

ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

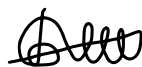
Факультет компьютерных наук
Образовательная программа «Прикладная математика и информатика»

УДК 550.343.6

Отчет об исследовательском проекте на тему:
Визуализация и анализ гелио- и гео- информационных данных

Выполнил:

студент группы БПМИ215
Бугаев Егор Петрович



(подпись)

25.05.23

(дата)

Принял руководитель проекта:

Попов Виктор Юрьевич
Доктор физико-математических наук
Заведующий лабораторией моделирования и управления сложными системами
факультета компьютерных наук НИУ ВШЭ



(подпись)

Москва 2023

Содержание

Аннотация	3
1 Постановка задачи	4
2 Очистка данных	6
3 Разбиение на кластеры и проверка гипотезы.	8
3.1 Разбиение с помощью KMeans	8
3.2 Разбиение с помощью DBSCAN	11
4 Оценка качества кластеризации	13
4.1 Average silhouette score	13
4.2 Adjusted Rand Index	16
5 Заключение. Дальнейшая работа	20
Список литературы	21

Аннотация

Методы анализа данных и машинного обучения уже прочно вошли практически во все области прикладных наук. В данной работе рассматривается применение набора методов обработки, анализа и визуализации данных к исследованию наборов геоданных, в частности карты землетрясений.

Работа носит ознакомительный характер и призвана показать, что методы анализа данных можно легко применять к геоданным (в частности, сейсмологическим).

Здесь приведен пример интересного использования простых алгоритмов кластеризации (KMeans, DBSCAN) для очистки данных и разбиения информации на группы, визуализации с помощью библиотек языка Python (в частности, Plotly) для постановки гипотез и методов математической статистики (t-test) для их проверки.

Для оценивания успешности кластеризации выбраны и применены несколько метрик качества разбиения (Average silhouette score, Adjusted Rand Index), строятся наглядные графики.

Данная работа покрывает только малейшую часть возможностей анализа данных в сфере геологии и географии, дальнейшие идеи по применению алгоритмов кластеризации и визуализации в контексте анализа и предсказания землетрясений указаны в конце.

Ключевые слова

Датасет (англ. **Dataset**) - набор данных (в данной работе чаще всего - структурированная таблица).

Kaggle - интернет-платформа с большим количеством датасетов и соревнованиями по анализу данных, откуда были взяты данные для анализа.

GitHub - платформа (удаленный репозиторий), где храниться код и прочие данные проекта.

Магнитуда землетресения - величина, характеризующая энергию, выделившуюся при землетрясении в виде сейсмических волн.

Кластеризация - процесс группировки данных в похожие классы (кластеры).

KMeans, DBSCAN - алгоритмы кластеризации.

1 Постановка задачи

Одной из самых частых причин землетрясений являются столкновения и движения литосферных плит, в результате которых происходят подземные толчки и колебания земной поверхности, которые мы и называем землетрясениями. Известно, что значительная часть землетрясений происходит на стыках литосферных плит - изучением детальных причин возникновения таких сдвигов занимается наука сейсмология.

В данной работе изучается вопрос мощности землетрясений вдоль некоторых границ литосферных плит, в частности, Евразийской. Евразийская плита является одной из самых больших, ее южная сторона контактирует с Африканской, Аравийской¹, Индостанской и Австралийской плитами. На востоке же она граничит с Северо-Американской, Тихоокеанской и Филиппинской плитой.

Кроме того, на стыке Евразийской, Филиппинской, Северо-Американской и Тихоокеанской плит образуется уникальное место, расположенное на так называемом Тихоокеанском вулканическом огненном кольце.

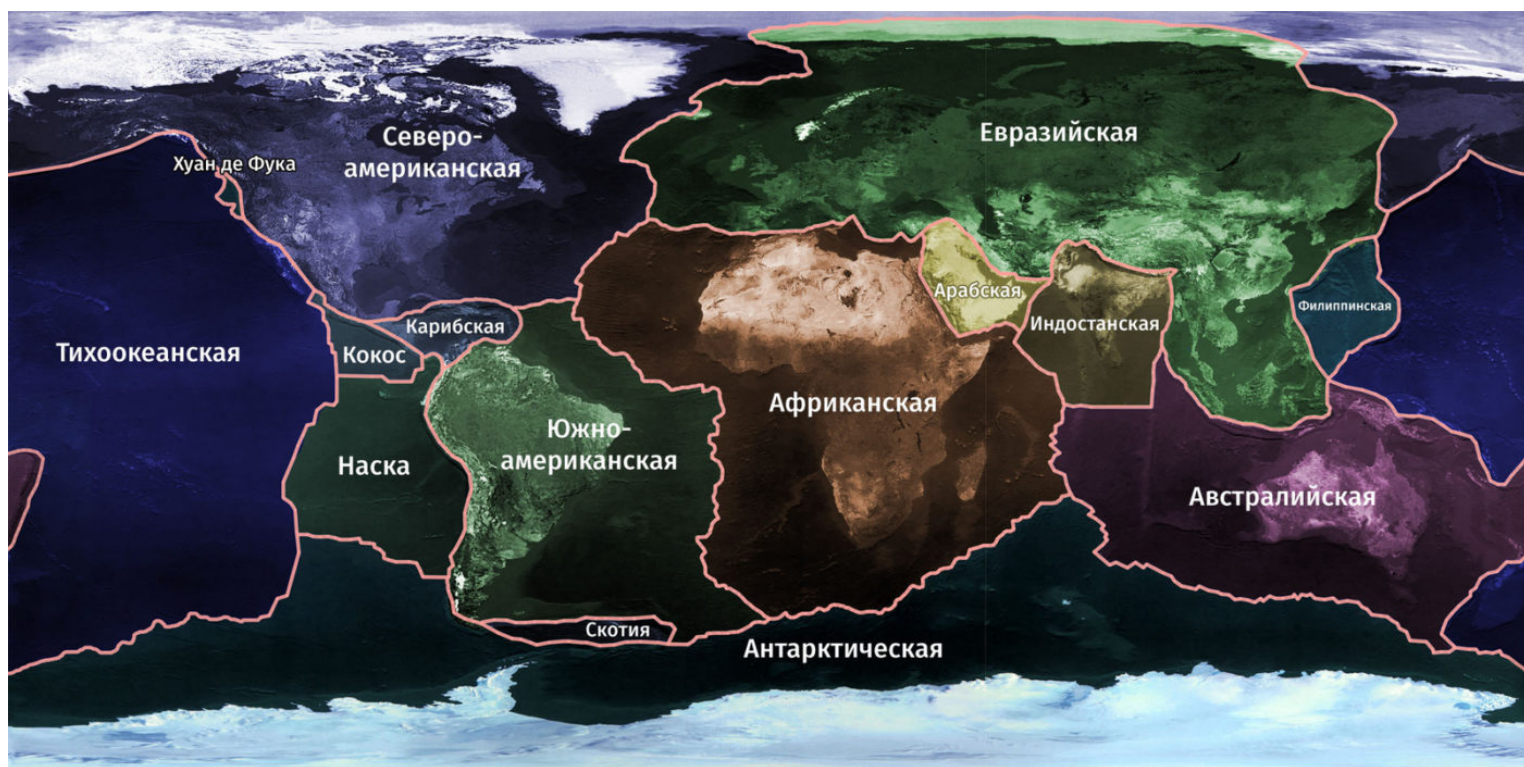


Рис. 1.1: Примерные границы литосферных плит на карте мира

Для анализа силы землетрясений на стыках литосферных плит было выбрано два набора данных с сайта Kaggle. Первый набор данных [2] содержит большое количество записей о землетрясениях в период с -2150 года до н. э. до самых современных землетрясений. Несмотря на обильное количество информации в датасете, в данной работе используются только локации и магнитуды землетрясений.

Так же в работе используется еще один датасет [5], содержащий локации границ тектонических плит (границы содержатся в виде точек на маленьком расстоянии друг от друга). Эти данные используются для очистки информации о землетрясениях (убираем далекие от стыков участки), для более наглядного отображения результатов, а так же при подсчете метрик качества кластеризации.

¹Аравийская плита так же называется Аравской из-за своего английского названия (Arabian plate)

При взгляде на карту литосферных плит сразу бросается в глаза стык четырех плит на востоке от Евразийской плиты: с двух сторон Филиппинскую плиту поджимают огромные Евразийская и Тихоокеанская, а сверху расположена Северо-Американская плита. Напрашивается предположение, что в этой зоне землетрясения будут наиболее сильными, даже сравнительно других границ Евразийской плиты.

Для того, что первоначально оценить наше предположение, с помощью библиотеки Plotly языка Python нанесем на график границы литосферных плит, а так же обозначим землетрясения с их магнитудой по шкале Рихтера.

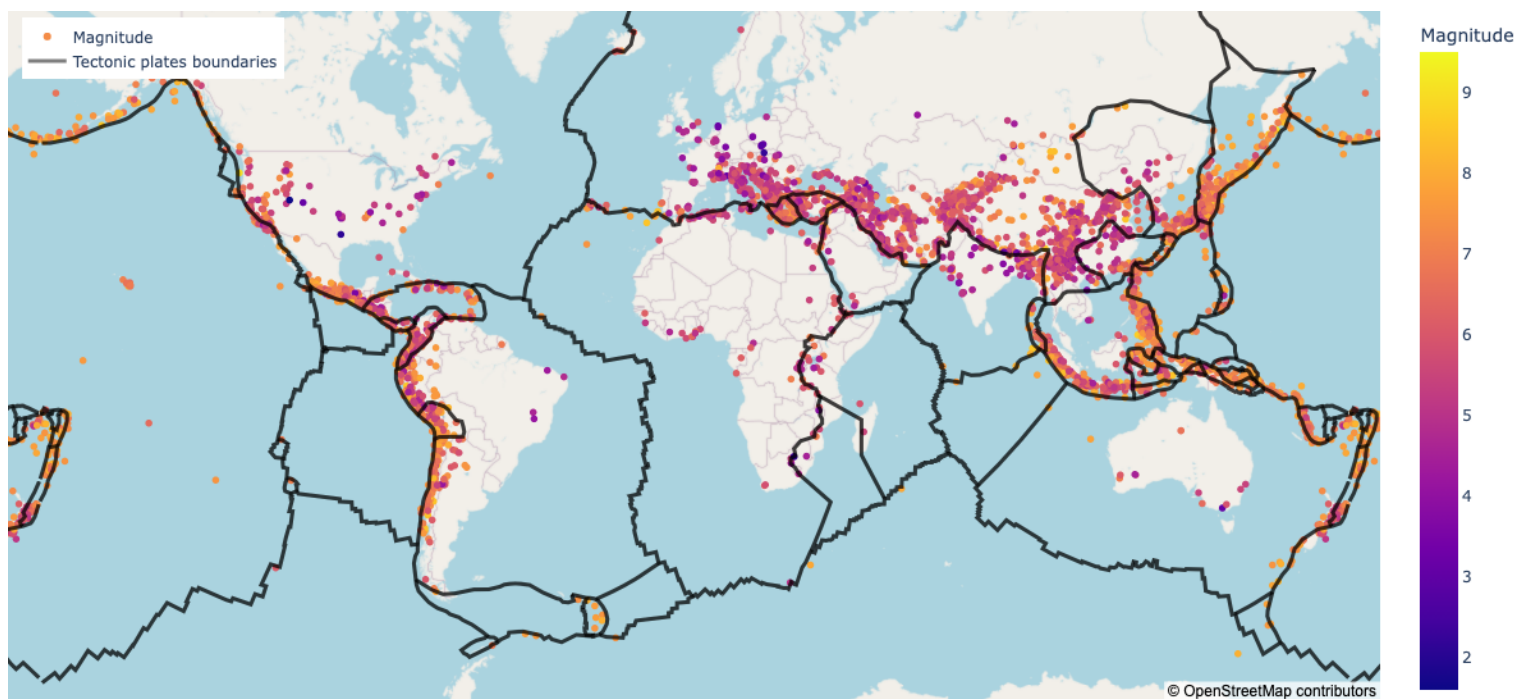


Рис. 1.2: Землетрясения на карте мира

Посмотрим на Рисунок 1.2 и 1.3. Здесь черными линиями обозначены границы литосферных плит (учтено больше малых плит и платформ, чем на Рисунке 1.1, но общие границы легко проследить). Цветом, в соответствии со шкалой справа, обозначены магнитуды землетрясений.

В частности, нас снова интересует сектор на стыке четырех плит (восток Евразийской литосферной плиты). На нашем график достаточно отчетливо видно, что там (как и в целом вдоль всей Тихоокеанской плиты, но это нас в данной работе не так интересует) в среднем землетрясения происходят сильнее, чем вдоль южной границы Евразийской плиты (ее стыках с Африканской, Аравийской, Индостанской).

Выдвинем гипотезу, что на этом участке (восточная граница Евразийской плиты) средняя магнитуда сильнее, чем вдоль ее южной границы. Однако заметим, что сейчас в данных присутствует большое количество землетрясений вдали от границ литосферных плит. Они нам для этой гипотезы не интересны, попробуем убрать их из данных.

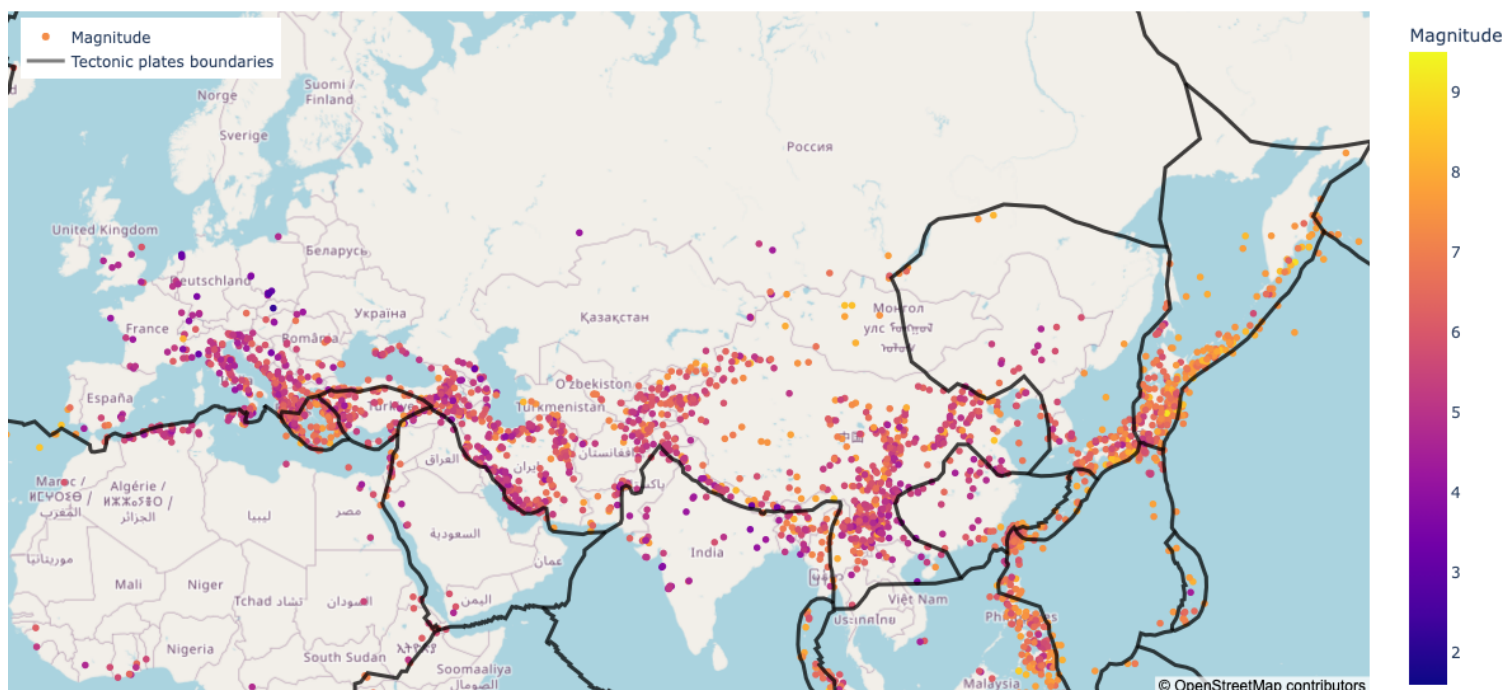


Рис. 1.3: Землетрясения у Евразийской плиты

2 Очистка данных

По графикам выше легко заметить, что в данных большое количество землетрясений вдалеке от границ тектонических плит (центральные части Северной Америки, Африки, и, наиболее важное для нас, в Евразии на удалении от границы ее литосферной плиты).

Хотим оставить только скопления землетрясений вдоль границ литосферных плит.

Попробуем добиться следующей цели - собрать все землетрясения вокруг границ в один кластер (группу), остальные землетрясения исключить.

Для данных целей отлично подходит алгоритм кластеризации DBSCAN (разбор алгоритма приводится в [8]).

Действительно, алгоритм позволяет объединять объекты в кластера необычной формы. Большинство алгоритмов кластеризации тянутся к круглым кластерам, для нашей задачи кластеризации вдоль линии разломов это малоприменимо.

Алгоритм зависит от двух параметров, **eps** и **minPts**. В основе алгоритма DBSCAN лежит концепция **core points**: это точки, вокруг которых будут формироваться кластера. Для того, чтобы алгоритм сформировал кластер вокруг одной из точек (расположение землетрясения на карте), на расстоянии не более **eps** от нее должны находиться хотя бы **minPts** других точек. Дальше соседи этой точки по некоторым правилам присоединяются к этому кластеру (для лучшего понимания стоит прочитать [8]), в частности, все точки на расстоянии **minPts** от **corePoints** конкретного кластера входят в этот кластер.

Таким образом, увеличивая параметр **minPts**, мы можем требовать более и более скучкованные кластера. Воспользуемся следующим трюком: вспомним, что границы литосферных плит в наших данных представлены непрерывной цепочкой из точек вдоль границы (точки находятся на маленьком расстоянии друг от друга). Добавим эти точки вдоль границ в наши наборы землетрясений (можем оставить пустыми поле магнитуды, оно нам не пригодятся).

Повторим операцию добавления точек в датасет немного меньше раз², чем мы выставили minPts (15 это minPts, добавим 10 раз). Тогда получаем, что теперь наши данные содержат большие скопления точек ровно на границах. При запуске DBSCAN все эти скопления будут объединены в несколько крупных кластеров - действительно, так как каждая точка на границе включена в датасет большое количество раз, она почти автоматически становится corePoint, ей нужно всего лишь несколько точек рядом. Мы получаем несколько больших кластеров вдоль границ литосферных плит, как мы и хотели.

Кроме того, все точки в досягаемости кластеров около границ подсоединяются к их кластерам как достижимые из corePoints. Если точек около границы много, они сами тоже становятся corePoints и увеличивают достижимую область. Результатом становится, что в наши большие кластера входят границы и существенные скопления рядом.

Точки, которые ни в какие кластера не вошли, из датасета уберем. Они удалены от границ литосферных плит и нас не интересуют. Результаты алгоритма показаны ниже (график после кластеризации с помощью DBSCAN и оставшиеся после удаления точки).

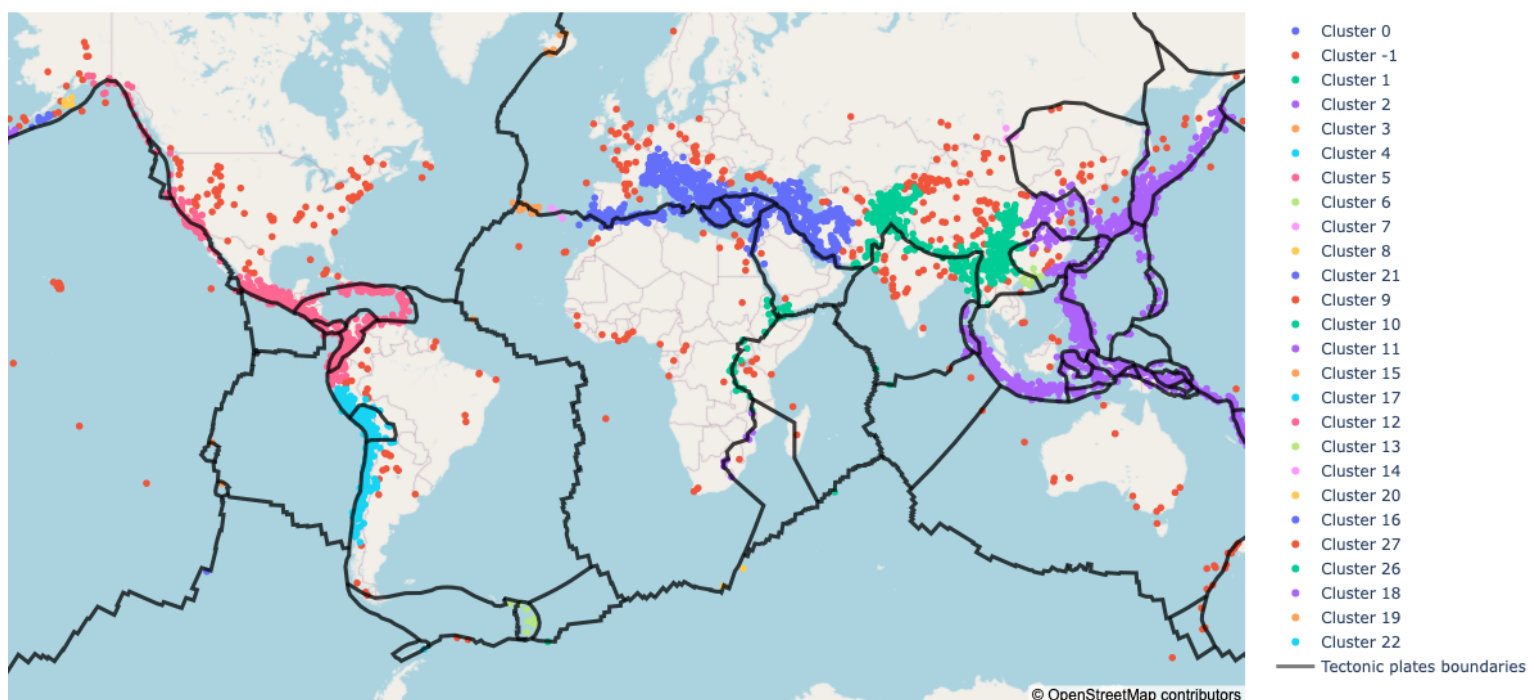


Рис. 2.1: Разбиение на кластеры с DBSCAN после добавления точек вдоль плит

Красным на Рисунке 4.2 отмечены точки, которым не нашлось кластера. Заметим что почти все точки ближе к центрам плит как раз кластера не нашли. Немного точек ближе к центру осталось на территории Европы, но они все еще сравнительно близки к восточной границе Евразийской плиты и интересны для изучения. Удалим из данных все точки без кластера, а так же добавленные искусственные точки вдоль границ плит.

²Исходно был рассмотрен вариант с добавлением ровно minPts раз, но кластер получался слишком большим, распространялся почти на всю карту и выбросы уходили хуже

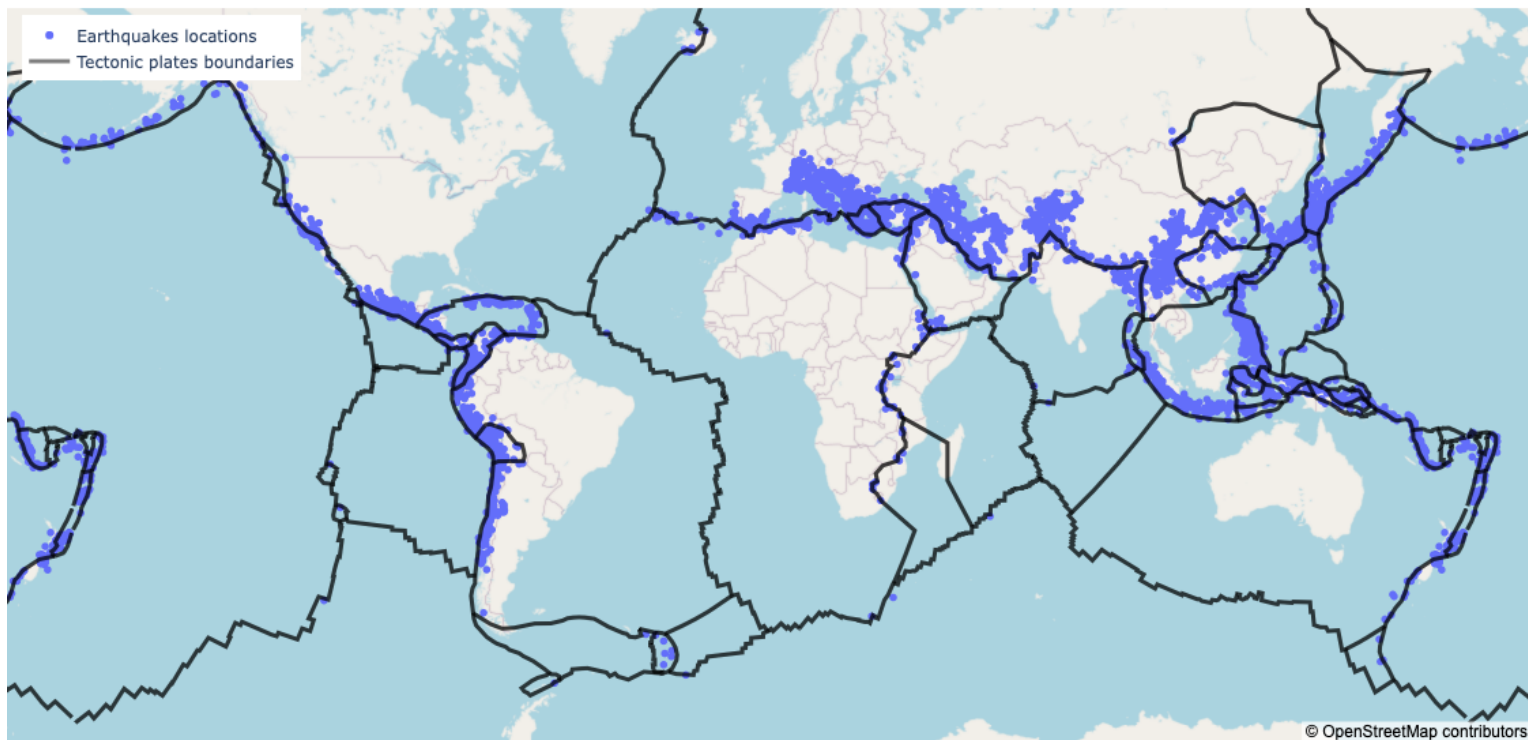


Рис. 2.2: Результаты после удаления точек, которым не нашлось кластера

3 Разбиение на кластеры и проверка гипотезы.

Вспомним поставленную нами гипотезу: на восточной границе Евразийской плиты, а именно у места ее контакта с Филиппинской, Тихоокеанской и Северо-Американской, средняя магнитуда землетрясений сильно выше, чем у ее южной границы. Попробуем кластеризовать землетрясения так, чтобы выделились группы вдоль южной границы и группа около восточной.

3.1 Разбиение с помощью KMeans

Одним из наиболее популярных алгоритмов кластеризации является **KMeans**. Это итерационный алгоритм: исходно он случайным образом выбирает k центров кластеров из всех элементов, затем на каждом шаге присваивает каждый элемент к ближайшему к нему центру.

Происходит обновление центров так, чтобы среднее расстояние от элемента до центра в новоопределенных кластерах было минимально. Дальше шаг с присоединением к ближайшему центру повторяется. Подробно алгоритм описан здесь [3].

Из-за минимизации среднего расстояния до центра внутри каждого шага алгоритм склонен формировать круглые кластера, что в нашем случае скорее является недостатком. Однако практические тесты показывают, что несмотря на это, отделение землетрясений в интересующих нас зонах все еще прошло успешно.

Дополнительно алгоритм требует вручную задать количество кластеров, на которые нужно разбить, выбрано 15 как количество тектонических плит (дает наилучшее разбиение из обозримого набора опций).

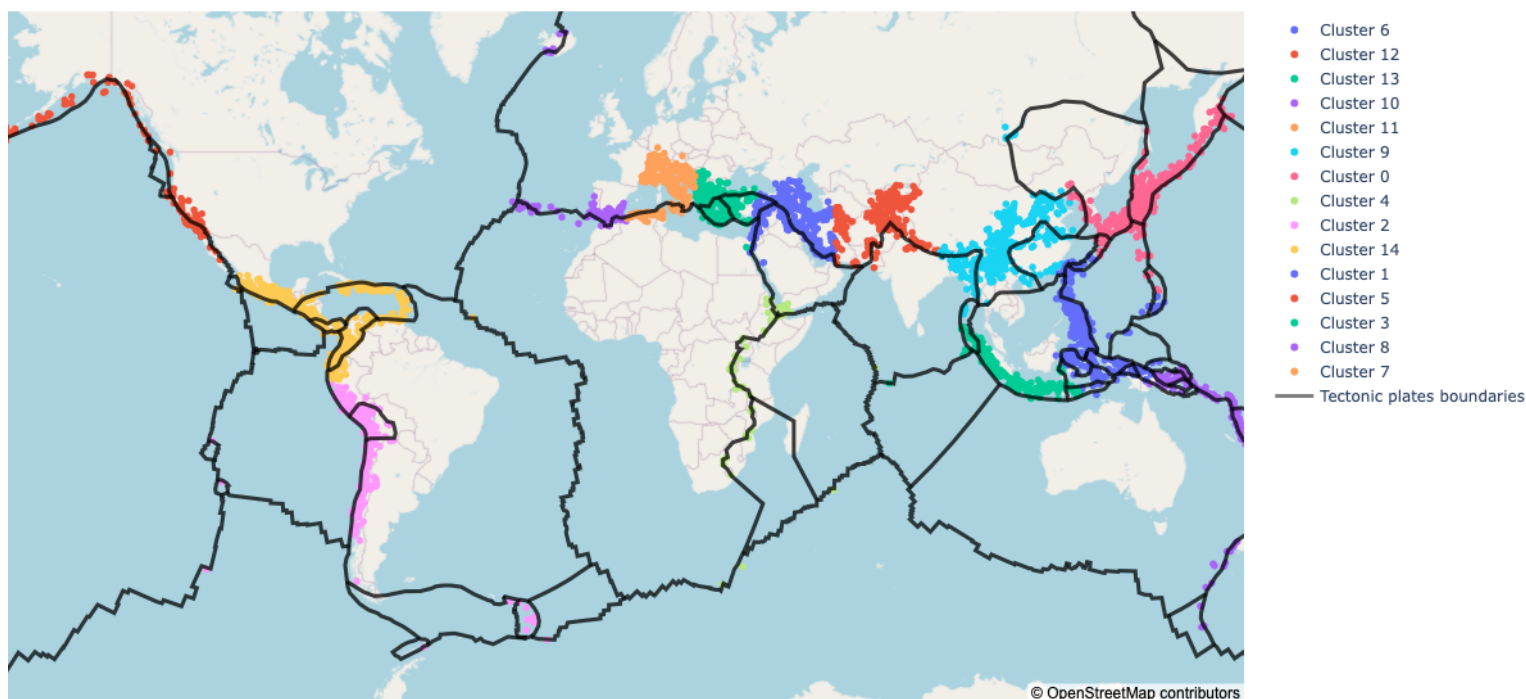


Рис. 3.1: Результаты после применения KMeans с 15 кластерами

Заметим, что алгоритм сработал достаточно успешно: на стыке Южно-Американской плиты с ее западными соседями, плитой Наска и Карибской плитой алгоритм правильно выделил границы землетресений. Однако есть и недочеты - алгоритм не совсем точно определил границы кластера около Индостанской плиты, он немного залезает на границу с Аравийской (на Рисунке 3.1 красные точки, кластер 12).

Перейдем к интересующим нас областям:

- **Кластер 0** (розовый): Стык Евразийской, Тихоокеанской, Североамериканской и Филлипинской плиты.
- **Кластеры 10, 11, 13** (пурпурный, оранжевый и салатовый слева направо): Стык Евразийской и Африканской плиты. Далее рассматриваем их как одну группу.
- **Кластер 6** (фиолетовый): Стык Евразийской и Аравийской плиты.
- **Кластер 12** (красный): Стык Евразийской и Индостанской плиты.

Посчитаем средние значения в каждом кластере, попробуем проверить идею, что на востоке Евразийской плиты средняя магнитуда выше, чем в остальных местах. Для проверки существенности различия воспользуемся двухвыборочным t-критерием для независимых выборок [6] (или его вариацией, применяемой, когда у выборок различная дисперсия [7]).

Предварительно посчитаем в каждой из групп выше среднееквадратичное отклонение (ско) и среднее значение (нужно для выбора, какой t-test применить, но и несет интересную информацию само по себе).

Действительно, сразу можем заметить, что в нашем кластере на востоке Евразийской плиты среднее значение магнитуды на один балл по шкале Рихтера выше, чем во всех других группах. Применим t-test, чтобы показать значимость этого отличия (используем версию [6] если разница между ско меньше 0.02, иначе используем [7]), посчитаем p-value.

Таблица 3.1: Расчет среднего и ско для магнитуд землетрясений внутри выбранных групп

	Восток Евразийской	Евр. и Африканская	Евр. и Аравийская	Евр. и Индостанская
Среднее	7.002	6.0244	5.8307	6.0972
СКО	0.7943	0.9056	0.8414	0.9065

Посчитаем по всем парам из наших групп p-value для гипотезы, что данные в выбранных двух кластерах имеют одинаковое среднее. Мы выбираем уровень значимости в 0.01, поэтому все значения меньше 0.0001 будем считать равными нулю (они в любом случае опровергают нашу гипотезу о равенстве средних).

Таблица 3.2: Расчет p-value для пар кластеров с помощью t-test'ов

p-value	Восток Евразийской	Евр. и Африканская	Евр. и Аравийская	Евр. и Индостанская
Восток Евразийской	-	0	0	0
Евр. и Африканская	0	-	0.0008	0.2679
Евр. и Аравийская	0	0.0008	-	0.0001
Евр. и Индостанская	0	0.2679	0.0001	-

Видим, что Восточная граница Евразийской плиты действительно значимо отличается по средней магнитуде землетрясений от ее Южной границы (от всех областей). Мы так же дополнительно выявили, что землетрясения около стыка Евразийской и Аравийской плиты в среднем имеют меньшую магнитуду, чем около других частей восточной границы Евразийской.

Одной из причин более сильных землетрясений на восточной границе Евразийской плиты может быть столкновение с Тихоокеанской, которая крайне массивная и порождает вокруг себя упомянутое выше Тихоокеанское вулканическое огненное кольцо, однако это утверждение требует дополнительных исследований.

Таким образом, даже самый простой алгоритм кластеризации KMeans позволил нам проверить нашу геологическую гипотезу. Попробуем перепроверить ее, разбив на кластера с помощью уже известного нам DBSCAN.

3.2 Разбиение с помощью DBSCAN

Применим уже описанный в разделе 2 алгоритм DBSCAN для другой кластеризации. Вспомним, что DBSCAN лучше выделяет кластера необычной формы, что должно положительно сказаться на качестве кластеризации в нашем случае.

Может показаться, что снова выгодно добавить точки на границах в наш датасет как искусственные землетрясения, чтобы алгоритм лучше кластеризовал вдоль границ литосферных плит. Однако реальные тесты показывают, что это практически всегда приводит к излишнему объединению кластеров, и становится невозможным анализ, например, только восточной границы Евразийской плиты, т.к. она объединяется с другими участками.

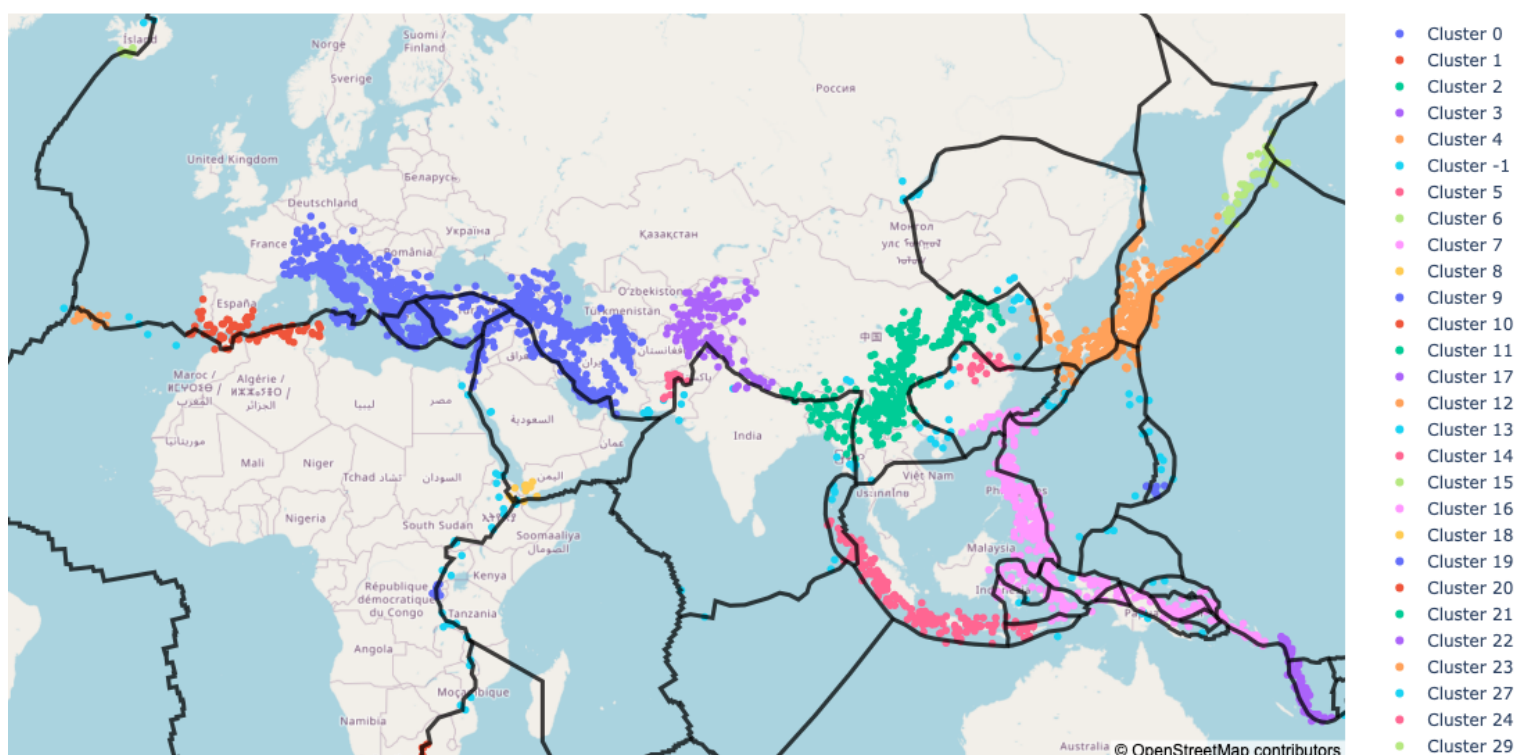


Рис. 3.2: Результаты после применения DBSCAN с 8 minPts и 2 eps

Заметим, что DBSCAN определил мало точек вне кластеров, что говорит о малом количестве выбросов (если алгоритм обнаруживает точку с малым количеством соседей на расстоянии меньше eps, он определяет ее в кластер -1).

Снова сравним концентрацию землетрясений на пересечении 4 плит и южную границу Евразийской плиты на значимые отличия в средней силе землетрясений.

Перейдем к интересующим нас областям:

- **Кластер 4** (оранжевый): стык 4 тектонических плит (включая восток Евразийской).
- **Кластер 1 и 0**: (красный и синий) стыки на южной границе Евразийской плиты с Африканской и Аравийской (здесь эти две области объединены в один кластер).
- **Кластер 3** (фиолетовый): стык Евразийской и Индостанской плиты

Таблица 3.3: Расчет среднего и ско для магнитуд землетрясений внутри выбранных групп

	Восток Евразийской	Евр. и Африканская/Аравийская	Евр. и Индостанская
Среднее значение	6.9447	5.9557	6.0789
СКО	0.7787	0.8779	0.8099

Посчитаем среднее значение магнитуды по шкале Рихтера и ско магнитуды в обозначенных областях. Снова видим, что среднее значение в области на востоке Евразийской плиты (и граничащих с ней там плит) практически на единицу по шкале Рихтера больше, чем в остальных рассматриваемых областях. Проверим заключение аналогично тесту для кластеризации с KMeans [3.1](#)

Таблица 3.4: Расчет среднего и ско для магнитуд землетрясений внутри выбранных групп

p_value	Восток Евразийской	Евр. и Африканская/Аравийской	Евр. и Индостанская
Восток Евразийской	-	0	0
Евр. и Африканская/Аравийской	0	-	0.0715
Евр. и Индостанская	0	0.0715	-

И тогда при уровне значимости в 0.01 снова подтверждаем нашу гипотезу, что на востоке Евразийской плиты (ее стыке с тремя плитами) землетрясения в среднем имеют большую магнитуду.

Таким образом оба алгоритма кластеризации разбивают наши точки на достаточно удобные для анализа зоны. Несмотря на это, нам хотелось бы каким-то образом оценивать качество кластеризации. Для этого существует целый набор метрик, часть из которых будут рассмотрены далее.

4 Оценка качества кластеризации

Попробуем оценить качество разбиения на кластеры с помощью DBSCAN и KMeans.

Для оценивания качества кластеризации выберем следующие метрики. Заметим сразу, что здесь показатели метрик не будут являться достоверным знаком оптимальности того или другого разбиения на кластеры, так как мы хотим разбить на кластеры нестандартных форм. Тем не менее они все еще интересны для изучения:

- **Average silhouette score:** метрика схожести объектов внутри одного кластера в сравнении с объектами из других кластеров. Чем больше эта метрика, тем лучше мы разбили на кластера. Принимает значения от -1 до 1, где 0 показывает, что кластера накладываются, а -1 - что точки определены в неверные кластера и для них есть вариант лучше.
- **Adjusted Rand Index:** эта метрика кластеризации требует знания заранее известных эталонных кластеров для части данных (для нас это границы тектонических плит, мы уже знаем, что кусочки из одной границы в идеале должны попасть в один кластер). При отсутствии корреляции между эталонными и предсказанными метриками будет получать нулевое значение, максимально же возможное - 1.

Специально не берем метрики, основанные на дисперсии внутри кластера, так как наши кластера часто имеют вытянутую форму (а значит большую дисперсию по одному из измерений), что нивелирует смысл метрик. Примером метрики из такого класса будет метрика Davies-Bouldin, про которую можно прочесть в статье ее автора [1]. В нашем случае она малоприменима, так как не очень хорошо оценивает вытянутые кластера, ведь дисперсия в вытянутом кластере вдоль одного из измерения точно будет большой.

4.1 Average silhouette score

Метрика вычисляется по следующей формуле:

$$\theta = \frac{\sum_{i \in P} \frac{b-a}{\max(a,b)}}{|P|} \quad (1)$$

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 .

Посчитаем эту метрику для результата KMeans с различными параметрами искомого количества кластеров.

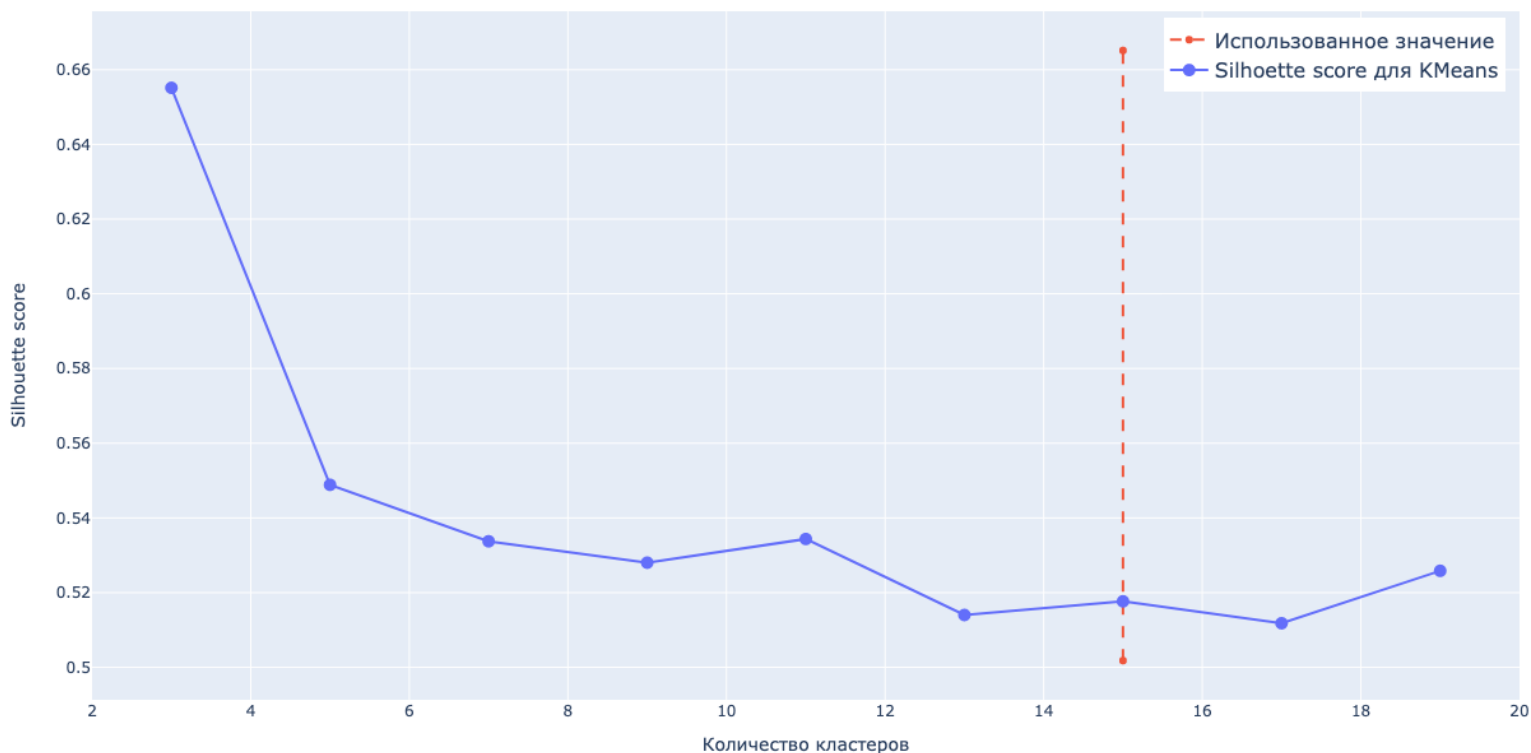


Рис. 4.1: Метрика Average silhouette score для запусков KMeans

Заметим, что использованное в главе 3.1 значение далеко не является оптимальным с точки зрения метрики. Действительно, наиболее оптимальные значения достигаются на малом количестве кластеров. Но при таком количестве кластеров мы не сможем проанализировать интересующие нас участки.

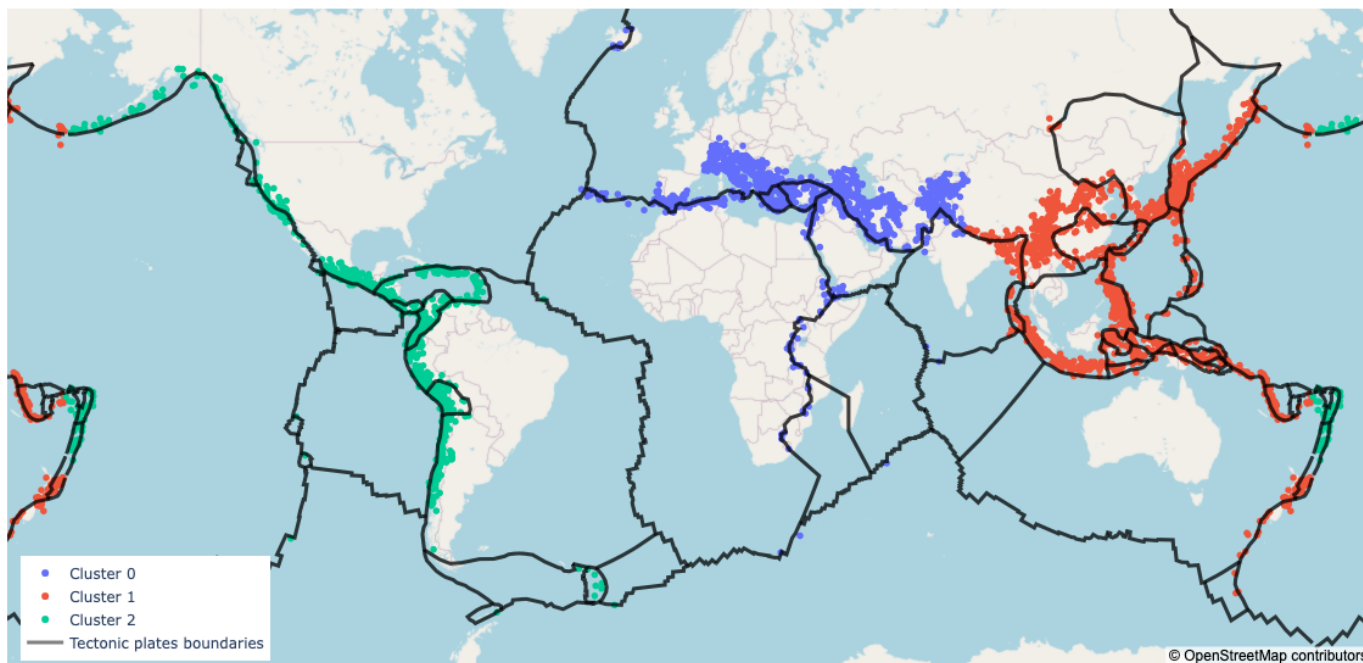


Рис. 4.2: KMeans при 3 кластерах. Можем анализировать только на масштабе континентов

Тем не менее используемое в работе значение 15 все еще отражает достаточно адекватную кластеризацию - значение 0.5 показывает, что для большинства точек расстояние до их центра кластера ближе, чем до других.

Теперь посчитаем значение той же метрики для алгоритма DBSCAN. Здесь у нас есть уже два параметра: `eps`, определяющий максимальное расстояния между связанными точками, и `minPts`, определяющий минимальное количество точек для формирования кластера. При этом заметим, что DBSCAN может не сопоставить части точек никакой кластер. Такие точки перед подсчетом метрики мы убираем.

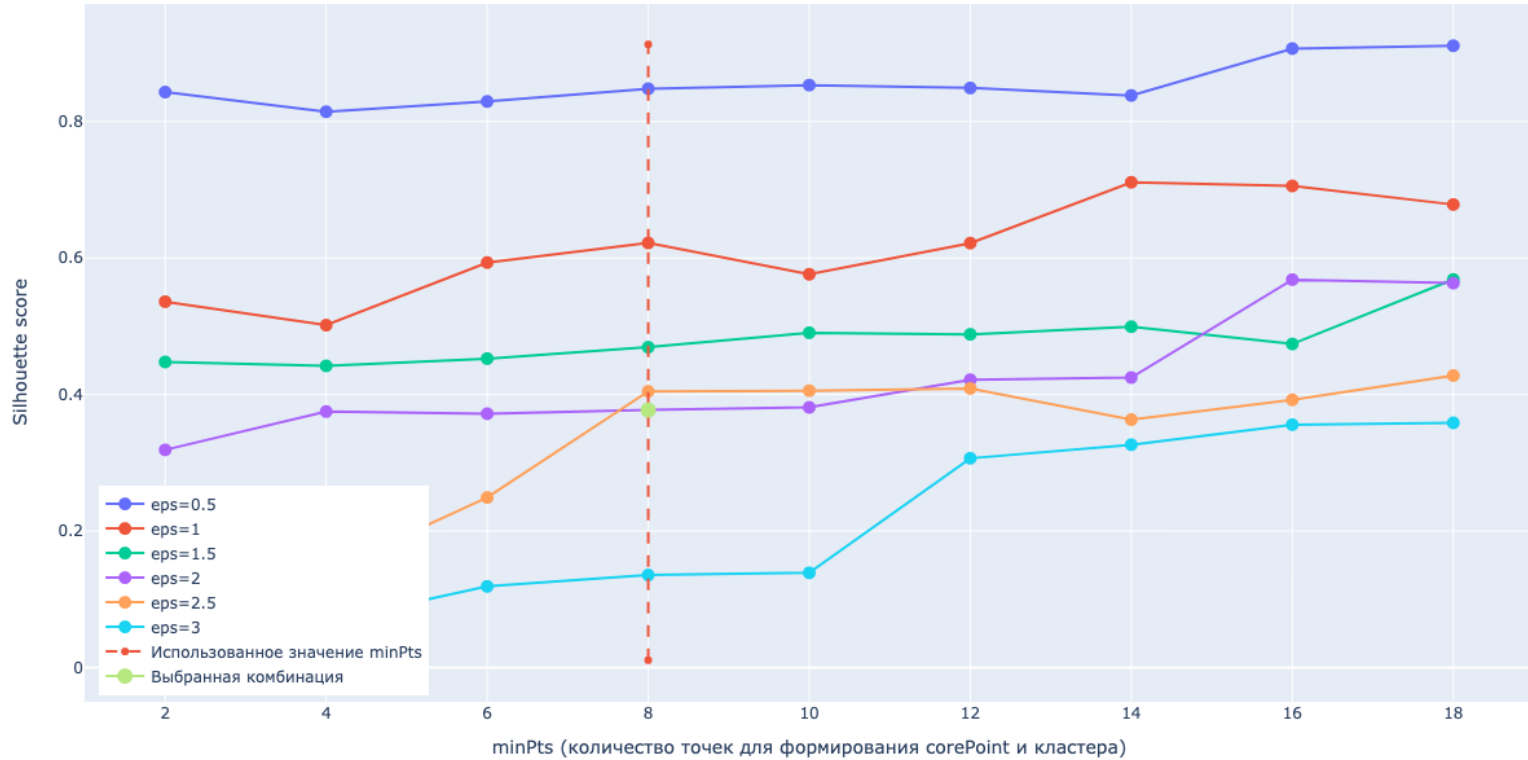


Рис. 4.3: Метрика Average silhouette score для запусков DBSCAN

Заметим, что есть четкая зависимость значения метрики от `eps`. Это легко понять - чем меньше `eps`, тем больше точек без четкого кластера DBSCAN просто удалит, оставив только наиболее яркие скопления. Вдобавок при уменьшении `eps` будет сильно расти количество кластеров, что тоже не является хорошим признаком - алгоритм становится слишком чувствительным и разделяет объекты из одной области (например, в нашем случае, из исследуемой восточной границы Евразийской плиты).

Тем не менее видим, что выбранная нами комбинация (пересечение пунктирной и фиолетовой линии на Рисунке 4.5) имеет результат около 0.4, что хотя и говорит о не очень хорошей кластеризации, все еще показывает наличие смысла в разбиении. Как видно по Рисунку 4.4, она еще и сохраняет большую часть данных.

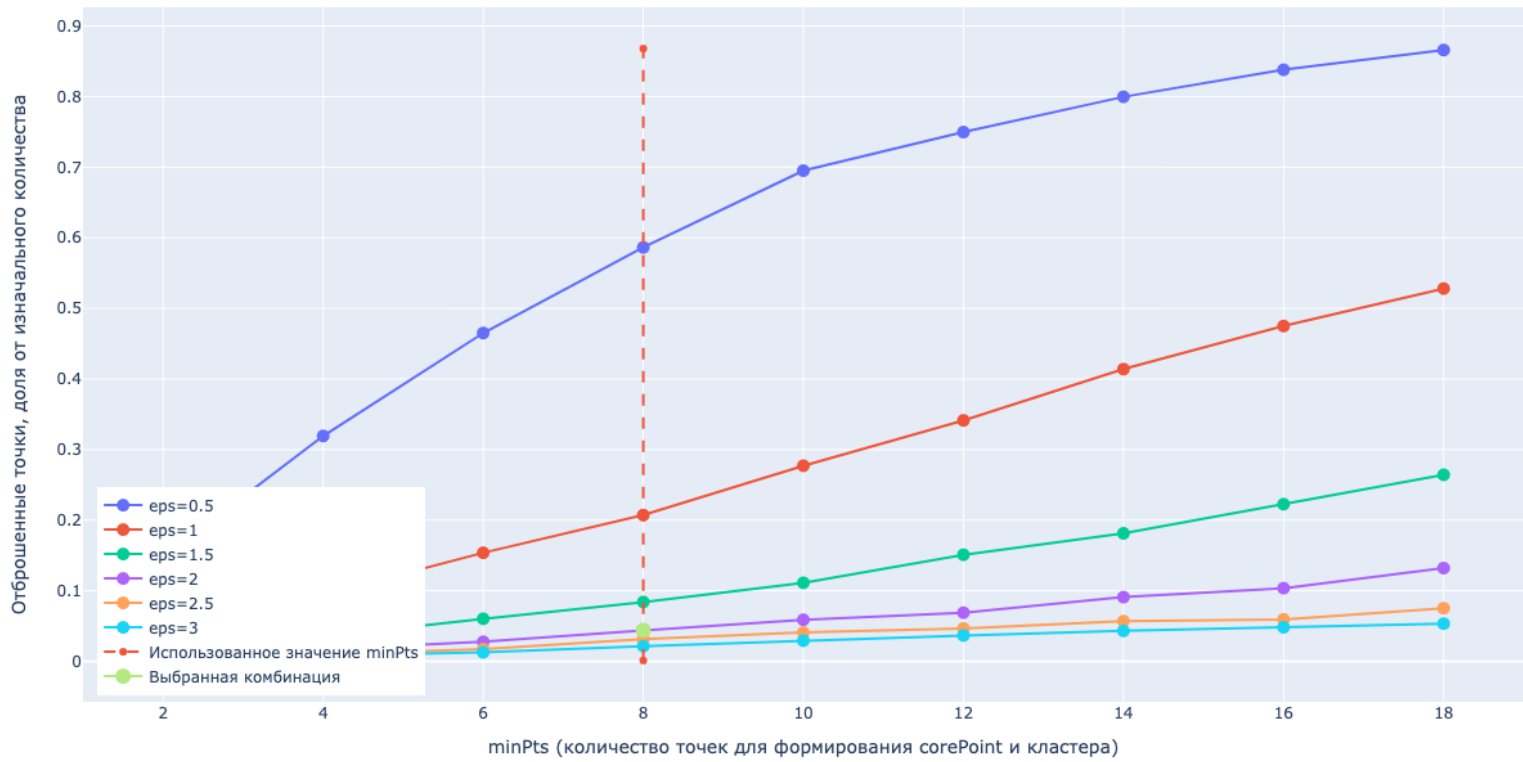


Рис. 4.4: Доля отброшенных точек для запусков DBSCAN

4.2 Adjusted Rand Index

Эта метрика требует наличия исходной информации о правильном разбиении кластеров. В нашем случае мы добавим часть точек, которые расположены на границах, и которые в датасете [5] уже размечены по литосферным плитам, которым они принадлежат. Подробно метрика разобрана в главе курса [4].

При том что у нас есть набор точек, для которых мы знаем оптимальную кластеризацию и предсказанную кластеризацию, мы можем попробовать оценить совпадение предсказанной с оптимальной.

Чтобы посчитать эту метрику, сначала потребуется ввести метрику **Rand Index**:

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

где n – количество точек в наборе данных, k – количество пар, для которых мы правильно установили принадлежность/не принадлежность одному кластеру.

Adjusted Rand Index вносит поправку в эту метрику, добавляя свойство стремления к нулю у случайно выставленного набора меток (предыдущий вариант этим не обладает). А именно он выглядит как:

$$\theta = \frac{RI - E[RI]}{\max(RI) - E[RI]} \quad (3)$$

Для того, чтобы добавить предварительно размеченный набор точек к нашим данным, было проделано следующее:

- Из данных границ плит были убраны границы всех больших плит (Евразийская, Тихоокеанская, и другие), так как их точки с одной меткой разбросаны по всей карте и сбивают метрику.

- Были оставлены только плиты около Евразийской литосферной плиты, чтобы оценить качество разбиения на кластера именно в этом участке.
- Среди оставленных точек на границах были выбраны случайные 10% и добавлены в наш датасет землетрясений (от него они составляют около 5 процентов и на кластеризацию не влияют).
- Уберем дубликаты по локации: в наших данных каждая граница включена два раза, относительно одной и другой литосферной плиты. Отсюда появляются две точки с одинаковой локацией и разными метками. Отсортируем все точки по плите, которой они принадлежат, и из каждой пары с одинаковой локацией выберем нижнюю в табличке. Тем самым каждая граница останется только для одной литосферной плиты, а для другой целиком удалится.

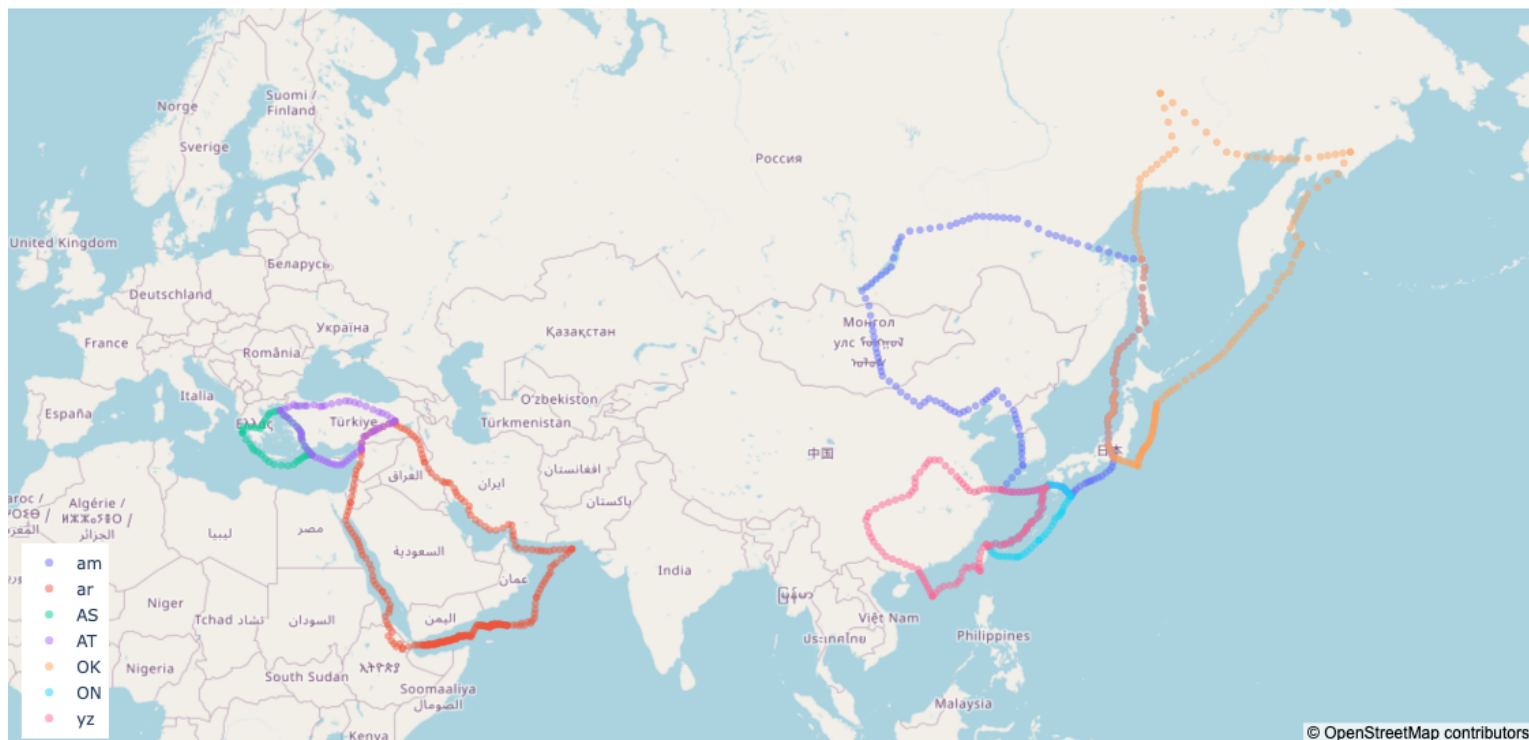


Рис. 4.5: Границы, точки вдоль которых попадут в случайную выборку при добавлении

Оценим качество алгоритма KMeans с помощью ARI. По Рисунку 4.6 видно, что оптимум достигается на значениях около 8, а выбранное нами значение 15 на самом деле находится в некотором минимуме. Построим разбиение с 9 кластерами и посмотрим, действительно ли оно лучше группирует нужные нам сектора.

Видим на Рисунке 4.8, что на самом деле в таком разбиении в фиолетовый сектор, который находится на востоке Евразийской плиты, попадают лишние землетрясения с границы Тихоокеанской плиты (расположены на карте южнее). Они помешали бы нам в наших исследованиях, а значит несмотря на лучшее значение метрики, это разбиение подходит нам меньше.

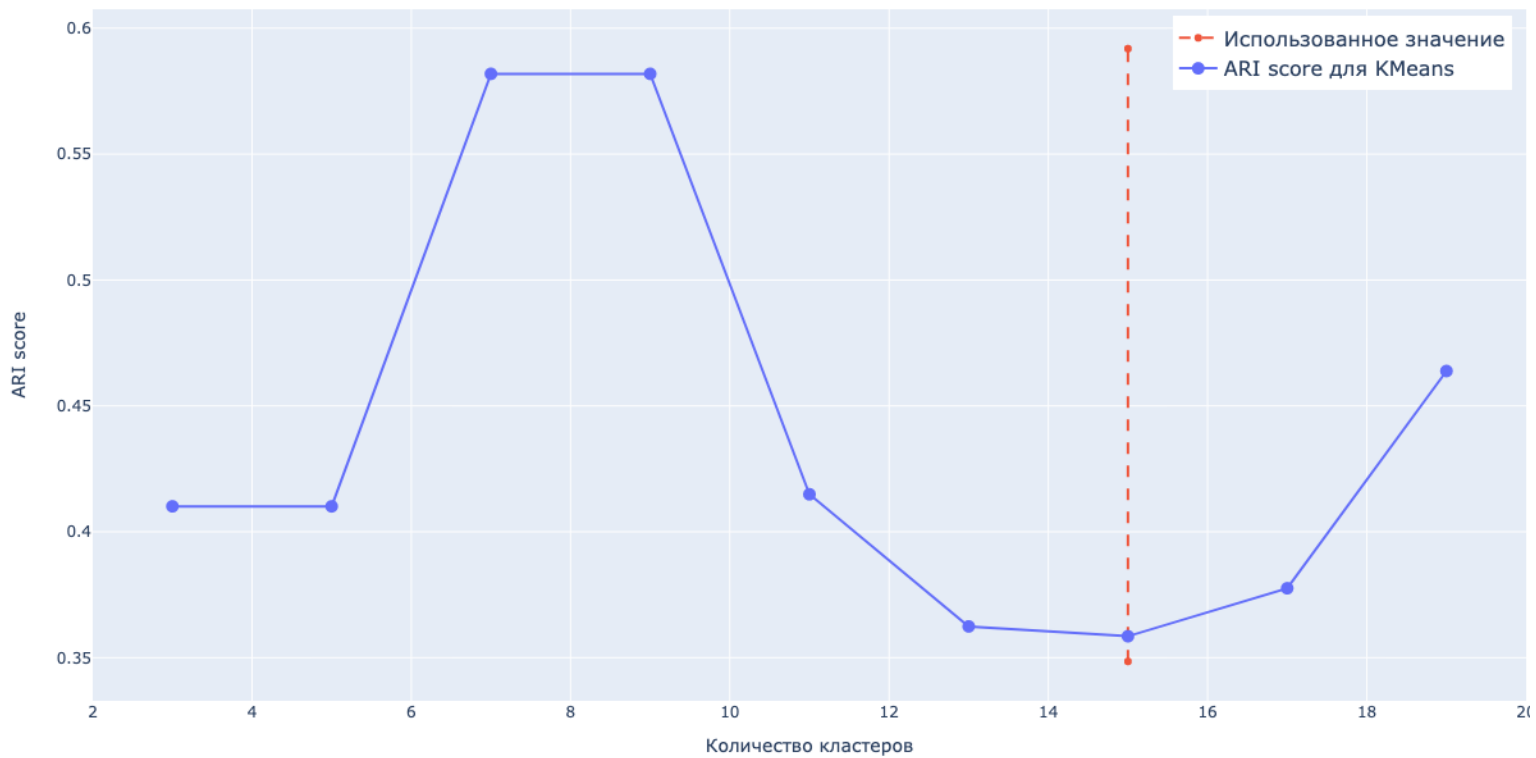


Рис. 4.6: ARI score для KMeans с различными параметрами

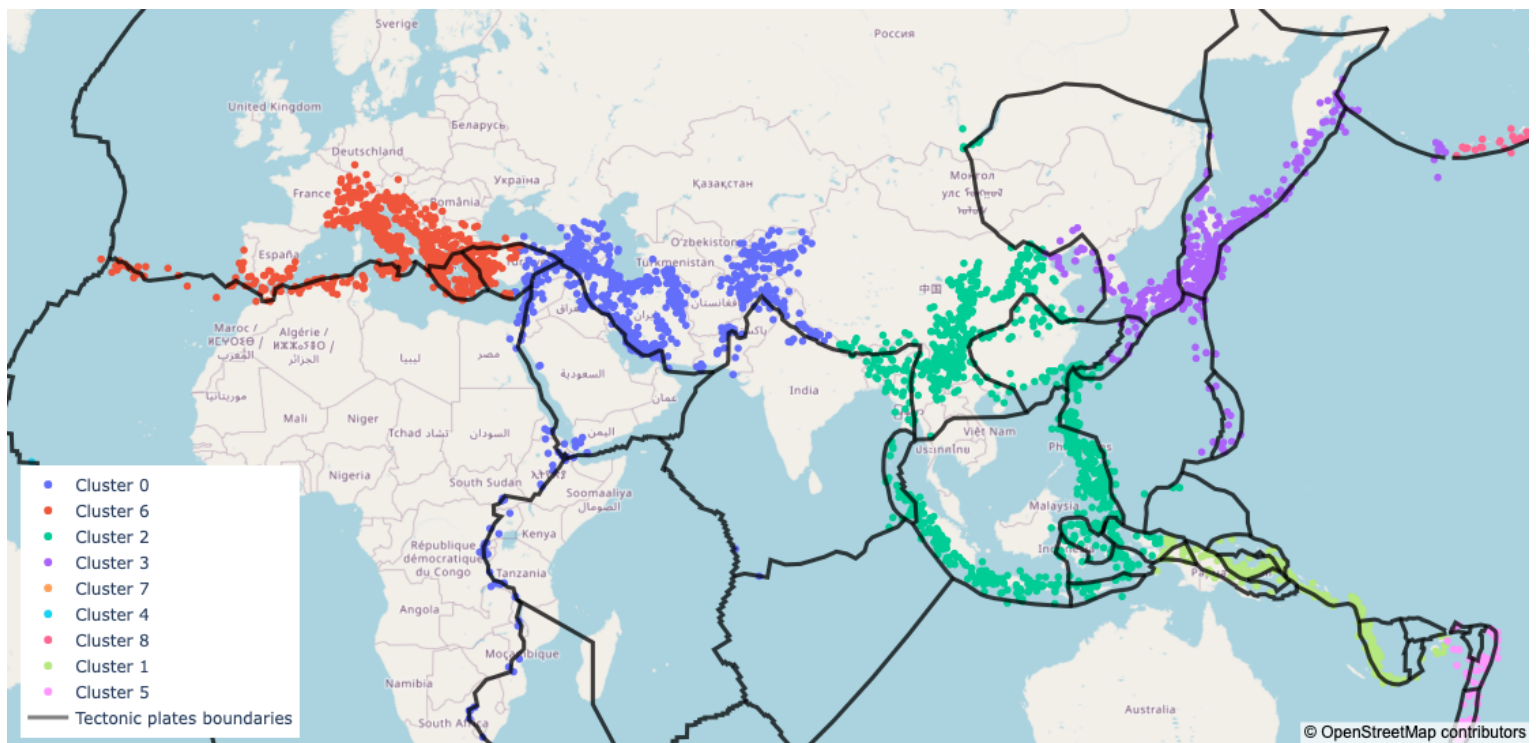


Рис. 4.7: KMeans с 9 кластерами

Заметим, что метрика демонстрирует неплохие значения - мы сильно удалены от нуля и приближаемся к 0.5, а значит получилось сравнительно неплохо разбить на группы.

Наконец, применим нашу метрику к алгоритму DBSCAN. Снова перед применением метрики удаляем точки, которым алгоритм не нашел кластера. Примерная картина доли удаленных точек в зависимости от параметров алгоритма отражена на Рисунке 4.4 (точный график потерь можно найти в приложении 5).

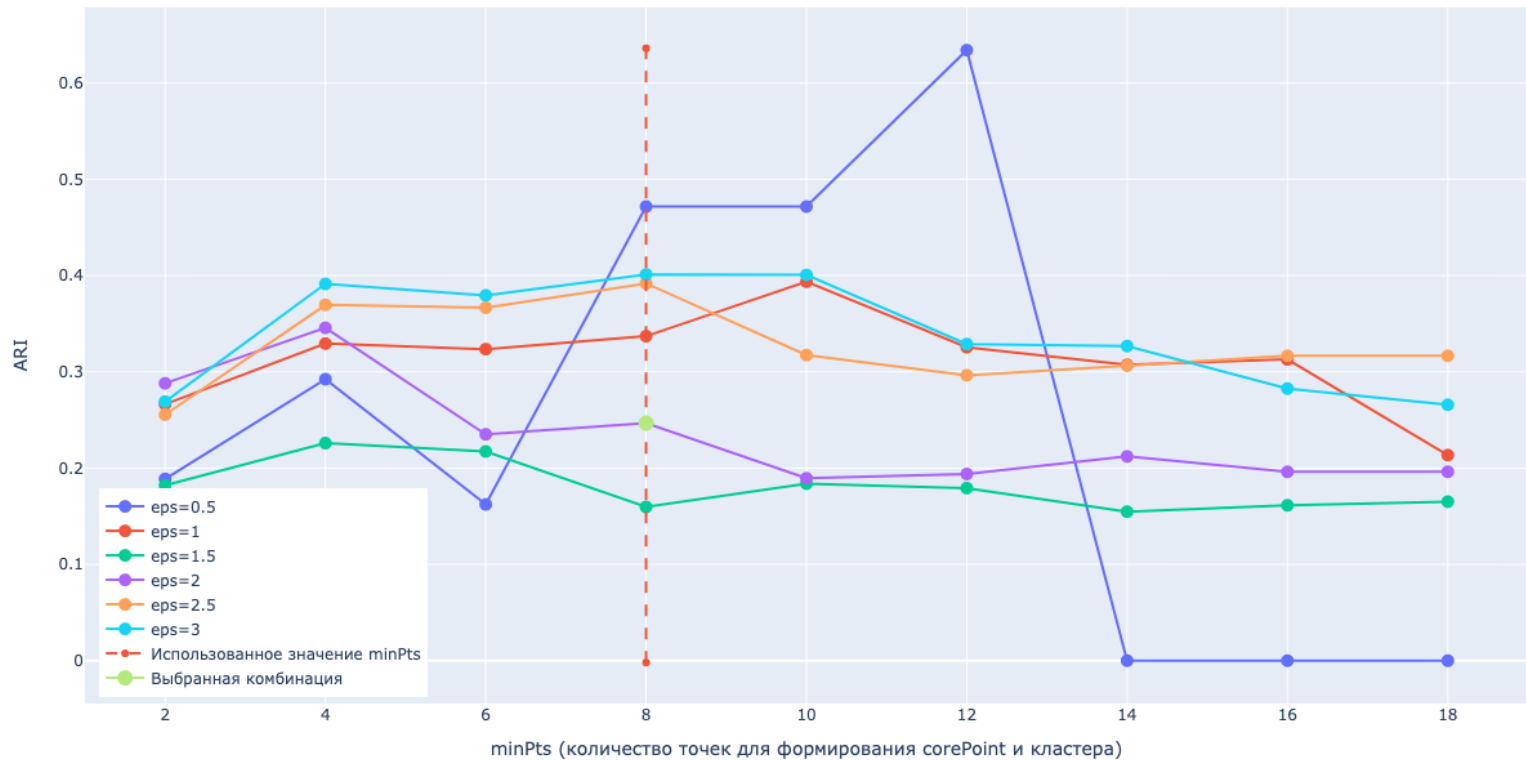


Рис. 4.8: ARI score для DBSCAN с различными параметрами

Заметим, что здесь результаты получаются ниже, чем у KMeans. Это обусловлено вытянутостью кластеров DBSCAN, из-за чего точки из несколько различным литосферных плит на границе Евразийской плиты попадают в один кластер и учитываются метрикой как некорректные. Это можно проследить, например, сравнив графики 3.2 и 3.1 - KMeans четче разделяет кластера на границе Евразийской с Африканской и Аравийской плитами.

Аналогично метрике Average silhouette score получаем лучшие значения при меньшем eps, но снова при этом теряем большую часть данных (не определяем их ни в какой кластер). В частности, отсюда падает в ноль график с $eps = 0.5$: с таким параметром и большим minPts алгоритм не находит ни одного кластера и удаляет все точки. Тем не менее, снова для всех запусков алгоритма значения отличимы от шума, то есть алгоритм все-таки может с некоторым успехом отличать границы плит, однако для детальных исследований ему требуется ручная корректировка (как мы и делали в данной работе).

5 Заключение. Дальнейшая работа

Методы анализа данных уже широко применяются в географии и геологии, и, в частности, в сейсмологии. Результаты данной работы призваны продемонстрировать, что поле для применения анализа данных очень широко, и позволяет рассматривать даже вопросы континентального масштаба.

В частности, в рамках этой работы была поставлена и подтверждена гипотеза, что наиболее высокая по средней магнитуде землетрясений часть Евразийской плиты находится на востоке, на месте ее стыка с Тихоокеанской, Филиппинской и Северно-Американской.

В процессе были применены два алгоритма кластеризации, проведено детальное сравнение результатов. Для численного анализа успеха кластеризации были изучены и применены две метрики. Обе метрики показали отличие результатов алгоритма от шума, но при этом большой простор для улучшений.

Несмотря на весь потенциал анализа данных, он так же содержит в себе множество 'опасностей'. Необходимо тщательно очищать данные, иногда применяя нестандартные решения, как, например, в данном случае при удалении землетрясений вдалеке от разломов 2. Тем не менее, после детальной очистки данные представляют собой настоящую кладезь для исследований.

В качестве дальнейшей работы можно рассмотреть попытку применения других алгоритмов к задаче кластеризации. Или можно взять кардинально другой подход и поставить задачу классификации. Необходимо взять набор уже подготовленных точек (в нашем случае, например, границ тектонических плит) и натренировать на них модель. Затем отдать этой модели остальные координаты землетрясений и проанализировать, насколько хорошо она смогла сопоставить их ближайшим разломам. Результаты можно сравнить с результатами данной работы.

Приложение

Больше тестов и полные результаты численных экспериментов, а также код алгоритмов можно найти по ссылке в репозитории:

https://github.com/EgorBugaev/earthquakes_clusterization

Список литературы

- [1] David L. Davies и Donald W. Bouldin. “A Cluster Separation Measure”. В: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.2 (1979), с. 224—227. DOI: [10.1109/TPAMI.1979.4766909](https://doi.org/10.1109/TPAMI.1979.4766909).
- [2] Bharath Pose. *Earthquakes -2150 BC – 2023 AD around the world*. URL: <https://www.kaggle.com/datasets/bharathposa/earthquakes-from-2150bc-2023-ad-around-the-world> (дата обр. 10.05.2013).
- [3] Arif R. “Step by Step to Understanding K-means Clustering and Implementation with sklearn”. В: *Data Folks Indonesia* (4.10.2020). URL: <https://medium.com/data-folks-indonesia/step-by-step-to-understanding-k-means-clustering-and-implementation-with-sklearn-b55803f519d6> (дата обр. 10.05.2023).
- [4] Open Data Science. *Открытый курс машинного обучения. Тема 7. Обучение без учителя: PCA и кластеризация*. URL: <https://habr.com/ru/companies/ods/articles/325654/> (дата обр. 10.05.2023).
- [5] Curtis Thompson. *Tectonic Plate Boundaries*. URL: <https://www.kaggle.com/datasets/cwthompson/tectonic-plate-boundaries> (дата обр. 10.05.2023).
- [6] Wikipedia. *Independent two-sample t-test*. URL: https://en.wikipedia.org/wiki/T-test#Independent_two-sample_t-test (дата обр. 10.05.2023).
- [7] Wikipedia. *Welch’s test*. URL: https://en.wikipedia.org/wiki/Welch%27s_t-test (дата обр. 10.05.2023).
- [8] Soner Yildirim. “DBSCAN Clustering — Explained”. В: *Towards Data Science* (22.04.2020). URL: <https://towardsdatascience.com/dbscan-clustering-explained-97556a2ad556> (дата обр. 10.05.2023).