Визуализация и анализ гелио- и гео-информационных данных

Факультет компьютерных наук, НИУ ВШЭ

Титульный лист

Курсовая работа второго курса программы ПМИ ФКН НИУ ВШЭ:

1. Тема: Визуализация и анализ гелио- и гео- информационных данных
2. Topic: Analysis and Visualization of Helio- and Geo- Data
3. Работу выполнил: **Бугаев Егор Петрович**, студент БПМИ 215
4. Руководитель проекта: **Попов Виктор Юрьевич**
    - Доктор физико-математических наук
    - Заведующий лабораторией моделирования и управления сложными системами ФКН НИУ ВШЭ

Faculty of Computer science

Анализ данных в сейсмологии

- **Сейсмология** является одним из важнейших направлений геологических наук
- Анализ данных уже используется, но далеко не во всех подразделах
- Изучение землетрясений представляет крайний интерес (в том числе практический)

- Проводятся соревнования с денежными призами для выявления новых подходов и сравнения существующих [1]

Цель и задачи работы

Цель: изучение применимости простых методов анализа и визуализации данных в сейсмологии

Задачи:

- Изучение и применение простых методов анализа данных в специфических для сейсмологии задачах
- Оценка успешности примененных методов
- Создание примеров визуализации геоданных

Содержание

01

Выбор темы

Факультет компьютерных наук, НИУ ВШЭ

Постановка вопроса

Одним из главных факторов землетрясений является движение **литосферных плит**.

Гипотезой, проверяемой в данной работе, будет различие в магнитуде землетрясений вдоль границы Евразийской литосферной плиты.

В частности, будем сравнивать восточную границу Евразийской плиты с южной границей.

Границы литосферных плит

Южная

Южная граница Евразийской литосферной плиты

Восточная

Восточная граница Евразийской плиты

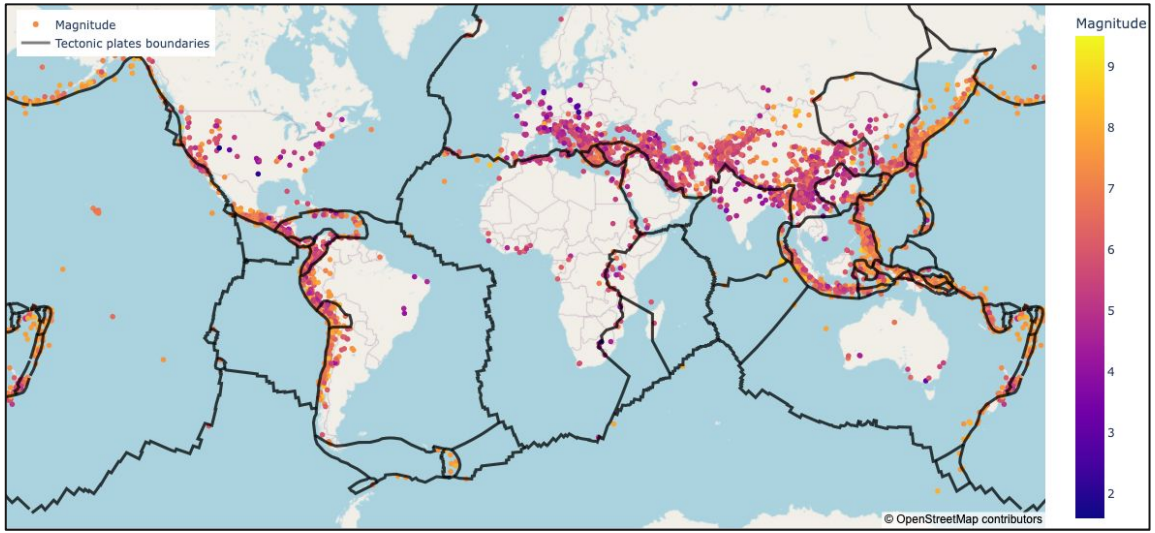
Выбор данных

Используются два набора данных [2] и [3].

Набор данных [2] содержит сведения о землетрясениях с ~2000 года до н.э. по современные землетрясения. В данном исследовании используются геолокации и магнитуды этих землетрясений.

Набор данных [3] содержит границы литосферных плит на карте в удобном формате: точками вдоль границ плит.

Предварительная оценка гипотезы



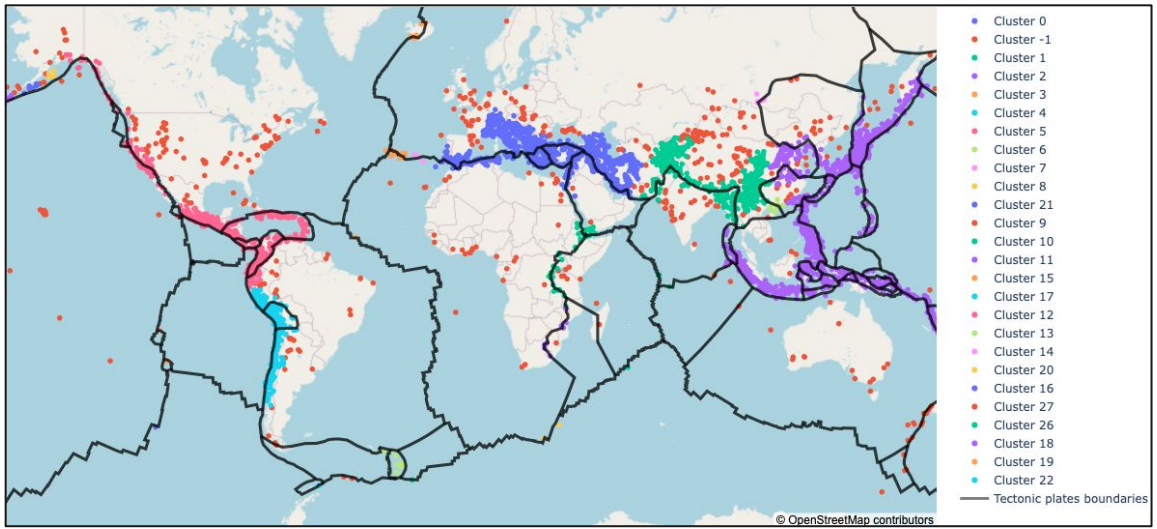
Можем заметить, что в восточной области Евразийской плиты визуально землетрясения в среднем большей магнитуды

02

Очистка данных

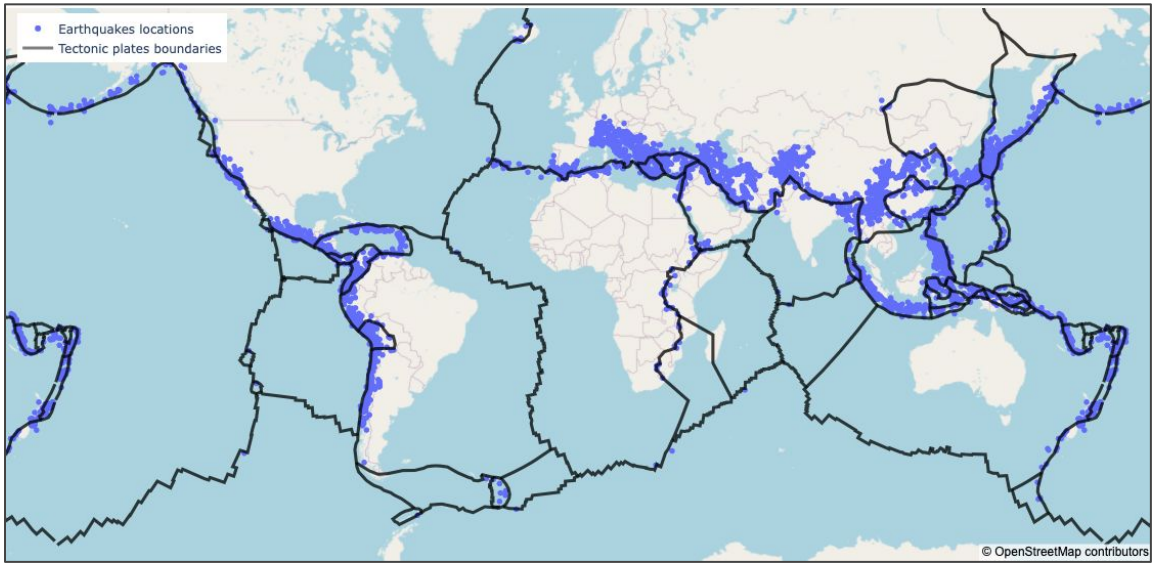
Факультет компьютерных наук, НИУ ВШЭ

Оставляем релевантные данные



Много точек вдалеке от границ литосферных плит. Добавляем границы плит в данные и применяем **DBSCAN**. Оставляем только точки с кластерами (удаляем красные).

Оставляем релевантные данные



Оставшиеся без кластеров точки отбрасываем.

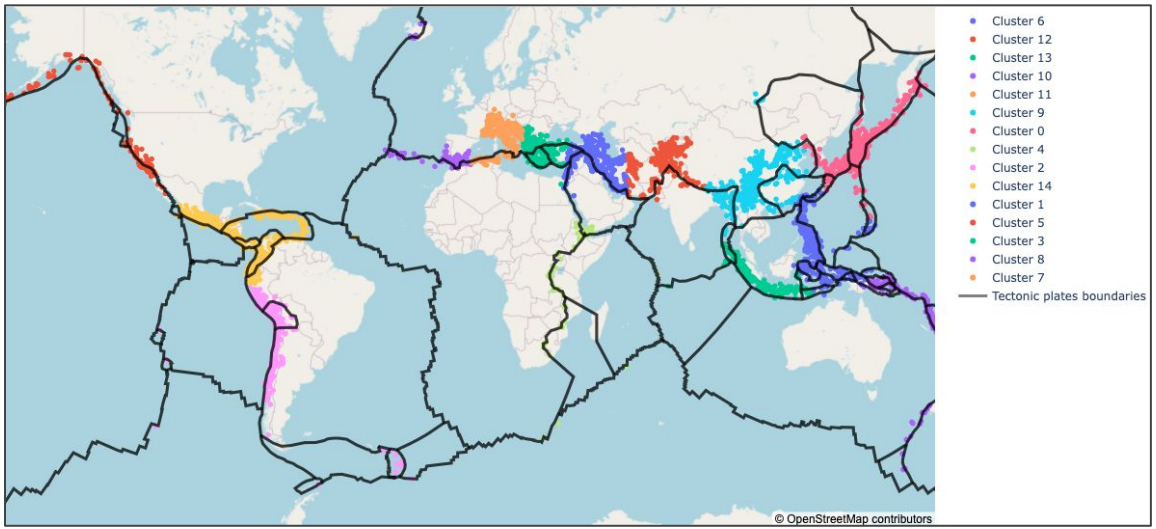
Получаем большую близость точек к границам.

03

Кластеризация

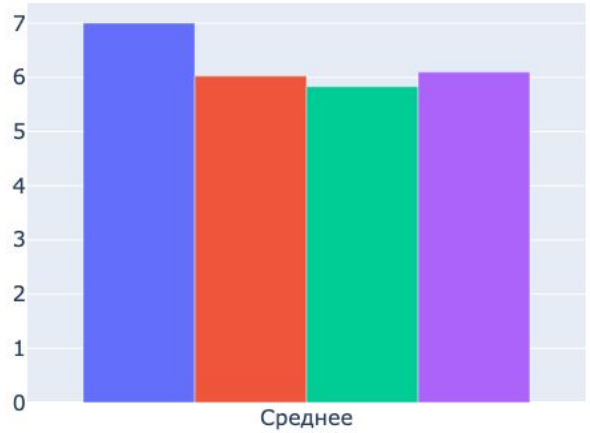
Факультет компьютерных наук, НИУ ВШЭ

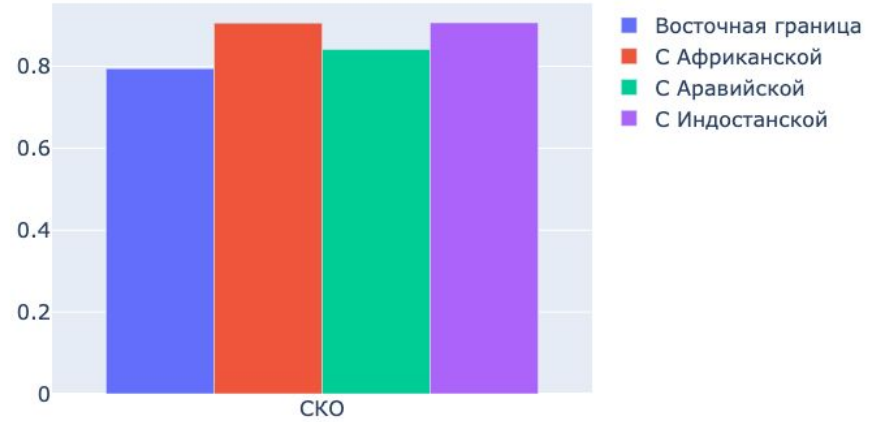
Применяем KMeans. Кластеризация



- 15 кластеров
- Стартовые вершины выбираются случайно
- Результаты схожи при разных запусках, кластеризация стабильна

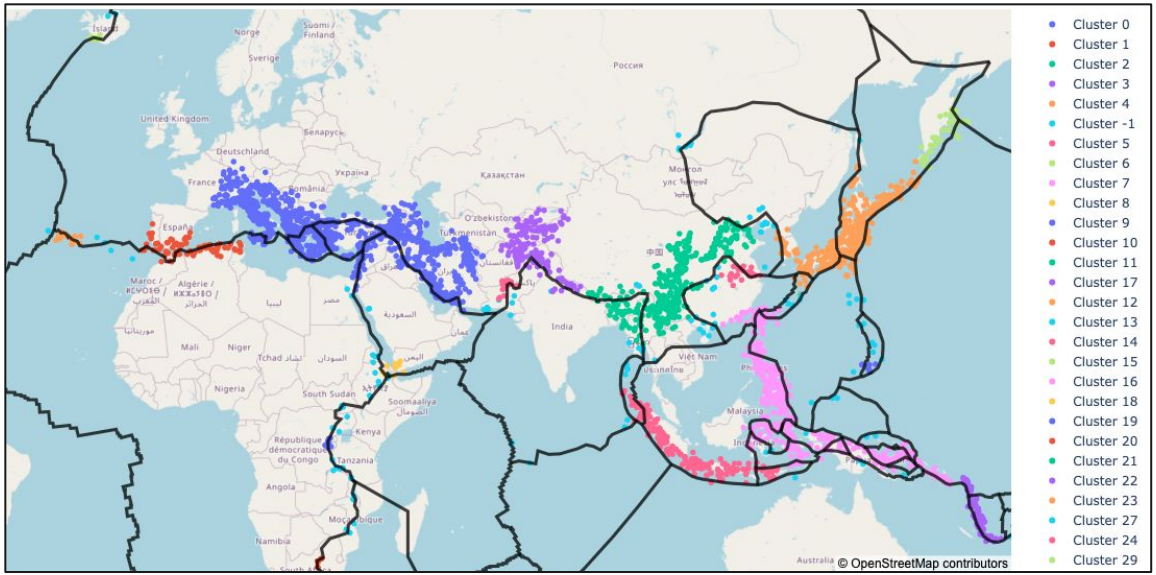
Применяем KMeans. Результаты





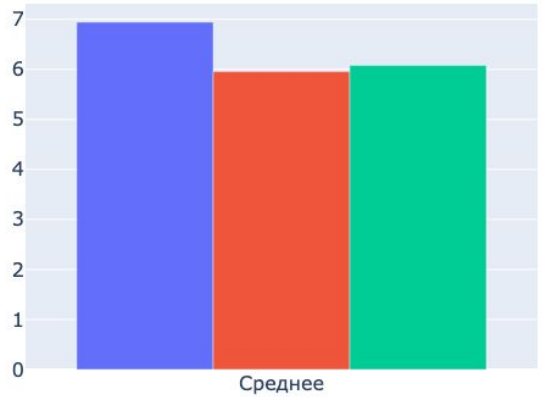
Значимость отличий в среднем значении у восточной границы относительно остальных подтверждается проверкой с помощью t-test с уровнем значимости 0.05

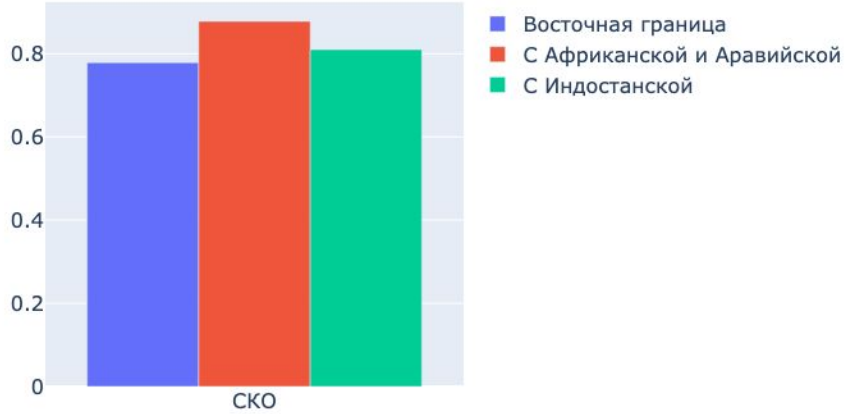
Применяем DBSCAN. Кластеризация



- minPts = 8
- eps =2 (считаем в градусах)
- Получаем более вытянутые кластера из-за специфики алгоритма

Применяем DBSCAN. Результаты





Значимость отличий в среднем значении у восточной границы относительно остальных подтверждается проверкой с помощью t-test с уровнем значимости 0.05

04

Оценка кластеризации

Факультет компьютерных наук, НИУ ВШЭ

Silhouette score

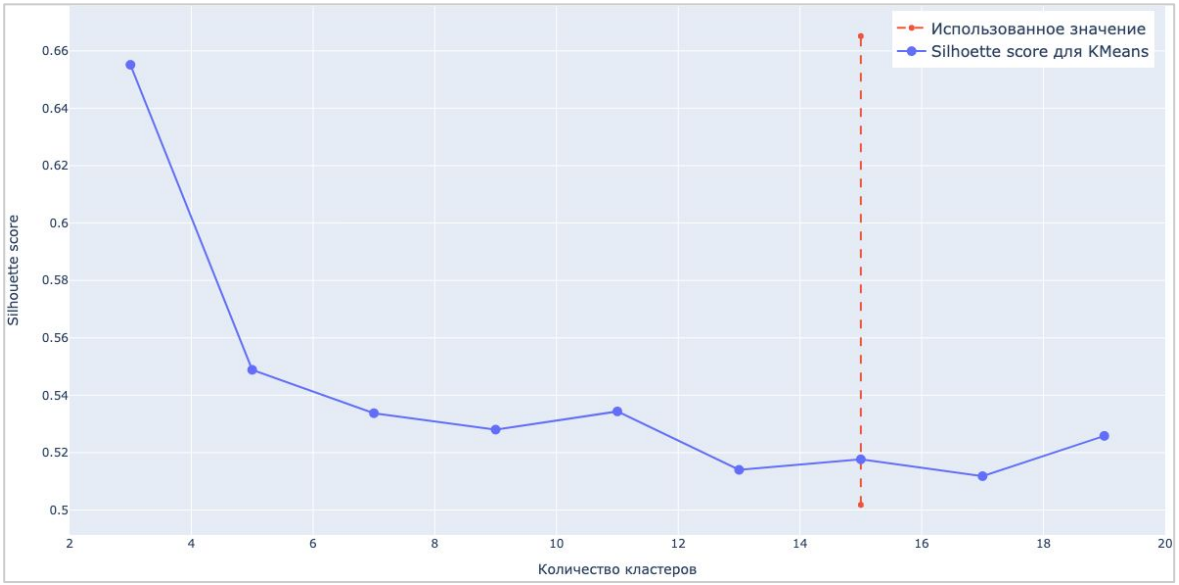
$$\theta = \frac{\sum_{i \in P} \frac{b-a}{\max(a,b)}}{|P|}$$

где для каждой точки **i** из всех точек в данных **P**:

- **a** - среднее расстояние между точками в кластере, которому принадлежит **i**.
- **b** - расстояние от **i** до точки в ближайшем кластере, не равном кластеру **i**.

При хорошем разбиении на кластера (не пересекаются) стремится к 1, при плохом (точки определены в неверные кластера) к -1.

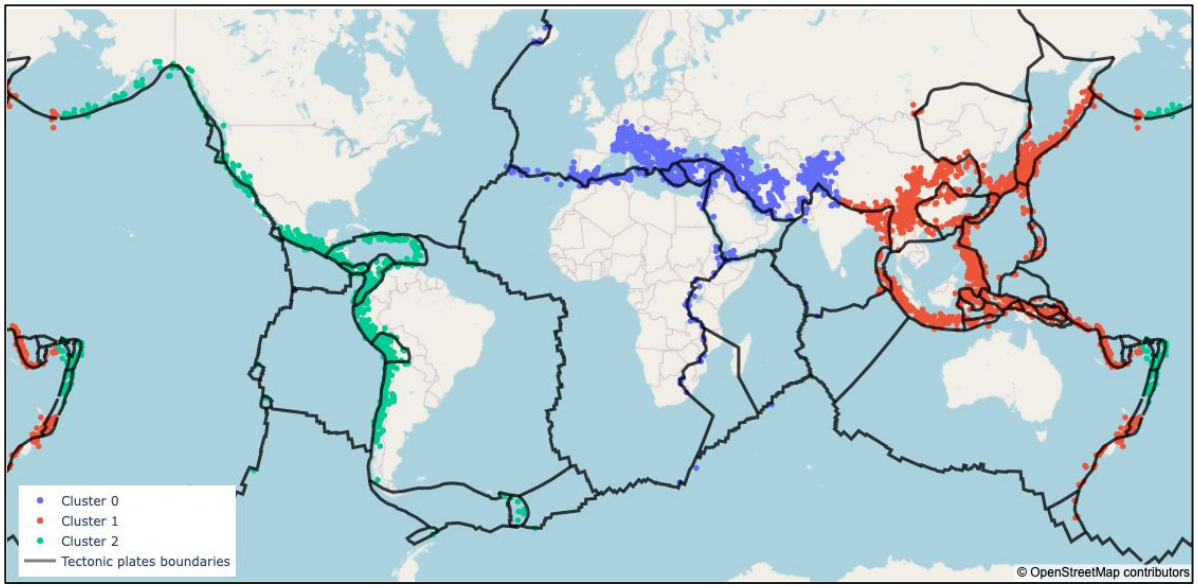
Silhouette score. KMeans



Оптимальное значение метрики для больших кластеров.

В выбранной нами кластеризации значение около **0.5.**

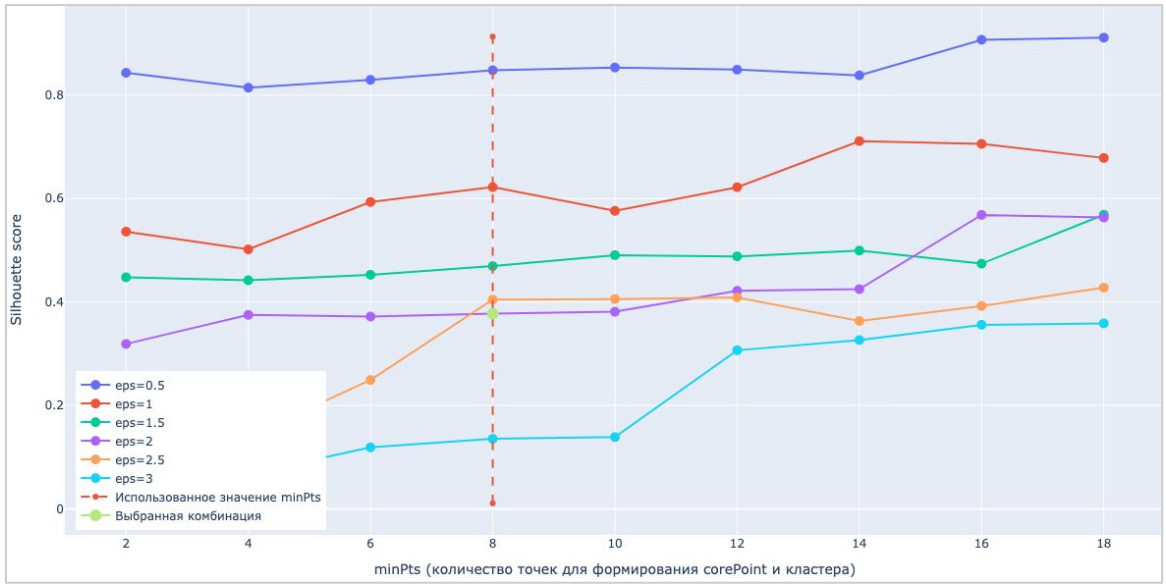
KMeans с 3 кластерами



Оптимальное значение метрики Silhouette score.

Очень крупные кластера, непригодно для анализа.

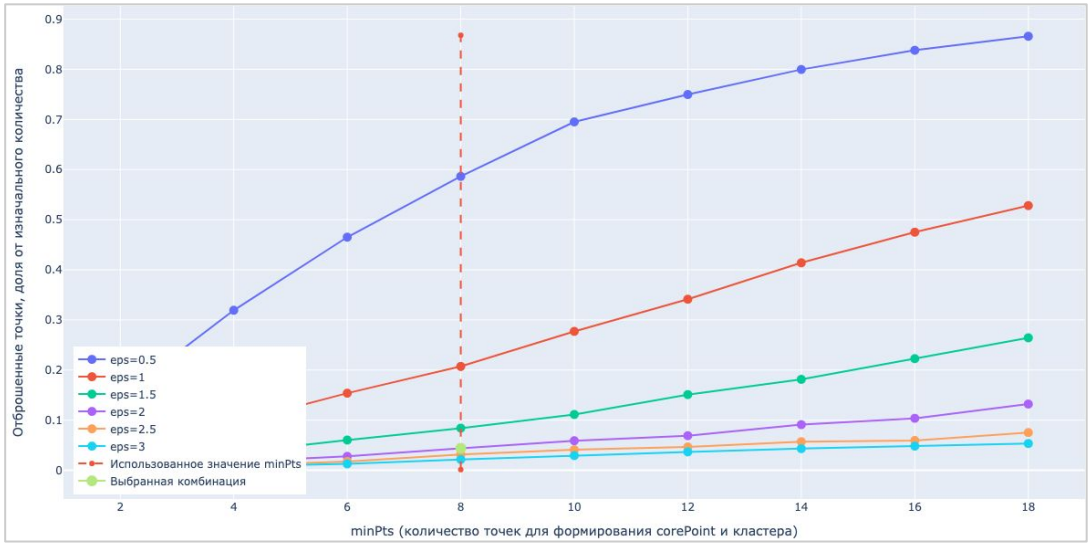
Silhouette score. DBSCAN



Значения почти не зависят от minPts.

При уменьшении eps метрика растет (но становится ли лучше кластеризация?)

Silhouette score. DBSCAN



При маленьком eps теряется много точек (особенно при увеличении minPts)

DBSCAN **не находит** им кластера и такие точки перед подсчетом метрики отбрасываются.

Adjusted Random Index Score

$$\theta = \frac{2 \cdot k}{n(n-1)}$$ - Random Index

где:

- **k** - количество пар, правильно определенных в один или разные кластера.
- **n** - количество предварительно размеченных точек в наборе данных.

Требуется набор точек с заранее размеченными **эталонными** кластерами. Сравниваем, насколько близка наша кластеризация к эталонной.

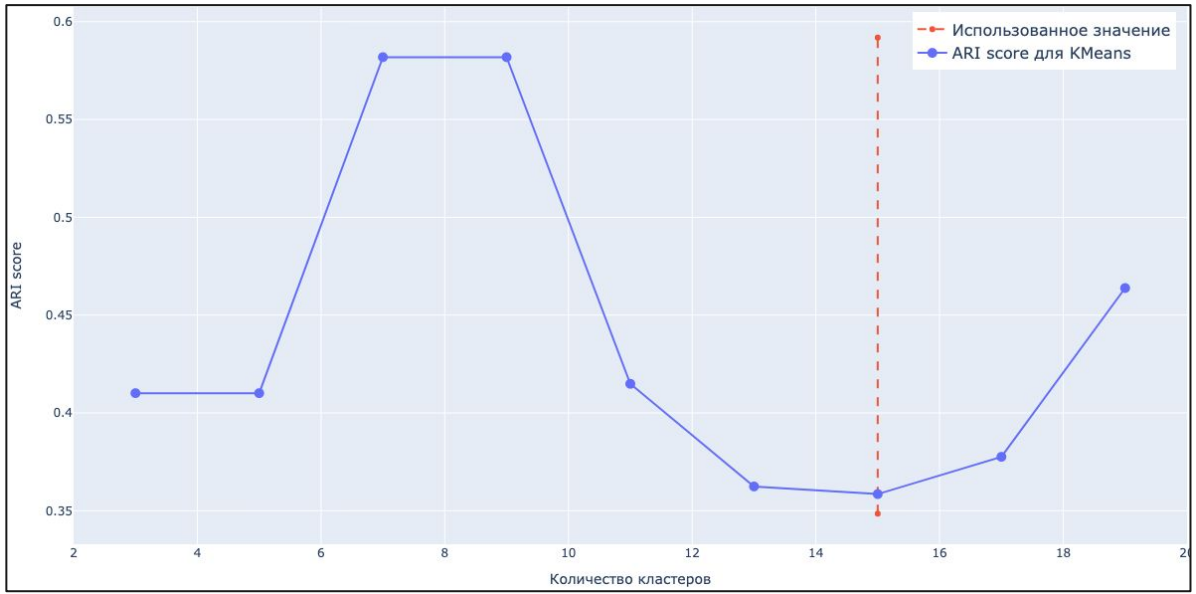
Adjusted RI считаем с помощью несложных преобразований, получаем 0 при случайной кластеризации (исходно RI не обладает этим свойством).

Готовим эталонные кластера

Выбираем границы литосферных плит около Евразийской, точки на них размечены по плите, которой принадлежат.

На каждой границе оставляем точки только одной плиты.

ARI Score. KMeans

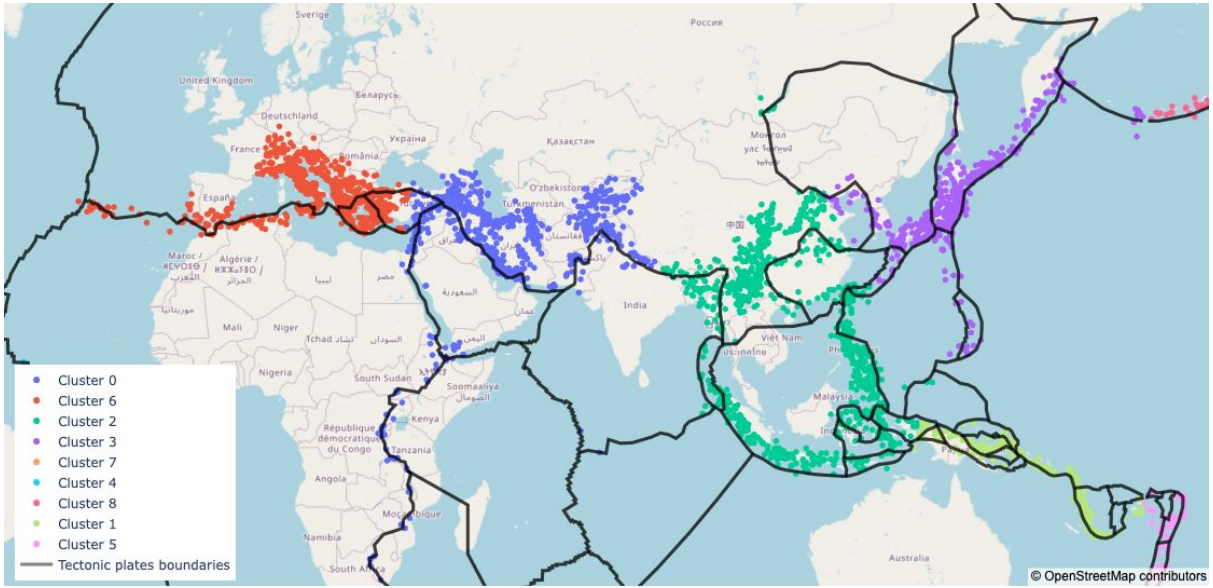


Избавились от тенденции к укрупнению кластеров.

Оптимально другое разбиение - с 8 кластерами.

Наше находится в локальном минимуме.

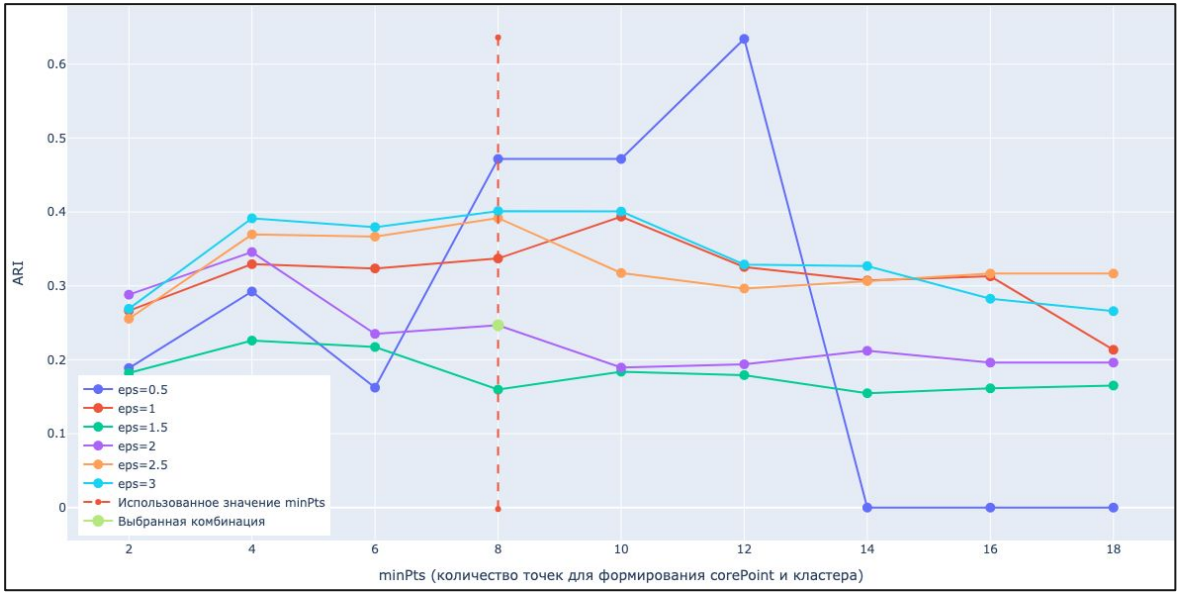
KMeans с 9 кластерами



Плохо сгруппировали восточную границу Евразийской плиты, важную для исследования.

В остальном кластеризация похожа на выбранную.

ARI Score. DBSCAN



Снова явная зависимость качества кластеризации от eps (но теряем больше точек).

Выбранное значение соблюдает некий баланс.

Нулевая ошибка при большом minPts и малом eps - не остается точек.

05

Заключение

Факультет компьютерных наук, НИУ ВШЭ

Заключение

- Методы кластеризации пригодны для исследований, но часто требуют ручной доработки/проверки
- Метрики показывают отличие кластеризации от шума, но не все одинаково хорошо применимы для оценки алгоритмов
- Проводить первичную оценку гипотез и дальнейшую проверку можно даже простыми алгоритмами

Дальнейшая работа

- Применять алгоритмы классификации вместо кластеризации: использовать заранее размеченные точки на границах
- Проводить более детальный анализ (меньше кластера)
- Оценивать изменения средней магнитуды с течением времени

Использованные источники

- LANL Earthquake Prediction (Kaggle ML Competition)
- Earthquakes -2150 BC – 2023 AD around the world, Bharath Pose (Kaggle dataset)
- Tectonic Plate Boundaries, Curtis Thompson (Kaggle Dataset)
- Step by Step to Understanding K-means Clustering and Implementation with sklearn, Data Folks Indonesia, Arif R.
- DBSCAN Clustering — Explained, Towards Data Science, Soner Yıldırım

Подробнее можно посмотреть в pdf курсовой работы

Использованное ПО

- Plotly, NumPy, Pandas libraries (для вычислений и визуализации)
- Google Collab (для вычислений и представления работы)

Ссылка на репозиторий с работой:

https://github.com/EgorBugaev/earthquakes_clusterization

Спасибо за внимание!

Контактная информация:

egor07072003@gmail.com
@epbugaev (tg)

Faculty of Computer science