

# Постановка проблемы

→ ПАО «Газпром нефть» при выполнении Региональных Исследовательских работ получает для оценки и анализа данных широкий перечь данных об исследуемых и потенциально интересных месторождениях. Часть данных поступает на бумажных носителях: это ретро-карты, содержащие информацию о глубине залегания целевых (карта кровли) геологических слоев.

**Карта глубины залегания** – это представление земной коры с цветовым кодированием глубины залегания

## Данные на бумажных носителях:

- недоступны для импорта в ГИС-системы (JPG/PDF-сканы)
- требуют ручной оцифровки (несколько часов на одну карту)
- теряют информацию при субъективном переводе в цифровой формат

# Цель и задачи



**Цель исследования:** Разработать подход оцифровки геологических данных для преобразования неструктурированных данных в машиночитаемый формат.

## Задачи:

1. Определить особенности исходных геологических данных и предварительно их обработать;
2. Разработать алгоритм выделения цветовой палитры и распознавания глубинных отметок;
3. Сопоставить цвету глубину залегания и реализовать назначение глубин залегания для каждого пикселя;
4. Провести верификацию и оценку метрик качества оцифрованных данных.

# Обзор аналогичных решений/исследований

Подходы к оцифровке карт, описанные в научной литературе:

Инструмент / Решение	Архитектура	Предобработка изображений	Распознавание текста	Сегментация цветов	Особенности и ограничения
Ручная оцифровка (Legacy)	-	Визуальная	Ручное	Ручная	Высокая точность, но низкая скорость
USGS Automated Digitization (DIGMAPPER)	U-Net / Transformers	CLAHE, нормализация	Ограниченнное	Семантическая	Ориентирован на топографию (линии, полигоны), требует огромных размеченных датасетов (DARPA-USGS)
ArcGIS	CNN (базовая)	Стандартная нормализация	Ручное	Полуавтоматическая трассировка	Коммерческое решение, дорогие лицензии, требует опыта ГИС-специалиста, так как требуется ручное распознавание текста
Deep Learning Segmentation	CNN (U-Net)	Аугментация	Нет	Pixel-wise	Отлично выделяет границы (лес/вода), но плохо работает с легендами карт и привязкой "цвет-глубина" без OCR
Scan2CAD	Object-based vectorization (не нейросеть)	Бинаризация, истончение линий, очистка от "мусора"	Есть (распознаёт ориентацию и шрифты)	Ограниченнная (опирается только на линии)	Платное ПО. Отлично векторизует чертежи (CAD), но плохо справляется с заливками геологических карт. Требует много ручной настройки параметров даже для извлечения изолиний
Решение от ITMO Student	K-Means + EasyOCR + Tesseract	CLAHE, удаление шума, морфология	EasyOCR + Tesseract	Кластеризация	Специализирован под геокарты, работает на связке «цвет - глубина», работает на малых данных

# Методы и инструменты исследования

## → Методы:

кластеризация (обучение без учителя): KMeans

распознавание глубинных отметок: Deep Learning OCR

методы компьютерного зрения

## → Инструменты:

Python как основной язык программирования

OpenCV для предобработки изображений

EasyOCR, Tesseract для распознавания текста

# Потенциал использования полученных результатов в индустрии



## Области применения:

- нефтегазовая отрасль
- горнодобывающая промышленность
- государственные геологические архивы



## Потенциальный эффект:

- ускорение оцифровки в 25 раз (вместо 4 часов вручную – 10 минут автоматически);
- исключение человеческого фактора и субъективных ошибок преобразований неструктурированных данных;
- универсальность выходного формата, так как выходные данные имеют стандартный машиночитаемый вид, что позволяет как мгновенный импорт в профильное ПО, так и простую автоматизированную предобработку для специализированных задач.

Ссылка на описание индустриального проекта в свободной форме:

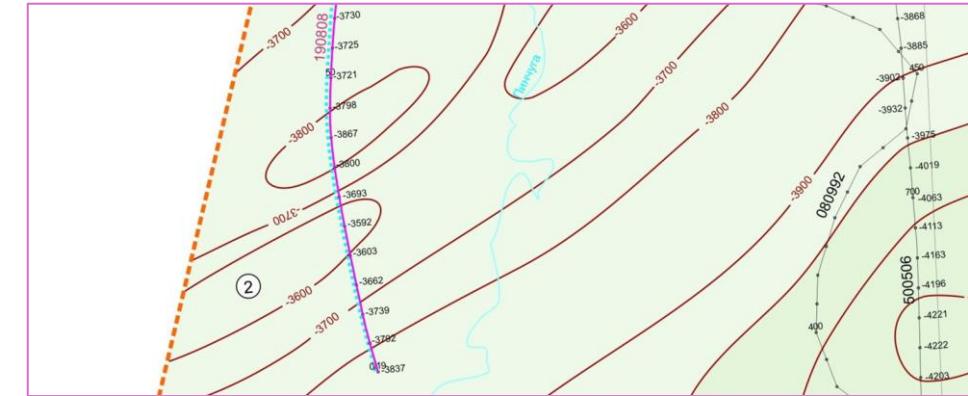
[https://docs.google.com/document/d/1xfOM1gsMGSa47dI0w4afdT6etkezXRF1Fndxm\\_H6RLg/edit?usp=sharing](https://docs.google.com/document/d/1xfOM1gsMGSa47dI0w4afdT6etkezXRF1Fndxm_H6RLg/edit?usp=sharing)

# Ход решения

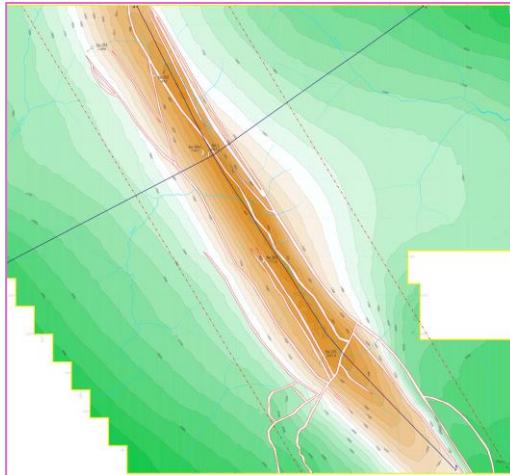
Оценка качества исходных бумажных материалов



Пример 1. Фрагмент исходной карты №1



Пример 2. Фрагмент исходной карты №2



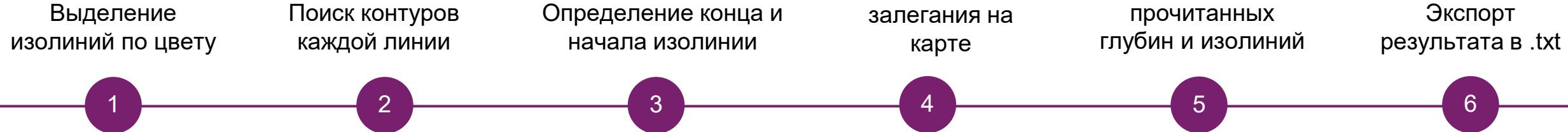
Пример 3. Фрагмент исходной карты №3

## Критические проблемы исходных материалов:

- Наложение гидросети на геологические контуры
- Высокая плотность изолиний в зонах крутых склонов
- Нерегулярное распределение глубинных отметок и вариативность их ориентации
- Искажения изображений, вызванные процессом сканирования

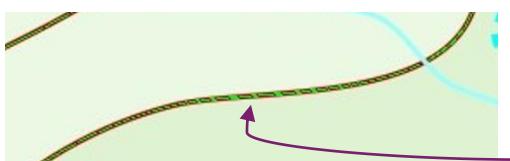
# Ход решения

В ходе работы были рассмотрены два подхода. Идея **первого подхода**:



## Ограничения подхода:

- Изолинии на одной карте могут иметь разные цвета из-за других объектов
- Прерывистый морфологический каркас объектов
- Не все изолинии замкнуты, они могут прерываться тектоническими нарушениями
- Сложности привязки надписей глубин залегания к изолиниям по критерию минимального расстояния



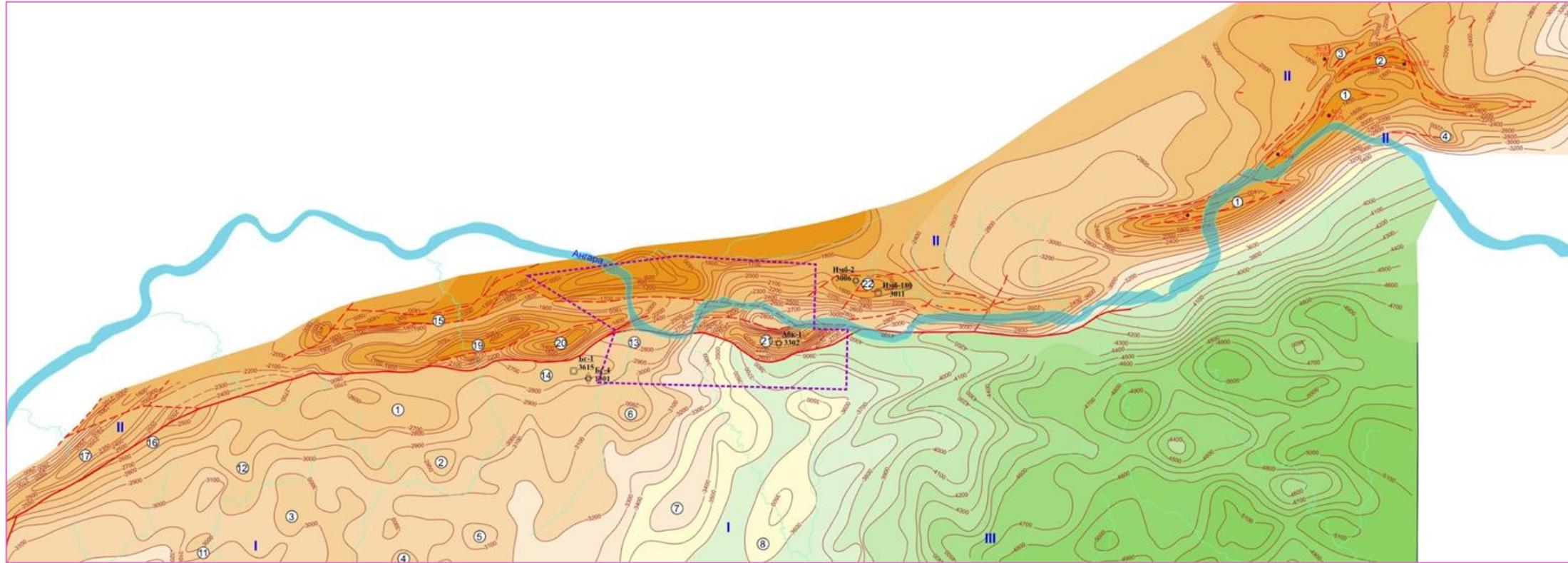
прерывистый каркас изолинии  
(зеленые пиксели)

# Ход решения

**Второй подход** основан на кластеризации цветового пространства карты и сопоставлении геозон с отметками глубины залегания. Процесс обработки карты состоит из 4 шагов.

## Шаг 1

Загрузка карты, анализ её особенностей



Исходная карта

# Ход решения

Шаг 2

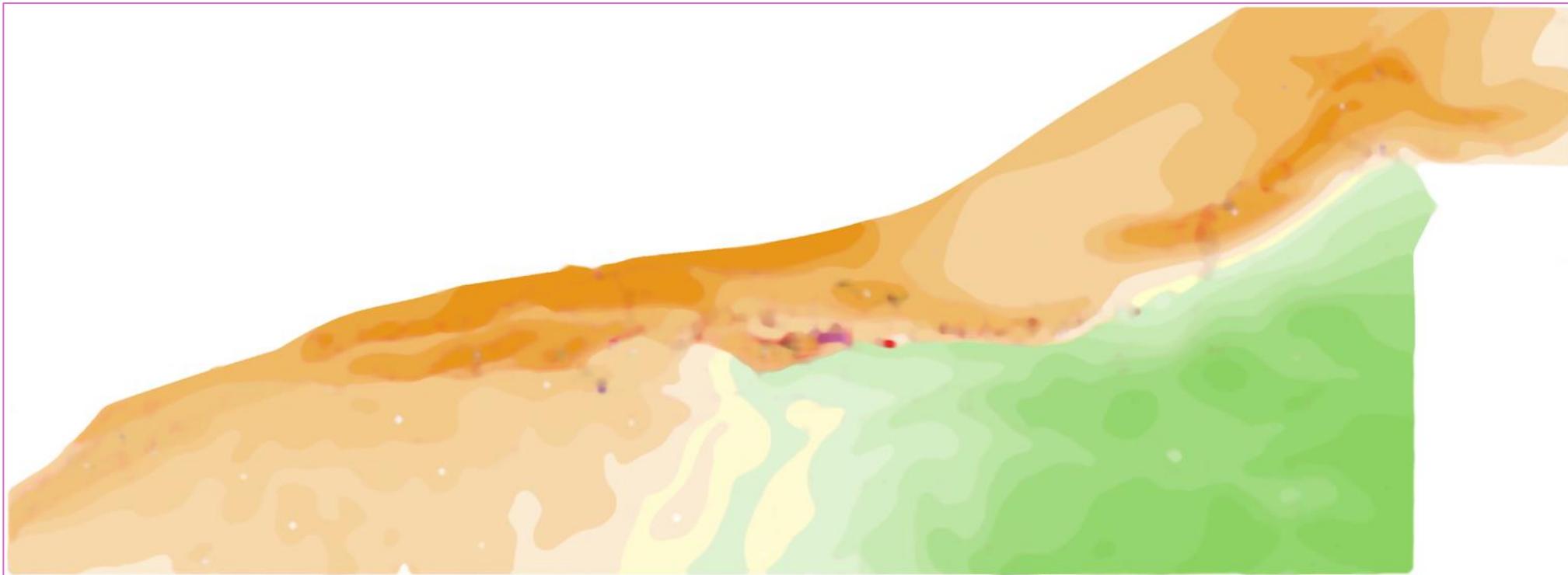
Очистка карты от шума, выделение всех RGB кластеров, группировка близких цветов

Задача	Функция / итог
Удаление тонких линий	cv2.medianBlur()
Сглаживание текстур с сохранением границ	cv2.bilateralFilter()
Линеаризация массива изображения	reshape()
Стохастическая выборка пикселей для ускорения расчетов	np.random.choice()
Кластеризация цветового пространства (поиск центроидов)	cv2.kmeans()
Сохранение эталонной палитры геозон	<p>RGB: (245, 211, 161) #f5d3a1 RGB: (163, 219, 131) #a3db83 RGB: (233, 156, 50) #e99c32 RGB: (252, 243, 216) #fcf3d8 RGB: (23, 23, 23) #171717 RGB: (238, 180, 101) #eeb465 RGB: (240, 195, 129) #f0c381 RGB: (194, 230, 171) #c2e6ab RGB: (218, 239, 202) #daefca RGB: (145, 212, 105) #91d469</p> <p>...</p>

# Ход решения

Шаг 2

Очистка карты от шума, выделение всех RGB кластеров, группировка близких цветов



Очищенная и сглаженная карта (цвета центроидов K-means)

# Ход решения

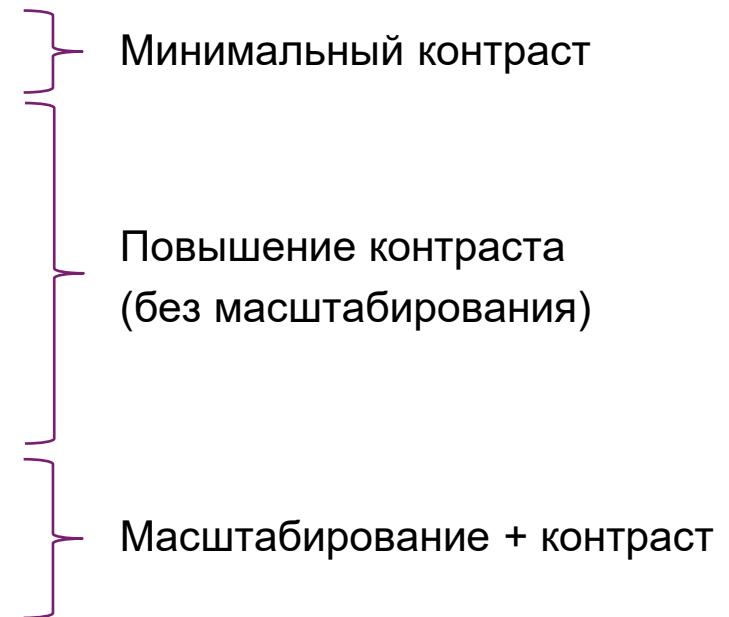
Шаг 3

Распознавание числовых значений и их сопоставление с локальным цветовым контекстом

1

Генерируется несколько вариантов одной и той же карты:

1. CLAHE (контраст 1.0×, масштаб 1×)
2. CLAHE (контраст 2.0×, масштаб 1×)
3. CLAHE (контраст 2.5×, масштаб 1×)
4. CLAHE (контраст 3.0×, масштаб 1×)
5. CLAHE (контраст 4.0×, масштаб 1×)
6. CLAHE + масштабирование (контраст 2.5×, масштаб 1.5×)
7. CLAHE + масштабирование (контраст 2.5×, масштаб 2.0×)



2

Применяем **EasyOCR** и **Tesseract** к каждому варианту обработанного изображения

# Ход решения

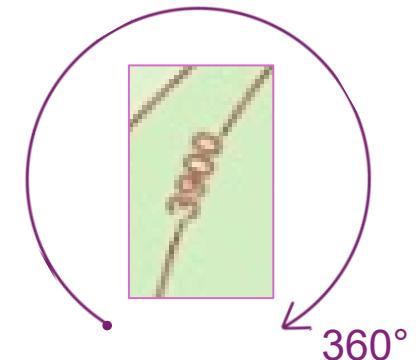
## Шаг 3

Распознавание числовых значений и их сопоставление с локальным цветовым контекстом

3

Для каждого детектированного текстового фрагмента:

- вырезаем патч вокруг найденной цифры
- вращаем патч полный оборот ( $0^\circ$ – $360^\circ$ , шаг  $20^\circ$ )
- распознаем текст при каждом угле поворота, выбираем по голосованию



4

Зная, что диапазон глубин от -500 до -8000 метров:

- отклоняем все 2-значные числа
- для 3-значных устанавливаем порог уверенности 0,3
- для 4-значных устанавливаем порог уверенности 0,15

5

Так как один и тот же текст может быть найден несколько раз (разные движки, разные варианты предобработки), группируем близкие детекции ( $50 \times 50$  px) и оставляем одну (с максимальной уверенностью)

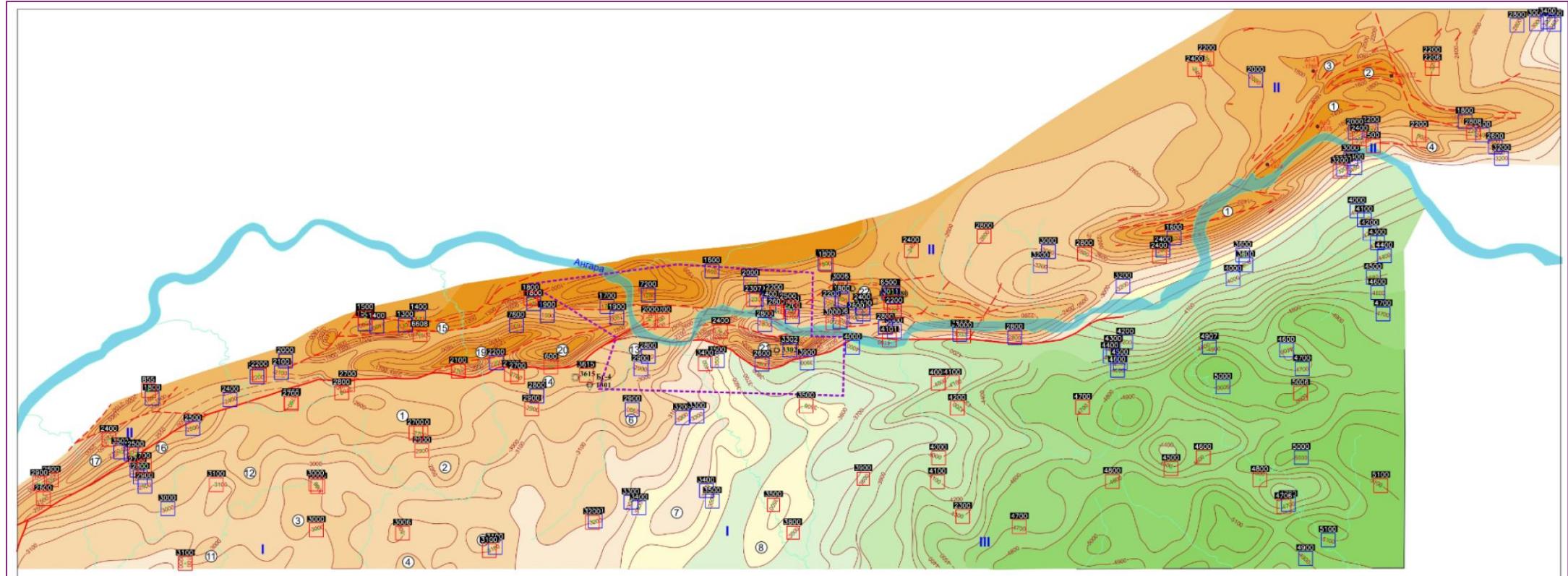
# Ход решения

Шаг 3

Распознавание числовых значений и их сопоставление с локальным цветовым контекстом

6

Определение цветового контекста вокруг каждой распознанной цифры с ранее очищенной и сглаженной карты.  
Сохранение словаря: {цвет - глубина}



# Ход решения

Для оценки качества распознавания введены следующие метрики:

- $\text{accuracy}_{\text{распознавания}} = \frac{\text{количество верно распознанных глубин}}{\text{общее количество детектированных глубин}}$
- $\text{accuracy}_{\text{общая}} = \frac{\text{количество верно распознанных глубин}}{\text{общее количество глубин}}$

Оценка качества распознавания отметок глубины залегания по картам

Карта	Всего на карте	Детектированные глубины	Верно распознанные глубины	Accuracy распознавания	Accuracy общая
Карта 1	223	163	132	0,81	0,59
Карта 2	85	68	51	0,75	0,60
Карта 3	141	86	72	0,84	0,61

- Алгоритм достаточно точно распознаёт найденные глубины залегания, но корректно восстанавливает лишь около половины всех глубин на карте, что требует доработки детекции путем оптимизации пороговых значений и расширения обучающей выборки.

# Ход решения

## Шаг 4



### На входе:

- очищенная и сглаженная карта в цветах геозон
- маска фона
- словарь соответствия {RGB-цвет – глубина}



### Процесс получения конечного результата:

- Для каждого пикселя очищенной карты определяем цвет
- По словарю соответствия {RGB-цвет – глубина} подставляем глубину залегания
- Записываем точки в текстовый файл в формате (x y z)

itog\_neg.txt - Блокнот

Файл Правка Формат Вид Справка

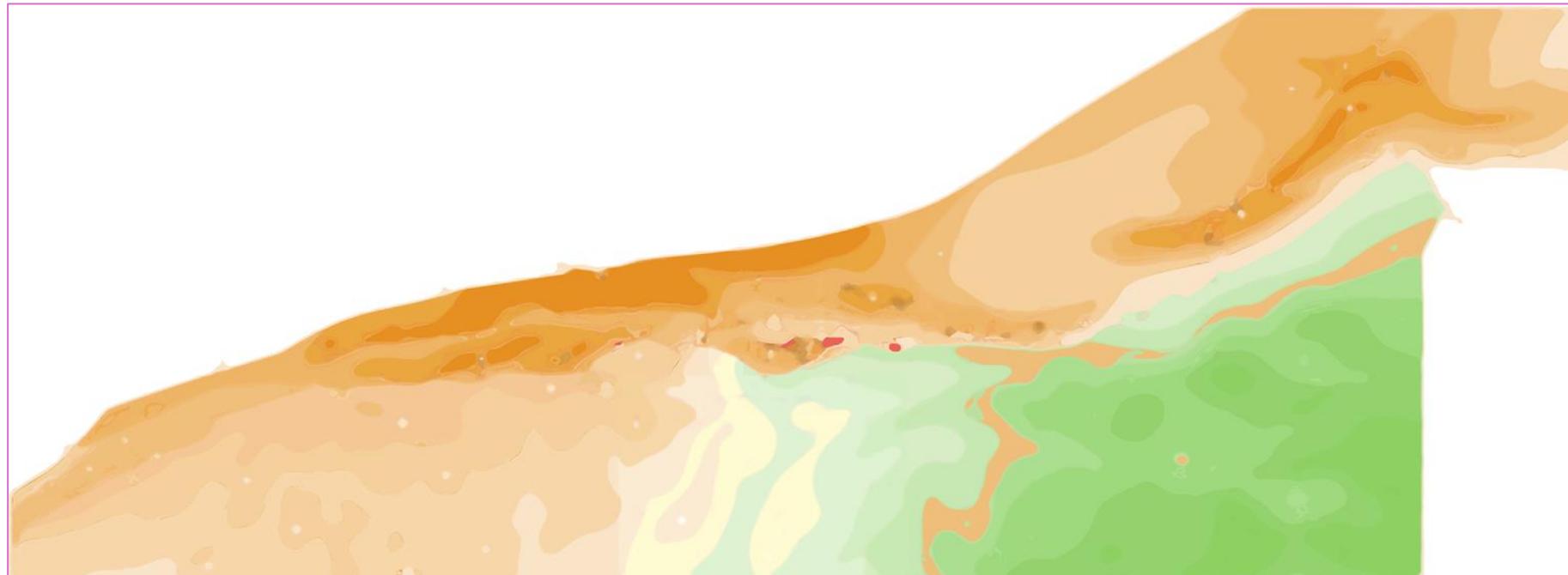
4647,1927,-4500.0  
4648,1927,-4500.0  
4649,1927,-4500.0  
4650,1927,-4600.0  
4651,1927,-4600.0  
4652,1927,-4600.0  
4653,1927,-4600.0  
4654,1927,-4600.0  
4655,1927,-4600.0  
4656,1927,-4600.0  
4657,1927,-4600.0  
4658,1927,-4500.0  
4659,1927,-4700.0  
4660,1927,-4700.0  
4661,1927,-4700.0  
4662,1927,-4700.0  
4663,1927,-4700.0  
4664,1927,-4700.0  
4665,1927,-4700.0  
4666,1927,-4700.0  
4667,1927,-4700.0  
4668,1927,-4700.0  
4669,1927,-4700.0  
4670,1927,-4700.0  
4671,1927,-4700.0  
4672,1927,-4700.0  
4673,1927,-4700.0  
4674,1927,-4700.0  
4675,1927,-4700.0  
4676,1927,-4700.0  
4677,1927,-4700.0  
4678,1927,-4700.0  
4679,1927,-4700.0  
4680,1927,-4700.0  
4681,1927,-4700.0

Стр 31, столб 13 100% Windows (CRLF) UTF-8

# Ход решения

Проверка корректности оцифрованных данных:

- по итоговому файлу восстанавливается растровая карта: для каждой строки ( $x, y, z$ ) по значению глубины залегания  $z$  выбирается цвет из словаря {глубина → RGB} и закрашивается соответствующий пиксель



Карта, построенная по выходному файлу алгоритма

**Совпадение структуры геозон и границ подтверждает корректность оцифровки.**

# Основные результаты

- Доля верно распознанных глубин залегания среди детектированных составляет в среднем 0,80
- Среднее значение доли верно распознанных глубин от общего числа глубин залегания на карте составляет 0,60
- Восстановленные геозоны и границы похожи по структуре и контурам на исходные карты
- Файлы в формате (x, y, z) подходят для импорта в ГИС или CAD-системы
- Получен положительный отзыв партнера (ПАО «Газпромнефть»)

→ Ссылка на открытый репозиторий с кодом реализации поставленной научной задачи и проведения экспериментов: <https://github.com/granatic12/Digitization-of-geological-maps>

# Заключение

Разработан прототип решения для оцифровки геологических карт, объединяющий компьютерное зрение и машинное обучение без учителя. Такой подход позволяет обрабатывать карты в 25 раз быстрее ручной работы и получать данные, подходящие для интеграции с ГИС-системами в стандартном формате (x, y, z).

**Цель работы достигнута.** Неструктурированный растр успешно преобразован в структурированный массив координат и глубин залегания, алгоритм решения протестирован на реальной карте, а верификация подтвердила корректность алгоритма.

## Перспективы развития включают:

- масштабирование решения на большие массивы карт и автоматическую пакетную обработку архивов геоданных;
- совершенствование алгоритмов обнаружения числовых подписей глубин залегания на карте;
- добавление интерфейса для эксперта-геолога, позволяющего интерактивно корректировать цвета зон и глубины залегания перед экспортом.