

Малыгин И.В.

*Институт физики Земли им. О.Ю. Шмидта РАН
Лаборатория геоинформатики*

Методы машинного обучения в геофизических задачах с дефицитом данных

диссертация на соискание ученой степени кандидата технических наук

25.00.10 – Геофизика, геофизические методы поисков полезных ископаемых

научный руководитель – к.ф.-м.н. И.М. Алешин

1. Цель исследования

Цель исследования – продемонстрировать эффективность базовых методов машинного обучения в геофизических исследованиях при дефиците входной информации на примере задач анализа результатов измерений с двумерным и трехмерным пространственным распределением, а также коротких временных рядов.

Для достижения цели получено решение следующих **задач**:

1. Разработка алгоритма построения трехмерной цифровой модели заданной области при проведении электромагнитного просвечивания на основе базового метода машинного обучения к ближайшим соседям. Программная реализация алгоритма, практическая апробация метода на реальных данных.
2. Разработка алгоритма построения двумерных цифровых моделей (карт) на основе регрессионного анализа и решения задачи классификации по результатам инверсии приёмных функций и смежных данным на основе базового метода машинного обучения к ближайшим соседям. Программная реализация алгоритма, практическая апробация метода для Фенноскандии.
3. Разработка метода прогноза заторообразования на основании короткого временного ряда гидрологических и метеорологических признаков на основе комбинаторно-логического подхода теории распознавания образов. Программная реализация и апробация прогнозной системы заторообразования для участка р. Северная Двина.

2. Дефицит данных

Методы машинного обучения

Дефицит данных

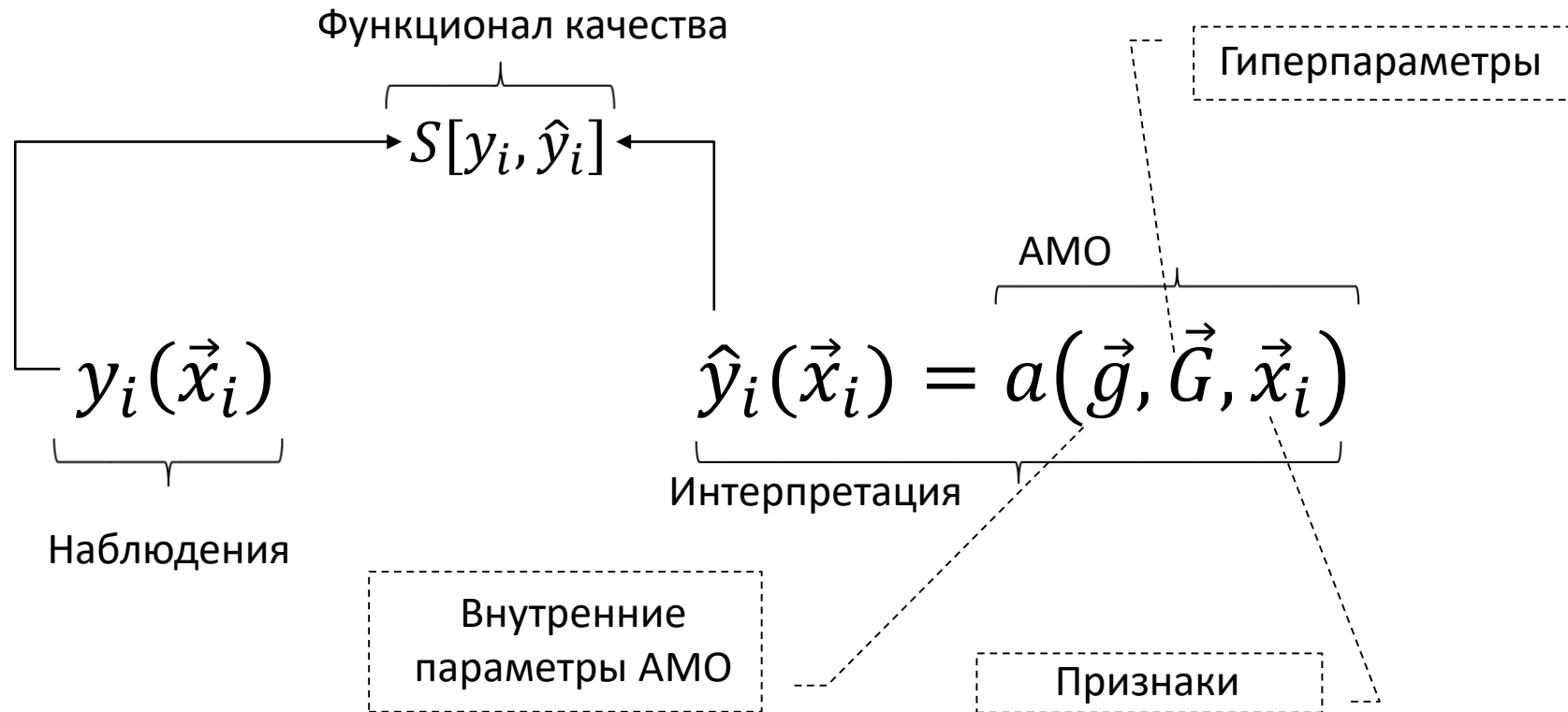
Большие данные

Классические статистические методы

Стат. значимость

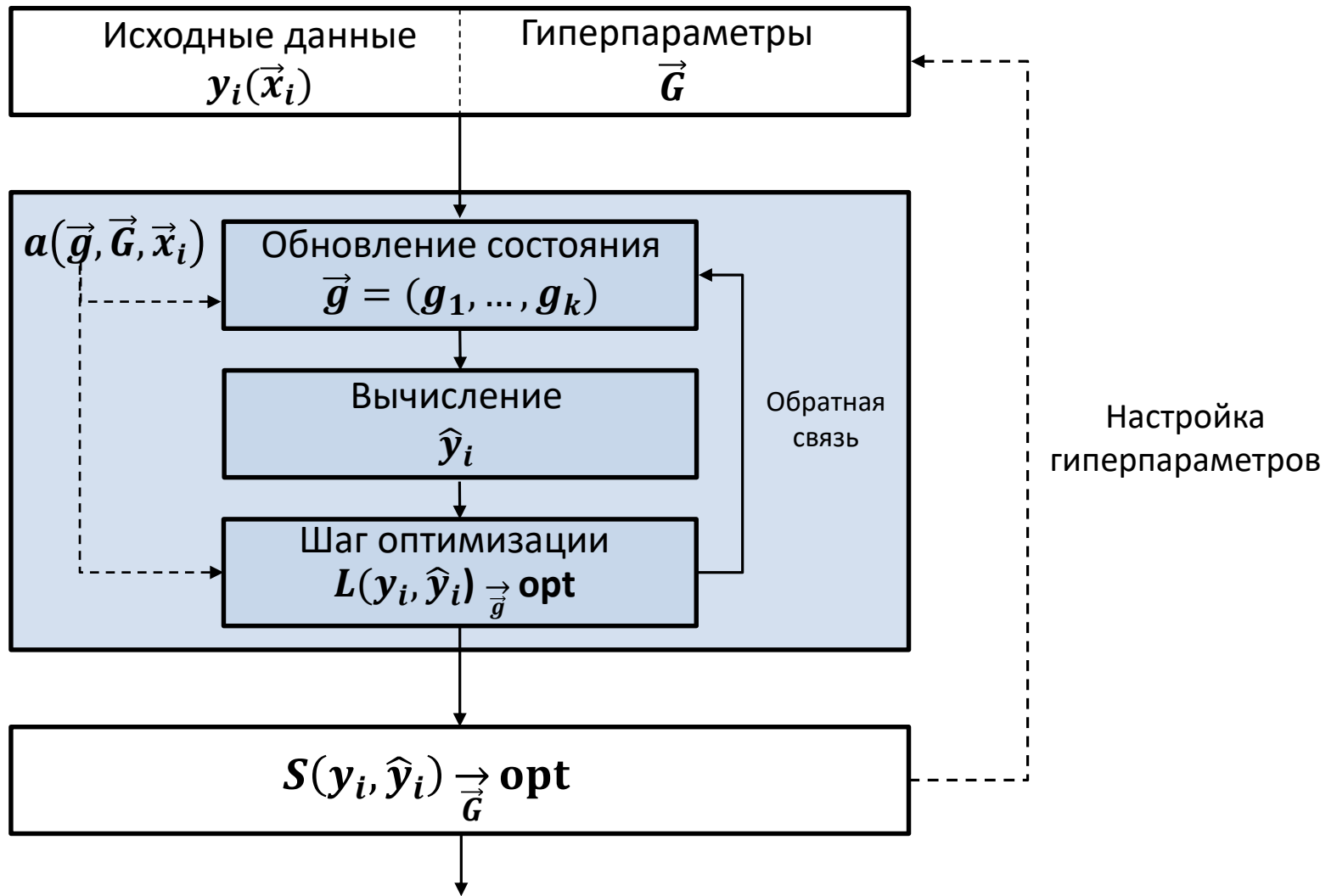
Производительность

3. Машинное обучение



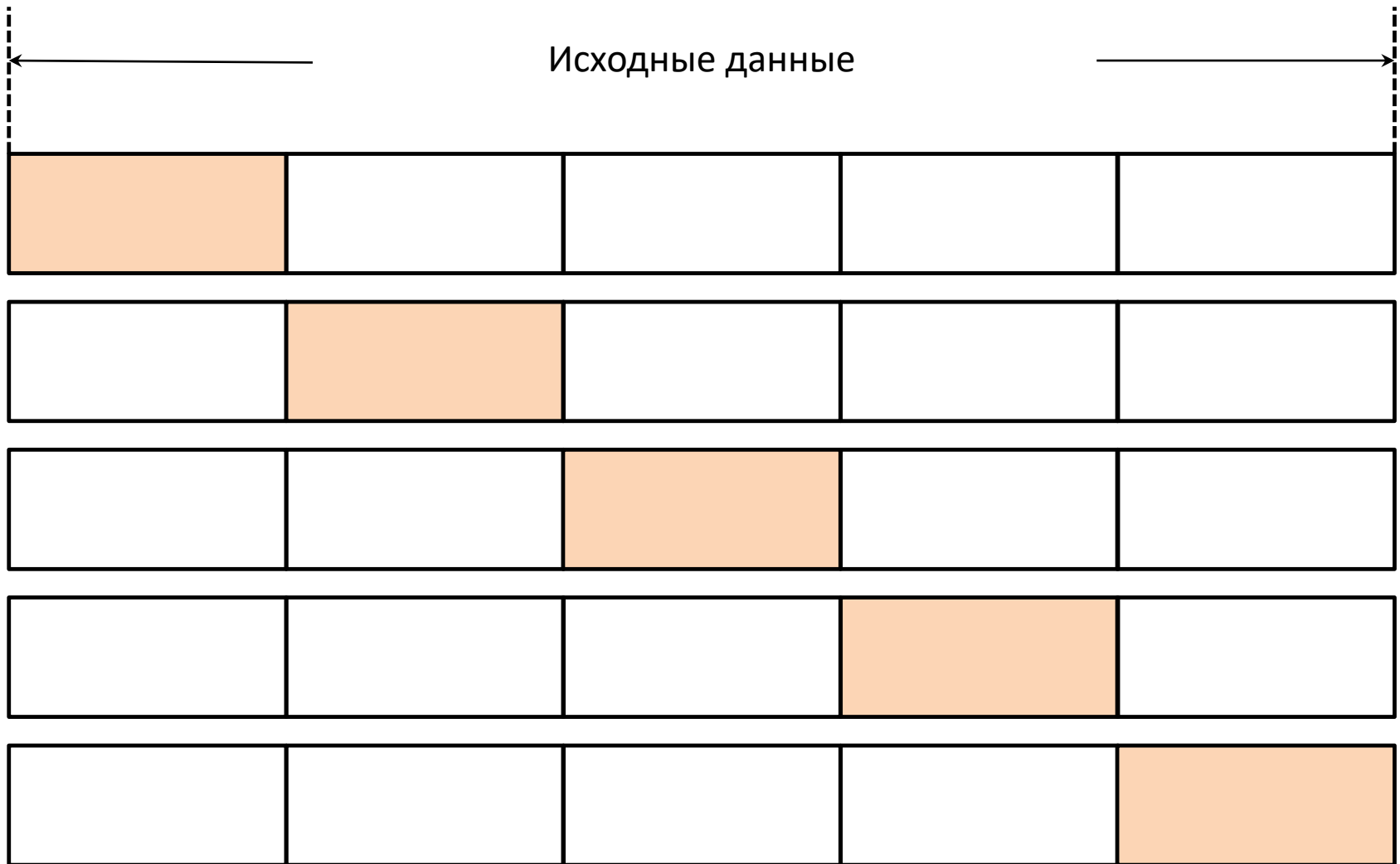
МО – Алгоритм **М**ашинного **О**бучения

4. Процедура обучения



Процедура обучения – выбор параметров $\vec{g}$ по исходным данным

5. Кросс-валидация



Обучающая выборка
(участвует в обучении)



Контрольная выборка
(не участвует в обучении)

6. Функционалы качества

Регрессия

$$MAE(a, X) = \frac{1}{l} \sum_{i=1}^l |a(x_i) - y_i| \quad R^2(a, X) = 1 - \frac{\sum_{i=1}^l (a(x_i) - y_i)^2}{\sum_{i=1}^l (y_i - \frac{1}{l} \sum_{i=1}^l y_i)^2}$$

Бинарная классификация

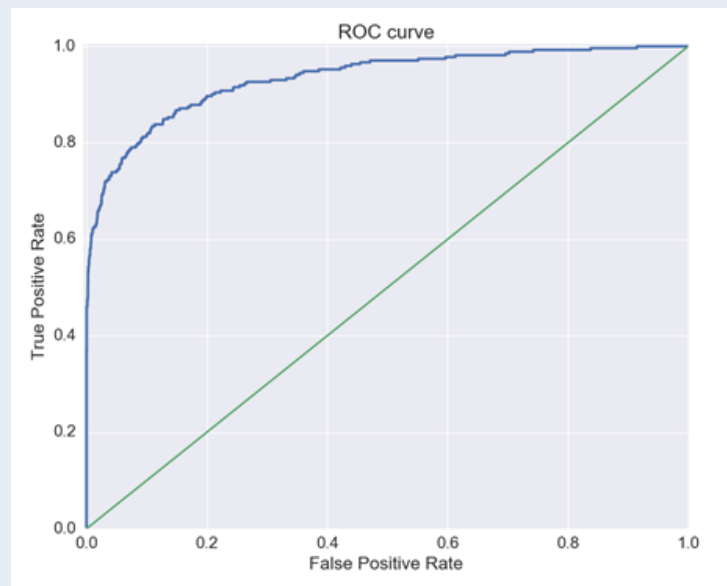
$$\text{accuracy}(a, X) = \frac{1}{l} \sum_{i=1}^l [a(x_i) = y_i]$$

ROC-кривая (оценка принадлежности в рамках бинарной классификации)

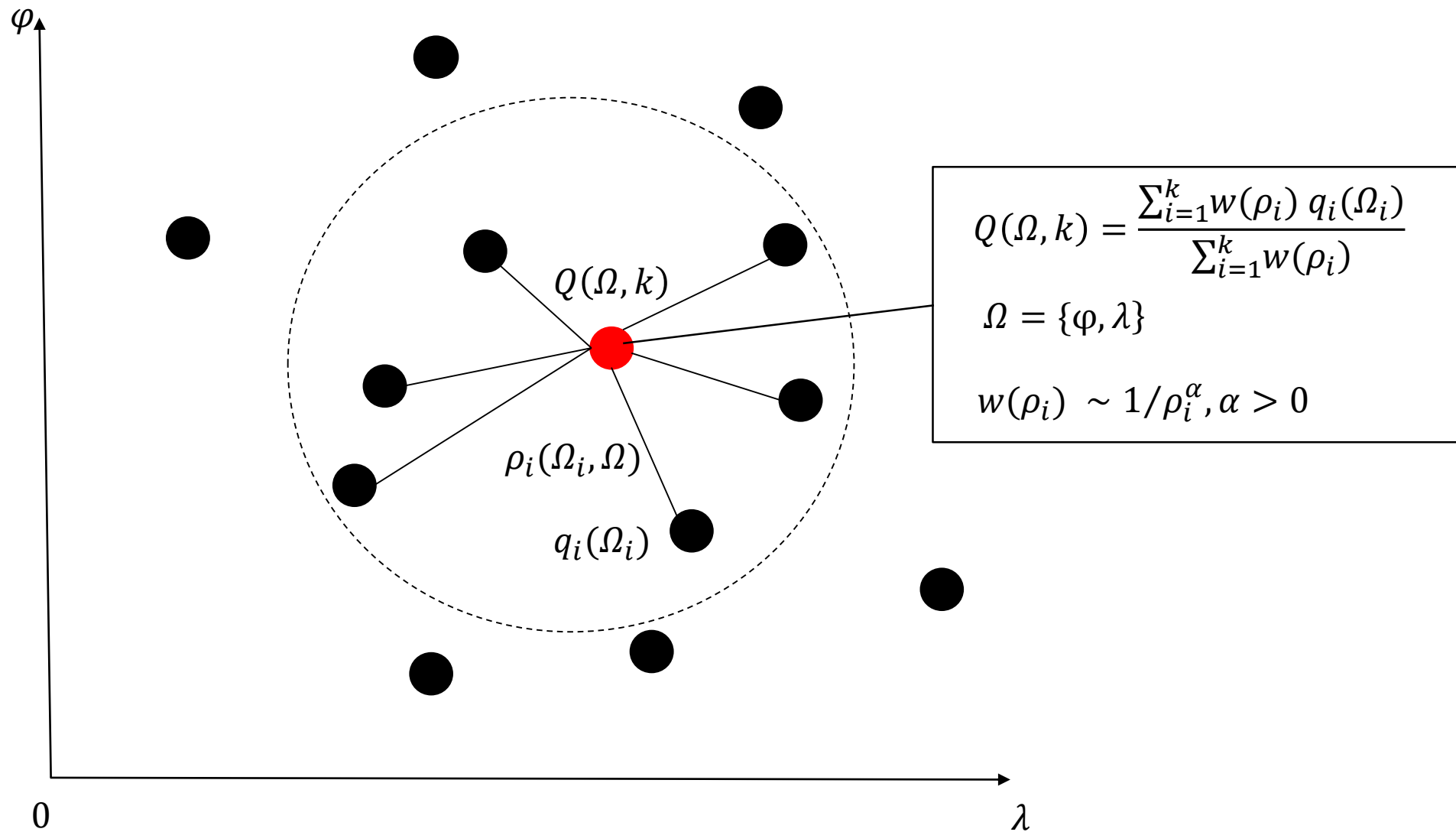
	$y = 1$	$y = 0$
$a(x) = 1$	True Positive (TP)	False Positive (FP)
$a(x) = 0$	False Negative (FN)	True Negative (TN)

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

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



7. kNN -интерполяция



8. 3D-модель среды при электромагнитном просвечивании

Постановка задачи:

Построить 3D-модель среды при проведении межскважинных исследований

Исходные данные:

Источник данных: результаты электромагнитного просвечивания;

Признаки: трехмерные координаты;

Целевая переменная: коэффициент затухания электромагнитных волн;

Объем доступных измерений: ~2300 точек.

Метод машинного обучения:

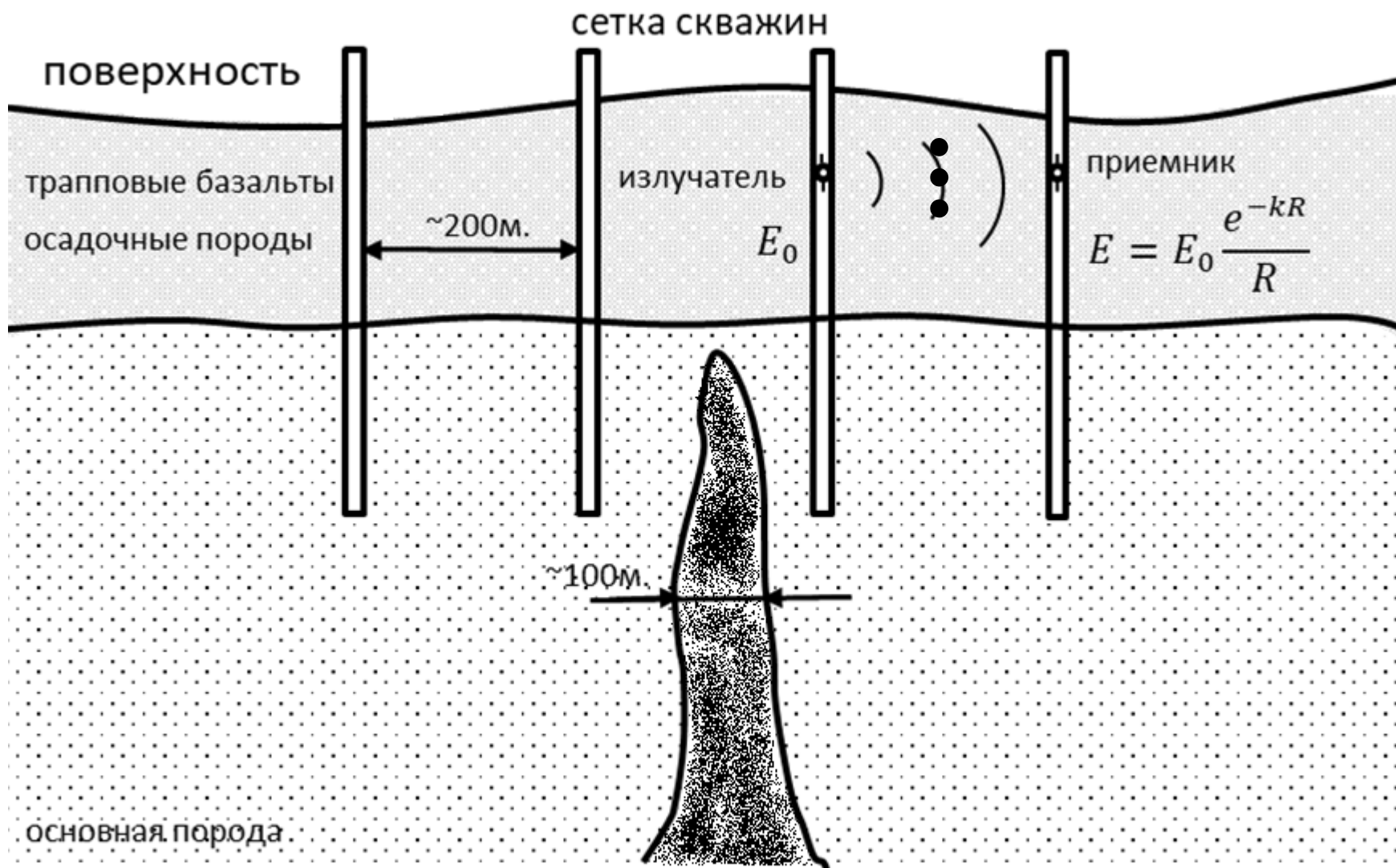
Регрессия на основе kNN

Настройка обучения и гиперпараметров:

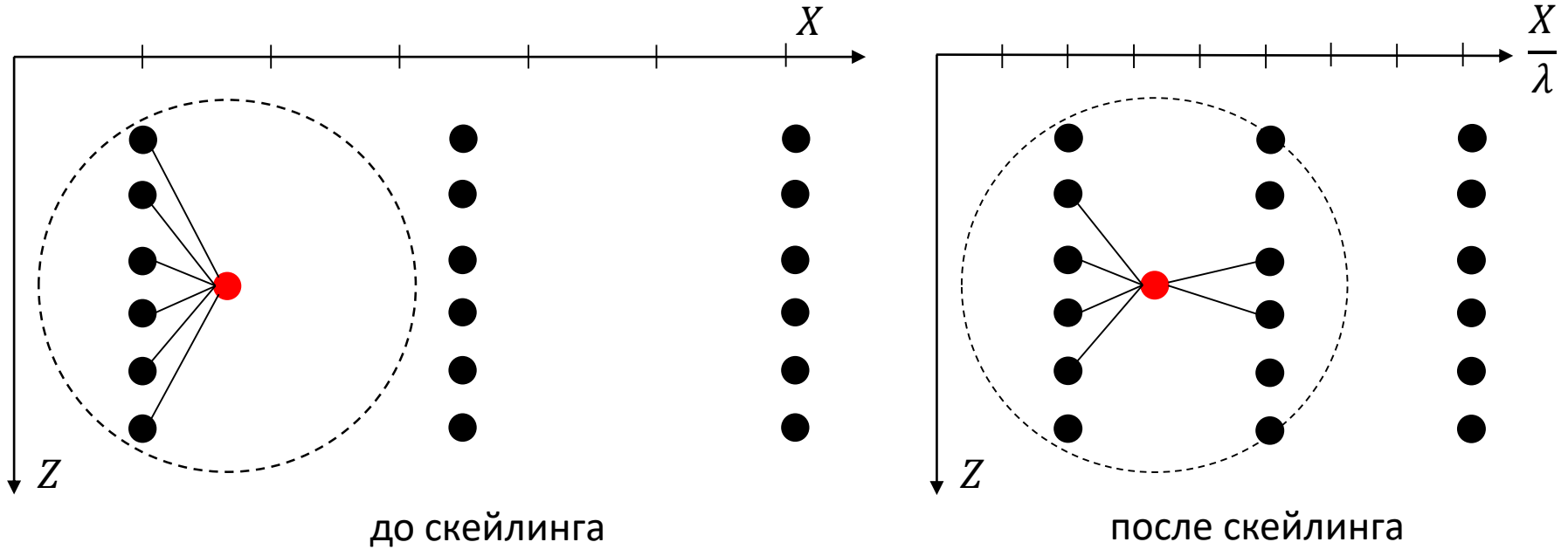
Функционал качества: R^2

Стратегия кросс-валидации: 5-fold

9. Схема просвечивания



10. Горизонтальный скейлинг и kNN -регрессия



Измерения группируются вдоль линейных объектов (скважин) – это порождает пространственную анизотропию

Скейлинг

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

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

kNN -регрессия

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

$$\sum_{i=1}^k w_i = 1; \quad w_i(\vec{r}, \vec{r}_i) \sim \frac{1}{\rho'(\vec{r}, \vec{r}_i)}$$

11. Подбор гиперпараметров модели

Функционал качества:

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

Оптимальная область:

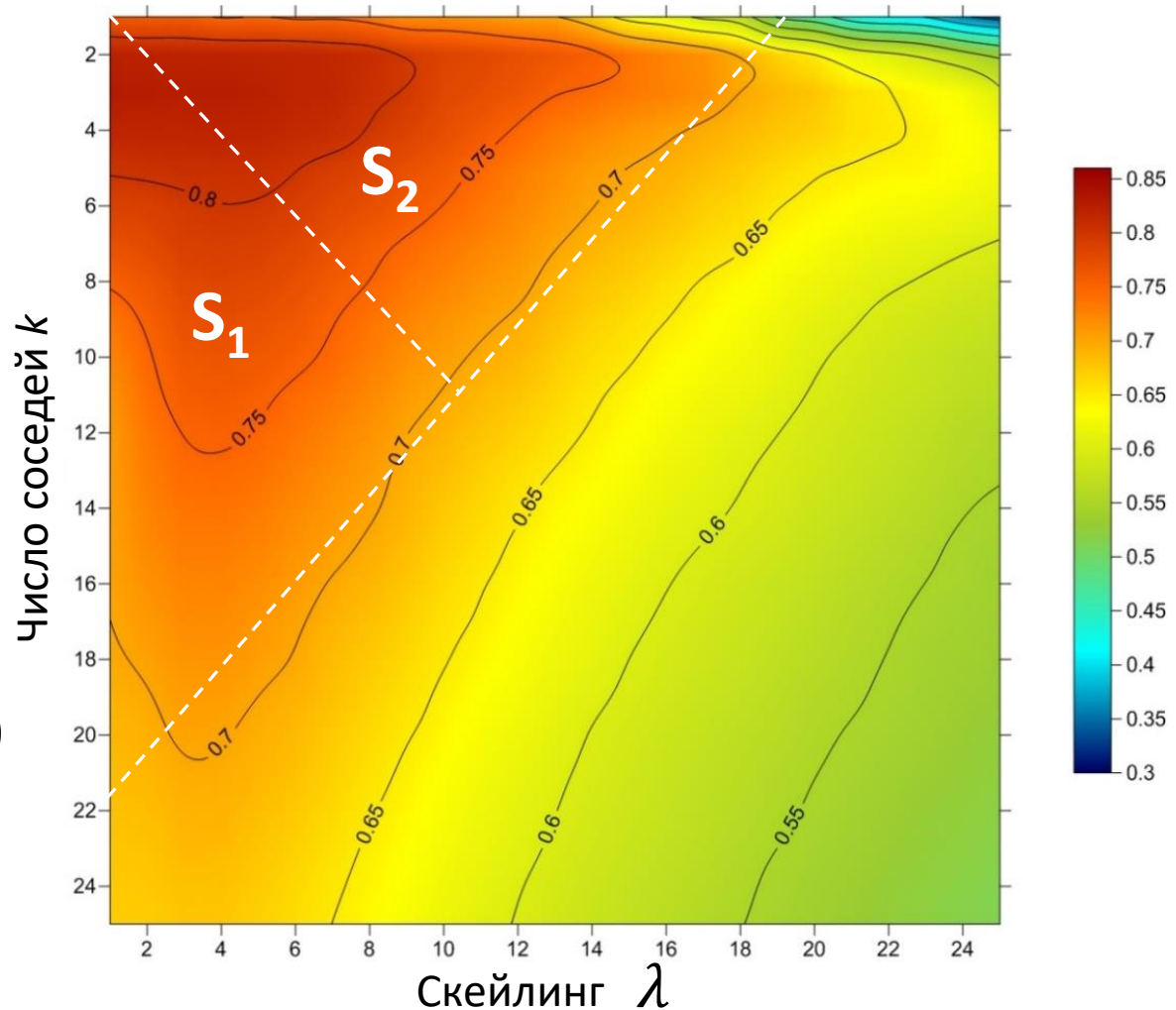
$$R^2 \geq 0.7$$

$S_1 = S_2$ (правило Парето)

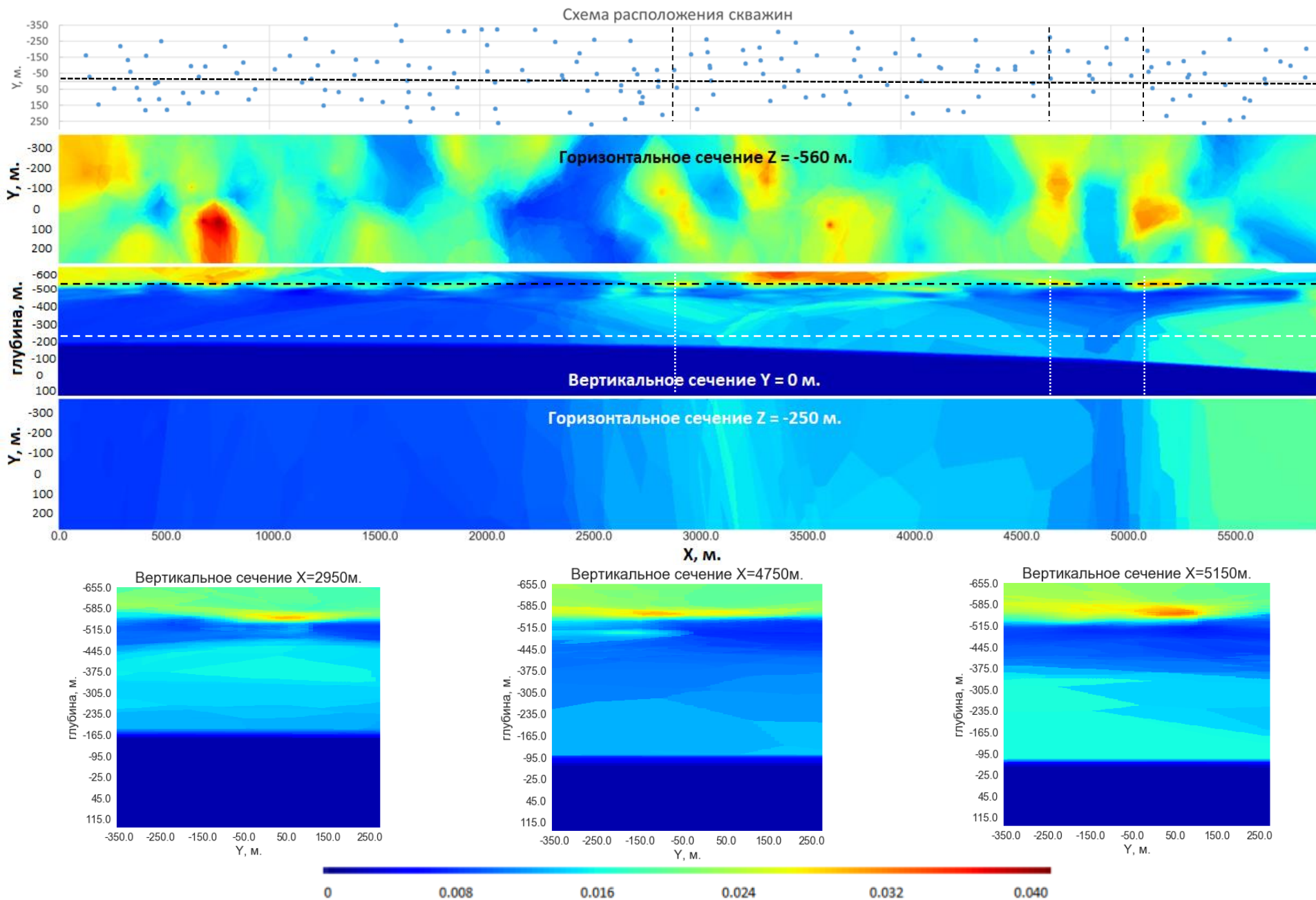
Итоговый выбор:

$$\lambda = 10$$

$$k = 11$$



12. 3D-модель электропроводности среды



13. Выводы

- Применение методов МО позволило построить модель среды с учетом 3D-распределения исходных данных;
- Получено контрастное изображение областей повышенного затухания электромагнитных волн с малыми горизонтальными размерами;
- Эффект пространственной анизотропии учтён через введение коэффициента скейлинга.

14. Построение карты глубины Мохоровичича

Постановка задачи:

Построить карту глубины поверхности Мохоровичича для северной части Балтийского щита (Финляндия, Карелия)

Исходные данные:

Исходные данные: инверсия приемных функций;

Признаки: двумерные координаты;

Целевая переменная: глубина поверхности Мохоровичича;

Объем доступных измерений: 61 точка.

Метод машинного обучения:

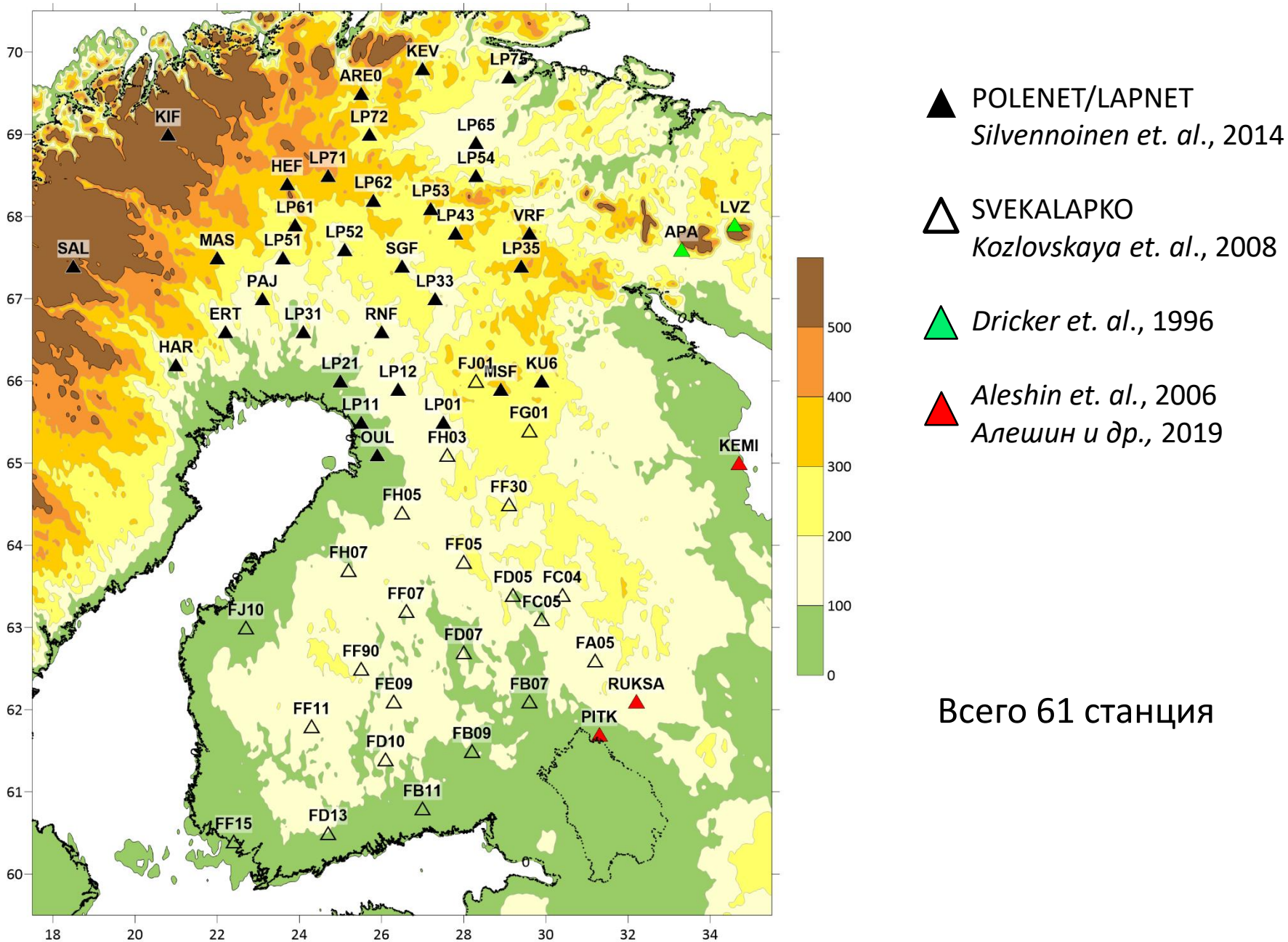
Регрессия на основе kNN

Настройка обучения и гиперпараметров:

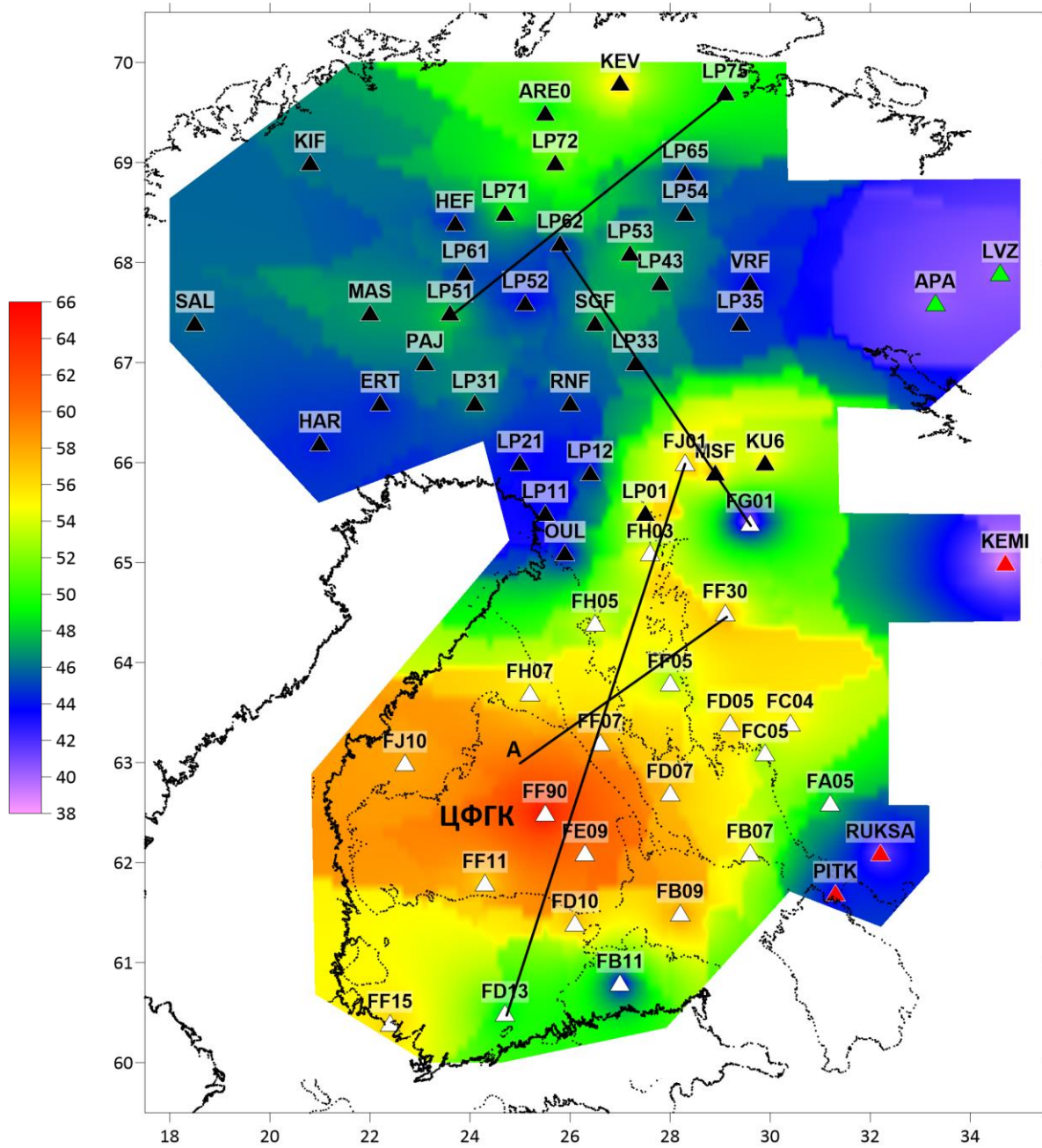
Функционал качества: MAE

Стратегия кросс-валидации: Leave-One-Out

15. Схема станций в регионе



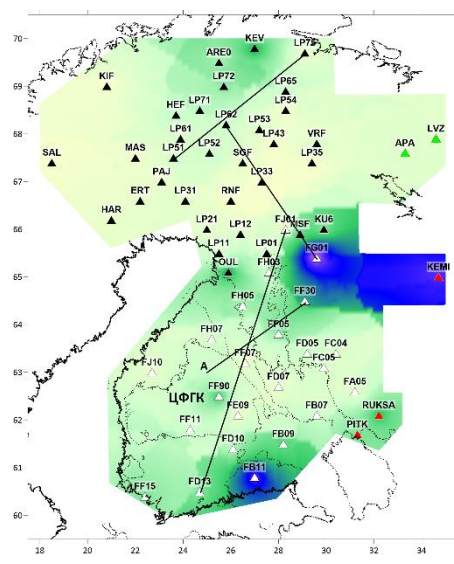
16. Карта глубины Мохоровичича



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



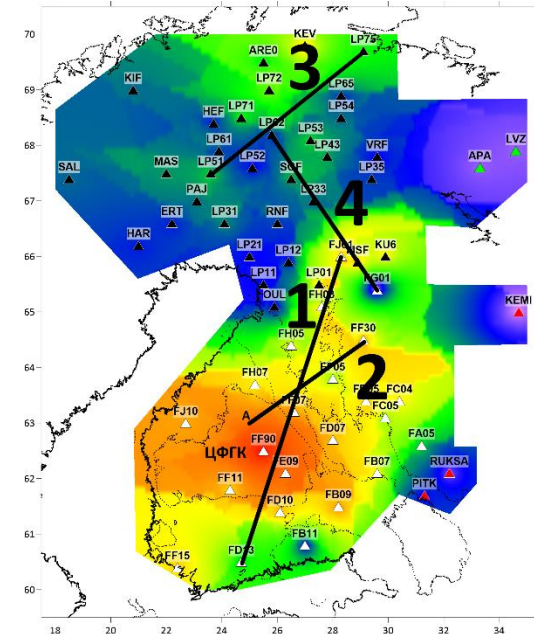
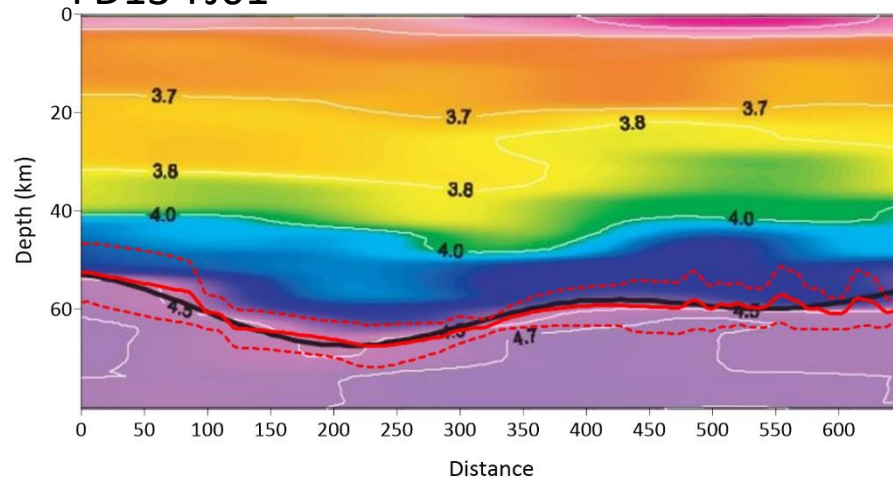
средняя ошибка
3.7 км.



17. Сравнение профилей: южная часть

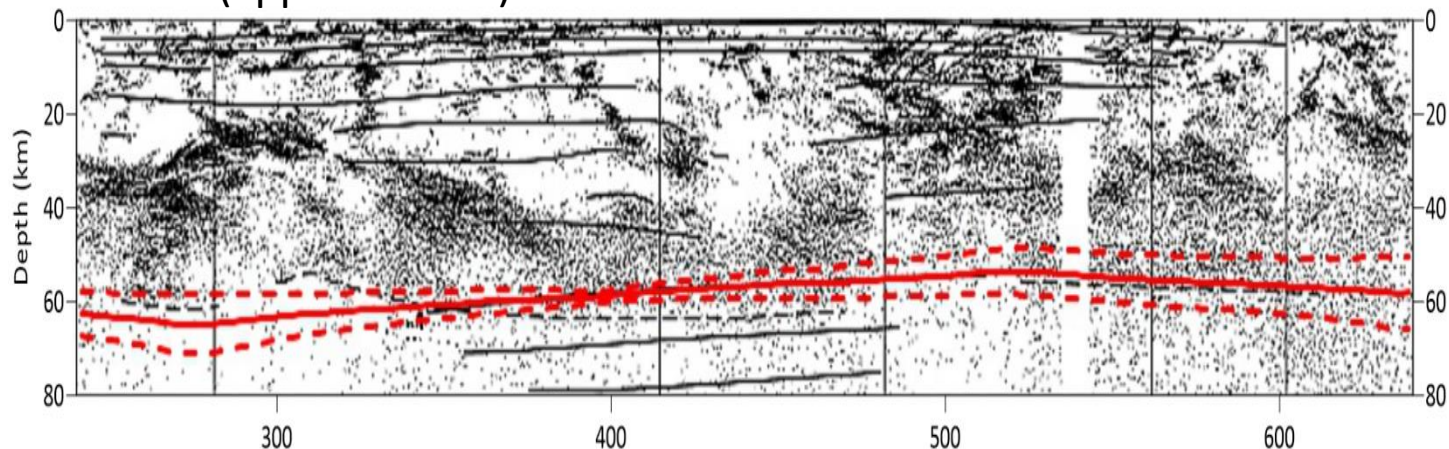
1. *Kozlovskaya et al., 2008*

FD13-FJ01



2. *Janik et al., 2007*

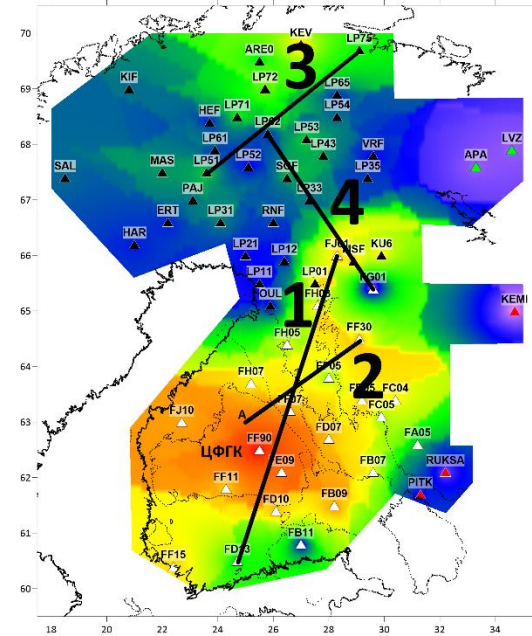
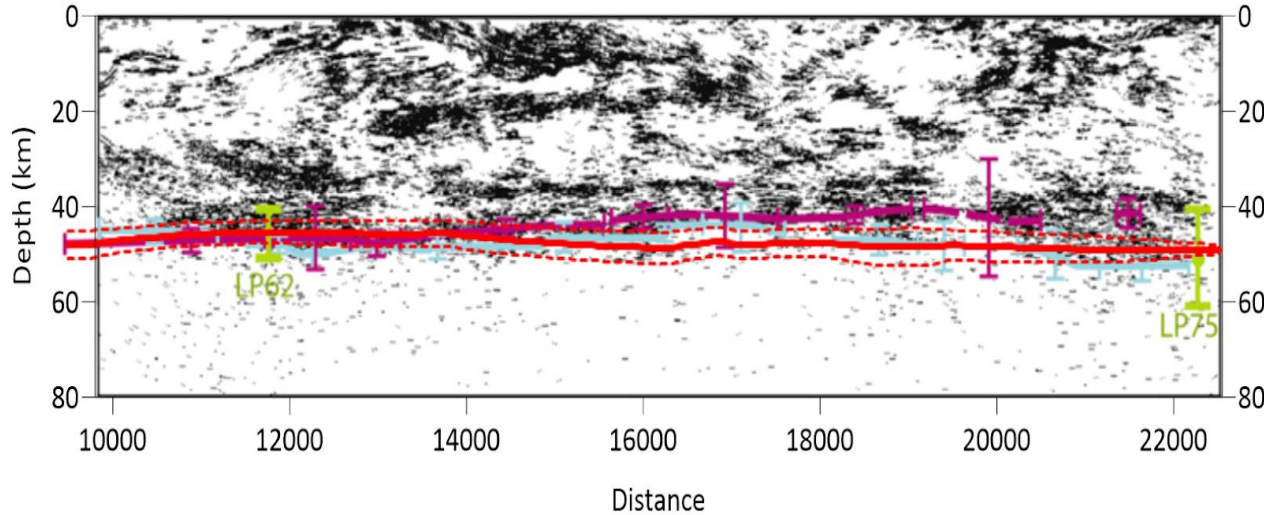
A-FF30 (approx. FIRE1)



18. Сравнение профилей: северная часть

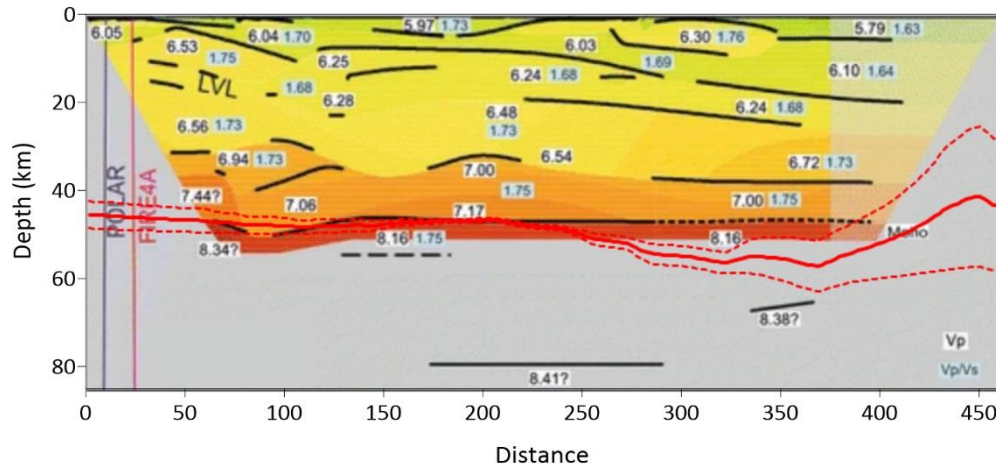
3. *Silvennoinen et. al., 2014*; *Janik et. al., 2009*; *Janik et. al., 2009*

LP51-LP75



4. *Tiira et. al., 2014*

LP62-FG01 (approx. HUKKA2007)



19. Выводы

- По данным приемных функций (S-волны) методом машинного обучения *kNN* построена карта границы Мохоровичича всего региона;
- Построенная форма границы Мохоровичича хорошо коррелирует с основными геологическими структурами в регионе;
- Построенная форма границы Мохоровичича согласуется с полученными ранее построениями, в том числе, на основе данных другого типа (Р-волны: ОГТ, ГСЗ).

20. Построение карты слоя низких скоростей V_s

Постановка задачи:

Построить карту поверхностного слоя низких скоростей поперечных сейсмических волн V_s для северной части Балтийского щита (Финляндия, Карелия)

Исходные данные:

Источник данных: данные обработки приемных функций и ГСЗ;

Признаки: двумерные координаты;

Целевая переменная: бинарная разметка (1 – наличие слоя, 0 – отсутствие);

Объем доступных измерений: 60 точек.

Метод машинного обучения:

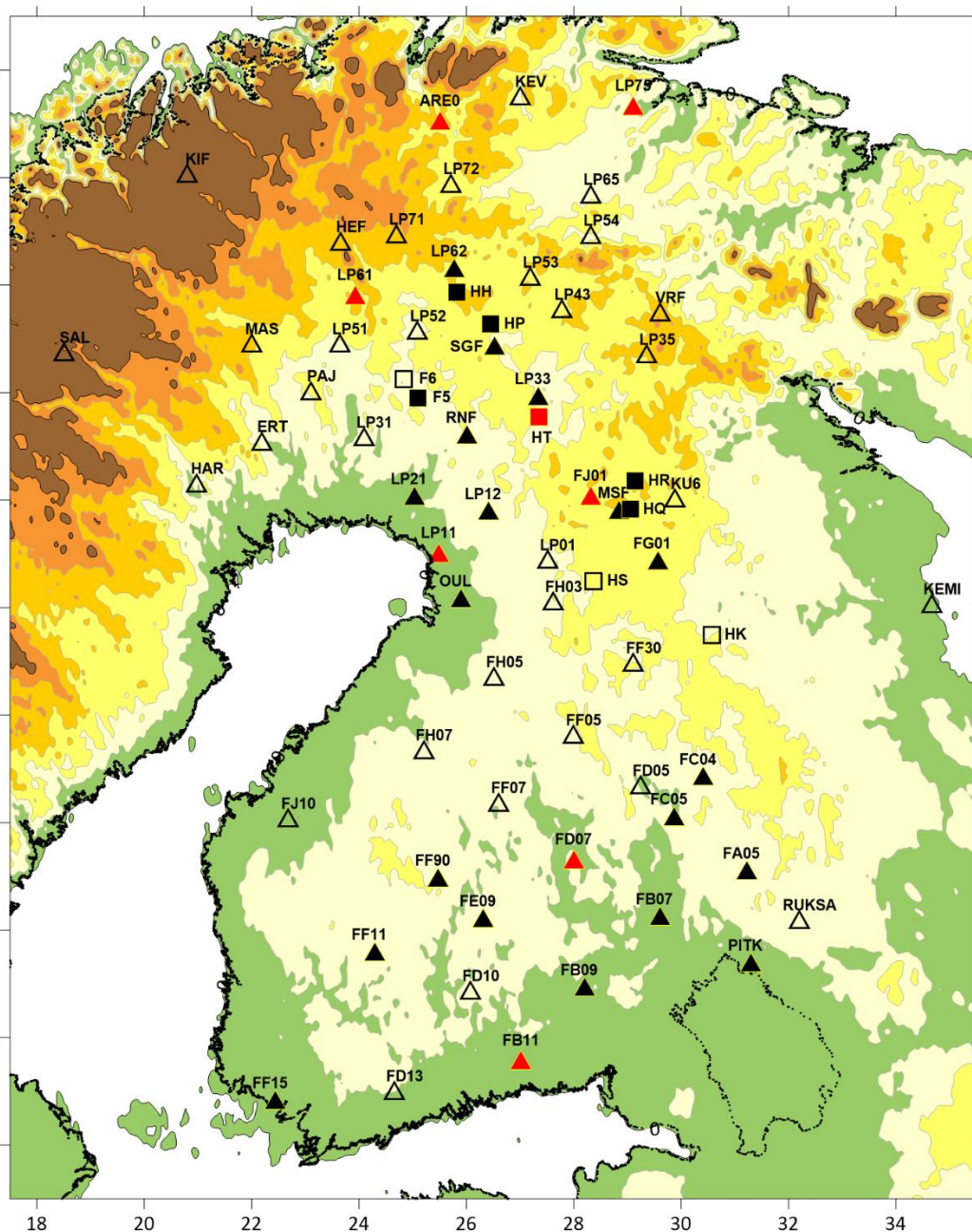
Оценка принадлежности в рамках бинарной классификации методом kNN

Настройка обучения и гиперпараметров:

Функционал качества: ROC-AUC

Стратегия кросс-валидации: Leave-One-Out

21. Разметка исходных данных



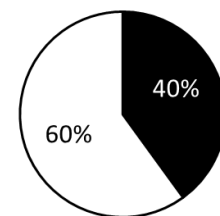
▲ Слой отсутствует
(класс 0)

△ Слой присутствует
(класс 1)

▲ Нельзя определить
(класс ?)

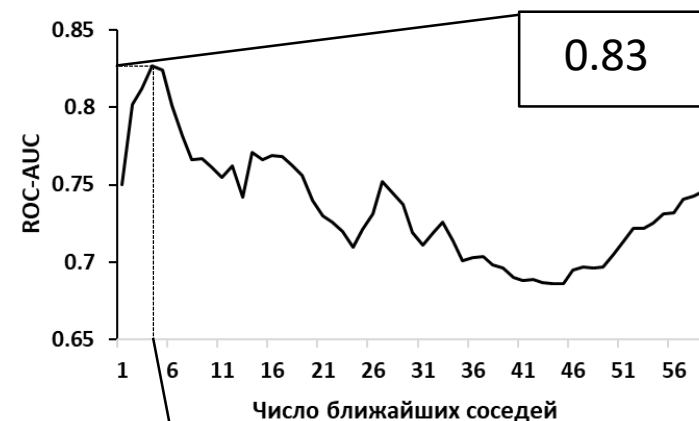
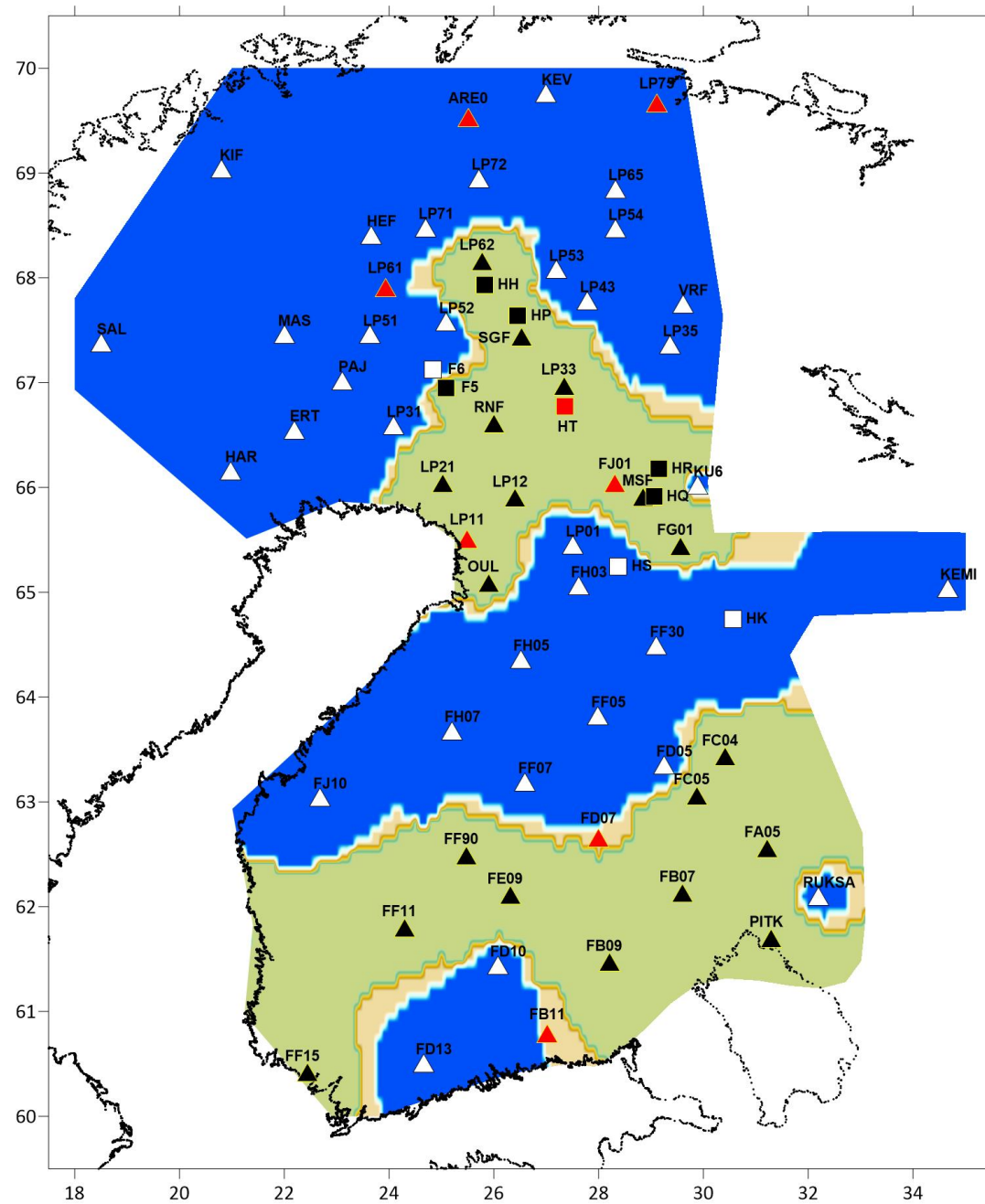
△ Метод приемных функций

□ Метод ГСЗ



Всего:
60 размеченных точек
8 не размеченных

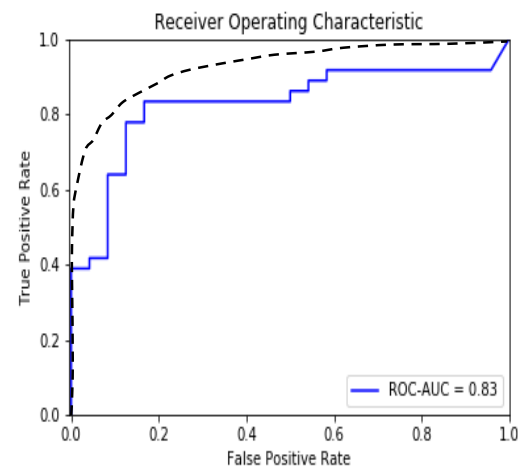
22. Карта слоя пониженной скорости



$k = 4$

Вероятность p присутствия слоя

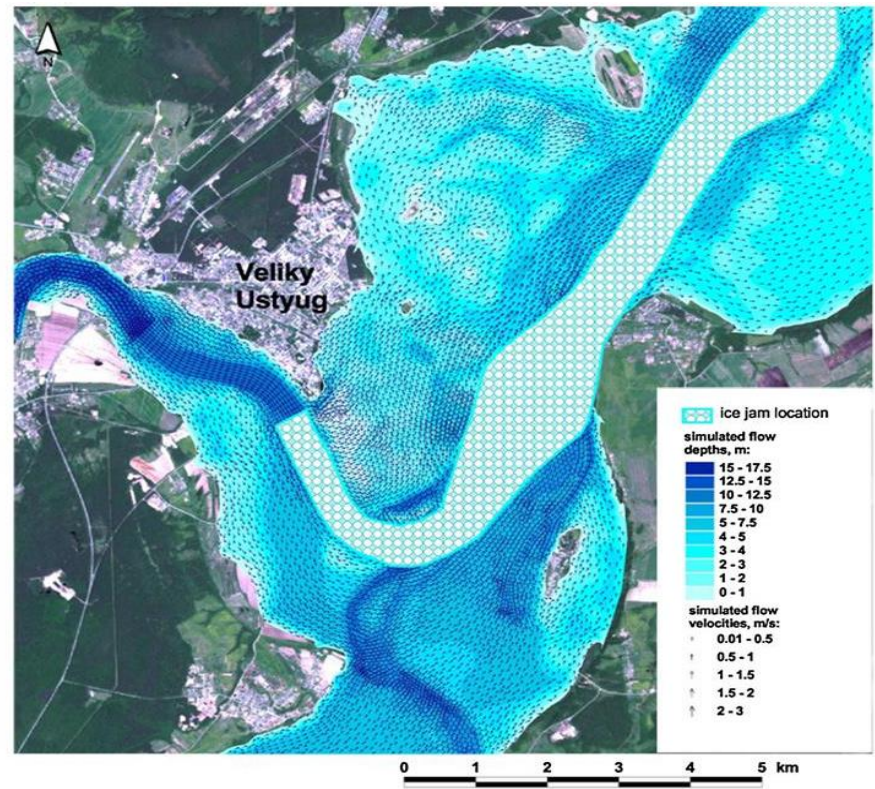
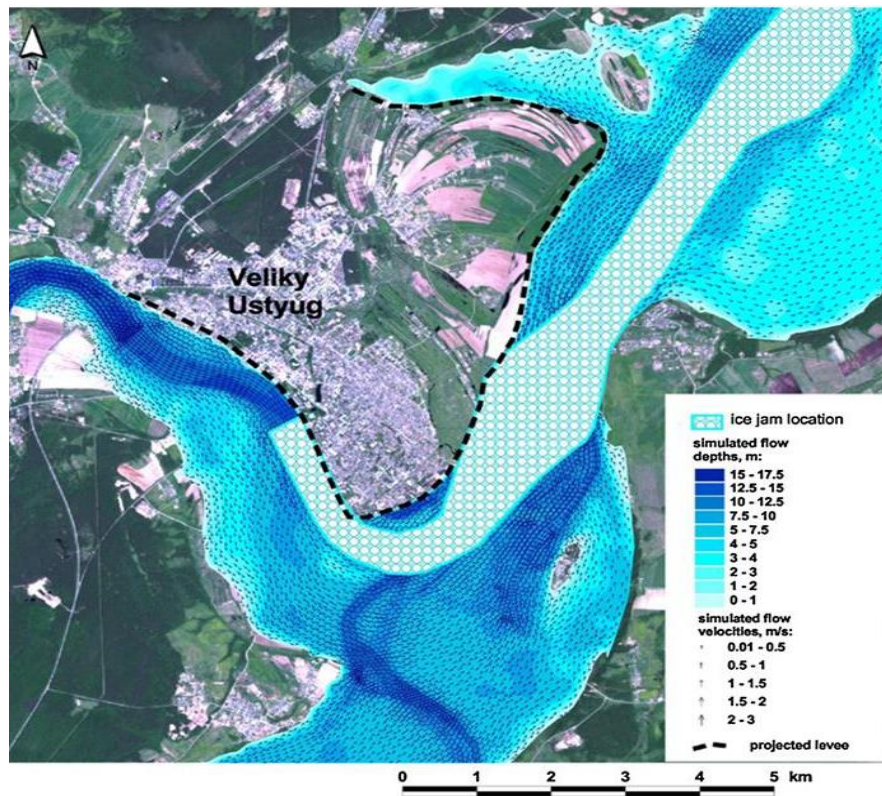
- $p < 0.45$
- $p > 0.55$
- $0.45 \leq p \leq 0.55$



23. Выводы

- Применение методов МО позволило оконтурить области, содержащие поверхностный низкоскоростной слой;
- На основе совместной интерпретации данных метода приемных функций и данных метода ГСЗ построена карта наличия поверхностного низкоскоростного слоя;
- Низкоскоростной слой в южной и северной части региона обусловлен особенностями геологического строения (юг: рапакиви, север: осадочные породы; слой в центральной части региона предположительно образовался из-за неравномерного движения в результате пост-ледникового поднятия).

24. Ледовые заторы как причина наводнений



Смоделированные глубина и скорость течения при наводнении, вызванном ледовым затором*

*S. A. Agafonova, N. L. Frolova, I. N. Krylenko, A. A. Sazonov, P. P. Golovlyov
Dangerous ice phenomena on the lowland rivers of European Russia, Natural Hazards, 2016

25. Разработка прогнозной системы заторообразования

Постановка задачи:

Прогнозирование ледового заторообразования на участке р. Северная Двина

Исходные данные:

Признаки: набор из 11 гидрометеорологических факторов;

Целевая переменная: бинарная (1 – затороопасный сезон, 0 – иначе);

Объем доступных измерений: 26 сезонов.

Метод машинного обучения:

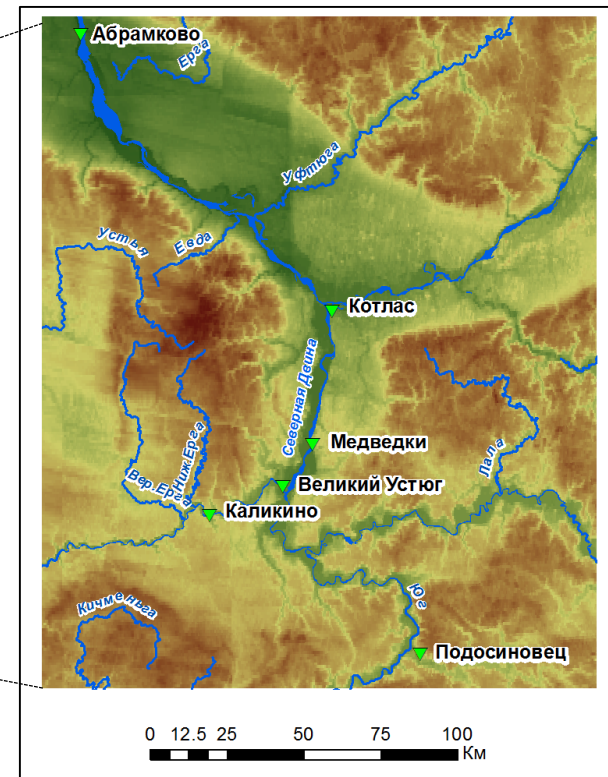
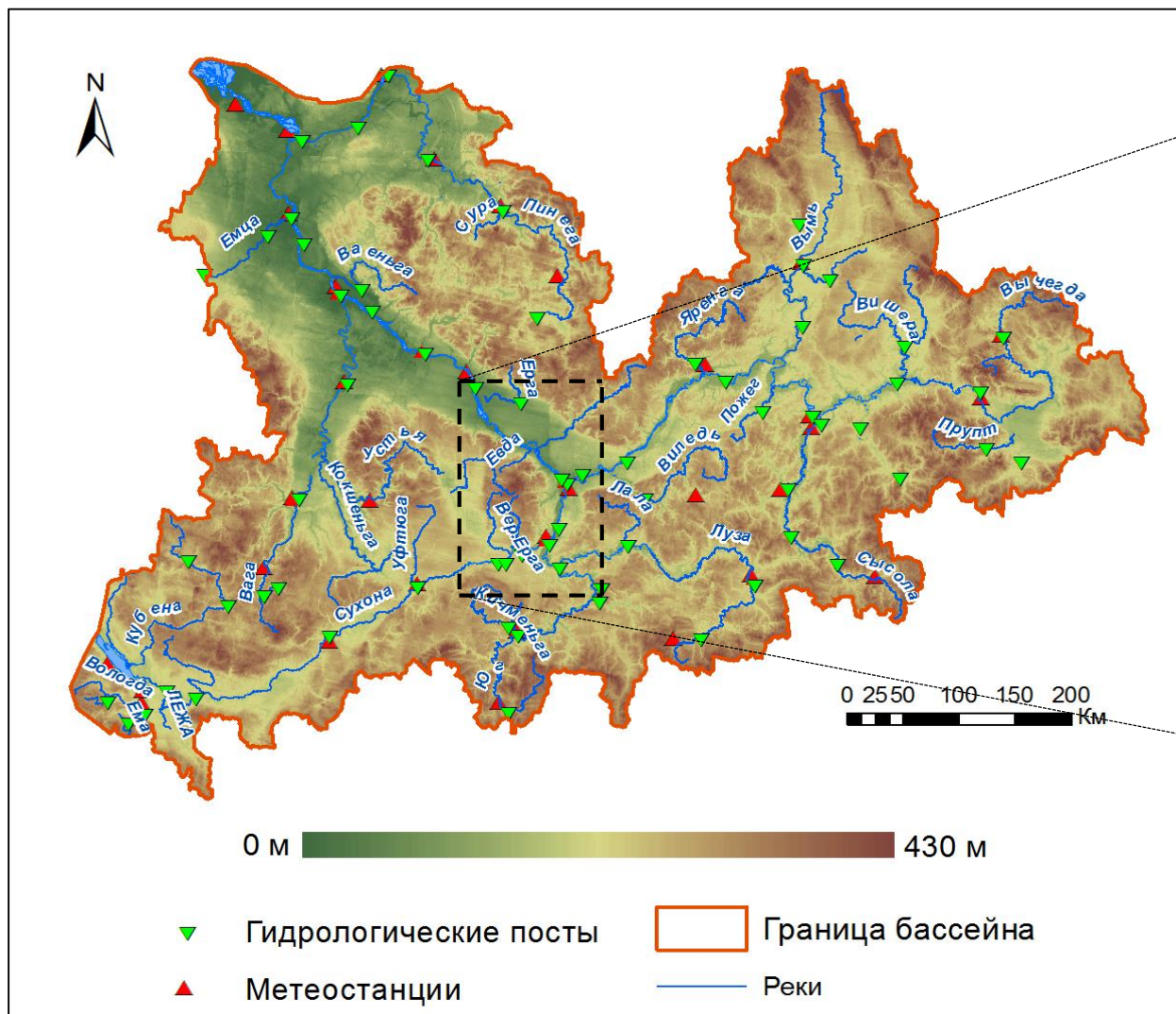
Бинарная классификация методом голосования и методом вычисления оценок

Настройка обучения и гиперпараметров:

Функционал качества: accuracy

Стратегия кросс-валидации: Leave-One-Out

26. Схема речного бассейна



27. Интегральные признаки заторообразования

№	Название признака	Единицы измерения	Мин. значение	Макс. значение
1	Предлежавший уровень воды	см.	4	322
2	Продолжительность осеннего ледохода	сут.	5	20
3	Наличие зажоров	0/1	0	1
4	Особенности температурного режима в период замерзания: факт перехода температуры воздуха через ноль	сут.	3	22
5	Сумма отрицательных температур воздуха за холодный период	°C	200	700
6	Сумма положительных температур воздуха за холодный период	°C	5	15
7	Количество дней с положительными температурами воздуха за холодный период	сут.	6	17
8	Сумма твердых осадков	мм.	20	60
9	Особенности температурного режима в период вскрытия: факт перехода температуры воздуха через ноль	сут.	4	16
10	Толщина льда перед вскрытием	см.	3	11
11	Интенсивность роста уровней и расходов воды в период подвижек	см./сут.	8	200

гидрологические признаки

метеорологические признаки

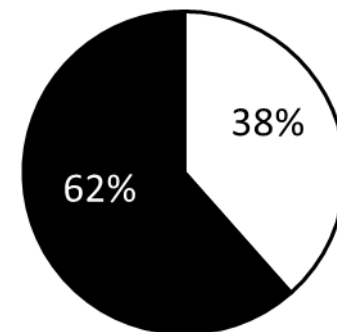
28. Классификация сезонов

Исходная выборка

1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
1	1	0	0	0	1	0	1	0	0	1	1	0	1	1	0	0	1	1	1

Дополнительная выборка

2011	2012	2013	2014	2015	2016
1	0	1	1	1	1

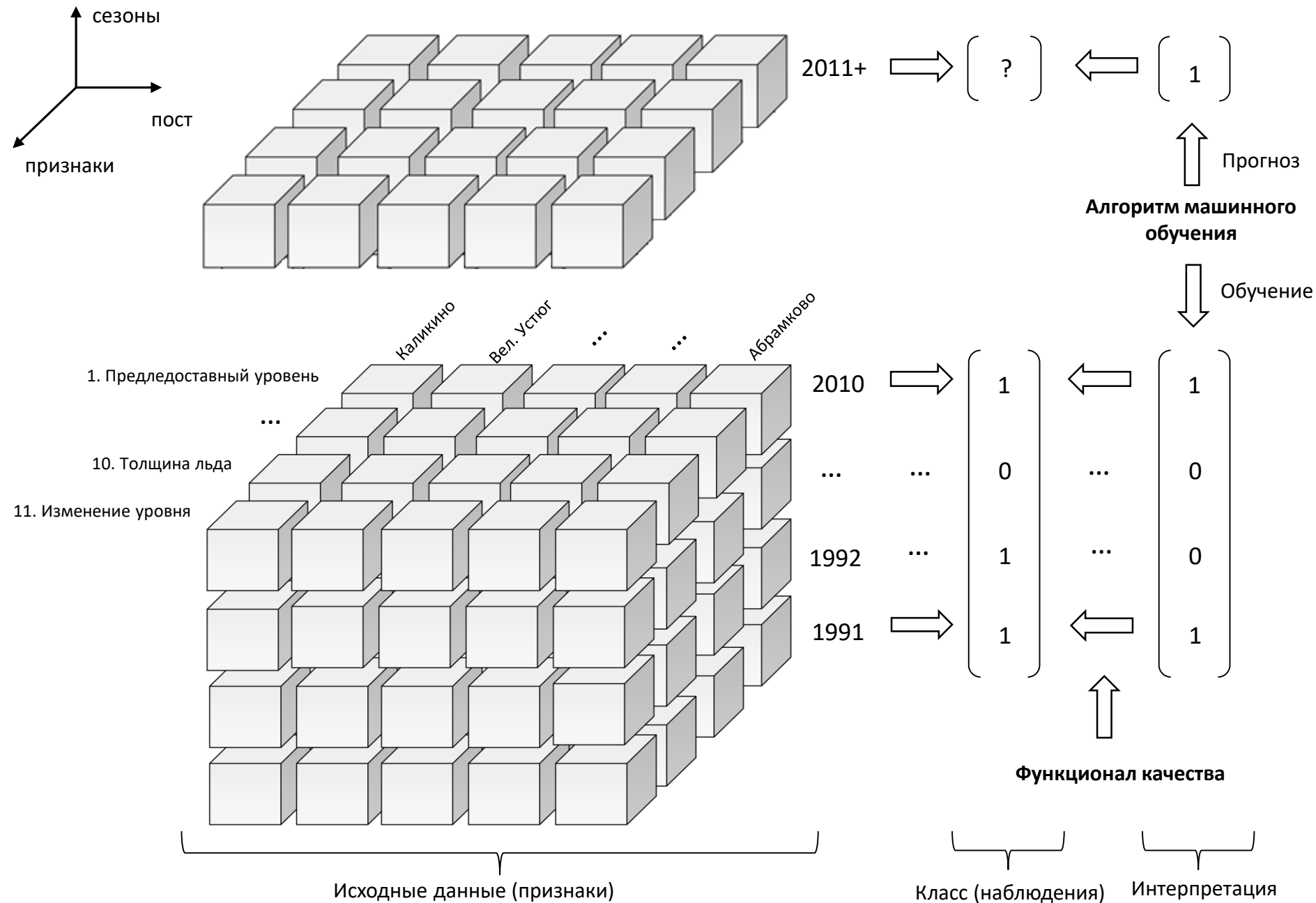


1 – наличие сильных заторов на исследуемом участке

0 – отсутствие



29. Прогнозная задача



30. Логика прогнозной системы



Алгоритм голосования

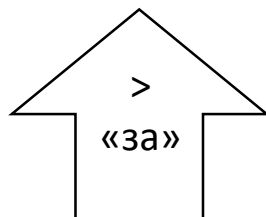
$$g(X, \Gamma_i, t) = \prod_{p=1}^n (1 - t_p \theta(B_p(X, \Gamma_i) - \delta_p))$$

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

$$F_1(X) = j_0$$

$$G_{j_0} = \max_j G_j$$

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



Алгоритм вычисления оценок

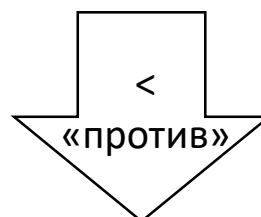
$$v(X, \Gamma_i, t) = \sum_{p=1}^n t_p \theta(B_p(X, \Gamma_i) - \delta_p)$$

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

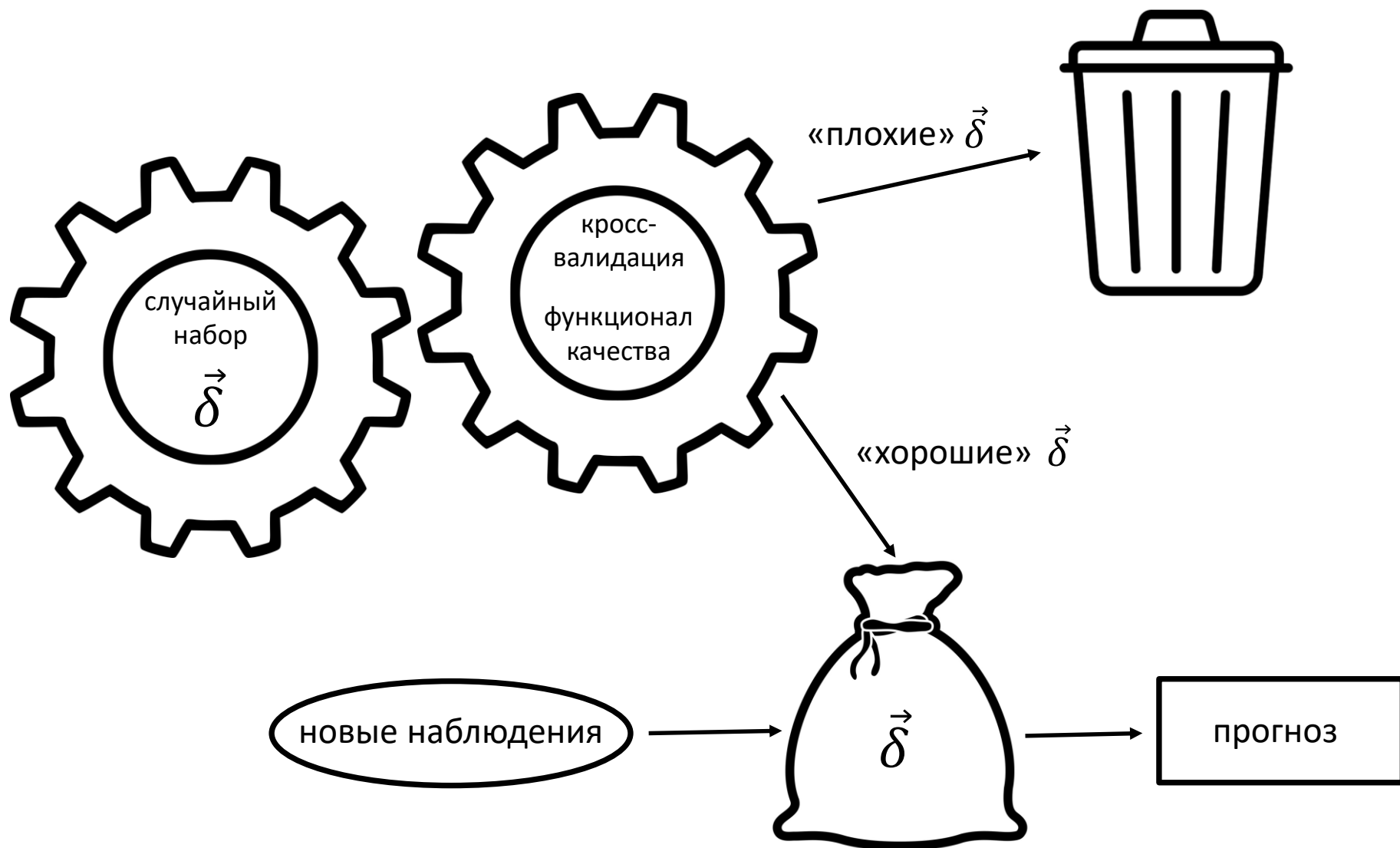
$$F_2(X) = j_0$$

$$V_{j_0} = \min_j V_j$$

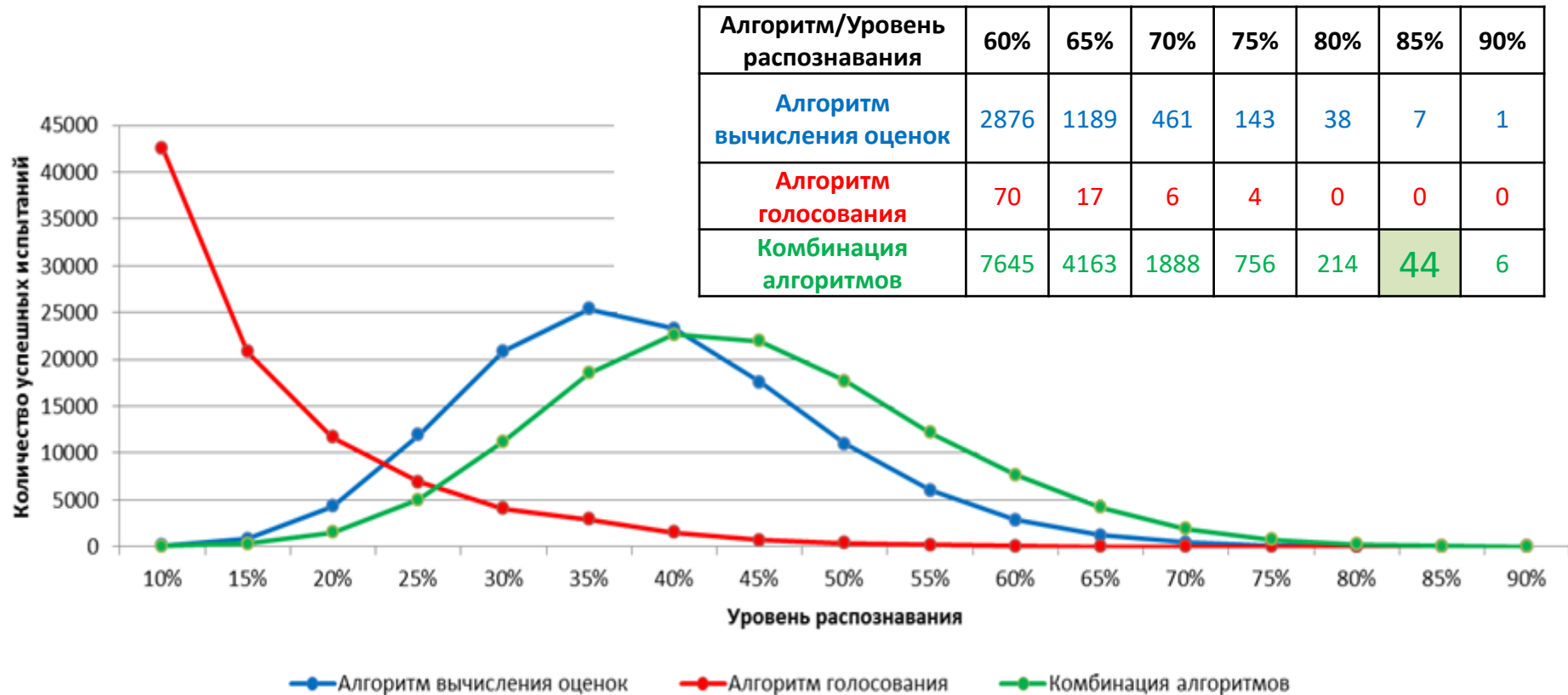
$$\theta(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}$$



31. Схема численного моделирования



32. Результат обучения системы



128 000 испытаний Монте-Карло

33. Результаты прогнозирования и валидации

Сезон	Класс	Кросс валидация	2011	2012	2013	2014	2015	2016	Весь период
1991	1	1	1	1	1	1	1	1	1
1992	1	1	1	1	1	1	1	1	1
1993	0	1	0	0	0	0	0	0	0
1994	0	0	0	0	0	0	0	0	0
1995	0	0	0	0	0	0	0	0	0
1996	1	1	1	1	1	1	1	1	1
1997	0	0	0	0	0	0	0	0	0
1998	1	1	1	1	1	1	1	1	1
1999	0	0	0	0	0	0	0	0	0
2000	0	0	0	0	0	0	0	0	0
2001	1	1	1	1	1	1	1	1	1
2002	1	1	1	1	1	1	1	1	1
2003	0	0	0	0	0	0	0	0	0
2004	1	1	1	1	1	1	1	1	1
2005	1	1	1	1	1	1	1	1	1
2006	0	1	0	0	0	0	0	0	0
2007	0	0	0	0	0	0	0	0	0
2008	1	0	1	1	1	1	1	1	1
2009	1	1	1	1	1	1	1	1	1
2010	1	1	1	1	1	1	1	1	1
2011	1		1	1	1	1	1	1	1
2012	0			0	0	0	0	0	0
2013	1				1	1	1	1	1
2014	1					1	1	1	1
2015	1						1	1	1
2016	1							1	1

1. Кросс-валидация
1991-2010:

accuracy = 0.85

2. Кросс-валидация
1991-2016:

accuracy = 0.85

3. Обучение 1991-
2010, прогноз
2011-2016:

accuracy > 0.85

Обучение

Прогноз

34. Важность признаков (сумма 100%)



35. Выводы

- На основе специализированных методов МО построена система прогноза заторообразования на участке р. Северная Двина при существенном недостатке (дефиците) данных;
- Построенная система обеспечивает качество прогноза 85%;
- Применение методов МО позволяет оценить вклад признаков в итоговый результат явления в условиях дефицита данных;
- Основной вклад вносят гидрологические и гидродинамические признаки, полученные с минимальной заблаговременностью (изменение уровней, толщина льда);
- Среди метеорологических признаков наиболее важными являются температурные характеристики.

36. Защищаемые положения

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%.

37. Научная новизна

1. Разработан новый метод интерполяции на основе базового метода машинного обучения (метод ближайших соседей). Метод допускает различные постановки задачи машинного обучения (регрессия, классификация).
2. На основе метода построена карта границы Мохоровичича Северной и Южной Финляндии по данным приёмных функций (S-волны).
3. Разработан новый метод вероятностной оценки наличия поверхностного низкоскоростного слоя поперечных сейсмических волн. Построена соответствующая карта региона. Такие слои имеются на севере, юге и в центре региона.
4. В южной и северной частях региона наличие низкоскоростного слоя обусловлено геологическим строением: на юге – рапакиви, на севере – наличие осадочного покрытия.
5. В центральной части региона нет корреляции наличия низкоскоростного слоя с возрастом. Его образование может быть связано с постледниковым поднятием.
6. Разработан новый метод построения модели проводимости среды на основе данных межскважинного электромагнитного просвечивания.
7. 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.
8. Предложенный подход позволяет учесть эффект пространственной анизотропии распределения исходных данных.
9. Для задачи анализа коротких временных рядов предложен оригинальный подход на основе специализированных методов теории машинного обучения и распознавания образов (алгоритм голосования, алгоритм вычисления оценок).
10. Разработана система краткосрочного прогнозирования ледового заторообразования для участка р. Северная Двина.
11. С помощью разработанной системы проведен анализ заторообразования за 20 сезонов и построен прогноз на предстоящие 6 сезонов. Проведена валидация построенных прогнозов. Оцененная достоверность составила 85%.
12. В условиях ограниченного набора исходных данных оценена важность вклада каждого из рассматриваемых признаков в формирование заторообразования.
13. Показано, что наибольший вклад вносит группа признаков, отражающих гидрологические и гидродинамические характеристики участка реки с наименьшей заблаговременностью (скорость изменения уровней, толщина льда перед вскрытием).

38. Список основных публикаций

- ★ ★ ★ 1) Алешин И.М., **Малыгин И.В.** Интерпретация результатов радиоволнового просвечивания методами машинного обучения // Компьютерные исследования и моделирование. — 2019. — Т. 11, № 4. — С. 675–684.
- ★ ★ 2) 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.
- ★ ★ ★ 3) 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.
- ★ ★ 4) Алешин И.М., **Малыгин И.В.** Верификация экспертной системы прогноза заторообразования на Северной Двине // Геофизические процессы и биосфера. — 2018. — Т. 17, № 2. — С. 48–60.
- ★ 5) **Малыгин И.В.** Логический подход к созданию экспертных систем прогнозирования опасных природных явлений // Естественные и технические науки. — 2015. — № 2. — С. 102–112.
- ★ 6) **Малыгин И.В.** О задаче прогнозирования ледовых заторов // Интеллектуальные системы. Теория и приложения. — 2014. — Т. 18, № 3. — С. 73–80.
- ★ ★ ★ 7) **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.

Свидетельства о регистрации прав на ПО

8) **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.

9) **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.

10) **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.

Спасибо за внимание!