На правах рукописи

Малыгин Иван Вячеславович

АДАПТАЦИЯ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ К ГЕОФИЗИЧЕСКИМ ЗАДАЧАМ С ДЕФИЦИТОМ ДАННЫХ

Специальность 25.00.10

Геофизика, геофизические методы поисков полезных ископаемых

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Москва – 2022

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, главный научный сотрудник лаборатории геоинформатики Федерального государственного бюджетного учреждения науки Института физики Земли им. О.Ю. Шмидта РАН;

Официальные оппоненты: **Антоновская Галина Николаевна**
доктор технических наук, заместитель директора по научной работе Федерального государственного бюджетного учреждения науки Федерального исследовательского центра комплексного изучения Арктики имени академика Н.П. Лаверова Уральского отделения РАН;
Кислов Константин Викторович
кандидат физико-математических наук, старший научный сотрудник Федерального государственного бюджетного учреждения науки Института теории прогноза землетрясений и математической геофизики РАН.

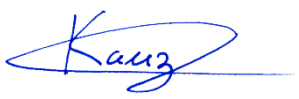
Ведущая организация: **Федеральное государственное бюджетное учреждение науки Институт динамики геосфер Российской академии наук (ИДГ РАН), г. Москва.**

Защита диссертации состоится **22 сентября 2022 г. в 14:00** 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 «____» августа 2022 г.

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

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы

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, перерывами в регистрации данных из-за технических неполадок оборудования. 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 алгоритмов, разработанных в рамках теории машинного обучения и распознавания образов. В практических исследованиях выделяются две группы методов для решения задач с пропусками 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 широкое 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. 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. 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). 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), улучшить прогнозирование мест возникновения сильных землетрясений (алгоритм Барьер). 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 возможных подходов к решению подобной задачи классификации является использование в процессе обучения интегральных характеристик, построенных по этим рядам с предварительной интерполяцией пропусков.

Цель диссертационного исследования – разработка компьютерных систем и методов обработки данных в условиях ограниченного количества данных, недостаточных для проведения классического статистического анализа, на основе методов машинного обучения, применение таких систем для построения геолого-геофизических моделей и решения задач охраны окружающей среды.

Основные задачи исследования:

1. Определение круга задач с ограниченным набором данных;
2. Разработка метода решения задач с ограниченным набором данных с применением алгоритмов машинного обучения;
3. Разработка способов определения гиперпараметров для задач каждого типа;
4. Применение разработанного метода к решению прикладных геофизических задач.

На защиту выносятся следующие положения:

1. Метод пространственно-временной интерполяции нерегулярно распределённых геофизических данных, основанный на использовании базовых методов машинного обучения, и применимый в задачах с дефицитом данных и сильной анизотропией пространственного распределения измерений.

2. Интеллектуальная система для краткосрочного прогноза образования ледовых заторов в весенний период на северных реках, основанная на анализе данных, полученных по ограниченному набору наблюдений на гидропостах и метеостанциях, с применением специализированных методов теории машинного обучения и распознавания образов.

Методика исследований

Основные результаты исследования получены с применением алгоритмов машинного обучения и алгоритмов теории распознавания образов. Реализация метода обработки данных для пространственной интерполяции выполнена на языке Python 3. Реализация логического алгоритма прогнозной системы выполнена на языке программирования C/C++ в среде Microsoft Visual Studio. 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

1. На основе алгоритмов машинного обучения разработан новый метод построения трёхмерного распределения проводимости среды по данным межскважинного электромагнитного просвечивания.
2. На основе алгоритмов машинного обучения разработан новый метод расчета трёхмерной сейсмической модели по набору одномерных скоростных разрезов.
3. В результате применения разработанного метода к данным сейсмических экспериментов SVEKALAPKO и POLENET/LAPNET построена карта Мохо центральной части Фенноскандии по S-волнам и оконтурены области с низкими приповерхностными значениями скоростей поперечных сейсмических волн. Показано, что низкоскоростной слой поперечных волн в центральной части Финляндии может быть обусловлен особенностями процесса постледниковой релаксации региона.
4. Разработанная интеллектуальная система краткосрочного прогнозирования ледового заторообразования реализована для участка р. Северная Двина с оцененной достоверностью прогнозирования 85%.
5. Получены количественные оценки влияния гидрометеорологических факторов на процесс ледового заторообразования. На основании анализа данных показано, что основное влияние оказывает группа гидрологических факторов.

Практическая значимость результата работы

Разработанный метод интерпретации геофизических данных алгоритмами машинного обучения является достаточно общим. Метод применим для решения широкого круга задач интерполяции геофизических измерений. Метод может использоваться для построения пространственных распределений измеряемых величин, например, в геомагнитных исследованиях.

Разработана интеллектуальная система для осуществления краткосрочного прогнозирования мощности процесса заторообразования для участка реки Северная Двина, что является важной частью прогноза наводнений для данной территории. Разработанная интеллектуальная система может быть применена на данных других регионов, а также для прогнозирования других опасных природных явлений с аналогичной структурой исходных данных о событиях. Реализована

функциональность, которая позволяет применять интеллектуальную систему в качестве инструмента анализа данных: проверять гипотезы относительно влияния признаков на трудноформализуемый исследуемый процесс, количественно оценивать величину вклада конкретного признака на итоговую мощность явления в условиях дефицита данных.

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: Научная конференция молодых ученых и аспирантов ИФЗ РАН (2021, 2020, 2019, 2018, 2017); Information Technologies in Earth Sciences and Applications for Geology, Mining and Economy (ITES&MP-2019); Всероссийская конференция с международным участием II Юдахинские чтения «Проблемы обеспечения экологической безопасности и устойчивое развитие арктических территорий», Архангельск, 2019; 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); IV Международная научно-практическая конференция «Научные перспективы XXI века. Достижения и перспективы нового столетия», Новосибирск, 2014.

Структура работы

Диссертация состоит из введения, трех глав, заключения, списка литературы из 168 наименований, четырех приложений. Текст диссертации изложен на 154 страницах машинописного текста, содержит 16 таблиц и 39 рисунков.

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.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность работы, сформулированы цель и задачи исследования, показаны научная новизна, практическая значимость, 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 алгоритмов 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. Сделан обзор основных методов решения задач с пропуском данных.

Построена трехмерная модель среды при проведении межскважинных исследований на основе адаптации алгоритмов машинного обучения к данным электромагнитного просвечивания, позволяющая более точно выделить границы слоев по сравнению с методами, используемыми ранее (кригинг). Такая задача возникает при поиске кимберлитовых тел прямым методом бурения, чтобы с одной стороны, минимизировать количество скважин, и, одновременно с этим, минимизировать вероятность пропуска трубки между скважинами. На Рис. 1 представлена схема электромагнитного просвечивания.

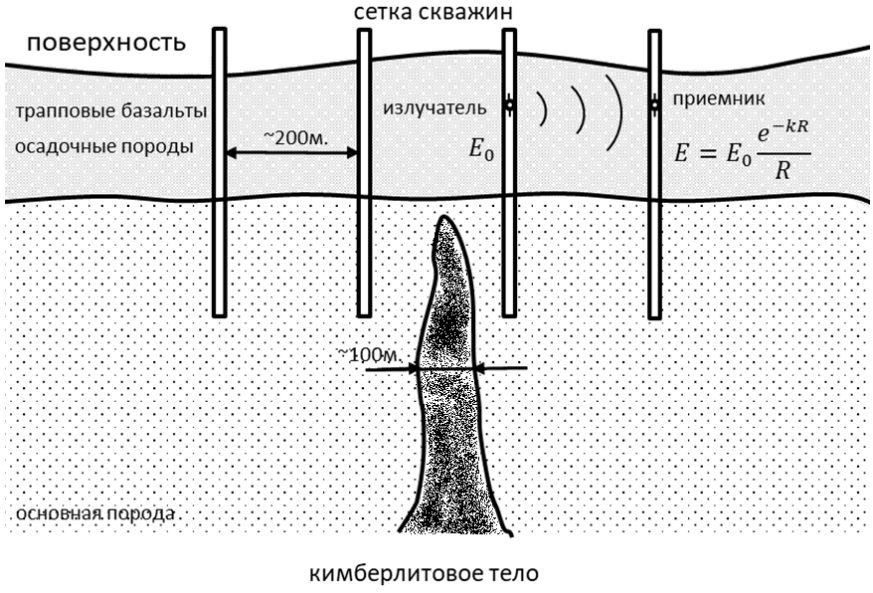


Рис. 1. Схема процесса радиоволнового просвечивания

Суть метода заключается в измерении затухания электромагнитной волны при ее прохождении через межскважинное пространство в точках, соответствующих середине отрезка между источником и приемником, синхронно погружаемых в соседние скважины. Далее строится интерполяция коэффициента затухания

электромагнитных волн в межскважинной среде и оконтуриваются области с высокими значениями коэффициента затухания.

Задача построения трехмерной модели среды сведена к постановке задачи регрессии в терминах машинного обучения: в качестве признаков использованы трехмерные координаты точки, в качестве целевой переменной коэффициент затухания электромагнитных волн, алгоритм машинного обучения – регрессия на основе k ближайших соседей (kNN), функционал качества R^2.

В данной задаче присутствует существенная особенность: исходные данные измерений группируются вдоль линейных объектов – скважин, что порождает пространственную анизотропию и дефицит равномерно распределенных данных. В предложенном способе решения этот эффект сглаживается с помощью введения коэффициента скейлинга λ так, чтобы алгоритм kNN захватывал данные с соседних групп измерений (Рис. 2). Скейлинг является вторым гиперпараметром модели, помимо k – количества ближайших соседей.

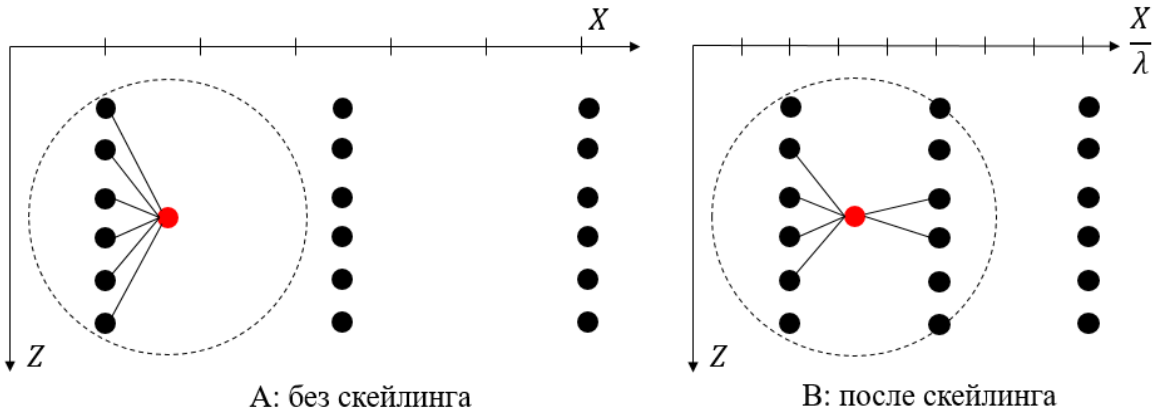


Рис. 2. Влияние масштабного множителя λ на распределение ближайших точек

Формулы (1) и (2) задают преобразование расстояний до скейлинга и после скейлинга соответственно:

$$\rho(\vec{r}_1, \vec{r}_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \quad (1)$$

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

Для определения гиперпараметров модели использован метод кросс-валидации. Исходные данные разбиваются на M=5 групп. Каждая из этих групп поочередно удаляется из процедуры обучения, и используется для проверки. Оценка качества решения производится по функционалу качества R^2 (коэффициент детерминации) – доле дисперсии целевой переменной, объясненной моделью. Коэффициент детерминации был рассчитан на сетке $1 \le k \le 25,\ 1 \le \lambda \le 25$ с единичным шагом по каждому из параметров. Результат приведен на Рис. 3.

Полученное распределение имеет вид, типичный для задач многопараметрической оптимизации. Для выбора значений гиперпараметров использован уровень значений коэффициента $R^2 = 0.7$. Выбранные значения гиперпараметров $k = 11$ и $\lambda = 10$ соответствуют пересечению медианы

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

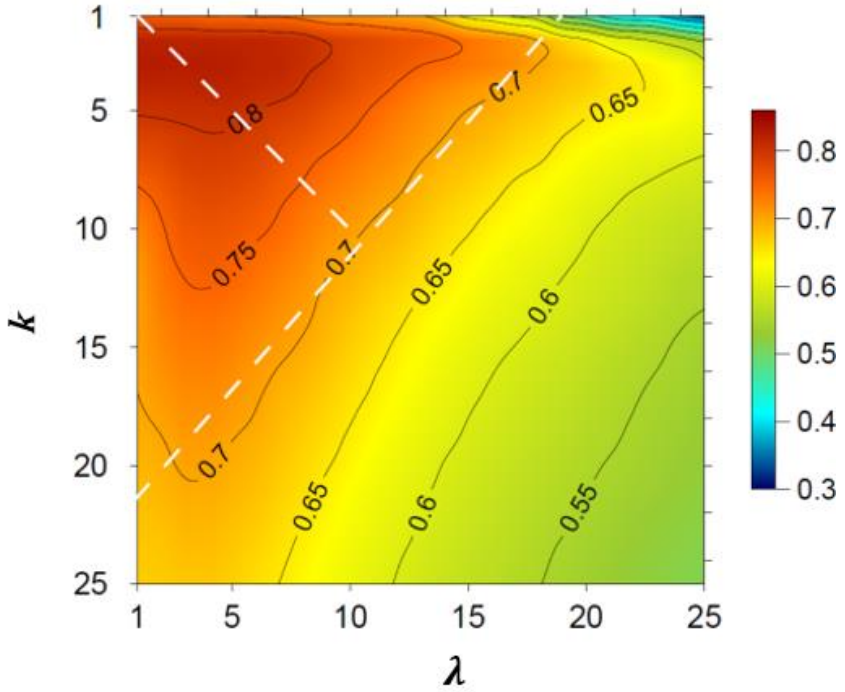


Рис. 3. Значения коэффициента детерминации, рассчитанные на сетке параметров

После того, как значения гиперпараметров определены, задача построения контрастного образа состоит в определении интересующей величины – коэффициента затухания методом *kNN* с модифицированным расстоянием (2) в узлах трехмерной решетки. На Рис. 4 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. 4 приведены также короткие вертикальные сечения, соответствующие этим линиям, на которых соответствующие области диаметром менее 100 м. также отчетливо видны.

Использованный подход позволяет получить контрастное изображение неоднородных областей, что позволяет выделить объекты, чьи геометрические размеры меньше расстояния между скважинами. 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. Модель может быть

улучшена, если привлечь дополнительные данные (геологические, сейсмические, магнитные) для их совместной интерпретации.

Таким образом, возможно адаптировать подход к моделированию на основе алгоритмов машинного обучения к геофизическим приложениям с дефицитом данных в следующий **метод**:

1. Постановка задачи моделирования в терминах машинного обучения, включая
   - выбор типа обучения (с учителем или без);
   - выбор типа решаемой задачи (регрессия, классификация, кластерный анализ, распознавание изображений и т.д.);
   - определение целевой переменной;
   - определение списка потенциальных признаков;
   - подготовка исходных данных (включая работу с пропусками и качеством данных).
2. Выбор алгоритма машинного обучения
   - это математическая модель интерпретации данных наблюдений. В зависимости от специфики исходных данных, одни модели могут быть более предпочтительны, чем другие.
   - идентификация прочих ограничения, включая требования к интерпретируемости результатов анализа, вычислительной сложности и скорости обучения.
3. Выбор функционала качества построенной математической модели
   - В зависимости от типа и объема доступных данных устанавливается функционал, позволяющий количественно сравнить реальное наблюдение и восстановленное моделью синтетическое значение моделируемой величины.
4. Обучение алгоритма машинного обучения состоит из:
   - процедуры настройки гиперпараметров;
   - выбора стратегии валидации и ее параметров (кросс-валидация, отложенная выборка);
   - выбора критерия остановки (например, при достижении оптимального значения функционала качества обучения);
   - финального обучения модели на всем доступном объеме данных.
5. Применение обученной модели к новым данным состоит из:
   - подготовки новых данных, для которых необходимо смоделировать значение целевой переменной в формате аналогичном данным обучения;
   - применения обученной модели на новых данных с целью восстановить значение моделируемой целевой переменной;
   - анализа и интерпретации итоговых результатов (включая, например, картографирование).

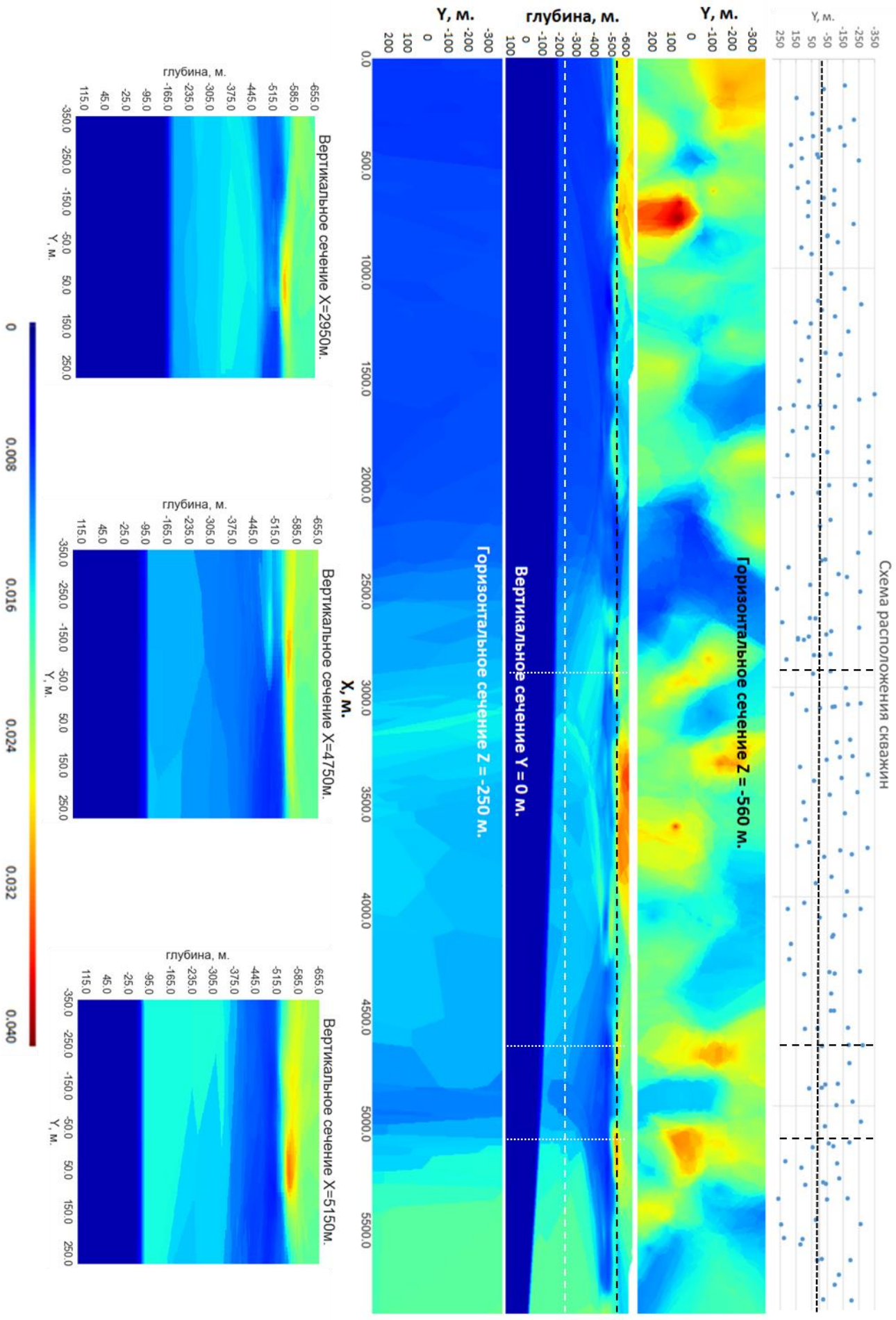


Рис. 4. Результаты трехмерного моделирования

Выводы по главе 1

1. Приведен обзор применения алгоритмов машинного обучения в различных геофизических задачах: поиск полезных ископаемых, анализ месторождений, прогнозирование опасных явлений. Рассмотрен класс задач с пропусками данных.
2. Предложена схема моделирования на основе алгоритмов машинного обучения на примере построения трехмерной модели среды при проведении межскважинного электромагнитного просвечивания.
3. Предложенная схема обобщена в новый метод пространственно-временной интерполяции геофизических характеристик и полей в условиях дефицита данных.

Во **второй главе** представлены результаты применения разработанного метода для геофизического анализа региона Фенноскандия на основании данных экспериментов POLENET/LAPNET и SVEKALAPKO. Для анализа доступны данные трех типов, все наборы характеризуются дефицитом наблюдений. Набор A – двумерные данные волновых форм, по которым возможно определить наличие или отсутствие поверхностного слоя пониженной скорости *Vs* для 60 станций в регионе. Набор B – данные скоростных разрезов под станциями для 23 станций южной части региона. Набор C – двумерные данные для 61 станций региона, для которых по данным инверсии приемных функций определены глубины Мохо.

Разработанный метод на основе машинного обучения применен к набору данных C. Построенная карта глубины Мохо в регионе изображена на Рис. 5. В качестве алгоритма машинного обучения использовалась регрессия на основе *kNN*, функционал качества – средняя абсолютная ошибка, настройка гиперпараметров осуществлялась по кросс-валидации, значение k=4 минимизирует среднюю ошибку построения на уровне 3.7 км.

В южной части региона форма границы Мохо хорошо коррелирует с основными геологическими структурам на поверхности. В частности, наиболее толстая кора находится в районе Центрального финского гранитоидного комплекса.

Полученные новым методом результаты интерпретации данных инверсии приемных функций в регионе сопоставлены с результатами построений другими методами по четырем разрезам, отмеченным на карте черными линиями. Наблюдается хорошее соответствие по всем профилям. Профиль LP51–LP75 из северной части региона приведен на Рис. 6. Видно, что результаты нового построения (красная сплошная линяя – основное построение, красные пунктирные линии – оценка ошибки построения) находятся в границах доверительных интервалов предыдущих построений по всей длине разреза.

Важной особенностью строения данного региона является также наличие относительно тонкого поверхностного слоя с низкими скоростями поперечных сейсмических волн *Vs*. Разработанный метод на основе машинного обучения применен к набору данных A. Построенная карта поверхностного слоя низких скоростей поперечных сейсмических волн *Vs* изображена на Рис. 7. В качестве алгоритма машинного обучения использовалась классификация на основе *kNN*, функционал качества – площадь под *ROC*-кривой, настройка гиперпараметров

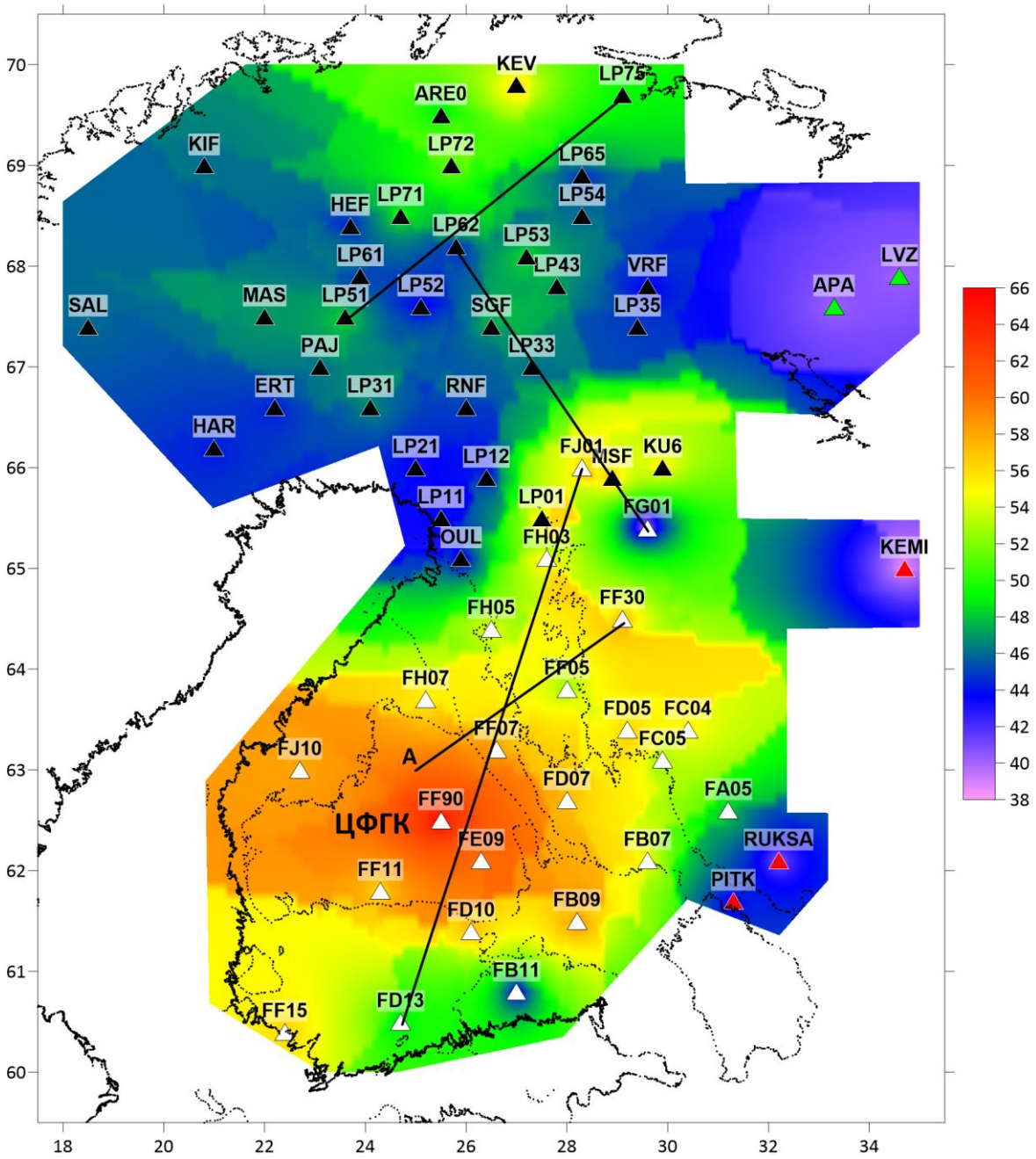


Рис. 5. Карта глубины границы Мохо

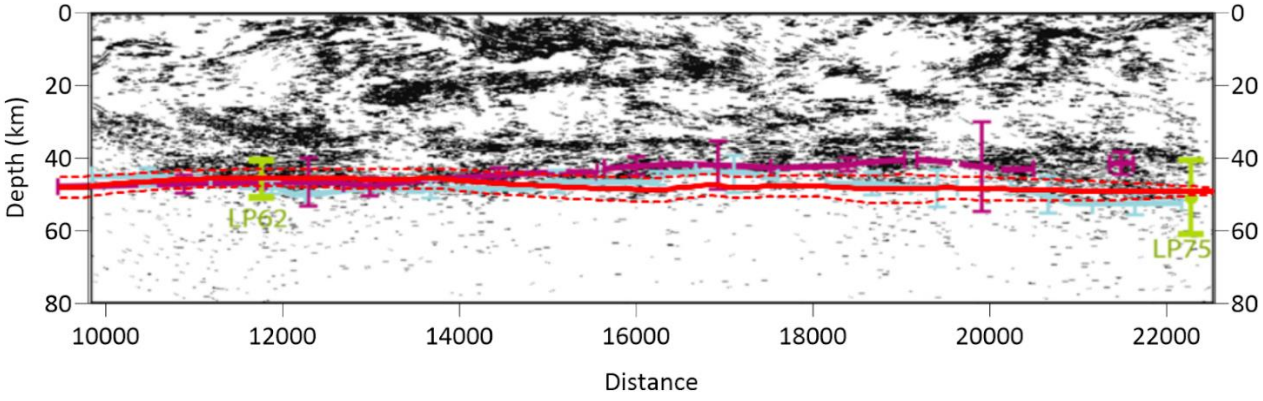


Рис. 6. Глубина Мохо вдоль профиля LP51–LP75

осуществлялась по кросс-валидации. Значение k=4 максимизирует функционал качества построения на уровне порядка 0.83 (значение 1 соответствует идеальной модели). На карте синим цветом выделены области, в которых вероятность присутствия низкоскоростного приповерхностного слоя превышает 55%, а оливковым — где эта вероятность меньше 45%; бежевым цветом отмечены области с «промежуточными» вероятностями от 45% до 55%. Видно, что низкоскоростной слой имеется в северной, центральной и южной частях региона. В северной и южной частях наличие слоя связано с геологическим строением. В северной части наличие такого слоя связывают с протерозойскими осадочными породами, покрывающими архейский фундамент. В южной части известно, что присутствуют рапакиви, обладающих другими скоростными свойствами, чем твердые породы. В центральной части региона, нет пространственной корреляции с границей архейских и протерозойских пород. Наличие низкоскоростного слоя здесь может быть объяснено неравномерным движением пород в ходе последней постледниковой релаксации.

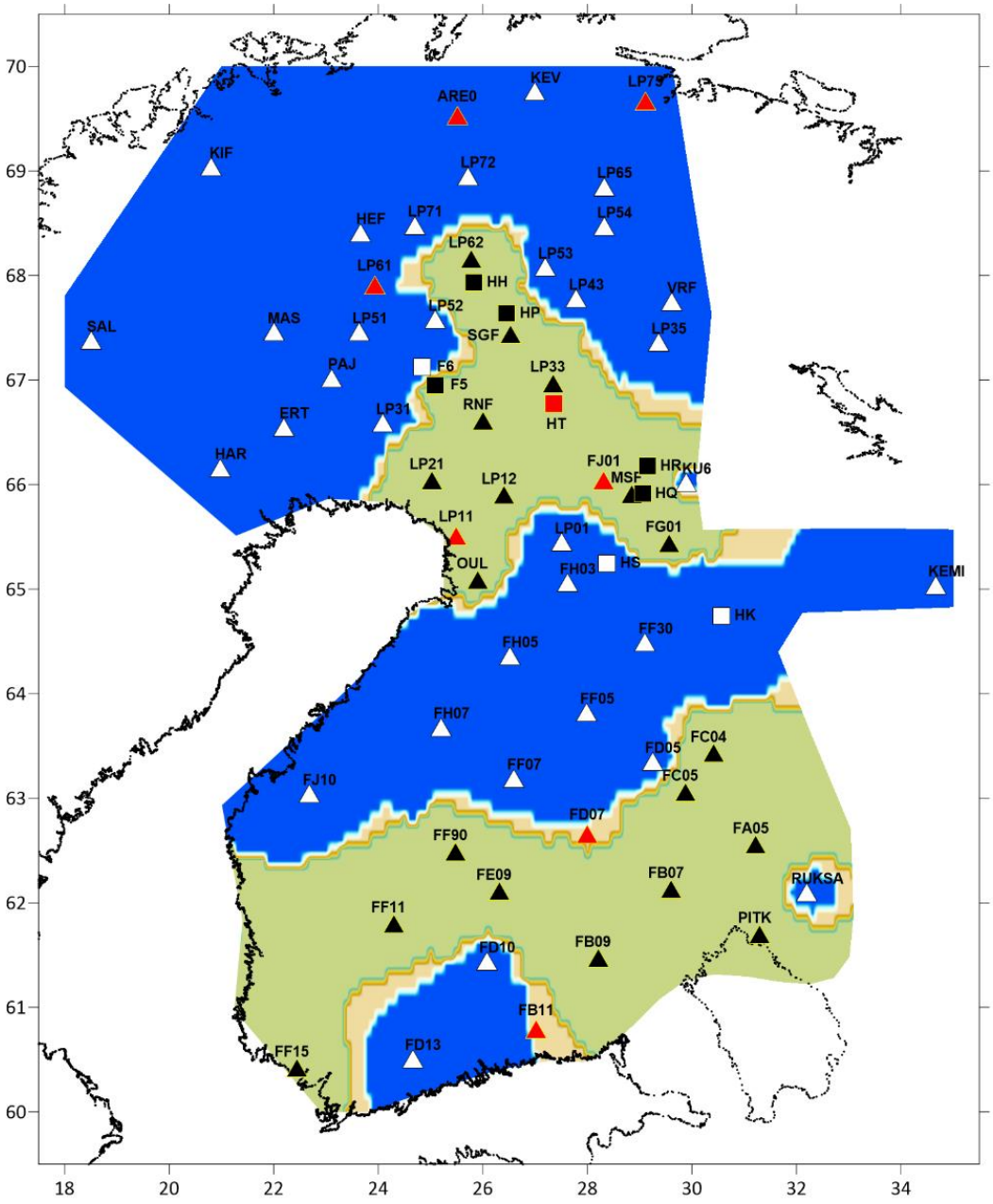


Рис. 7. Карта поверхностного низкоскоростного слоя Vs

Для более детального исследования центральной части региона построена трехмерная скоростная модель. Разработанный метод на основе машинного обучения применен к набору данных B. Построенная трехмерная сейсмическая модель изображена на Рис. 8. В качестве алгоритма машинного обучения использовалась регрессия на основе *kNN*, функционал качества – среднеквадратичная ошибка, настройка гиперпараметров осуществлялась по кросс-валидации. В данной задаче присутствует пространственная анизотропия, поэтому был использован дополнительный гиперпараметр – скейлинг λ. Значение $k = \lambda = 42$ соответствует уровню ошибки в 0.05 км/с – половине точности измерения *Vs*.

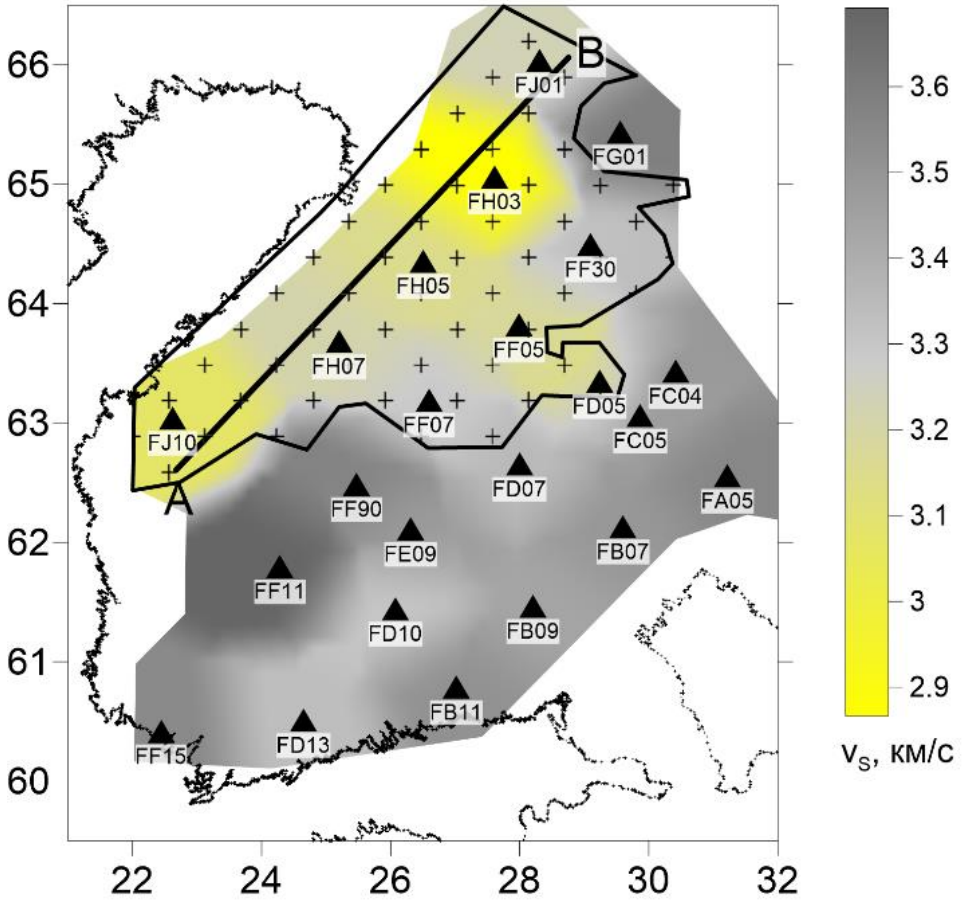


Рис. 8. Скоростная модель региона

На Рис. 8 знаками «+» замощена область, соответствующая контурам низкоскоростного слоя по предыдущему построению (Рис. 7). Вертикальный разрез по профилю, перпендикулярному границе архейской и протерозойской областей приведен на Рис. 9. Как отмечалось ранее, понижение скорости *S*-волн в верхней части коры связано с наличием водонасыщенных трещин, образованных постледниковой релаксацией. В отличие от качественного анализа, построенная цифровая модель позволяет определить не только область, содержащую низкоскоростной слой, но также его строение и пространственную структуру. На разрезе [*A, B*] видно, что толщина слоя и контрастный переход скорости в нем в архейских породах (справа) значительно выше, чем в протерозойских (слева). Этот вывод согласуется с ранее известными сведениями о том, что параметр трещиноватости в архейской части региона почти в два раза выше, чем в

протерозойской. Минимальные понижения скорости и толщины слоя в данной области относятся к переходной зоне архей-протерозой.

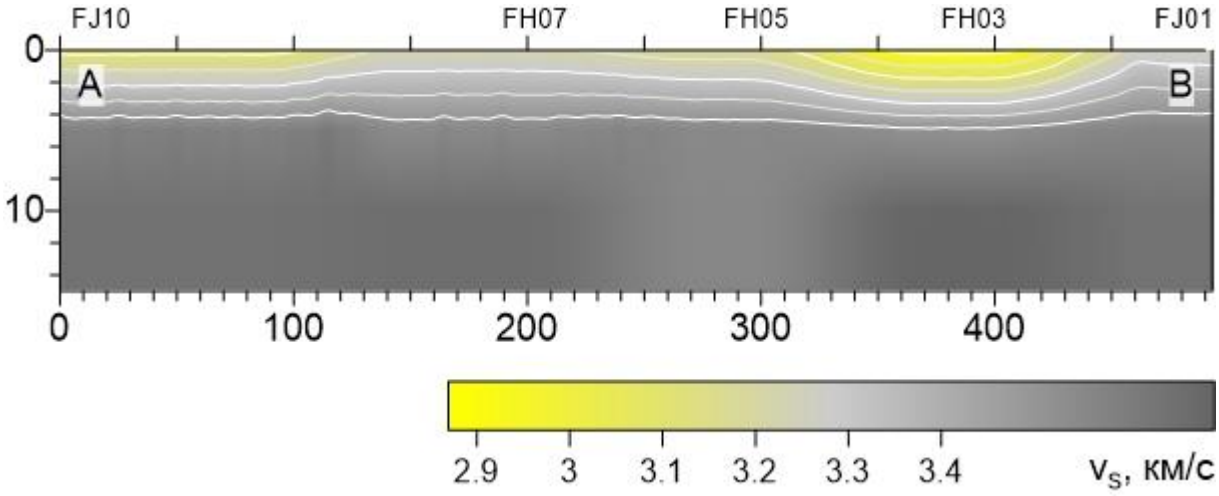


Рис. 9. Вертикальное сечение [A, B]

Выводы по главе 2

1. С помощью разработанного метода на основе машинного обучения построена уточненная карта границы Мохоровичича для региона Фенноскандия с использованием метода *k* ближайших соседей. Вертикальные сечения построенной модели хорошо согласуются с полученными ранее профилями.
2. С помощью разработанного метода на основе машинного обучения построена карта слоя с низкими значениями скорости поперечных сейсмических волн *Vs* для региона Фенноскандия с использованием метода *k* ближайших соседей. Показано, что слой низких сейсмических скоростей на поверхности присутствует на значительной части региона, включая области с протерозойскими породами.
3. Разработанный метод применен к данным инверсии волновых форм приемных функций для построения цифровой модели южной части региона. Это позволило детальнее проанализировать структуру верхней части коры под низкоскоростным слоем.
4. Проведенный 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 **третьей главе** рассматривается задача построения интеллектуальной системы на основе коротких временны*ы*х рядов наблюдений для прогноза ледового заторообразования на участках северных рек. Система разрабатывалась на данных р. Северная Двина, в дальнейшем была применена на данных другого региона – бассейне р. Лена.

Существуют довольно сложные гидродинамические модели прогноза наводнений и, в общем, речного стока, основанные на численных решениях различных форм системы дифференциальных уравнений Навье-Стокса. Важной частью входных данных этих моделей являются параметры, отражающие факт заторообразования на данном участке. Места заторов, в основном, стационарны, т.к. существенно зависят от геометрической формы русла, поэтому достаточно предсказать мощность явления, чтобы существенно улучшить прогноз наводнений на таких участках. Разработанный метод на основе машинного обучения применен для для создания прогнозной интеллектуальной системы для двух регионов участка р. Северная Двина и р. Лена. В качестве признаков доступны измерения на гидропостах и метеостанциях на основе которых рассчитан ряд гидрометеорологических факторов, целевая переменная бинарная (1 – затороопасный сезон, 0 – иначе), объём доступных измерений: 26 сезонов по С. Двине и 34 сезона по Лене, классификация на основе алгоритмов голосования и вычисления оценок, функционалом качества являлась точность (доля верно классифицированных элементов обучающей выборки). Процесс разработки системы проведен на участке р. Северная Двина. На Рис. 10 приведена схема исследуемого участка. Использованы данные шести гидропостов и одной метеостанции. Выделены 11 интегральных факторов возникновения сильного затора. Факторы делятся на гидрологические и метеорологические, также доступны с различной заблаговременностью: от нескольких дней до нескольких месяцев до начала весеннего вскрытия, в Табл. 1 приведен их список.

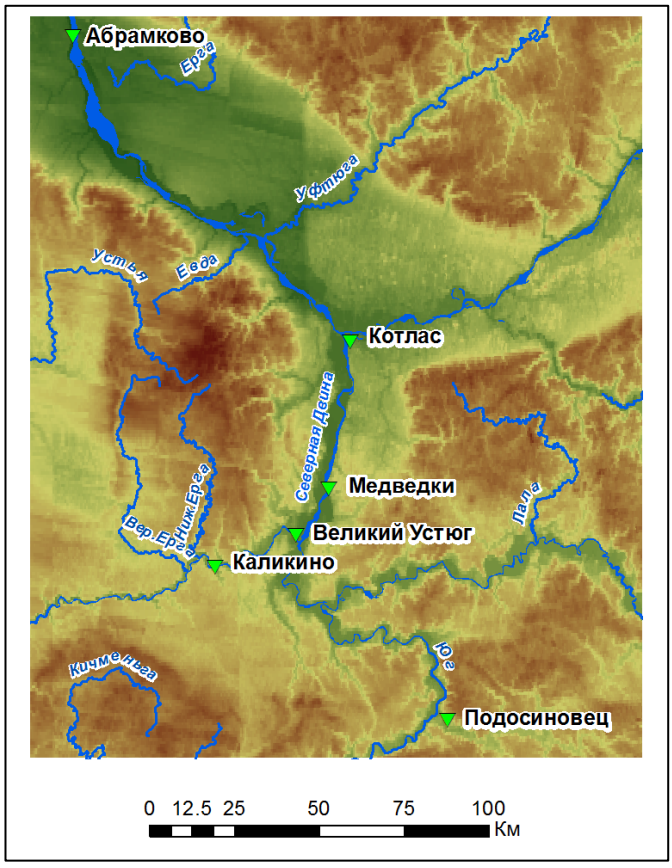


Рис. 10. Схема расположения речных постов на р. С. Двина и ее притоках

Доступна классификация уже произошедших событий. Первоначально разработка системы проводилась на выборке из 20 сезонов, чуть позже была добавлена дополнительная выборка из 6 сезонов. Исходные данные представляют собой трехмерный массив, имеющий три размерности: временную (сезоны), пространственную (посты и метеостанции) и признаки (интегральные характеристики). Каждому слою этого массива поставлено в соответствие наблюдавшееся состояние, т.е. проведена классификация.

Прогнозная система обучается на этих данных и, получив на вход новый слой данных, возвращает прогнозный класс мощности события.

Проведено три эксперимента по оценке качества и прогнозированию. На кросс-валидации на первоначальной выборке из 20 сезонов получено 3 ошибки. Режим прогнозирования на дополнительной выборке из 6 сезонов: система каждый раз полностью переобучалась на всем доступном объеме предшествующих данных,

на вход поступали новые данные (которых система не видела до этого), строился прогноз и сравнивался с фактическим реализовавшимся состоянием. Проведено прогнозирование всей дополнительной выборки из 6 сезонов в виде последовательности. Все 6 сезонов классифицированы верно. Таким образом, оцененное по кросс-валидации качество прогнозирования, составило 85%.

Табл. 1. Признаковое пространство для бассейна р. С. Двины

№	**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

Интеллектуальная система позволяет оценить важность влияния каждого из признаков на итоговый результат классификации – это порождает связь с физическим процессами и дает инструмент проверки гипотез в условиях дефицита данных, когда

напрямую по исходным данным достаточно трудно оценить стандартные метрики статистических связей, например, корреляции. На Рис. 11 приведена интерпретация важности признаков, использующая понятие информационного веса (приведено к 100%). Наибольшее влияние оказывает гидродинамический признак, отражающий скорость изменения уровней и гидрометеорологические признаки, отражающие наименьшую заблаговременность до периода начала вскрытия. Полученные оценки хорошо согласуются с теоретическими знаниями о факторах заторообразования.

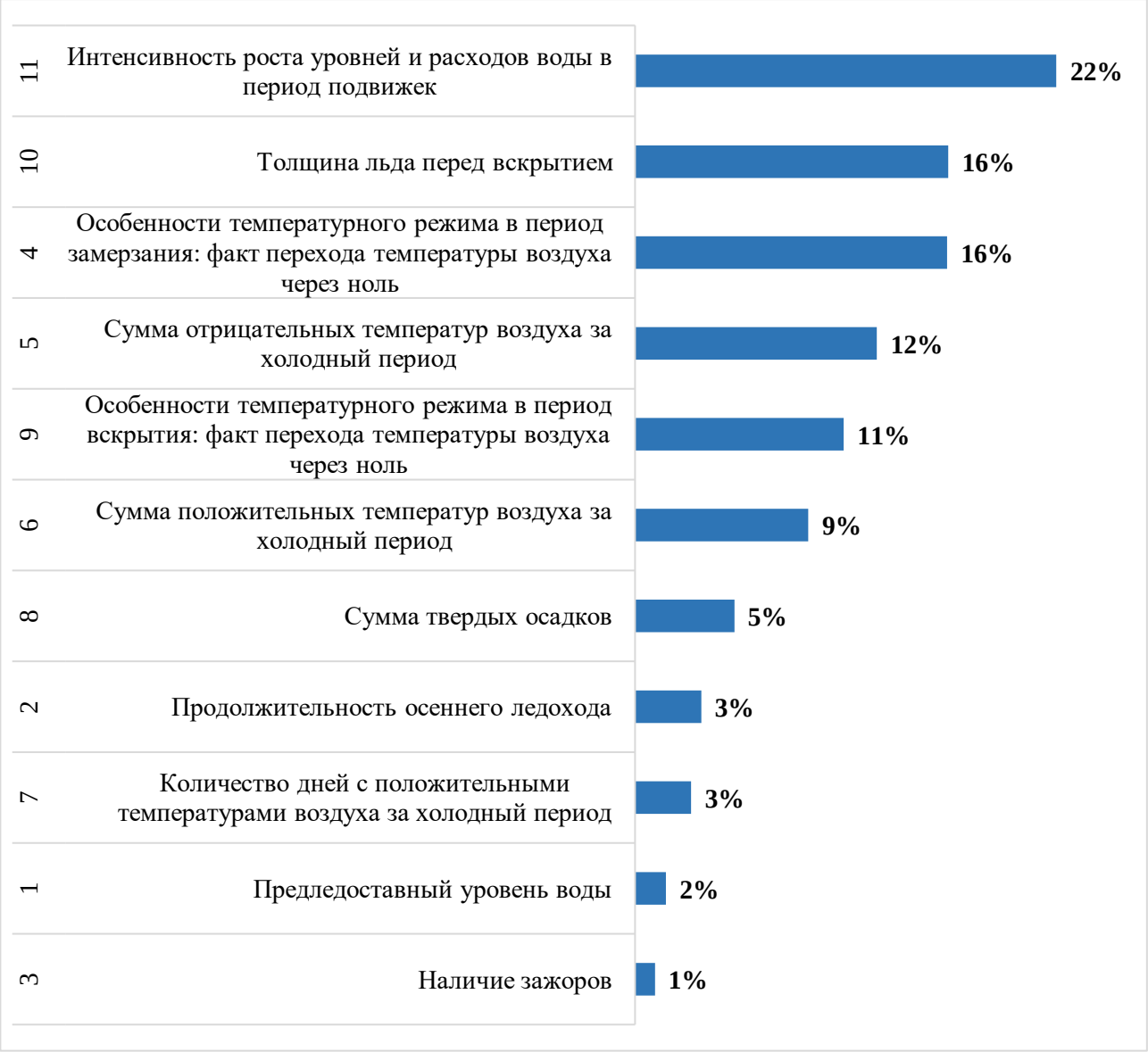


Рис. 11. Вклад признаков в итоговый результат классификации

Разработанная прогнозная интеллектуальная система применена на данных нового региона – бассейна р. Лена. Применение разработанной технологии к данным другого региона обусловлено схожей постановкой задачи прогнозирования. Возможно сформулировать гипотезы о зависимости ряда интегральных показателей (признаков) на итоговый результат явления (целевую переменную) по данным измерений гидрологических и метеорологических показателей на гидропостах и метеостанциях в бассейне р. Лена. В качестве исходных данных использованы два набора гидрологических и метеорологических данных в регионе за период 1985-2019

гг. Указанные наборы содержат данные по 27 гидропостам и 38 метеостанциям в регионе. В качестве основных факторов возникновения заторообразования выбрано 7 интегральных признаков, построенных на имеющихся исходных данных: среднесуточный уровень воды, среднесуточная температура воды, толщина льда перед вскрытием, среднесуточный расход воды, сумма твердых осадков, запас воды в снеге, среднесуточная температура воздуха. На основании анализа статистических данных доступна итоговая интегральная классификация сезонов по критерию наличия (класс 1) или отсутствия (класс 0) мощных и продолжительных заторов в регионе в указанный сезон.

Особенностью данной задачи является необходимость адаптации метеоданных под координаты гидропостов. Поскольку пространственная привязка измерений на метеостанциях и гидропостах различная, возникает задача по переносу измерений с метеостанций на гидропосты, координатами которых индексированы входных данные для прогнозирования. Задача была решена с помощью интерполяции на основе машинного обучения. Использован алгоритм случайного леса. Проведено обучение системы, оценена точность с использованием кросс-валидации. Лучший результат дал оценку точности классификации порядка 76% по функционалу качества «точность», т.е. 26 из 34 сезонов корректно классифицированы системой к своим классам.

Проведена оценка важности факторов. Основное влияние оказывает группа гидрологических факторов, что соответствует ранее полученным результатам для Северной Двины. Среди гидрологических факторов наименьшую оценку важности получил среднесуточный расход воды. Вероятно, это связано с существенной нестационарностью этой характеристики по времени, выявленной ранее при анализе многолетних рядов гидрологических наблюдений в регионе. Группа метеорологических факторов показала достаточно слабый отклик на мощность заторного явления.

Выводы по главе 3

1. Разработана интеллектуальная система, предназначенная для краткосрочного прогноза образования заторов в весенний период на северных реках на основе анализа коротких временных данных с применением специализированных методов теории машинного обучения.
2. Система позволяет проверять гипотезы о влиянии признаков на итоговую мощность явления в условиях ограниченного набора исходных наблюдений на гидропостах и метеостанциях. Построено ранжирование используемых признаков по степени влияния на результат. 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 для региона бассейна р. Северная Двина составляет 85%.
4. Разработанная интеллектуальная система применена к другому региону – бассейну р. Лена. Оцененная достоверность прогнозирования для нового региона составляет 76%. Основные качественные выводы соответствуют полученным ранее результатам.

ЗАКЛЮЧЕНИЕ

В результате проведенных исследований и практических разработок **была достигнута цель диссертационного исследования** – осуществлена разработка компьютерных систем и методов обработки данных в условиях ограниченного количества данных, недостаточных для проведения классического статистического анализа, на основе методов машинного обучения, 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. Показаны основные способы решения задачи о восстановлении пропусков в данных.

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 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 результате сочетания новых разработок для трехмерного локального и двумерного регионального моделирования геофизических полей был разработан метод пространственной интерполяции нерегулярно распределённых геофизических измерений с пропусками данных, основанный на применении теории машинного обучения. Метод применим для данных с экстремальной анизотропией пространственного распределения. Разработанный метод является универсальным, и не зависит от способа получения одномерных моделей. Необходимые вычисления выполняются непосредственно для нужных сечений, при этом форма сечений может быть произвольной. В качестве его реализации была построена трехмерная региональная модель коры для южной Фенноскандии.

Разработан метод для анализа многомерных временных рядов ограниченной длины. Техническая реализация метода выполнена на основе прогнозной интеллектуальной системы для процесса ледяного заторообразования на северных реках. 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 участках р. Северная Двина и р. Лена. Система также допускает применение в качестве инструмента проверки гипотез относительно влияния тех или иных факторов на итоговые состояния опасных процессов в условии дефицита данных, когда результатам традиционных методов анализа данных не хватает статистической значимости. Проведена валидация системы на новых, недоступных в момент первоначальной разработки, исторических данных. После этого, система применена на новом, отличном от первоначального использованного для разработки системы, регионе – бассейне р. Лена. Итоговая оцененная достоверность прогнозирования составила от 76% до 85%.

ОСНОВНЫЕ ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации, входящие в перечень рецензируемых научных журналов из списка ВАК:

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

Публикации в изданиях, входящих в международные реферативные базы данных и системы цитирования Wos и Scopus:

3. 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.
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.

Публикации в изданиях, входящих в международную реферативную базу данных и систему цитирования Scopus:

5. Алешин И.М., Козловская Е.Г., **Малыгин И.В.** Применение методов теории машинного обучения в томографии приемных функций // Геофизические исследования. — 2022. —Т.23, №1. — С. 49–61.
6. 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.
7. Алешин И.М., **Малыгин И.В.** Интерпретация результатов радиоволнового просвечивания методами машинного обучения // Компьютерные исследования и моделирование. — 2019. — Т. 11, № 4. — С. 675–684.

Прочие публикации:

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

Свидетельства о государственной регистрации программ для ЭВМ:

1. **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.
2. **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.
3. **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.

Подписано в печать __.07.2022
Формат 64×84/16. Объем: 1 п.л.
Тираж 100 экз. Заказ № __
Отпечатано в типографии «Реглет»
119526, г. Москва,
пр-т Вернадского, д.39
+7 (495) 363-78-90
www.reglet.ru