направление является одним из перспективных в плане оперативного анализа больших объемов данных.

В этой связи представляется целесообразным создание методологических основ обработки неоднородной пространственно-временной информации как парадигмы обработки и анализа данных на основе их пространственного распределения, базирующейся на совокупности согласованных методов и подходов формирования единого информационного пространства в виде множества восстановленных пространственных данных заданного формата, а также результатов их графической и аналитической обработки.

Степень разработанности темы. В ряде работ российских и зарубежных ученых (О.И. Ларичев, А.Б. Петровский, О.В. Тиханычев, А.Д. Оспанов, Ч.А. Найданов, Н.И. Юсупова, Л.Р. Черняховская, М.Б. Гузаиров, С.В. Павлов, Дж. Гупта, М. Мора, Б. Маклин, И. Чен, Э. Беллуччи и др.) дается широкое представление о подходах к разработке систем поддержки принятия решений, обсуждаются вопросы создания моделей и методов извлечения из данных распределенных источников, которые необходимы лицам, принимающим решения. Также в ряде работ (Е.А. Карпова, И.Б. Зайчикова, А. А. Зацаринный, Э.В. Киселев, С.В. Козлов, К.К. Колин, В.А. Зеленцов и др.) рассматриваются подходы к созданию единых информационных хранилищ, обсуждаются подходы к федерализации и консолидации источников данных. Применительно к сфере техносферной безопасности основные научные исследования (В.В. Галасюк, Л.М. Зейналова, В.А. Акимов, В.Д. Новиков, Н.Н.Радаев, В.М. Дозорцев, Н. Чура, К. Канаэртс, М.Пьютч и др.) сопряжены с конкретными прикладными областями и ориентируются на базовые принципы построения информационных систем. Наконец, во многих работах отечественных и зарубежных ученых (Ю. В. Афанасьев, Ю. М. Яновский, А. Д. Гвишиани, В. Г.

Петров, А. А. Соловьев, В. В. Любимова, А. Н. Зайцев, С. П. Гайдаш, А. В. Белов, Н. Г. Птицына, Р. А. Рахматуллин, Любимов В.В., В. Х. Кириаков, В. Н. Бобров, Ю. Г. Астраханцев, П. Рипка, Д. Ванг, А. Типек, А. Томсон, Дж. Виллорези и др.) дается представление о методах и подходах, обеспечивающих регистрацию и оценку параметров магнитного поля внутриземных источников и их вариаций, а также о типовых средствах, используемых для сбора, обработки и интерпретации геомагнитных данных. Однако ввиду отсутствия единой концепции объединения разнородных источников данных, невозможности агрегирования данных в соответствии с особенностями их пространственно-временного распределения, высокими вычислительными затратами и распространенной невозможностью реализации в условиях ограниченных вычислительных ресурсов перечисленные подходы не решают проблему информационной избыточности систем поддержки принятия решения, особенно в области техносферной безопасности (и мониторинга геомагнитных возмущений в частности).

Цель работы – повышение эффективности процессов обработки информации в системах поддержки принятия решений посредством разработки единых методологических основ обработки, анализа и визуализации больших объемов пространственных данных, полученных из территориально распределенных гетерогенных источников. Для достижения поставленной цели необходимо решить следующие **задачи**.

1. Анализ проблемы обработки пространственно-зависимой информации из распределенных гетерогенных источников и разработка методологических основ обработки неоднородной пространственно-временной информации в системах поддержки принятия решений на основе теоретико-множественного, теоретико-информационного и статистического подходов, обеспечивающих достоверность результатов применяемых методов и подходов в процессе формирования единого информационного пространства.

2. Разработка комплекса моделей и методов обработки информации при интеграции гетерогенных источников данных в гибридное хранилище систем поддержки принятия решений, обеспечивающих доступность разнородных территориально распределенных источников информации для оперативной обработки и анализа больших объемов предоставляемых ими данных.

3. Разработка методов восстановления временных рядов данных (на примере геомагнитной информации) в информационных системах на основе принципов машинного обучения и информационного резервирования источников информации, обеспечивающих возможность импутации пропусков с ошибкой в пределах допустимого нормативами отклонения.

4. Разработка модели хранения данных в системах поддержки принятия решений, обеспечивающего сокращение вычислительных затрат на их физическое размещение и повышение вычислительной скорости обработки запросов к данным.

5. Разработка алгоритма визуализации пространственно-временного распределения данных на примере геомагнитной информации с возможностью варьирования уровня детализации представления земного и околоземного пространства, что позволит повысить вычислительную скорость процедуры рендеринга геопространственного изображения в веб-ориентированной среде.

6. Анализ эффективности предложенных методологических основ, методов, алгоритмов на основе разработанного прототипа веб-ориентированного инструментально-программного средства обработки геомагнитной информации.

Объектом исследования выступают, таким образом, информационные системы поддержки принятия решений, ориентированные на обработку и анализ пространственно-зависимых данных, а **предметом** – методологические основы интеграции их гетерогенных источников, нормализации, физического хранения и визуальной интерпретации.

Методы исследования. В работе применяются методы интеграции и конгломерации массивов данных, статистического анализа случайных величин, машинного обучения с использованием размеченных данных, вычислительной математики, системного анализа, теория магнитного поля магнитосферных токов, элементы теории множеств и реляционной алгебры, теории реляционных, иерархических и колончатых баз данных, методы и алгоритмы научной визуализации, пространственного анализа, веб-программирования и создания аппаратно-зависимой компьютерной графики.

Научная новизна результатов заключается в следующем.

1. Методологические основы обработки неоднородной пространственно-временной информации отличаются тем, что, с целью повышения оперативности доступа к информации, необходимой для принятия решений, выделяются критерии пространственной зависимости и пространственной гетерогенности для групп источников данных и на этой основе подстраиваются процессы сбора, анализа и визуализации информации (пп. 4, 8 Паспорта специальности).

2. Комплекс моделей и методов обработки информации при интеграции гетерогенных источников данных в гибридные хранилище систем поддержки принятия решений отличается тем, что, с целью повышения вычислительной скорости сбора и обработки данных, преобразование оперативной информационной составляющей в постоянную определяется адаптированной моделью старения информации Бартона–Кеблера с исключенным динамическим компонентом (пп. 4, 13 Паспорта специальности).

3. Методы восстановления временных рядов данных (п. 4 Паспорта специальности), включающие:

– индуктивный метод, отличающийся тем, что, с целью повышения точности и скорости восстановления данных, наиболее вероятные значения определяются на базе статического сходства между массивом, образованным

предшествующими и последующими за пропущенным фрагментом значениями, и массивами, построенными аналогично из известных значений;

– метод информационного резервирования источников данных, отличающийся тем, что, с целью обеспечения полноты временных рядов, наиболее вероятные значения определяются посредством формирования доверительного списка на основании оценки пространственной гетерогенности и зависимости синхронно регистрируемых данных, а также сравнительной оценки фрагментов рядов, зарегистрированных в предшествующий момент.

4. Модель хранения данных отличается тем, что, с целью повышения реактивности программных средств и сокращения затрат физической памяти, реляционная, иерархическая и колончатая модели данных объединены на базе правил ссылочной целостности, а также комбинирования текстового и бинарного форматов описания как собственно данных, так и их метаданных (п. 12 Паспорта специальности).

5. Алгоритм визуализации пространственно-временного распределения геомагнитных данных отличается тем, что для клиентского веб-рендеринга больших пространственных данных учитывается их пространственная анизотропия посредством комбинирования подходов, демонстрирующих наилучшие показатели реактивности в соответствующих пространственных областях (п. 12 Паспорта специальности).

Достоверность полученных научных результатов обеспечивается строгостью применяемого математического аппарата, результатами математического и компьютерного моделирования, подтверждается при обработке и анализе оригинальных геомагнитных данных, предоставляемых территориально распределенными гетерогенными источниками на примере магнитных станций и обсерваторий международной сети INTERMAGNET, а также результатами обработки и визуализации геомагнитных данных в рамках разработанного прототипа веб-ориентированного инструментально-программного средства Geomagnet (https://www.geomagnet.ru). Предложенные математические модели, методы и алгоритмы обработки данных (на примере геомагнитной информации) исследованы на базе ФБУ «Государственный региональный центр стандартизации, метрологии и испытаний в Республике Башкортостан». Разработанный прототип веб-ориентированного инструментально-программного средства Geomagnet используется в «Центре прогнозов космической погоды» ФГБУН ИЗМИРАН, что подтверждается соответствующей ссылкой на информационном сайте организации (http://spaceweather.izmiran.ru/rus/links.html).

Теоретическая значимость результатов. В работе предложен подход к построению подсистем извлечения данных для информационных систем поддержки принятия решений (преимущественно ориентированных на данные), обобщающий результаты, полученные для отдельных моделей, методов и алгоритмов решения задач обработки, анализа и визуализации

пространственно-зависимых данных, и позволяющий получить качественно новые решения для некоторых из них. При этом, во-первых, предложены методологические основы обработки неоднородной пространственно-временной информации, необходимой для принятия решения, включающие в себя особенности теоретико-множественных, статистических и теоретико-информационных зависимостей как между элементами самих данных, так и их источниками. Во-вторых, предложенные модели и методы обработки информации при интеграции гетерогенных источников данных в системах поддержки принятия решений обеспечивают сокращение вычислительных затрат и повышение оперативности выполнения запросов к источникам данных, что имеет особенно важно значение в области техносферной безопасности. В-третьих, предложенный метод восстановления одномерных и многомерных временных рядов (на примере геомагнитных данных) обеспечивает высокую точность импутации пропущенных значений, что снимает ограничения на проведение амплитудно-частотного и спектрального видов анализов любых срезов геомагнитных данных независимо от интенсивности геомагнитной активности. В-четвертых, предложенная модель хранения данных в системах поддержки принятия решений позволяет формально описывать любые данные с пространственно-временной зависимостью сочетанием трех моделей данных и обеспечивает улучшение технических характеристик созданных на их основе систем. Наконец, представленный подход к веб-ориентированной визуализации пространственно-временного распределения данных позволил выйти на новый уровень эффективности решения конкретных прикладных задач, например, для геомагнитной информации – это учет параметров геомагнитного поля при проведении геофизических работ, проведении калибровочных испытаний магниточувствительной аппаратуры и пр.

Практическая значимость результатов. Полученные научные результаты были проанализированы на примере задач обеспечения техносферной безопасности и позволили достигнуть следующих показателей:

1. Согласно результатам проведенных экспериментов, вычислительная скорость получения информации о геомагнитной обстановке при определении месторасположения лаборатории по поверке магниточувствительной аппаратуры с использованием предложенных моделей и методов обработки информации при интеграции гетерогенных источников составила в среднем 32.4 % от времени, затрачиваемого при получении данных по практикуемым в настоящее время подходам.

2. При использовании предложенных методов восстановления временных рядов данных были с точностью с точностью от 0.01 до 0.5 нТл импутированы пропущенные фрагменты геомагнитных данных, необходимые для расчета специальных индексов геомагнитной активности, которые используются для прогнозирования уровня геоиндуцированных токов в линиях электропередач арктического региона.

3. В процессе принятия решений в случае возникновения опасности нарушения в работе систем автоматики высокоширотных железных дорог применение предложенной модели хранения данных обеспечивает повышение скорости обработки данных в ~ 4 раза (и, как результат, скорость принятия решений), а также сокращение вычислительных затрат на их хранение в ~5 раз по сравнению с известными подходами.

4. При использовании предложенного алгоритма визуализации пространственно-временного распределения данных возможно в веб-ориентированной среде (с увеличенной на 18 % скоростью клиентского рендеринга) оценить глобальную картину пространственного распределения параметров геомагнитного поля для выявления потенциально опасных для появления геоиндуцированных токов пространственных областей.

Реализация и внедрение результатов исследования. Результаты работы внедрены и использованы в Институте геологии УФИЦ РАН, НИИ ТС Пилот, АО Уфимское агрегатное производственное объединение, ООО «Алгоритм», учебном процессе ФГБОУ ВО Уфимский государственный авиационный технический университет, ФГБОУ ВО «Московский государственный технический университет гражданской авиации», ООО «Информационно-технологическая сервисная компания», ООО «Комплекс Проект».

Основные положения, выносимые на защиту.

1. Методологические основы обработки неоднородной пространственно-временной информации.
2. Комплекс моделей и методов обработки информации при интеграции гетерогенных источников данных в гибридные хранилище систем поддержки принятия решений.
3. Методы восстановления временных рядов данных на основе их статистической обработки, принципов машинного обучения и информационного резервирования.
4. Модель хранения данных, основанная на сочетании реляционного и нереляционного подходов.
5. Алгоритм визуализации пространственно-временного распределения геомагнитных данных на основе геоинформационных методов визуальной интерпретации информации.
6. Результаты анализа эффективности предложенных методологических основ, методов, алгоритмов, а также специального математического и алгоритмического обеспечения на основе разработанного прототипа веб-ориентированного инструментально-программного средства.

Апробация. Материалы диссертации докладывались и обсуждались на научных семинарах Геофизического центра РАН, British Geological Survey (BGS), а также на международных конференциях: 14th International Multidisciplinary Scientific Geoconference, Albena (Bulgaria), 2014; 16th International Conference on Computer Science and Information Technologies,

Sheffield (United Kingdom), 2014; 1st International Conference on Geographical Information Systems Theory, Applications and Management, Barcelona (Spain), 2015; 15th International Multidisciplinary Scientific Geoconference, Albena (Bulgaria), 2015; 1st International Conference on Computer Science and Computational Intelligence, Jakarta (Indonesia), 2015; 2nd International Conference on Geographical Information Systems Theory, Applications and Management, Rome (Italy), 2016, 17th International Multidisciplinary Scientific Geoconference, Albena(Bulgaria), 2017; Международный мультидисциплинарный симпозиум «Интеграция текущих исследовательских задач и решение глобальных вызовов», Ставрополь (Россия), 2018; VII Всероссийской научной конференции с международным участием «Информационные технологии интеллектуальной поддержки принятия решений», Уфа (Россия), 2019 и др.

Исследования выполнены в рамках грантов РФФИ №№ 13-07-00011, 14-07-00260, 14-07-31344, 15-07-02731, 15-17-20002, 16-07-00239, 16-17-20062, 19-07-00682, 20-07-00011, государственного задания № FEUE-2020-0007.

Публикации. По теме диссертации всего опубликовано более 100 работ, в том числе 33 статьях в рецензируемых журналах из списка ВАК; 22 статьях в изданиях, индексируемых международными системами Scopus / Web of Science; 3 монографиях, изданных в России и за рубежом; 2 патентах на изобретение; 5 свидетельствах о государственной регистрации программы для ЭВМ, трудах конференций и др.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Введение посвящено обоснованию актуальности, практической и теоретической ценности диссертационной работы. Здесь формулируются цель и задачи работы; представлены положения, выносимые на защиту; изложены краткая характеристика и сведения об апробации работы

Глава 1 посвящена анализу проблемы обработки информации из распределенных гетерогенных источников в процессе принятия решений и обсуждению предложенных методологических основ подхода к анализу и обработке пространственного распределения данных в информационных системах поддержки принятия решений.

Проведенный сравнительный анализ подходов к проектированию информационных систем поддержки принятия решений позволил выявить компонент, слабо зависящий от решаемых системой прикладных задач. Таким компонентом является подсистема извлечения данных, предназначенная главным образом для сбора и интеграции данных из множества источников. Современный экспоненциальный рост объемов данных во многих прикладных областях делает подсистему извлечения данных одним из наиболее уязвимых компонент систем поддержки принятия решения и создает условия информационной перегруженности для лиц, принимающих решения.

В информационных системах поддержки принятия решений, ориентированных на данные, проблема информационной перегруженности пользователей проявляется в наиболее явном виде.

Одной из прикладных областей, в которых оперативное получение достоверной информации имеет критически важное значение, является техносферная безопасность. Возможность непрерывного мониторинга данных из многих, зачастую разнородных источников, является залогом успешного предотвращения или снижения негативных эффектов со стороны окружающей среды на объекты и системы техносферы.

Среди факторов окружающей среды, воздействующих на техносферу, особо следует выделить геомагнитные возмущения, которые в зависимости от величины соответствующих параметров геомагнитного поля характеризуются негативным влиянием на целый ряд сложных технических объектов и систем.

Результаты анализа проблемной области представлены в виде интеллект-карты, раскрывающей необходимость решения четырех укрупненных задач, связанных соответственно с интеграцией источников данных, обработкой, хранением и визуализацией получаемых при этом данных. На примере геомагнитной информации CATWOE-характеристика информационной системы поддержки принятия решений позволила выделить такие ее основные компоненты, как, в первую очередь, исходные данные (результаты наблюдений параметров магнитного поля Земли) и результирующие данные (единое информационное пространство геомагнитных данных), связанные друг с другом процессом трансформации. Владельцами данных выступают при этом регистрирующие и распространяющие их магнитные обсерватории и вариационные станции, а потребителями – лица, принимающие решения в области техносферной безопасности. Здесь же выделены ограничения (внешние факторы) системы и нормативы, определяющие формализацию данных в ней (стандарты и спецификации по регистрации и расчету данных). Показано, что аналогичная характеристика может быть формализована в CATWOE-терминах и для других задач в области техносферной безопасности.

Проведенный анализ позволил обозначить несколько основных направлений исследований, результаты которых позволят достичь поставленной цели, а также критерии оценки их эффективности. Так, требуется разработка подхода к аналитической обработке пространственной зависимости данных в системах поддержки принятия решений, критерием эффективности которого является формализованное описание данных по выделенным критериям. Для интеграции гетерогенных распределенных источников данных необходимо разработать соответствующие модели и методы, эффективность которых предлагается оценить посредством анализа изменений вычислительной скорости обработки и анализа больших объемов данных. Далее на этапе предварительной обработки данных требуется обеспечить максимально возможное восстановление временных рядов с учетом внешних ограничений,

а оценить эффективность методов восстановления возможно расчетом имеющей при этом место погрешности. На следующем шаге данные необходимо сохранить в составе информационной системы, для чего требуется разработка моделей, эффективность которых можно оценить численными характеристиками повышения скорости обработки данных и сокращения вычислительных затрат на их хранение. Наконец, требуется разработка алгоритма веб-ориентированной визуализации данных как инструмента анализа данных в процессе принятия решений, критерием эффективности которого выступает численное изменение скорости клиентского рендеринга пространственного изображения.

Исследования показали, что во многих сферах деятельности требующие мониторинга внешние факторы отличаются пространственной зависимостью, анализ которой позволит получить знания, необходимые для эффективной поддержки принятия решений. Указанная характеристика проявляется в том числе и в характере представления пространственных данных как совокупности атрибутивной и пространственной информации. Предлагается порядок оценки пространственной зависимости данных: выделение локальных, региональных и глобальных групп источников данных; парный анализ пространственной зависимости взаимной информации источников данных; статистический анализ пространственно-временной зависимости в источниках данных.

На первом этапе все множество источников данных разделяется на группы, в зависимости от покрытия ими пространственной области: глобальные A^g, региональные A^r и локальные A^l:

$$A = \left\langle A^g, A^r, A^l \right\rangle, A^g = \left\langle a_1^g, ..., a_n^g \right\rangle,$$

$$A^r = \left\langle a_1^r, ..., a_n^r \right\rangle, A^l = \left\langle a_1^l, ..., a_n^l \right\rangle,$$

Тогда для множеств-элементов групп D^g, D^r, D^l:

$$\exists\, D^l = \left\langle d_1^l, ..., d_n^l \right\rangle \in a_i^l : a_i^l \in A^l,\ D^l \subset a_i^g \in A^g \parallel D^l \subset a_i^r \in A^r,$$

$$\exists\, D^r = \left\langle d_1^r, ..., d_n^r \right\rangle \in a_i^r : a_i^r \in A^r,\ D^r \subset a_i^g \in A^g,$$

$$\exists\, D^g = \left\langle d_1^g, ..., d_n^g \right\rangle \in a_i^l : a_i^l \in A^l,\ D^g \in a_i^r : a_i^r \in A^r \parallel\ D^g \in a_i^l : a_i^l \in A^r,$$

При этом общее число атомарных источников данных по формуле включений и исключений определяется как:

$$N = \left| D^g \cup D^r \cup D^l \right| = \left| D^g \right| + \left| D^r \right| + \left| D^l \right| - \left| D^g \cap D^r \right| - \left| D^g \cap D^l \right| - \left| D^r \cap D^l \right| + \left| D^g \cap D^r \cap D^l \right|.$$

Тогда исходя из пространственных запросов на получение данных при принятии решений целесообразно обращаться к одной или нескольким группам источников данных, имеющих соответствующую географическую привязку. Так, к примеру, в процессе поддержки принятия решений при выборе размещения лаборатории по поверке магниточувствительной аппаратуры могут

быть использованы данных тех источников, чьи группы соответствуют анализируемому на предмет геомагнитной обстановки полигону.

Регистрируемые и поступающие в источник данные $X = \{x_1,\ldots,x_n\}$ проходят k этапов обработки, в результате чего формируется множество $X^k = \left\{x_1^k,...,x_n^k\right\}$. При этом вектор состояния задается как $C_i = \left\langle X_1,k_1;...;X_n,k_n\right\rangle$, а множество всех возможных векторов состояний $C = \left\{C_i, i = \overline{1,|C|}\right\}$ составляет полное множество состояний данных, которое можно представить в виде: $|C| = \prod_i k_i$.

Состояния данных неравнозначны. Так, в случае геомагнитных данных для обработки предпочтительны окончательные и квазиокончательные геомагнитные данные. К примеру, для расчета специальных индексов, которые используются для прогнозирования уровня геоиндуцированных токов в линиях электропередач арктического региона, во избежание коллизий должны быть использованы только окончательные данные. Неоднозначность состояния может быть описана в виде функции ограничения, которая в общем случае представляет собой отображение полного множества состояний $f_0 : C \to P$, где P – некоторое заданное множество.

Функция ограничения f_0 для отображения в интервале наблюдения T множества моментов времени измерений, примененных на множестве наблюдаемых состояний C^: $f_0 : C\text{^} \to T,\ T \to C\text{^}$. Представленное отображение является однозначным, поскольку состояние данных в момент наблюдения детерминировано. Тогда функция ограничения примет вид:

$$t = t_i, C = Ci \Rightarrow f_0 = 1;\ t = t_i, C \neq Ci \Rightarrow f_0 = 0.$$

Поскольку $|C| \le |T|$, то используется функция возможностей, которая задается в виде: $W = \left[W_k, k = \left\{1,k\right\}\right]$: $W_k = N_k / \max N_k, i \in |C|$, где N_k – число наблюдаемых состояний C_k.

Множество данных может быть представлено совокупностью кортежей. Теоретико-множественное описание кортежа (на примере геомагнитных данных) может быть задано в виде (в соответствии с правилами расчета полного вектора магнитного поля Земли, Рисунок 1, где x_i – компоненты вектора поля):

$$\bigcup_k \left(x_1,x_2,..,x_7\right)^k = A,\ \forall k \neq 0\ \exists \left(x_1,x_2,..,x_7\right)^k :$$

$$x_7 = \sqrt{x_1^2 + x_2^2 + x_3^2} = \sqrt{x_4^2 + x_3^2},$$

$$x_4 = x_7 \times \cos x_6,\ x_3 = x_7 \times \sin x_6,$$

$$x_1 = x_4 \times \cos x_5,\ x_2 = x_4 \times \sin x_5.$$

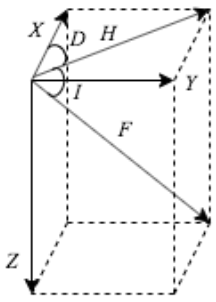


Рисунок 1 – Взаимосвязь компонент вектора магнитного поля Земли, выраженного геомагнитными данными

При этом проекция $\mathrm{pr}_{i1,\ldots,ik}$ множества геомагнитных данных D на оси $i_1,\ldots,i_n$ может быть определена следующим образом:

$$D = \left\{\left(t, x_1, x_2, ..., x_7, \mathrm{id}\right)_i\right\}_{i=1}^{n},\ \mathrm{pr}_{i_1,\ldots,i_k}(D) = \left\{\left(t, x_1, x_2, ..., x_7, \mathrm{id}\right)_i\right\}_{i=1}^{n},$$

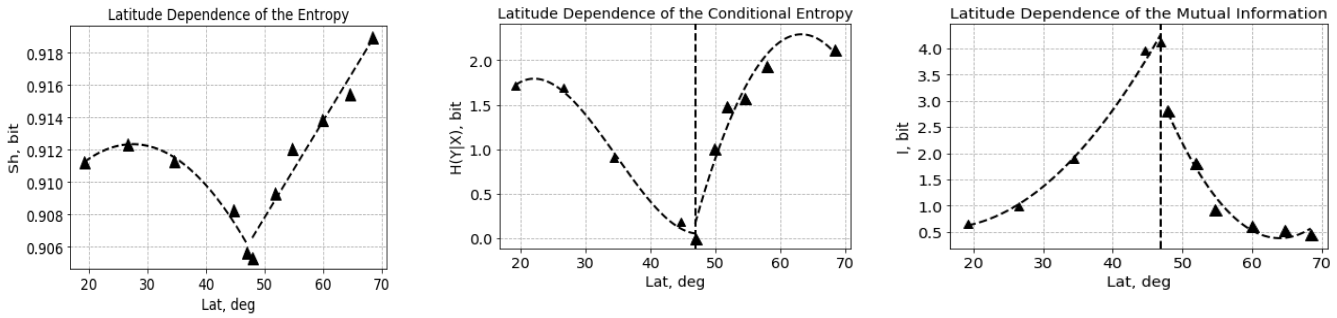
где n – количество источников данных, k – число осей для проекции.

Теоретико-множественные соотношения могут быть использованы для восстановления значений элементов кортежей, например, при анализе геомагнитной обстановки в процессе принятия решений при проведении инклинометрических работ.

На следующем шаге оценивается усредненная информативность источников данных – информационная энтропия Шеннона для x событий с n возможными состояниями вероятности $p(i)$ появления:

$$S(x) = -\sum_{i=1}^{n} p(i)\log_2 p(i) = \sum_{i=1}^{n} p(i)\log_2 \frac{1}{p(i)}$$

Необходима оценка трех энтропийных параметров: информационной энтропии, условной энтропии и взаимной энтропии (последние два параметра подразумевают попарное сравнение источников). На Рисунке 2 представлены результаты анализа этих параметров на примере источников геомагнитных данных. Так, расчет информационной энтропии Шеннона для каждого набора геомагнитных данных показал, что наименьшая неопределенность наблюдается в районе средних широт. Энтропия возрастает по направлению к высоким и низким широтам (Рисунок 2, *а*).



а *б* *в*

Рисунок 2 – Пространственная (широтная) зависимость теоретико-информационных характеристик источников данных (на примере геомагнитной информации): *а* – информационная энтропия, *б* – условная энтропия, *в* – взаимная информация

Физическое объяснение полученных результатов теоретико-информационного анализа заключается в том, что значения параметров геомагнитного поля и его вариаций на средних широтах зависят от наименьшего количества факторов. По мере приближения к экватору и полюсам число факторов, влияющих на результаты наблюдений, возрастает (к примеру, кольцевые токи на экваторе и суббури в полярных областях), что приводит

к росту неопределенности данных о состоянии геомагнитного поля, но повышает информативность каждой итерации мониторинга в процессе принятия решений, например, при анализе областей на предмет возможности возникновения критических ситуаций, связанных с наведением геоиндуцированных токов в системах автоматики железных дорог.

Известно, что одним из приложений информационной энтропии является оценка степени связности процессов. Здесь имеет место единственный источник процесса, т.е. геомагнитные данные в различных точках Земли демонстрируют общую составляющую на магнитометрах. При этом результаты долготной зависимости информационной энтропии тех же данных свидетельствуют о высокой связности процессов, определяющих регистрируемые значения на различных географических долготах и одном диапазоне широт.

Анализ также показал, что условная энтропия не зависит от пространственной привязки источника геомагнитных данных и увеличивается по мере удаления парного источника данных, поскольку синхронная регистрация геомагнитных данных фиксированным источником не влияет на снижение неопределенности в сравнении с достигнутой при этом энтропией в масштабах одного и того же временного интервала (Рисунок 2, *б*). Аналогичные результаты были достигнуты и при исследовании взаимной информации пар источников (Рисунок 2, *в*). Наилучшие результаты показали источники, разделенные в средним 1^{o} независимо от направления. Из этого можно сделать вывод о совместной встречаемости регистрируемых ими временных рядов, а в дальнейшем – об их взаимозаменяемости при заданном масштабировании.

Далее исследуются статистические характеристики используемых в информационной системе данных, главным образом, степень автокорреляции временных рядов. Так, для геомагнитных данных на основании анализа коррелограмм и теста Квенилле установлен максимальный лаг автокорреляции в среднем 8-9 значений. Проведение вычислительного эксперимента применительно к геомагнитным данным с привязкой к высоким географическим широтам показало изменение характера корреляционной связи между разделенными лагом значениями временного ряда: для высокоширотных магнитных обсерваторий максимальная длина лага для автокорреляции регистрируемых временных рядов уменьшается по сравнению со среднеширотными и приэкваториальными магнитными обсерваториями.

Важными аспектами при построении моделей геопространственных данных являются определение параметров пространственной неоднородности и пространственной зависимости, предполагающие наличие пространственной корреляции (положительной или отрицательной) между пространственными наблюдениями. Последняя может быть определена на основании расчета индекса Морана I_G, применяемого в геостатистике:

$$I_G = \frac{\sum_i \sum_j w_{ij}(x_i - \mu)(x_j - \mu)}{\sum_i (x_i - \mu)^2},$$

где x_i, x_j – значения параметра x в пространственных точках i и j, μ – среднее значение параметра x, w – экспертный весовой коэффициент. В зависимости от соотношения значений I_G и $I(E)$ (где для n точек $I(E) = -1 / (n - 1)$) возможно определить, являются ли значения в соседних пространственных регионах подобными.

Таким образом, предложенные методологические основы обработки неоднородной пространственно-временной информации в системах поддержки принятия решений позволяют: разделить источники данных на пространственные группы и привязать их к задачам принятия решений; оценить совместную встречаемость данных и оценить степень связности описываемых ими процессов. Результаты могут быть использованы при аналитической обработке данных, необходимых для принятий решений, в частности, при восстановлении используемых временных рядов.

Во второй главе обсуждаются модели и методы обработки данных при интеграции гетерогенных источников, целью реализации которых является повышение оперативности получения данных, необходимых для принятия решений, в частности, в области техносферной безопасности.

Одним из основных требований к информационному обеспечению систем поддержки принятия решений является возможность своевременного получения необходимых данных, что определяется как форматом их представления, так и вычислительной скоростью выполнения запросов. При этом ввиду неединственности источника данных задача получения информации усложняется необходимостью интеграции результатов запросов к ним. Одним из вариантов решения такой задачи является создание единого информационного пространства, которое представляет собой совокупность гетерогенных источников данных, а также инфокоммуникационных технологий их интеграции, обработки, анализа и визуализации, функционирующих на основе единых принципов и обеспечивающих информационное взаимодействие поставщиков и потребителей данных, равно как и удовлетворение их информационных потребностей при решении прикладных и научно-исследовательских задач (Рисунок 3).

Пусть единое информационное пространство использует множество из n территориально распределенных гетерогенных источников данных: $A = \left[A_1, ..., A_n\right]$, каждый из которых предоставляет множество кортежей данных виде $D^{A_i} = \left\{d_1^{A_i}, ..., d_m^{A_i}\right\}$. Все источники данных сгруппированы и ранжированы на

основании предложенных теоретических основ обработки по пространственной привязке, теоретико-информационным и статистическим характеристикам.

На длительных временных периодах (несколько лет) объем данных (на примере геомагнитной информации) увеличивается в соответствии с мальтусовской экспоненциальной моделью роста, предполагающей скорость роста функции пропорционально ее значению и выражаемой отношением: $V_t = V_{t_0} e^{rt}$, где V_t – объем геомагнитных данных информации в момент времени t, V_{t0} – объем информации в начальный момент времени t_0, r – мальтузианский параметр, t – время. При этом флуктуация каждого потока определяется как:

$$\sigma(t_n) = \sqrt{\frac{1}{n}\sum_{i=0}^{n}\left[y(t_i) - \left(y(t_0) + v(t_i - t_0)\right)\right]^2}.$$

К примеру, по результатам экспериментов, проведенных для годовых архивов геомагнитных данных доступных магнитных обсерваторий, было установлено, что флуктуация изменяется пропорционально квадратному корню от времени. Это свидетельствует о том, что каждый рассматриваемый при этом информационный процесс является процессом с независимыми приращениями. Это, в свою очередь, свидетельствует о сильной корреляционной связи между последовательными сообщениями в информационном потоке.

Основной идеей информационного пространства является объединение разнородных информационных потоков данных в централизованное хранилище под управлением единого метода доступа. Множество кортежей данных, доступных в едином информационном пространстве, можно обозначить как: $D = \{d_1,...,d_k\}$. При этом единое информационное пространство есть результат объединения кортежей данных из доступных источников:

$$D = \{d_1,...,d_k\} = \bigcup_{i=1}^{n} D^{A_i}.$$

Предлагается условное разделение на две части – стабильную (аналитические данные) D^a и динамическую (оперативные данные) D^o, которые имеют различные характеристики своего развития, где стабильная составляющая содержит так называемые архивные данные, предшествующие некоторому временному интервалу T, а динамическая – обновляемые данные за некоторый период T (например, T – текущие календарные сутки): $D = \{D^a, D^o\}$.

Формирование централизованного хранилища данных реализуется методом, являющимся комбинацией принципов консолидации (ETL, Extract-Transform-Load) и федерализации (EII, Enterprise Information Integration) данных, первый из которых реализует ведение стабильной, а второй – динамической составляющих информационного пространства (Рисунок 3):

$$\text{EII. } D^o = \bigcup_{i=1}^{n} \bigcup_{t=t_0}^{T} \left\{ d_1^{A_i}(t), ..., d_m^{A_i}(t) \right\} : \left\{ d_1^{A_i}, ..., d_m^{A_i} \right\} \subset D^{A_i}, \left\{ d_1^{A_i}, ..., d_m^{A_i} \right\} \not\subset D.$$

$$\text{ETL. } D^a = \bigcup_{i=1}^{n} \bigcup_{t=0}^{t_0} \left\{ d_1^{A_i}(t), ..., d_m^{A_i}(t) \right\} \xrightarrow{f} \left\{ d_1^a, ..., d_p^a \right\} : d_i^{A_s}(t) \xrightarrow{f} d_j^a;$$

$$\left\{ d_1^a, ..., d_p^a \right\} \subset D, \ \forall A^k \left\{ d_1^{A_i}, ..., d_m^{A_i} \right\} \not\subset D^{A_k}.$$

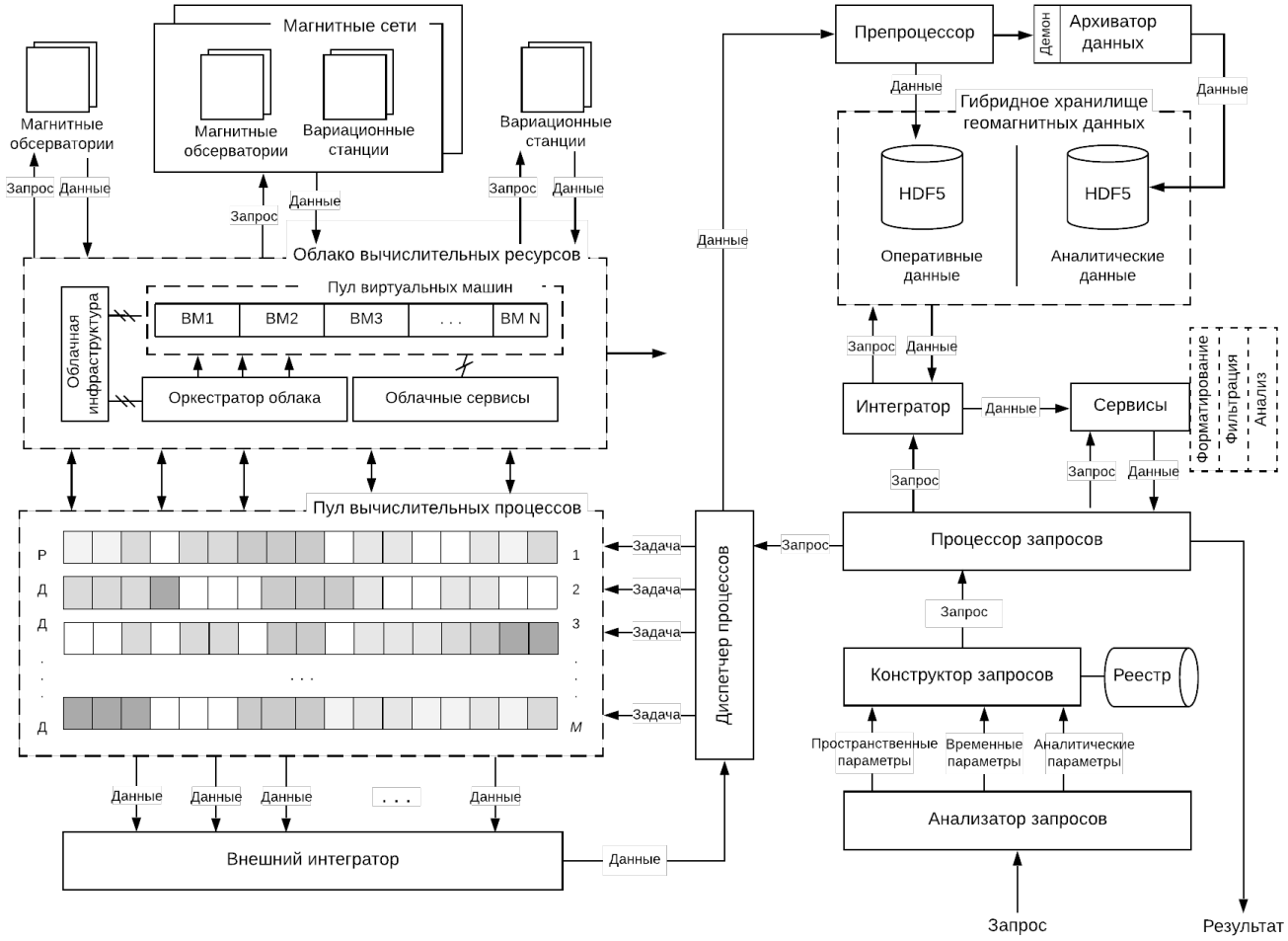


Рисунок 3 – Концептуальная схема единого информационного пространства на основе разнородных территориально распределенных источников (на примере геомагнитной информации)

Процесс трансформации оперативных данных в архивные задается отображением вида $f : D^o \to D^a$, является триггерным (событием считается завершение временного периода длительностью T) и предлагается представлять в терминах старения информации на основе адаптированной модели Бартона-Кеблера с исключенным динамическим компонентом:

$m(t) = 1 - ae^{-T} - be^{-2T} \Rightarrow 1 - ae^{-T}$, где $m(t)$ – часть оперативной информации в общем потоке через время T, ae^{-T} – архивные данные (нормированный объем в долях от единицы, единица – полный объем данных), be^{-2T} – данные, полученные за период времени T (нормированный объем в долях от единицы, единица – полный объем данных), но не помещаемые в основной объем данных.

Аналитические данные в предложенной архитектуре являются физически хранимыми, в отличие от оперативных данных, заданных только виртуальным представлением. Заполнение стабильной компоненты информационного пространства осуществляется вычислительными процессами-демонами, выполняемыми по принципу Cron. Один из компонентов архитектуры единого пространства данных – процессор запросов – отвечает за формирование результирующего набора данных для пользователя. Программные интерпретаторы, получив управление от процессора запросов, обращаются к гибридному хранилищу и извлекают из него искомые данные, подключая интегратор для объединения результатов. Высокая вычислительная скорость обращения к распределенным источникам данных обеспечивается облачной инфраструктурой, обслуживающей ряд виртуальных машин, каждой из которых выделяется один или несколько вычислительных процессов по получению и обработке массивов данных.

Таким образом, предлагаемые модели и методы обработки информации при реализации предполагают интеграцию данных из разнородных источников и позволяют ускорить доступ пользователей к данным без необходимости предварительной загрузки, обработки и фильтрации из больших наборов.

Третья глава посвящена обсуждению предложенных методов восстановления временных рядов данных (на примере геомагнитной информации) в информационных системах на основе принципов машинного обучения и информационного резервирования источников информации, обеспечивающих возможность импутации пропусков с ошибкой в пределах допустимого нормативами отклонения.

Одной из проблем на пути автоматической обработки и визуализации данных в едином информационном пространстве является наличие в поступающих от источников данных временных рядах многочисленных пропусков и выбросов. При этом особенности пространственной гетерогенности, а также взаимные энтропийные характеристики источников данных, задействованных в принятии решений, позволяют расширить возможности известных методов импутации пропущенных значений. Основным элементом таких решений является ранжирование и агрегирование источников данных по результатам анализа пространственной зависимости теоретико-множественных, теоретико-информационных и статистических характеристик. Выявленные признаки совместной встречаемости временных рядов данных и степень связности описываемых ими процессов служат основанием для

формирования доверительного списка для каждого задействованного в едином информационном пространстве источника данных. Например, для геомагнитных данных в доверительный список попадают обсерватории с близкими к 1 показателями взаимной информации и условной энтропии, со значениями информационной энтропии, отличающимися не более, чем на 0.01, а также относящиеся к одной пространственной группе и описываемые одним статистическим законом распределения.

Проведенные вычислительные эксперименты показали (на примере геомагнитной информации), что подобные доверительные списки могут быть составлены не для всех задействованных в едином информационном пространстве источников данных. Так, при оценке геомагнитной обстановки в пространственном регионе с заданными границами для определения возможности калибровки магниточувствительной аппаратуры в полярных областях крайне неравномерное распределение магнитных обсерваторий препятствует использованию метода информационного резервирования.

Однако для тех источников, для которых результаты предварительного анализа данных показали лучшие результаты, применим предложенный метод восстановления временных рядов, основанный на принципах информационного резервирования. В этом случае попавшие в один доверительный список источники данных считаются резервирующими друг друга. На примере геомагнитной информации исследования, показали, что для отсутствующего фрагмента временного ряда обсерватории наиболее вероятными значениями являются уровни фрагмента временного ряда (с масштабированием), синхронно зарегистрированного ближайшей обсерваторией из списка.

Для источников данных, для которых невозможно использовать метод информационного резервирования, предложен метод импутации пропусков во временных рядах, в основе которого лежат принципы машинного обучения с использованием размеченных данных. Метод получил название индуктивного. Его идея заключается в предположении, что если пара непоследовательных фрагментов ряда, разделенных отсутствующим фрагментом, оказываются близкими к паре фрагментов, разделенных известным фрагментом, то промежуточные между ними значения в соответствии с теоремой Такенса будут отличаться статистически незначительно. При этом признаковым описанием фрагмента временного ряда выступает пара предшествующего и следующего за ним фрагментов того же ряда, образующих обучающую выборку для поиска недостающего фрагмента по набору его признаков с последующим линейным масштабированием для восстановления исходного тренда информационного сигнала (Рисунок 4). Основанием для применения индуктивного метода восстановления временных рядов являются их статистические характеристики, в частности, показатели степени автокорреляции и максимальное количество допустимых при этом лагов. Последнее определяет максимально возможную длительность восстанавливаемого фрагмента временного ряда.

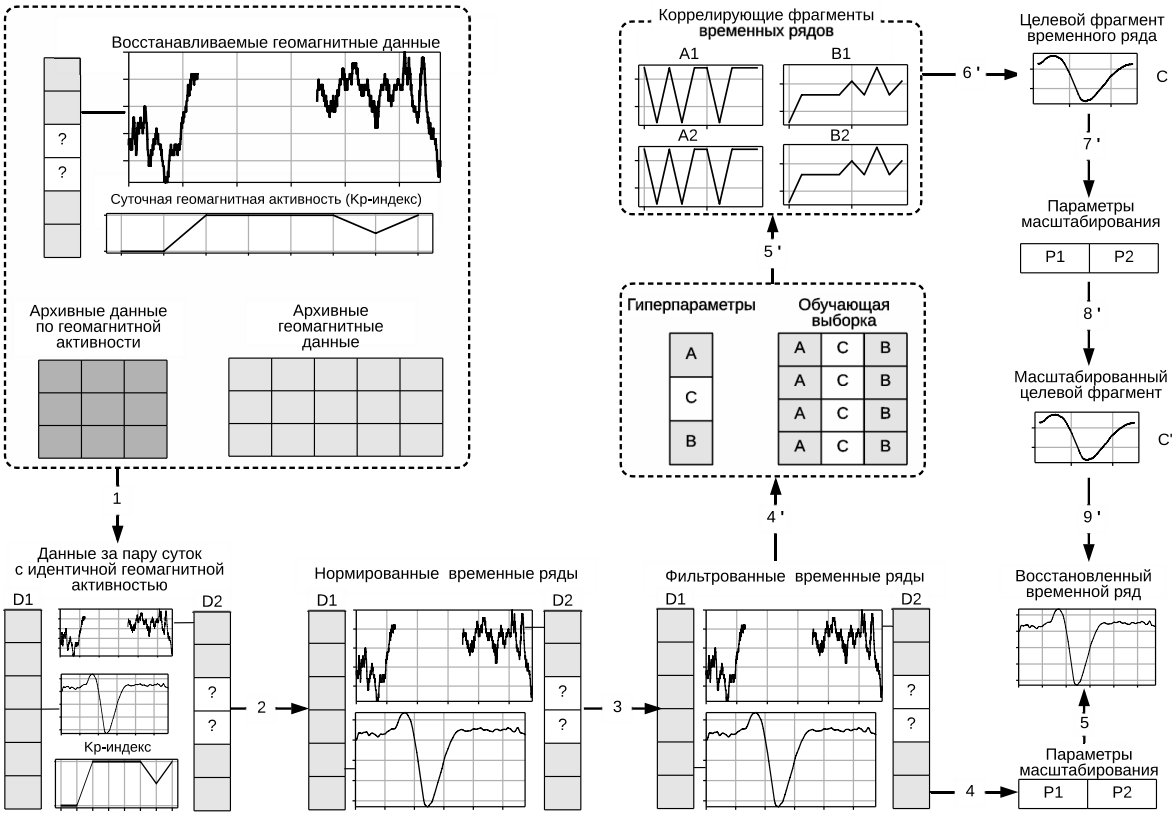


Рисунок 4 – Схема индуктивного метода восстановления временных рядов

Пусть задан временной ряд $y(t)$ суточных наблюдений $y(t_1)$, $y(t_2)$,…$y(t_n)$ параметра геомагнитного поля, полученных в последовательные моменты времени: $y(t) = \{(y_1(t),...y_k(t),...,y_l(t),...,y_m(t)\}$, где $\{y_1(t),\ldots, y_{k\text{-}1}(t)\}$, $\{y_{l\,+\,1}(t),\ldots, y_m(t)\}$ – наблюдаемые значения уровней временного ряда; $\{y_k(t),\ldots, y_l(t)\}$ – пропущенные значения уровней временного ряда. Тогда шаблон поиска T состоит из трех последовательных фрагментов временного ряда, один из которых (T_2) представлен набором пропущенных значений, а два других – фрагментами, предшествующим (T_1) и следующим (T_3) за первым и равными ему по размерности:

$$T = \left\{y_{2k-l}(t),...,y_k(t),...,y_l(t),...,y_{2l-k}(t)\right\} = \left\{T_1,T_2,T_3\right\};$$

$$T_1 = \left\{y_{2k-l}(t),...,y_k(t)\right\}, T_2 = \left\{y_k(t),...,y_l(t)\right\}, T_3 = \left\{y_l(t),...,y_{2l-k}(t)\right\},$$

При этом пространство объектов X обучающей выборки задано множеством пар непоследовательных фрагментов временного ряда, разделенных набором значений, число которых равно размерности пропущенного фрагмента исследуемого временного ряда: $X = (a_i,b_i)_{i=1}^{l}$, где l – размер обучающей выборки, a, b – фрагменты временного ряда.

Значимой проблемой реализации индуктивного метода является большой объем данных, который используется для формирования обучающей выборки

и последующего перебора пар ее значений. На примере геомагнитной информации для повышения вычислительной скорости метода восстановления предлагается сократить объем обучающей выборки за счет использования временных рядов данных, зарегистрированных магнитной обсерваторией в сутки с магнитной активностью, идентичной наблюдаемой в исследуемые сутки. Так, сутки, в течение которых наблюдалась та же геомагнитная активность, что и в восстанавливаемые, могут быть определены как:

$$j : K^j = \left(k_k^j\right)_{k=1}^n \in \left\{K^i\right\}_{i=1}^l,\ \left(k_k^j\right)_{k=1}^n = \left(k_k\right)_{k=1}^n,\ K = \left(k_k\right)_{k=1}^n.$$

Тогда исходными данными для формирования обучающей выборки метода выступают результаты наблюдений, полученные магнитной обсерваторией в *j*-е сутки: $y^j(t) = \{(y^j(t_1),...,y^j(t_m)\}$, где $\{y^j(t_1),\ldots,\ y^j(t_m)\}$ – минутные значения уровней временного ряда $y_f^j(t)$, m – количество минутных значений в сутки (m = 1440). Также для исключения неоднозначности результатов сравнения данных из полученных временных рядов исключается низкочастотная составляющая применением фильтра низких частот (ФНЧ) Баттерворта, обладающего гладкой амплитудно-частотной характеристикой как в полосе пропускания, так и в полосе задержки.

Для простоты пара переменных каждого элемента из пространства объектов интегрирована в один, что достигается структурным сдвигом соответствующих фрагментов временного ряда и формированием новых массивов значений уровня: $X = \left\{x_i\right\} : x_i = a_i \to b_i\ R,\ i = \overline{1,l}$, где x_i – экземпляр пространства объектов, a_i, b_i – исходные фрагменты временного ряда, R – квантор сдвига фрагмента временного ряда вправо.

Пространство объектов Y обучающей выборки задано множеством фрагментов временного ряда (целевых переменных), образованных значениями уровней, число которых равно количеству пропущенных значений в исследуемом временном ряду: $Y = \left\{y_i\right\},\ i = \overline{1,l}.$ Сравнение объектов обучающей выборки с шаблоном поиска предполагает нормализацию последнего путем устранения фрагмента T_2, структурного сдвига фрагментов T_1 и T_3 временного ряда и формирования нового массива значений уровня T_0:

$$T_0 = T1 \cup T3 = \left\{y(t_{2k-l}),...y(t_k), y(t_l),...,y(t_{2l-k})\right\}.$$

Мера близости между фрагментом T_0, составленным из предшествующих и последующих за отсутствующим фрагментом значений, и фрагментами из обучающей выборки вычисляется на основе коэффициента корреляции. Наиболее близким к шаблону поиска принимается тот фрагмент x_t пространства объектов X, которому соответствует максимальное значение коэффициента

корреляции. Для аппроксимации данных применительно к восстанавливаемым рядам выполняется их нормализация с помощью метода наименьших квадратов.

Восстановленные одним из предложенных способов данные размещаются в едином информационном пространстве и могут быть использованы для обработки, анализа и визуализации в процессе принятия решений.

В **четвертой главе** обсуждаются особенности предложенной модели хранения данных в системах поддержки принятия решений, обеспечивающего сокращение вычислительных затрат на их физическое размещение и повышение скорости обработки запросов к данным.

Предложена модель хранения данных, представленная совокупностью трех компонент и отличающаяся тем, что использует правила ссылочной целостности для объединения реляционной, иерархической и колончатой моделей данных, применяемых для описания метаданных и непосредственно данных, а также реализует комбинацию текстового и бинарного форматов представления информации с целью повышения реактивности программных средств обработки данных, с одной стороны, и сокращения затрат требуемого объема физической памяти, с другой (Рисунок 5). Предлагаемый формат используется для представления данных в гибридном хранилище в составе описанного выше единого информационного пространства.

Раздел метаданных, используемый для хранения общей информации об источниках данных (название, идентификатор и пр.), представляет собой набор связанных реляционных таблиц, где родительские сущности задаются в виде $R_p(A_1,\ldots, A_n) = \{A_1,\ldots, A_n\}$ ($A_1,\ldots, A_n$ – имена атрибутов сущности), а дочерние – в виде n-арных отношений типа: $R_c(A_1,\ldots A_n) = \left\{A_i\right\}_{i=1}^{n}$. Тогда кортеж-отображение r_c схемы R_c на домен значений D_c:

$$r_c = \left\{t_1,\ldots,t_p\right\} : R_c\{A_i\}_{i=1}^{m} \to D_c\{D_{c_i}\}_{i=1}^{m} : t_k(A_i) \in D_i,\ \forall k = \overline{1,n};\ \forall i = \overline{1,m}.$$

Обобщенная временная составляющая данных может быть представлена посредством иерархической XML-структуры, позволяющей хранить результаты наблюдений, сгруппированных по календарным годам и датам. Тогда для каждого источника данных создается XML-документ O с тремя составляющими, представленными множествами лет Y, месяцев M и дней D:

$$O = \left\{Y\right\}, Y = \left\{y_1,\ldots,y_n\right\} : \forall y \in Y\ \exists M \subset y, M = \left\{m_1,\ldots,m_l\right\}, l \le 12;$$

$$\forall m \in M\ \exists d \subset m, D = \left\{d_1,\ldots,d_p\right\}, p \le 31.$$

Непосредственно данные рекомендуется хранить в сжатом иерархическом формате типа Parquet, обеспечивающем оптимизацию хранения и доступа к данным большого объема. В соответствии с требованиями иерархического формата данные по каждому источнику данных разбиваются на страницы P (Page). Далее каждый параметр выделяется в один столбец C (Column), где представлены значения Ch (Chunk) и метаданные столбца M (Metadata):

$$P = \{C\}, C = \{c_1, ..., c_n\} : \forall c = \{Ch, M\}.$$

Предлагаемая модель хранения данных подразумевает смешанную логическую и физическую интеграцию предложенных выше компонент (Рисунок 5). Образуется иерархия структур данных, корневым элементом которой выступает реляционная структура с метаданными источников данных. Непосредственно данные физически размещаются в каталоге, в котором каждому источнику данных выделен XML-документ с именем, содержащим уникальный код. В составе XML-документа соответствующий году данных элемент содержит блок CDATA, где размещается набор данных в бинарном формате Parquet. При этом анализатор запросов в архитектуре единого пространства геомагнитных данных обеспечивает проверку ссылочной целостности как по уникальному коду, так и по указанным временным меткам.

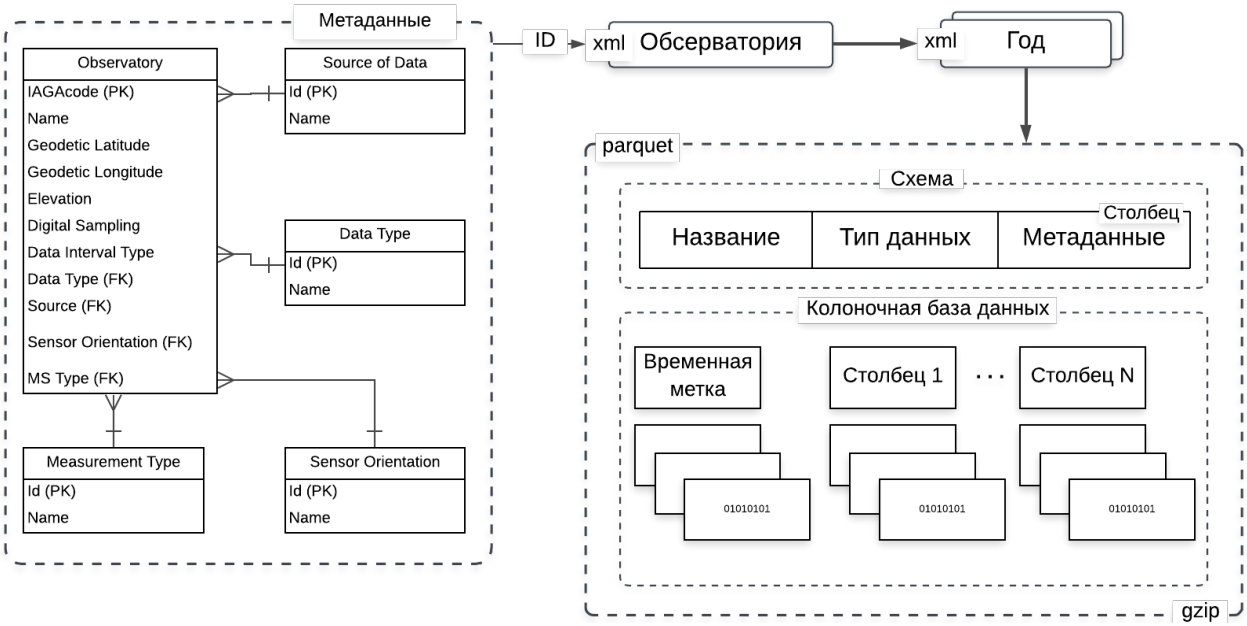


Рисунок 5 – Модель хранения данных в информационной системе поддержки принятия решений (на примере геомагнитной информации)

Взаимодействие с хранилищем данных осуществляется в соответствии с иерархической структурой. По коду из реляционной структуры выгружаются метаданные. Далее тот же код используется для обращения к XML-файлу источника данных, а оттуда посредством XPath-запроса выбирается секция CDATA с искомыми данными. При необходимости выполняется фильтрация, группирование и агрегирование результатов наблюдений с использованием языка запросов SQL.

Ожидается, что предложенная модель хранения данных обеспечит как сокращение вычислительных затрат при физическом хранении данных, так и повышение вычислительной скорости обработки запросов к ним.

В **пятой главе** на примере геомагнитных данных рассматривается алгоритм визуализации пространственно-временного распределения данных на основе компьютерных методов обработки информации.

В рамках задачи одним из значимых вопросов является разработка подхода для локального анализа пространственного распределения данных в заданном регионе произвольной формы. Предлагается метод построения локальной координатной сетки, суть которого заключается в том, что ограниченное регионом пространство заполняется пространственными объектами, разделенные равными интервалами и описанные вдоль собственных границ упорядоченным множеством размещенных с одинаковым разрешением пространственных точек. В результате метод локального анализа пространственного распределения данных сводится к выделению региона некоторой формы, построению локальной координатной сетки и расчету / отображению значений данных в каждой ее точке.

Предложен алгоритм визуализации пространственно-временного распределения геомагнитных данных в двух- и трехмерном режимах в веб-среде (Рисунок 6). Такой способ представления данных необходим для того, чтобы на основании данных из единого информационного пространства оценить распределение параметров геомагнитного поля и выделить опасные для наведения геоиндуцированных токов регионы.

Основным препятствием на пути веб-визуализации больших объемов геомагнитных данных являются высокие вычислительные затраты. Предлагается усовершенствовать известные подходы к визуальному представлению геопространственных данных применительно к рендерингу геомагнитных данных посредством разработки адаптивного алгоритма, формирующего изображение на основании пространственной привязки данных.

Установлено, что наименьшая пространственная анизотропия геомагнитных данных характерна для приэкваториальных и среднеширотных областей земной поверхности, в отличие от полярных областей, где наблюдается крайняя неоднородность данных, сопоставимая с изменениями геомагнитной обстановки в периоды высокой геомагнитной активности. Результаты экспериментов показали, что наилучшие результаты (по скорости рендеринга в веб-среде) демонстрируют метод сканирования применительно к полярным областям и метод шагающих квадратов для визуализации геомагнитных данных в остальных областях (Рисунок 6).

Предложенный алгоритм разделяет всю координатную решетку на области: полярные, приэкваториальные, среднеширотные. Для них формируются новые регулярные координатные решетки, каждая из которых представляет собой множество пространственных точек, являющееся подмножеством исходного множества точек, описывающих координатную сетку для всей земной поверхности. Далее в зависимости от типа выделенной области к координатной решетке применяется один из методов расчета и

визуализации линий уровня. Результаты выполнения по каждой из областей интегрируются для распределения значений по всей земной поверхности.

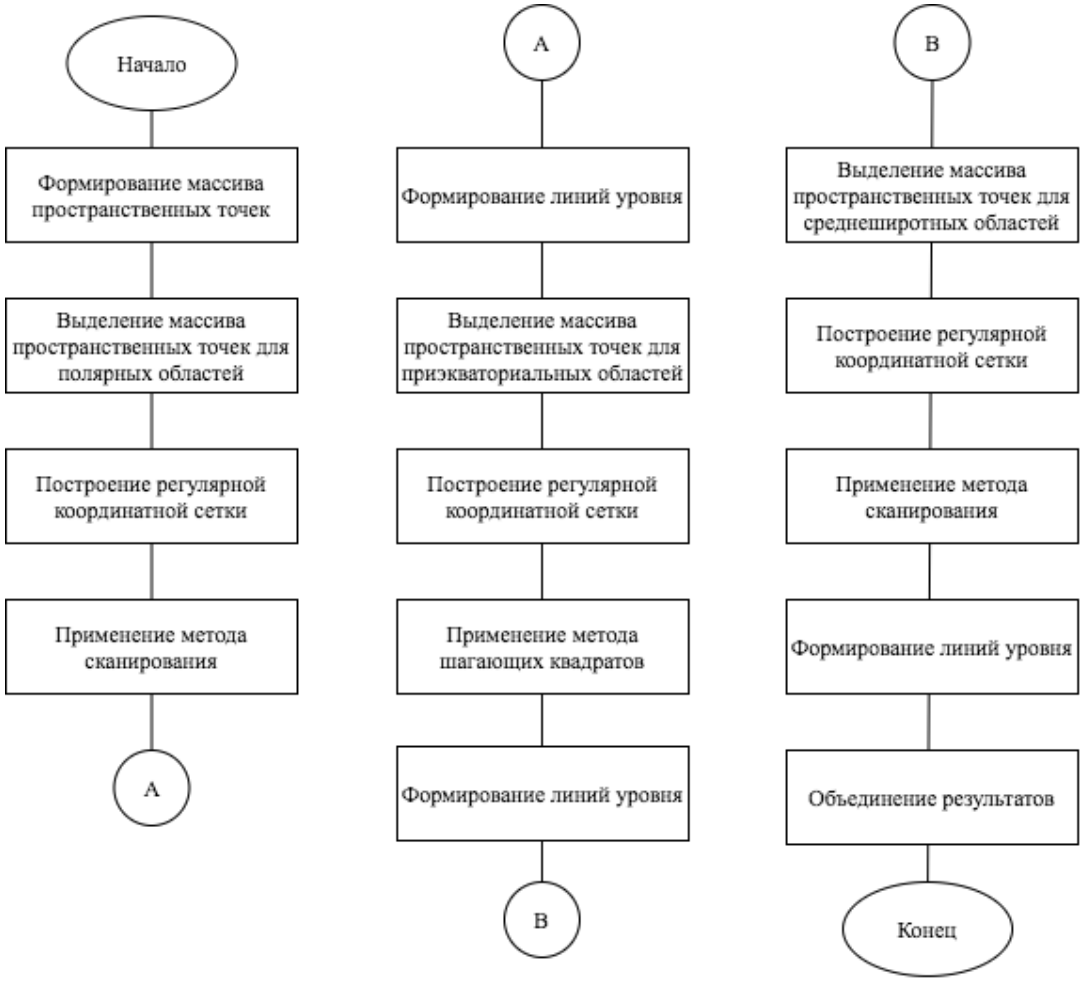


Рисунок 6 – Алгоритм применения гибридного метода визуализации распределения геомагнитных данных

В шестой главе приводятся результаты эффективности предложенных теоретических основ, методов, алгоритмов, а также специального математического и алгоритмического обеспечения на основе разработанного прототипа веб-ориентированного инструментально-программного средства и выделенных показателей эффективности.

Для оценки работоспособности предложенных решений на основе геоинформационных и веб-технологий по архитектуре MVC разработан исследовательский прототип информационной системы GEOMAGNET (https://www.geomagnet.ru), обеспечивающий обработку, анализ и визуализацию геомагнитных данных из разнородных распределенных источников (Рисунок 7).

Преимущества решений показаны на примере ряда задач поддержки принятия решений на основе геомагнитной информации, предоставляемой распределенными гетерогенными источниками геомагнитных данных. Одна из

них – предупреждение критических ситуаций, связанных с возникновением геоиндуцированных токов, что реализуется операторами, обслуживающими системы электроэнергоснабжения (Рисунок 8).

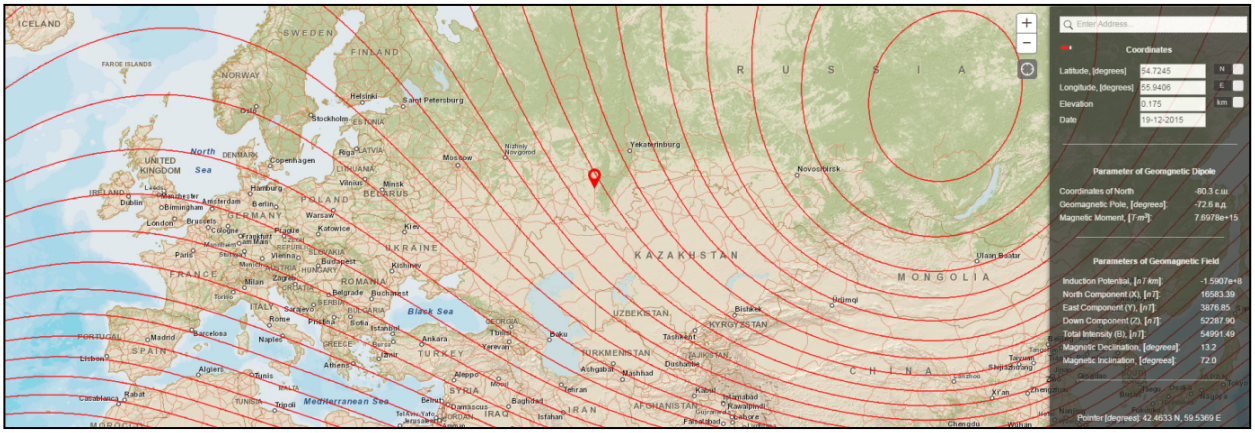


Рисунок 7 – Пример экранной формы исследовательского прототипа информационной системы

Применение предложенных решений позволит повысить оперативность принимаемых при управлении системой электроэнергоснабжения решений за счет предоставления единого доступа к геомагнитным данным. При этом лицо, принимающее решения, получает данные в уже обработанном виде (нормализованными и восстановленными с заданным шагом дискретизации и в заданных единицах измерения), что позволяет применить к ним аналитические инструменты, спрогнозировать появление геоиндуцированных токов и принять решение о проведении мероприятий в электроэнергетической системе.

Аналогичные преимущества могут быть использованы в системах поддержки принятия решений при проведении буровых работ с помощью магнитометрических инклинометров, а также при организации поверки магниточувствительного оборудования.

По существу исследование эффективности предложенных решений по выделенным ранее критериям показало следующие результаты.

а) Оценка эффективности предложенных методологических основ подхода к анализу пространственного распределения данных, а также моделей и методов обработки информации при интеграции гетерогенных источников данных в системах поддержки принятия решений выполнена на базе запроса к суточным данным (в качестве примера была взята дата 12.12.2016) к трем магнитным обсерваториям, принадлежащим различным магнитным сетям (INTERMAGNET и IMAGE), необходимым для предварительной оценки геомагнитной обстановки в пространственных точках, где планируется проведение поверки и калибровки магниточувствительной аппаратуры. Вычислительный эксперимент был проведен в два этапа.

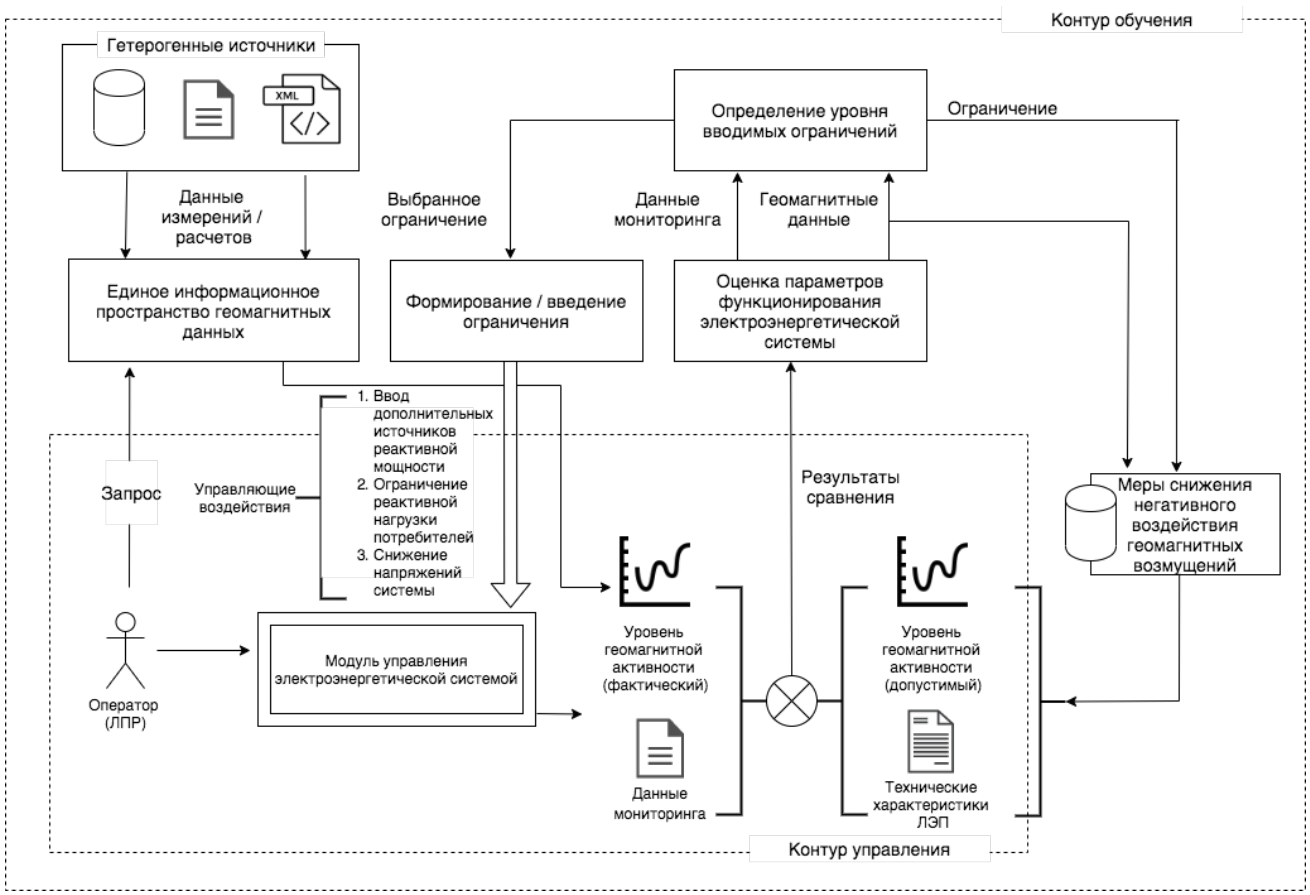


Рисунок 8 – Схема поддержки принятия решений при управлении системами электроэнергоснабжения на основе предложенных методологических основ обработки неоднородной пространственно-временной информации

На первом из них запрос был выполнен непосредственным обращением к источникам данных, доступным посредством веб-сервисов. На втором этапе запрос на получение тех же данных был выполнен с использованием разработанного исследовательского прототипа информационной системы. Анализ результатов эксперимента показал, что выполнение запроса к данным в условиях применения персонального компьютера со средней производительностью (процессор с частотой 1.6 ГГц, 2 ядра, оперативная память 4 Гб, скорость интернет-соединения 342.7 Мбит/с) в соответствии с предложенными решениями занимает около 54 с, что меньше времени отклика при использовании практикуемого способа, составляющего 158 с (~34.2 % от исходного времени).

б) При оценке эффективности методов восстановления временных рядов данных в информационных системах для минимизации влияния на результат сторонних факторов исследования были проведены на геомагнитных данных, полученных в среднеширотной обсерватории Grocka (GCK) в условиях спокойной магнитосферы. Пропущенный фрагмент представлен 10 последовательными значениями полного вектора геомагнитного поля за период

11.00-11.10. Восстановление указанного фрагмента предлагаемым методом обеспечивает значение среднеквадратической ошибки, равное 0,006 нТл, что меньше предельно допустимой стандартами погрешности измерений (1 нТл). Форма восстановленного сигнала также незначительно отличается от исходной, что свидетельствует о целесообразности применения метода (Рисунок 9). Аналогичные результаты получены в результате восстановления данных магнитной обсерватории, данные который использованы для расчета специализированных индексов для прогнозирования геоиндуцированных токов.

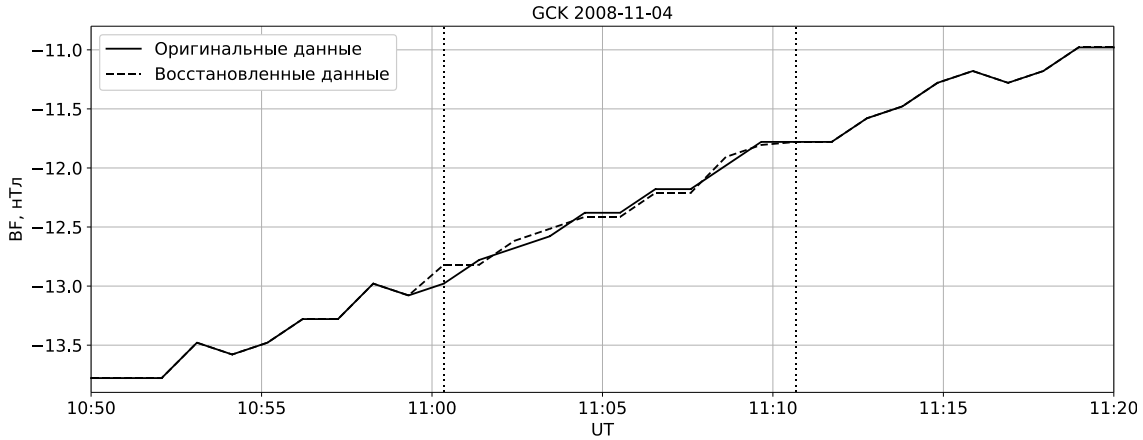


Рисунок 9 – Результаты восстановления 10-минутного фрагмента временного ряда геомагнитных данных

в) Оценка эффективности предложенной модели хранения данных выполнена на основании сравнительного анализа распространенных форматов данных. Критериями оценки эффективности модели хранения данных определены реактивность программной обработки данных и объем требуемого для их размещения дискового пространства. Исследование выполнено на примере получения выборки из годового архива минутных наблюдений станции с IAGA-кодом BOX за период 01-06.03.2018.

Экспериментальные исследования показали, что минимальное время отклика программного сценария обработки геомагнитных данных достигается при использовании для их хранения предложенной модели хранения данных (2.2 с), что примерно в 4,3 раза меньше, чем для IAGA2002/CSV. Пропорционально сокращается время, необходимое для принятия решения по геомагнитным данным для предотвращения опасной ситуации ввиду наведения геоиндуцированных токов на системы автоматики железных дорог.

Применение предложенной модели для хранения геомагнитных данных позволяет сократить требования к объему дискового пространства. Так, по сравнению с IAGA2002/CSV, для хранения годового архива наблюдений одной станции требуется примерно в 5,2 раза меньше объема дискового пространства.

г) Оценка эффективности предложенного алгоритма визуализации пространственно-временного распределения геомагнитных данных выполнена на основании сравнительного анализа известных методов построения пространственных изолиний. Критерием оценки эффективности определена скорость рендеринга при полной визуализации геомагнитных данных.

Исследование выполнено на примере визуализации выборки геомагнитных данных, рассчитанных на основе предложенного выше метода для каждой точки земной поверхности с шагом в 1^{o} методами триангуляции Делоне, шагающих квадратов, сканирования, а также предложенным гибридным алгоритмом. Результаты эксперимента показали, что применение предложенного алгоритма визуализации геомагнитных данных позволяет повысить скорость процедуры рендеринга геопространственного изображения в веб-ориентированной среде в среднем на 18% по сравнению с аналогами. При использовании существующих подходов допустима только локальная визуализация, что существенно усложняет процедуру принятия решений.

В **заключении** приводятся основные результаты и выводы по проведенной работе.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

В ходе диссертационного исследования были получены следующие результаты:

1. Методологические основы обработки неоднородной пространственно-временной информации на основе теоретико-множественного, теоретико-информационного и статистического подходов. Отличаются тем, что, с целью повышения оперативности доступа к информации, необходимой для принятия решений, выделяются критерии пространственной зависимости и пространственной гетерогенности для групп источников данных и на этой основе подстраиваются процессы сбора, анализа и визуализации информации.

2. Комплекс моделей и методов обработки информации при интеграции гетерогенных источников данных в гибридные хранилище систем поддержки принятия решений на основе принципов консолидации и федерализации. Отличается тем, что, с целью повышения вычислительной скорости сбора и обработки данных, преобразование оперативной информационной составляющей в постоянную определяется адаптированной моделью старения информации Бартона–Кеблера с исключенным динамическим компонентом.

3. Методы восстановления временных рядов данных на основе их статистической обработки, принципов машинного обучения и информационного резервирования, включающие:

– индуктивный метод, отличающийся тем, что, с целью повышения точности и скорости восстановления данных, наиболее вероятные значения определяются на базе статического сходства между массивом, образованным

предшествующими и последующими за пропущенным фрагментом значениями, и массивами, построенными аналогично из известных значений;

– метод информационного резервирования источников данных, отличающийся тем, что, с целью обеспечения полноты временных рядов, наиболее вероятные значения определяются посредством формирования доверительного списка на основании оценки пространственной гетерогенности и зависимости синхронно регистрируемых данных, а также сравнительной оценки фрагментов рядов, зарегистрированных в предшествующий момент.

4. Модель хранения данных, основанная на сочетании реляционного и нереляционного подходов. Отличается тем, что, с целью повышения реактивности программных средств и сокращения затрат физической памяти, реляционная, иерархическая и колончатая модели данных объединены на базе правил ссылочной целостности, а также комбинирования текстового и бинарного форматов описания как собственно данных, так и их метаданных.

5. Алгоритм визуализации пространственно-временного распределения геомагнитных данных на основе геоинформационных методов визуальной интерпретации информации. Отличается тем, что для клиентского веб-рендеринга больших пространственных данных учитывается их пространственная анизотропия посредством комбинирования подходов, демонстрирующих наилучшие показатели реактивности в соответствующих пространственных областях.

6. Прототип веб-ориентированной информационной системы, реализующий предложенные концепцию, модели, методы, алгоритмы применительно к геомагнитным данным, в ходе экспериментальных исследований с которым установлено, что:

– методологические основы обработки неоднородных пространственно-временных данных, а также модели и методы обработки информации при интеграции гетерогенных источников данных позволяют повысить скорость сбора и обработки больших геомагнитных данных в среднем в 2,9 раза;

– методы восстановления временных рядов позволяют реконструировать пропущенные фрагменты геомагнитных данных с точностью 0,01–0,5 нТл, в зависимости от их длительности, геопространственной привязки и геомагнитной активности;

– модель хранения данных обеспечивает в условиях ограниченных вычислительных ресурсов повышение скорости геомагнитных данных в ~ 4 раза и сокращение вычислительных затрат на их хранение в ~5 раз по сравнению с известными подходами;

– алгоритм визуализации пространственно-временного распределения данных позволяет повысить скорость рендеринга геопространственного изображения в веб-ориентированной среде в среднем на 18% по сравнению с известными аналогами.

ОСНОВНЫЕ ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

В рецензируемых изданиях из перечня, рекомендуемого ВАК

1. Миронов В.В., **Шакирова Г.Р.** Концепция динамических XML-документов // Вестник Уфимского государственного авиационного технического университета. – 2006. – Т. 8. № 5. – С. 58-63.

2. Миронов В.В., **Шакирова Г.Р.** Интерпретация XML-документов со встроенной динамической моделью // Вестник Уфимского государственного авиационного технического университета. – 2007. – Т. 9. № 2. – С. 88-97.

3. Миронов В.В., **Шакирова Г.Р.** Программно-инструментальное средство для создания и ведения динамических XML-документов // Вестник Уфимского государственного авиационного технического университета. – 2007. – Т. 9. № 5. – С. 54-63.

4. Воробьев А. В., **Шакирова Г.Р.** Автоматизированный анализ невозмущенного геомагнитного поля на основе технологии картографических веб-сервисов // Вестник УГАТУ. – 2013. – Т. 17. № 5(58). – С. 177-187.

5. Воробьев А. В., **Шакирова Г.Р.**, Иванова Г.А., Попкова Е.Е. Анализ и исследование частных геомагнитных вариаций // Современные проблемы науки и образования. – 2014. – № 2.

6. Миловзоров Г.В., Воробьев А. В., **Шакирова Г.Р.**, Кильметов Э.А. Исследование и анализ амплитудно-частотных характеристик геомагнитной псевдобури, возникающей в процессе авиаперелета воздушных судов различного целевого назначения // Вестник УГАТУ. – 2014. – Т. 18, №3(64). – С. 132-141.

7. Vorobev A.V., **Shakirova G.R.** Geoinformation system of geomagnetic pseudostorm parameters registration and analysis // Вестник Уфимского государственного авиационного технического университета. – 2014. – Т. 18. № 5 (66). – С. 62-67.

8. Воробьев А. В., **Шакирова Г.Р.**, Иванова Г.А. Исследование и анализ естественных факторов, воздействующих на метрологические характеристики магнитометрических инклинометров // Вестник УГАТУ. – 2015. – Т. 19, № 1(67). – С. 105-113.

9. Воробьев А. В., **Шакирова Г.Р.** Расчет и анализ динамики параметров геомагнитного поля внутриземных источников за период 2010–2015 гг // Геоинформатика. – 2015. №1. – С. 37-46.

10. Воробьев А. В., **Шакирова Г.Р.**, Иванова Г.А. Система принятия решения для гибридной инклинометрической системы на основе технологии картографического веб-сервиса // Фундаментальные исследования. – 2015. – № 5. – С. 260-264.

11. Воробьев А. В., **Воробьева Г.Р.** Применение геоинформационных систем для мониторинга и аналитического контроля параметров космической

погоды, геомагнитного поля и его вариаций // Информация и космос. – 2016. – № 1. – С. 121-128.

12. **Воробьева Г.Р.** Принципы федерализации и консолидации геомагнитных данных в едином информационном пространстве // Перспективы науки. – 2019. – №11(122). – С. 64-67.

13. **Воробьева Г.Р.**, Воробьев А.В. Подход к повышению производительности программных процессов обработки и хранения больших объемов геомагнитных данных // Вестник Томского гос. ун-та. Управление, вычислительная техника и информатика. – 2020. – № 50. – С. 23–30.

В рецензируемых изданиях, индексируемых международными системами цитирования Scopus и Web of Science

14. Vorobev A. V., **Shakirova G. R.** Web-Based Information System for Modeling and Analysis of Parameters of Geomagnetic Field // Procedia Computer Science. ELSEVIER. – 2015. – No. 59. – P. 73 – 82.

15. Vorobev A. V., **Vorobeva G. R.** Web-based geoinformation system for exploring geomagnetic field, its variations and anomalies // Advances in Intelligent Systems and Computing. – 2016. – Vol. 582. – P. 22-35.

16. Vorobev A.V., **Vorobeva G.R**. Analytical information system for control and spectral analysis of geomagnetic field and space weather parameters // Russian Journal of Earth Sciences. – 2016. – Vol. 16, no. 4. – P. 1-10.

17. Vorobev A.V., **Vorobeva G.R.** Weboriented 2D/3Dvisualization of geomagnetic field and its variations parameters // Scientific Visualization. – 2017. – Vol. 9, Issue 2. – P. 94-101.

18. Vorobev A.V., **Vorobeva G.R.** Information system for automated multicriterial analytical control of geomagnetic field and space weather parameters // Communications in Computer and Information Science. – 2017. – No. 741. – P. 109-121.

19. Vorobev A.V., **Vorobeva G.R.** Geoinformation system for amplitude-frequency analysis of geomagnetic variations and space weather observation data // Computer Optics. – 2017. – No. 41. – P. 963-972.

20. Vorobev A.V., **Vorobeva G.R.** Evaluation of the Influence of Geomagnetic Activity on Metrological Characteristics of Inclinometric Information Measuring Systems // Measurement Techniques. – 2017. – No 6. – P. 546-551.

21. Vorobev A.V., **Vorobeva G.R**. Inductive method of geomagnetic data time series recovering // SPIIRAS Proceedings. – 2018. – Vol. 2, no. 57. – P. 103-133.

22. Vorobev A.V., **Vorobeva G.R.** Correlation Analysis of Geomagnetic Data Synchronously Recorded by the INTERMAGNET Magnetic Laboratories // Geomagnetism and Aeronomy. – 2018. – Vol. 58, no. 2. – P. 178-184.

23. Vorobev A.V., **Vorobeva G.R.** Approach to Assessment of the Relative Informational Efficiency of Intermagnet Magnetic Observatories // Geomagnetism and Aeronomy. – 2018. – Vol. 58, no. 5. – P. 625-628.

24. Vorobev A.V., **Vorobeva G.R.** Visualization of geomagnetic variations in time-frequency area of information signal // Scientific Visualization. – 2019. – Vol. 1, no. 2. – P. 143-155.

25. Vorobev A.V., **Vorobeva G.R.**, Yusupova N.I. Conception of geomagnetic data integrated space // SPIIRAS Proceedings. – 2019. – Vol. 18, Issue 2. – P. 390-415.

26. **Vorobeva G.R.** Approach to the recovery of geomagnetic data by comparing daily fragments of a time series with equal geomagnetic activity // Computer Optics. – 2019. – No. 43. – P. 1053-1063.

Монографии

27. Миронов В.В, Юсупова Н.И., **Шакирова Г.Р.**. Иерархические модели данных: концепции и реализация на основе XML. – М: Машиностроение, 2011. – 411 с.

28. Kolios, S., Vorobev, A.V., **Vorobeva, G.R.,** Stylios, C. GIS and Environmental Monitoring. Applications in the Marine, Atmospheric and Geomagnetic Fields. – Springer, 2017. – 174 p.

29. Воробьев А.В., **Воробьева Г.Р.** Метеоинформатика. Геомагнитные вариации и космическая погода. – М.: Инновационное машиностроение, 2017. – 140 с.

Патенты на изобретение

30. Воробьев А. В., **Воробьева Г.Р.** Способ обработки и многослойной визуализации данных с геопространственной привязкой: Патент на изобретение №2568274 . – М.: РосАПО, – 2015.

31. Воробьев А. В., **Воробьева Г.Р.** Способ оценки влияния геомагнитной активности на метрологические характеристики инклинометрического и навигационного оборудования: Патент на изобретение №2644989. М.: РосАПО, 2016.

Свидетельства о государственной регистрации программы для ЭВМ

32. Воробьев А. В., **Воробьева Г.Р.** [GIMS] array _analyzer v1.0: Св-во об офиц. регистрации программы для ЭВМ №2014615627. – М.: РосАПО, 2014.

33. Воробьев А. В., **Воробьева Г.Р.**, Иванова Г.А. Расчет допустимых геомагнитных вариаций при проведении инклинометрических исследований: Св-во об офиц. регистрации программы для ЭВМ №2015614394. – М.: РосАПО, 2015.

Прочие публикации

34. Yusupova N., Vorobev A., Groumpos P., **Vorobeva G.** Web-based solutions in modeling and analysis of geomagnetic field and its variations // CEUR Workshop Proceedings. – 2018. – Vol. 2254. – P. 282-289.

Соискатель  Г.Р. Воробьева