Документация, содержащая описание функциональных характеристик экземпляра программного комплекса, предоставленного для проведения экспертной проверки

Программа для ЭВМ

«M-Vision Pro»

Оглавление

1) Выделение аномалий на основе метода кластеризации (ClusterModel) 3

2) Выделение аномалий с на основе характеристик сигнала (XGBoost) 8

3) Определения типа аномалии и степени ее опасности 10

4) Выявление некорректных данных на основе корреляции 11

5) Классификация степени опасности аномалий с использованием автоматического машинного обучения 15

6) Классификация степени опасности аномалий с использованием базы данных аномальных зон 15

1) Выделение аномалий на основе метода кластеризации (ClusterModel)

Первый этап детектирования аномалий – это DataTransformer. Как видно на рисунке 1, входные данные достаточно неоднородны и зашумлены.

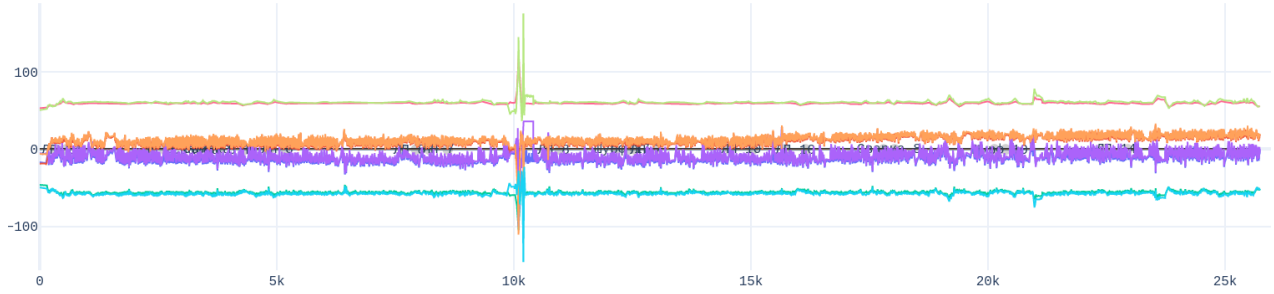

Рисунок 1 - Пример необработанных данных

DataTransformer выполняет с данными несколько действий – сглаживание. Производится несколько циклов сглаживания, их число указывается в файле конфигурации, сдвиг рядов к единому нулевому среднему значению, очистка начала и конца каждого ряда (на практике почти всегда находятся очень большие аномалии данных, которые появляются при включении или выключении приборов) и нормализация данных.

Помимо этого, на этом этапе подготовленные данные преобразуются в один ряд. Это происходит с по следующему алгоритму (см. рис. 2).

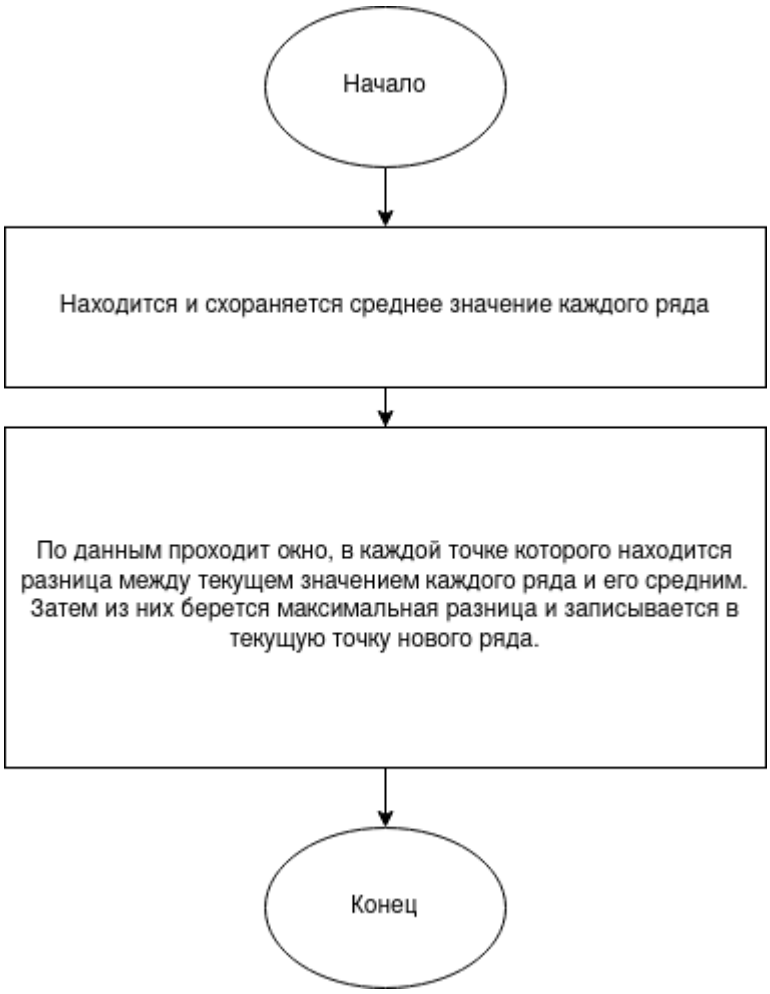

Рисунок 2 – Алгоритм преобразования данных

Пример преобразованных данных приведен на рис 3.

Рисунок 3 – Преобразованные данные. Синий график снизу показывает результат трансформации всех верхних рядов в один ряд, который отражает аномальное поведение рядов.

Следующим этапом является применение детектора зон (ZonesDetector). Этот детектор разделяет данные на аномальные зоны. Это происходит при помощи обрезания трансформированного ряда по граничному значению – все точки, которые имею значение меньше этого значения устанавливаются в 0.

Однако существует неопределенность в том, какое значение выбрать. Каждый файл данных имеет свою семантику, поэтому установка одного значения не даст приемлемой точности. Вместо этого используется адаптивный подбор граничного значения по следующему алгоритму (см. рис. 4).

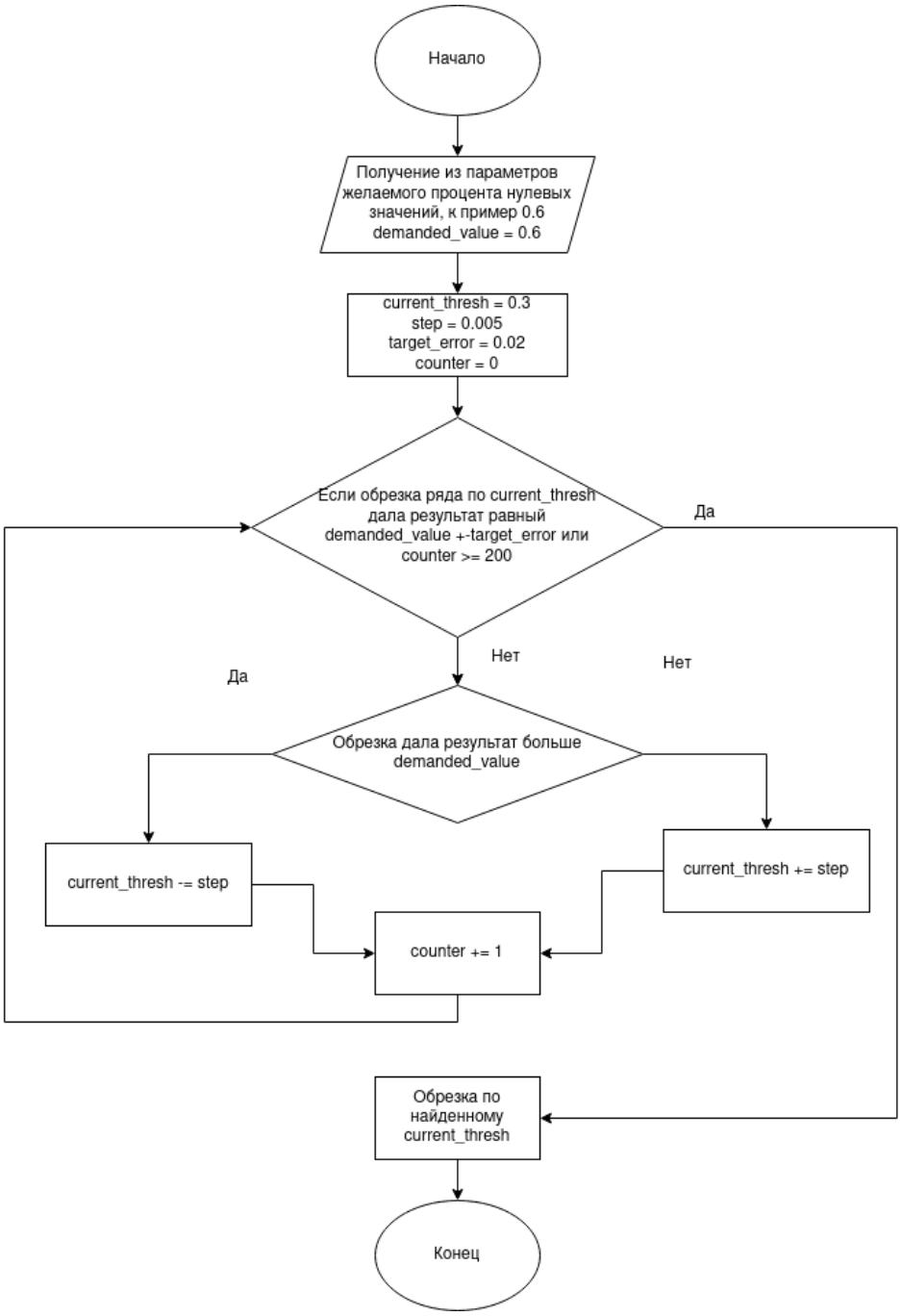

Рисунок 4 – Алгоритм подбора граничного значения

Результатом обрезки по найденному граничному значению является следующий график (см. рис. 5). Его легко поделить на отдельные зоны, которые охватывают только интересующие части данных.

Рисунок 5 – Вид данных после образки по граничному значению и нормализации

После деления на зоны, в каждую зону копируются части каждого временного ряда, которые она захватывает (см. рис. 6).

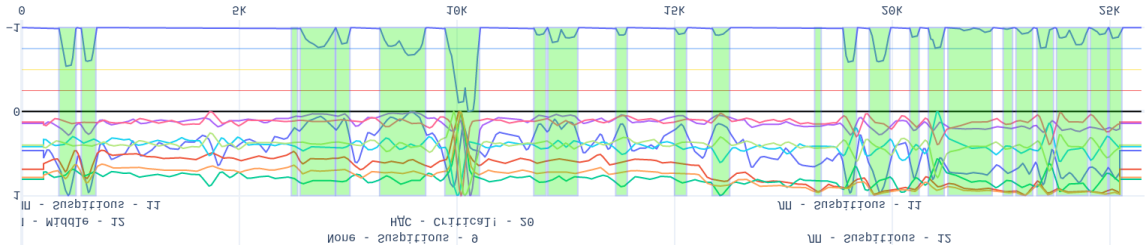

Рисунок 6 – Выделенные аномалии

Класс AnomaliesSplitte отвечает за разделение длинных аномалий на короткие. Принцип работы данного элемента основан на сканировании аномалий по небольшим последовательным частям. Среди них легко найти части, где главный график аномалии не претерпевает каких-либо изменений. Пример такого разбиения показан на рис. 7.

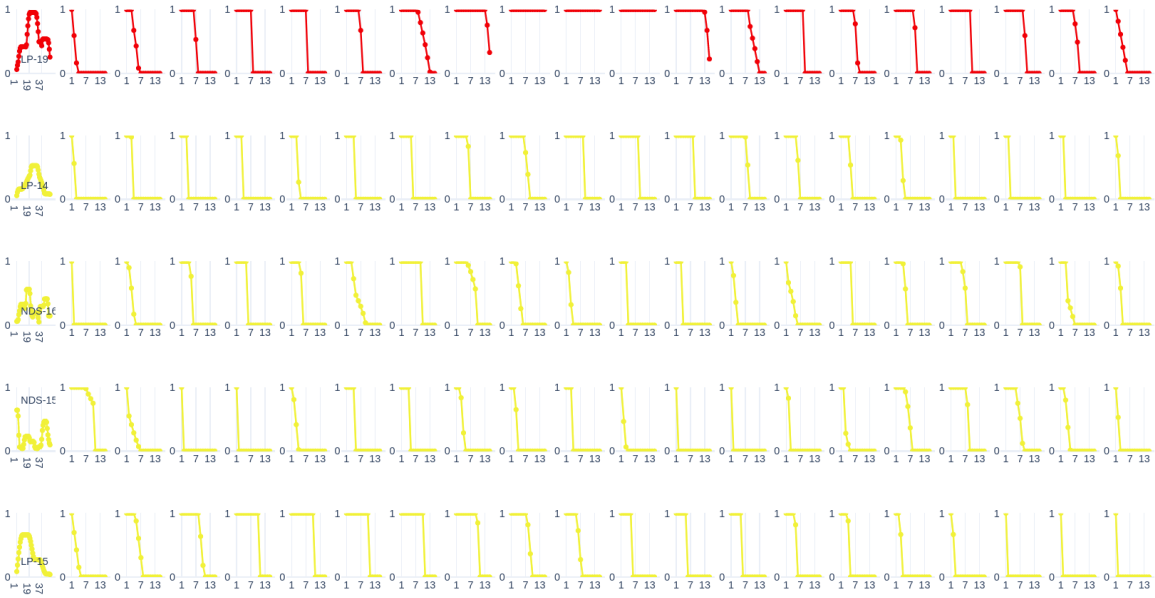

Рисунок 7 – Пример разбиения аномалий (крайний левый график) на части (остальные графики). График каждой части отображает то, насколько сильны перепады высот на соответствующей части основного графика.

Затем области с наименьшими перепадами группируются вместе и аномалии разбиваются как в примере (см. рис. 8).

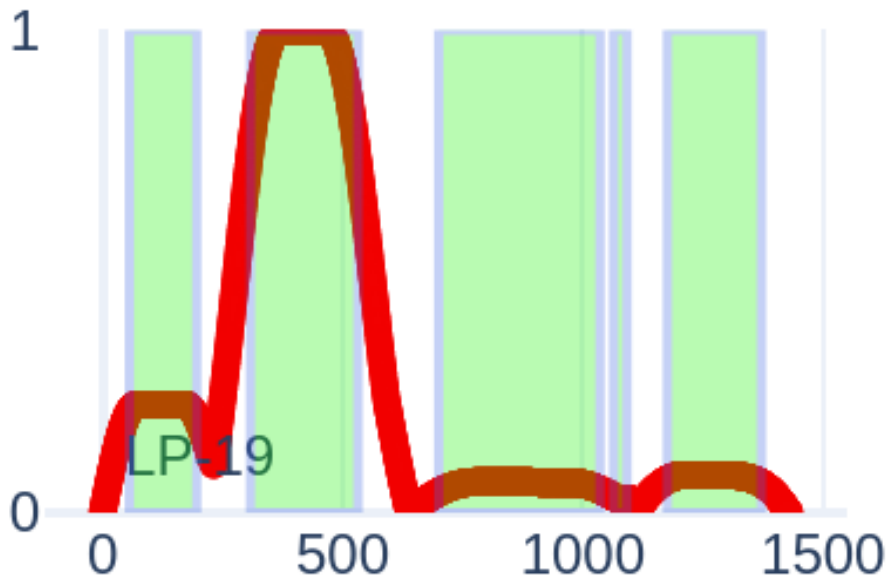

Рисунок 8 – Пример разбиения аномалии. Зеленые зоны – новые аномалии

Данные операции применяются лишь к длинным аномалиям (>3000 единиц отсчета). Единственной задачей этапа извлечения признаков (FeatureExtractor) – это извлечение признаков из главного временного ряда каждой аномалии. Кроме стандартных признаков, были использованы и дополнительные (список признаков можно найти в файле /constants/Features.py):

- LENGTH_OF_TIME_SERIES – длина аномалии.
- MAX_DELTA – дистанция между минимум и максимум ряда.
- MAX – максимум ряда
- COMPRESSING_TO_10_POINTS – сжатие ряда до десяти точек через максимум.
- SUM_OF_DELTAS_OF_TS_COMPRESSED_TO_40_POINTS – ряд сжимается до 40 точек, затем находится сумма последовательных дельт по оси y для всех точек.
- SMART_MEAN – среднее значение ряда с поправкой на значение первого элемента.
- MEAN – среднее значение ряда.
- SMART_MEDIAN – медиана ряда с поправкой на первое значение ряда.
- MEDIAN – медиана ряда.
- SMART_MEAN_MEDIAN_DIST – дистанция между средним и медианой ряда с поправками на первое значение.

– MEAN_MEDIAN_DIST – дистанция между средним и медианой ряда.

– SMART_MAX – максимум ряда с поправкой на первое значение.

– SMART_SUM – сумма ряда с поправкой на первое значение.

– SMART_LEN_SUM – длина ряда/сумма ряда с поправкой на первое значение.

– SMART_LEN_MAX – длина ряда/максимум ряда с поправкой на первое значение.

– SUM – сумма ряда.

– Q95 – 95-ый квантиль ряда.

Далее метод FeaturesReducer редуцирует вектора признаков аномалий до 2D-размерности с помощью редусера, который загружается из базы данных.

Первый метод распознавания аномалий ClusterizationPredictor – кластеризация по критичности аномалий. Редуцированные вектора признаков используются как координаты для проверки на принадлежность к зонам критичности (см. рис. 9).

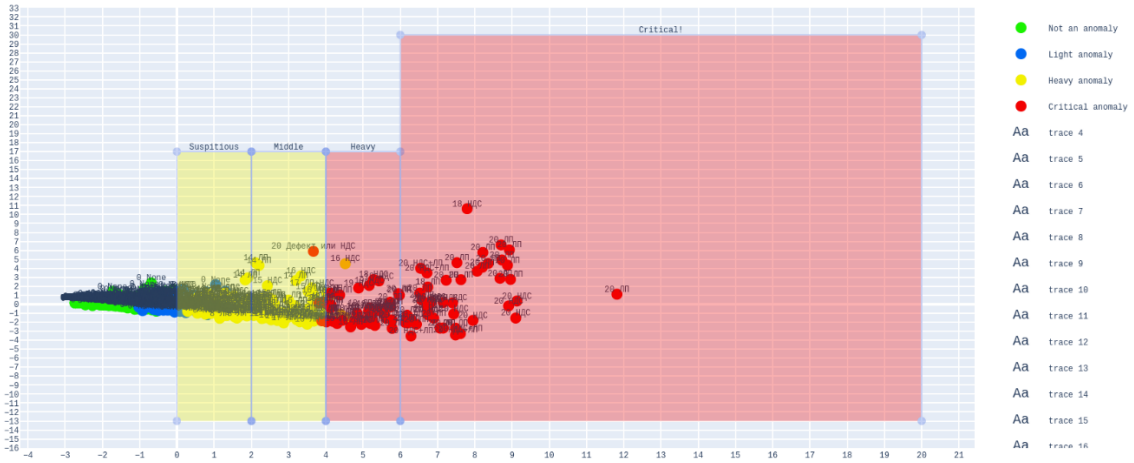

Рисунок 9 – Пример разбиения аномалии. Зеленые зоны – новые аномалии

Второй метод распознавания аномалий DatasetPredictor – проверка по базе данных. Проверяется расстояние редуцированного вектора к векторам из базы данных. Аномалии присваиваются тяжесть, тип и комментарий самой близкой аномалии из базы данных.

На следующем этапе происходит комбинирование двух методов, которые ансамблируются в итоговое предсказание (Ensambler).

Далее метод AnomaliesFilter отсеивает аномалии по степени критичности. Элемент Metrics извлекает метрики точности при наличии разметки. Расчет завершается вызовом ResultWriter, который записывает результаты работы всего *workflow* в файл.

2) Выделение аномалий с на основе характеристик сигнала (XGBoost)

Для решения задачи был выбран алгоритм градиентного бустинга и его реализация XGBoost из-за высокой эффективности в задачах классификации. Для выделения характеристик был применен метод выделения зон относительно 25% и 75% квантилей. Чтобы избежать чрезмерного пересечения для 25% квантили в анализ идут полигоны, у которых пик находится выше значения 25% квантили, а для 75% квантили полигоны, у которых пик находится ниже значения 75% квантили (рисунок 10).

Далее рассчитываются характеристики выделенных зон, которые и будут переданы в модель XGBoost в качестве признаков:

- высота пика,
- левая и правая амплитуды (относительно полигона слева и полигона справа соответственно),
- длина в количестве значений (X), длина в метрах (DIST, опционально),
- площадь полигона (X), площадь полигона (DIST, опционально),
- стандартное отклонение пика,
- отношение левой амплитуды к пику,
- отношение правой амплитуды к пику
- среднее между пиком, левой амплитудой и правой амплитудой.

Рисунок 10 – Пример выделения полигонов относительно 25 и 75 квантили

Используя экспертную разметку аномалий, полигоны помечались как аномальные или не аномальные. Если полигон был аномальным, то также в данных указывался тип и степень опасности (рисунок 11).

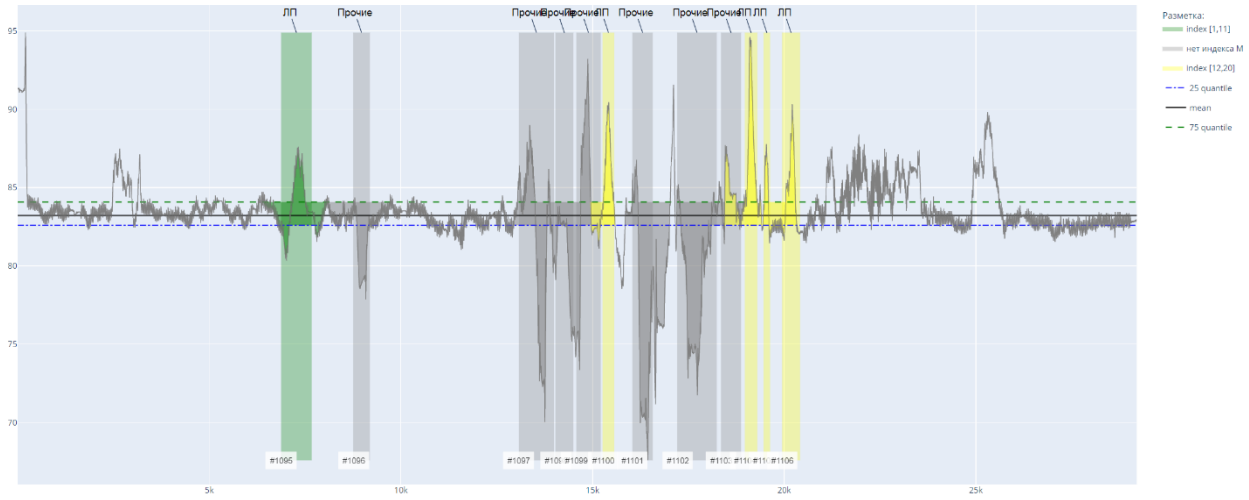

Рисунок 11 – Пример полигонов с учетом разметки

Также для настройки модели данные разделяются на обучающие и тестовые выборки (70% и 30% соответственно).

3) Определения типа аномалии и степени ее опасности

Также были реализованы две дополнительные модели основанные на XGBoost для определения типа аномалии и степени ее опасности. В таблице 1 представлены наименования типов аномалий и ранги опасностей. В качестве признаков используются те же характеристики, что описаны в методе выделения аномалий. Для обучения модели использовались только учтенные аномалии, для которых в данных были определены тип и степень опасности (см. рис. 12).

Таблица 1 – Типы аномалий и степень их опасности

Степень опасности	Тип аномалии
Index M [1,11]	ЛП
	НДС
Index M [12,20]	НДС+ЛП
	Сварка
Index M >20	Шурф
	Прочие

Так как данные содержат неравномерное количество аномалий каждого типа или степени, для улучшения результатов данные были перебалансированы, чтобы количество элементов с типами и степенями опасности в данных для обучения было примерно одинаковое.

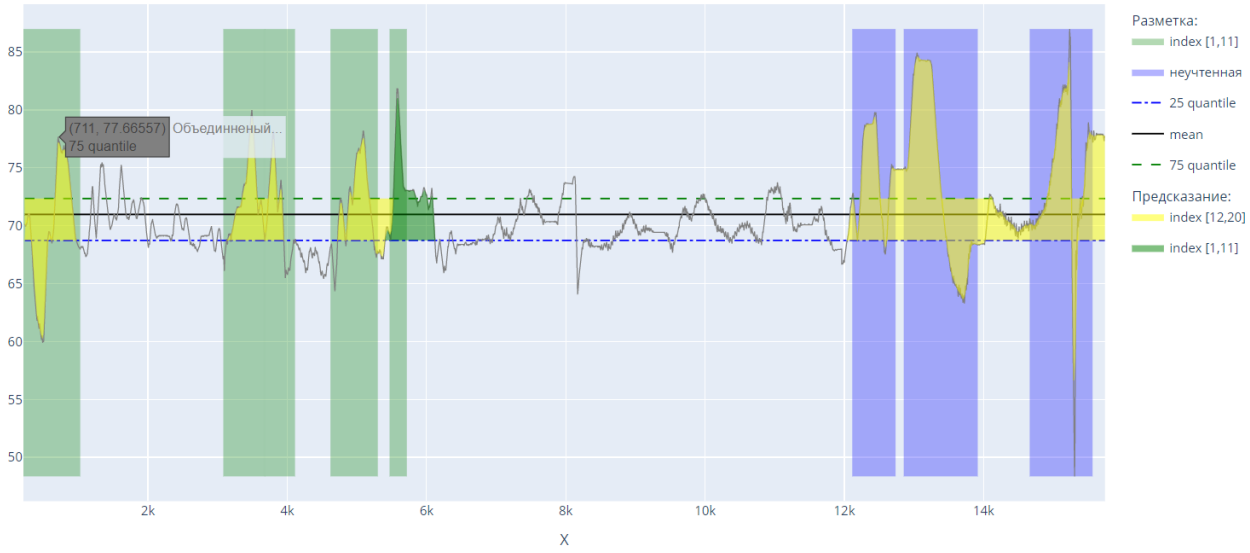

Рисунок 12- Результаты работы метода определения типа аномалий и степени опасности для файла 151118_0813

4) Выявление некорректных данных на основе корреляции

В рамках задачи выявления некорректных данных был выбран корреляционный подход. Так как сигналы Vu и Vd связаны друг с другом, их статистическая взаимосвязь должна быть высокой. Но иногда по причине внешнего воздействия или технической ошибки сигналы могут вести себя по-разному, например, начнут пересекать, в таких случаях коэффициент корреляции поможет выявить некорректность данных.

Для этого по каждому из сигналов проходимся окном определенной длины $delta$:

$$u_j = [Vu_{ij}, Vu_{ij+1}, Vu_{ij+2}, \dots, Vu_{ij+delta}],$$

$$u_{j+1} = [Vu_{ij+1}, Vu_{ij+2}, \dots, Vu_{ij+delta}, Vu_{ij+delta+1}],$$

$$u_{j+2} = [Vu_{ij+2}, Vu_{ij+3}, \dots, Vu_{ij+delta+1}, Vu_{ij+delta+2}],$$

$$\dots$$

$$d_j = [Vd_{ij}, Vd_{ij+1}, Vd_{ij+2}, \dots, Vd_{ij+delta}],$$

$$d_{j+1} = [Vd_{ij+1}, Vd_{ij+2}, \dots, Vd_{ij+delta}, Vd_{ij+delta+1}],$$

$$d_{j+2} = [Vd_{ij+2}, Vd_{ij+3}, \dots, Vd_{ij+delta+1}, Vd_{ij+delta+2}],$$

...

Далее для каждой пары сигналов рассчитываем коэффициент корреляции $r_{u_j d_j}$. Таким образом мы можем отслеживать как меняется поведение двух сигналов на протяжении всей длины. И устанавливая порог приемлемости (в нашем случае 0.5) мы можем выделять участки с некорректными данными. На рисунке 13 представлены сигналы Vu и Vd из файла «151117_1233». Для данного примера длина окна $delta$ = 200 значений.

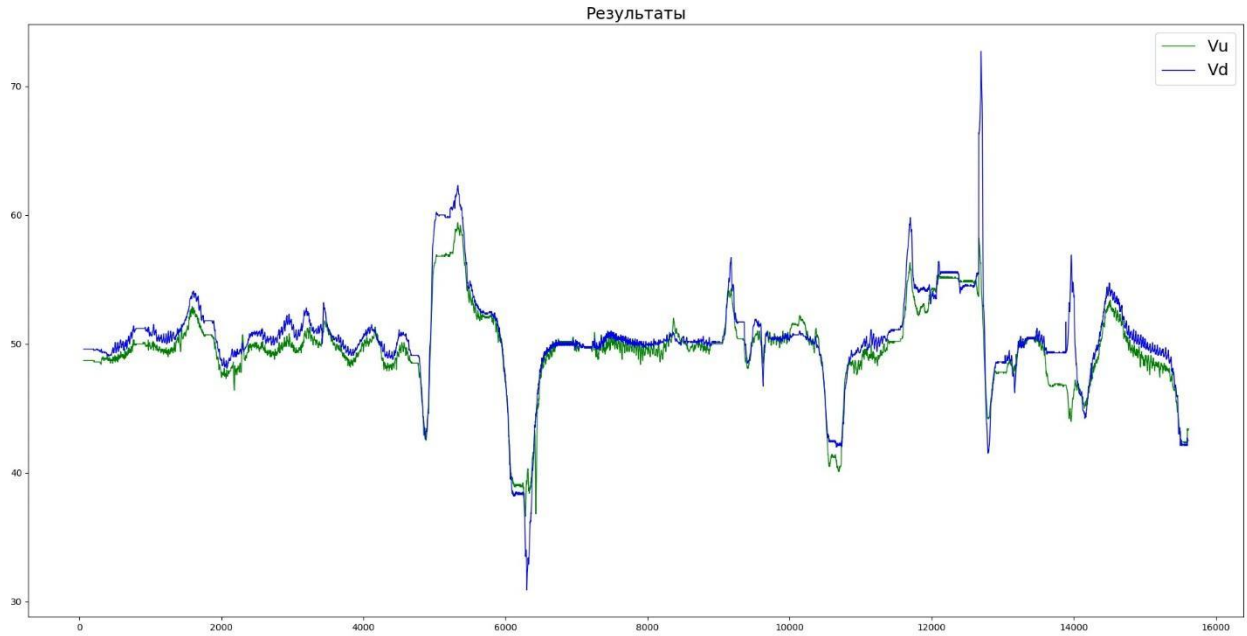

Рисунок 13 – Сигналы Vu и Vd из файла «151117_1233»

На рисунке 14 представлены результаты работы метода. Из рисунка можно увидеть, в каком месте сигналы начали себя вести подозрительно. Таким образом в файле «151117_1233» мы выделили три участка, где корреляция сигналов меньше 0.5.

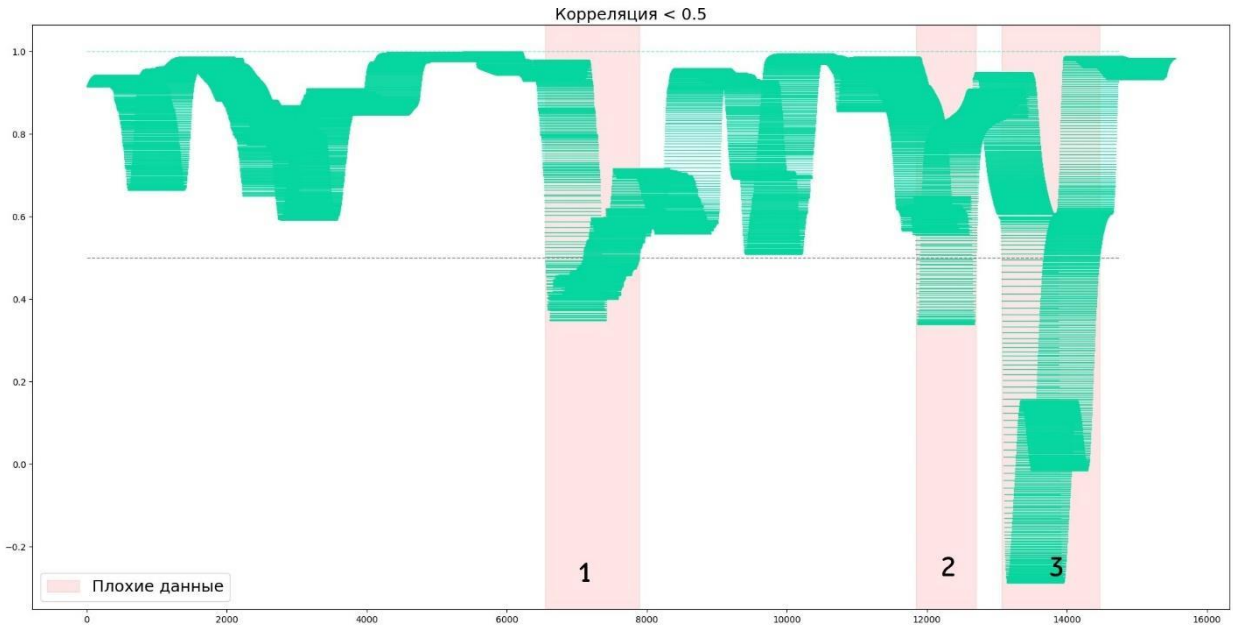

Рисунок 14 – Результаты корреляционного анализа

На рисунке 15 представлены участки, где корреляция была меньше 0.5, из графиков видно, что в какой-то момент сигналы стали вести себя странно. Таким образом данные участки определяются как некорректные и при обучении моделей из обучающей выборки их следует убрать.

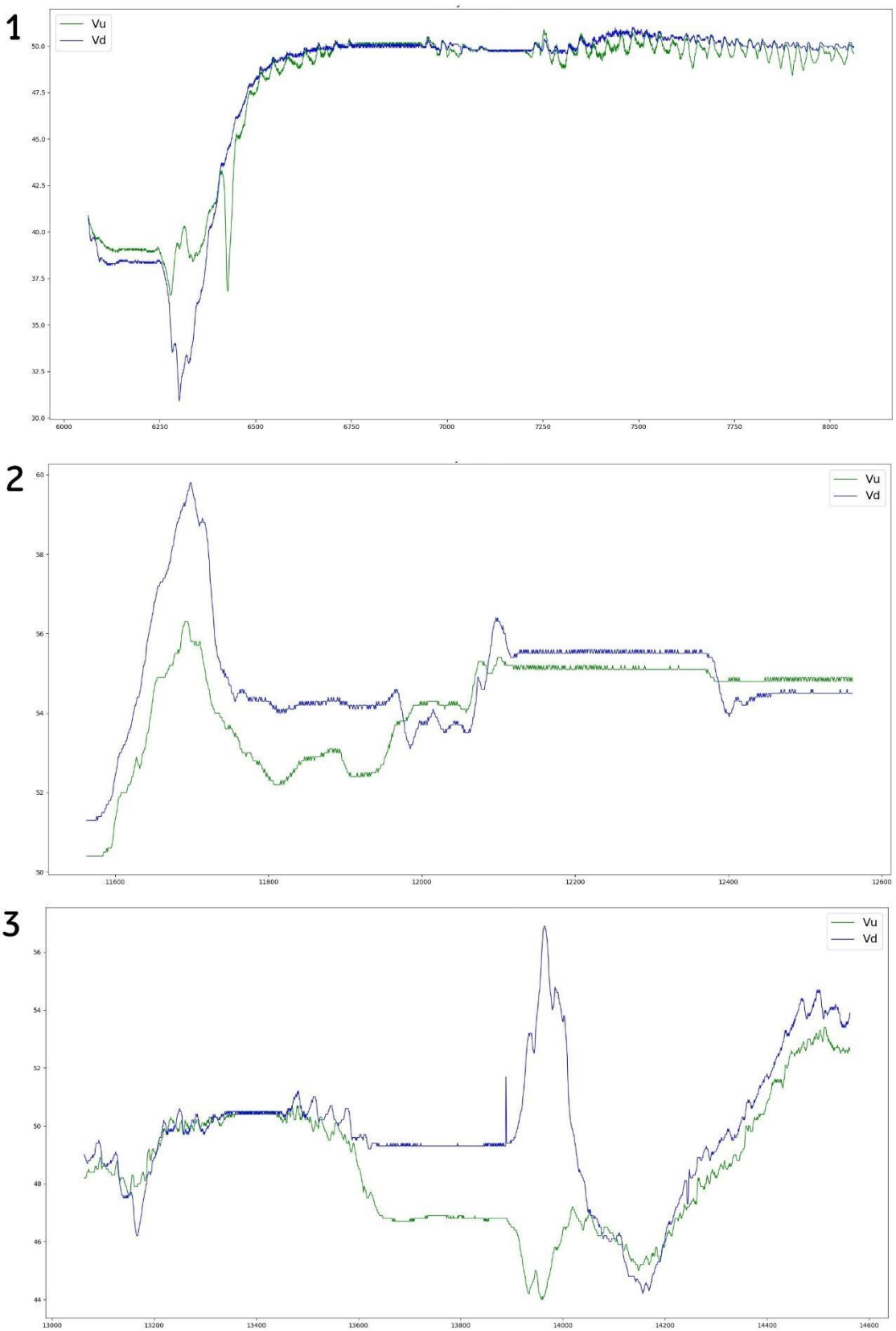

Рисунок 15 – Сигналы Vu и Vd на участках с низкой меньше 0.5

5) Классификация степени опасности аномалий с использованием автоматического машинного обучения

В качестве базового генератора признаков для классификации типа аномалии используется генератор квантильных признаков. Идея метода проста: для каждой аномалии временного ряда формируется его вектор признаков, который включает значения статистических признаков, вычисленных по всей длине аномалии. Статистические признаки, приведенные ниже, используются для формирования вектора признаков каждого временного ряда. Если дисперсия полученных статистических признаков равна или почти равна нулю, то такой признак удаляется из конечного вектора признаков. Этот метод может быть изменен с помощью оконных методов, которые могут позволить отслеживать изменчивость объектов с течением времени. В качестве характеристики используются следующие статистические значения: среднее и медианные значения, значения стандартного отклонения и дисперсии, минимальные и максимальные значения, количество ненулевых значений, 5%, 25%, 75%, и 95% квантилей. В качестве метрики качества используется ROC-AUC для бинарной классификации и F1 для многоклассовой классификации. В качестве модели мы используем подход, основанный на автоматическом машинном обучении с использованием фреймворка FEDOT. Пример модели показан на рисунке 16.

Рисунок 16 – Пример модели машинного обучения, полученной с помощью фреймворка FEDOT

6) Классификация степени опасности аномалий с использованием базы данных аномальных зон

Для использования метода описанного в конце раздела 3 была введена система классификации аномалий и создана база данных аномальных зон.

Аномалии были разделены на 4 класса, основываясь как на представленной разметке, так и на схожести конфигураций графиков в неразмеченных областях.

Погрешности/фон - области с не являющиеся аномалиями. Помехи от работы приборов, движения и подобных незначительных вещей, которые составляют основной фон ряда.

Небольшие дефекты - области с небольшими возмущениями, которые могут быть вызваны как незначительными внешними факторами, так и незначительными дефектами

трубопровода. Их конфигурация похожа на показанное на рисунке 17, с той лишь разницей, что их амплитуда больше и они могут быть длиннее.

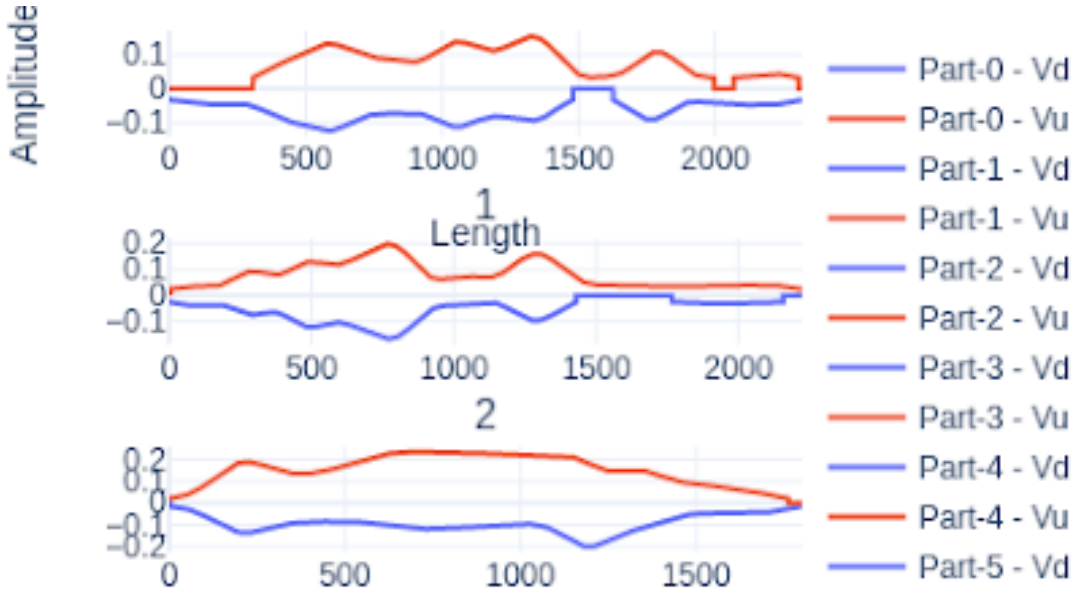

Рисунок 17 – Пример помех

Ниже показан пример нескольких подобных аномалий:

Заметные дефекты/докритические аномалии - области с выраженными скачками графиков, которые являются аномалиями, чьи свойства, однако же, не делают их критическими. Их пример приведен ниже (рисунок18):

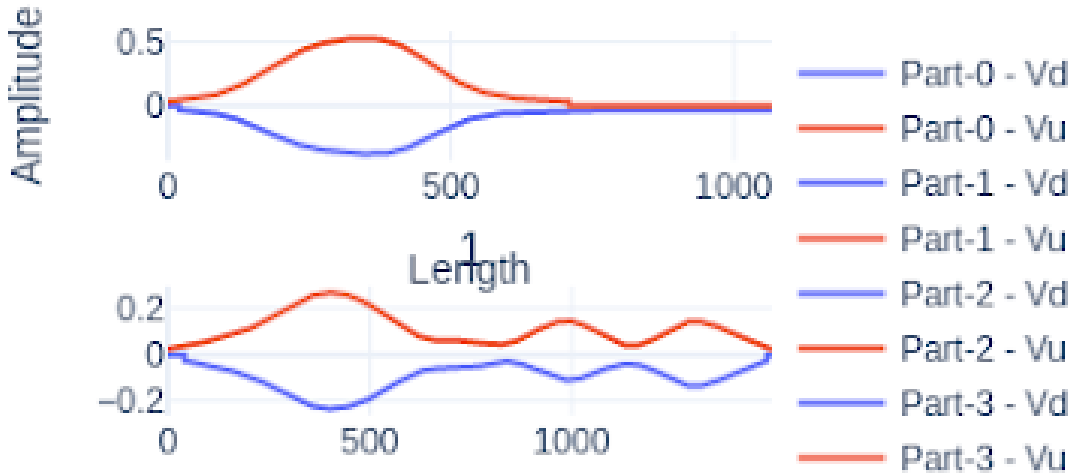

Рисунок 18 - Докритические аномалии

Критические аномалии - искомые аномалии, критические зоны, которые сильно отличаются от остальных зон и имеют сильные перепады высот, большие продолжительности или и то, и другое одновременно. Поиск и работа с этими зонами приоритетна (рисунок 19):

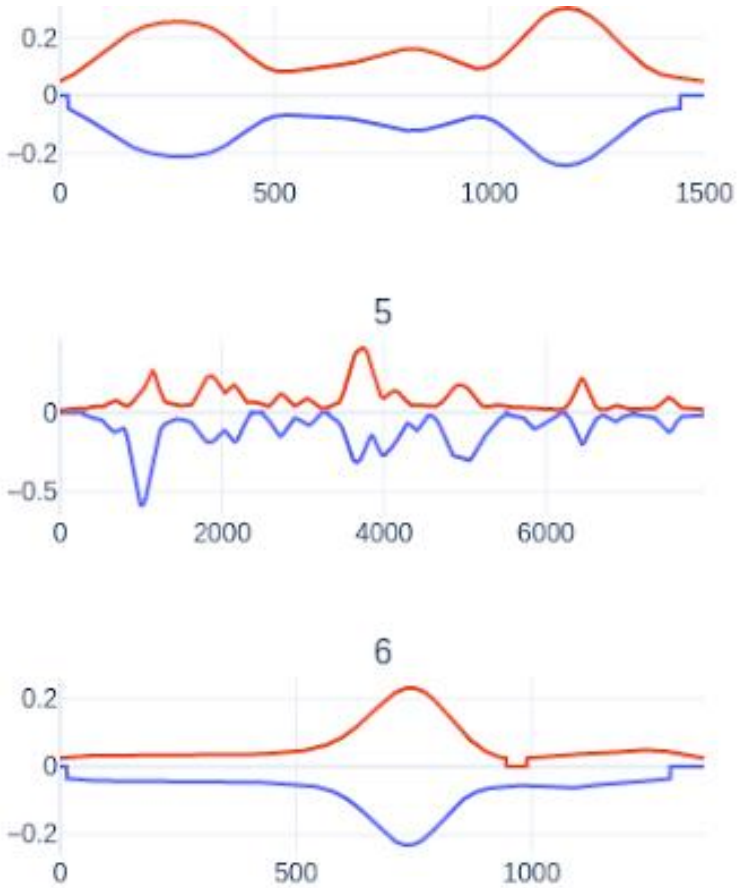

Рисунок 19 - Критические аномалии

Имея базу данных аномалий, можно весьма эффективно оценивать подобие найденных зон и аномальных зон из базы. Метод основан на сравнении метрик временных рядов. В первой версии базы были сохранены метрики каждого ряда, однако это показало себя не лучшим образом. Второй вариант созданной базы включает в себя участки содержащие аномалии, взятые из исходных данных без предобработки.

Это позволяет гибко настраивать как набор метрик, который характеризует аномалии. Так же это позволяет сравнивать данные по любому доступному ряду (при условии, что ряд есть и в базе, и в данных - некоторые файлы содержат дополнительные временные ряды), извлекая любой набор метрик из сохраненных рядов в базе и рядов из найденной аномальной зоны и оценивая расстояние между векторами метрик.

Комбинирование этого способа и кластеризации видится самым перспективным путём решения поставленной задачи по нескольким причинам.

Как видно на рисунке 20 большая часть аномалий принадлежит классу 0 (погрешности/фон), который обозначен красным цветом. Левая и центральная область почти полностью состоит из аномалий такого рода. Зеленые аномалии класса 1 (небольшие

дефекты) составляют половину правой области. Докритические и критические аномалии сгруппированы в правой области с небольшими включениями в центральную.

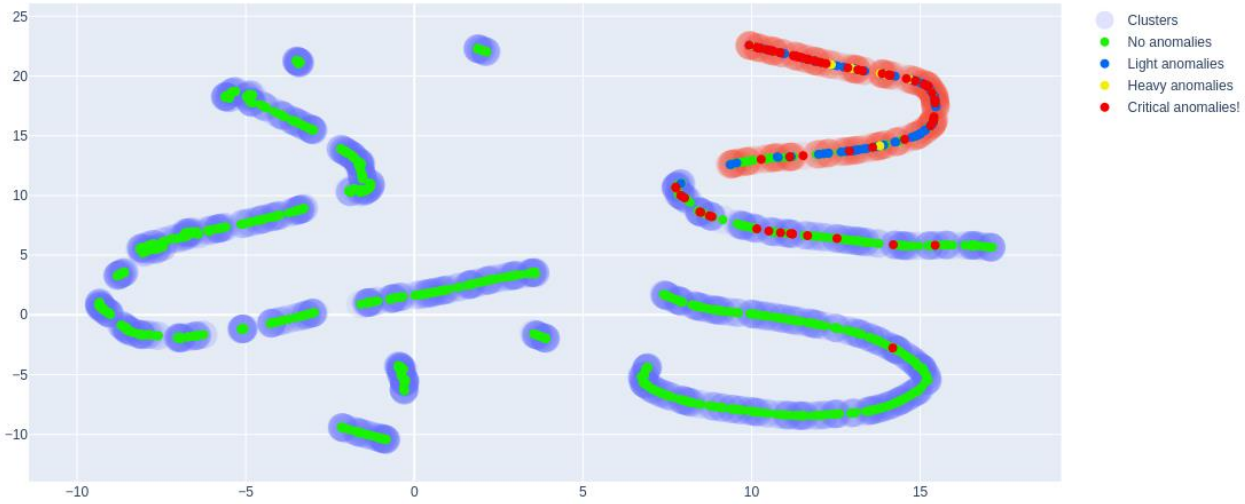

Рисунок 20 - Визуализация относительного положения векторов аномальных зон

Подобное распределение позволяет комбинировать и уточнять предсказания. Так как искомыми классами аномалий являются докритические и критические, то алгоритм поиска таких аномалий может быть представлены в виде следующих этапов:

- Кластеризация полученных аномальных зон из новых рядов вместе с зонами из базы данных.
- Разделение кластеров на два типа - аномальные и не аномальные, базируясь на количестве зон помех в кластере. На рисунке 9 не аномальными будут являться левый, центральный и два небольших кластера между ними. Кластер справа будет помечен как аномальный.
- Зоны, который были определены как докритические/критические путем сравнения с зонами из базы данных, но оказались в некритических кластерах сохраняются для последующего анализа.
- Аномальный кластер разбивается на кластеры меньшего размера. Кластеры отмечаются как некритические, докритические и сверхкритические. Зоны из этих кластеров объединяются с сохраненными зонами из пункта 3 и отправляются на дальнейший анализ.

Такой метод позволит находить аномальные зоны тех конфигураций, которые раньше не встречались, так как при даже небольшой схожести с уже найденными критическими зонами, кластеризация поместит новые зоны в кластер с уже известными аномалиями.

Как видно на рисунке 20 - все легкие (менее опасные) и тяжелые (более опасные) аномалии выделяются в один кластер, почти все критические аномалии также лежат в этом же кластере.

Соответственно этот метод позволяет отделить аномальные зоны от не аномальных с точностью 90.5%. Для легких и тяжелых аномалий точность составляет порядка 96%, а для критического порядка 85%. Дальнейшая обработка этих результатов позволит повысить точность разделения зон.

Классификация зон по степени тяжести и типу аномалий возможна с использованием повторной кластеризации. Как видно на рисунке 21 аномалии внутри аномального кластера распределяются на новые кластера, причем большинство критических аномалий сгруппировано в один кластер - левый, а остальные составляют легкие и тяжелые аномалии, предположительно разных типов и/или опасности. Работа с ними и создание датасета для оценки тяжести и типов аномалий является дальнейшим направлением работы.

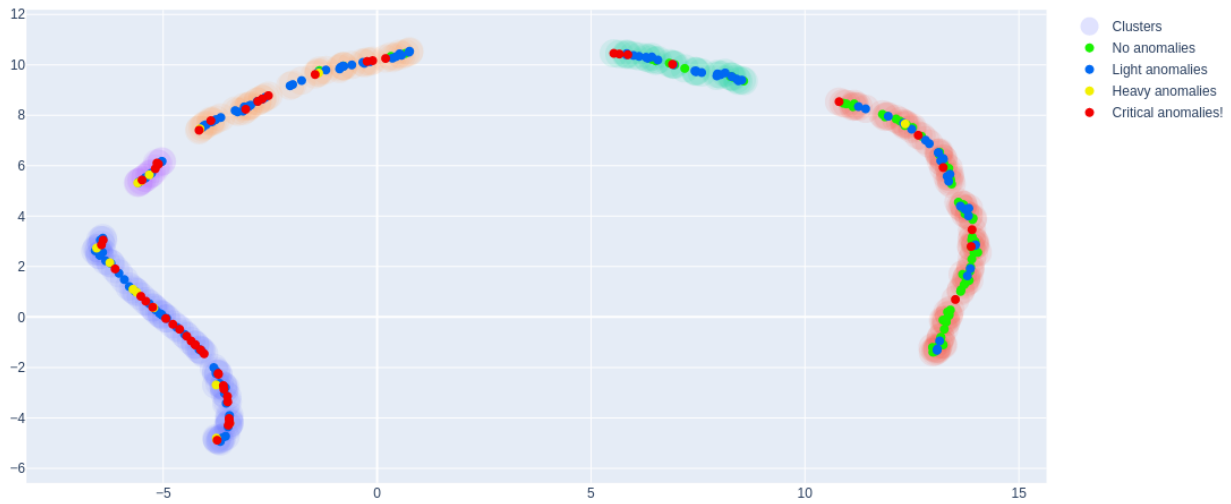

Рисунок 21 - Кластера внутри аномального кластера с рисунка 9