ИНСТИТУТ ФИЗИКИ ЗЕМЛИ ИМ. О.Ю. ШМИДТА РАН

ЛАБОРАТОРИЯ ГЕОИНФОРМАТИКИ

На правах рукописи

МАЛЫГИН Иван Вячеславович

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ В ГЕОФИЗИЧЕСКИХ ЗАДАЧАХ С ДЕФИЦИТОМ ДАННЫХ

25.00.10 – геофизика, геофизические методы поисков полезных ископаемых

Диссертация на соискание ученой степени

кандидата технических наук

Научный руководитель:

кандидат физико-математических наук

Алешин Игорь Михайлович

Москва – 2021

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность темы

Данная диссертация посвящена исследованию нескольких классических задач с недостатком данных: пространственная интерполяция (двумерная и трехмерная) и классификация на основе временных рядов. Задачи с пропуском данных являются традиционными для геофизических исследований. Это связано с недостаточным количеством измерений, непродолжительным промежутком времени наблюдений, большой протяженностью объектов. Подобная ситуация дефицита данных затрудняет обработку и интерпретацию результатов измерений геофизических полей. Несмотря на то, что в последнее время большое внимание уделяется задачам с большими данными (Big Data), большинство геофизических данных по-прежнему не являются таковыми [Гвишиани, 2019].

Применение классических методов статистического анализа позволяет получать достоверные результаты при наличии большого объема наблюдений. Недостаток данных в геофизических исследованиях требует применения специальных методов анализа. С точки зрения математической постановки такие проблемы являются классическим случаем задач с пропуском данных [Журавлев, Никифоров, 1971; Горелик, Скрипкин, 1977; Dempster, Laird, Rubin, 1977; Гафуров, Краснопрошин, Образцов, 2007; Luengo, Garcia, Herrera, 2012]. Одним из подходов при решении подобных задач, является группа методов, разработанных в рамках теории распознавания образов и машинного обучения [Вапник, Червоненкис, 1980; Журавлев, Рязанов, Сенько, 2006; Flach, 2012]. В практических исследованиях выделяются два подхода к решению задач с недостатком данных: базовые и специализированные.

Базовые методы основаны на универсальных алгоритмах машинного обучения. Они могут применяться для анализа многомерной и разнородной информации, включая геофизические данные. Такие алгоритмы для обучения используют простые характеристики исходных данных, например, вычисляют

расстояния между точками, строят линейные приближения, производят операции с множествами. Эти алгоритмы начали разрабатываться с 1950-х годов XX в., имеют теоретическое и эмпирическое обоснование. В качестве примера задачи, для решения которой эффективно использовать базовые методы теории машинного обучения, можно привести задачу интерполяции. Она возникает при построении двумерных и трехмерных моделей физических характеристик горных пород и геофизических полей в ограниченной области по измеренным значениям. В таких исследованиях, как правило, измерения проводятся на нерегулярной сетке в небольшом количестве точек. Это обусловлено значительным расстоянием между точками наблюдений и особенностями методик измерений. Для построения распределений в таких случаях широко используются интерполяционные процедуры. Наибольшее распространение получил кригинг – метод разработанный Д. Криге [Krige, 1951]. Применение кригинга имеет ряд недостатков. Этот метод часто приводит к излишнему сглаживанию пространственного распределения исследуемых величин. Поэтому одним из актуальных направлений является разработка методов, которые позволяют улучшить контрастность получаемых образов. Эту задачу в постановке машинного обучения возможно рассматривать как задачу пространственной классификации: необходимо отнести значения интерполянта в промежуточных точках к одному из заданных классов, что позволяет построить выраженные границы в пространстве.

Специализированные методы разрабатываются для решения конкретной прикладной задачи. В их основе лежат общие алгоритмы машинного обучения, существенно адаптированные под специфику решаемой задачи. Центральной частью применения специально адаптированных методов машинного обучения является более сложная процедура обучения. Машинное обучение – систематическое обучение алгоритмов, в результате которого их знания и качество работы возрастают по мере накопления опыта [Flach, 2012]. Под

обучением понимается автоматизированная настройка параметров алгоритма на данных конкретной прикладной задачи. На практике это чаще всего означает численное решение некоторой оптимизационной задачи: конечный прикладной результат исследования представляют в виде формализованного функционала качества, который оптимизируется в процессе обучения на объектах исходных данных. Процедура обучения активно используется, например, в одном из методов прогнозирования мест сильных землетрясений EPA (Earthquake-Prone Areas recognition) [Гельфанд И.М. и др., 1972; Гельфанд и др., 1976]. Здесь обучение используется для поиска критериев, обеспечивающих наилучшее согласование с известными результатами наблюдений. Использование дополнительной информации позволяет получать карты распределения вероятности возникновения землетрясения в рассматриваемой сейсмоактивной территории (Карта Ожидаемых Землетрясений) [Завьялов, 2004], улучшить прогнозирование мест возникновения сильных землетрясений (алгоритм Барьер) [Гвишиани и др., 2017; Дзебоев и др., 2019]. Другим примером специализированных методов является классификация на основе временных рядов. Задачи с временными рядами особо актуальны при мониторинге катастрофических процессов (магнитные бури, гидрометеопроцессы). Особенностью этой задачи в ситуации недостатка данных являются короткие временные ряды однородных наблюдений на исследуемой территории, возможно, содержащие пропуски в данных измеряемых значений. Одним из подходов к решению подобной задачи классификации является использование в процессе обучения интегральных характеристик, построенных по этим рядам.

Цель исследования

Цель исследования – разработка методов решения геофизических задач с дефицитом данных, на основе теории машинного обучения.

В диссертации рассмотрены задачи двух типов:

1) Построение двухмерного и трехмерного распределения геофизических величин по данным полевых измерений на нерегулярной сетке с помощью базовых методов теории машинного обучения.

2) Построение прогнозной системы, основанной на рядах коротких временных данных, для чего применялся специализированный метод, опирающийся на комбинаторно-логический подход теории распознавания образов.

Основные задачи исследования:

1. Построение карты границы Мохоровичича по данным исследований сейсмических волн.
2. Расчет кажущегося сопротивления среды на основе данных электропросвечивания.
3. Создание прогнозной системы для определения мощности ледового заторообразования.

На защиту выносятся:

1. Разработан и реализован метод построения региональной двумерной и трехмерной цифровой модели на основе небольшого количества исходных данных. Метод основан на применении базовых методов машинного обучения, модифицированных с учетом специфики входных данных. Метод применим в ситуации, когда объем данных невелик, и они имеют сильно анизотропное пространственное распределение.
2. Разработана прогнозная система, предназначенная для осуществления краткосрочного прогноза образования заторов весенний период на Сев. Двине. Система позволяет осуществлять прогноз и анализировать данные в условиях ограниченного набора исходных наблюдений на гидропостах и метеостанциях. Применение разработанной системы позволяет достигнуть точности прогнозирования до 85%.

Методика исследований

Основные результаты исследования получены с применением методов машинного обучения и методов теории распознавания образов. Реализация алгоритмов обработки данных для пространственной интерполяции выполнена на языке Python 3 (глава 2). Реализация логического алгоритма прогнозной системы выполнена на языке программирования C/C++ в среде Microsoft Visual Studio (глава 3). Составление карт (пп. 2.1, 2.2.) проводилось в среде GoldenSoftware Surfer 15, визуализация данных прогнозной системы (Приложение 3) проводилась в среде ArcGIS 10. Источниками информации являются научные публикации, справочные издания, тематические электронные ресурсы, экспертные знания.

Научная новизна

Предложено решение научно-технической задачи, имеющей влияние на развитие геофизических и геоинформационных технологий, методов прогнозирования сложных, трудно-формализуемых процессов, методов анализа и обработки временно*й* и пространственно-распределенной геофизической информации.

Показано, что применение базового метода машинного обучения (метод ближайших соседей) в задачах по построению пространственных распределений двумерных геофизических величин с ограниченным объемом исходных данных позволяет достичь лучших результатов в части уточнения границ объектов по сравнению с использовавшимся ранее методом кригинга на основе линейной регрессии. Метод ближайших соседей позволил лучше определять нелинейные зависимости в пространственном распределении геофизических величин, лучше выделять области пространственной неоднородности.

Применена сферическая метрика для уточнения работы метода ближайших соседей для существенно удаленных на поверхности Земли опорных точек интерполяции.

Показано, что в задаче построения трехмерной модели среды при проведении межскважинных исследований метод ближайших соседей позволяет оконтурить малые объекты даже при использовании синхронной схемы измерений.

Влияние пространственной анизотропии распределения данных можно исключить, если модифицировать пространственную метрику, определяющую расстояние между данными. Это достигается введением коэффициента скейлинга, который изменяет масштаб в горизонтальном направлении. Использованный подход позволяет получить контрастное изображение неоднородных областей, что позволяет выделить неоднородности, чьи геометрические размеры меньше расстояния между скважинами.

Разработанная технология является универсальной: процесс построения трехмерной модели не зависит от физической модели, использованной для интерпретации измерений.

Для задачи анализа временных рядов ограниченной длины предложен оригинальный подход на основе специализированных методов теории распознавания образов. Разработанный метод создания прогнозной системы сочетает используемые на практике принципы классификации явлений с экспертными и математическими методами прогнозирования. Разработанная на основе методов машинного обучения прогнозная система является универсальной в части требований к исходным данным, алгоритмического обеспечения задачи прогноза и анализа результата его достоверности.

Выполнен прогноз ледового заторообразования для участка р. Северная Двина на несколько сезонов. Проведена валидация прогнозов системы. Оцененная достоверность составила 85%.

Проведен факторный анализ, построенных на основе коротких временных рядов, характеристик процесса заторообразования. Результаты анализа подтвердили выдвинутые ранее теоретические гипотезы о важности признаков.

Практическая значимость результата работы

Представленные в работе результаты анализа геофизических данных могут применяться на практике для решения ряда геофизических задач.

Построена уточненная карта границы Мохоровичича для региона Фенноскандия. Толщина коры, определяемая как расстояние от поверхности до этой границы, является основной характеристикой при анализе строения региона, а также при изучении структуры европейской литосферы. Построенная карта может применяться для дальнейших исследований строения литосферы северной части Балтийского щита, изучения строения мантии северной и южной Финляндии, построения трехмерных сейсмических моделей южной Финляндии.

Построенная карта слоя с низкими скоростями поперечных сейсмических волн может применяться для продолжения исследований природы его возникновения. Приведенный анализ сейсмических данных показал эффективность методов машинного обучения для их анализа и обобщения. Достоинства такого подхода связаны с универсальностью применяемых методов. Особенно ярко преимущества алгоритмов теории машинного обучения проявляются в условиях недостатка данных, типичных для многих геофизических исследований.

Построена трехмерная модель проводимости среды при проведении межскважинных исследований. Использованный метод машинного обучения (метод ближайших соседей) позволяет построить трехмерную модель проводимости среды между скважин даже при использовании синхронной схемы измерений.

Разработана прогнозная система для осуществления краткосрочного прогнозирования мощности процесса заторообразования для участка реки Северная Двина от г. Котлас до г. Великий Устюг, что является важной частью прогноза наводнений для данной территории.

Достигнута точность прогнозирования на уровне 85%, результаты подтверждены проведенной валидацией прогнозов.

Реализована функциональность, которая позволяет применять прогнозную систему в качестве инструмента анализа данных: проверять гипотезы относительно влияния признаков на исследуемый процесс, оценивать величину вклада конкретного признака в итоговый результат явления.

Соответствие паспорту специальности

Работа содержит решение задач, имеющих научно-практическую значимость в части совершенствования способов обработки и интерпретации данных измерений геофизических полей, интегрированного анализа многомерной, многопараметрической и разнородной информации, включающей геофизические данные, а также применение геофизических методов в решении задач охраны окружающей среды и соответствует пунктам №№ 14, 18, 25 Паспорта специальности ВАК 25.00.10 «Геофизика, геофизические методы поисков полезных ископаемых» (технические науки).

Апробация работы

Работа и отдельные результаты обсуждались на научных семинарах ИФЗ РАН, МГУ им. М.В. Ломоносова, а также на следующих конференциях:

- Information Technologies in Earth Sciences and Applications for Geology, Mining and Economy (ITES&MP-2019) – Moscow, 2019;
- Всероссийская конференция с международным участием II Юдахинские чтения «Проблемы обеспечения экологической безопасности и устойчивое развитие арктических территорий» – Архангельск, 2019;
- Научная конференция молодых ученых и аспирантов ИФЗ РАН (2019) – Москва, 2019;
- VI Международная научно-практическая конференция «Индикация состояния окружающей среды: теория, практика, образование» – Москва, 2018;

- Научная конференция молодых ученых и аспирантов ИФЗ РАН (2018) – Москва, 2018;
- IV Школа-семинар «Гординские чтения» – Москва, 2017;
- Научная конференция молодых ученых и аспирантов ИФЗ РАН (2017) – Москва, 2017;
- Международный молодежный научный форум «ЛОМОНОСОВ-2015» – Москва, 2015;
- Международный молодежный научный форум «ЛОМОНОСОВ-2014» – Москва, 2014;
- IV Международная научно-практическая конференция «Научные перспективы XXI века. Достижения и перспективы нового столетия» – Новосибирск, 2014.

Получены свидетельства о государственной регистрации программ для ЭВМ:

- **Малыгин И.В.** Свидетельство о государственной регистрации программы для ЭВМ №2014614960 Экспертная система прогнозирования ледового заторообразования. Дата гос. регистрации в Реестре программ для ЭВМ 14.05.2014.
- **Малыгин И.В.**, Алешин И.М. Свидетельство о государственной регистрации программы для ЭВМ № 2020617961 Программа расчета и построения региональных карт геофизических свойств методом k-ближайших соседей. Дата гос. регистрации в Реестре программ для ЭВМ 15.07.2020.
- **Малыгин И.В.**, Алешин И.М. Свидетельство о государственной регистрации программы для ЭВМ № 2020617962 Программа расчета и построения трехмерной модели проводимости среды по данным межскважинных измерений методом k-ближайших соседей. Дата гос. регистрации в Реестре программ для ЭВМ 15.07.2020.

Публикации

По материалам диссертации опубликовано 12 работ, в том числе 8 статей в ведущих рецензируемых изданиях, рекомендованных ВАК РФ.

Структура работы

Диссертация состоит из введения, трех глав, заключения, списка литературы из 129 наименований, трех приложений. Текст диссертации изложен на 124 страницах машинописного текста и содержит 12 таблиц, 32 рисунка и 3 приложения.

Автор выражает благодарность научному руководителю к.ф.-м.н. Игорю Михайловичу Алешину (ИФЗ РАН) за поддержку на всех этапах проведения работы, а также коллективу лаборатории геоинформатики ИФЗ РАН.

ГЛАВА 1. Методы машинного обучения в задачах с дефицитом данных

В первой главе рассмотрены примеры геофизических задач в ситуации недостатка данных, в которых применение современной теории машинного обучения и методов распознавания образов привело к новым результатам. Введены понятия и определения теории машинного обучения, рассмотрена общая постановка задачи обучения с учителем, приведены основные функционалы качества. В задачах построения 2D-модели региона и построения 3D-модели среды рассмотрена ситуация недостатка исходных пространственных данных измерений, предложен способ решения на основе метода ближайших соседей. В задаче прогнозирования опасных геофизических явлений с ограниченным объемом временны*х* данных предложен способ решения на основе комбинаторно-логического подхода теории распознавания образов.

1.1. Ситуация дефицита данных в геофизических задачах

В практических задачах часто отсутствуют полные и одинаково структурированные исходные данные, собранные на большой территории, либо за большой промежуток времени. Возникает ситуация дефицита данных, когда для используемых методов обработки (в основном, вероятностных) недостаточно информации для достижения статистической значимости выводов при проверке гипотез, а, следовательно, и для подтверждения качества построенной модели.

Недостаток данных в геофизических приложениях обуславливается рядом факторов. Районы проведения геофизических исследований могут располагаться в труднодоступных, удаленных местах. Часто, требуется проведение дорогостоящего сбора данных на местности. Например, при проведении скважинных исследований, стараются экономить на количестве и регулярности скважин, а также на самом способе межскважинных измерений. Появляется возможность использования в исследованиях новых классов данных, часто ограниченных сроком начала их сбора. Примером этого являются данные ГНСС,

объем которых увеличивается по мере появления новых систем. Например, система BeiDou планируемая к глобальному запуску в 2020 г. [BeiDou Navigation Satellite System]. Использование данных ГНСС помогает получить пространственно-временную привязку геофизических процессов, однако часто это невозможно из-за отсутствия необходимой временной глубины данных.

В задаче прогноза землетрясений требуется построить прогноз места, силы и времени возникновения события. Работы в этом направлении начались в середине XX в. [Герасимов, 1947], постановка задачи была формулирована В.И. Кейлис-Бороком [Ранцман, 2001]. Одним из первых алгоритмов, использующих теорию распознавания образов, был КОРА-3 [Бонгард, 1967; Гельфанд и др., 1976], который решал статическую задачу, без использования временной составляющей. Этот недостаток был устранен в методе ФОП [Вапник, Червоненкис, 1980]. Развитием первоначальных подходов стали комплексы алгоритмов КН, М8, MSc [Аллен и др., 1984; Аллен и др., 1986; Кейлис-Борок, Кособоков 1984; Кейлис-Борок, Кособоков 1986; Kosobokov et al., 1990], однако эти алгоритмы не позволяют получить карты распределения вероятности возникновения сильного землетрясения в рассматриваемой сейсмоактивной территории. Для их построения был разработан алгоритм КОЗ [Завьялов, 2004]. Развитием прогнозирования мест возникновения сильных землетрясений является алгоритм Барьер [Гвишиани и др., 2017; Дзебоев и др., 2019].

Подобная прогнозная задача возникает и при исследовании ледовой обстановки на северных реках. Аналогично землетрясениям, требуется определить мощность опасного явления. При этом места возникновения ледовых заторов, обладают высокой повторяемостью [Агафонова, Фролова, 2007], поэтому достаточно построить прогноз мощности опасного явления с необходимой заблаговременностью. Задача является актуальной для ряда районов европейской территории России и решается, как правило, по данным гидрометеорологических наблюдений [Бузин, Зиновьев, 2009]. При этом

некоторые районы обеспечены наблюдениями за длительный период, например, по р. Неве имеются структурированные данные за период более 200 лет [Бузин, 1997], однако есть районы, наблюдения по которым охватывают период 15-30 лет. Эта разница имеет существенное значение при выборе методики прогноза, как для методической ее части, так и для оценки практической достоверности. В гидрологии разработаны различные подходы к прогнозу ледового режима рек [Шуляковский, 1951; Шуляковский, Еремина, 1952; Берденников, Шматков, 1974; Agafonova et al., 2017]. В главе 3 представлен способ прогнозирования мощности ледового заторообразования в ситуации недостатка данных, когда многолетние одинаково структурированные полные данные труднодоступны, либо отсутствуют.

Другим примером геофизической задачи с ограниченным объемом данных, в которых применение методов распознавания помогло достичь поставленных результатов, является задача об оценке масштабов оруднения месторождений полезных ископаемых [Константинов, Королева, 1973; Константинов, Королева, Кудрявцев, 1976]. Было задано три группы месторождений, содержащих: ртуть, бокситы, флюорит. В каждой группе в соответствии с запасами содержащегося сырья все месторождения были поделены на два класса: крупные (пригодные к промышленной разработке) и мелкие (не пригодные к промышленной разработке). Все месторождения были описаны набором геологических признаков: 92 признака для ртутных месторождений, 80 признаков для месторождений бокситов и 40 признаков для флюоритовых месторождений. Требовалось на основании сделанных проб внутри каждой группы месторождений определить является ли оно крупным, либо мелким.

Нетрудно видеть, что задача в такой постановке является задачей бинарной классификации с недостатком данных: обучающие выборки в данной задаче содержали всего от 8 до 14 объектов (известных месторождений), а

контрольные выборки от 10 до 12 месторождений. Решение этой задачи основано на комбинаторно-логическом подходе теории распознавания образов [Журавлев, Никифоров, 1971; Яблонский и др., 1971]. Авторам удалось достичь итоговой точности классификации порядка 80%, в практическом плане результатом стали рекомендации о постановке геологоразведочных работ ряда малоизученных месторождений (богатство которых подтвердилось при проведении разведки), а также выдвинуты гипотезы о степени важности тех или иных геологических признаков на итоговый масштаб оруднения месторождения.

Еще одним примером задачи с ограниченным объемом исходных данных является задача интерполяции для построения плоских (2D) и трехмерных (3D) пространственных распределений и карт различных геофизических характеристик (например, карты границы Мохоровичича или трехмерного распределения коэффициента затухания радиоволн при проведении скважинных исследований). Стандартный подход, встречающийся в геостатистике [Isaaks, Srivastava, 1989] и геоинформатике [Кошель, Мусин, 2001] – применение метода кригинга (ordinary kriging), имплементированного в большинстве современных геоинформационных программных продуктов (ESRI ArcGIS, Goldensoftware Surfer). Метод кригинга основан на обобщении метода линейной регрессии, и при построении количественного распределения пространственных данных (вариограмм) отличается нечеткими границами между областями с постоянным значением интерполируемой величины. Возникает задача более точного оконтуривания слоев при построении интерполяции по сравнению с методом кригинга. Ряд методов теории машинного обучения позволяет существенным образом учесть пространственную и геометрическую структуру оконтуриваемых объектов. Более подробно эта задача будет рассмотрена в главе 2 на конкретных примерах.

1.2. Основные понятия теории машинного обучения

Машинным обучением называется систематическое обучение алгоритмов, в результате которого их знания или качество работы возрастают по мере накопления опыта [Flach, 2012]. Характерной чертой методов машинного обучения является не прямое решение задачи, а *обучение* в процессе применения решению множества сходных задач. Под обучением понимается автоматизированная настройка параметров алгоритма на данных конкретной прикладной задачи.

Предположим, моделируется некоторый геофизический процесс. Имеется множество объектов (исходных ситуаций) и множество возможных ответов (состояний процесса). Выдвигается гипотеза о существовании зависимости между ответами и объектами, но она заранее неизвестна. Известна только конечная совокупность пар «объект, ответ», называемая *обучающей выборкой*. На основе этих данных требуется восстановить неявную зависимость, то есть построить алгоритм, способный для любого возможного входного объекта выдать достаточно точный классифицирующий ответ. Ответ алгоритма часто называют целевой функцией или целевой переменной (target).

Для измерения точности ответов вводится *функционал качества* – функция, зависящая от ответов алгоритма и истинных значений, и выдающая некоторую оценку качества (score) построенного алгоритма. По-другому, это количественная оценка способности построенного алгоритма устанавливать соответствие между входными данными и наиболее вероятным значением целевой переменной.

Объекты представляются в виде структур данных, чаще всего таблиц, содержащих их признаковое описание. *Признаком* (feature) называется любое доступное измеримое свойство или характеристика объекта. Признаки бывают вещественными (например, измеренная физическая характеристика, показатели датчиков, наблюдения и т.д.) и категориальными (например, тип, цвет, форма и т.д.). Прочие виды признаков, извлеченные из сложно-структурированных

данных (изображения, текст, метаданные и т.д.), как правило, можно свести к вещественным или категориальным.

Приведенная постановка является классической задачей обучения по прецедентам, или, часто говорят, *обучения с учителем* (supervised learning) [Кудрявцев, Гасанов, Подколзин, 2006; Flach, 2012]. Основные типы таких задач зависят от вида целевой переменной. В случае, если целевой переменной является дискретная величина с ограниченным набором возможных значений, говорят о задаче *классификации* или *распознавания*[1]. В случае, когда целевая переменная представлена вещественным числом – о задаче *регрессии*.

Отдельного внимания заслуживает задача *бинарной классификации*: в этом случае целевая переменная может принимать всего два значения (чаще всего их обозначают метками 0 и 1). Соответствующий такой постановке алгоритм машинного обучения называется бинарным классификатором. Задача бинарной классификации является базовой как на теоретическом, так и на практическом уровне, поскольку классификацию с большим числом классов можно свести к задаче бинарной классификации с помощью способа «один против всех» («one vs. all»), когда строится последовательность бинарных классификаций, на каждой итерации которой последовательно отделяется один класс (с меткой 1) от всех остальных (с меткой 0). У задачи бинарной классификации также существует очень важная вероятностная постановка, так называемая задача *оценки принадлежности*, когда целевая переменная представлена не бинарными метками 0 и 1, а вещественным числом p из диапазона [0;1], которое интерпретируется как «вероятность такого события, что объект X принадлежит классу с меткой 1». В ряде задач такая постановка является более универсальной, чем классическая бинарная классификация.

[1] В настоящее время этот термин применяется, в основном, к задачам распознавания зрительных образов (цифровых изображений, космических снимков, фотографий). Однако, представителями отечественной школы кибернетики [Журавлев и др.] этот термин применяется как синоним задачи классификации.

Процесс проверки качества построенного алгоритма называется *валидацией*. Обычно под этим понимается вычисление количественной оценки качества через определенный функционал качества. Для этого необходимо отделить часть исходных данных от обучающей выборки и оставить их для оценки качества. Такая выборка называется *отложенной* или *контрольной* (часто также используется понятие *валидационной* выборки).

Однако чаще, для оценки качества, а также для настройки свободных параметров алгоритма (*гиперпараметров*) используется метод *скользящего контроля* или *кросс-валидации* (cross-validation). Эта техника является стандартной в задачах машинного обучения, она основана на проведении последовательности численных экспериментов. На каждом шаге исходные данные делятся на два блока. Первый блок используется в качестве обучающей выборки, второй – в качестве контрольной. Разница между результатами расчетов и известными значениями вычисляется на основе заданного функционала качества.

Разделение данных на два блока необходимо во избежание так называемой «утечки данных». Такая ситуация возникает в том случае, когда и для определения значений свободных параметров, и для последующего контроля качества используются одни и те же данные, что и приводит к завышению числовых оценок качества в процессе контроля.

Стратегий разбиения данных на обучающую и контрольную выборки существует много. На практике чаще всего используют следующую. Исходные данные случайным образом разбивают на p блоков одинакового размера. Каждый блок последовательно выполняет роль контрольной выборки, а совокупность оставшихся p-1 блоков – роль обучающей выборки. Основное преимущество такой стратегии обусловлено тем, что все измерения, доступные в исходных данных, используются и для обучения, и для проверки качества. При

этом каждый элемент исходных данных используется для проверки качества ровно один раз.

Остается вопрос с выбором конкретного значения p применительно к доступным данным. Если доступно большое количество исходных данных (десятки, сотни тысяч объектов и более), число p выбирают, как правило, из диапазона 3-20. Например, в различных фреймворках, содержащих алгоритмы машинного обучения, в качестве параметров по умолчанию используются небольшие значения p=3 или p=5 [Pedregosa et al., 2011]. Это обусловлено необходимостью p раз выполнять процедуру обучения, что на большом количестве данных может привести к вычислительным трудностям. В задачах с недостатком данных, когда для анализа доступно всего несколько десятков измерений, значение p можно выбрать равным числу измерений в исходных данных. В англоязычной литературе такая стратегия получила название Leave-one-Out кросс-валидации [Molinaro, Simon, Pfeiffer, 2005]. Небольшое количество доступных для анализа данных позволяет использовать этот подход и в настоящей работе.

В качестве функционалов качества возможно рассматривать различные величины. Выбор из них в конкретной прикладной задаче обуславливается типом постановки задачи (классификация, регрессия, оценка принадлежности), физическим смыслом и распределением целевой переменной на исходных данных. Ниже приведены некоторые функционалы качества для трех типов задач, которые понадобятся в гл. 2 и гл. 3.

Для определения функционалов качества необходимо ввести следующие обозначения:

a – алгоритм машинного обучения;

X – контрольная выборка данных;

l – мощность контрольной выборки (число объектов);

x_i – признаковое описание i-го объекта контрольной выборки;

y_i – истинное (*a priori* известное) значение целевой переменной i-го объекта контрольной выборки;

$\bar{y}$ – среднее значение целевой переменной y на контрольной выборке;

$[a(x_i) = y_i]$ – индикатор события: если условие в скобках верно, то 1, иначе 0.

Тогда во введенных обозначениях средняя квадратичная ошибка в задаче регрессии имеет вид

$$MSE(a, X) = \frac{1}{l}\sum_{i=1}^{l}(a(x_i) - y_i)^2.$$

Такой функционал является гладким, поэтому его достаточно легко оптимизировать, используя метод градиентного спуска. Этот функционал сильно штрафует за большие ошибки, так как отклонения возводятся в квадрат. Это приводит к тому, что штраф на выбросе будет очень сильным, и алгоритм машинного обучения будет настраиваться на выбросы.

Другой похожий функционал в задаче регрессии – средняя абсолютная ошибка

$$MAE(a, X) = \frac{1}{l}\sum_{i=1}^{l}|a(x_i) - y_i|.$$

Такой функционал уже не является гладким, поэтому его сложнее оптимизировать, но у такого функционала больше устойчивость к выбросам, так как штраф за сильное отклонение гораздо меньше.

Еще один широко используемый функционал качества регрессии – коэффициент детерминации

$$R^2 = 1 - \frac{\sum_{i=1}^{l}(a(x_i) - y_i)^2}{\sum_{i=1}^{l}(y_i - \bar{y})^2}$$

Коэффициент детерминации показывает долю разнообразия ответов в целевой переменной, которую построенный алгоритм смог объяснить.

В задаче классификации естественным функционалом качества является точность

$$\text{accuracy}(a, X) = \frac{1}{l}\sum_{i=1}^{l}[a(x_i) = y_i]$$

Точность показывает долю правильно распознанных элементов выборки. Эта простая метрика и очень часто используется, однако она имеет существенные недостатки в случае несбалансированной выборки, либо в случае необходимости учета различных цен ошибок.

Для определения функционалов качества в задаче оценки принадлежности, необходимо определить вспомогательные понятия: *матрицу ошибок* (confusion matrix) и ее компоненты (Табл. 1.1.). Она помогает удобно разделить случаи, как соотносятся между собой результат работы алгоритма и истинный ответ.

Таблица 1.1. Матрица ошибок бинарного классификатора

	$y = 1$	$y = 0$
$a(x) = 1$	True Positive (TP)	False Positive (FP)
$a(x) = 0$	False Negative (FN)	True Negative (TN)

Когда алгоритм относит объект к классу 1, говорят, что алгоритм срабатывает. Если алгоритм сработал и объект действительно относится к классу 1, имеет место верное срабатывание (true positive), а если объект на самом деле относится к классу 0, имеет место ложное срабатывание (false positive).

Если алгоритм дает ответ 0, говорят, что он пропускает объект. Если имеет место пропуск объекта класса 1, то это ложный пропуск (false negative). Если же алгоритм пропускает объект класса 0, имеет место истинный пропуск (true negative). Таким образом, существуют два вида ошибок: ложные срабатывания и ложные пропуски. Для каждого из них нужен свой функционал качества, чтобы измерить, какое количество ошибок каждого типа совершается.

Точность показывает насколько можно доверять классификатору в случае срабатывания

$$\text{precision}(a,\, X) = \frac{TP}{TP + FP}.$$

Полнота показывает на какой доле истинных объектов первого класса алгоритм срабатывает

$$\text{recall}(a,\, X) = \frac{TP}{TP + FN}.$$

В некоторых задачах есть ограничения на один из этих функционалов, тогда как по второму будет производиться оптимизация. Но в некоторых случаях необходимо максимизировать и точность, и полноту одновременно. В таких случаях обычно используется одна из агрегатных функций (среднее, минимум и т.д.).

Возвращаясь к функционалам качества в задаче оценки принадлежности важно отметить, что многие алгоритмы бинарной классификации устроены следующим образом: сначала вычисляется некоторое вещественное число $b(x)$, которое сравнивается с порогом t:

$$a(x) = [b(x) > t],$$

где $b(x)$ – оценка принадлежности классу 1. Иначе, $b(x)$ является некоторой оценкой уверенности, что объект x принадлежит классу 1. В практических задачах часто необходимо оценить качество именно оценки принадлежности, а порог выбирается позже из соображений на точность или полноту.

В практических задачах оценки принадлежности широкое распространение получила величина AUC-ROC (Area Under Curve - Receiver Operating Characteristic) – площадь под ROC-кривой [Fawcett, 2006]. Значение метрики ROC-AUC в задаче бинарной классификации отражает вероятность того, что случайно выбранный объект класса 1 имеет оценку принадлежности к классу 1 выше, чем случайно выбранный объект класса 0.

ROC-кривая, строится в осях False Positive Rate (FPR – ось x):

$$FPR = \frac{FP}{FP + TN}$$

и True Positive Rate (TPR – ось y):

$$TPR = \frac{TP}{TP + FN}.$$

ROC-кривая строится итеративно: постепенно рассматриваются случаи различных значений порогов и отмечаются точки на графике. Кривая стартует из точки (0,0) и приходит в точку (1,1). При этом, если существует идеальный классификатор, кривая должна пройти через точку (0,1). Чем ближе кривая к этой точке, тем лучше будут оценки, а площадь под кривой будет характеризовать качество оценок принадлежности к первому классу. Такой функционал качества и называется AUC-ROC, или площадь под ROC-кривой. На Рис. 1.1 приведен пример построения ROC-кривой в реальной практической задаче.

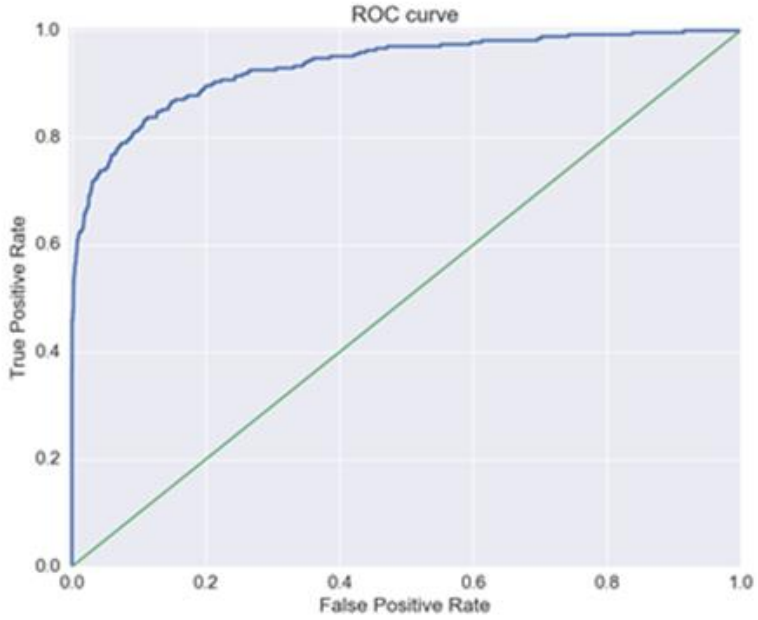

Рис. 1.1. Пример ROC-кривой

1.3. Метод ближайших соседей для пространственной интерполяции

В теории методов машинного обучения существует несколько классов алгоритмов, которые имеют преимущества (перед другими алгоритмами) при

работе с пространственными данными и геометрией объектов. Прежде всего это современная теория глубокого обучения, которая предоставляет классы рекуррентных и сверточных нейронных сетей [Николенко, Кадурин, Архангельская, 2018]. Такие алгоритмы демонстрируют выдающиеся результаты в задачах с большим количеством данных (big data), однако совершенно неприменимы в условиях ограниченного объема исходных данных [Goodfellow, 2016]. В этом используются метрические алгоритмы распознавания, и самый известный из них – метод ближайших соседей (*k Nearest Neighbors* или сокращенно *kNN*) [Altman, 1992; Журавлев, Рязанов, Сенько, 2006; Hastie, Tibshirani, Friedman, 2009].

Алгоритм *kNN* относится к группе так называемых «ленивых» алгоритмов (lazy learning). Обучение здесь сводится к расчету матрицы расстояний между объектами исходных данных, то есть, фактически, к запоминанию параметров обучающей выборки. Для составления предсказания для новой точки необходимо определить ближайшие k объектов обучающей выборки.

В методе *kNN* значение интерполируемой величины Q в точке с координатами $\Omega = \{\phi, \lambda\}$ (ϕ, λ — географические широта и долгота соответственно) определяется по набору известных значений $\{q_i\}$, заданных в точках Ω_i, формулой

$$Q(\Omega, K) = \sum_{i=1}^{k} w(\rho_i)\, q_i(\Omega_i) \,/\, \sum_{i=1}^{k} w(\rho_i)\,.$$

Здесь $w(\rho_i)$ — весовая функция, зависящая от расстояния от i-й точки, суммирование ведется по K точкам, ближайшим к Ω. Как правило, зависимость весовой функции от обратного расстояния между точками выбирают в виде степенного закона:

$$w(\rho_i) \sim 1/\rho_i^{\alpha}, \alpha > 0.$$

Важной особенностью применения метода ближайших соседей в геофизических задачах является выбор метрики: функционала, который задает

расстояние между двумя точками с измеренными координатами. Часто объекты находятся на сфере (поверхности Земли) на значительном удалении друг от друга (более 100 км.). В таком случае, для повышения точности интерполяции необходимо использовать соответствующее расстояние на сфере, вычисленное как длина дуги геодезической линии (большого круга). В этом случае, формулы длины ортодромии имеют вид:

$$\rho(\Omega_1, \Omega_2) = 2R_E \arcsin\sqrt{\sin^2(\Delta\phi/2) + \cos\phi_1 \cos\phi_2 \sin^2(\Delta\lambda/2)},$$

$$\Delta\lambda = \lambda_2 - \lambda_1, \Delta\phi = \phi_2 - \phi_1.$$

Число ближайших соседей k является свободным параметром задачи (часто используется термин *гиперпараметр*). Для его определения обычно использует один из подходов, основанных на методе отложенной выборки. Часть данных исключается из рассмотрения и используется для проверки. Функционалом качества интерполяции μ может служить, например, MAE – среднее абсолютное отклонение рассчитанных для этих точек значений от реальных:

$$\mu(k) = 1/M \sum_{m=1}^{M} |Q(\Omega_m, k) - q_m|$$

$\{q_m \equiv q(\Omega_m)\}$ – набор из M «отложенных» исходных данных q_m, относящихся к точкам Ω_m. Величина $\mu = \mu(k)$зависит от числа соседей k, как от параметра. Поэтому оптимальное значение k можно определить из условия минимума этой зависимости.

1.4. Методы создания систем прогнозирования опасных геофизических явлений

Задача прогнозирования состоит в описании изменения объекта, среды или явления во времени при наличии данных о поведении за некоторый предшествующий период. В рамках такой постановки ее можно рассматривать

как задачу распознавания. Для ее решения в теории распознавания существует два основных подхода: вероятностный и комбинаторно-логический. Выбор из них в конкретных исследованиях определяется объемом данных за предшествующий период. В случае наличия временных рядов за длительный период наблюдения целесообразно применять вероятностные методы, которые обеспечивают необходимую достоверность прогноза, иначе целесообразно использовать комбинаторно-логический подход теории распознавания образов. Применение вероятностных методов в задачах распознавания с малообъемными выборками не обеспечивает необходимую достоверность прогноза.

В задачах прогнозирования, решаемых вероятностными методами, все возможные состояния объекта исследования образуют пространство, в котором с вероятностью $P(x)$ случайно и независимо появляются события x, которые надо отнести к одному из K состояний. Это отнесение делается согласно условной вероятности $P(\omega | x)$, где ω – номер класса. Пусть определено множество классификаторов $\Omega = \{F(x,\alpha)\}$, где α – номер конкретного классификатора. В случае дискретного пространства событий $X = \{x_1,...,x_n\}$ определяется вероятность ошибки (функционал качества):

$$P(\alpha) = \sum_{\omega=1}^{K} \sum_{i=1}^{n} (\omega - F(x_i,\alpha))^2 P(x_i) P(\omega | x_i).$$

Среди всех функций $F(x,\alpha)$ надо найти такую $F(x,\alpha_0)$, которая минимизирует ошибку или находится близко к ней. Однако, совместное распределение $P(x,\omega) = P(x_i)P(\omega | x_i)$ исследователю не известно. Поиск ведется только с использованием обучающей выборки $\{x_1\omega_1,...,x_l\omega_l\}$ длины l.

Таким образом, при использовании вероятностного подхода, общая постановка задачи классификации выглядит следующим образом: для любой функции $P(x,\omega)$ среди класса функций $F(x,\alpha)$ найти по обучающей выборке фиксированной длины l такую функцию $F(x,\alpha_*)$, о которой с достоверностью не

меньше $1-\eta$ можно утверждать, что ее качество $P(\alpha_*)$ отличается от качества лучшей функции $F(x,\alpha_0)$ на величину не превосходящую ε. В такой постановке задача относится к классу задач о минимизации величины среднего риска. Чаще других, используются три способа решения.

Первый способ предполагает восстановление по значениям в конечном наборе точек (обучающей выборке) функции распределения $P(x,\omega)$. Наиболее известные алгоритмы – байесовское правило минимизации среднего риска и метод максимального правдоподобия.

Второй способ связан с организацией итерационных процедур по параметру α. Наиболее известные методы – метод стохастической аппроксимации и метод потенциальных функций [Вапник, Червоненкис, 1980].

Третий способ использует замену функционала качества $P(\alpha)$ на эмпирическую оценку, вычисленную на обучающей выборке – алгоритм минимизации эмпирического риска [Vapnik, 1992].

Все перечисленные выше подходы и алгоритмы имеют математическое обоснование и известные границы применения, включающие условия на объем обучающей выборки для достижения необходимой достоверности [Вапник, Червоненкис, 1980].

Общая постановка задачи классификации в рамках комбинаторно-логического подхода формулируется следующим образом. Пусть M – множество объектов наблюдения. Множество M может быть разбито на непересекающиеся подмножества – классы $K_1,...,K_l$. Целиком само разбиение неизвестно, однако в каждом классе есть подмножество элементов, о которых полностью известны их принадлежность и описание (характеристики). Совокупность таких подмножеств всех классов образует обучающую выборку: $T_1,...,T_l,\ T_i \subset K_i,\ i=1,...,l$. Элементы обучающей выборки называются *эталонами*. Каждый элемент множества M характеризуется набором n

признаков, образующих *признаковое пространство*. Каждый признак принимает либо числовое значение, либо набор числовых значений.

Для элементов множества M не из обучающей выборки принадлежность к классу не известна. Для распознавания (классификации) предъявляется элемент множества M не входящий в обучающую выборку. Требуется классифицировать этот элемент, то есть отнести его к одному из существующих классов, представленных обучающей выборкой.

Наиболее часто при решении задач распознавания применяется случай двух классов. С алгоритмической точки зрения это является минимальной задачей, ее решение обеспечивает распознавание для произвольного количества классов путем применения итеративной процедуры. В практических задачах количество классов часто ограничено исторической глубиной обучающей выборки, для достижения требуемой достоверности необходима ее репрезентативность.

Важным этапом развития теории распознавания было появление комбинаторно-логических методов. Их сильной стороной является использование для решения задачи распознавания минимальных требований к описаниям объектов. От множеств описаний не требуется каких-либо свойств вероятностного или метрического характера. Основой этих методов является понятие теста, введенное С.В. Яблонским [Яблонский, 1955]. *Тестом* называется такой набор признаков, что для любой пары элементов обучающей выборки из разных классов имеется различие между этими элементами хотя бы по одному признаку из этого набора.

Для булевских векторов $x=(x_1,..x_n)$ и $y=(y_1,..y_n)$ определим операцию $x\circ y=(x_1y_1,...,x_ny_n)$ – поразрядное умножение. Тестом для обучающей выборки $T_1,...,T_l$, называется такой булевский вектор $t=(t_1,..t_n)$,что для любой пары $x\in T_i$, $y\in T_j$, $i,\ j=1,...,l,\ i\neq j$ имеет место условие

$$t\circ x\neq t\circ y.$$

Пусть $W = W(T_1,...,T_l)$ – множество всех тестов для обучающей выборки $T_1,...,T_l$. Вектор $w = (w_1,...w_n)$, для которого

$$w = \frac{1}{|W|}\sum_{t \in W} t$$

называется *вектором информационных весов* для соответствующей обучающей выборки [Алешин, 1996].

Информационный вес задает меру *важности признака*. Если признак входит в наибольшее количество тестов, то удаление этого признака приводит к наибольшей потере информации о различии состояний процесса. Следовательно, чем больше информационный вес, тем существеннее данный признак для распознавания [Константинов, Королева, Кудрявцев, 1976].

Тест различает эталоны, принадлежащие разным классам. Его использование для оценки принадлежности распознаваемого элемента одному из классов приводит к процедурам *голосования* – тестовым алгоритмам распознавания образов.

В общем случае процедура голосования требует выполнения следующей последовательности действий.

1. Определение понятия *близости* между двумя однородными признаками и построение процедуры сравнения

Значения признака представляет собой действительное число, либо набор чисел. Необходимо задать условия, при которых эти значения эксперт считает одинаковыми. Например, если абсолютная величина разности сравниваемых признаков не превышает порога, то полагается, что признаки одинаковые в смысле проявления их физической природы в рассматриваемой задаче. Эксперт может задать пороговое значение в явном виде, или может указать только диапазоны изменения порогового значения, а его автоматический выбор производится в рамках работы прогнозной системы по определенному критерию качества распознавания.

Таким образом, для работы алгоритма распознавания эксперт вырабатывает логическое правило сравнения любых элементов множества M по признаку p – *функцию близости* $B_p(Y,\ Z)$, $p = 1,...,n$, где Y, Z – элементы M. Если эта величина превышает пороговое эвристическое значение, то полагается, что различие по признаку p в этой паре элементов есть, в противном случае оно отсутствует:

$$B_p(Y,\ Z) \geq \delta_p \text{ – различие есть;}$$

$$B_p(Y,\ Z) < \delta_p \text{ – различия нет,}$$

где δ_p – пороговое значение по *p*-му признаку.

2. Определение множества тестов, по которым будет производиться голосование – *опорного множества*

В классическом подходе [Андреев, Гасанов, Кудрявцев, 2007] теории распознавания в качестве опорного множества выбирают либо вообще все тесты, построенные для обучающей выборки, либо специальные подмножества тестов (например, тесты одинаковой длины). В существенной степени это зависит от числа признаков и объема обучающей выборки: если число признаков большое (порядка 100), то обычно используется одно из специальных подмножеств; для небольшого числа признаков (порядка 10) целесообразно использовать полное множество тестов.

3. Выбор алгоритма распознавания

Выбор алгоритма распознавания обусловлен различными способами определения *числа голосов* за класс (в том числе и по специальным подмножествам) и, наконец, выбором решающего правила. Логическая процедура голосования имеет различные модификации по решающему правилу: от «жесткого» – идеальное соответствие распознаваемого элемента эталону, до «мягкого» – неполное соответствие эталону; прогнозная система предлагает пользователю выбор характера процедуры.

Первая модификация представлена алгоритмом голосования по тестам [Кудрявцев, 2006]. Фиксируется тест t и эталон Γ_i из обучающей выборки. Единичные значения координат теста определяют набор признаков, по которым сравнивается распознаваемый элемент X и эталон Γ_i. Если в каждой единичной координате теста значение признака распознаваемого элемента «совпало», то есть они равны в смысле процедуры сравнения с признаками эталона, то полагается присвоить один голос тому классу, к которому принадлежит эталон:

1. $g(X,\Gamma_i,t)=1$, если для каждого p-го признака такого, что p-я координата t равна 1, p-й признак распознаваемого элемента X «совпал» с p-м признаком эталона Γ_i;

2. $g(X,\Gamma_i,t)=0$, если хотя бы для одного p-го признака такого, что p-я координата t, равная 1, p-й признак распознаваемого элемента X «не совпал» с p-м признаком эталона Γ_i.

Производится суммирование голосов по всем тестам и всем элементам из обучающей выборки T_j класса K_j, $j=1,...,l$:

$$G_j=\frac{1}{|T_j|}\sum_{t\in W}\sum_{\Gamma_i\in T_j} g(X,\Gamma_i,t),$$

где

$$g(X,\Gamma_i,t)=\prod_{p=1}^{n}\left(1-t_p\theta(B_p(X,\Gamma_i)-\delta_p)\right),$$

$$\theta(x)=\begin{cases}0, & x<0\\ 1, & x\geq 0\end{cases}$$ – функция Хэвисайда,

n – число признаков, $t=(t_1,...,t_n)$ – тест, X – распознаваемый элемент [Малыгин, 2014 «*О задаче…*»].

Решающее правило F_1 алгоритма заключается в отнесении текущего распознаваемого элемента к тому классу, за который подано больше голосов:

$$F_1(X) = j_0, \text{ где } G_{j_0} = \max_j G_j.$$

Вторая модификация представлена алгоритмом вычисления оценок [Журавлев, Никифоров, 1971]. Как и выше, фиксируется тест t и эталон Γ_i из обучающей выборки. Единичные значения координат теста определяют набор признаков, по которым сравнивается распознаваемый элемент X и эталон Γ_i. Считается количество единичных координат теста таких, что значение признака распознаваемого элемента «не совпало», то есть они отличаются в смысле процедуры сравнения с признаками эталона. Таким образом, определяется общее число «не совпадений» $v(X, \Gamma_i, t)$ эталона Γ_i и распознаваемого элемента X по фиксированному тесту t. По смыслу $v(X, \Gamma_i, t)$ выражает число голосов «против» принадлежности распознаваемого вектора X к классу, которому принадлежит эталон Γ_i. Далее, производится суммирование значений v по всем тестам и всем элементам из обучающей выборки T_j класса K_j, $j = 1,...,l$:

$$V_j = \frac{1}{|T_j|} \sum_{t \in W} \sum_{\Gamma_i \in T_j} v(X, \Gamma_i, t),$$

где

$$v(X, \Gamma_i, t) = \sum_{p=1}^{n} t_p \theta(B_p(X, \Gamma_i) - \delta_p).$$

Решающее правило F_2 алгоритма заключается в отнесении текущего распознаваемого элемента к тому классу, за который подано меньше голосов «против»:

$$F_2(X) = j_0, \text{ где } V_{j_0} = \min_j V_j.$$

На техническом уровне представленной теоретической базе хорошо соответствует класс *экспертных систем*[2] *(ЭС)*, т.е. специализированных компьютерных приложений, способных частично заменить человека-эксперта при принятии решения в конкретной предметной области [Элти, Кумбс, 1987; Уотермен, 1989; Harmon, Sawyer, 1990; Джексон, 2001; Джарратано, Райли, 2007].

Разработка прогнозных систем с использованием комбинаторно-логического подхода теории распознавания образов основана на представлении исследуемого процесса как элемента одного из выделенных классов. Совокупность классов является полным множеством развития процесса. Функционирование прогнозной системы – моделирование решения задачи классификации так, как это делал бы человек-эксперт (группа экспертов) в рамках установленного признакового пространства.

Необходимым условием функционирования системы является ее обучение. Обучение системы рассматривается как автоматизированный поиск наборов параметров, с помощью которых алгоритм классификации принимают верные решения на известных данных (эталонах). Процедура нахождения одного набора параметров выглядит следующим образом [Малыгин, 2013].

Формируются пороговые значения признаков: человек-эксперт указывает диапазоны изменения пороговых значений признаков $(\delta_{i\ \min,}\ \delta_{i\ \max})$. Сами пороговые константы δ_i выбираются либо полным перебором возможных вариантов, либо случайным образом из интервала изменения *i*-го признака с шагом, кратным единице измерения признака. Полный перебор по всем возможным пороговым числовым значениям $(\delta_1,...,\delta_n)$ с заданным шагом h_i потребует следующее количество итераций:

[2] В настоящее время термин «экспертные системы» считается устаревшим, в данной работе наряду с ним используется и более широкий термин «прогнозные системы».

$$\left[\prod_{i=1}^{n}\left(\frac{\delta_{i\ \max}-\delta_{i\ \min}}{h_i}+1\right)\right].$$

Величину шага требуется выбирать достаточно малой для дискретной аппроксимации признакового пространства, поэтому количество итераций может быть настолько большим, что время, необходимое для полного перебора узлов построенной сетки, выведет прогнозную систему за рамки реального применения. По этой причине в практических задачах поиск пороговых значений признаков целесообразно производить методом Монте-Карло, количество испытаний которого является входным параметром прогнозной системы.

Далее, производится проверка корректности найденных пороговых значений признаков методом скользящего поиска. На каждой итерации из обучающей выборки последовательно удаляется каждый эталон, который подается на распознавание. Остальные эталоны образуют новую обучающую выборку. Производится распознавание (с использованием тестовых алгоритмов) удаленного эталона и соотнесение его с фактической классификацией. После обработки всех эталонов вычисляется оценка достоверности, выраженная отношением правильно распознанных эталонов к общему их числу.

Далее проверяется критерий оптимальности набора пороговых значений. В качестве таких критериев могут использоваться максимум достоверности, достижение установленного уровня достоверности. Оптимальные наборы $\Delta_j = (\delta_{j1},...,\delta_{jn})_{\text{opt}},\ j = 1,...,s$ сохраняются и используются при дальнейшем функционировании прогнозной системы. После этого процедура обучения переходит к следующему случайному испытанию. Результатом работы всей процедуры обучения является определение наборов оптимальных пороговых значений признаков.

Следующим шагом работы системы является определение сценария исследуемого процесса. Производится цикл локальных процедур распознавания. На каждом шаге проводится тестовая процедура с использованием найденного

оптимального набора пороговых значений признаков. По каждому найденному набору параметров происходит процедура голосования за принадлежность процесса к одному из сценариев. Итоговым результатом распознавания является интегральная характеристика локальных распознаваний – моделирование принятия решения экспертной группой.

Анализ показывает функциональное различие блоков обучения и распознавания. При фиксированной обучающей выборке процедура обучения проводится однократно с целью построения набора оптимальных пороговых значений. Процедура определения сценария только использует этот набор, поэтому при вариации исходных данных распознаваемого процесса обучение проводить не требуется. Переобучение системы проводится при изменении обучающей выборки.

В качестве локальных решающих правил прогнозной системы использовалась решающие правила алгоритма голосования, алгоритма вычисления оценок и их комбинации.

Алгоритм голосования является универсальным, т.к. всякая булевская функция правильно распознается с его помощью [Алешин, 1996; Андреев, Гасанов, Кудрявцев, 2007]. Это значит, что если в процессе скользящего поиска последовательно не удалять каждый эталон из обучающей выборки, то вся выборка распознается верно. Многие эвристические алгоритмы распознавания не удовлетворяют условию правильного распознавания материала обучения. Для положительного голосования по тесту и эталону необходимо абсолютное совпадение распознаваемого элемента с эталоном в смысле функции близости. Этот факт является достаточно сильным условием, приводящим к тому, что эталон не всегда голосует по построенным тестам; это может приводить к вычислению одинакового числа голосов за различные классы. Однако, если алгоритм голосования обеспечил принятие решения для одного класса, то ввиду его сильных требований, такое решения является более качественным.

Алгоритм вычисления оценок задает линейное относительно функции близости решающее правило [Алешин, 1996], что накладывает существенные ограничения на структуру множества эталонов для его правильного распознавания. В случае двух классов эталоны как точки многомерного пространства признаков должны быть линейно отделимы гиперплоскостью. Эта особенность алгоритма преодолевается построением набора гиперплоскостей, порожденных различными векторами Δ_i, каждый из которых обеспечивает требуемый уровень распознавания.

В случае комбинации алгоритмов, в качестве основного был выбран алгоритм голосования. Однако, в случаях, когда алгоритм голосования не может произвести распознавание, используются результаты алгоритма вычисления оценок, который на практике позволяет минимизировать количество не классифицированных элементов.

При проведении процедуры обучения к каждому случайному набору $\delta_i = (\delta_{i1},...,\delta_{in})$ применяются три описанных подхода. В каждом из них из обучающей выборки последовательно удаляются эталоны, и производится их распознавание. Определяется количество правильно и неправильно распознанных эталонов посредством сравнения мер голосов за каждый класс. В случае совпадения на эталоне из обучающей выборки мер голосов за классы, такой эталон не относится ни к числу правильно распознанных, ни к числу неправильно распознанных.

В случае распознавания, если алгоритмы или их комбинация не могут принять решение на оптимальных векторах Δ_i, выбор предоставляется эксперту, либо производится принудительное отнесение к одному из классов.

Моделирование принятия решения экспертной группой – интегральная характеристика локальных решающих правил осуществляется процедурой определения большинства голосов экспертов, классифицировавших распознаваемый элемент. Такой общепринятый подход предполагает

увеличение достоверности принятия решения экспертной группой с ростом количества экспертов, т.е. принятие решения экспертной группой путем подсчета голосов не ухудшает достоверность итогового результата распознавания по сравнению с достоверностью каждого отдельного участника. В терминах прогнозной системы это значит, что в случае генерации в процессе обучения n оптимальных векторов Δ_i с достоверностью каждого не менее, чем Level $(\mathrm{Level} \geq 50\%)$, итоговое принятие решение произойдет также с достоверностью не менее, чем Level.

ГЛАВА 2. Построение 2D и 3D моделей региона методами машинного обучения

Во второй главе приведены результаты, полученные с помощью методов машинного обучения в применении к задаче пространственной интерполяции геофизических полей. Рассмотрено три примера геофизических приложений.

В задаче анализа строения коры северной части Балтийского щита построена уточненная карта поверхности Мохоровичича. Основу исследования составляют данные, полученные методом приемных функций. Были использованы сведения, полученные в предыдущих исследованиях этого региона, дополненные новыми расчетами и измерениями. Исходные данные представляют собой набор зависимостей сейсмической скорости от глубины, рассчитанных для более чем 60 постоянных и временно действующих геофизических станций. С точки зрения машинного обучения, данная задача является задачей регрессии. Для восстановления регрессионной зависимости глубины Мохо от двумерных координат был использован метод k ближайших соседей с необходимой адаптацией в части выбора метрики.

Еще одним результатом в данной задаче стало построение карты слоя с низкими значениями скорости поперечных сейсмических волн V_S. В исследуемом регионе практически отсутствует осадочный слой. Несмотря на это, имеются области, в которых присутствует слой с низкими значениями скорости V_S. Относительно низкие значения V_S обычно объясняют наличием в слое большого количества водонасыщенных трещин. Присутствие такого слоя не зависит от возраста пород. Эта задача относится к оценке принадлежности в рамках задачи бинарной классификации. Для небольшого количества сейсмических станций (порядка 20) известно наличие или отсутствие слоя низких скоростей. С помощью метода k ближайших соседей в каждой точке исследуемого региона оценена вероятность наличия слоя низких скоростей. Построенная карта включает в себя классификацию по принципу наличия или отсутствия слоя низких скоростей, а также буферную область, в которой на

основании имеющихся данных нельзя сделать однозначный вывод. Показано, что слой низких сейсмических скоростей на поверхности присутствует на значительной части региона, включая области с протерозойскими породами. В южной части Финляндии положение низкоскоростной области коррелирует с относительно низким значением толщины коры.

В задаче построения трехмерной модели среды при проведении межскважинных исследований предложена новая интерпретация данных радиоволнового просвечивания, позволяющая более точно выделить границы слоев по сравнению с методами, используемыми ранее (кригинг). В настоящий момент значительно возросла глубина работ по разведке кимберлитовых тел и рудных месторождений. Традиционные геологические методы поиска оказались неэффективными. На практике единственным прямым методом поиска является бурение системы скважин до глубин, которые обеспечивают доступ к вмещающим породам. Из-за высокой стоимости бурения возросла роль межскважинных методов. Они позволяют увеличить среднее расстояние между скважинами без существенного снижения вероятности пропуска кимберлитового или рудного тела. Среднее расстояние между ближайшими скважинами составляет 200 м.

Метод радиоволнового просвечивания особенно эффективен при поиске объектов, отличающихся высокой контрастностью электропроводящих свойств. Физическую основу метода составляет зависимость распространения электромагнитной волны от проводящих свойств среды распространения. Источником и приемником электромагнитного излучения является электрический диполь. При измерениях они размещаются в соседних скважинах, а расстояние между источником и приемником известно. Поэтому измерив величину уменьшения амплитуды электромагнитной волны при ее распространении между скважинами можно оценить коэффициент поглощения среды. Породе с низким электрическим сопротивлением соответствует высокое

поглощение радиоволн. Поэтому данные межскважинных измерений позволяют оценить эффективное электрическое сопротивление породы.

Обычно источник и приемник синхронно погружаются в соседние скважины. Измерение величину амплитуды электрического поля в приемнике позволяет оценить среднее значение коэффициента затухания на линии, соединяющей источник и приемник. Измерения проводятся во время остановок, приблизительно каждые 5 м. Расстояние между остановками значительно меньше расстояния между соседними скважинами. Это приводит к значительной пространственной анизотропии в распределении данных. При проведении разведочного бурения скважины покрывают большую площадь. Задача состоит в построении трехмерной модели распределения электрических свойств межскважинного пространства на всем участке по результатам совокупности измерений. Анизотропия пространственного распределения препятствует использованию стандартных методов геостатистики.

Основными направлениями, за счет которых стало возможным получить новые результаты, стали применение и адаптация метода ближайших соседей в ситуации дефицита исходных данных измерений, а также учет пространственной анизотропии геофизических процессов.

2.1. Построение карты толщины коры северной части Балтийского щита

Проведение пассивных сейсмических экспериментов SVEKALAPKO [Bock et al., 2001] и POLENET/LAPNET [Kozlovskaya et al., 2006] послужило основой большого количества работ по изучению строения литосферы северной части Балтийского щита [Bruneton, Farra, Pedersen, 2002; Alinaghi et al., 2003; Aleshin et al., 2006; Алешин и др., 2007; Vecsey et al., 2007; Kozlovskaya et al., 2008; Grad, Tiira, 2012; Uski et al., 2012; Pedersen et al., 2013; Silvennoinen et al., 2014; Vinnik et al., 2016]. Значительная часть этих исследований основана на применении метода приемных функций [Alinaghi et al., 2003; Aleshin et al., 2006; Алешин и др., 2007; Kozlovskaya et al., 2008; Grad, Tiira, 2012; Vinnik et al., 2016]. Вместе с широким покрытием сейсмическими станциями, это позволило

получить ряд важных результатов о строении мантии северной и южной Финляндии [Alinaghi et al., 2003; Frassetto, Thybo, 2013; Vinnik et al., 2016], построить трехмерную сейсмическую модель южной Финляндии [Kozlovskaya et al., 2008]. Кроме сейсмических исследований были получены распределения электрического сопротивления [Korja et al., 2002] и аномалий магнитного поля [Lahtinen, Korja, Nironen, 2005].

Среди исследований важное место занимает изучение формы границы Мохоровичича [Kosarev, Makeyeva, Vinnik, 1987; Alinaghi et al., 2003; Kozlovskaya et al., 2008; Silvennoinen et al., 2014]. Толщина коры, определяемая как расстояние от поверхности до этой границы, является основной характеристикой при анализе строения региона, а также при изучении структуры европейской литосферы [Grad, Tiira, 2012; Uski et al., 2012]. Существуют исследования, в которых этот вопрос исследовался на основе данных сейсмических экспериментов с управляемым источником [Grad, Luosto, 1992; Janik, Kozlovskaya, Yliniemi, 2007; Uski et al., 2012; Минц и др., 2018]. В статье [Silvennoinen et al., 2014] расчет топографии границы Мохо под северной частью Финляндии выполнен с учетом данных обоих типов. В данной задаче, используются данные, полученные исключительно на основе метода приемных функций. Это обусловлено тем, что разные методы сейсмических исследований могут привести к отличающимся значениям искомой величины. Во-первых, в исследованиях с управляемым источником наиболее надежно определяются значения скорости продольных волн, а функции приемника, наоборот, наиболее чувствительны к скоростям поперечных волн. Во-вторых, в этих методах существенно отличается частотный состав сейсмического сигнала. В настоящее время отсутствует надежно обоснованный метод совместной интерпретации приемных функций и экспериментов с управляемым источником. Также отличие может быть связано с сейсмической анизотропией верхней мантии региона, которая может составлять 3-5% [Яновская, Лыскова, Королева, 2019].

При построении пространственных распределений не использовались традиционные линейные методы интерполяции. В задачах такого рода использование кригинга, как это описано в статье [Kozlovskaya et al., 2008], или интерполяция сплайнами [Horspool, Savage, Bannister, 2006], нельзя считать оптимальным уже потому, что решение получается излишне сглаженным, что может исказить реальную форму поверхности. Был использован метрический классификатор, реализованный на основе алгоритма *k*-ближайших соседей. Функционирование этого алгоритма подробно приведено в п. 1.3.

Для построения карты глубины Мохо использованы значения, полученные методом приемных функций. Основу составляют результаты, полученные в рамках проектов SVEKALAPKO и POLENET/LAPNET, статьи [Kozlovskaya et al., 2008] и [Silvennoinen et al., 2014] соответственно. Кроме того, использованы сведения для станций APA и LVZ из статьи [Dricker et al., 1996]. Значения глубины границы Мохо для станций PITK и KEMI было определено в рамках отдельного исследования [Алешин и др., 2019], а для станции RUKSA использовались более ранние результаты [Aleshin et. al., 2006]. Карта региона представлена на рис. 2.1.

Был произведен расчет и последующая инверсия *P*- и *S*-приемных функций. В табл. 2.1 Приложения 1 приведены численные значения толщины коры, при этом координаты сейсмической станции могут не совпадать с координатами, приведенными в таблице, так как последние содержат поправки за наклонное падение волн. В табл. 2.1 Приложения 1 приведены также сведения о наличии или отсутствии тонкого низкоскоростного слоя под станциями. В методе приемных функций находящийся под станцией слой пониженной скорости поперечных волн проявляется наличием фазы, вступающей в течении первой секунды после прихода основной *P*-волны. Так, например, в статье [Kozlovskaya et al., 2008] на рис. 2 такие фазы помечены черными кругами. Простое, визуальное выделение сигнала от неглубокой границы возможно лишь при использовании системы координат *L*-*Q*, введенной в работе [Vinnik, 1977].

Переход к этим компонентам от вертикальной и радиальной осуществляется поворотом координат в плоскости падения так, чтобы вертикальная ось совпала с основным смещением в падающей *P*-волне (ось *L*). Тогда перпендикулярная ей ось *Q* будет оптимальной, для выделения поперечных волн. Без такого преобразования координат обменная волна, образованная на мелкой границе, будет

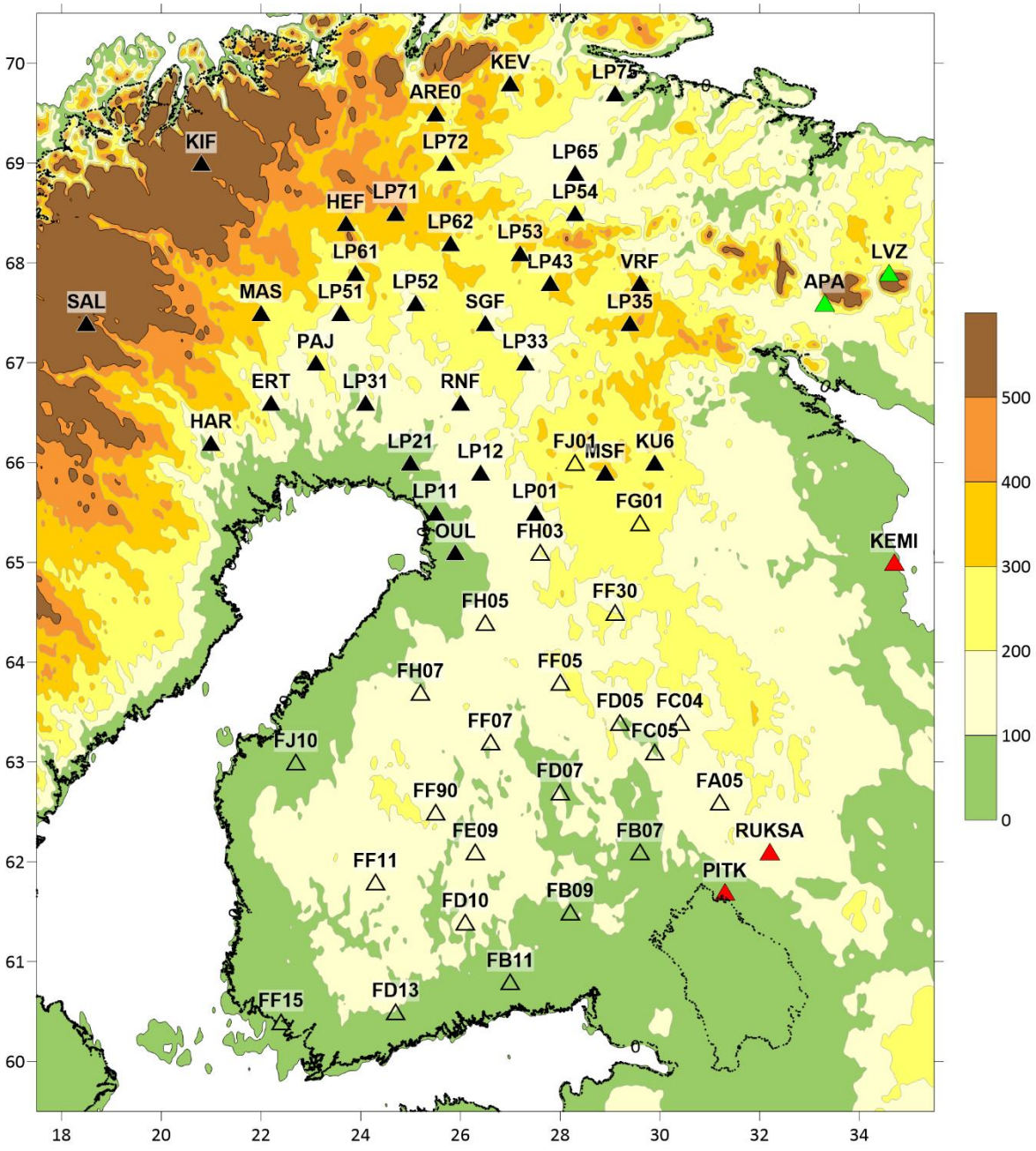

Рис. 2.1. Карта исследуемого региона. Цветовая шкала отражает высоту над уровнем моря в метрах. Треугольниками обозначены сейсмические станции. Цвет

станции соответствует использованному литературному источнику: черный – POLENET/LAPNET Silvennoinen et. al., 2014; белый – SVEKALAPKO Kozlovskaya et. al., 2008; зеленый – Dricker et. al., 1996; красный – Aleshin et. al., 2006; Алешин и др., 2019.

замаскирована проекцией вступления исходной P-волны. Для трех станций – FB11, FD07 и FF05 – не удается сделать определенного вывода в пользу наличия или отсутствия соответствующей фазы, а значит, и слоя. В результате анализа исходных данных получен набор точек с заданными значениями булева типа, представленных в табл. 2.1 Приложения 1.

Для построения пространственного распределения толщины коры использован метод k ближайших соседей (п. 1.3.). Построение пространственного распределения глубины границы Мохоровичича возможно рассматривать как задачу нелинейной регрессии, где признаками (зависимыми переменными) являются 2D-координаты геофизических станций (например, географические широта и долгота), а целевой переменной – глубина поверхности Мохо, выраженная в километрах.

Для решения описанных выше задач с помощью алгоритма kNN, нужно определить значение свободных параметров. В данной задаче этот параметр только один – число ближайших соседей k. Для расчета карты толщины коры имеются данные измерений для A=61 станций. Для определения оптимального числа ближайших соседей была использована техника Leave-one-Out (п. 1.2.): при фиксированном значении параметра k из исходных данных удалялось одно значение q_a, измеренное в точке Ω_a. Оставшиеся данные использовались для расчета значения глубины Мохо в точке $Q_a = Q_a(\Omega_a, k)$, которая была удалена из выборки. Затем вычислялось абсолютное значение разницы вычисленного и наблюденного значений

$$\mu_a(k) = |Q(\Omega_a, k) - q_a|.$$

Эта процедура повторялась с каждым из A значений. Качество интерполяции определялось функционалом качества *MAE* как среднее абсолютное отклонение:

$$\mu(k) = 1/A \sum_{a=1}^{A} \mu_a(k).$$

На рис. 2.2. приведена зависимость $\mu(k)$, рассчитанная для использованных данных.

Рис. 2.2. На рисунке показана зависимость средней абсолютной ошибки, полученной в процедуре кросс-валидации, от числа ближайших соседей *k*. Из графика видно, что ошибка минимальна, когда *k*=4. При этом ошибка интерполяции составляет 3.7 км.

Из рисунка видно, что оптимальное значение числа соседей равно четырем. Средняя ошибка интерполяции при этом составляет 3.7 км. Соответствующая карта толщины земной коры приведена на рис. 2.3.

Из построенной карты (рис. 2.3) видно, что в северо-восточной части региона имеется область с относительно ровной тонкой корой порядка 45 км [Silvennoinen et al., 2014]. Использованные дополнительные данные позволяют утверждать, что эта область простирается еще восточнее с уменьшением мощности коры до величин менее 40 км. Максимальная толщина коры достигается в центре южной части региона в соответствии с результатами, изложенными в [Alinaghi et al., 2003; Kozlovskaya et al., 2008]. В этих работах

отмечалась также корреляция поверхности Мохо с формой Центрального финского гранитоидного комплекса. Нелинейная природа процедуры, использованной при построении, позволяет выявить эту корреляцию более отчетливо.

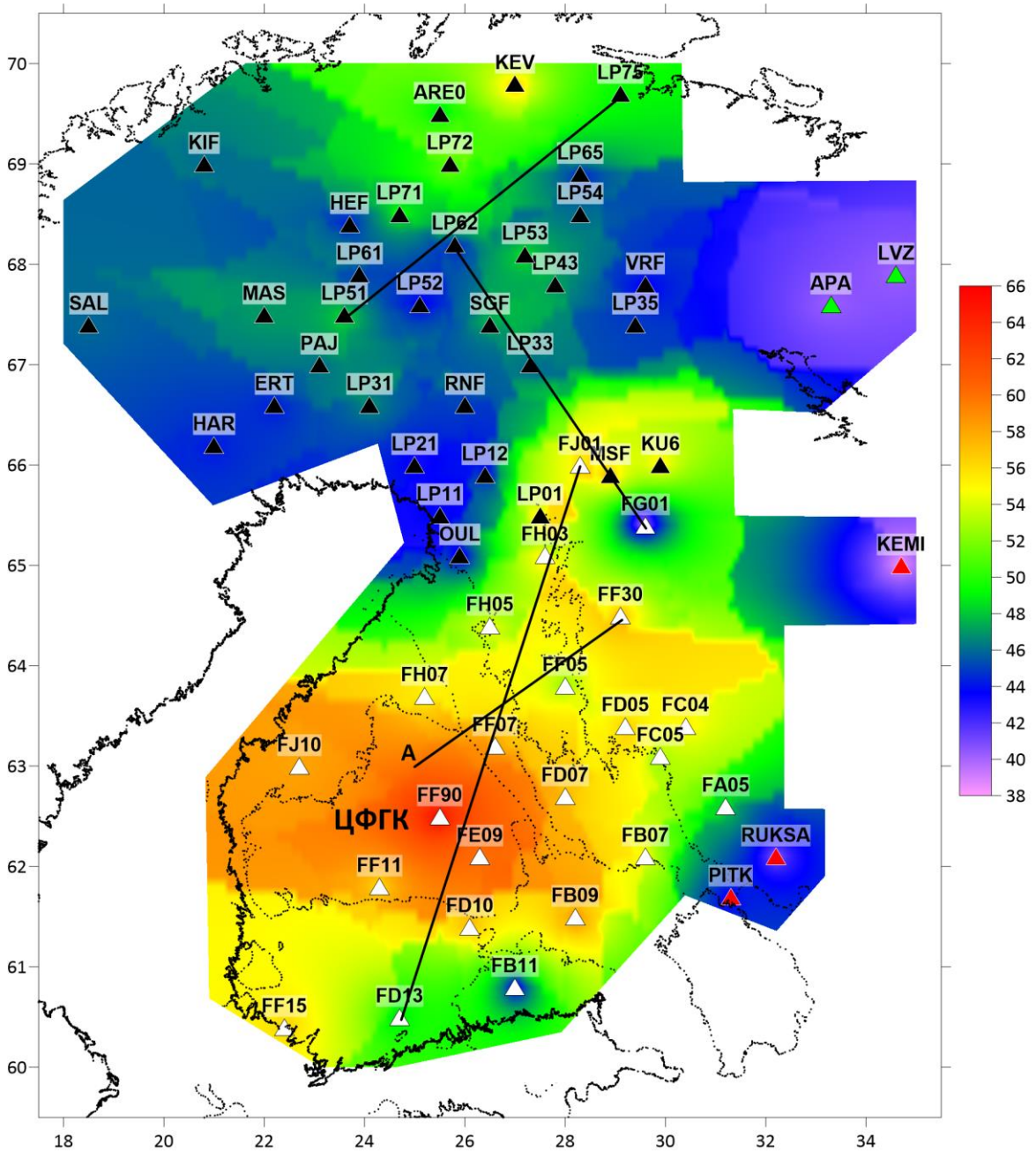

Рис. 2.3. Карта глубины границы Мохоровичича. Треугольниками отмечены точки, в которых положение границы было измерено. Буквенные коды означают названия сейсмических станций. Аббревиатура ЦФГК означает Центральный

финский гранитоидный комплекс. Линиями показаны сейсмические профили, сравнение с которыми приведено в настоящей главе ниже.

Дополнительно было проведено построение пространственного распределения ошибки интерполяции (рис. 2.4).

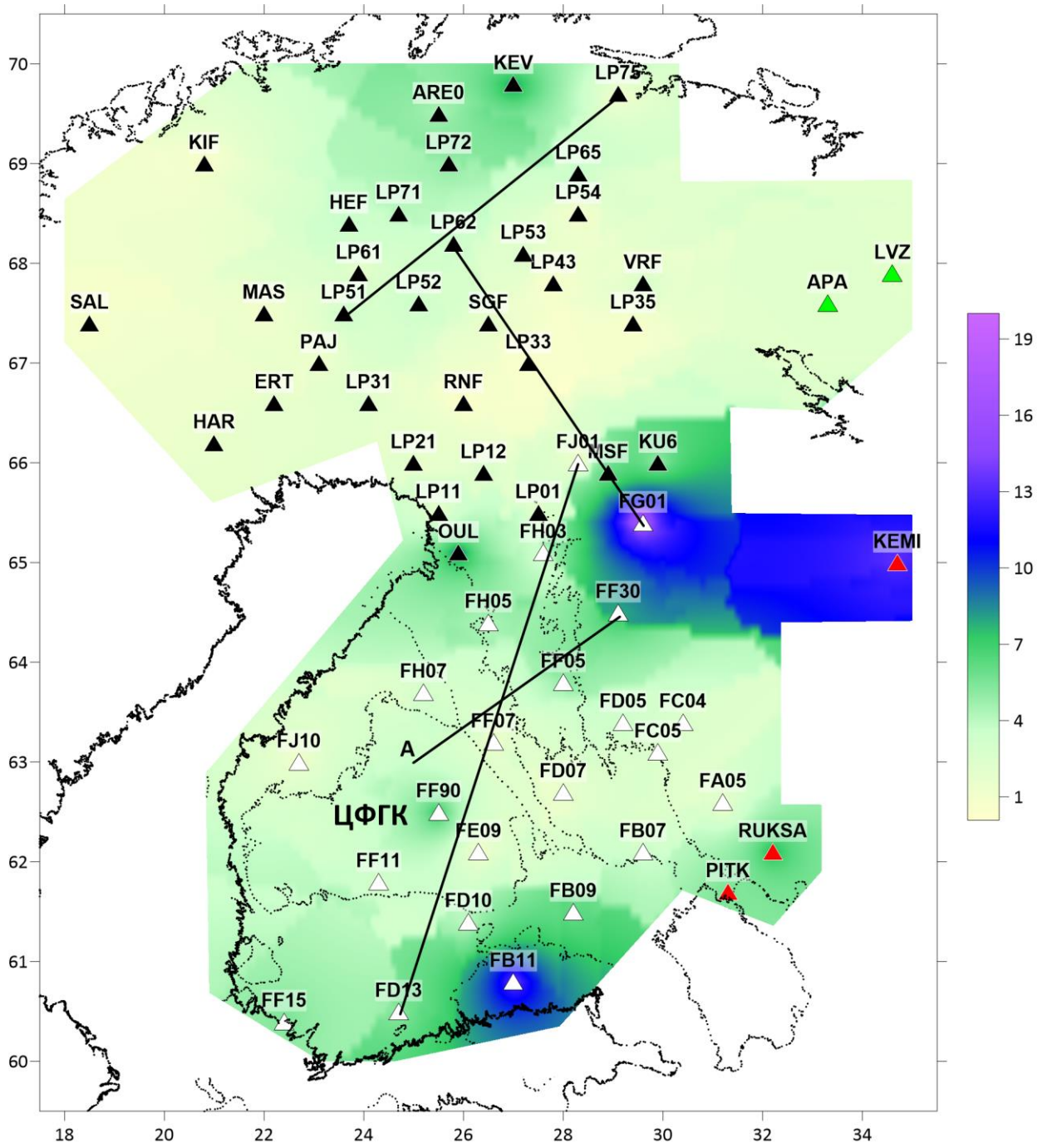

Рис. 2.4. Пространственное распределение ошибки интерполяции.

Визуализируется две области с повышенной ошибкой интерполяции: окрестность

станции FB11 на юге региона, аномальная область вокруг станции FG01 в центре региона и KEMI на востоке.

Проведено графическое сравнение глубины Мохо, построенной в рамках настоящего диссертационного исследования, с ранее полученными результатами по четырем профилям: по два профиля в южной и северной части исследуемого региона соответственно. Эти профили изображены черными прямолинейными отрезками на рис. 2.3 или 2.4: LITHOSCOPE (FD13-FJ01), FIRE 1 (FF30-A), HUKKA2007 (LP62-FG01), POLAR (LP51–LP75). На представленных далее (на рис. 2.5–2.8) сравнениях результаты настоящего исследования показаны красным цветом; красными пунктирными линиями дополнительно показаны оцененные границы ошибки построения вдоль каждого профиля.

На рис. 2.5. приведено сравнение полученных результатов по профилю FD13-FJ01 и результатов из работы [Kozlovskaya et al., 2008]. В цитируемой работе использовались также данные только приемных функций, однако применялся другой метод интерполяции – кригинг. Видно, что полученные в рамках настоящего исследования методом kNN результаты достаточно хорошо соответствуют предыдущим результатам, однако позволяют выявить локальные неоднородности и, в целом, более следуют за скоростным распределением, что хорошо заметно в левой части профиля.

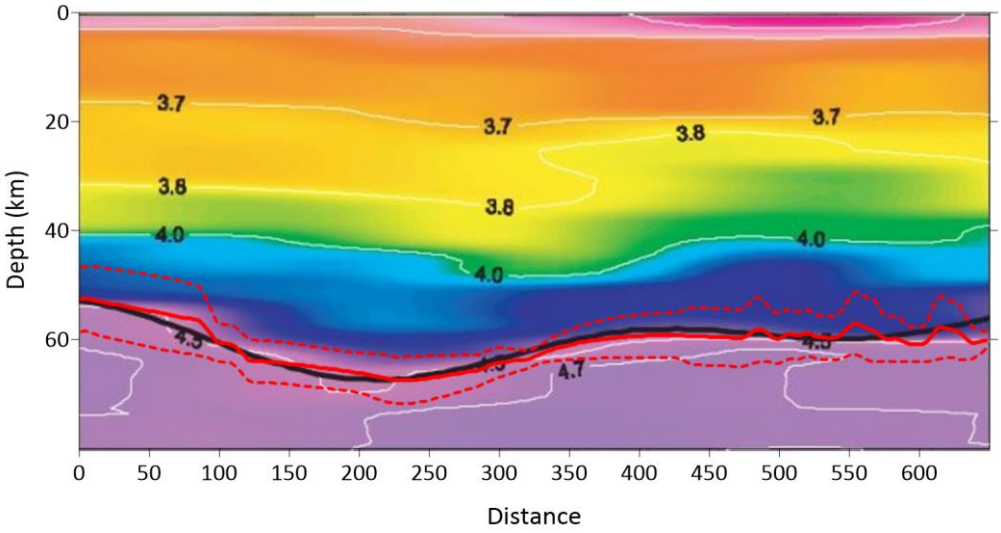

Рис. 2.5. Профиль FD13-FJ01. На профиль границы Мохо из работы [Kozlovskaya et al., 2008] (черная линия) нанесена граница Мохо, построенная методом *kNN* в рамках настоящего исследования (красная линия). Красными пунктирными линиями дополнительно показаны оцененные границы ошибки построения вдоль профиля.

На следующем разрезе (рис. 2.6.) точками показано распределение рассеивающих центров в эксперименте с вибрирующим источником, черной сплошной линией – граница Мохоровичича, определенная методом общей глубинной точки - ОГТ, черной пунктирной линией – результаты построения границы Мохоровичича методом глубинного сейсмического зондирования (ГСЗ) для профиля FIRE 1 на рисунке 11a из работы [Janik, Kozlovskaya, Yliniemi, 2007]. Красной линией, по-прежнему, результаты, полученные в рамках настоящего диссертационного исследования. Профиль FIRE 1 аппроксимирован разрезом по линии FF30-A. Видно, что метод ОГТ не всегда дает результаты, как в правой части представленного разреза. Здесь нужно обращать внимание на падение плотности распределения точек с ростом глубины. С результатами метода ГСЗ достигнуто хорошее соответствие.

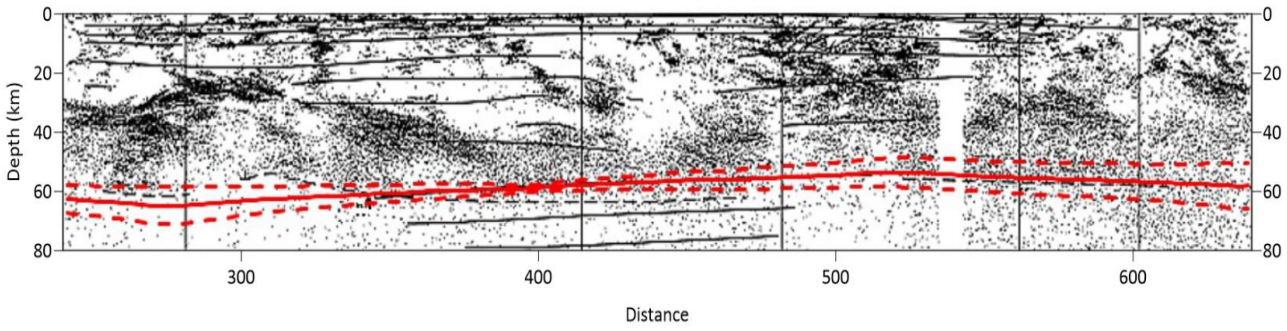

Рис. 2.6. Профиль FF30-A (approx. FIRE 1). На профиль границы Мохо из работы [Janik et al., 2007], черная пунктирная линяя – метод ГСЗ, черная сплошная линяя – метод ОГТ, нанесена граница Мохо, построенная методом *kNN* в рамках настоящего исследования – красная линия. Красными пунктирными линиями дополнительно показаны оцененные границы ошибки построения вдоль профиля.

Далее проведено сравнение результатов из работы [Silvennoinen et al., 2014] с полученными в рамках настоящей работы результатами. Построена глубина Мохо вдоль профиля POLAR, как это сделано на рис. 6 в цитируемой статье. Взято сечение поверхности, построенной методом *kNN*, вдоль профиля POLAR (LP51–LP75), и построена зависимость глубины Мохо от расстояния вдоль этой плоскости. Результат графически наложен на иллюстрацию из статьи (рис. 2.7.). На данном профиле результаты вибросейсмического эксперимента уже позволяют выделить четкую границу – голубая линия на иллюстрации. Результаты метода ГСЗ нанесены малиновой линией. Также авторами исходной публикации были дополнительно рассчитаны доверительные интервалы для каждого построения. Следует отметить, что полученный профиль хорошо согласуется с полученными ранее результатами. По всей длине профиля результаты интерпретации методом *kNN* попадают в границы доверительных интервалов предыдущих построений.

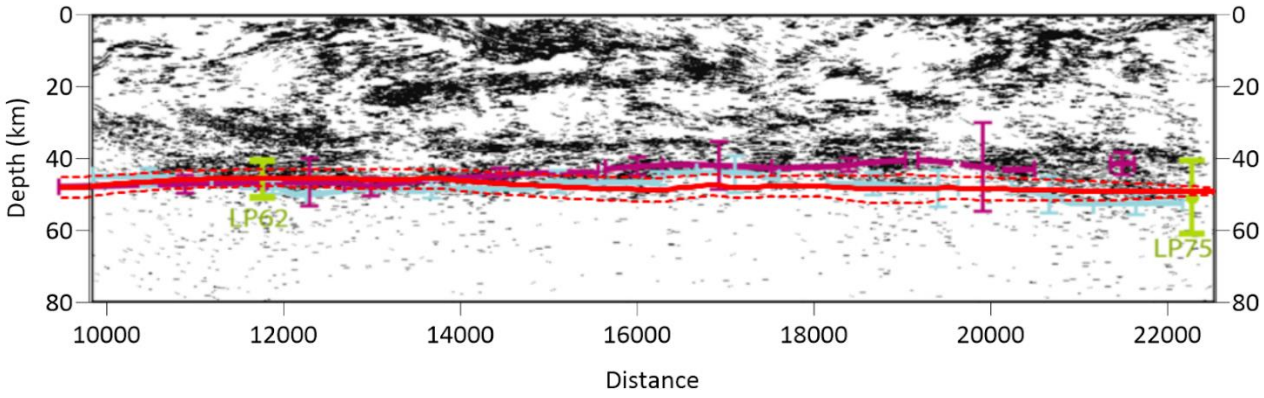

Рис. 2.7. Глубина Мохо вдоль профиля POLAR (LP51–LP75). На рис. 6 из статьи [Silvennoinen et. al., 2014] красной линией наложена глубина Мохо, полученная в рамках настоящего исследования методом *kNN*. Красными пунктирными линиями дополнительно показаны оцененные границы ошибки построения вдоль профиля.

На рис 2.8. показан еще один профиль – HUKKA2007 из работы [Tiira et. al., 2014], аппроксимированный линией LP62-FG01. Результаты построения границы Мохо в рамках настоящего исследования показаны красной линией. На левой части профиля наблюдается достаточно точное соответствие ранее построенной границе. В правой части профиля ранее полученные результаты уточнены с помощью построения методом *kNN*.

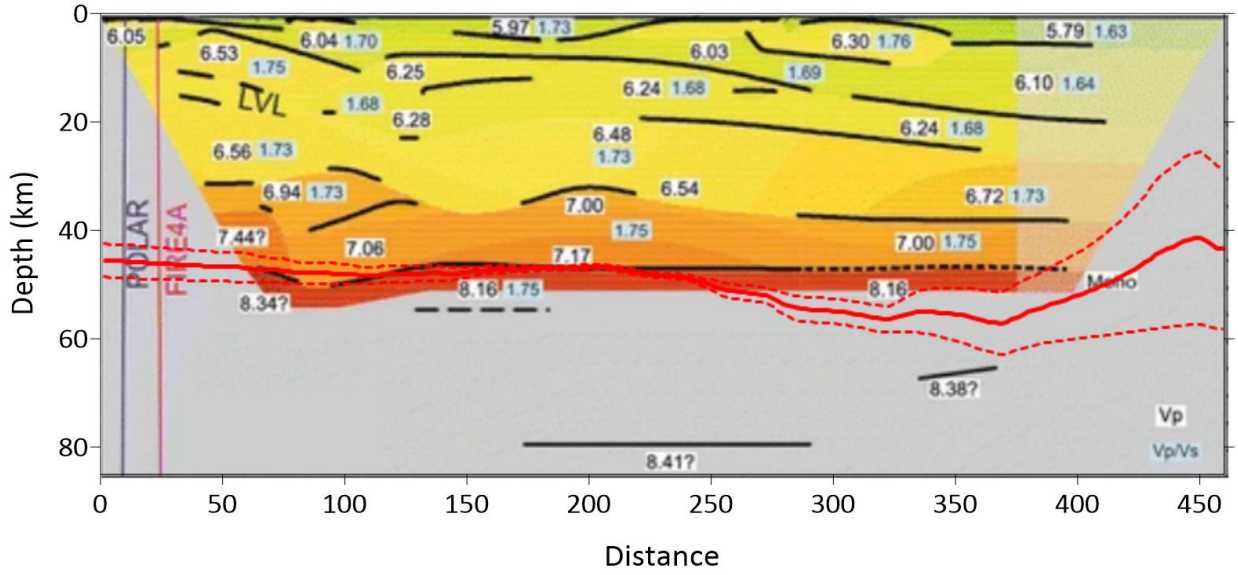

Рис. 2.8. Глубина Мохо вдоль профиля HUKKA2007 (approx. LP62–FG01). На рис. 11 из статьи [Tiira et. al., 2014] красной линией наложена глубина Мохо, полученная в рамках настоящего исследования методом *kNN*. Красными

пунктирными линиями дополнительно показаны оцененные границы ошибки построения вдоль профиля.

2.2. Построение карты слоя с низкими значениями скорости поперечных сейсмических волн для северной части Балтийского щита

Еще одной важной особенностью строения региона является наличие относительно тонкого поверхностного слоя с низкими скоростями поперечных волн *Vs*. Толщина слоя варьируется около значения 1 км, значения *Vs* в нем на 15-30% меньше средних по коре значений. Впервые наличие такого слоя было отмечено в работах [Pedersen, Campillo, 1991; Grad, Luosto, 1992; Grad, Luosto, 1994] при исследовании затухания сейсмических волн. По мнению авторов цитированных работ, природа слоя пониженной скорости обусловлена значительным изменениями механических свойств гнейсов в архейской области Фенноскандии [Grad, Luosto, 1994]. Наличие такого слоя в архейской части региона было подтверждено позднее в работе [Aleshin et al., 2006]. В работе [Kozlovskaya et al., 2008] показано, что наличие такого слоя характерно не только для большей части архейской части южной Фенноскандии. Он присутствует в значительной части региона, сформированного в протерозойский период. Карта исследуемого региона представлена на рис. 2.9.

Построение пространственного распределения слоя пониженной скорости относится к задаче бинарной классификации. Присвоим станциям метки классов. Наличие под станцией слоя пониженной скорости означает, что она принадлежит к классу 1. Если такой слой отсутствует, то станция относится к классу 0 (табл. 2.1 Приложения 1). Задачу можно сформулировать как расчет вероятности принадлежности произвольной точки на поверхности к классу 1. Окончательная классификация осуществляется с помощью заданного порогового значения. Если вероятность превышает порог, то точка относится к классу 1. В противном случае точка классифицируется значением 0. В такой постановке в качестве метрики вместо абсолютного или среднеквадратического

отклонения удобнее использовать величину ROC-AUC (Receiver Operating Characteristic - Area Under Curve) – площадь под ROC-кривой (п. 1.3.). Значение метрики ROC-AUC в задаче бинарной классификации отражает вероятность того, что случайно выбранный объект класса 1 имеет оценку принадлежности к классу 1 выше, чем случайно выбранный объект класса 0. В задачах с малым количеством данных к лучшему результату приводит использование аналога статистических доверительных интервалов – буферной зоны между двумя классами [Macskassy et al., 2003]. Вместо одного порогового значения используется два. Точка относится к классу 1, если соответствующее ей значение вероятности выше большего из порогов. Если это значение ниже меньшего порога, то точка относится к классу 0. Промежуточные точки не классифицируются.

Для построения карты низкоскоростного поверхностного слоя была рассчитана вероятность наличия такого слоя на исследуемой территории. Для определения оптимального числа ближайших соседей была выбрана техника Leave-one-Out (аналогично п. 2.1.). В качестве функционала качества был использован критерий ROC-AUC.

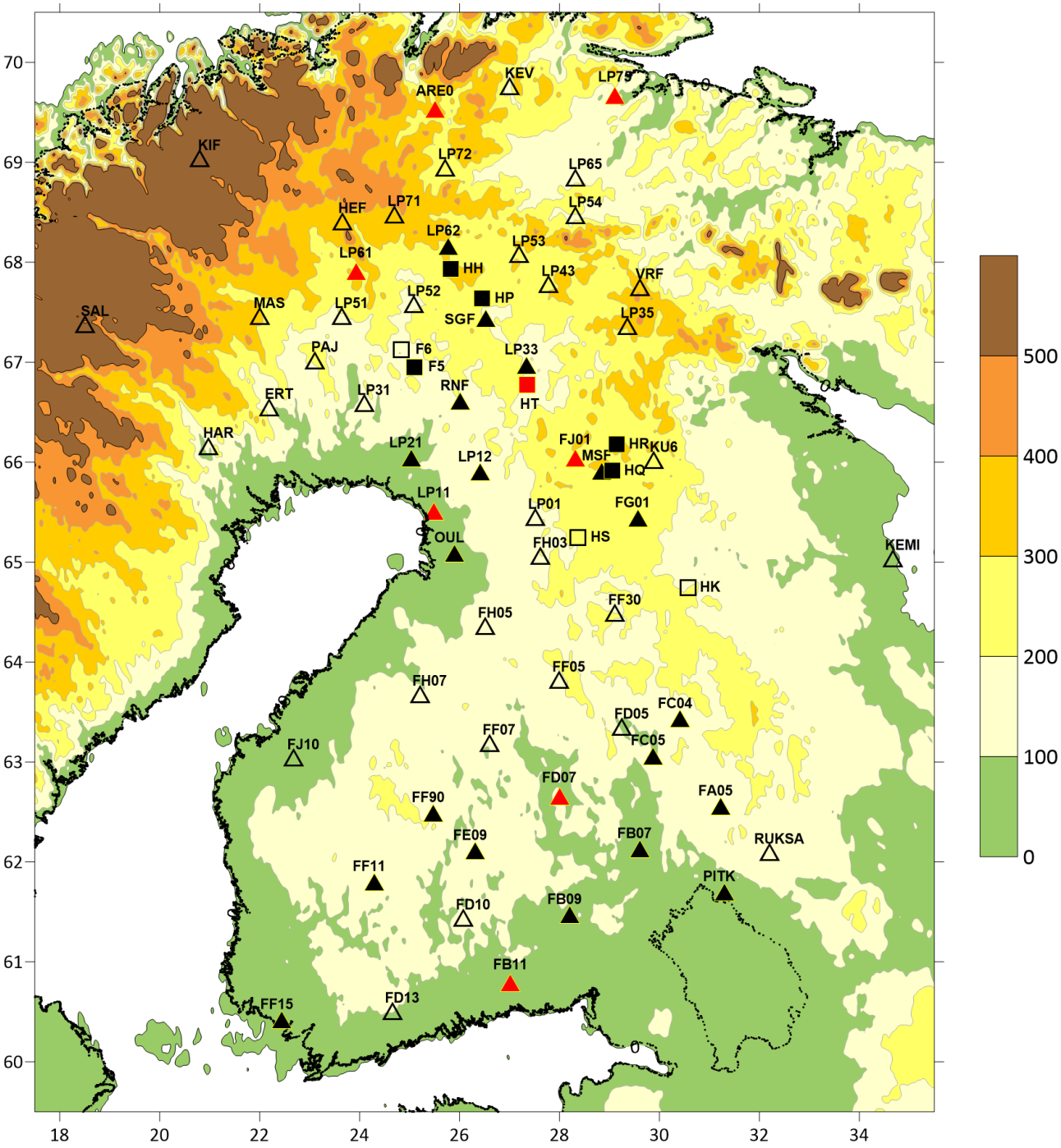

Рис. 2.9. Карта исследуемого региона. Цветовая шкала отражает высоту над уровнем моря в метрах. Треугольниками обозначены сейсмические станции, где данные получены методом приемных функций, и квадратами, где данные получены методом ГСЗ. Цвет станции соответствует классу: черный – слой пониженной скорости отсутствует (класс 0), белый – слой присутствует (класс 1), красный – нельзя достоверно определить.

Из рис. 2.10. следует, что максимальное значение качества 0.83 также, как и в предыдущем примере, достигается при числе ближайших соседей k=4.

Рис. 2.10. Зависимость критерия ROC-AUC, характеризующего качество бинарной классификации от числа ближайших соседей k. Оптимальное значение k соответствует максимальному значению критерия 0.83, которое достигается при k=4.

На рис. 2.11. синим цветом выделены области, в которых вероятность присутствия слоя превышает 55%. Оливковым цветом закрашены области, в которых эта вероятность меньше 45%. Область, в которой вероятность наличия слоя лежит между этими пределами залита промежуточным бежевым цветом. Результаты проведенных расчетов подтверждают существование низкоскоростного слоя в LP61, ARE0, LP79; отсутствие слоя – в LP11, HT, FJ01. Еще две станции FD07 и FB11 по-прежнему определить нельзя, они попадают в буферную зону, разделяющую два класса.

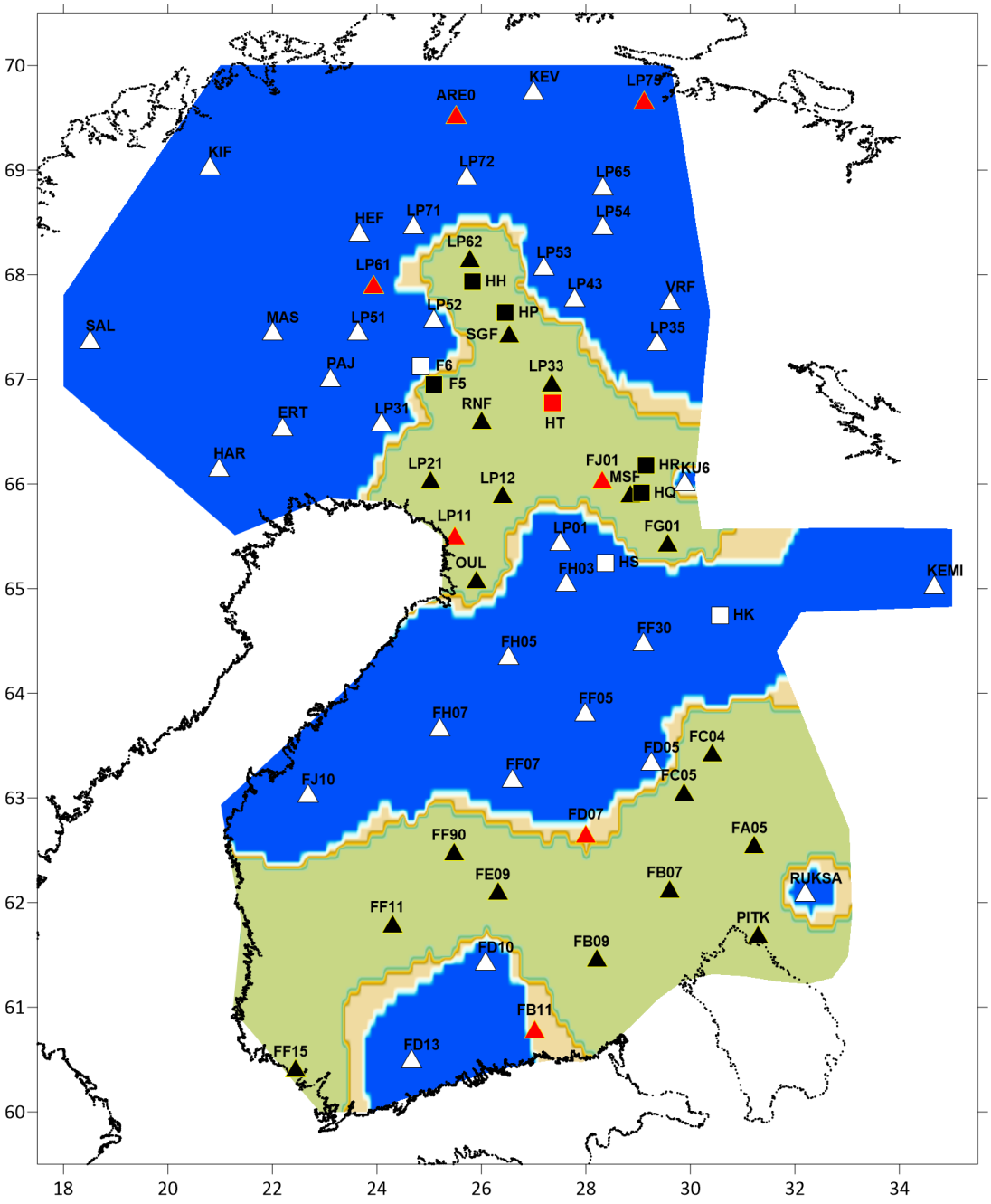

Рис. 2.11. Карта поверхностного слоя низкой скорости поперечных сейсмических волн, полученная в результате расчета по исходным данным. Синим цветом отображены области, в которых вероятность наличия слоя превышает 0.55, оливковым цветом окрашены области, в которых слой отсутствует (вероятность меньше 0.45). Промежуточным значениям вероятности соответствуют области бежевого цвета.

Наличие тонкого поверхностного слоя низких скоростей поперечных сейсмических волн в этом регионе впервые упоминается в статье [Pedersen,

Campillo, 1991]. Анализ поверхностных волн на сейсмическом профиле LITHOSCOPE, целиком расположенного в архейской части (профиль 1 на (рис. 2.4.)), показал очень низкую добротность пород в верхнем слое, вплоть до глубин порядка километра. С ростом глубины добротность очень быстро и сильно возрастает (практически на порядок величины). Природа возникновения слоя в рамках данной задачи не исследовалась. В работе [Grad, Luosto, 1992] рассмотрены данные профиля SVEKA, расположенные по обе части Ладожско-Ботнического пояса. Было показано, что поверхностный слой, характеризующийся большим поглощением сейсмической энергии, присутствует по обе стороны границы. Однако породы разных возрастов отличаются плотностью. Поперечная скорость сейсмических волн в слое имеет низкие значения, величина параметра *Vp/Vs* большое, порядка ~2.0. Учитывая низкую электрическую проводимость пород, авторы [Grad, Luosto, 1992] делают вывод о том, что снижение скорости поперечных волн обусловлено наличием в нем большого количества трещин. Предполагается, что трещины заполнены водой, но изолированы друг от друга.

Выводы работ [Pedersen, Campillo, 1991] и [Grad, Luosto, 1992] основаны на анализе данных поверхностных волн, что приводит к пространственному усреднению значений характеристик среды. Использование метода приемных функций позволяет получить более точную, локальную оценку. В работе [Aleshin et. al., 2006] показано, что минимальное значения *Vs* в поверхностном слое составляет 2.4 км/с, что существенно меньше величины 2.8 км/с, полученной в [Grad, Luosto, 1992]. Этот факт может быть объяснен тем, что трещины могут быть частично связаны [Aleshin et. al., 2006]. Это утверждение, в целом, согласуется с данными об электропроводности среды. Наличие тонкого проводящего поверхностного слоя вдоль профиля SVEKA отмечено в работе [Korja, Koivukoski, 1994]. Одно из возможных объяснений наличия такого слоя на северо-западе Финляндии может быть связано с постледниковым изостатическим поднятием региона [Лукк и др., 2019]. Присутствие такого слоя

в южной части региона, в Приладожье, подтверждается результатами электропрофилирования на постоянном токе [Жамалетдинов и др., 2018]. Трехмерные модели проводимости среды, построенные по результатам электромагнитного зондирования, приведены в статьях [Korja et al., 2002; Варенцов и др., 2006; Минц и др., 2018]). Однако использованный в цитированных работах подход не обеспечивает надлежащего разрешения для выделения тонкого поверхностного проводящего слоя.

Полученные в настоящей работе результаты являются уточнением проведенных ранее исследований. Выполнена совместная интерпретация результатов прежних исследований, дополненных данными для нескольких новых станций, с использованием современного математического аппарата. Приведенный анализ сейсмических данных показал эффективность методов машинного обучения для их анализа и обобщения. Достоинства такого подхода связаны с универсальностью применяемых методов. Особенно ярко преимущества алгоритмов теории машинного обучения проявляются в условиях недостатка данных, типичных для многих геофизических исследований. В использованных данных только 5 из 61 станций расположены на территории России. Дальнейшее продвижение в изучении региона затрудняет отсутствие данных на значительной части российской территории.

2.3. Построение 3D-модели среды по данным радиоволнового просвечивания

Ввиду того, то в Западной Якутии практически исчерпаны месторождения алмазов, непосредственно доступных с поверхности, работы по поиску кимберлитовых тел осуществляются на территориях, где традиционные геолого-геофизические исследования оказались неэффективными [Шмаков, 2017]. Для площадей, перекрытых осадочными породами, а также в местах развития траппов, единственным прямым методом поиска кимберлитов является бурение по сети. Чтобы снизить стоимость работ, расстояние между скважинами желательно увеличить, однако это увеличивает риск пропуска мелких

кимберлитовых трубок. Чтобы это избежать, используются методы межскважинного зондирования, в частности, радиоволновые методы [Кеворкянц и др., 2005]. Методика радиоволнового просвечивания (РВП) была разработана в середине прошлого века [Петровский, 1971] и активно применяется в настоящее время [Истратов и др., 2006] причем, не только при поиске кимберлитовых трубок [Толстов и др., 2018], рудных [Кузнецов, 2008] и нефтяных [Истратов и др., 2000] месторождений и пр., так и для мониторинга состояния природных объектов [Черепанов, 2017] и технологических процессов [Истратов и др., 2009]. Схема процесса радиоволнового просвечивания приведена на рис. 2.12.

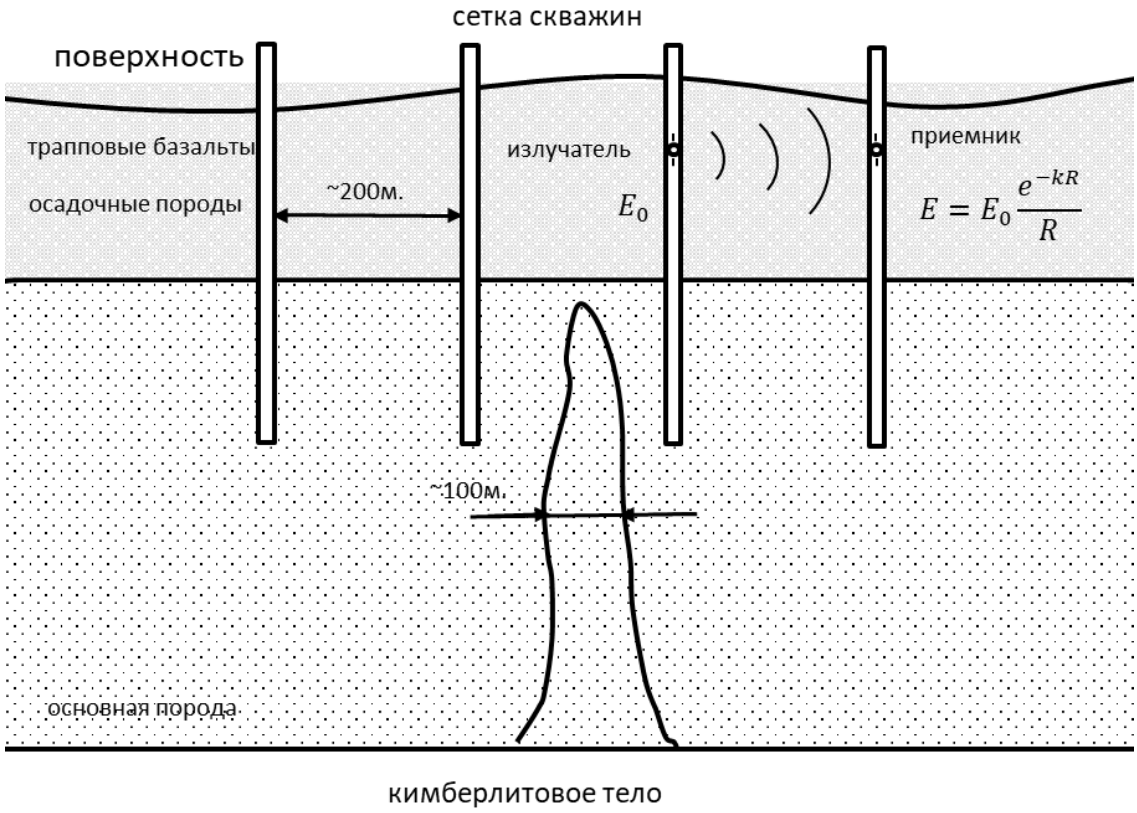

Рис. 2.12. Процесс радиоволнового просвечивания

Идея метода заключается в оценке затухания электромагнитной волны при ее прохождении через межскважинное пространство. Источник и приемник электромагнитного поля помещается в соседние скважины для измерения ослабления электрического поля. Породы, обладающие более низким

сопротивлением, характеризуются более высоким поглощением радиоволн, поэтому пространственное распределение коэффициента поглощения при фиксированной частоте пропорционально распределению электропроводности среды [Петровский, 1971]. Таким образом, радиоволновое просвечивание является частным случаем электроразведки [Жданов, 1986]. Его отличия от классических методов электроразведки, применяемых на этапе первоначального поиска кимберлитов [Поспеева и др., 2004], проявляются в диапазоне используемых частот, мощности сигнала, взаимным расположением приемника и источника и другими параметрами используемых электромагнитных волн.

В идеальном случае межскважинное просвечивание выполняется веерным способом (рис. 2.13. А): положение излучателя изменяется с заданным шагом, на каждом из которых его положение фиксируется, и приемник, находящейся в другой скважине, фиксирует измерения по всему рабочему интервалу скважины. После этого приемник перемещается в следующую позицию. Перемещение источника осуществляется дискретно, с заданным шагом. Такая схема измерений позволяет получить детальную картину электрических свойств межскважинного пространства в плоскости, проходящей через обе скважины. Совместная интерпретация совокупности полученных таким образом разрезов позволяет получить трехмерный образ электрических свойств среды [Кузнецов, 2012; Mishra et al., 2019].

При веерной съемке число измерений равно n^2, где n – количество стоянок в скважине. Поэтому на практике, при радиоволновой съемке зачастую ограничиваются синхронным (рис. 2.13. В) погружением источника и приемника в соседние скважины. В этом случае количество измерений пропорционально n, что существенно снижает объем измерений, но, вместе с тем, информативность измерений также снижается. Фактически, все, что становится возможным получить при такой схеме измерений, это среднее значение кажущегося коэффициента затухания, соответствующее середине отрезка, соединяющего источник и приемник.

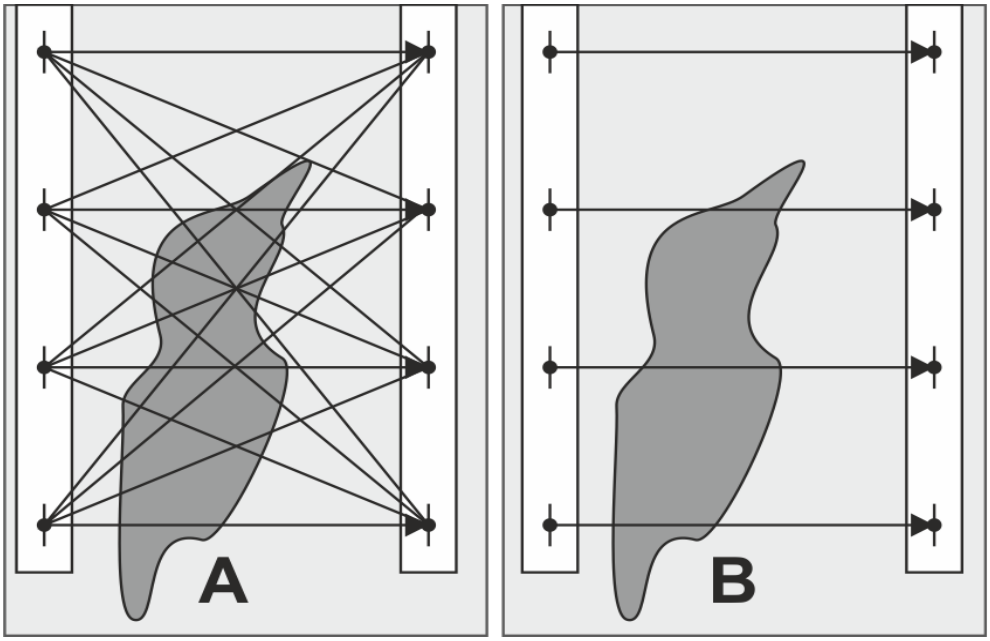

Рис. 2.13. Схемы измерений, используемых в радиоволновых исследованиях. А – веерная, положение источника фиксируется, приемник перемещается по всему рабочему диапазону с заданным шагом. Затем источник смещается в следующую позицию и измерения повторяются и т. д. В – синхронные измерения, источник и приемник перемещаются по скважине одновременно.

Для построения трехмерной модели среды томографический метод становится неприменим, поэтому необходимо использовать принципиально иную интерполяционную процедуру. В работе [Aleshin, Zhandalinov, 2009] горизонтальные сечения модели строятся на основе регрессионной модели – методом кригинга [Isaaks, Srivastava, 1989]. Параметры интерполяции выбираются исходя из согласования ошибок интерполяции с погрешностью задания входных данных. Однако при построения действительно трехмерной модели непосредственное применение регрессионного подхода невозможно из-за экстремальной анизотропии распределения исходных данных. Кроме того, даже в двумерном случае решение получается излишне сглаженным, в то время как основная задача состоит в получении максимально контрастного образа.

Одна из возможных альтернатив состоит в использовании методов машинного обучения для проведения интерполяционной процедуры [Aleshin,

Malygin, 2019; Алешин, Малыгин, 2019]. Ниже описана процедура построения трехмерного распределения коэффициента затухания на основе метрического классификатора. В качестве реализации был использован алгоритм k ближайших соседей (kNN). В данной работе для иллюстрации были использованы данные АО «АЛРОСА» – результаты радиоволнового просвечивания, выполненного на одном из якутских участков.

Для анализа данных межскважинных измерений необходимо выбрать модель пространственного изменения электрического поля волны, излучаемой источником. Так как имеющиеся данные получены в результате синхронных измерений, то, возможно оценить лишь среднее значение коэффициента поглощения среды вдоль прямой, соединяющей источник и приемник. Поэтому используется модель распространения волны, определяемой формулой поля излучения электрического диполя в однородной изотропной среде

$$E = E_0\ exp\ (-q/R)\ /R\ \ sin\ (\theta).$$

Здесь E – полярная компонента электрического поля, E_0 – амплитуда излучаемой волны, R – расстояние от источника. При синхронных измерениях источник и приемник находятся приблизительно на одной глубине, поэтому полярный угол θ можно положить равным $\pi/2$. Тогда искомый коэффициент поглощения равен

$$q = -\ ln\ (R\ E/E_0)/R.$$

Обозначим через $Q = \{q_n\}$ набор исходных данных – значения кажущегося коэффициента затухания, измеренные в N точках. Координаты точек $\vec{r}_n = \{x_n, y_n, z_n\}$ соответствуют середине отрезка, соединяющего источник и приемник. Здесь x, y определяют положение точки в горизонтальной плоскости, а z – глубина, отсчитываемая от уровня моря. В алгоритме kNN значение величины q в произвольной точке $\vec{r} = \{x, y, z\}$ определяется основной формулой

$$q(\vec{r}) = \sum_{i=1}^{k} w_i(\vec{r}, \vec{r}_i) q_i\ , \sum_{i=1}^{k} w_i = 1,$$

суммирование проводится по k точкам, ближайшим к $\vec{r}$. Число k является свободным параметром алгоритма (гиперпараметром), требующим дополнительного определения. Иногда вместо числа соседей в качестве

гиперпараметра используют радиус сферы с центром в точке $\vec{r}$. Тогда соседними считаются все точки, находящиеся внутри сферы. В качестве расстояния ρ в нашем случае естественно выбрать евклидово расстояние между точками:

$$\rho(\vec{r}_1, \vec{r}_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}.$$

Величина $w_i(\vec{r}, \vec{r}_i)$ – весовая функция, зависящая от расстояния текущей точки $\vec{r}$ до соответствующей точки с заданным значением, в качестве которой обычно используют величину, обратно пропорциональную расстоянию:

$$w_i(\vec{r}, \vec{r}_i) \sim 1/\rho(\vec{r}, \vec{r}_i).$$

Если пространственное расположение точек не учитывать, то веса одинаковы для всех точек $w_i = 1/k$.

Как уже отмечалось ранее, распределение данных, полученных методом РВП, сильно анизотропно: шаг по глубине имеет величину 5 м при длине скважины порядка 500 м, в то время, как расстояние между ближайшими скважинами составляет, порядка 200 м. Это приводит к тому, что при построении трехмерной модели нельзя использовать классические методы геостатистики [Isaaks, Srivastava, 1989]. Чтобы воспользоваться методом ближайших соседей также требуется его модификация. Чтобы нивелировать разницу в горизонтальном и вертикальном масштабах, переопределим метрику, введя безразмерный масштабный множитель λ (формула 2.2.):

$$\rho'(\vec{r}_1, \vec{r}_2) = \sqrt{(x_1 - x_2)^2 / \lambda^2 + (y_1 - y_2)^2 / \lambda^2 + (z_1 - z_2)^2}.$$

Можно ожидать, что подходящий выбор значения параметра λ позволит компенсировать анизотропию данных (см. рис. 2.14.), однако критерий, позволяющий сделать этот выбор отсутствует. Поэтому масштабный множитель λ, наряду с числом соседей k, является еще одним гиперпараметром задачи.

Для определения гиперпараметров использован метод кросс-валидации (п. 1.2.). Исходные данные разбиваются на M групп ($M \sim 5$). Каждая из этих групп поочередно устраняется из процедуры обучения, и используется для проверки. Оценки качества решения производится по функционалу качества R^2

(коэффициент детерминации) – доля дисперсии зависимой переменной, объясняемая моделью, который определяется формулами

$$R^2(k,\lambda) = (1/M)\sum_{m=1}^{M} R^{2^{(m)}}(k,\lambda),$$

$$R^{2^{(m)}}(k,\lambda) = 1 - \sum_{i=1}^{N/M}\left(q_i^{(m)} - q(\vec{r}_i;k,\lambda)\right)^2 / \sum_{i=1}^{N/M}\left(q_i^{(m)} - \underline{q}^{(m)}\right)^2,$$

$$\underline{q}^{(m)} = \sum_{i=1}^{N/M} q_i^{(m)}.$$

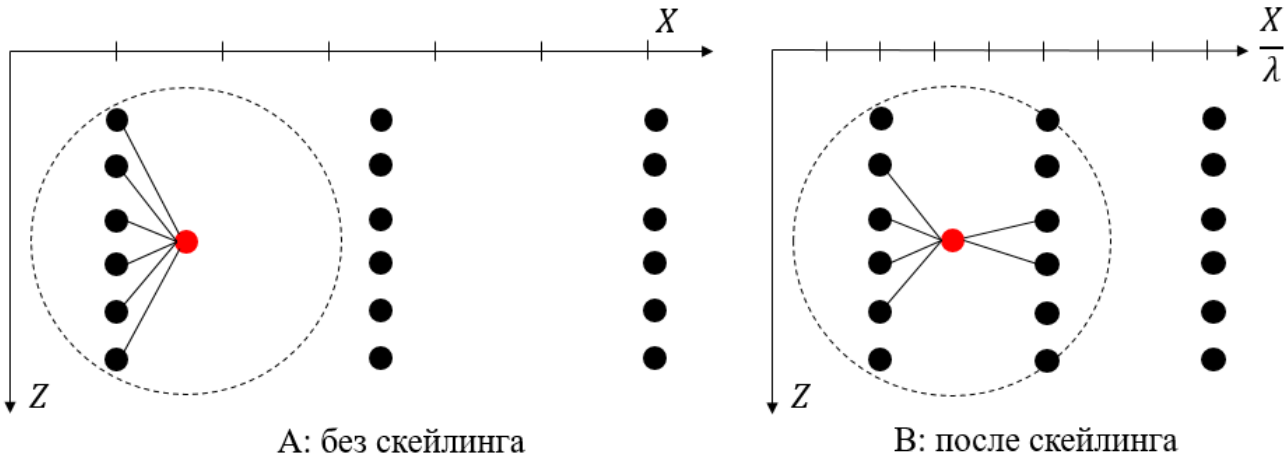

Рис. 2.14. Влияние масштабного коэффициента λ на распределение ближайших точек. Особенность пространственного распределения данных приводит к тому, что практически для всех точек пространства ближайшими оказываются данные, относящиеся к одной группе измерений (A). Масштабирование горизонтальных осей позволяет исправить эту ситуацию (B).

Интерполянт $q(\vec{r}_i;k,\lambda)$ вычисляется по основной формуле kNN с метрикой ρ' без учета исключенных данных. Коэффициент детерминации был рассчитан на сетке $1 \le k \le 25$, $1 \le \lambda \le 25$ с единичным шагом по каждому из параметров. Результат расчетов приведен на рис. 2.15. Полученное распределение имеет вид, типичный для задач многопараметрической оптимизации. Для выбора значений гиперпараметров использован уровень значений коэффициента $R^2 = 0.7$, что соответствует приблизительно 80% корреляции модели и исходных данных. Выбранные значения гиперпараметров $k = 11$ и $\lambda = 10$ соответствуют

пересечению медианы треугольника, образованного осями координат и прямой, аппроксимирующей 70% уровень значений коэффициента детерминации.

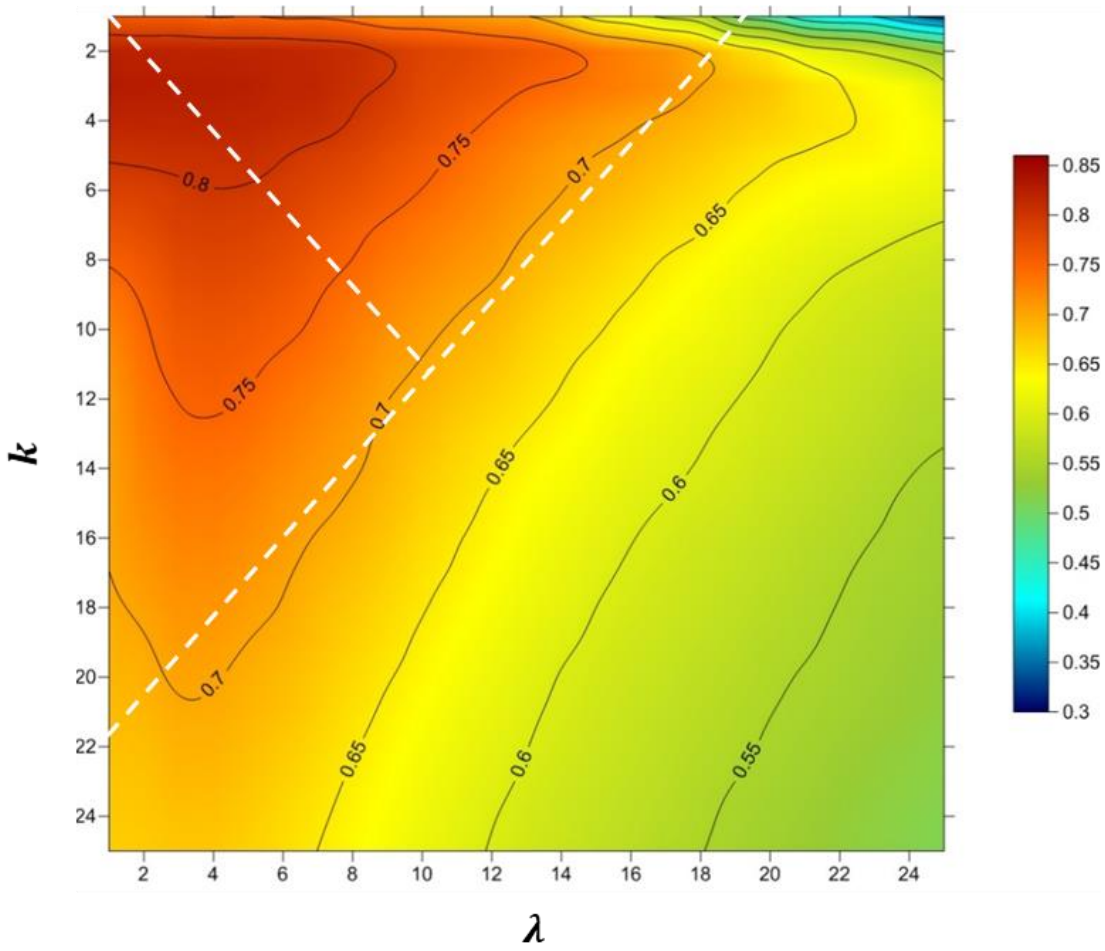

Рис. 2.15. Значения коэффициента детерминации $R^2(k, \lambda)$, рассчитанные на сетке параметров. В качестве приемлемых значений гиперпараметров k (число ближайших соседей), и λ (масштабный множитель) выбрана область, в которой R^2 превышает значение 0.7. Для построения модели выбраны значения, определяемые точкой пересечения медианы и гипотенузы треугольника, образованного осями координат и прямой приближающей изолинию $R^2 = 0.7$.

После того, как значения гиперпараметров определены, задача построения образа сводится к вычислению интересующей величины – коэффициента затухания – по формулам kNN с модифицированной метрикой ρ' в узлах трехмерной решетки. Построение модели, расчет горизонтальных и

вертикальных сечений реализовано на языке программирования Python с использованием коллекции пакетов scikit-learn [Pedregosa et al., 2011].

На рисунке 2.16. приведены результаты моделирования: набор вертикальных и горизонтальных сечений пространства. Глубина отложена от уровня моря, положение горизонтальных осей согласовано с геометрией участка. Расположение вертикального разреза соответствует линии $Y = 0$ на схеме расположения скважин. Горизонтальные сечения соответствуют глубинам $Z = -560$ и $Z = -250$ метров (черная и белая пунктирные линии на вертикальном разрезе). Из рисунка видно, что построенная модель позволяет локализовать объекты, чьи горизонтальные размеры существенно меньше расстояния между скважинами. В качестве примера можно привести области повышенных значений коэффициента затухания, расположенные на глубине -560 м., и горизонтальными координатами $X = 2950$, $X = 4750$ и $X = 5150$ метров.

Для наглядности, на рисунке 2.16. приведены также короткие вертикальные сечения, соответствующие этим линиям, на которых соответствующие области диаметром менее 100 м. также отчетливо видны.

Таким образом, можно сделать вывод, что использованный метод kNN позволяет построить трехмерную модель проводимости среды между скважин даже при использовании синхронной схемы измерений. Влияние анизотропии распределения данных можно исключить, если модифицировать пространственную метрику, определяющую расстояние между данными. Это достигается введением коэффициента скейлинга, который изменяет масштаб в горизонтальном направлении. Использованный подход позволяет получить достаточно контрастное изображение неоднородных областей, что позволяет выделить неоднородности, геометрические размеры которых меньше расстояния между скважинами.

Следует заметить также, что процесс построения модели не зависит от физической модели, использованной для интерпретации измерений. Уточнение физической модели процесса распространения радиоволн между скважинами

позволит улучшить качество построения образа. Кроме того, модель может быть улучшена, если привлечь дополнительные данные (геологические, сейсмические, магнитные) для их совместной интерпретации.

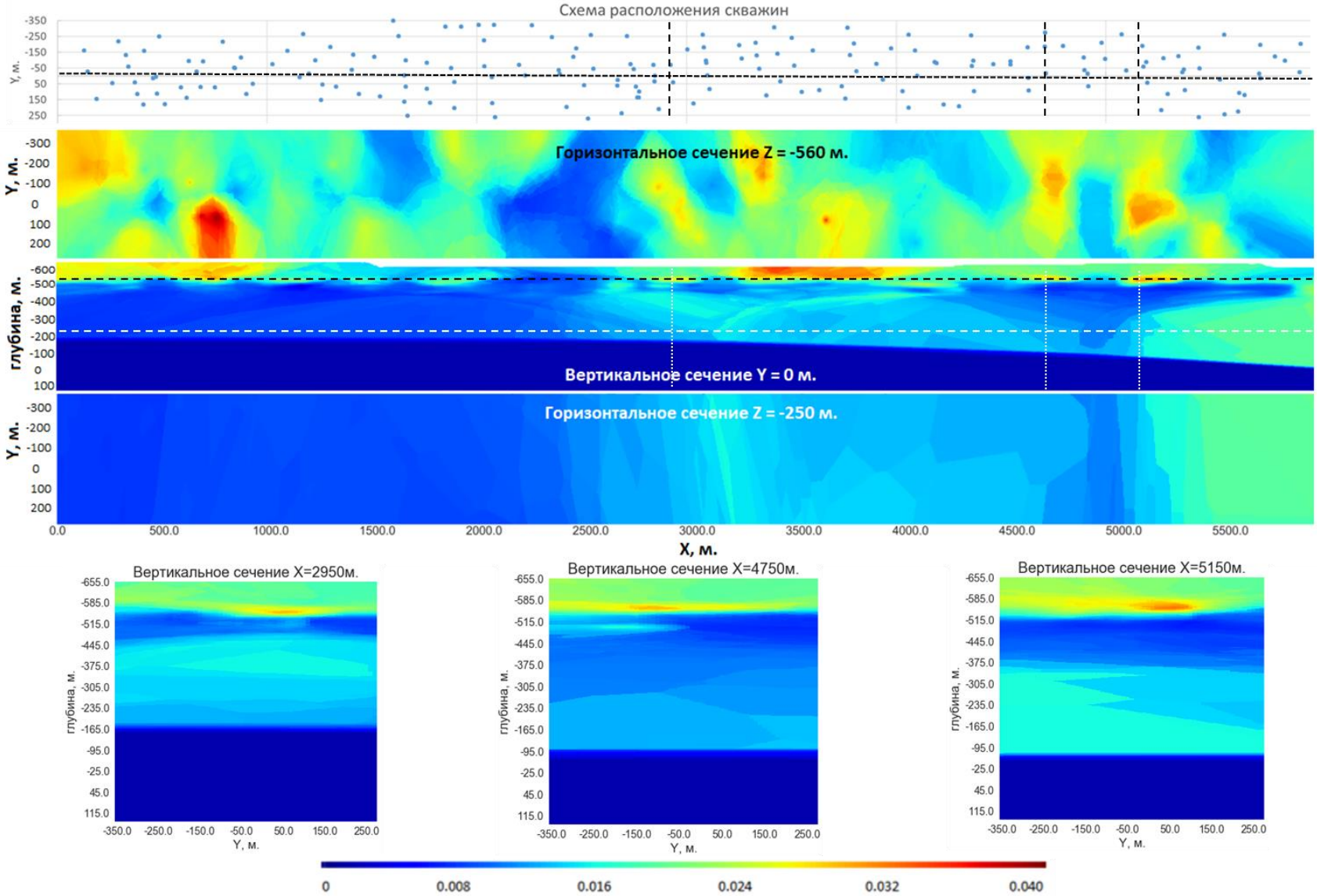

Рис. 2.16. Результаты 3D-моделирования

ГЛАВА 3. Система прогноза заторообразования на Северной Двине

К наиболее опасным природным явлениям относятся высокие уровни воды (при половодьях, дождевых паводках, ледовых заторах). Это уровни, при которых происходит подтопление населенных пунктов, земель сельскохозяйственного использования, объектов транспортной или промышленной инфраструктуры. В ряде районов образование ледовых заторов является основной причиной возникновения наводнений, а на реках, например, бассейна Северной Двины, еще и совпадает по времени с максимумом подъема воды во время весеннего половодья.

Места возникновения заторов часто обладают высокой повторяемостью, т.к. существенным образом зависят от геометрии русла реки, наличии островов и препятствий, поэтому задача прогнозирования ледового затора на реке состоит в определении вероятной мощности возникновения опасного явления с необходимой заблаговременностью (для краткосрочного прогноза 3-7 суток). Явление заторообразования в настоящее время до конца не изучено, однако выдвинуты гипотезы о влиянии ряда гидрометео-показателей на итоговую мощность события. Эта ситуация также осложняется недостатком исходных данных для проведения классических статистических исследований. В-основном, это касается данных, собираемых на гидрологических постах непосредственно на р. Северная Двина и ее притоках (для метеостанций эта проблема стоит в меньшей степени). В наиболее критичных для составления прогноза местах набор полных, одинаково структурированных временны*х* исходных данных доступен за период менее 30 лет (сезонов осень-весна).

В данной задаче применены методы теории распознавания образов, потребовавшие проведения специальной адаптацию под конкретное приложение: доработана возможность использования непрерывных признаков (в оригинальных алгоритмах – только бинарные), предложен способ автоматизированного нахождения множества решающих правил (обучения системы), применена комбинация алгоритмов вычисления оценок и голосования

для достижения лучшего качества классификации. На техническом уровне решение представлено экспертной системой с функциями обучения, прогнозирования и оценки вклада каждого признака в итоговую мощность явления. Система разрабатывалась на данных 1991-2010 гг. (20 сезонов) в 2012 г., в дальнейшем в 2018 г. дополнительно проведена ее валидация на периоде новых данных 2011-2016 гг. Оцененная точность прогнозирования составила 85%.

3.1. Разработка прогнозной системы

Согласно А.И. Чеботареву под *ледовым затором* понимается нагромождение льдин в русле реки во время ледохода, вызывающее стеснение живого сечения и связанный с этим подъем воды [Чеботарев, 1978]. Преимущественно заторы наблюдаются во время весеннего ледохода, при осеннем ледоходе массы льда обычно не столь значительны, чтобы вызвать образование мощных заторов. Явление заторообразования на реках России изучалось с начала XX века. По данным экспедиций и полевых исследований были сделаны первые выводы о некоторых морфометрических факторах заторообразования [Близняк, 1916]. Было сформулировано определение ледового затора, соответствующее современным представлениям о явлении [Иогансон, 1927]. В середине и второй половине XX века исследования носят системный описательный подход. Результатом работы стали труды М.К. Федорова, Л.Г. Шуляковского [Шуляковский, 1951; Шуляковский, Еремина, 1952; Федоров, 1956; Шуляковский, 1960]. Р.В. Донченко получены закономерности появления заторов на реках СССР [Донченко, 1987].

Установлено, что образование ледовых заторов зависит как минимум от двух групп факторов: тепловых и механических [Бузин, 2004]. Группа тепловых факторов отражает толщину и прочность ледяных образований, а группа механических факторов отражает то, как происходит взлом и разрушение целостности ледяного покрова, его движение вниз по реке. Затор образуется, если имеется недостаток кинетической энергии речного потока для взлома

ледяного покрова. Характеристикой сопротивляемости ледяного покрова вскрытию может служить произведение относительной прочности льда (по отношению к прочности в начале периода таяния) на толщину ледяного покрова [Бузин, Зиновьев, 2009]. Были созданы классификации и типизации заторов льда по условиям их прогнозирования [Деев, Попов, 1978; Чижов 1975]. Теоретическое представление о заторах через построение натурных физических моделей представлено в работах В.П. Берденникова, Д.Ф. Панфилова [Панфилов, 1968; Берденников, 1974; Берденников, Шматков 1976]. Традиционная методика прогнозирования наводнений, обусловленных заторами льда, представлена в работах В.А. Бузина и состоит в следующем [Бузин, 1997; Бузин, 2000; Бузин 2004]. Прогноз ледового заторообразования целесообразен для мест с высокой повторяемостью явления, где ледовые скопления наблюдаются ежегодно. Для предсказания заторных максимумов уровня используются эмпирические зависимости уровней от факторов (признаков), определяющих процесс заторообразования. Такие зависимости устанавливаются по данным многолетних гидрометеорологических наблюдений. В таком случае, сформированное признаковое пространство, для каждого исследуемого участка должно быть свое. Для прогноза наводнений, обусловленных образованием заторов льда, обычно ограничиваются интегральными признаками, например, такими как максимальный предледоставный уровень, расходом воды у перемещающейся по течению кромки ледяного покрова и т.д. Существующие методики оценки наводнений преимущественно основаны на гидродинамических моделях речного стока, в которых требуется полная формализация пространственно-временного описания процесса. Учет влияния заторообразования в существующих методиках оценки стока затруднен из-за отсутствия явного вида математических зависимостей мощности процесса от его исходных параметров [Агафонова, Фролова, 2007].

Рассматривается задача прогноза ледовой обстановки для района с малой территорией и с коротким периодом наблюдения, на примере участка реки

Северная Двина от г.Котлас до г.Великий Устюг. Места образования заторов в бассейне Северной Двины чаще всего стационарны. Повторяемость заторов на отдельных участках достигает 86% [Агафонова, Фролова, 2007]. На Северной Двине заторы обычно формируются на участках вблизи г. Великий Устюг, г. Котлас, д. Орленцы, с. Холмогоры. В 57% случаев максимальный уровень воды в районе г. Великий Устюг обусловлен заторами льда. Это самое высокое для бассейна Северной Двины значение, для остальных участков заторы определяют максимальный уровень воды весной в 12 – 50% случаев [Агафонова, Фролова, 2007].

Возможным подходом к разрешению проблемы заторообразования является ее исследование в рамках построения прогнозных систем, основанных на алгоритмах теории распознавания образов и теории машинного обучения (п. 1.4.) [Малыгин, 2014 «*Методика…*»; Малыгин, 2014 «*О задаче…*»;].

Исходными данными являются наблюдения гидрологических постов и метеостанций различной временной глубины. Имеется статистика по явлению, т.е. фактический результат проявления (было ли явление, какой мощности). Экспертная система представляет прогноз по исследуемому явлению в будущий момент времени.

Для исследования ледовой обстановки экспертным образом в качестве признаков выбран ряд гидрологических и метеорологических показателей [Агафонова, Василенко, Фролова, 2016]. Общий список этих признаков представлен в табл. 3.1. Признаки делятся на два типа: гидрологические – вычисляемые по измеренным на гидропостах данным, и метеорологические – измеренные по данным метеостанций. К гидрологическим признакам относятся разнообразные водные характеристики, описывающие состояния до начала процесса образования заторов. Например, уровни воды, продолжительность ледохода, толщина льда. Важной частью являются гидродинамические признаки, обладающие наименьшей заблаговременностью, например, скорость роста уровня воды перед вскрытием. К метеопризнакам относятся интегральные

(по времени) характеристики, описывающие состояние окружающей среды в интересующий период времени. Например, это осадки, температура воздуха, характеристики перехода температуры через 0 градусов и т.д. В цитированных работах учеными-гидрологами выдвинуты гипотезы о влиянии этих показателей на итоговый процесс заторообразования. Дальнейшие исследования в рамках построения прогнозной системы подтвердили эти гипотезы (п. 3.2.).

Приведенные в таблице 3.1. гидрологические признаки рассчитаны на основании доступных на момент разработки системы данных речных постов в районе наблюдения: Каликино, Великий Устюг, Медведки, Котлас, Абрамково, Подосиновец (рис. 3.1.).

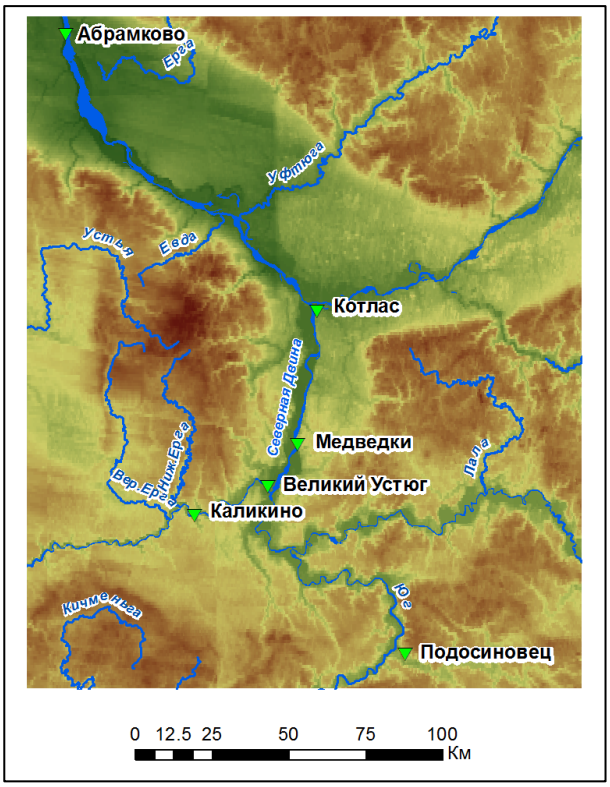

Рис. 3.1. Схема расположения используемых гидрологических постов на р. Северная Двина и ее притоках

Табл. 3.1 Экспертные признаки заторообразования

№	Название признака	Характеристика признака	Единицы измерения
1	Предледоставный уровень воды	Гидрологический признак	сантиметры
2	Продолжительность осеннего ледохода	Гидрологический признак	сутки
3	Наличие зажоров	Гидрологический признак	есть (1) – нет (0)
4	Особенности температурного режима в период замерзания	Метеорологический признак Факт перехода температуры воздуха через ноль	количество суток с 1 сентября
5	Сумма отрицательных температур воздуха за холодный период	Метеорологический признак	градусы Цельсия
6	Сумма положительных температур воздуха за холодный период	Метеорологический признак	градусы Цельсия
7	Количество дней с положительными температурами воздуха за холодный период	Метеорологический признак	сутки
8	Сумма твердых осадков	Метеорологический признак	миллиметры
9	Особенности температурного режима в период вскрытия	Метеорологический признак Факт перехода температуры воздуха через 0.	количество суток с 1 марта
10	Толщина льда перед вскрытием	Гидрологический признак	сантиметры
11	Интенсивность роста уровней и расходов воды в период подвижек	Гидрологический признак	сантиметры в сутки

Исходные данные для проведения исследования содержатся в реляционной базе данных. Подробное описание и ее схема представлены в

Приложении 1 (схема базы – рис. 3.2. в Приложении 1). Структура реляционной БД обеспечивает функционирование ядра прогнозной системы, которое формирует оценку прогнозных значений ряда параметров опасного явления. Набор таблиц и связей БД сформирован с учетом расширения функционала системы и степени детализации исходных данных прогнозирования.

Входными данными для логического модуля прогнозной системы являются представления над основной таблицей БД (DATA). В каждом представлении содержатся значения всех признаков для каждого поста за один год. Пример приведен в Таблице 3.2.

Табл. 3.2 Пример исходных данных за 1991г.

1991 год Пост	Номер признака										
	1 м	2 сутки	3 да/нет	4 сутки	5 C^0	6 C^0	7 сутки	8 мм	9 сутки	10 см	11 см/сутки
Каликино	337	13	0	59	-1449	3,9	6	192.4	34	41	76
Вел. Устюг	121	10	0							57	76
Медведки	135	5	0							67	42
Котлас	214	22	0							50	80
Абрамково	112	23	0							64	59
Подосиновец	94	1	0							48	61

Для группы метеорологических признаков №№ 4–9 использована аппроксимация по пространственной составляющей фактическими значениями по данным метеостанции г. Великий Устюг. Алгоритм визуализации признаков прогнозной системы средствами ГИС приведен в Приложении 3.

Согласно экспертной классификации принимаются два возможных сценария ледохода:

1) наличие заторов с различными мощностью и продолжительностью на участке г. Великий Устюг – г. Котлас;

2) отсутствие заторов, либо их несущественное проявление на участке г. Великий Устюг – г. Котлас (в этот класс попадают и ситуации, когда затор произошел выше или ниже по течению, чем исследуемый участок).

Указанные сценарии ледохода определяют классы K_1, K_2 периода наблюдения (Табл. 3.3.).

Табл. 3.3. Экспертная классификация периода наблюдения

Год	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10
№ класса	1	1	2	2	2	1	2	1	2	2	1	1	2	1	1	2	2	1	1	1

Ситуация недостатка данных ограничивает возможность разделения на большее число классов, соответствующих более подробной классификации исследуемого явления.

К логическим правилам прогнозной системы относятся процедура сравнения однородных признаков и алгоритм подбора числовых параметров сравнения (п. 1.4.).

Для работы алгоритма распознавания необходимо уметь сравнивать числовые значения однородных признаков за разные годы. Если различие в числовых значениях находится в определенном допуске, то полагается, что два сравниваемых года по этому признаку одинаковые, в противном случае – различные. Для определения различия признаков в алгоритме вычисляется сумма изменений значений по всем постам:

$$B_k(\Gamma_n,\ \Gamma_m) = \frac{1}{6}\sum_{j=1}^{6}\left| p_{knj} - p_{kmj} \right|,$$

где $B_k(\Gamma_n,\ \Gamma_m)$ – функция близости, т.е. величина различия *n*-го и *m*-го годов по *k*-му признаку, p_{knj} и p_{kmj} – числовые значения *k*-го признака на *j*-м посте в *n*-й и *m*-й годы.

Сравнение пар лет (эталонов) происходит по принципу, описанному в п. 1.4.: если значение функции близости превышает пороговое эвристическое значение, то полагается, что различие по исследуемому признаку в этой паре лет есть, в противном случае оно отсутствует.

Целью работы блока обучения является построение группы векторов числовых параметров $\Delta_j = (\delta_{j1},...,\delta_{j11})_{\text{opt}},\ j = 1,...,s$, обеспечивающих прогнозирование характера ледовой обстановки с наилучшей достоверностью.

На любом векторе из этой группы происходит оптимальное отнесение элементов обучающей выборки к своим классам в рамках процедуры скользящего поиска (использована техника Leave-one-Out, п. 2.1). Из всего периода наблюдения последовательно удаляются данные по каждому году, остальные принимаются в качестве обучающей выборки. Удаленный год подается в соответствующий блок в качестве распознаваемого, в результате производится его отнесение к одному из классов. Этот результат может совпадать или отличаться от фактической классификации. Необходимо выбрать такие параметры $\Delta_j = (\delta_{j1},...,\delta_{j11})_{\text{opt}}$, чтобы количество правильных прогнозов соответствовало критерию оптимальности.

Для каждой пары лет из разных классов с использованием процедуры сравнения однородных признаков определяется обобщенный вектор различия этой пары лет по всем признакам. Вектор формируется следующим образом: если в результате работы процедуры сравнения установлено различие в паре лет по *p*-му признаку, то в качестве координаты с номером *p* обобщенного вектора принимается 1, в противном случае 0, следовательно, это булевский вектор:

$$U(\Gamma_q,\ \Gamma_r) = (u_1,...,u_{11}),$$

$$u_p = 1, \text{ если } B_p(\Gamma_q,\ \Gamma_r) \geq \delta_p,$$

$$u_p = 0, \text{ если } B_p(\Gamma_q,\ \Gamma_r) < \delta_p,$$

$$p = 1,...,11.$$

Этот вектор не является нулевым, так как используются элементы из разных классов. Равенство нулю этого вектора означало бы, что выбранные пороговые значения велики, т.е. эталоны из разных классов при их сравнении не различаются. Совокупность всех таких ненулевых векторов составляет таблицу сравнения классов.

Далее необходимо сформировать наборы признаков (координат булевских векторов), по которым различаются все пары лет из разных классов, так называемых тестов. Как было сказано в п. 1.4., тестом является такой набор признаков, что для любой пары лет из разных классов имеется различие между

этими годами хотя бы по одному признаку из этого набора. По этому набору формируется булевский вектор: если признак принадлежит набору, то в соответствующую координату ставится 1; если признак не принадлежит набору, то в соответствующую координату ставится 0.

Вектора этого множества хранят информацию о том, насколько отличается один год из одного класса от другого года из другого класса. На этом этапе хорошо видно не только корреляцию признаков, но и корреляцию групп признаков.

Для определения числовых параметров прогнозирования используется метод Монте-Карло. Для установленного пользователем числа испытаний метода Монте-Карло производится случайный выбор $(\delta_1,...,\delta_{11})$. Исходной точкой случайного поиска является системное время и дата, что исключает повторение выборок при последующих запусках обучения.

Из всего периода наблюдения последовательно удаляется каждый элемент, который подается на распознавание. Остальные 19 элементов образуют обучающую выборку.

На каждом шаге для оставшихся 19-ти элементов строится таблица сравнения классов и множество тестов.

Производится распознавание удаленного элемента и соотнесение его с фактической классификацией. После обработки всех 20-ти элементов получается оценка качества по функционалу accuracy: отношение правильно распознанных элементов к общему их числу (выраженное в процентах). Далее алгоритм переходит к следующему случайному испытанию.

Из результатов работы метода Монте-Карло выбираются наборы $\Delta_j = (\delta_{j1},...,\delta_{j11})_{\text{opt}}$, на которых выполняется критерий оптимальности. Таких наборов может быть несколько, все они поступают на вход блока распознавания. В качестве критерия оптимальности доступны несколько стратегий: использование только тех наборов векторов, на которых достигнуто

максимальное качество; либо возможно использовать все найденные наборы, оценка качества которых превышает наперед заданный пользователем порог.

Проведено сравнение работы описанных в п.1.4. локальных решающих правил: алгоритма вычисления оценок, алгоритма голосования и их комбинации. Выполнено 128000 испытаний Монте-Карло (из них количество успешных 125856). Для каждого успешного испытания для обучения системы применены три описанных выше локальных решающих правила. Распределение числа успешных испытаний по уровням распознавания (процентное отношение количество верно распознанных эталонов к общему их числу) для трех алгоритмов показано в Табл. 3.5. и на Рис. 3.2. Анализ результатов показывает, что обучение системы по алгоритму голосования позволяет достичь сравнительно невысокого уровня при малом количестве найденных оптимальных векторов. При использовании алгоритма вычисления оценок увеличивается как уровень, так и мощность множества Δ_i. Наилучшие результаты на исследуемых данных достигаются при использовании комбинации алгоритма голосования и алгоритма вычисления оценок.

В качестве примера, Табл. 3.4. содержит результаты кросс-валидации на основании одного набора параметров, найденного по комбинации алгоритмов голосования и вычисления оценок.

Табл. 3.4. Проверка прогнозирования (один набор)

Год	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
№ класса	1	1	2	2	2	1	2	1	2	2	1	1	2	1	1	2	2	1	1	1
Результат прогнозирования (один эксперт)	1	1	1	2	2	1	2	1	2	2	1	1	2	1	1	1	2	2	1	1

Целью работы блока распознавания является определение класса наличия заторов, к которому относится распознаваемый год, то есть прогноз. В качестве входных данных блок принимает элементы базы данных (обучающая выборка), логические правила сравнения, данные по распознаваемому году, набор векторов числовых параметров прогнозирования, полученных в результате работы блока обучения. Последовательно производится распознавание текущего

года по каждому вектору числовых параметров прогнозирования. Для этого по всему периоду наблюдения, предшествующему распознаваемому году, строится таблица сравнения классов и формируется множество тестов. Далее, работает один из алгоритмов распознавания: реализованы алгоритм вычисления оценок и алгоритм голосования.

Результатом работы является числовая характеристика, определяющая принадлежность распознаваемого года к каждому из классов экспертной классификации. Производится суммирование таких характеристик, полученных для каждого вектора числовых параметров прогнозирования. Экстремальное значение этой суммы определяет принадлежность распознаваемого года к одному из классов. Для алгоритма вычисления оценок выбирается минимальное значение, для алгоритма голосования – максимальное.

На основании приведенных в Табл. 3.5. данных видно, что устойчивое качество классификации при использовании комбинации алгоритма вычисления оценок и алгоритма голосования составило 85%: найдено 44 разделяющих наборов пороговых значений признаков. При этом, максимальная точность составила 90% при найденных 6 наборах.

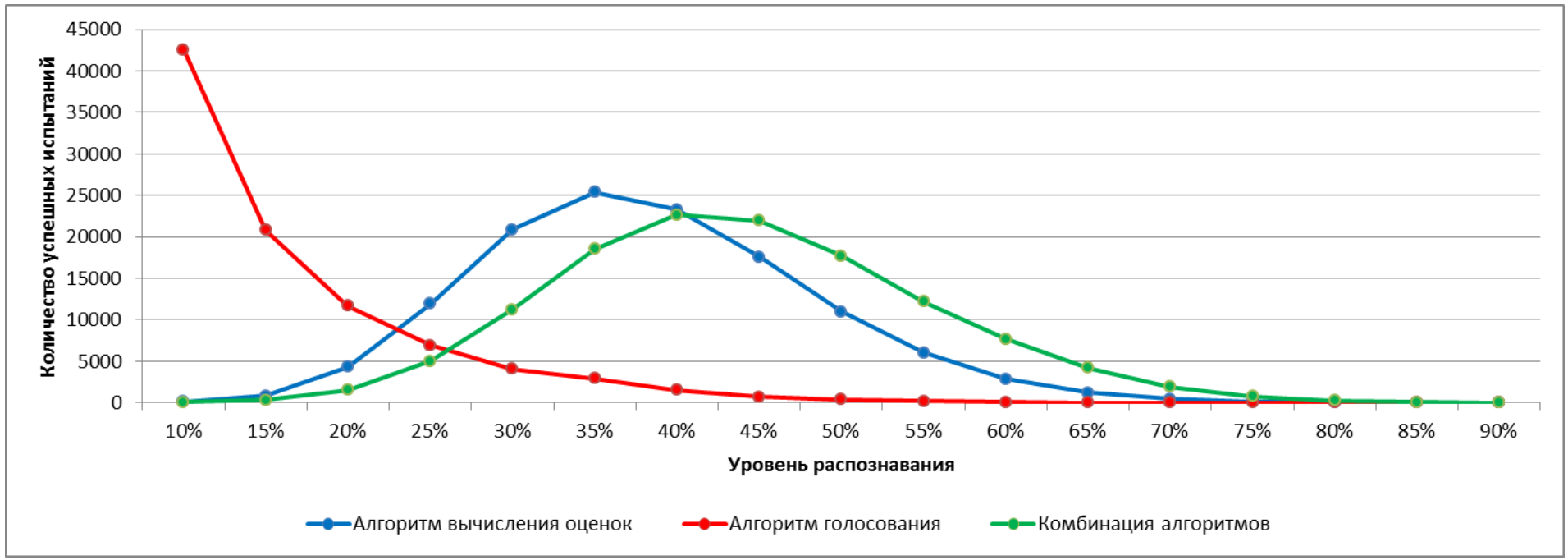

Рис. 3.2. График распределения числа успешных испытаний по уровням распознавания

Табл. 3.5. Распределение числа успешных испытаний по уровням распознавания

Алгоритм/Уровень распознавания	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**	**75%**	**80%**	**85%**	**90%**
Алгоритм вычисления оценок	105	798	4277	11949	20846	25360	23268	17560	10942	6030	2876	1189	461	143	38	7	1
Алгоритм голосования	42555	20786	11639	6937	4025	2928	1532	665	341	156	70	17	6	4	0	0	0
Комбинация алгоритмов	47	317	1535	5027	11247	18512	22678	21922	17701	12150	7645	4163	1888	756	214	44	6

3.2. Прогнозная система как инструмент проверки гипотез

Построенная система позволяет не только прогнозировать мощность опасного явления, но и является также инструментом для научных исследований в части определения важности признаков явления, влияющих на итоговый результат [Малыгин, 2015]. Прямое сравнение признаков в сложных задачах с небольшим объемом доступных наблюдений не отражает реальной картины зависимости. Признаки могут задавать различные характеристики, могут быть выражены в несравнимых единицах измерения, отражать более сложные зависимости, чем набор логических следствий «если-то».

Для решения вышеуказанных проблем используется естественная мера значимости признака по вкладу в итоговый результат классификации – информационный вес признака (п. 1.4.). Упорядочивание информационных весов позволяет разбить признаки на группы: ведущие, значимые, незначимые. Для каждого эксперта независимо строится таблица тестов и, следовательно, определяется свой вектор информационных весов. Репрезентативная группа найденных пороговых наборов позволяет качественно оценить меру вклада признаков в итоговый результат и их взаимосвязи.

Ниже представлен пример исследования признакового пространства задачи прогнозирования ледового заторообразования на реке Северная Двина. Для этого проведено обучение системы по критерию: $\mathrm{accuracy} \geq 75\%$. При этом лучший результат составил 85%. В результате обучения системы найдено 169 векторов и соответствующие им векторы информационных весов. Итоговый результат нормирован и приведен к 100%. Ранжирование признаков по убыванию вклада в итоговый результат классификации приведено на рис. 3.3. Хорошо видно, что сами признаки распадаются по значимости на несколько уровней-орбит. В порядке убывания значимости можно выделить пять уровней: №11; №№10, 4; №№5, 9, 6; №№8, 2, 7; №№1, 3.

Аналогичным образом исследованы зависимости признаков: по всем векторам и в каждом случае по всем тестам построены матрицы линейных

коэффициентов корреляции признаков. Итоговый результат сведен в матрицу, она симметричная, поэтому покажем только верхний треугольник (Табл. 3.6). В каждой клетке итоговой матрицы находится доля экспертов, у которых корреляция между соответствующими признаками составила от 0.1 до 0.4. Линейный коэффициент корреляции более 0.4 у найденных наборов не встречается.

Рис. 3.3. Вклад признаков в итоговый результат классификации.
Приведена интерпретация важности на основе расчета
информационных весов признаков.
Сумма по всем признакам составляет 100%

В результате видно, что, хотя и есть некоторые корреляции отдельных признаков, признаковое пространство подобрано хорошо. Все признаки влияют на итоговый результат, нет сильно зависимых признаков. Проделана качественная экспертная работа, результаты согласуются с гипотезами, выдвинутыми в [Агафонова, Фролова, 2007; Агафонова, Василенко, Фролова, 2016]. Признаки 11, 10, 9, 4, 5, 6 вносят наибольший вклад в итоговый результат, что, в целом, соотносится с теоретическими исследованиями.

Табл. 3.6. Итоговая матрица корреляции признаков

Признаки	1	2	3	4	5	6	7	8	9	10	11
1		0	0	0.01	0	0	0	0	0	0	0.02
2			0	0.01	0	0	0	0	0	0	0.02
3				0	0	0	0	0	0	0	0
4					0.04	0.08	0	0	0.01	0.13	0
5						0	0	0.01	0	0.17	0
6							0.01	0	0	0.01	0
7								0	0	0	0
8									0	0	0.01
9										0.05	0.14
10											0.35
11											

Описанный способ функционирования прогнозной системы для работы с данными не утверждает, что приведенное признаковое пространство является полным, т.е. что других признаков, влияющих на явление, не существует. Однако, в случае выдвижения экспертом гипотезы о значимости другого признака, не входящего в сформированное пространство, позволяет эффективно ее проверить.

На доступных данных были проведены эксперименты с целью проверки корректности работы прогнозной системы, определены числовые показатели, подтверждающие возможность применения разработанной прогнозной системы в реальных задачах прогнозирования. Экспериментально исследован логический процесс используемого алгоритма. Экспертная система позволяет изучить признаковое пространство и исходные данные конкретного процесса, т.е. допускает применение в качестве инструмента научных исследований. Список проведенных экспериментов представлен в Табл. 3.7.

Табл. 3.7. Список проведенных экспериментов на периоде разработки

№	Название эксперимента	Число случайных испытаний	Процент успешных испытаний	Время[3] (час.)	Результат эксперимента
1	Сравнение алгоритмов обучения	128000	98,34%	30	Исследована зависимость количества оптимальных параметров обучения для различных уровней достоверности распознавания. Результаты представлены в п. 3.1.
2	Оценка продолжительности процесса обучения ЭС	128000	98,34%	30	Получено время работы (для конкретных исходных данных).
3	Оценка достоверности прогнозирования ЭС	32000	98,30%	8	Для обучения системы комбинацией алгоритмов голосования и вычисления оценок получен максимальный уровень достоверности прогнозирования 85%. Приведены эталоны, на которых происходят ошибки обучения. Результаты представлены в п. 3.1.
4	Исследование признакового пространства задачи прогнозирования с помощью ЭС	128000	98,32%	30	Для уровня достоверности обучения не менее 75% найдены оптимальные наборы параметров обучения. Для каждого набора построена таблица тестов. Корреляционный анализ таких таблиц позволил произвести разделение признаков по группам информационной значимости. Результаты представлены в п. 3.1.

[3] Время работы измерено для однопоточного режима вычислений на стандартном процессоре Intel Core i5 2450M @ 2.5 GHz

3.3. Валидация прогнозной системы

Для решения проблемы прогнозирования заторообразования использован подход, основанный на комбинации алгоритмов, разработанных в рамках теории распознавания образов и машинного обучения. Подробное описание метода применительно к изучаемому региону содержится в п. 3.1. К моменту разработки экспертной системы были доступны данные за период времени с 1991 по 2010 гг. Разработка и настройка экспертной системы, а также, собственно, прогнозирование проводилась на этой выборке, при этом оценка достоверности прогнозирования составила 85%. В настоящее время стали доступны новые метеорологические и гидрологические данные пунктов наблюдения, полученные за период 2011-2016 гг. [ВНИИ гидромет. инф.]. Это позволяет провести независимую оценку точности прогноза, что и является предметом валидации [Алешин, Малыгин, 2018; Aleshin, Malygin, 2018].

Основное назначение экспертной системы состоит в получение краткосрочного прогноза образования сильных ледяных заторов на изучаемом участке Северной Двины. Как и в классическом подходе, прогноз основывается на имеющихся данных, однако эти данные используются для обучения системы, т.е. подбора векторов пороговых значений свободных параметров. Подобранные значения этих параметров позволяют построить краткосрочный прогноз образования затора в текущем сезоне.

В качестве результата работы системы используется логическая переменная, принимающая два значения: «1» соответствует наличию сильных заторов на исследуемом участке, «0» – затор небольшой мощности или их полное отсутствие. При этом, прогноз дополняется оценкой точности предсказания. Термин «мощность затора», по своей природе, величина эвристическая и не имеет точного определения. В зависимости от задачи для ее формализации могут использоваться различные параметры или их комбинации. Так как в рассматриваемом случае главное интересующее нас следствие образования затора – возможность наводнения, то как основу для оценки

мощности естественно использовать вызванный им подъем уровня воды. Считается, что затор сильный, если максимальный уровень подъема воды на каком-либо посту превысил половину исторического максимума (по всему доступному периоду с 1991 г.), что, в целом, согласуется с оценками специалистов. Итоговая классификация представлена в Табл. 3.8.

Входная информация прогнозной системы представляет собой ряд гидрологических и метеорологических признаков, фиксируемых на речных постах. Название признаков, их краткое описание и единицы измерений приведены в Табл. 3.1. Одна часть из них описывает гидрометеорологическую ситуацию в осенне-зимний период, другая – в весенний. Очевидно, что последние играют более важную роль в процессе заторообразования, однако в отсутствие адекватной физической модели процесса система должна обрабатывать все признаки единым образом. Тем не менее, получение реалистичного прогноза возможно только после того, как получена вся совокупность «весенних» данных. Таким образом, прогноз, полученный на основе предлагаемой экспертной системы, является краткосрочным, с заблаговременностью 3-7 дней.

Таблица 3.8. Классификация сезонов. Ячейки с номерами лет, в которых наблюдались сильные заторы (класс «1») выделены серым фоном

Сезоны			
1991	1998	2005	2012
1992	1999	2006	2013
1993	2000	2007	2014
1994	2001	2008	2015
1995	2002	2009	2016
1996	2003	2010	
1997	2004	2011	

Общая схема метода такова. Совокупность значений признаков на момент составления прогноза определяет состояние системы и является элементом множества. Это состояние разбивается на несколько непересекающихся подмножеств. В нашем случае таких подмножества всего два: наличие или отсутствия сильных заторов. Наша задача состоит в выработке процедуры классификации, т.е. отнесении элемента к одному из классов.

Как и ранее в п. 3.1., процедура построения решения задачи состоит из двух этапов. На первом этапе производится обучение системы. Для этого используется обучающая выборка – набор объектов, для которых результирующее состояние известно. Затем составляется решающее правило, на основании которого каждый набор может быть отнесен к тому или иному классу состояний. Решающее правило содержит зависимость от набора свободных параметров, и процесс обучения представляет процедуру определения таких значений этих параметров, которые обеспечивают наилучшую классификацию наборов состояний. На втором этапе решающее правило с подобранными на основе обучения параметрами используется для классификации наборов, не входящих в обучающую выборку. Оценка качества осуществляется с помощью Leave-one-Out кросс-валидации (п. 1.2.).

Первоначально, описанная система была апробирована на периоде наблюдений 1991-2010, оценка качества прогнозирования составила 85%. При получении новых данных за 2011-2016 гг. были проделаны два эксперимента имитирующие применение системы в реальных условиях. В первом случае для составления прогноза на каждый из шести добавленных сезонов использовал набор параметров, полученных ранее (при обучении на данных 1991-2010 гг.). Во втором случае для составления прогноза на добавленный сезон производилось полное переобучение системы: используя описанные в п. 1.4 и п. 3.1 процедуры заново строились наборы параметров, обеспечивающих необходимое качества прогноза. Затем процесс повторялся для следующего

периода (Табл. 3.9.). Оба подхода дали одинаковый результат, состоящий в успешном прогнозировании образования заторов для всех шести новых сезонов.

Таблица 3.9. Результаты эксперимента

Сезон	2011	2012	2013	2014	2015	2016
Реализовавшийся сценарий	1	2	1	1	1	1
Прогнозируемый сценарий	1	2	1	1	1	1

Следует отметить, что, хотя представленные в табл. 3.9 результаты прогнозирования совпадают с реальными реализовавшимися результатами, однако, малая мощность отложенной выборки позволяет утверждать только то, что полученный результат не хуже оценки качества в 85%, полученной ранее в [Малыгин, 2014 *«Методика…»*]: в результате оценки качества с использованием кросс-валидации на всех доступных данных было получено, что оценка метрики качества составила 84.6%.

На этапе валидации системы были построены распределения найденных пороговых значений признаков (рис. 3.5). По оси ординат отложены диапазоны изменения абсолютных значений признаков, по оси абцисс – количество значений порогов, находящихся в соответствующем диапазоне (бине). На Рис 3.5 видно, что признаки имеют мультимодальные распределения, а у признаков 11, 4, 6 и 7 есть несколько максимальных значений, соответствующих различным бинам в разбиении числовой шкалы признаков. Это указывает на сложность и многофакторность исследуемого процесса.

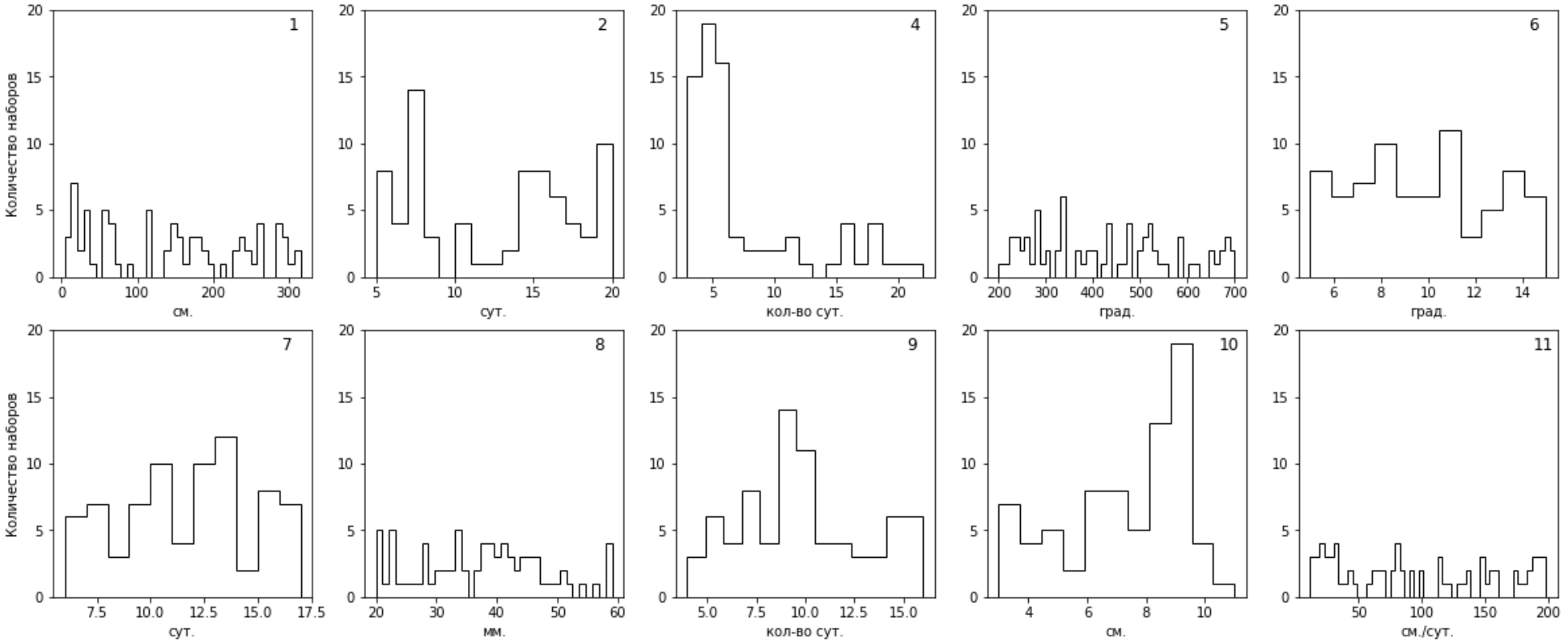

Рис 3.5. Распределение пороговых значений признаков. На рисунке представлены 10 из 11 признаков (кроме признака №3 принимающего бинарные значения). По оси ординат отложены диапазоны изменения абсолютных значений признаков, по оси абцисс – количество значений порогов, находящихся в соответствующем диапазоне (бине).

Применение описанной экспертной системы к новым данным наблюдений позволяет говорить о ее состоятельности и эффективности. Прогнозная экспертная система реализована на языке C++ в виде консольного приложения Windows, на которое получено свидетельство о регистрации ПО [Малыгин, 2014 *«Свидетельство…»*]. При этом время расчета прогноза при полном анализе данных составляет несколько часов. При необходимости, это время может быть существенно снижено за счет использования современных высокопроизводительных библиотек для параллельных вычислений и вычислений на GPU.

ЗАКЛЮЧЕНИЕ

В результате проведенных исследований и практических разработок **была достигнута цель диссертационного исследования** – разработка компьютерных систем и методов обработки данных в условиях ограниченного количества данных, не достаточных для проведения классического статистического анализа, на основе методов машинного обучения, включая интерпретацию результатов измерений геофизических полей и анализа многопараметрической информации, применение таких систем для построения геолого-геофизических моделей и решения задач охраны окружающей среды.

Проведен анализ методов машинного обучения в аспекте применения к геофизическим задачам: использование базовых методов, разработка более глубоких подходов. Приведены примеры применения методов машинного обучения в геофизических задачах с недостатком данных. Введены базовые понятия и алгоритмы машинного обучения. Проведена их необходимая адаптация для возможности применения к исследуемым геофизическим примерам.

Разработан метод анализа пространственных данных для построения двумерных и трехмерных (на основе набора 1D-профилей) изображений на основе алгоритма *k* ближайших соседей. Разработанный метод применен в задаче построения двумерных моделей строения коры северной части Балтийского щита. Построена уточненная карта поверхности Мохоровичича. Основу исследования составляют данные, полученные методом приемных функций. Были использованы сведения, полученные в предыдущих исследованиях этого региона, дополненные новыми расчетами и измерениями. Исходные данные представляют собой набор зависимостей сейсмической скорости от глубины, рассчитанных для более чем 60 постоянных и временно действующих геофизических станций. С точки зрения машинного обучения, данная задача является задачей регрессии. Для восстановления регрессионной

зависимости глубины Мохо от двумерных координат был использован метод k ближайших соседей с необходимой адаптацией в части выбора метрики.

Еще одним результатом в данной задаче стало построение карты слоя с низкими значениями скорости поперечных сейсмических волн V_S. В исследуемом регионе практически отсутствует осадочный слой. Несмотря на это, имеются области, в которых присутствует слой с низкими значениями скорости V_S. Относительно низкие значения V_S обычно объясняют наличием в слое большого количества водонасыщенных трещин. Присутствие такого слоя не зависит от возраста пород. Эта задача относится к оценке принадлежности в рамках задачи бинарной классификации. Для небольшого количества сейсмических станций (порядка 20) известно наличие или отсутствие слоя низких скоростей. С помощью метода k ближайших соседей в каждой точке исследуемого региона оценена вероятность наличия слоя низких скоростей. Построенная карта включает в себя классификацию по принципу наличия или отсутствия слоя низких скоростей, а также буферную область, в которой на основании имеющихся данных нельзя сделать однозначный вывод. Показано, что слой низких сейсмических скоростей на поверхности присутствует на значительной части региона, включая области с протерозойскими породами. В южной части Финляндии положение низкоскоростной области коррелирует с относительно низким значением толщины коры.

В задаче построения трехмерной модели среды при проведении межскважинных исследований предложена новая интерпретация данных радиоволнового просвечивания, позволяющая более точно выделить границы слоев по сравнению с методами, используемыми ранее (кригинг). Использованный метод k ближайших соседей позволяет построить трехмерную модель проводимости среды между скважин даже при использовании синхронной схемы измерений. Влияние анизотропии распределения данных можно исключить, если модифицировать пространственную метрику, определяющую расстояние между данными. Это достигается введением

коэффициента скейлинга, который изменяет масштаб в горизонтальном направлении. Использованный подход позволяет также получить достаточно контрастное изображение неоднородных областей, что позволяет выделить неоднородности, чьи геометрические размеры меньше расстояния между скважинами. Процесс построения трехмерной модели фактически не зависит от физической модели, использованной для интерпретации измерений. Уточнение физической модели процесса распространения радиоволн между скважинами позволит улучшить качество построения образа. Модель может быть улучшена, если привлечь дополнительные данные (геологические, сейсмические, магнитные) для их совместной интерпретации.

Разработан метод для анализа многомерных временных рядов ограниченной длины на основе прогнозной системы. Создание прогнозных систем, включающих в себя методы теории искусственного интеллекта, является актуальным развитием геоинформационных систем, а задача прогнозирования опасных природных явлений является востребованной в любой отрасли хозяйственной деятельности человека. Прогнозная система, основанная на разработанной методике, качественно проявила себя при решении трудно-формализуемой задачи прогнозирования опасного природного явления – образование заторов льда на участке р. Северная Двина. Система также допускает применение в качестве инструмента проверки гипотез относительно влияния тех или иных факторов на итоговые состояния опасных процессов. Проведена валидация системы на новых, недоступных в момент первоначальной разработки, данных. Итоговая оцененная достоверность прогнозирования составила 85%.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Агафонова С.А., Фролова Н.Л. Особенности ледового режима рек бассейна Северной Двины // Водные ресурсы. — 2007. — Т. 34, № 2. — С. 141–149.

2. Агафонова С.А., Василенко А.Н., Фролова Н.Л. Факторы образования ледовых заторов на реках бассейна Северной Двины в современных условиях // Вестник Московского университета. Серия 5: География. — 2016. — № 2. — С. 82–90.

3. Алексеевская М.А., Габриэлов А.М., Гвишиани А.Д., Гельфанд И.М., Ранцман Е.Я. Морфоструктурное районирование горных стран по формализованным признакам // Вычислительная сейсмология. Вып.10. Распознавание и спектральный анализ в сейсмологии / Под ред. Кейлис-Борока В.И. —М., 1977. — С.33−49.

4. Алешин И.М., Малыгин И.В. Верификация экспертной системы прогноза заторообразования на Северной Двине // Геофизические процессы и биосфера. — 2018. — Т. 17, № 2. — С. 48–60.

5. Алешин И.М., Ваганова Н.В., Косарев Г.Л., Малыгин И.В. Свойства коры Фенноскандии по результатам kNN-анализа инверсии приемных функций // Геофизические исследования. — 2019. —Т.20, №4. — С. 25–39.

6. Алешин И.М., Малыгин И.В. Интерпретация результатов радиоволнового просвечивания методами машинного обучения // Компьютерные исследования и моделирование. — 2019. — Т. 11, № 4. — С. 675–684.

7. Алешин И.М., Косарев Г.Л., Ризниченко О. Ю., Санина И.А. Скоростной разрез земной коры под сейсмической группой Ruksa, Карелия // Геофизические исследования. — 2007. — №7. — С. 3–13.

8. Алешин С. В. Распознавание динамических образов. — изд-во МГУ Москва, 1996. — 97 с.

9. Ален К, Кейлис-Борок В.И., Кузнецов И.В. и др. Долгосрочный прогноз землетрясений и автомодельность сейсмических предвестников. Калифорния, M ≥ 6.4, M ≥ 7.0 // В кн.: Достижения и проблемы современной геофизики. — М.: Наука. — 1984. — С.152–165.

10. Ален К, Кейлис-Борок В.И., Ротвайн И.М., Хаттен К. Комплекс долгосрочных сейсмологических предвестников. Калифорния и некоторые другие регионы. // В сб.: Математические методы в

сейсмологии и геодинамике (Вычислительная сейсмология). — М.: Наука. — 1986, —вып.19. — С. 23–37.

11. Андреев А.Е., Гасанов Э.Э., Кудрявцев В.Б. Теория тестового распознавания. Монография. — М.: Физматлит, 2007. — 320 стр.

12. Берденников В.П. Модельные исследования механизма заторообразования для обоснования схемы ледозадержания на р. Днестре и определения ледовых нагрузок // Труды ГГИ. — 1974, — вып. 219. — С. 31–55.

13. Берденников В.П., Шматков В.А. Натурные и лабораторные исследования образования заторов льда // Труды IV Всесоюзного гидрологического съезда. — Т.6. — Ленинград. — 1976. — С. 361–370.

14. Близняк Е.В. Река Енисей от Красноярска до Енисейска, ч. II: зимнее состояние реки. — Санкт-Петербург, 1916. — 79 с.

15. Бонгард М.М. Проблема узнавания. — М.: Наука. — 1967. — 320 с.

16. Бузин В.А. О наводнениях на реках, вызванных заторами льда // Водные ресурсы. — 2000. — т.27. — №5. — С.524-530.

17. Бузин В.А. Условия и прогноз подвижек льда при замерзании р. Нева // Метеорология и гидрология. — 1997. — №8. — С.83-87.

18. Бузин В.А. Заторы льда и заторные наводнения на реках. — Санкт-Петербург, Гидрометеоиздат. — 2004. — 203 с.

19. Бузин В.А., Зиновьев А.Т. Ледовые процессы и явления на реках и водохранилищах. Методы математического моделирования и опыт их реализации для практических целей (обзор современного состояния проблемы). — Барнаул. — ООО «Пять плюс». — 2009. — 168 с.

20. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов. — Москва, Наука. — 1980. — 416 с.

21. Варенцов И.М., Корья Т., Пальшин Н.А., Смирнов М. Ю., Энгельс М., Рабочая группа BEAR Обобщенная объемная геоэлектрическая модель коры Балтийского региона и структура ее глубинных электромагнитных откликов // Строение и динамика литосферы Восточной Европы. Результаты исследований по программе EUROPROBE. — М.: Геокарт: ГЕОС. — 2006. — С. 95–106.

22. ВНИИ гидрометеорологической информации – мировой центр данных [электронный ресурс] // http://meteo.ru

23. Гафуров С.В., Краснопрошин В.В., Образцов В.А. Проблема неполноты информации в задаче распознавания образов // Вестник БГУ. Сер. 1, Физика. Математика. Информатика. — 2007. — № 2. — С. 113–115.

24. Гвишиани А.Д. Устойчивость по времени прогноза мест сильных землетрясений. I. Юго-Восточная Европа и Малая Азия // Известия Академии наук СССР. Физика Земли. —1982. — №8. — С.13–19.

25. Гвишиани А.Д., Агаян С.М., Дзебоев Б.А., Белов И.О. Распознавание мест возможного возникновения эпицентров сильных землетрясений с одним классом обучения // Докл. РАН. — 2017. — Т. 474, №1. — С. 86–92.

26. Гвишиани А.Д. Системный анализ в изучении Арктики. Доклад на конференции ITES&MP-2019 (video.sgm.ru/records/20/1/01), дата обращения 23.11.2019

27. Гельфанд И.М., Губерман Ш.А., Извекова М.Л., Кейлис-Борок В.И., Ранцман Е.Я. О критериях высокой сейсмичности // Доклады Академии наук СССР. — 1972. — Т. 202, №6. — С. 1317−1320.

28. Гельфанд И.М., Губерман Ш.А., Кейлис-Борок В.И., Кнопов Л., Пресс Ф.С., Ранцман Е.Я., Ротвайн И.М., Садовский А.М. Условия возникновения сильных землетрясений (Калифорния и некоторые другие регионы) // В сб.: Исследование сейсмичности и моделей Земли (Вычислительная сейсмология). — М.: Наука. — 1976. — Вып. 9. — С. 3–91.

29. Герасимов И.П. Опыт геоморфологической интерпретации общей схемы геологического строения СССР. // Проблемы физической географии. — М.; — Л.: Изд-во АН СССР. — 1947. — Вып.12. — С. 30–51.

30. Горелик А.Л., Скрипкин В.А. Методы распознавания. — М.: Высшая школа. — 1977. — 222 с.

31. Горшков А.И., Кособоков В.Г., Ранцман Е.Я., Соловьев А.А. Проверка результатов распознавания мест возможного возникновения сильных землетрясений с 1972 по 2000 г. // В сб.: Проблемы динамики литосферы и сейсмичности (Вычислительная сейсмология). — М.: Геос. — 2001. — Вып.32. — С. 48–57.

32. Деев Ю.А., Попов А.Ф. Весенние заторы льда в русловых потоках. Физические основы и количественный анализ. — Л.: Гидрометеоиздат. — 1978. — 110 с.

33. Джарратано Д., Райли Г. Экспертные системы: принципы разработки и программирование. — М.: Вильямс. — 2007. — 1152 с.

34. Джексон П. Введение в экспертные системы. — М.: Вильямс. — 2001. — 624 с.

35. Дзебоев Б.А., Гвишиани А.Д., Белов И.О., Агаян С.М., Татаринов В.Н., Барыкина Ю.В. Распознавание мест возможного возникновения сильных землетрясений на основе алгоритма с единственным сильным классом обучения: I Алтай – Саяны – Прибайкалье M ≥ 6.0 // Физика Земли. — 2019. — №4. — С. 33–47.

36. Донченко Р.В. Ледовый режим рек СССР. — Л.: Гидрометеоиздат. — 1987. — 248 с.

37. Жамалетдинов А. А., Колесников В. Е., Скороходов А. А., Шевцов А. Н., Нилов М. Ю., Рязанцев П. А., Шаров Н. В., Бируля М. А., Киряков, И. А. Результаты электропрофилирования на постоянном токе в комплексе с АМТЗ по профилю, пересекающему Ладожскую аномалию // Труды Карельского научного центра Российской академии наук. — 2018. —№2. — С. 91–110.

38. Жданов М.С. Электроразведка. — М.: Недра, 1986. — 316 с.

39. Журавлев Ю.И., Никифоров В.В. Алгоритмы распознавания, основанные на вычислении оценок // Кибернетика. — 1971. — №3. — С. 1–11.

40. Журавлев Ю.И., Рязанов В.В., Сенько О.В. Распознавание. Математические методы. Программная система. Практические применения.: — М.: Фазис, 2006. — 159 с.

41. Завьялов А.Д. Среднесрочный прогноз землетрясений: основы, методика, реализации / Монография. — Москва, 2004. — 260 с.

42. Иогансон Е.И. Зимний режим р. Волхова и оз. Ильмень // Материалы по исследованию р. Волхова и его бассейна. — Л.: Из-во строительства Волховской ГЭС, 1927, вып. 4. — С. 23–35.

43. Истратов В.А., Лысов М.Г., Чибрикин И.В., Матяшов С.В. Радиоволновая геоинтроскопия (РВГИ) межскважинного пространства на месторождениях нефти // Геофизика. — 2000. Спецвыпуск. — С. 59–68.

44. Истратов В.А., Скринник А.В., Перекалин С.О. Новая аппаратура для радиоволновой геоинтроскопии горных пород в межскважинном пространстве «РВГИ-2005» // Приборы и системы разведочной геофизики. — 2006. — №1. — С. 37-43.

45. Истратов В.А., Колбенков А.В., Перекалин Е.В., Лях С.О. Радиоволновой метод мониторинга технологических процессов в межскважинном

пространстве // Вестник КРАУНЦ. Серия: Науки о Земле. — 2009. — Т. 14. — С. 59–68.

46. Кеворкянц С.С., Абрамов В.Ю., Ковалев Ю.Д. Скважинный радиоволновой комплекс при поисках кимберлитовых трубок в Западной Якутии // Геофизика. — 2005. —Т. 3. — С.56–64.

47. Кейлис-Борок В.И., Кособоков В.Г. Комплекс долгосрочных предвестников для сильнейших землетрясений мира. // В кн.: Землетрясения и предупреждение стихийных бедствий. 27-ой международный геологический конгресс (СССР, Москва, 4-14 августа 1984). Коллоквиум 06. — М.: Наука. — 1984. — Т.61. — С. 56–66.

48. Кейлис-Борок В.И., Кособоков В.Г. Периоды повышения вероятности возникновения для сильнейших землетрясений мира. // В сб.: Математические методы в сейсмологии и геодинамике (Вычислительная сейсмология). — М.: Наука. — 1986. — Вып.19. — С. 48–58.

49. Константинов Р.М., Королева З.Е. Применение тестовых алгоритмов к задачам геологического прогнозирования // Распознавание образов. Тр. Международного симпозиума 1971г. по практическим применениям методов распознавания образов. — Москва, ВЦ РАН СССР. — 1973. — С. 194–199.

50. Константинов Р.М., Королева З.Е., Кудрявцев В.Б. Комбинаторно-логический подход к задачам прогноза рудоносности. // Проблемы кибернетики. — М.: Наука. — 1976. — Вып. 31. — С. 5–38.

51. Кошель С.М., Мусин О.Р. Методы цифрового моделирования: кригинг и радиальная интерполяция // Информационный бюллетень ГИС-Ассоциации. — 2001. — № 2(29)-3(30). — С. 23–24.

52. Кудрявцев В.Б., Гасанов Э.Э., Подколзин А.С. Введение в теорию интеллектуальных систем. — М.: Изд-во ф-та ВМиК МГУ, 2006. — 208 с.

53. Кудрявцев В.Б. Теория тестового распознавания // Дискретная математика. — 2006. — Т. 18, вып. 3. — С. 3–34.

54. Кузнецов Н.М. Опыт применения радиоволновой геоинтроскопии межскважинного пространства для разведки золотомедного месторождения // Разведка и охрана недр. — 2008. — № 12. — С. 27–29.

55. Кузнецов Н.М. Способ 3D-обработки данных радиоволнового просвечивания межскважинного пространства // Вестник КРАУНЦ. Сер. Науки о Земле. — 2012. — Т. 1. — С. 240–246.

56. Лукк А.А., Леонова В.Г., Сидорин А.Я. Еще раз о природе сейсмичности Фенноскандии // Геофизические процессы и биосфера. — 2019. — Т. 18. — №1. — С. 74–90.

57. Малыгин И.В. Логический подход к созданию экспертных систем прогнозирования опасных природных явлений // Естественные и технические науки. — 2015. — № 2. — С. 102–112.

58. Малыгин И.В. Методика прогноза образования ледовых заторов на реках на основе теории распознавания образов // Вестник Московского университета. Серия 5: География. — 2014. — № 3. — С. 43–47.

59. Малыгин И.В. О задаче прогнозирования ледовых заторов // Интеллектуальные системы. Теория и приложения. — 2014. — Т. 18, № 3. — С. 73–80.

60. Малыгин И.В. Свидетельство о государственной регистрации программы для ЭВМ №2014614960 Экспертная система прогнозирования ледового заторообразования. Дата гос. регистрации в Реестре программ для ЭВМ 14.05.2014.

61. Малыгин И.В. Формирование параметров обучения в прогнозных экспертных системах // Наука и мир. — 2013. — № 3. — С. 34–35.

62. Малыгин И.В., Алешин И.М. Свидетельство о государственной регистрации программы для ЭВМ № 2020617961 Программа расчета и построения региональных карт геофизических свойств методом k-ближайших соседей. Дата гос. регистрации в Реестре программ для ЭВМ 15.07.2020.

63. Малыгин И.В., Алешин И.М. Свидетельство о государственной регистрации программы для ЭВМ № 2020617962 Программа расчета и построения трехмерной модели проводимости среды по данным межскважинных измерений методом k-ближайших соседей. Дата гос. регистрации в Реестре программ для ЭВМ 15.07.2020.

64. Минц М.В., Соколова Е.Ю., рабочая группа Ладога Объемная модель глубинного строения свекофеннского аккреционного орогена по данным МОВ-ОГТ, МТЗ и плотностного моделирования // Труды Карельского научного центра РАН. — 2018. — № 2. — С. 34–61.

65. Николенко С.И., Кадурин А.А., Архангельская Е.О. Глубокое обучение. — СПб.: Питер, 2018. — 479 с.

66. Панфилов Д.Ф. Закономерности движения воды и льда в широком прямоугольном русле при сплошном ледоходе // Метеорология и гидрология. — 1968. — №8. — С. 41–44.

67. Петровский А.Д. Радиоволновые методы в подземной геофизике. — М.: Недра, 1971. — 224 с.

68. Поспеева Е.В., Матросов А.В., Манаков В.А. Геоэлектрическая неоднородность земной коры в связи с кимберлитовым магматизмом юга якутской алмазоносной провинции // Вестник Воронежского государственного университета. Сер. Геология. — 2004. — Т. 1. — С. 137–147.

69. Ранцман Е.Я. Места сильных землетрясений: постановка задачи и способ решения // В сб.: Проблемы динамики литосферы и сейсмичности (Вычислительная сейсмология). — М.: Геос. — 2001. — Вып. 32. — С. 43–47.

70. Толстов А.В., Зинчук Н.Н., Серов И.В. Основные результаты научно-исследовательских и опытно-методических работ НИГП АК «АЛРОСА» (ПАО) // Эффективность геологоразведочных работ на алмазы: прогнозно-ресурсные, методические, инновационно-технологические пути ее повышения: сборник. — 2018. — С. 12–30.

71. Уотермен Д. Руководство по экспертным системам. — М.: Мир. — 1989. — 388 с.

72. Федоров М.К. Заторные и зажорные явления и их развитие на реке Лене // Труды ААНИИ. — 1956. — Т. 204. — С. 62–95.

73. Чеботарев А.И. Гидрологический словарь. — Л.: Гидрометеоиздат. — 1978. — 308 с.

74. Черепанов А.О. Многочастотные радиоволновые измерения в скважинах для контроля за процессом оттаивания ММП (на примере месторождения нефти «Русское», Западная Сибирь) // Вестник КРАУНЦ. Сер. Науки о Земле. — 2017. — № 4. — С. 118–123.

75. Чижов А.Н. О механизме формирования заторов льда и их типизация // Труды ГГИ. — 1975. — Вып. 227. — С. 3–17.

76. Шмаков И.И. Проблемы научного сопровождения при геологоразведочных работах на алмазы // Геология и минерагения Северной Евразии: материалы совещания, приуроченного к 60-летию Института геологии и геофизики СО АН СССР. — 3–5 октября 2017. Новосибирск, Россия. — С. 265.

77. Шуляковский Л.Г. О заторах льда и заторных уровнях воды при вскрытии рек // Метеорология и гидрология. — 1951. — №7. — С. 45–49.

78. Шуляковский Л.Г., Еремина В.И. К методике прогноза заторных уровней воды // Метеорология и гидрология. — 1952. — №1. — С. 46–51.

79. Шуляковский Л.Г. Появление льда и начало ледостава на реках, озерах и водохранилищах. Расчеты для целей прогнозов. — Л.: Гидрометеоиздат. — 1960. — 216 с.

80. Элти Дж., Кумбс М. Экспертные системы: концепции и примеры. — М.: Финансы и Статистика. — 1987. — 191 с.

81. Яблонский С.В. О тестах для электрических схем // Успехи математических наук. — 1955. — Т.10. — Вып. 4(66). — С. 182-184.

82. Яблонский С.В., Демидова Н.Г., Константинов Р.М., Королева З.Е., Кудрявцев В.Б., Сиротинская С.В. Тестовый подход к количественной оценке геолого-структурных факторов и масштабов оруднения (на примере ртутных месторождений) // Геология рудных месторождений. — 1971. — Т.13. — №2. — С. 30–42.

83. Яновская Т.Б., Лыскова Е.Л., Королева Т.Ю. Радиальная анизотропия верхней мантии Европы по данным поверхностных волн // Физика Земли. — 2019. — №2. — С. 3–14.

84. Agafonova S.A., Frolova N.L., Krylenko I.N., Sazonov A.A., Golovlyov Dangerous ice phenomena on the lowland rivers of european russia // Natural Hazards. — 2017. — Vol. 88, №S1. — P. 171–188.

85. Aleshin I.M., Kosarev G.L., Riznichenko O. Yu., Sanina I.A. Crustal velocity structure under the RUKSA seismic array (Karelia, Russia) // Russian Journal of Earth Sciences. — 2006. — V. 8. — №1. — P. 1–8.

86. Aleshin I.M., Malygin I.V. Machine learning approach to inter-well radio wave survey data imaging // Russian Journal of Earth Sciences. — 2019. — V. 19, no. ES3003. — P. 1–6.

87. Aleshin I. M., Malygin I.V. Verification of an expert system for forecasting ice-block-formation: The case of the Northern Dvina river // Izvestiya - Atmospheric and Oceanic Physics. — 2018. — V. 54, №8. — P. 898–905.

88. Aleshin I.M., Zhandalinov V.M. Application of interpolation procedures for presentation of data electromagnetic wave lightning // Russian Journal of Earth Sciences. — 2009. — V. 11, №1. — P. 1–4.

89. Alinaghi A., Bock G., King R., Hanka W., Wylegalla K., TOR and SVEKALAPKO Working Groups Receiver function analysis of the crust and upper mantle from the North German Basin to the Archaean Baltic Shield // Geophysical Journal International. — 2003. — V. 155, №2. — P. 641–652.

90. Altman N.S. An introduction to kernel and nearest-neighbor nonparametric regression // The American Statistician. — 1992. — V. 46, №3. — P. 175–185.

91. BeiDou Navigation Satellite System [web-resource] // http://en.beidou.gov.cn/

92. Bock G. and Seismic Tomography Working Group Seismic Probing of Fennoscandian Lithosphere // EOS, Trans. AGU. — 2001. — V. 82, № 50. — P. 621–629.

93. Bruneton M., Farra V., Pedersen H.A. Non-linear surface wave phase velocity inversion based on ray theory // Geophysical Journal International. — 2002. — V. 151, №2. — P. 583–596.

94. Dempster A.P., Laird N.M., Rubin D.B. Maximum likelihood from incomplete data via the EM algorithm //Journal of the Royal Statistical Society. — 1977. — Vol. 39. — P. 1–38.

95. Dricker I.G., Roecker S.W., Kosarev G.L., Vinnik L.P. Shear-wave velocity structure of the crust and upper mantle beneath the Kola Peninsula // Geophysical Research Letters. — 1996. — V. 23, №23. — P. 3389–3392.

96. Fawcett T. An introduction to ROC analysis // Pattern Recognition Letters. — 2006. — V. 27. — P. 861–874.

97. Flach P. MACHINE LEARNING. The Art and Science of Algorithms that Make Sense of Data. — NY: Cambridge University Press, 2012. — 396 p.

98. Frassetto A., Thybo H. Receiver function analysis of the crust and upper mantle in Fennoscandia – isostatic implications // Earth and Planetary Science Letters. — 2013. — V. 381. — P. 234–246.

99. Grad M., Luosto U. Fracturing of the crystalline uppermost crust beneath the SVEKA profile in Central Finland // Geophysica. —1992. — V. 28, №1/2. — P. 53–66.

100. Grad M., Luosto U. Seismic velocities and Q-factors in the uppermost crust beneath the SVEKA profile in Finland // Tectonophysics. — 1994. — V. 230, №1–2. — P. 1–18.

101. Grad M., Tiira T. Moho depth of the European Plate from teleseismic receiver functions // Journal of Seismology. — 2012. — V. 16, №2. — P. 95–105.

102. Goodfellow I., Bengio Y., Courville A. Deep Learning. — MIT Press, 2016. — 785 p.

103. Harmon P., Sawyer B. Creating Expert Systems for Business and Industry — NY: John Wiley and Sons, 1990.

104. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. — Springer-Verlag, 2009. — 746 p.

105. Horspool N.A., Savage M.K., Bannister S. Implications for intraplate volcanism and back-arc deformation in northwestern New Zealand, from joint inversion of receiver functions and surface waves //Geophysical Journal International. — 2006. — V. 166, №3. — P. 1466–1483.

106. Isaaks E.H., Srivastava R.M. Applied Geostatistics. — NY: Oxford University Press, 1989. — 589 p.

107. Janik T., Kozlovskaya E., Yliniemi J. Crust-mantle boundary in the central Fennoscandian shield: Constraints from wide-angle P and S wave velocity models and new results of reflection profiling in Finland // Journal of Geophysical Research: Solid Earth. — 2007. — V. 112, №B4. — P. B04302.

108. Korja T., Engels M., Zhamaletdinov A.A., Kovtun A.A., Palshin N.A., Smirnov M.Yu., Tokarev A.D., Asming V.E., Vanyan L.L., Vardaniants I.L. and the BEAR Working Group Crustal conductivity in Fennoscandia — a compilation of a database on crustal conductance in the Fennoscandian Shield // Earth, Planets and Space. — 2002. — V. 54, №5. — P. 535–558.

109. Korja T., Koivukoski K. Crustal conductors along the SVEKA profile in the Fennoscandian (Baltic) Shield, Finland // Geophysical Journal International. — 1994. — V. 116, №1. — P. 173–197.

110. Kosarev G.L., Makeyeva L.I., Vinnik L.P. Inversion of teleseismic P-wave particle motions for crustal structure in Fennoscandia // Physics of the Earth and planetary interiors. — 1987. — V. 47. — P. 11–24.

111. Kossobokov V.G., Keilis-Borok V.I., Smith S.W. Localization of intermediate-term earthquake prediction. // Journal of Geophysical Research. — 1990. — V. 95, № 12. — P.19763–19772.

112. Kozlovskaya E., Poutanen M. and POLENET/LAPNET Working Group POLENET/LAPNET – a multidisciplinary geophysical experiment in northern Fennoscandia during IPY 2007-2008 // AGU Fall Meeting Abstracts. — 2006. — S41A–1311.

113. Kozlovskaya E., Kosarev G.L., Aleshin I.M., Riznichenko O.Yu., Sanina I.A. Structure and composition of the crust and upper mantle of the Archean-Proterozoic boundary in the Fennoscandian shield obtained by joint inversion of receiver function and surface wave phase velocity of recording of the SVEKALAPKO array // Geophysical Journal International. — 2008. — V. 175, №1. — P. 135–152.

114. Krige D. A statistical approach to some basic mine valuation problems on the Witwatersrand // Journal of the Chem., Metal. and Mining Soc. of South Africa. — 1951. — V. 52, № 6. — P. 119–139.

115. Lahtinen R., Korja A., Nironen M. Paleoproterozoic tectonic evolution // Developments in Precambrian Geology. — 2005. — P. 481–531.

116. Luengo J., Garcia S., Herrera F. On the choice of the best imputation methods for missing values considering three groups of classification methods // Knowledge and Information Systems. — 2012. — V. 32. — P. 77–108.

117. Macskassy S.A., Provost F.J., Littman M.L. Confidence Bands for ROC Curves // CeDER Working Paper IS-03-04, Stern School of Business, New York University, NY, — 2003.

118. Mishra S., Shrivastava C., Ojha A., Miotti F. Waterflood Surveillance by Calibrating StreamlineBased Simulation with Crosswell Electromagnetic Data // International Petroleum Technology Conference. — 26–28 March. — Beijing, China, 2019.

119. Molinaro A.M., Simon R., Pfeiffer R.M. Prediction error estimation: a comparison of resampling methods // Bioinformatics. — 2005. — V. 21 (15). — P. 3301–3307.

120. Pedersen H., Campillo M. Depth dependence of Q beneath the Baltic Shield inferred from modeling of short period seismograms // Geophysical Research Letters. — 1991. — V. 18, №9. — P. 1755–1758.

121. Pedersen H., Debayle E., Maupin V. and POLENET/LAPNET Working Group Strong lateral variations of lithospheric mantle beneath cratons – Example from the Baltic Shield // Earth and Planetary Science Letters. — 2013. — V. 383, Supplement C. — P. 164–172.

122. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. — 2011. — V.12. — P. 2825–2830.

123. Silvennoinen H., Kozlovskaya E., Kissling E., Kosarev G.L. and POLENET/LAPNET Working Group A new Moho boundary map for the northern Fennoscandian Shield based on combined controlled-source seismic and receiver function data // GeoResJ. — 2014. — V. 1–2, Supplement C. — P. 19–32.

124. Tiira T., Janik T., Kozlovskaya E., Grad M., Korja A., Komminaho K., Hegedus E., Kovacs C.A., Silvennoinen H., Bruckl E. Crustal architecture of

the inverted Central Lapland Rift along the HUKKA 2007 profile // Pure and Applied Geophysics. — 2014. — V. 171. — P. 1129–1152.

125. Uski M., Tiira T., Grad M., Yliniemi J. Crustal seismic structure and depth distribution of earthquakes in the Archean Kuusamo region, Fennoscandian Shield // Journal of Geodynamics. — 2012. — V. 53. — P. 61–80.

126. Vapnik V. Principles of Risk Minimization for Learning Theory // Advances in neural information processing systems. — 1992. — P. 831–838.

127. Vecsey L., Plomerová J., Kozlovskaya E., Babuska V. Shear wave splitting as a diagnostic of variable anisotropic structure of the upper mantle beneath central Fennoscandia // Tectonophysics. — 2007. — V. 438, №1–4. — P. 57–77.

128. Vinnik L.P. Detection of waves converted from P to SV in the mantle // Physics of the Earth and Planetary Interiors. — 1977. — V. 15, №1. — P. 39–45.

129. Vinnik L.P., Kozlovskaya E., Oreshin S.I., Kosarev G.L., Piiponen K., Silvennoinen H. The lithosphere, LAB, LVZ and Lehmann discontinuity under central Fennoscandia from receiver functions // Tectonophysics. — 2016. — V. 667. — P. 189–198.

ПРИЛОЖЕНИЯ

Приложение 1. Данные для расчетов

Таблица 2.1. Данные, использованные для расчетов

Код станции	Широта, град.	Долгота, град.	Глубина Мохо, км.	Слой низкой скорости	Источник
FA05	62.6	31.2	48.1	0	Kozlovskaya et. al., 2008
FB07	62.1	29.6	55.1	0	Kozlovskaya et. al., 2008
FB09	61.5	28.2	59.3	0	Kozlovskaya et. al., 2008
FB11	60.8	27.0	42.3	?	Kozlovskaya et. al., 2008
FC04	63.4	30.4	55.6	0	Kozlovskaya et. al., 2008
FC05	63.1	29.9	52.3	0	Kozlovskaya et. al., 2008
FD05	63.4	29.2	57.2	1	Kozlovskaya et. al., 2008
FD07	62.7	28.0	57.4	?	Kozlovskaya et. al., 2008
FD10	61.4	26.1	57.8	1	Kozlovskaya et. al., 2008
FD13	60.5	24.7	48.3	1	Kozlovskaya et. al., 2008
FE09	62.1	26.3	59.9	0	Kozlovskaya et. al., 2008
FF05	63.8	28.0	52.1	1	Kozlovskaya et. al., 2008
FF07	63.2	26.6	58.6	1	Kozlovskaya et. al., 2008
FF11	61.8	24.3	55.8	0	Kozlovskaya et. al., 2008
FF15	60.4	22.4	56.5	0	Kozlovskaya et. al., 2008
FF30	64.5	29.1	58.0	1	Kozlovskaya et. al., 2008
FF90	62.5	25.5	66.1	0	Kozlovskaya et. al., 2008
FG01	65.4	29.6	37.2	0	Kozlovskaya et. al., 2008
FH03	65.1	27.6	56.0	1	Kozlovskaya et. al., 2008
FH05	64.4	26.5	55.0	1	Kozlovskaya et. al., 2008
FH07	63.7	25.2	55.0	1	Kozlovskaya et. al., 2008
FJ01	66.0	28.3	55.0	?	Kozlovskaya et. al., 2008
FJ10	63.0	22.7	58.0	1	Kozlovskaya et. al., 2008
RUKSA	62.1	32.2	40.0	1	Aleshin et. al., 2006
PITK	61.7	31.3	43.0	0	Новые данные
KEMI	65.0	34.7	38.0	1	Новые данные
APA	67.6	33.3	40.0	-	Dricker et. al., 1996
LVZ	67.9	34.6	40.0	-	Dricker et. al., 1996
LP01	65.5	27.5	56.0	1	Silvennoinen et. al., 2014
LP11	65.5	25.5	44.0	?	Silvennoinen et. al., 2014
LP12	65.9	26.4	44.0	0	Silvennoinen et. al., 2014
LP21	66.0	25.0	43.0	0	Silvennoinen et. al., 2014
LP31	66.6	24.1	47.0	1	Silvennoinen et. al., 2014
LP33	67.0	27.3	46.0	0	Silvennoinen et. al., 2014
LP35	67.4	29.4	45.0	1	Silvennoinen et. al., 2014
LP43	67.8	27.8	47.0	1	Silvennoinen et. al., 2014
LP51	67.5	23.6	49.0	1	Silvennoinen et. al., 2014

LP52	67.6	25.1	43.0	1	Silvennoinen et. al., 2014
LP53	68.1	27.2	49.0	1	Silvennoinen et. al., 2014
LP54	68.5	28.3	46.0	1	Silvennoinen et. al., 2014
LP61	67.9	23.9	45.0	?	Silvennoinen et. al., 2014
LP62	68.2	25.8	44.0	0	Silvennoinen et. al., 2014
LP65	68.9	28.3	44.0	1	Silvennoinen et. al., 2014
LP71	68.5	24.7	50.0	1	Silvennoinen et. al., 2014
LP72	69.0	25.7	52.0	1	Silvennoinen et. al., 2014
LP75	69.7	29.1	49.0	?	Silvennoinen et. al., 2014
OUL	65.1	25.9	42.0	0	Silvennoinen et. al., 2014
SGF	67.4	26.5	48.0	0	Silvennoinen et. al., 2014
MSF	65.9	28.9	58.0	0	Silvennoinen et. al., 2014
RNF	66.6	26.0	45.0	0	Silvennoinen et. al., 2014
HEF	68.4	23.7	44.0	1	Silvennoinen et. al., 2014
VRF	67.8	29.6	43.0	1	Silvennoinen et. al., 2014
KEV	69.8	27.0	57.0	1	Silvennoinen et. al., 2014
KIF	69.0	20.8	46.0	1	Silvennoinen et. al., 2014
KU6	66.0	29.9	56.0	1	Silvennoinen et. al., 2014
ARE0	69.5	25.5	48.0	?	Silvennoinen et. al., 2014
SAL	67.4	18.5	46.0	1	Silvennoinen et. al., 2014
PAJ	67.0	23.1	46.0	1	Silvennoinen et. al., 2014
MAS	67.5	22.0	47.0	1	Silvennoinen et. al., 2014
ERT	66.6	22.2	44.0	1	Silvennoinen et. al., 2014
HAR	66.2	21.0	44.0	1	Silvennoinen et. al., 2014
HH	67.9	25.8	-	0	Новые данные
HP	67.6	26.5	-	0	Новые данные
HR	66.2	29.1	-	0	Новые данные
HQ	65.9	29.1	-	0	Новые данные
HS	65.2	28.4	-	1	Новые данные
HK	64.7	30.6	-	1	Новые данные
F5	67.0	25.1	-	0	Новые данные
F6	67.1	24.8	-	1	Новые данные
HT	66.8	27.4	-	?	Новые данные

Приложение 2. Структура базы данных прогнозной системы

На рис. 3.4 показана структура данных, описание таблиц и полей представлено ниже (СУБД MS Access 2010).

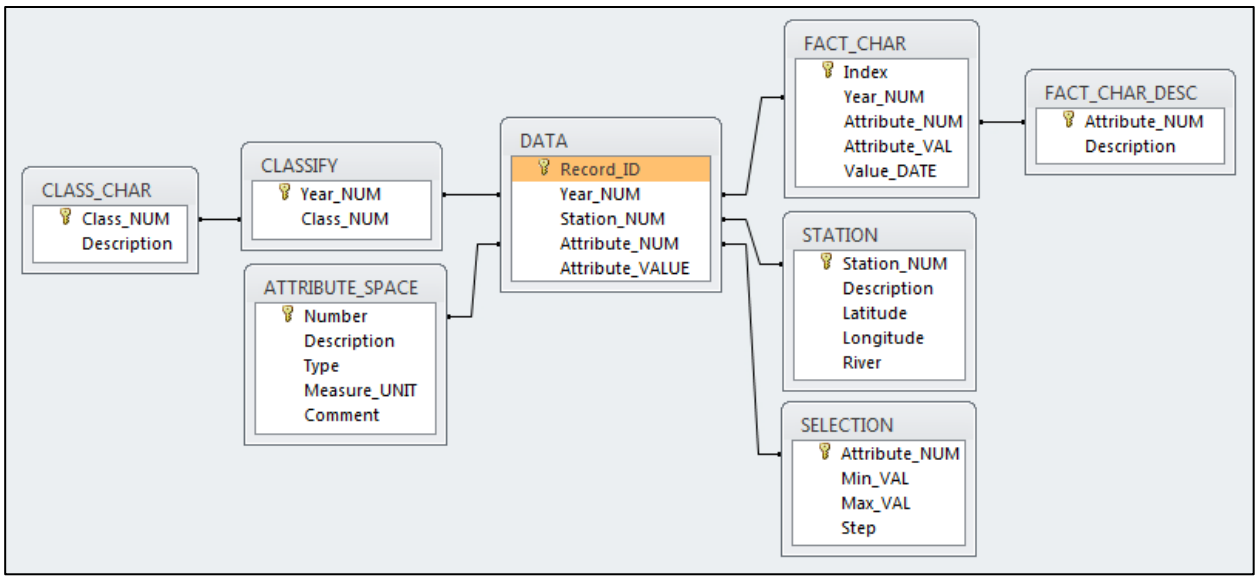

Рис. 3.4. Структура данных и связи в базе исходных данных

ATTRIBUTE_SPACE

Сформированное признаковое пространство – набор гидрологических и метеорологических параметров, влияющих на итоговый результат явления, т.е. на образование и мощность ледовых заторов в известных местах с высокой повторяемостью явления. Таблица БД содержит поля с названием, типом и единицей измерения признаков.

SELECTION

В таблице хранятся данные о минимальном и максимальном числовых значениях признаков и величине шага проведения испытаний методом Монте-Карло.

CLASSIFY

Таблица содержит экспертную классификацию явления по годам. В полях хранится информация о годе и номере класса. Классы являются результатом экспертной классификации объектов наблюдения (сезонов) по критерию мощности заторов.

CLASS_CHAR

Описание классов экспертной классификации.

STATION

Данные о гидрологических и метеорологических постах – точках наблюдения: название поста, название реки и географические координаты.

FACT_CHAR

История характеристик прогнозируемого явления. Например, максимальный заторный уровень воды.

FACT_CHAR_DESC

Описание характеристик прогнозируемого явления.

DATA

Основная таблица БД. Для каждого номера признака содержит числовые значения признака по временному измерению – год наблюдения и по пространственному – точка наблюдения.

Приложение 3. Визуализация факторов прогнозной системы

Современные ГИС содержат алгоритмы, позволяющие в автоматизированном режиме выделять речные бассейны. Исходными данными в этом случае являются цифровая модель рельефа и точка замыкающего створа. Последовательность процедур построения речного бассейна в среде ESRI ArcGIS 10.2.2 стандартными средствами геообработки (Spatial Analyst Tools – Hydrology и Conversion Tools) представлена на Рис 3.6.

Алгоритм построения бассейна содержит следующие шаги:

1. Загрузка исходных данных

   Использована цифровая модель рельефа GTOPO 30 с пространственным разрешением ~1км. [100, 101].
2. Заполнение пустых ячеек растра

   Исходная ЦМР содержит ряд неточностей и ошибок, поэтому необходимо провести коррекцию с использованием инструмента Fill.
3. Построение растра направления стока

   С помощью инструмента Flow Direction определяется направление стока из каждой ячейки растра. Опционально создается растр понижения (Output drop raster), показывающий отношение максимального изменения по высоте из каждой ячейки вдоль направления стока к расстоянию между центрами ячеек, выраженное в процентах [102].
4. Выделение водосбора

   С использованием инструмента Watershed, который в качестве входных данных принимает растр направления стока, построенный на предыдущем шаге, и точку замыкающего створа (Mouth), определяется водосборная область.
5. Построение итогового слоя

   Результат предыдущего шага преобразуется инструментом Raster to Polygon в shape-файл.

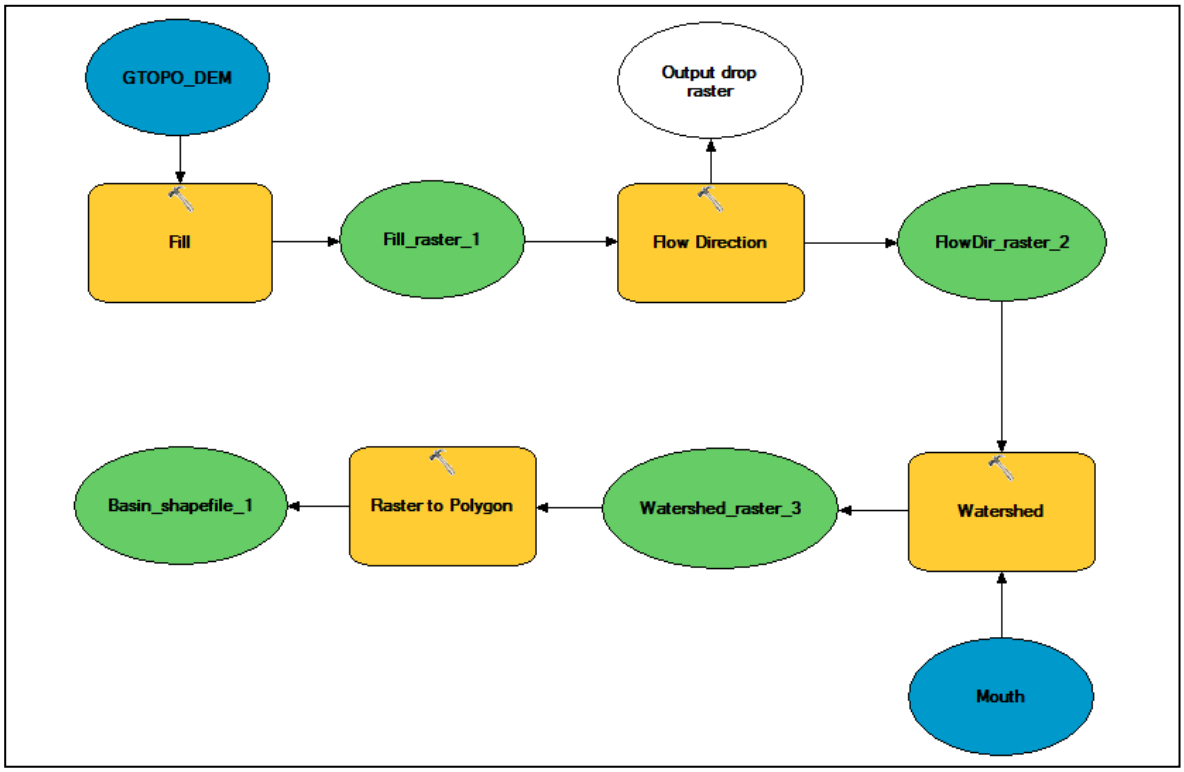

Рис. 3.6. Окно инструмента Model программного пакета ArcGis 10.2. Построение речного бассейна.

Исходные данные для прогнозной системы содержатся в базе геоданных Forecast_database (рис. 3.7). В блоке SD_Basin содержатся векторные слои исходных пространственных данных, полученные в результате построения речного бассейна, и координаты метеостанций и гидрологических постов, которые взяты по данным [ВНИИ гидрометеорологической информации]. Из исходного массива точек на территорию РФ был выделен массив на территорию бассейна р. Северная Двина при помощи инструмента Clip по векторному контуру бассейна – слои Hydro_stations и Meteo_stations. Итоговый результат визуализации изображен на рис. 3.8. В блоке Attributes содержатся координаты точек гидропостов и не интерполированные исходные

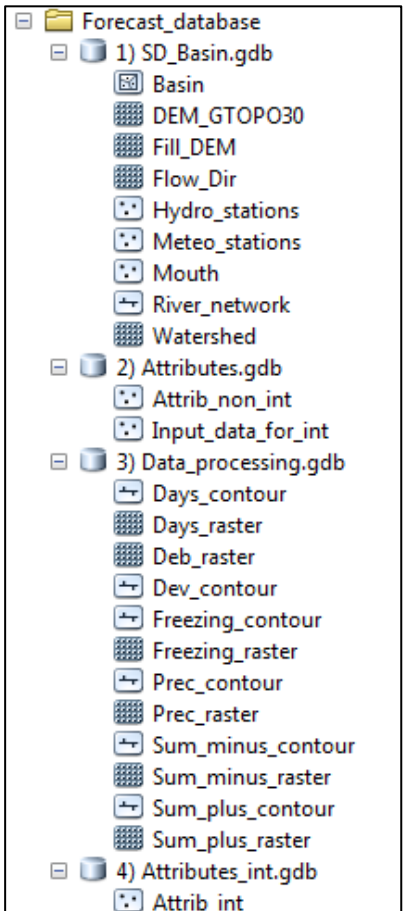

Рис. 3.7. База геоданных

данные для прогнозной системы – слой Attrib_non_int (соответствует таблицам типа табл. 3.2). Слой Input_data_for_int содержит метеорологические данные для интерполяции по 10 метеостанциям, расположенным в бассейне р. Северная Двина. В блоке Forecast_maps содержатся интерполированные растры и слои изолиний по каждому метеорологическому признаку. В блоке Attributes_int содержится слой со значениями интерполированного растра в точках гидропостов. Атрибутивная таблица соответствует таблице типа табл 3.3.

По фактическим значениям признаков был использован алгоритм интерполяции кригингом. Общая схема реализации в среде ESRI ArcGIS 10.2.2 представлена на Рис. 3.9. Данная процедура повторяется для всех метеорологических признаков.

Алгоритм состоит из следующей последовательности действий:

1. Загрузка точечного shape-файла, содержащего значения интерполируемых признаков в точках наблюдений.
2. Построение интерполированного растра алгоритмом Kriging.
3. Опционально создается дополнительный растр Out_variance_prediction_raster.
4. Из полученного интерполированного растра при помощи инструмента Clip вырезается нужная область по контуру бассейна р. Северная Двина.
5. При помощи инструмента Contour проводились изолинии с заданным шагом.
6. На вход инструменту Extract Values to Points подавался набор точек для извлечения интерполированных значений. В качестве точек использовались гидропосты (слой Attrib_non_int).

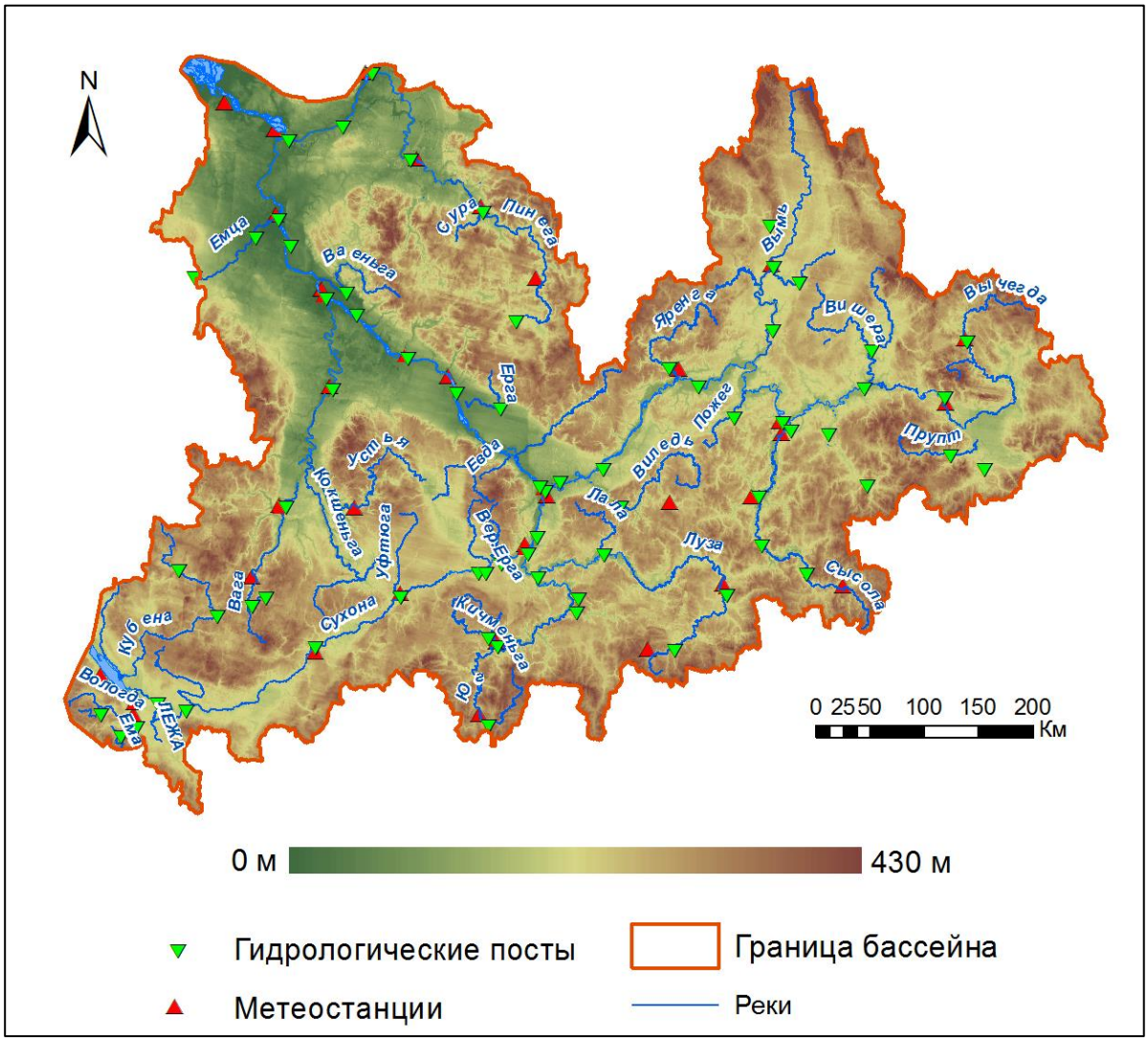

Рис. 3.8. Бассейн р. Северная Двина

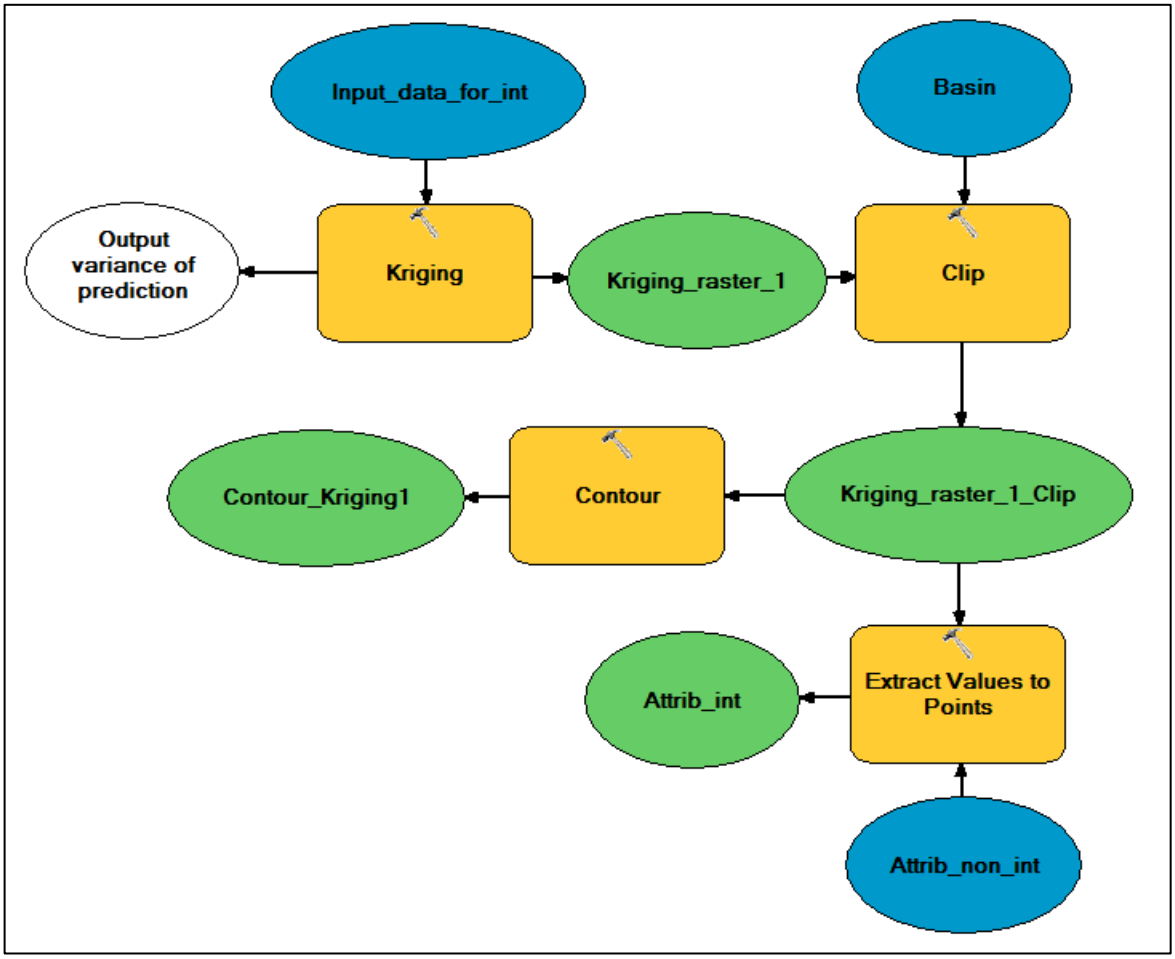

Рис. 3.9. Окно инструмента Model программного пакета ArcGis 10.2. Схема предобработки данных.

Например, для табл. 3.2 интерполяция метеорологических признаков по описанной схеме представлена на рис. 3.10 – Рис. 3.15. и дает следующий результат (табл. 3.10).

Табл. 3.10. Предобработка исходных данных за 1991г.

1991 год Пост	Номер признака										
	1 м	2 сутки	3 да/нет	4 сутки	5 C^0	6 C^0	7 сутки	8 мм	9 сутки	10 см	11 см/сутки
Каликино	337	13	0	60	-1421	4.4	7	185.2	34	41	76
Вел. Устюг	121	10	0	59	-1449	3.9	6	192.4	34	57	76
Медведки	135	5	0	59	-1449	3.8	6	190.4	34	67	42
Котлас	214	22	0	59	-1446	3.9	6	179.3	34	50	80
Абрамково	112	23	0	62	-1513	3.5	5	165.1	34	64	59
Подосиновец	94	1	0	56	-1421	3.0	7	199.8	34	48	61

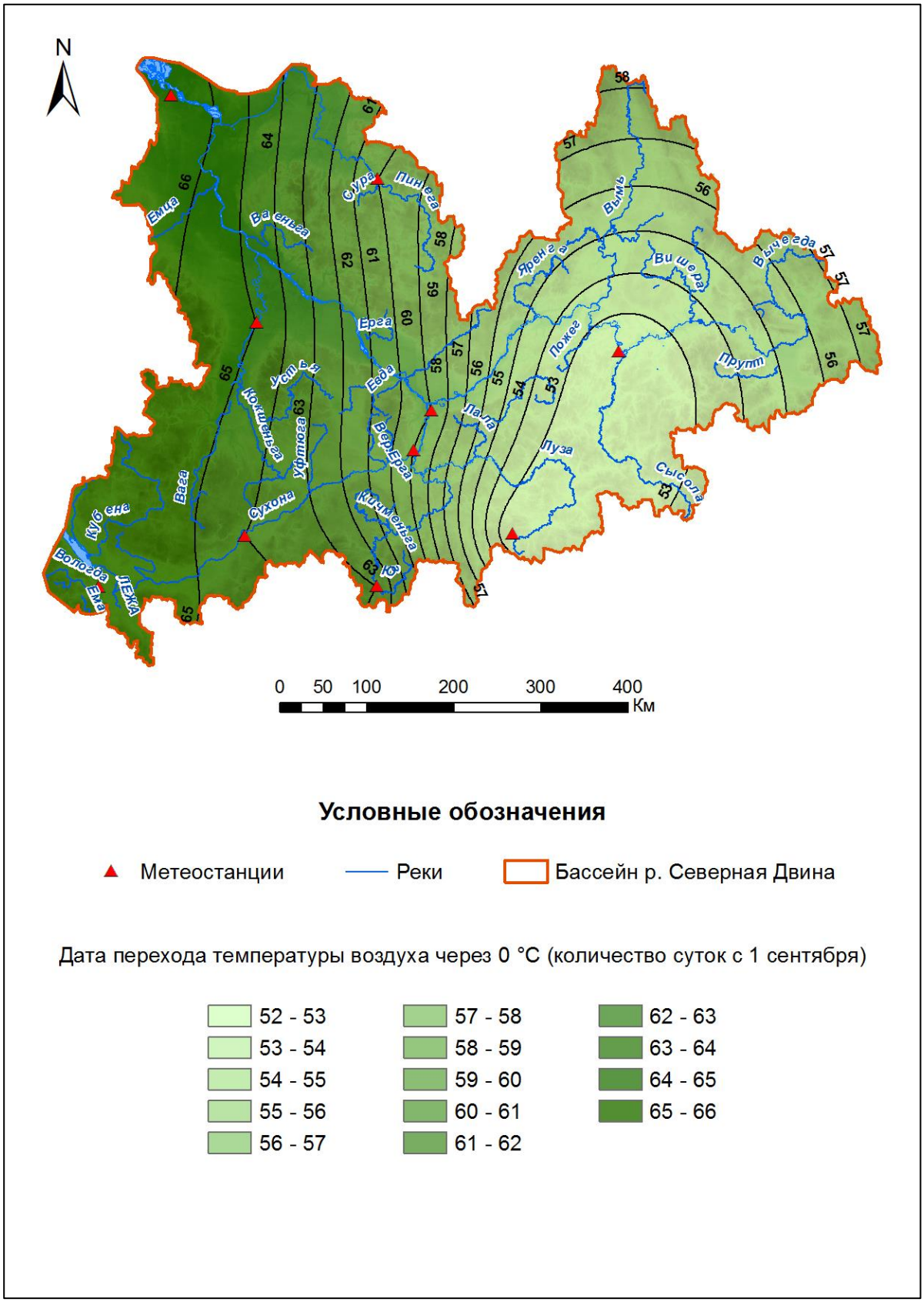

Рис. 3.10. Признак №4 «Особенности температурного режима в период замерзания»[4]

[4] Комментарий к легенде: здесь и далее каждый класс принимает числовое значение на интервале (a;b], где a и b нижняя и верхняя границы каждого класса соответственно.

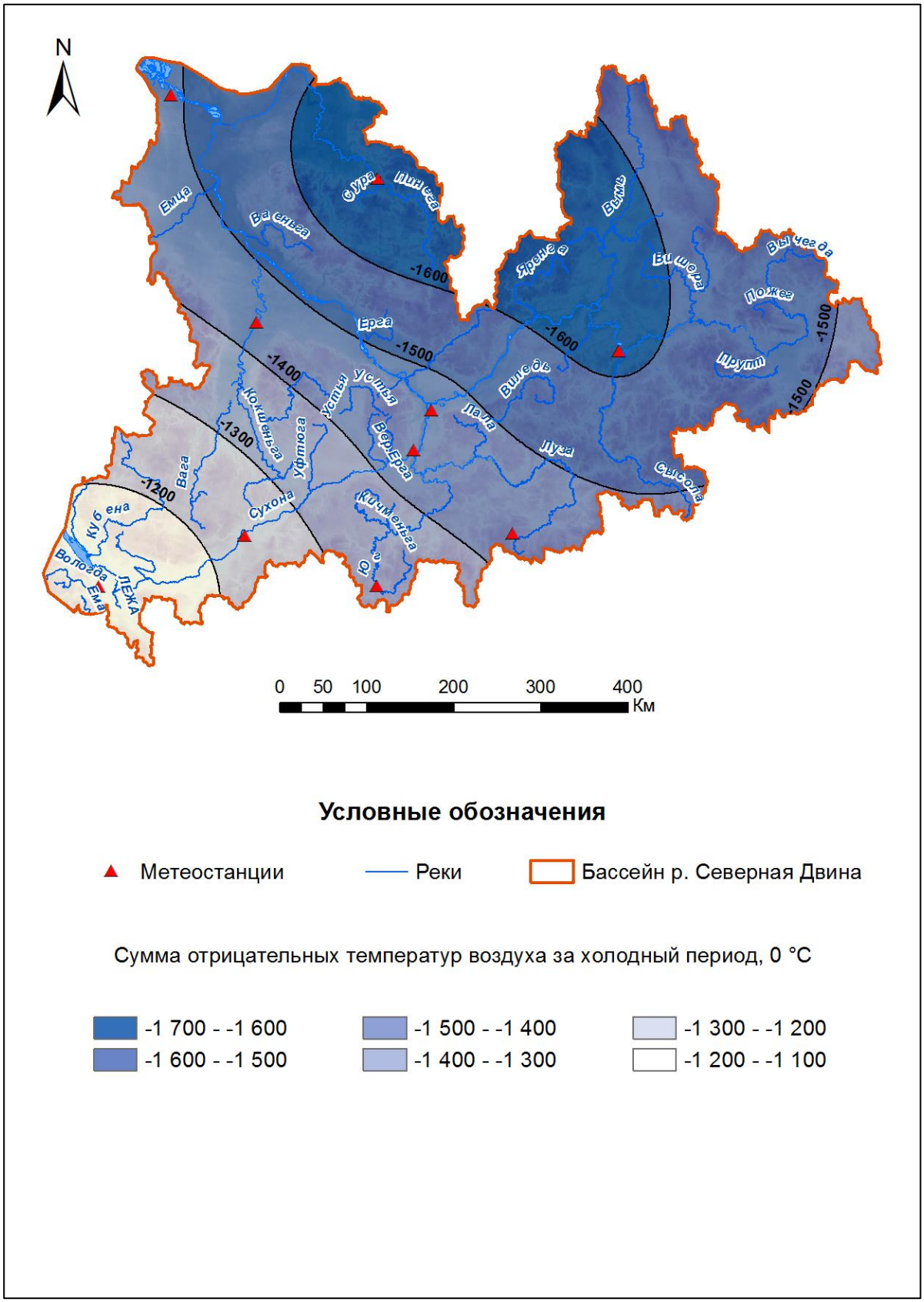

Рис. 3.11. Признак №5 «Сумма отрицательных температур воздуха за холодный период»

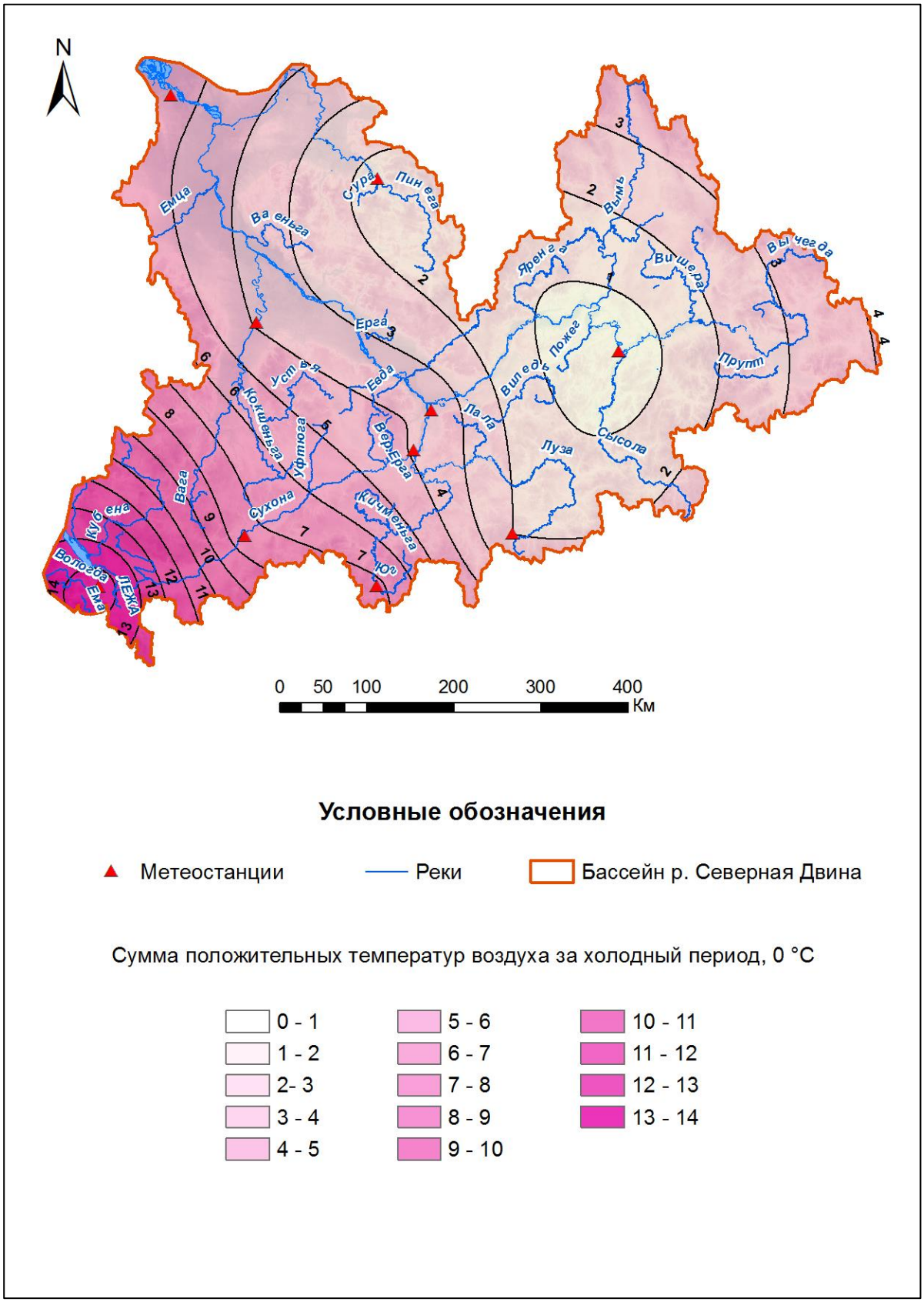

Рис. 3.12. Признак №6 «Сумма положительных температур воздуха за холодный период»

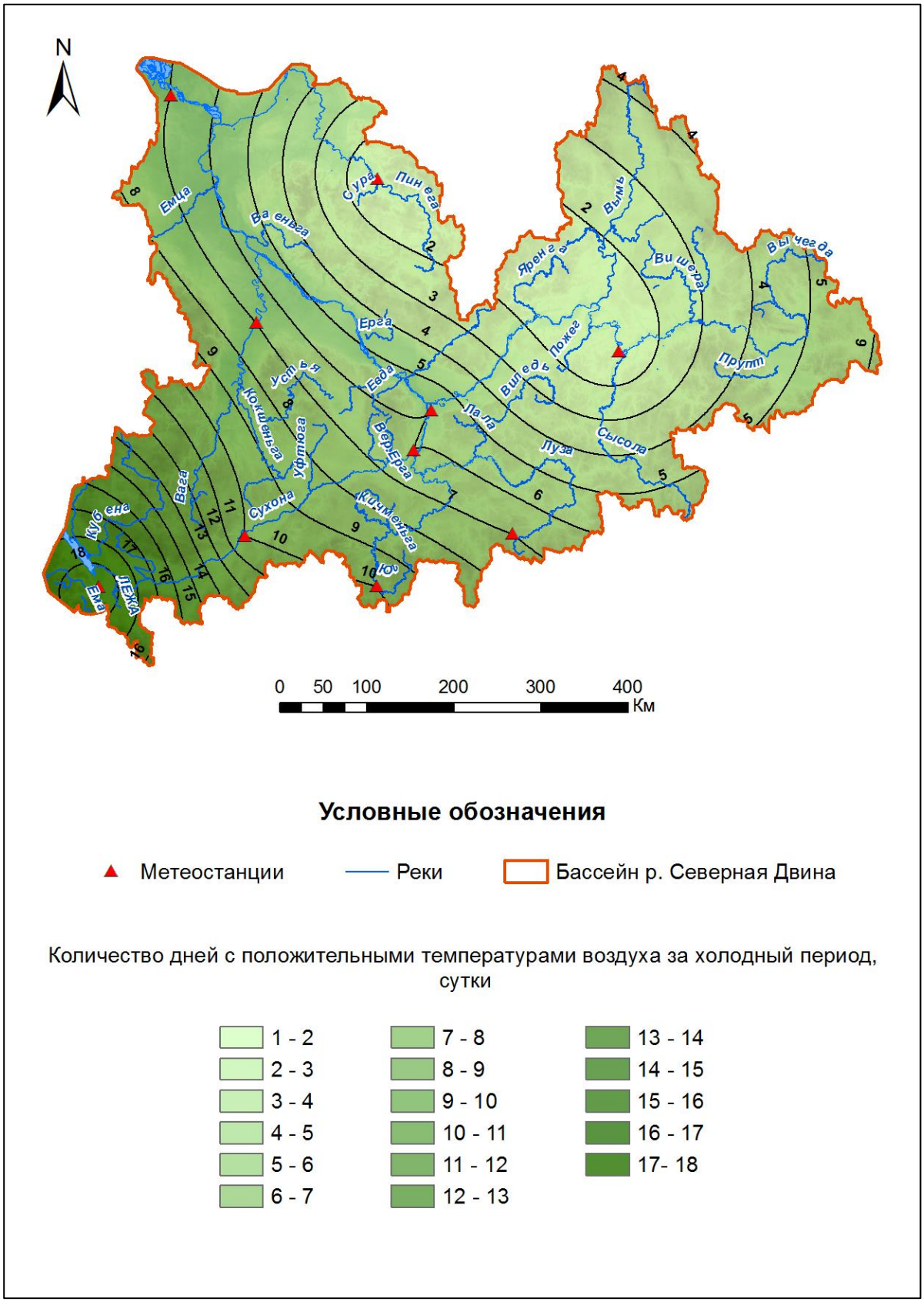

Рис. 3.13. Признак № 7 «Количество дней с положительными температурами воздуха за холодный период»

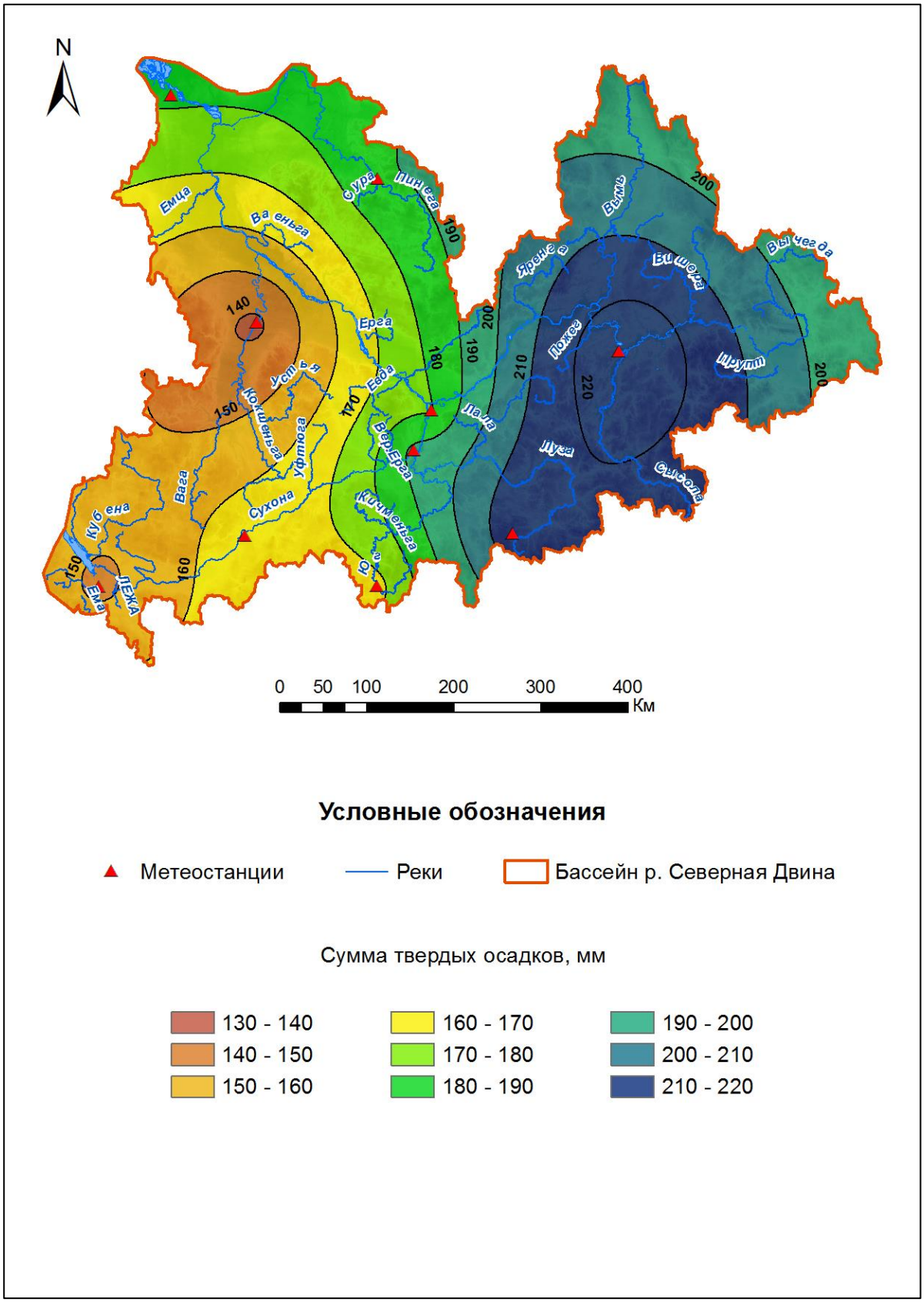

Рис. 3.14. Признак № 7 «Сумма твердых осадков»

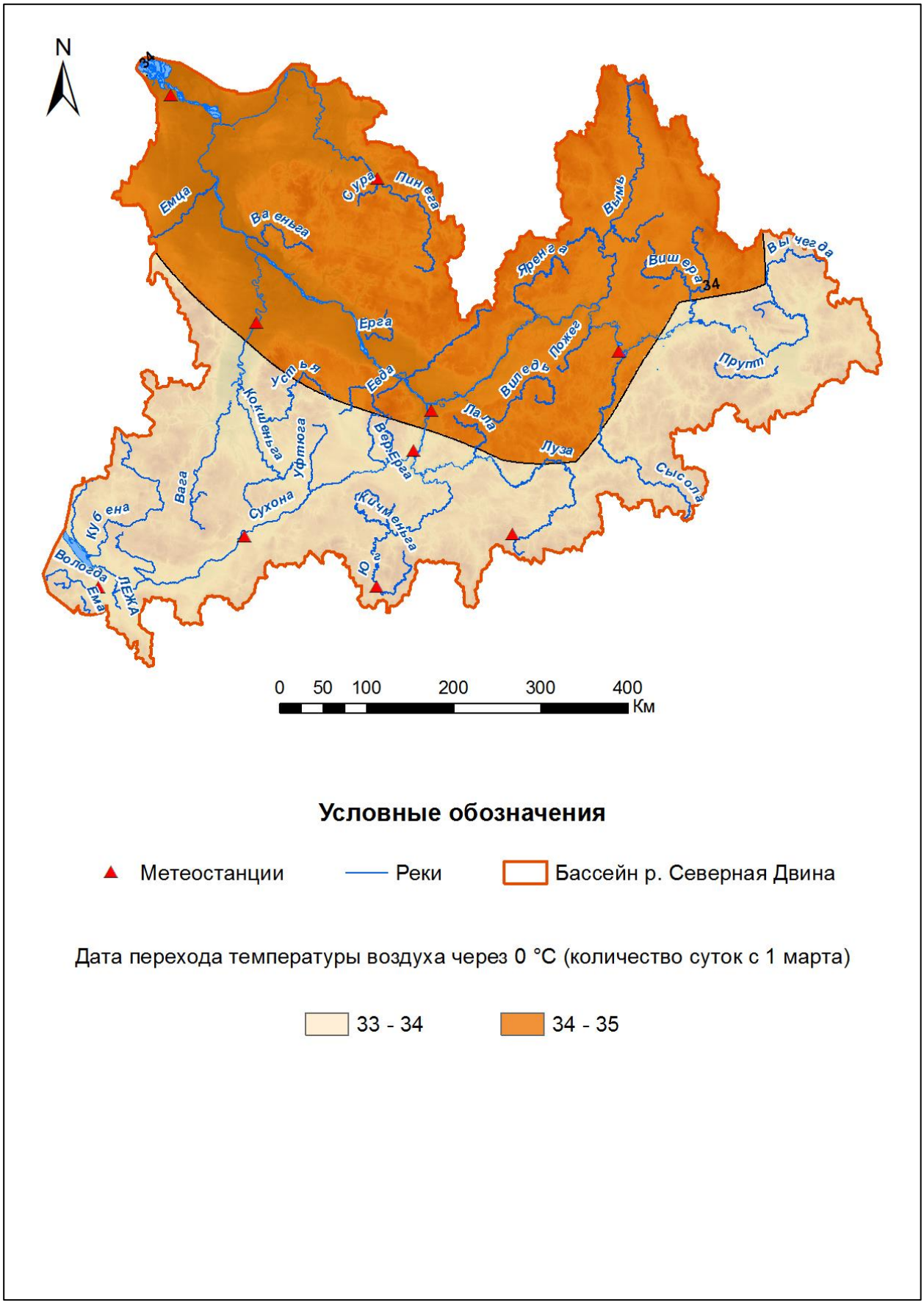

Рис. 3.15. Признак № 9 «Особенности температурного режима в период вскрытия»

Важно заметить, что приведенный в Приложении алгоритм используется исключительно в целях визуализации данных. Для обучения прогнозной системы и составления предсказаний используются неинтерполированные данные метеостанции в г. Великий Устюг, поскольку прогноз носит локальный характер.