Отчет

«Прогноз данных в межпрофильном пространстве с применением методов машинного обучения»

Выполнил: Зайцев Сергей Владимирович

Содержание

1. Постановка задачи

1.1 Общее описание проблемы

Целью данного проекта является восстановление структурного каркаса (геолого-геофизических границ, которые в подавляющем большинстве случаев определяются данными сейсморазведки) по данным потенциальных методов геофизики (гравиразведки и магниторазведки).

Актуальность данной задачи основывается на более высокой производительности несейсмических методов по сравнению с сейсморазведкой, а также меньшей стоимостью при сопоставимой детальности съемки.

Структурные границы, выделяемые по результатам сейсморазведки, являются границами раздела, на которых происходит скачок акустической жесткости. Акустическая жесткость является произведением скорости упругих волн на плотность и, как следствие, изменения в глубине структурных границ могут проявляться и в гравитационном поле. Поскольку источники аномалий гравитационного поля, располагающиеся на различных глубинах, проявляются в различных пространственных частотах аномалий гравитационного поля, значимыми для прогноза могут быть не только исходные аномалии силы тяжести, но и их высокочастотные и низкочастотные компоненты.

Кроме того, геологические слои, разделяемые структурными границами, также могут характеризоваться различными магнитными свойствами, поэтому аномалии магнитного поля и их трансформанты также могут считаться значимыми при прогнозировании глубин структурных границ.

Формальная постановка задачи заключается в следующем: имеется набор сейсморазведочных профилей (не обязательно регулярно расположенных) на которых определена глубина до структурного горизонта, требуется восстановить положение структурного горизонта в пространстве между профилями сейсморазведки (межпрофильное пространство). Классическое решение – методы интерполяции (обычно, методом Kriging). Такое решение не позволяет учитывать геологические особенности строения в межпрофильном пространстве.

Одним из возможных решений поставленной задачи является выявление корреляционных зависимостей между потенциальными полями и глубинами исследуемой границы. Однако, потенциальные поля нелинейно связаны с формой аномалообразующих объектов, в связи с чем в качестве приоритетного подхода к решению поставленной задачи выбрано машинное обучение, с помощью которого могут выявляться как линейные, так и более сложные зависимости между признаками (потенциальными полями) и целевыми значениями (глубинами границ).

Изменения глубин структурных границ может находить свое отражение в потенциальных полях и, следовательно, задача восстановления структурных границ на основе потенциальных полей может решаться с помощью алгоритмов машинного обучения.

Основной задачей данного проекта являлось распространение информации о глубинах структурного каркаса на всю территорию, где имеются данные несейсмических методов, на основе значений глубин границы, выделенных по результатом сейсморазведочных работ с неравномерным покрытием территории (Рисунок 1).

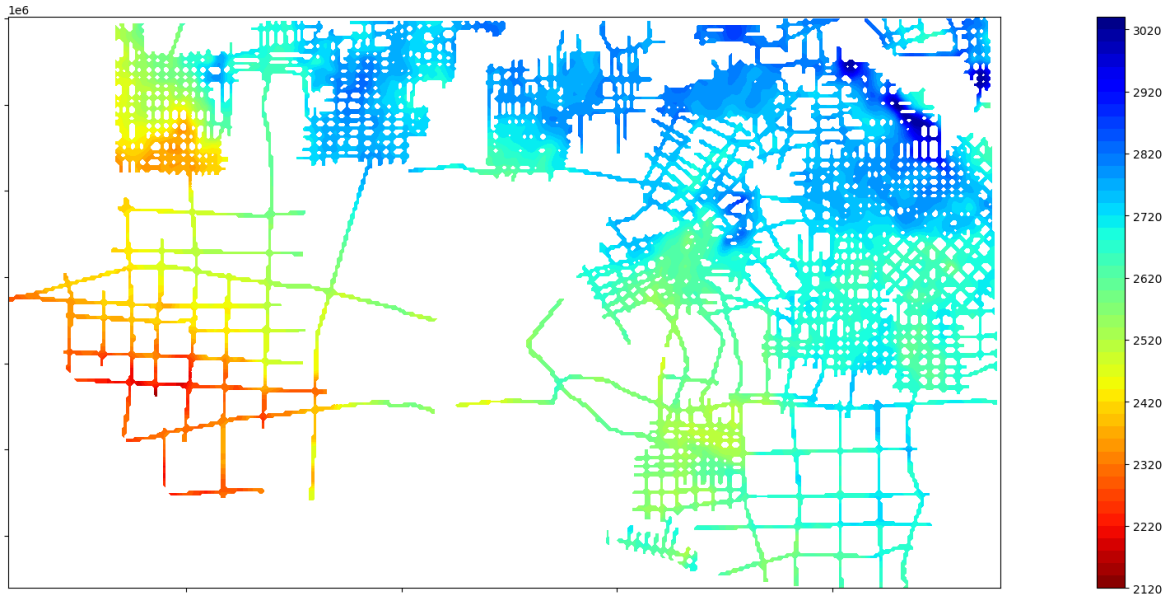

Рисунок 1 Карта глубин структурного каркаса, выделенных по данным сейсморазведки

Поскольку глубина границ является вещественным числом, задача восстановления структурных поверхностей на основе потенциальных полей может рассматриваться как задача регрессии. Для решения задачи регрессии можно выделить следующие алгоритмы машинного обучения:

- Линейная регрессия (LR);
- Линейная регрессия с L1 (Lasso), L2 (Ridge) или одновременно с L1 и L2 (ElasticNet) регуляризацией;
- Метод опорных векторов (SVM) с расширением пространства признаков на основе kernel trick;
- Метод k-ближайших соседей;
- Регрессия на основе гауссовского процесса (GPR – gaussian process regression);
- Метод случайного леса (RF - random forest);
- Метод градиентного бустинга (GB – Gradient boosting);
- Полносвязные нейронные сети (FCNN – fully connected neural network);
- Сверточные нейронные сети (CNN – convolutional neural network).

Для реализации большинства алгоритмов и методов использовались их реализации из библиотеки Scikit-Learn, которая является одной из наиболее популярных библиотек для машинного обучения языка программирования Python. Также для моделей градиентного бустинга помимо реализации из Scikit-Learn (далее GB), использовались также реализации из библиотек XGBoost (XGB) и CatBoost (CGB). Для построения полносвязных и сверточных нейронных сетей использовалась библиотека TensorFlow.

2. Результаты анализа исходных данных

2.1 Источники данных

Исходные данные, представляющие собой набор «грид» файлов формата Surfer 6 ASCII Grid. Исходный набор данных включал в себя:

- Аномалии гравитационного поля в редукции Буге;

- Аномалии магнитного поля;

- Низкочастотные и высокочастотные трансформанты аномалий гравитационного и магнитного поля, вычисленные с помощью двумерного фильтра Баттерворта 6-го порядка (ширина фильтра 1, 2, 5, 10, 15, 25 км)

- Первая и вторая вертикальная производная гравитационного и магнитного поля;

- Полный горизонтальный градиент гравитационного и магнитного поля. (Рисунок 2).

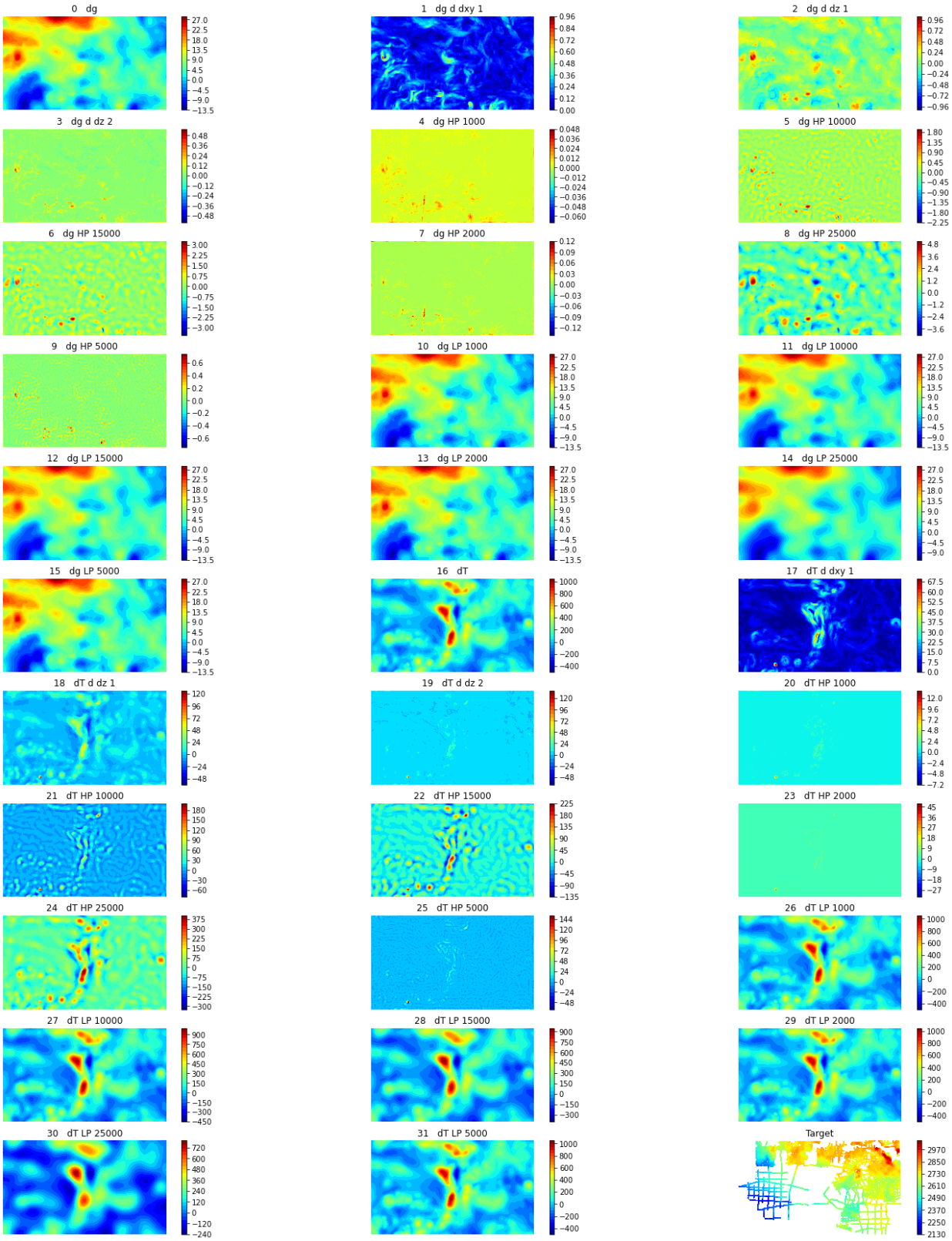

Рисунок 2 Карты потенциальных полей и их трансформант: dg – гравитационное поле, dT – магнитное поле, HP – высокочастотные трансформанты, LP – низкочастотные трансформанты, d dz – вертикалньая производная, d dxy – модуль горизонтального градиента.

Для анализа взаимосвязи между всеми данными были построены диаграммы рассеяния и корреляционная матрица.

По распределению точек на диаграммах рассеяния можно сделать вывод, что в областях, в которых отсутствует информация о глубинах границы, имеются точки, которые находятся за пределами распределения значений в обучающей выборке. В таких точках качество прогнозов может быть ниже по сравнению с точками, схожими с точками из обучающей выборки. Также на основе анализа диаграмм рассеяния можно сделать выводы об отсутствии простых функциональных связей между признаками и целевыми значениями, что также подтверждается относительно малыми значениями модуля коэффициента корреляции между признаками и целевыми значениями (не более 0.247).

Также стоит отметить, что перед дальнейшим использованием данных для машинного обучения была выполнена их нормировка путем вычитания из каждого соответствующего признака его среднего значения и деления результата на среднеквадратическое отклонение.

3. Описание решения задачи

3.1 Общая информация

3.1.1 Логика решения

Для решения задачи восстановления структурного каркаса по данными несейсмических методов были поставлены и решены следующие подзадачи:

- Анализ данных и предварительная отбраковка зависимых признаков;
- Оценка работы алгоритмов машинного обучения на обучающей выборке для предварительной оценки возможности их дальнейшего использования;
- Оценка качества прогнозов моделей на тестовой выборке;
- Оценка важности признаков по методу Шепли и корреляционным методом.

3.1.2 Общие метрики

Во всех экспериментах в качестве первостепенных метрик при ранжировании алгоритмов используются R^2 и RMSE, MPE рассматривается в качестве второстепенной метрики. Однако, как будет понятно позднее, MPE – крайне неэффективная метрика.

3.1.3 Сверточные нейронные сети – ограничения

Для решения задач проекта также была использована сверточная нейронная сеть архитектуры U-Net. Поскольку U-Net для своих прогнозов использует площадное распределение данных и в результате прогноза строит матрицу значений, имеющую такой же размер, как и матрица исходных данных (например, по участку с потенциальными полями размера 10x10 ячеек прогнозируется участок структурного каркаса размером 10x10 ячеек), при использовании U-Net возникло ограничение – чтобы не потерять информацию о глубинах границ, где располагается наиболее редкая сеть наблюдений, максимальный размер разбиения составляет 4x4 ячейки (1x1 км). Также такая особенность делает невозможным прореживание сетки в обучающей выборке.

Для устранения данного ограничения в рамках проекта была разработана модифицированная архитектура U-Net, представляющая собой первую половину исходной архитектуры и прогнозирующая одно значение глубины по участку с несейсмическими данными (например, по участку с потенциальными полями размера 11x11 ячеек прогнозируется одно

значение глубины в центральной ячейке). В последующем тексте данная архитектура будет именоваться как U-Net ½.

3.1.4 Подготовка данных

Подготовка в данном случае заключается в импорте данных из грид-файлов и формирование из них таблицы (для всех моделей, кроме сверточных нейронных сетей). Для сверточных нейронных сетей подготовка данных заключается в формировании двумерных матриц по каждому признаку, а также двумерных матриц целевых значений для нейронной сети U-Net и единичных целевых значений в центральной точке для нейронной сети U-Net ½.

3.2 Описание решения

По результатам обучения (*Таблица 1*) можно предположить, что модели GPR и GB переобучаются, так как они показывают нулевую ошибку на обучающей выборке. Из оставшихся моделей лучшими оказались нейронные сети и случайный лес, худшими оказались линейные алгоритмы. Стоит отметить, что линейные алгоритмы, использующие L1 регуляризацию, выдают константные прогнозы с параметрами регрессии по умолчанию.

Таблица 1 Оценка качества работы моделей машинного обучения на обучающей выборке

Алгоритм или архитектура	**R^2**	**RMSE (м)**	**MPE (%)**
GPR	1.00	0.00	100.0
GB	1.00	0.00	100.0
U-Net ½	1.00	4.64	99.88
RF	1.00	5.02	99.93
U-Net	1.00	6.46	99.87
FCNN	0.98	20.19	99.43
XGB	0.94	33.82	99.07
CGB	0.88	46.24	98.70
KNN	0.88	47.22	98.96
SVM	0.80	60.72	98.44
LR	0.21	120.29	96.36
Ridge	0.21	120.36	96.36
Lasso	0.00	135.57	95.97
ElasticNet	0.00	135.57	95.97

Для тестирования геопространственных алгоритмов на тестовой выборке не подходят стандартные методы оценки (например, K-fold), так как признаки и целевая функция распределена в пространстве и возможны потери части данных при исключении участка для тестирования.

Для этого был разработан следующий метод: карта разбивается случайным образом по регионам размерам 5x5 км, т.е. предварительно вся карта разбивалась на квадратные непересекающиеся участки размером 5x5 км, после чего для каждого такого участка случайным образом задавался номер одной из 5 частей выборки (*Рисунок 3*). В результате формируется 5 комбинаций карт с обучающей и валидационной выборкой. С помощью такого подхода искусственно моделируется отсутствие данных в межпрофильном пространстве.

Рисунок 3 Пример распределения пяти частей выборки при кросс-валидации по регионам

После проведения обучения и тестирования на тестовой выборке из начального списка исключены линейная регрессия без регуляризации и с регуляризацией, регрессия на основе гауссовского процесса (GPR), а также реализация градиентного бустинга из библиотеки Scikit-Learn (GB). Лучшими по результатам кросс-валидации по регионам стали сверточные нейронные сети.

Последний шаг в тестировании и выборе моделей – подбор гиперпараметров. Для большинства моделей качество прогнозов с оптимальными гиперпараметрами оказывается лучше, чем с параметрами по умолчанию (*Таблица 2*). Исключением является метод опорных векторов, для которого оптимальные гиперпараметрами совпали с исходными, а также метод k-ближайших соседей, для которого количество соседей было уменьшено с 5 до 3. Также стоит отметить незначительное изменение качества прогнозов для модели RF, по которой при подборе гиперпараметров также не наблюдалось существенного улучшения результатов. Для всех остальных моделей улучшение качества прогнозов составило 4 - 5 м, за исключением модели CGB, для которой среднеквадратическая ошибка прогнозов уменьшилась на 15.4 м.

Таблица 2 Результаты оценки качества прогнозов на тестовой выборке до и после подбора гиперпараметров

Алгоритм или архитектура	**RMSE с гиперпараметрами по умолчанию**	**RMSE с оптимальными гиперпараметрами**	**Разница («до» - «после»)**
U-Net ½	47.0	42.2	4.8
U-Net	63.8	58.9	4.9
FCNN	66.3	61.5	4.8
RF	80.4	80.1	0.3
XGB	85.8	81.6	4.2
SVM	86.9	86.9	0
CGB	90.7	75.3	15.4
KNN	93.6	94.8	-1.2

По итоговым значениям всех используемых в проекте метрик, вычисленным на основе прогнозов на тестовой выборке, можно сделать вывод, что с текущим набором трансформант модель U-Net ½ является наиболее подходящей для дальнейшего применения в проекте. Также близкие к «хорошим» значениям метрик показывают U-Net и FCNN.

Таблица 3 Итоговые значения метрик на тестовой выборке

Алгоритм или архитектура	**RMSE**	**R2**	**MPE**
U-Net ½	42.2	0.91	99.16
UNet	58.9	0.83	98.89
FCNN	61.5	0.82	98.87
CGB	75.3	0.72	98.28
XGB	80.1	0.69	98.37
RF	81.6	0.68	98.13
SVM	86.9	0.63	97.67
KNN	94.8	0.56	98.05

3.2.1 Оценка важности признаков

Задача оценки важности признаков заключается в оценке вклада каждого признака в итоговый прогноз и ранжирование признаков исходя из их вклада. Оценка важности признаков осуществлялась с помощью вычислений значений Шепли (библиотека SHAP), которые выражаются для каждого объекта и i-го признака следующим образом:

$$\varphi_i(p) = \sum_{i=1}^{N} \frac{|S|!(N!-|S|-1)!}{N!}(p(S \cup S_i) - p(S)),$$

где $\varphi_i(p)$ – значение Шепли для i-го признака для прогноза p, S – набор признаков без i-го, S_i – i-й признак, $p(S \cup S_i)$ – прогноз модели на признаках наборе признаков, включающем i-й, $p(S)$ – прогноз модели на признаках наборе признаков, не включающем i-й.

3.2.2 Постановка математической подзадачи и выбор метрик

Для вычисления значений Шепли используется предварительно обученная модель, а также некоторый набор данных, для которых происходит оценка влияния каждого признака. Для оценки важности признаков на основе значений Шепли использовалась модель градиентного бустинга CGB с оптимальными гиперпараметрами, поскольку процедура обучения и прогноза для данной модели осуществляется значительно быстрее, чем для сверточных и полносвязных нейронных сетей.

3.3 Оценка возможности экстраполяции данных

3.3.1 Постановка математической подзадачи и выбор метрик

Оценка возможности экстраполяции данных подразумевает обучение модели на данных с некоторой территории и прогнозе целевых значений (глубины структурного каркаса) за пределы обучающей территории.

Для оценки результатов данного эксперимента использовалась метрика RMSE.

3.3.2 Подготовка данных

Для данного эксперимента перед обучением модели машинного обучение происходило исключение данных по четверти территории, расположенной в одном из углов, а исключенные данные выступали в качестве тестовой выборки. Такая процедура повторялась для каждого угла. В остальном процедура подготовки данных идентична процедуре, которая рассматривалась в предыдущих экспериментах.

3.3.3 Описание решения

Оценка возможности экстраполяции осуществлялась для модели U-Net ½, показавшей лучшие результаты в предыдущих экспериментах. Оценка возможности экстраполяции производилась путем исключения четверти территории, расположенной в одном из углов территории (*Рисунок 4*), из обучающей выборки и последующем прогнозе глубины структурного каркаса на этой территории. Эксперимент был проведен для каждого из углов территории.

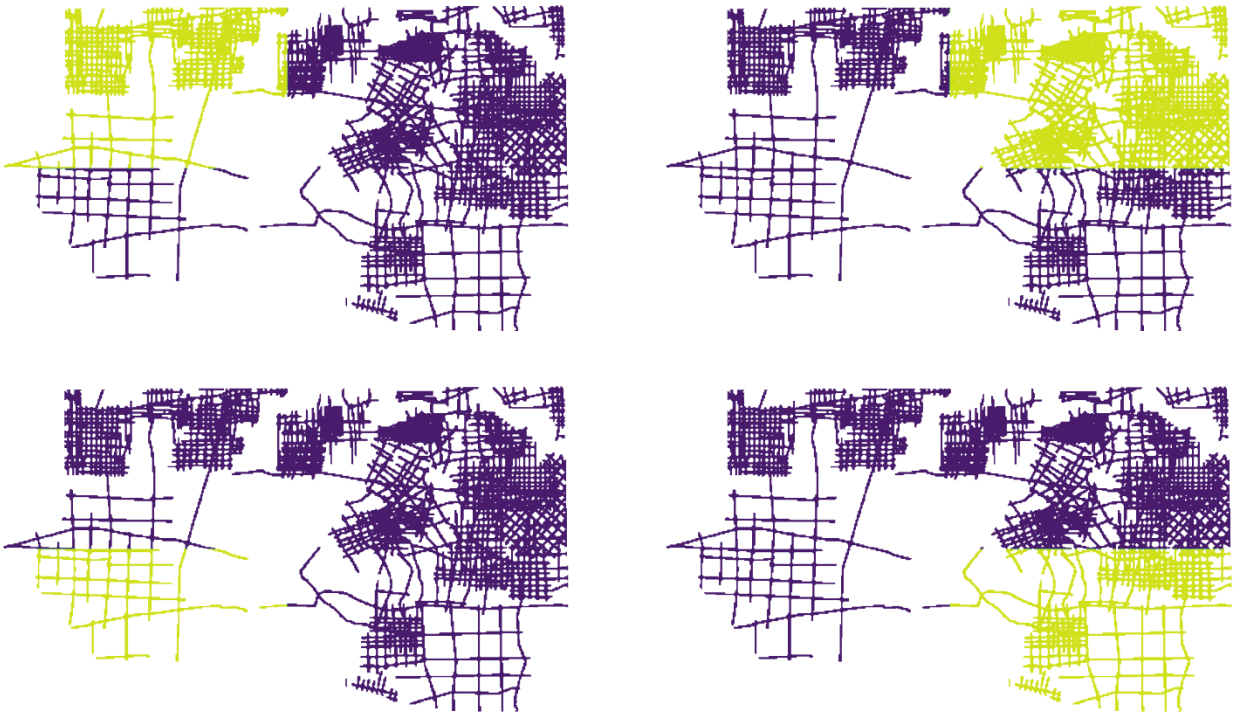

Рисунок 4 Распределение обучающей (фиолетовый цвет) и валидационной (желтый цвет) для оценки возможности экстраполяции данных

Стоит отметить, что распределение глубин структурного каркаса имеет существенную зональность, в частности для юго-западного угла территории характерны наименьшие глубины границы (до 2120 м), а для северо-восточного – наибольшие глубины (до 3032 м) (**Ошибка! Источник ссылки не найден.**, **Ошибка! Источник ссылки не найден.**).

Максимальная среднеквадратическая ошибка прогноза в данном эксперименте составила 223.2 м и характерна для юго-западного угла, минимальная – 119.9 и характерна для юго-восточного угла территории (*Таблица 4*). Стоит отметить, что при исключении областей с малыми глубинами границы (юго-западный угол), в таких областях происходит завышение глубин границ при прогнозе, а при исключении областей с большими глубинами границы (северо-восточный угол), в таких областях происходит занижение глубин границ при прогнозе (*Рисунок 5*).

Таблица 4 Значения RMSE при экстраполяции на разные углы территории

Угол территории	RMSE (м)
Юго-западный	223.2
Юго-восточный	119.9
Северо-западный	162.6
Северо-восточный	120.8

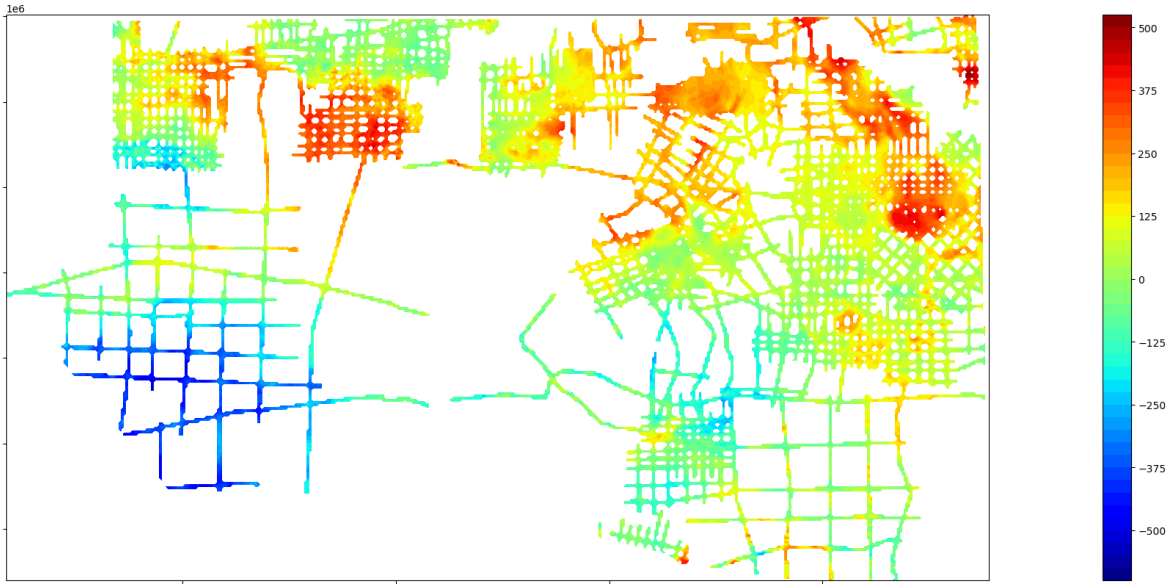

Рисунок 5 Разность между истинными и спрогнозированными глубинами структурного каркаса при экстраполяции

По результатам данного эксперимента можно сделать вывод, что качество прогнозов на внешней территории значительно уступает качеству в межпрофильном пространстве, однако максимальная величина ошибки не превысила 10% от среднего значения глубины структурного каркаса.

3.3.4 Выводы

Результаты данного эксперимента показывают значительное ухудшение качества прогнозов при экстраполяции данных (RMSE от 119.9 до 223.2 м) по сравнению с построением прогнозов в межпрофильном пространстве (RMSE на тестовой выборке 42.2 м). Это связано с особенностью прогнозирования геопространственных данных.

4. Заключение

По результатам оценки итоговых метрик можно сделать вывод, что с исходным набором трансформант модель U-Net ½ полностью удовлетворяет высокому уровню метрик, а модели U-Net, CGB и FCNN также показывают хорошие результаты. Основная область применений – интерполяция горизонтов в межпрофильном пространстве.