На правах рукописи

Малыгин Иван Вячеславович

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ В ГЕОФИЗИЧЕСКИХ ЗАДАЧАХ С ДЕФИЦИТОМ ДАННЫХ

Специальность 25.00.10 Геофизика,

геофизические методы поисков полезных ископаемых

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Москва – 2021

Работа выполнена в лаборатории геоинформатики Федерального государственного бюджетного учреждения науки Института физики Земли им. О.Ю. Шмидта Российской академии наук.

Научный руководитель:	**Алешин Игорь Михайлович** кандидат физико-математических наук, Федеральное государственное бюджетное учреждение науки Институт физики Земли им. О.Ю. Шмидта Российской академии наук, заведующий лабораторией геоинформатики
Официальные оппоненты:	
Ведущая организация:	

Защита диссертации состоится дд.мм.2021 г. в хх:хх часов на заседании диссертационного совета Д.002.001.01, созданном на базе Федерального государственного бюджетного учреждения науки Института физики Земли им. О.Ю. Шмидта Российской академии наук, по адресу 123242, г. Москва, ул. Большая Грузинская, д. 10, стр. 1, конференц-зал.

С диссертацией можно ознакомиться в библиотеке ИФЗ РАН и на сайте института www.ifz.ru. Автореферат размещен на официальном сайте Высшей аттестационной комиссии при министерстве образования и науки Российской Федерации www.vak.minobrnauki.gov.ru и на сайте ИФЗ РАН.

Отзывы на автореферат, заверенные печатью, в двух экземплярах, просьба направлять по адресу: 123242, Москва, Большая Грузинская ул., д. 10, стр.1, ИФЗРАН, ученому секретарю диссертационного совета Владимиру Анатольевичу Камзолкину.

Автореферат разослан «____» _____2021 г.

Ученый секретарь диссертационного совета,
кандидат геолого-минералогических наук В.А. Камзолкин

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы

Работа 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.

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. Одним из подходов при решении подобных задач, является группа методов, разработанных в рамках теории распознавания образов и машинного обучения. В практических исследованиях к решению задач с недостатком данных выделяются две группы методов: 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 начали разрабатываться с середины XX в., имеют теоретическое обоснование и эмпирическое подтверждение. Базовые методы теории машинного обучения применены в настоящей работе для решения задачи интерполяции. Она возникает при построении двумерных и трёхмерных 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. 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. 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. Другим примером специализированных методов является классификация на основе временных рядов. Задачи с временными рядами особо актуальны при мониторинге катастрофических процессов (магнитные бури, гидрометеорологические процессы). 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 признаков на основе 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 в весенний период на реке Северная Двина. 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. Реализация логического алгоритма прогнозной системы выполнена на языке программирования C/C++. Составление карт и визуализация данных проводились в средах GoldenSoftware Surfer 15 и Esri 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. Метод ближайших соседей позволил лучше определять нелинейные зависимости в пространственном распределении геофизических величин, лучше выделять области пространственной неоднородности.

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.

Показано, что влияние пространственной анизотропии распределения данных можно исключить, если модифицировать пространственную метрику, определяющую расстояние между данными. Использованный подход позволяет получить контрастное изображение неоднородных областей, что позволяет выделить неоднородности, геометрические размеры которых меньше расстояния между скважинами.

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

Предложен оригинальный подход для анализа временных рядов ограниченной длины на основе специализированных методов теории распознавания образов. Разработанный метод создания прогнозной системы сочетает используемые на практике принципы классификации явлений с 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, оценена достоверность результатов.

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

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 р. 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, 2018, 2017) – Москва; VI Международная научно-практическая конференция «Индикация состояния окружающей среды: теория, практика, образование» – Москва, 2018; IV Школа-семинар «Гординские чтения» – Москва, 2017; Международный молодежный научный форум «ЛОМОНОСОВ» (2015, 2014) – Москва. Автором получено свидетельство о государственной регистрации программы для ЭВМ (2014).

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 научное руководство и поддержку на всех этапах проведения работы, а также коллективу лаборатории геоинформатики ИФЗ РАН.

СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность работы, сформулированы цель и задачи исследования, показаны научная новизна, практическая значимость работы, представлены основные защищаемые положения.

В соответствии с целью исследования 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.

Выводы по главе 1

1. Приведены примеры использования методов машинного обучения в различных геофизических задачах: поиск полезных ископаемых, анализ месторождений, прогнозирование опасных явлений.
2. Проведен анализ методов машинного обучения. Предложены методы решения задач интерполяции на основе алгоритма ближайших соседей и анализа временных данных для создания прогнозной системы.

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 61 постоянных и временно действующих геофизических станций. С точки зрения машинного обучения, данная задача является задачей регрессии. Для восстановления регрессионной зависимости глубины Мохоровиича от двумерных координат был использован метод k ближайших соседей с необходимой адаптацией в части

выбора сферической метрики. На исходных данных проведен подбор гиперпараметров и обучение метода ближайших соседей. Оптимальное значение числа соседей равно четырем (k=4). Средняя ошибка интерполяции при этом составляет 3.7 км. Соответствующая карта толщины земной коры приведена на Рис. 1.

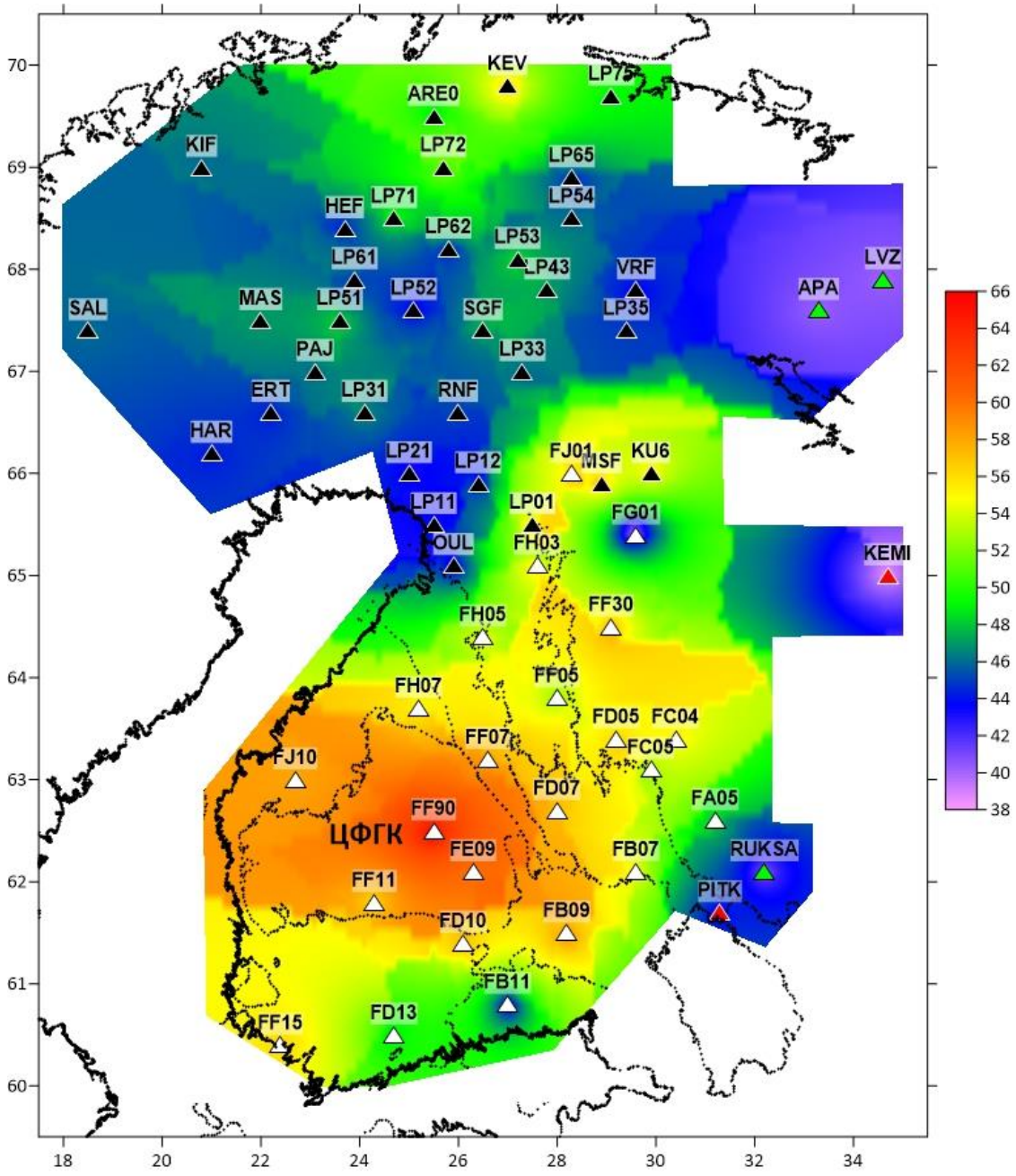

Рис. 1. Карта глубины границы Мохоровичича. Треугольниками отмечены точки, в которых положение границы было измерено. Буквенные коды означают названия сейсмических станций. Цвет треугольников определяется использованным литературным источником: черный — *Silvennoinen et. al.*, 2014 (эксперимент POLENET/LAPNET), белый — *Kozlovskaya et. al.*, 2008 (эксперимент SVEKALAPKO), зеленый — *Dricker et. al.*, 1996 и *Aleshin et. al.*, 2006, красный — данные получены в рамках настоящего исследования. Аббревиатура ЦФГК означает Центральный финский гранитоидный комплекс.

Проведено графическое сравнение ранее полученных результатов с результатами настоящей работы по четырем сейсмическим профилям. Построена глубина поверхности Мохоровичича вдоль соответствующих профилей. Взято сечение поверхности, построенной методом *kNN*, и построена зависимость глубины Мохоровичича от расстояния вдоль этой плоскости. Результаты графически наложены на иллюстрации из публикаций красной сплошной линией, красными пунктирными линиями показана оцененная ошибка интерполяции (рис. 2-5).

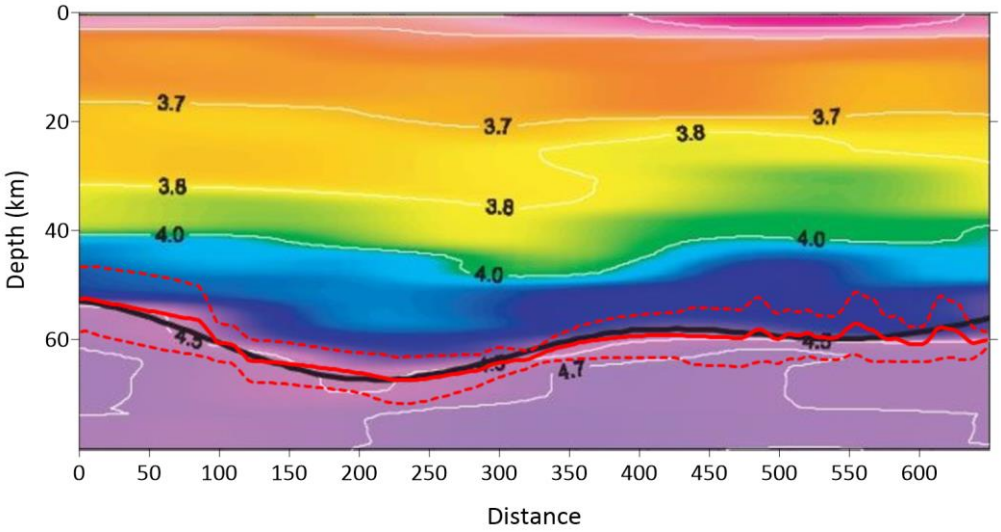

Рис. 2. Профиль FD13-FJ01. На профиль границы Мохо из работы [Kozlovskaya et al., 2008] (черная линия) нанесена граница Мохо, построенная методом *kNN* в рамках настоящего исследования (красная линия).

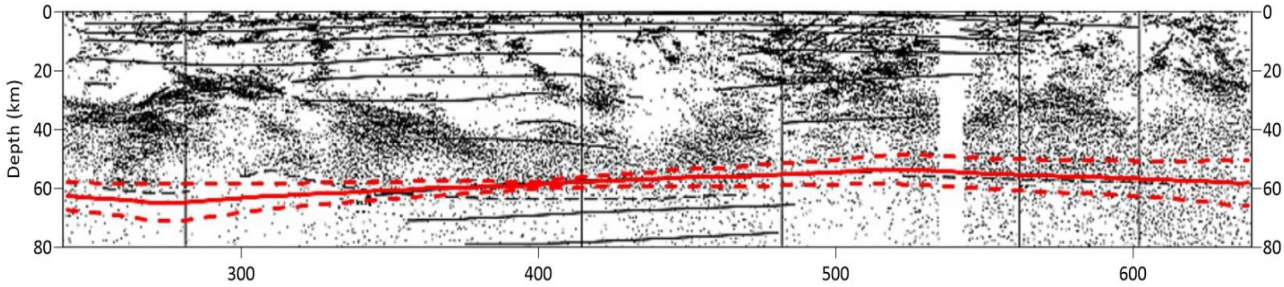

Рис. 3. Профиль FF30-A (approx. FIRE 1). На профиль границы Мохо из работы [Janik et al., 2007], черная пунктирная линяя – метод ГСЗ, черная сплошная линяя – метод ОГТ, нанесена граница Мохо, построенная методом *kNN* в рамках настоящего исследования – красная линия.

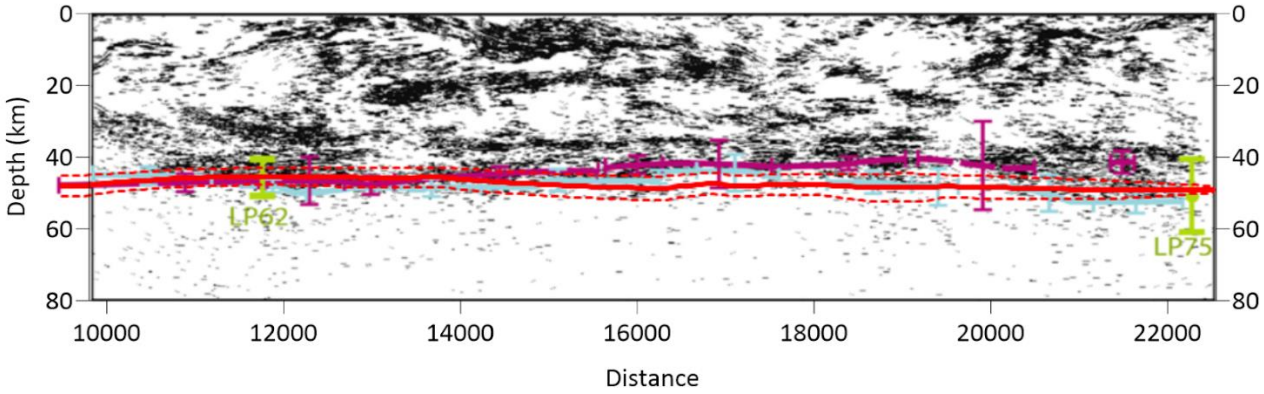

Рис. 4. Глубина Мохо вдоль профиля POLAR (LP51–LP75). На рис. 6 из статьи [Silvennoinen et. al., 2014] красной линией наложена глубина Мохо, полученная в рамках настоящего исследования методом *kNN*.

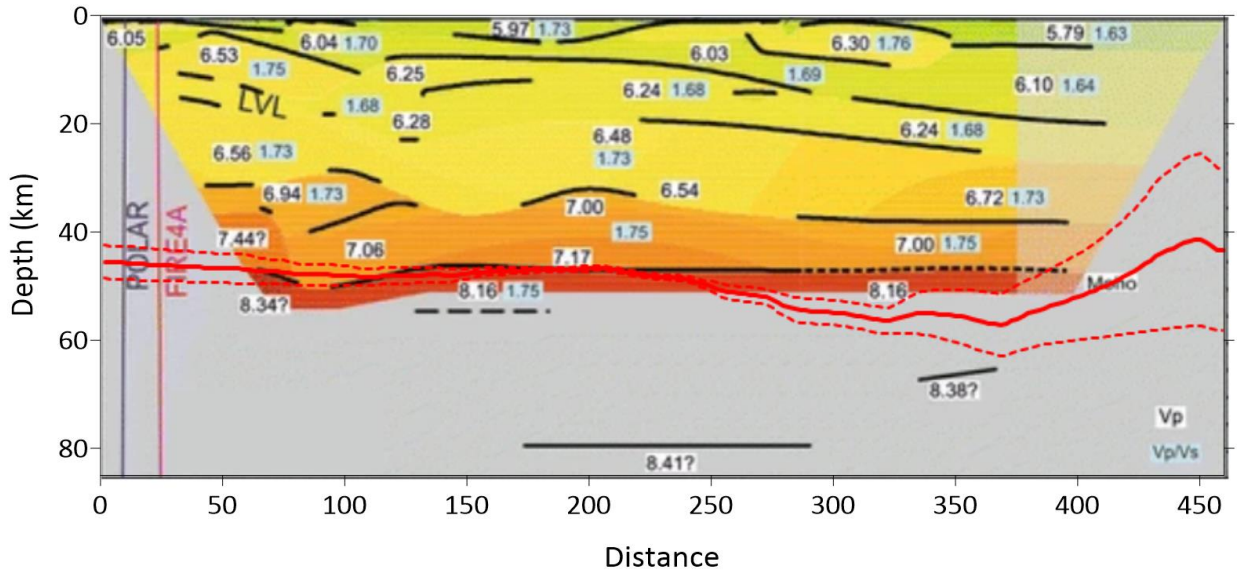

Рис. 5. Глубина Мохо вдоль профиля HUKKA2007 (approx. LP62–FG01). На рис. 11 из статьи [Tiira et. al., 2014] красной линией наложена глубина Мохо, полученная в рамках настоящего исследования методом *kNN*.

Построенные профили хорошо согласуются с полученными ранее результатами. По всей длине профилей результаты интерпретации методом *kNN* попадают в границы доверительных интервалов предыдущих построений.

Результатом решения данной задачи также 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 60 сейсмических станций известно наличие или отсутствие слоя низких скоростей. С помощью метода k ближайших соседей в каждой точке исследуемого региона оценена вероятность наличия слоя низких скоростей. На исходных данных проведен подбор гиперпараметров и обучение метода ближайших соседей. Оптимальное значение числа соседей равно, как и в предыдущей задаче, четырем (k=4). Карта слоя пониженных скоростей приведена на Рис. 6.

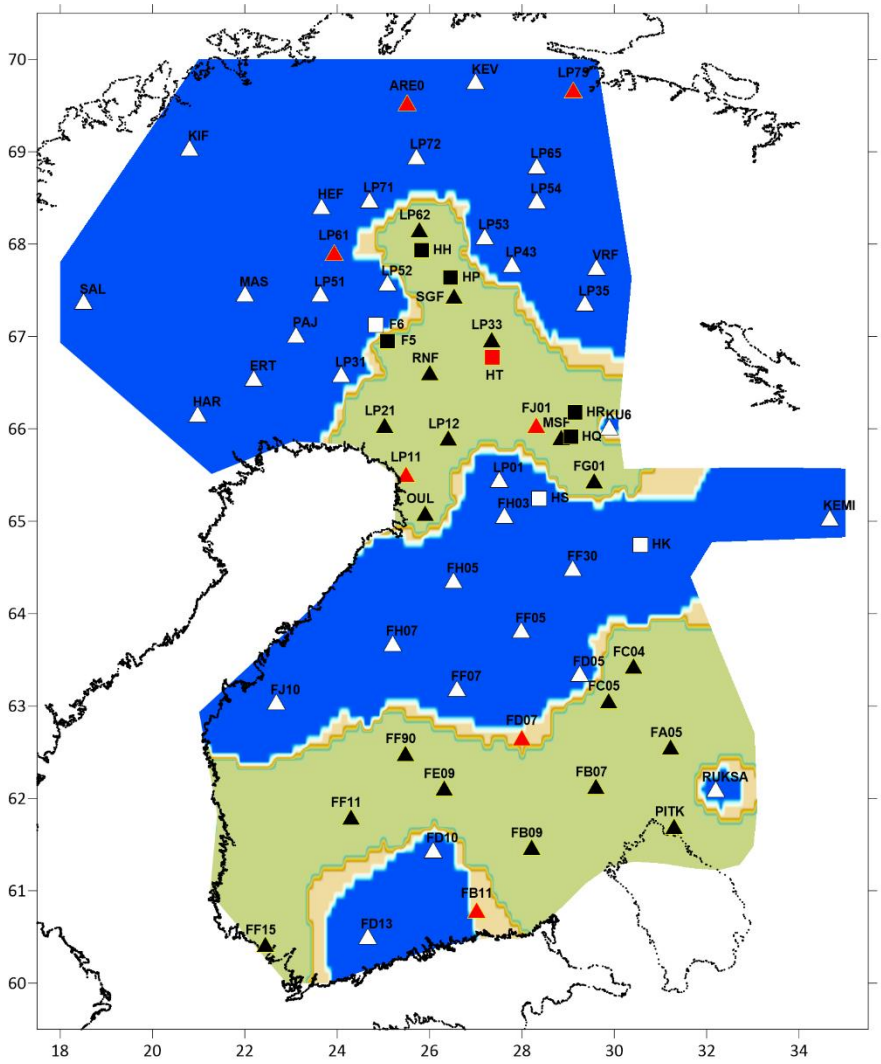

Рис. 6. Карта поверхностного слоя низкой скорости поперечных сейсмических волн, полученная в результате расчета по исходным данным. Синим цветом отображены области, в которых вероятность наличия слоя превышает 0.55, оливковым цветом окрашены области, в которых слой отсутствует (вероятность меньше 0.45). Промежуточным значениям вероятности соответствуют области цвета охра.

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 (Рис. 7.).

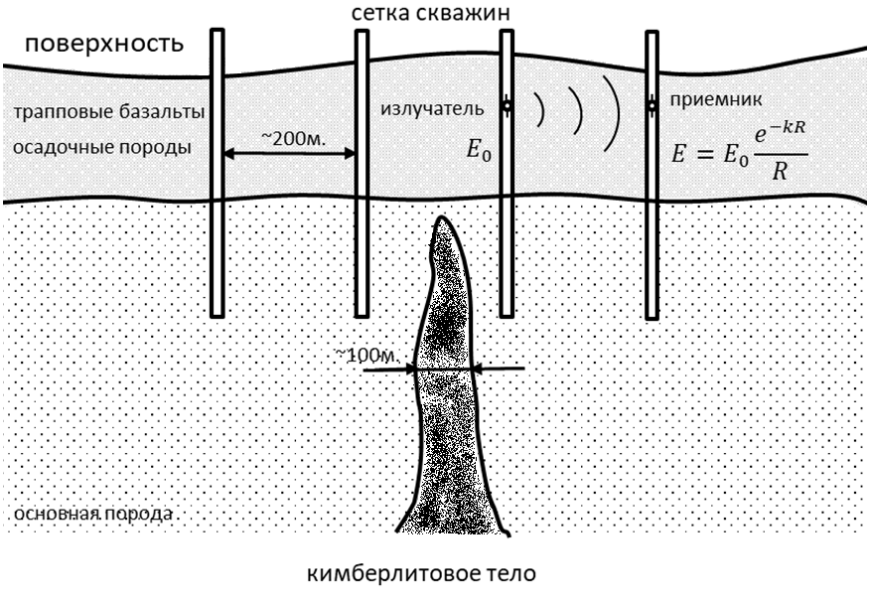

Рис. 7. Схема процесса радиоволнового просвечивания

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. Особенность пространственного распределения данных приводит к тому, что практически для всех точек пространства ближайшими оказываются данные, относящиеся к одной группе измерений. Масштабирование горизонтальных осей позволяет исправить эту ситуацию (Рис 8.).

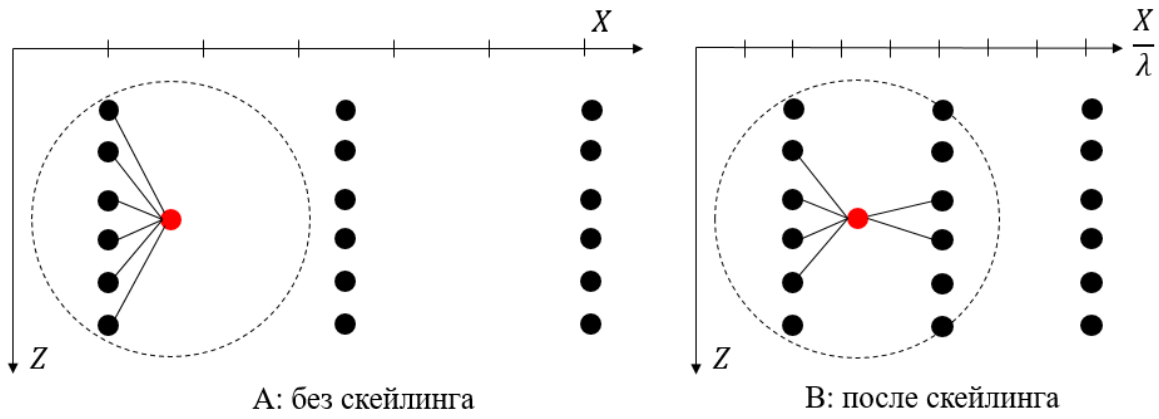

Рис. 8. Влияние масштабного коэффициента λ на распределение ближайших точек

Для определения гиперпараметров (масштабного множителя и числа ближайших соседей) использован метод кросс-валидации. Исходные данные разбиваются на M=5 групп. 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. Коэффициент детерминации был рассчитан на сетке $1 \leq k \leq 25,\ 1 \leq \lambda \leq 25$ с единичным шагом по каждому из параметров. Результат расчетов приведен на Рис. 9.

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.

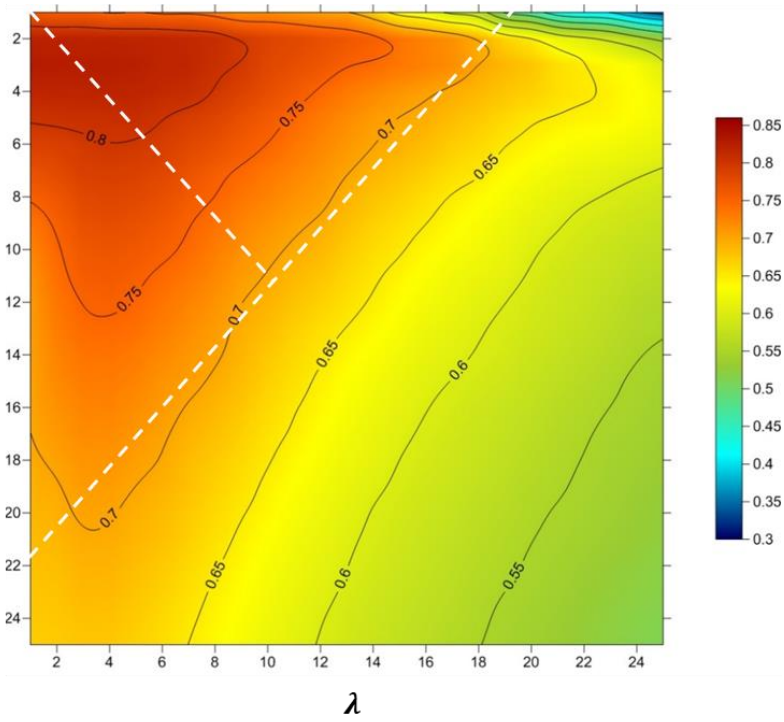

Рис. 9. Значения коэффициента детерминации $R^2(k, \lambda)$, рассчитанные на сетке параметров.

После того, как значения гиперпараметров определены, задача построения образа состоит в определении интересующей величины – коэффициента затухания методом kNN с модифицированной метрикой в узлах трехмерной решетки. На Рис. 10 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 Рис. 10 приведены также короткие вертикальные сечения, соответствующие этим линиям, на которых соответствующие области диаметром менее 100 м. также отчетливо видны.

Использованный подход позволяет получить достаточно контрастное изображение неоднородных областей, что позволяет выделить объекты, чьи геометрические размеры 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.

Выводы по главе 2

1. Построена уточненная карта границы Мохоровичича для региона Фенноскандия с использованием метода k ближайших соседей с необходимой адаптацией в части выбора сферической метрики. Построенный профиль хорошо согласуется с полученными ранее результатами.
2. 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.
3. 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. Влияние анизотропии распределения данных можно исключить, если модифицировать пространственную метрику, определяющую расстояние между данными, введением коэффициента скейлинга. Получено контрастное изображение неоднородных областей, позволяющее выделить неоднородности, геометрические размеры которых меньше расстояния между скважинами.

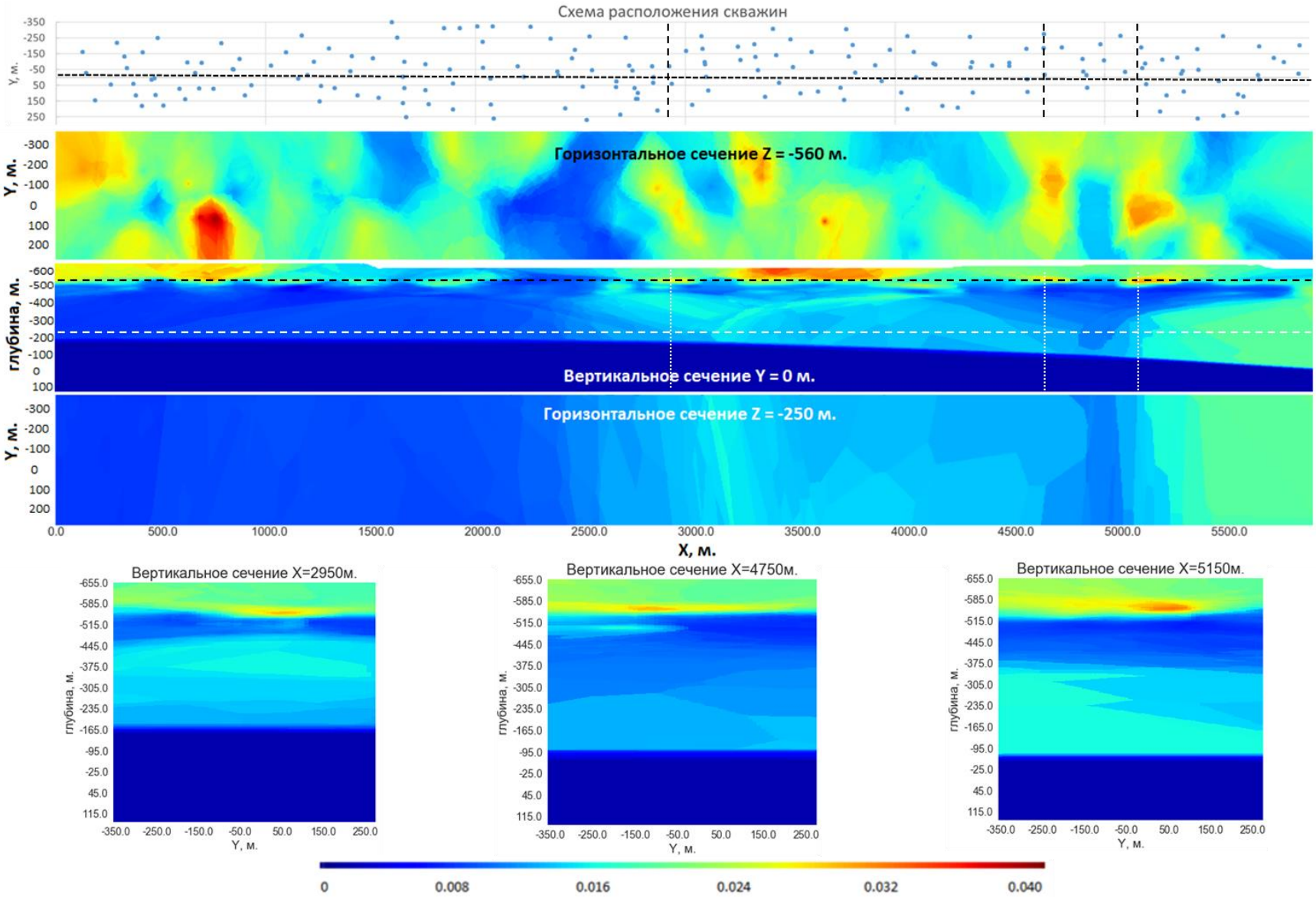

Рис. 10. Результаты 3D-моделирования

В **третьей главе** рассматривается задача построения прогнозной системы на основе коротких рядов наблюдений для прогноза ледовой обстановки района с малой территорией и с коротким периодом наблюдения, на примере участка реки Северная Двина от г.Котлас до г.Великий Устюг. Возможным подходом к разрешению проблемы заторообразования является ее исследование в рамках построения прогнозных систем, основанных на алгоритмах теории распознавания образов и теории машинного обучения.

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). Разработанная 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. Общий список этих признаков представлен в Табл. 1.

Табл. 1. Признаковое пространство

№	**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 Дата 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 Дата 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

На Рис. 11 представлена схема района наблюдения с расположением речных постов: Каликино, Великий Устюг, Медведки, Котлас, Абрамково, Подосиновец.

Под классификацией понимается 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. 2.). Ситуация недостатка данных ограничивает возможность разделения на большее число классов, соответствующих более подробной классификации исследуемого явления.

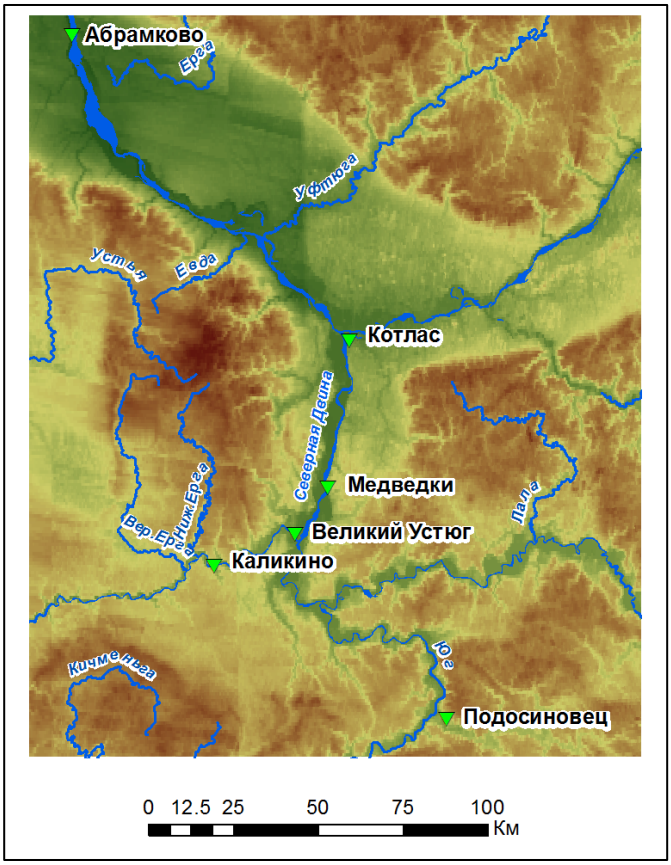

Рис. 11. Схема расположения речных постов на р. Северная Двина и ее притоках

Табл. 2. Классификация периода наблюдения

Сезон	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

Процедура построения решения задачи состоит из двух этапов. На первом этапе производится обучение системы – определение наборов пороговых значений признаков, обеспечивающих наилучшую классификацию. Для этого используется обучающая выборка – набор объектов, для которых результирующее состояние известно. Затем составляется решающее правило, на основании которого каждый новый, поступающий на вход системы, набор может быть отнесен к одному из классов 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, т.е. для прогнозирования. Оценка качества осуществляется с помощью Leave-one-Out кросс-валидации.

Первоначально, описанная система была апробирована на периоде наблюдений 1991-2010, оценка качества прогнозирования составила 85% (Табл. 3.).

Табл. 3. Оценка качества прогнозирования за период 1991-2010

Сезон	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	1	1	1	2	2	1	2	1	2	2	1	1	2	1	1	1	2	2	1	1

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 использован набор параметров, полученных ранее (при обучении на данных 1991-2010 гг.). Во втором случае для составления прогноза на добавленный сезон производилось полное переобучение системы: заново строились наборы параметров, обеспечивающих необходимое качества прогноза. Затем процесс повторялся для следующего периода. Оба подхода дали одинаковый результат, состоящий в успешном прогнозировании образования заторов для всех шести новых сезонов. Результат представлен в Табл 4.

Табл. 4. Оценка качества прогнозирования за период 2011-2016

Сезон	2011	2012	2013	2014	2015	2016
№ класса	1	2	1	1	1	1
Результат прогнозирования	1	2	1	2	2	1

Построенная система позволяет не только прогнозировать мощность опасного явления, но и является также инструментом для исследований в части определения важности признаков, влияющих на итоговый результат. Прямое сравнение признаков в сложных задачах с небольшим объемом доступных наблюдений не отражает реальной картины зависимости. Для решения подобных проблем используется естественная мера значимости признака по вкладу в итоговый результат классификации – информационный вес признака. Упорядочивание информационных весов позволяет разбить признаки на группы: ведущие, значимые, незначимые.

Проведено исследование признакового пространства задачи прогнозирования ледового заторообразования на реке Северная Двина. По результатам обучения системы произведен расчет информационных весов признаков. Итоговый результат нормирован и приведен к 100%. Ранжирование признаков по убыванию вклада в итоговый результат классификации приведено на рис. 12. Аналогичным образом исследованы зависимости признаков между собой. 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 ранее выдвинутыми гипотезами. Признаки 11, 10, 9, 4, 5, 6 вносят наибольший вклад в итоговый результат, что, в целом, соотносится с теоретическими исследованиями.

Рис. 12. Вклад признаков в итоговый результат классификации.

Выводы по главе 3

1. Разработана прогнозная система, предназначенная для краткосрочного прогноза образования заторов в весенний период на реке Северная Двина, на основе анализа коротких временных данных с применением специализированных методов теории распознавания образов.
2. Система позволяет проверять гипотезы о влиянии признаков на итоговую мощность явления в условиях ограниченного набора исходных наблюдений на гидропостах и метеостанциях. Построена иерархия используемых признаков по степени влияния результат, выделен группы ведущих и значимых признаков. Полученные результаты согласуются с ранее проведенными исследованиями.
3. Проведена оценка достоверности прогнозов на периоде разработки (20 сезонов), и валидация на расширенном периоде (26 сезонов). Оцененная достоверность прогнозирования для этих периодов согласуется и составляет 85%.

ЗАКЛЮЧЕНИЕ

В результате проведенных исследований и практических разработок **была достигнута цель диссертационного исследования** – разработаны методы решения геофизических задач с дефицитом данных, на основе теории машинного обучения.

В работе рассмотрены задачи двух типов. Построение двумерного и пространственного распределения геофизических величин по данным полевых измерений на нерегулярной сетке с использованием базовых методов теории машинного обучения. Разработка прогнозной системы по рядам коротких временных данных на основе специализированных методов комбинаторно-логического подхода теории распознавания образов.

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 приложениям.

Разработан метод анализа пространственных данных для построения двумерных и трехмерных 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. Построена карта слоя с низкими значениями скорости поперечных сейсмических волн V_S. Она 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.

Построена трехмерная модель среды при проведении межскважинных исследований. Предложена новая интерпретация данных радиоволнового просвечивания, позволяющая более точно выделить границы слоев по сравнению с методами, используемыми ранее (кригинг). Разработанный метод позволяет построить трехмерную модель проводимости среды между скважин даже при использовании синхронной схемы измерений и сильной анизотропии данных. Разработанный подход позволяет получить 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.

Разработан метод анализа многомерных временных рядов ограниченной длины для создания прогнозной системы. Разработана прогнозная система, предназначенная для краткосрочного прогноза образования заторов в весенний период на реке Северная Двина. Система позволяет оценивать влияние факторов на итоговый результат явления в условиях ограниченного набора исходных наблюдений. Оцененная достоверность прогнозирования составляет 85%.

ОСНОВНЫЕ ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в журналах, включенных в перечень российских рецензируемых научных журналов и изданий для опубликования основных научных результатов диссертации

1. 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.
2. Алешин И.М., **Малыгин И.В.** Интерпретация результатов радиоволнового просвечивания методами машинного обучения // Компьютерные исследования и моделирование. — 2019. — Т. 11, № 4. — С. 675–684.
3. 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.
4. **Малыгин И.В.** Логический подход к созданию экспертных систем прогнозирования опасных природных явлений // Естественные и технические науки. — 2015. — № 2. — С. 102–112.
5. **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.
6. **Малыгин И.В.** О задаче прогнозирования ледовых заторов // Интеллектуальные системы. Теория и приложения. — 2014. — Т. 18, № 3. — С. 73–80.
7. 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.
8. 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.

Авторские свидетельства

9. **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.
10. **Малыгин И.В.**, Алешин И.М. Свидетельство о государственной регистрации программы для ЭВМ № 2020617961 Программа расчёта и построения региональных карт геофизических свойств методом k-ближайших соседей. Дата гос. регистрации в Реестре программ для ЭВМ 15.07.2020.
11. **Малыгин И.В.**, Алешин И.М. Свидетельство о государственной регистрации программы для ЭВМ № 2020617962 Программа расчёта 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.

Прочие публикации

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