Анализ скорости счёта гамма-квантов, регистрируемых в подземном детекторе LVD, с целью предсказания сильных сейсмических событий

Студент:
Комлык Егор Романович

Научный руководитель:
Агафонова Наталья Юрьевна

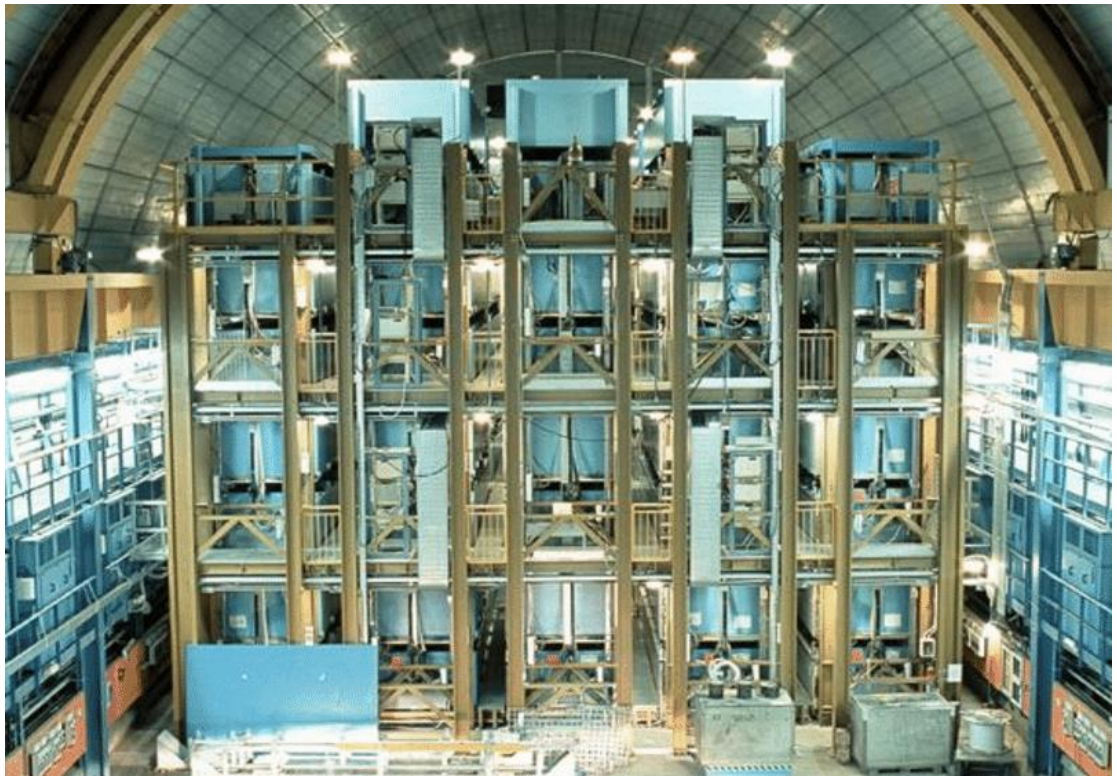
Цели

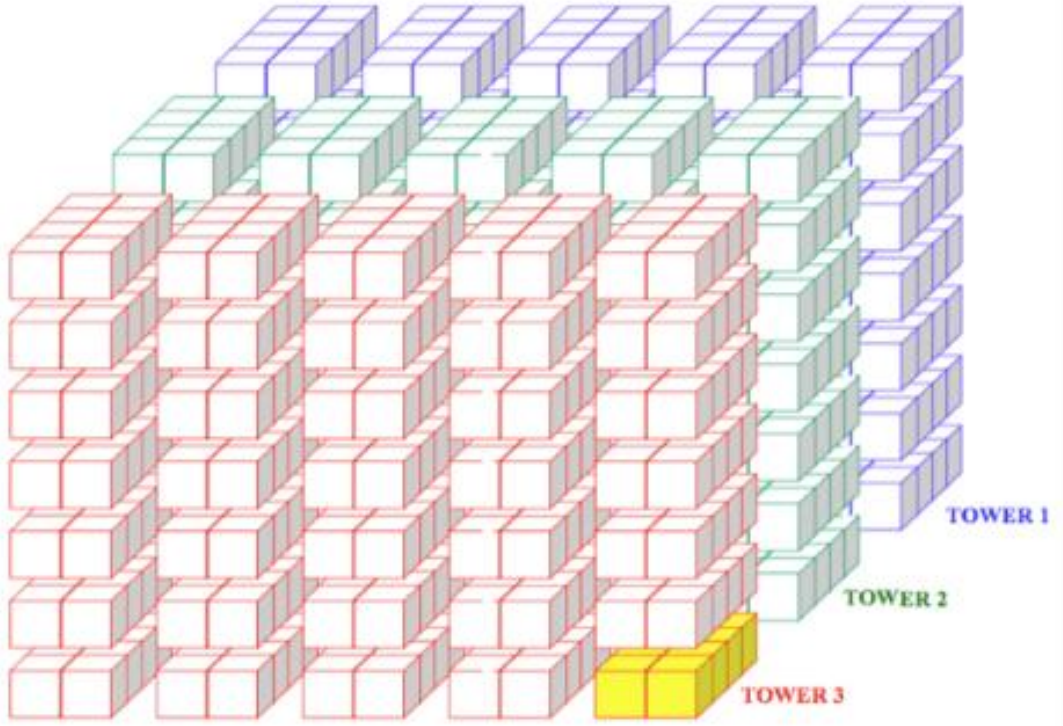
Целью работы является на основе экспериментальных данных по скорости счета гамма-квантов, регистрируемых детектором LVD, построить с применением машинного обучения модель для обнаружения предвестниковой фазы землетрясений.

Задачи

- Изучить работы, посвящённые исследованию отклика детектора LVD на сейсмические события.
- Проанализировать экспериментальные данные с целью выделения характерных для предвестников землетрясений форм сигнала.
- Найти подход к решению задачи по поиску предвестников землетрясений с точки зрения машинного обучения.

Описание LVD

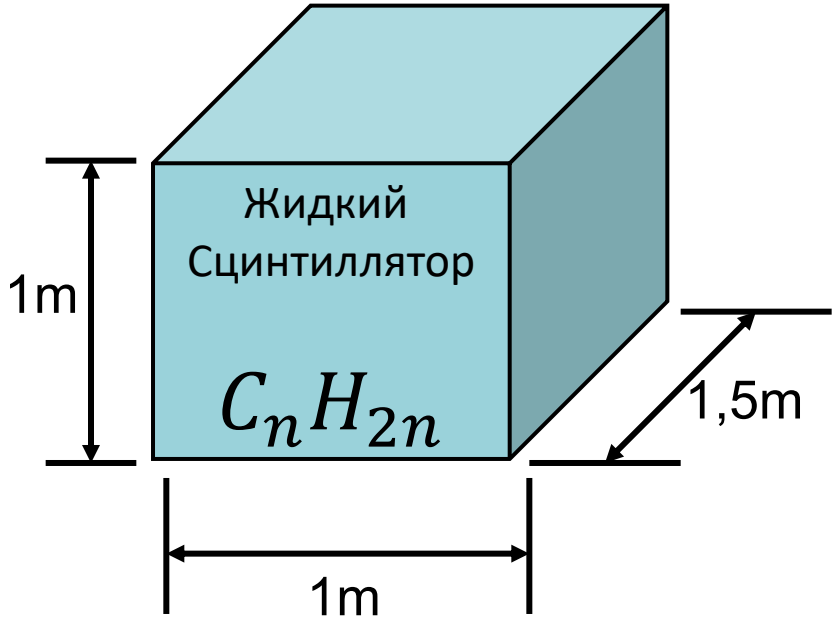




Глубина залегания – 3650 м в.э.

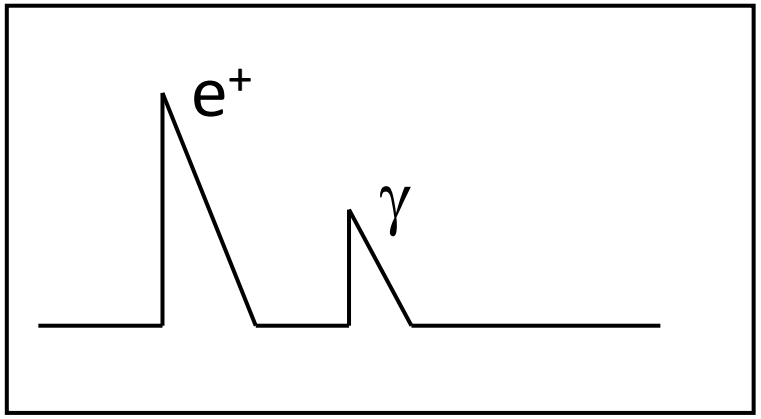
Назначение: поиск разных типов нейтрино от гравитационных коллапсов звездных ядер в нашей Галактике.

Длина	22.7 м
Ширина	13.2 м
Высота	10.0 м
Масса сцинтиллятора	1008 т
Число счётчиков	840
Число ФЭУ	2520



Основная реакция – обратный бета-распад (IBD):

$$\bar{\nu}_e + p \rightarrow e^+ + n$$



$$n + p \rightarrow \mathrm{d} + \gamma \qquad (E_\gamma = 2.2\ \text{МэВ})$$

$$n + {}^{56}_{26}Fe \rightarrow {}^{57}_{26}Fe + \Sigma\gamma \qquad (\Sigma E_\gamma = 10.16\ \text{МэВ})$$

Два порога регистрации:

Триггерный $E_{HET} = 4$ МэВ,
низкоэнергетичный $E_{LET} = 0.5$ МэВ

Фон: атмосферные мюоны (средняя энергия которых около 280 ГэВ, скорость счёта мюонов на счетчик ~10^{-4}сек$^{-1}$ счётчик$^{-1}$), и естественная радиоактивность скального грунта и материалов установки (для внутренних счетчиков первой башни (~45 сек$^{-1}$ счётчик$^{-1}$).

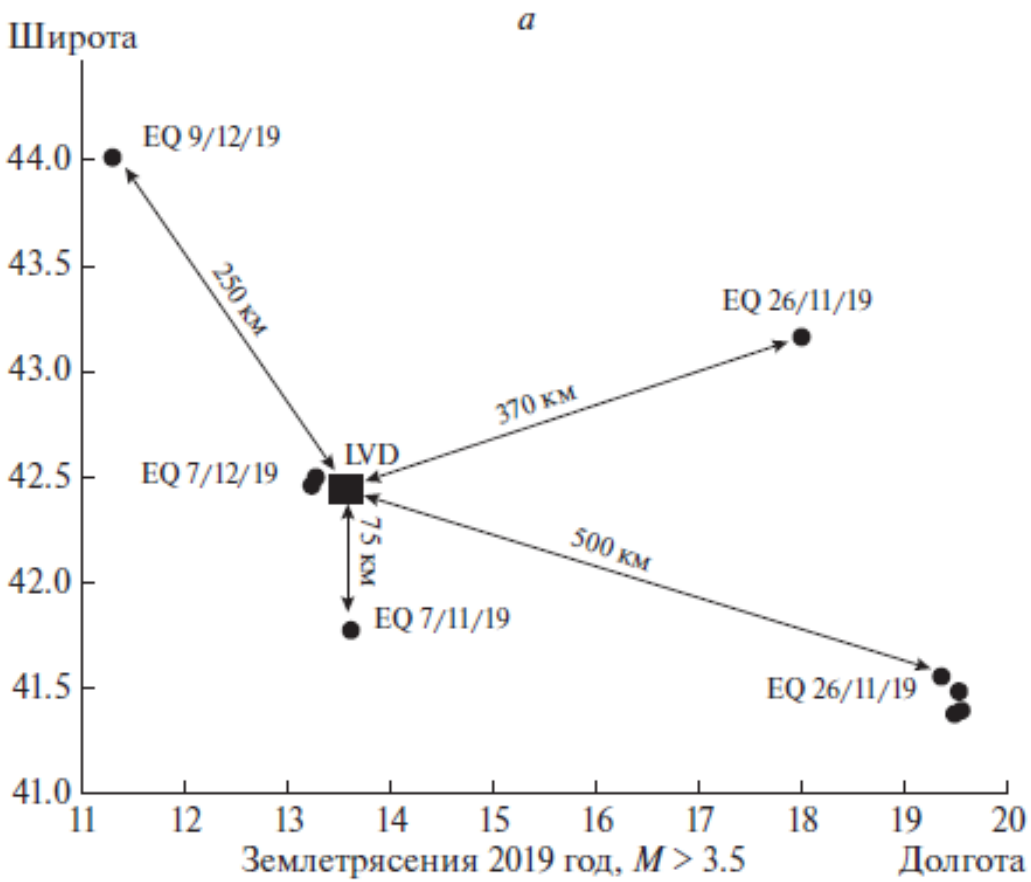
Радоновый фон и предвестники землетрясений

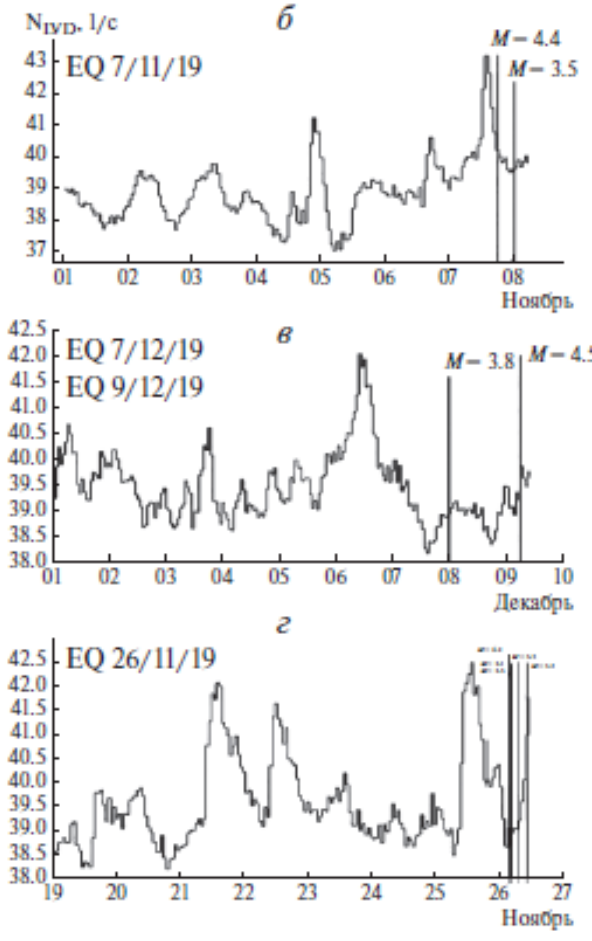
$$^{222}_{86}Rn \rightarrow^{\alpha} {}^{218}_{84}Po \rightarrow^{\alpha} {}^{214}_{82}Pb \rightarrow^{\beta-} {}^{214}_{83}Bi \rightarrow^{\beta-} {}^{214}_{84}Po \rightarrow^{\alpha} {}^{210}_{82}Pb$$

3.82 дня | 3.05 мин | 26.8 мин | 19.7 мин | 1.6e-4 сек | 22.3 года

Гамма–излучение создаётся в основном ядрами $^{214}_{83}Bi$, $E_\gamma = 0.6 - 2.5$ МэВ

Сейсмические события 2019 г. с магнитудой больше 3.5 и отклик установки LVD:
слева (*а*) – показаны эпицентры сильных толчков и их расположение относительно установки LVD;
справа (*б*, *в*, *г*) – данные установки по нижнему порогу (по оси абсцисс – дата, по оси ординат – темп счета в секунду на счетчик).
Линиями обозначены моменты сильных толчков.

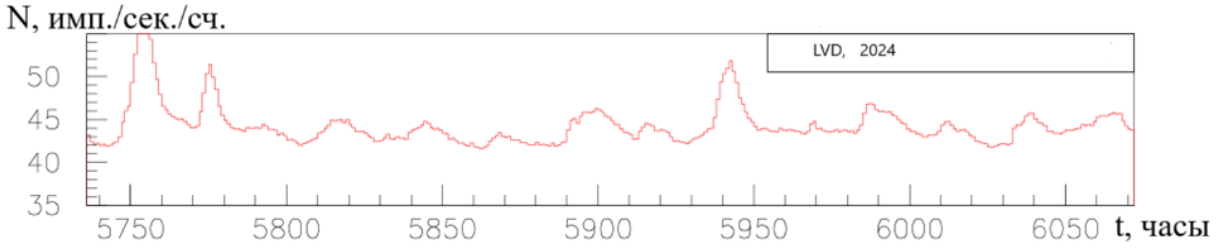




Описание данных

Лабораторные данные – временные ряды:

- Скорость счёта гамма-квантов
- Радон



Источники данных землетрясений:

1. Камчатский филиал ФИЦ ЕГС РАН (EMSD)
2. Глобальная база данных значительных землетрясений NCEI/WDS (NOAA)
3. Департамент землетрясений национального института геофизики и вулканологии Италии (INGV)
4. Геофизическая служба РАН (GSRAS)

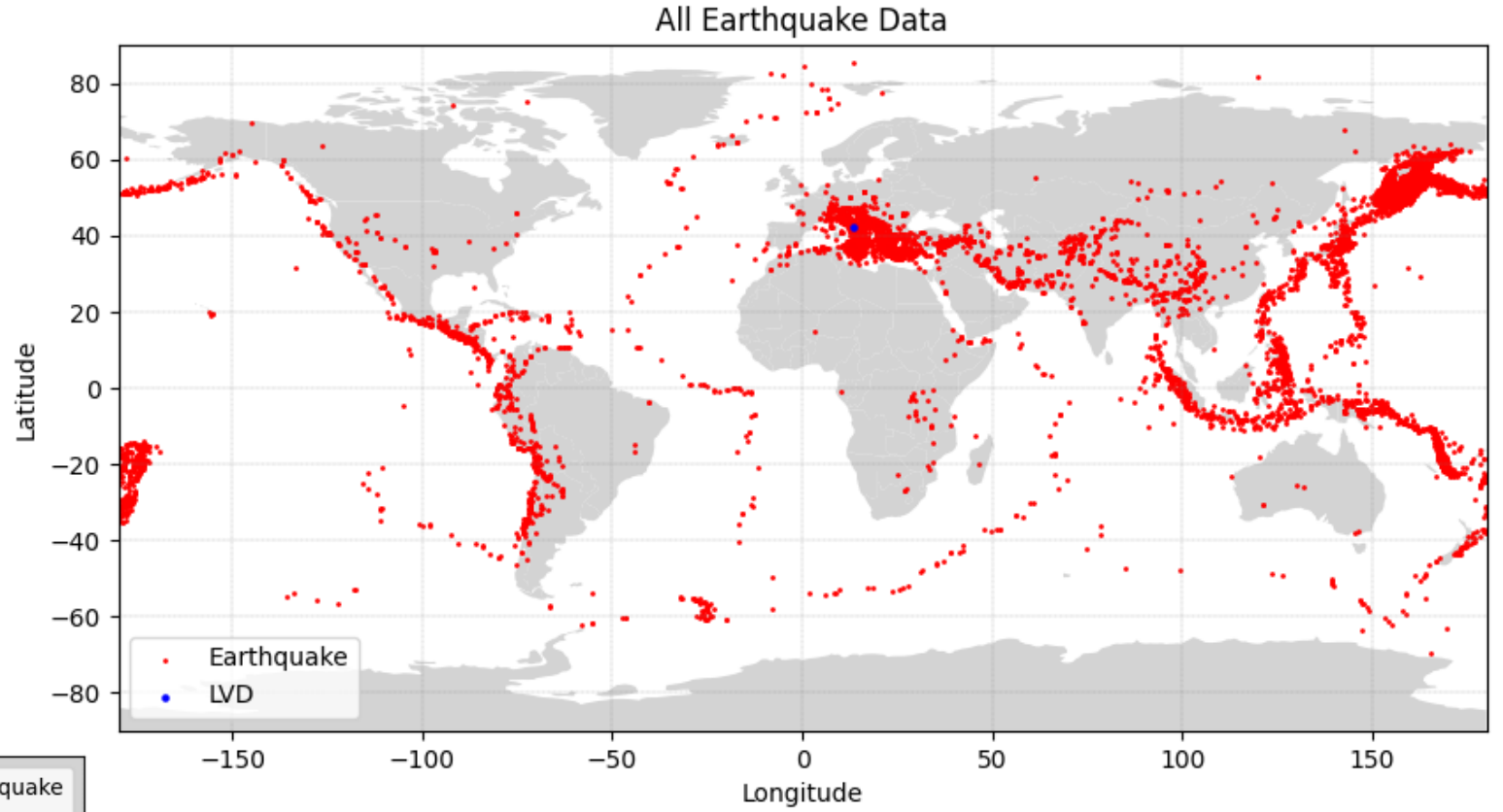
	Date	Count_rate_1	Count_rate_2	Count_rate_3	Pressure_1	Pressure_2	Pressure_3	Pressure_4
0	2004-01-10 19:00:00+00:00	54.193	54.480	82.111	760.69	760.44	760.83	760.92
1	2004-01-10 20:00:00+00:00	54.726	54.055	82.918	761.40	761.18	761.45	761.60
...	...	...	...	...	...	...	...	...
168961	2023-04-20 20:00:00+00:00	37.814	40.825	61.653	760.17	759.91	760.14	760.18
168962	2023-04-20 21:00:00+00:00	38.103	41.439	60.807	760.35	760.17	760.17	760.23

168963 rows × 8 columns

	Date	Latitude	Longitude	Depth	Magnitude	Distance_from_LVD	Intensity
0	2004-01-01 00:08:09+00:00	49.65	156.440	17.0	4.0	9184.202815	-4.870648
1	2004-01-01 20:59:31.900000+00:00	-8.31	115.788	45.0	5.8	11637.177401	-2.530478
...	...	...	...	...	...	...	...
34897	2025-03-05 05:21:51.300000+00:00	52.24	159.940	27.0	3.9	9030.266499	-4.994959
34898	2025-03-14 19:37:15+00:00	41.98	15.350	10.0	5.0	159.163411	2.790554

34899 rows × 7 columns

Фильтрация данных

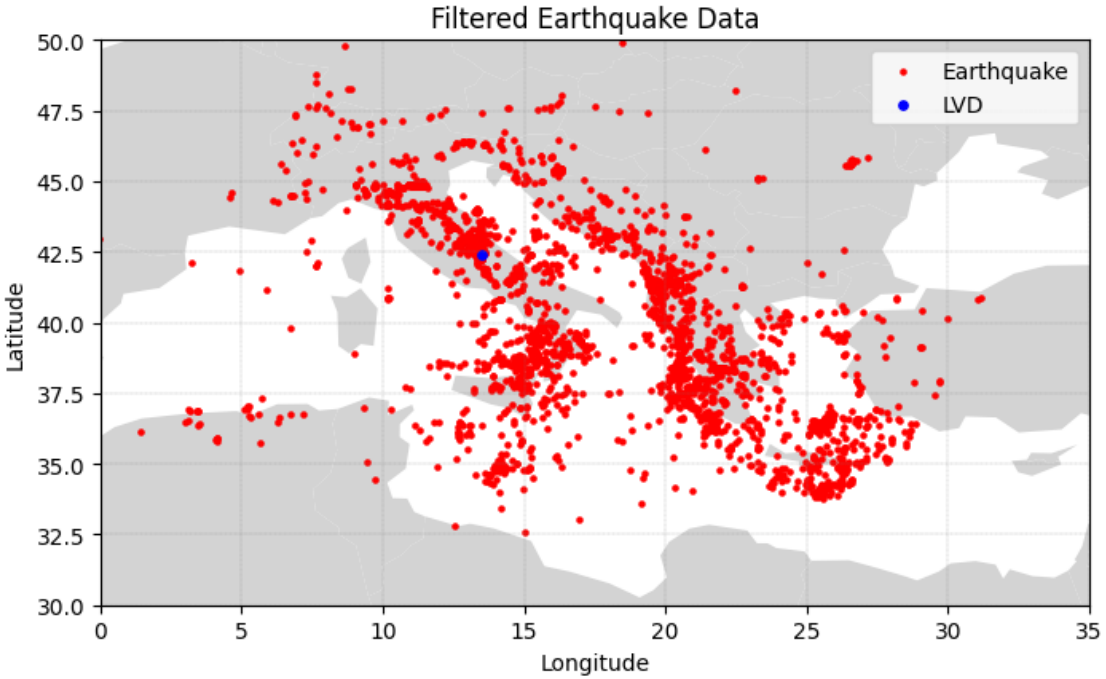


34899

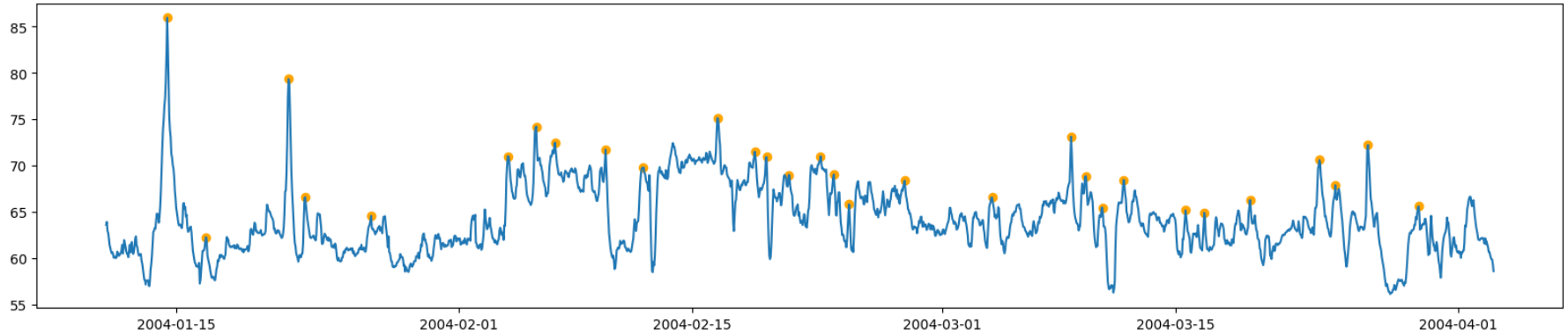
Магнитуда > 3.5

Расстояние от LVD ≤ 1500 км

2862



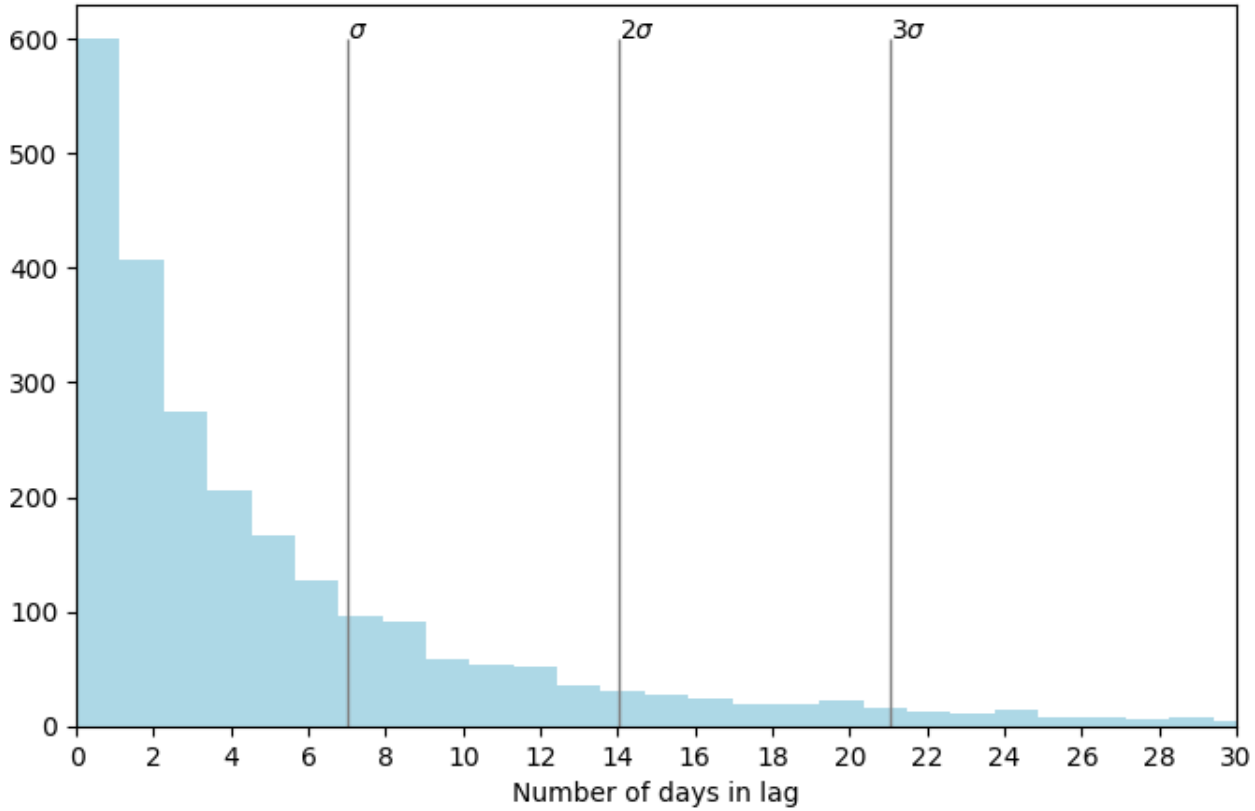
Определение интервала прогнозирования



Параметры обнаружения пиков:

- Минимальное расстояние – 12
- Минимальная амплитуда – σ/ 2

аномальное поведение радона перед землетрясением →
→ определение величины интервала между пиком и землетрясением →
→ ориентировочные значения для дальнейшего анализа

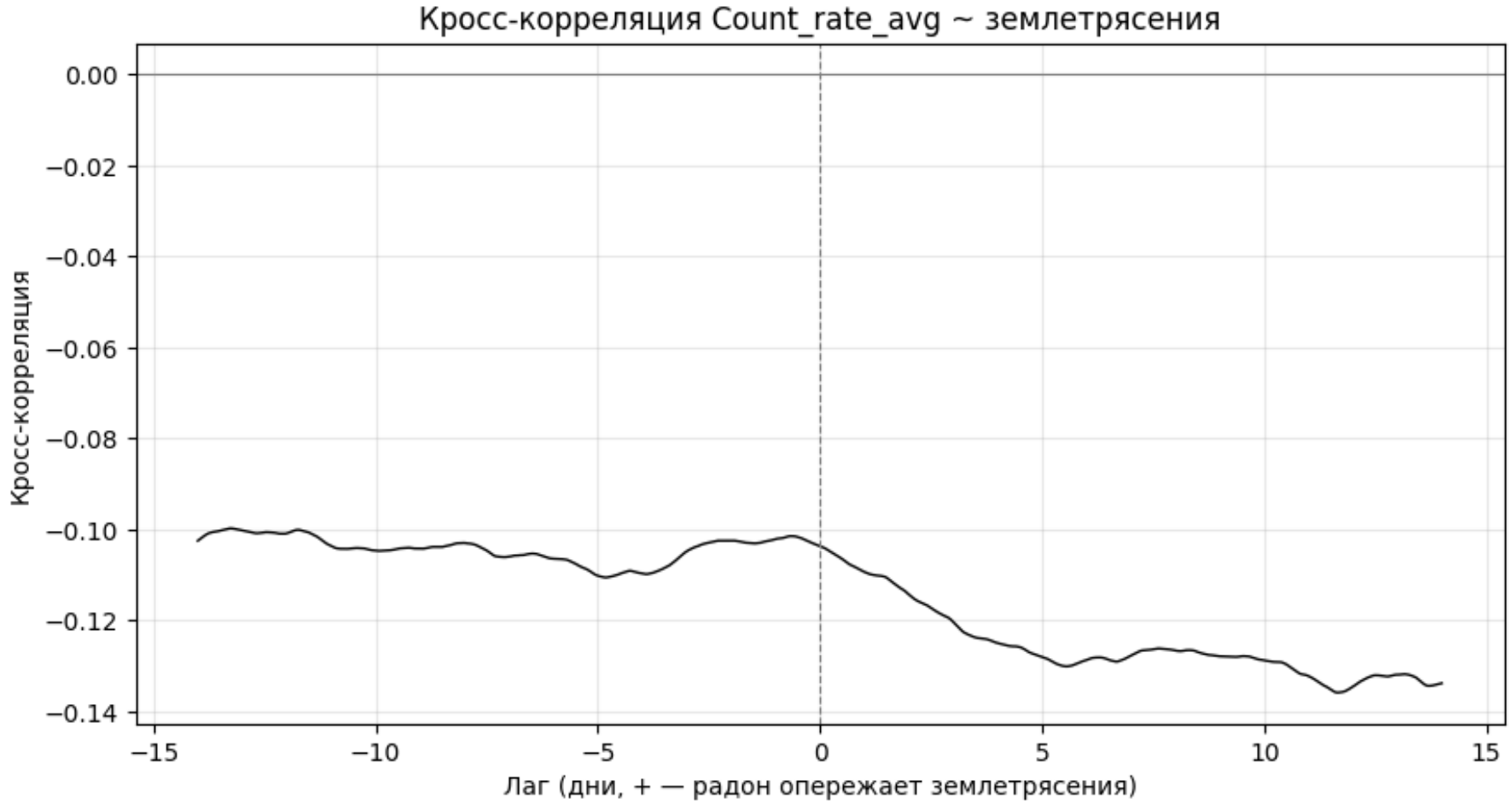


Среднее количество дней в лаге: 5.68

- σ – 7 дней – 73.9% лагов
- 2σ – 14 дней – 89.3% лагов
- 3σ – 21 день – 95.1% лагов

Поиск корреляций

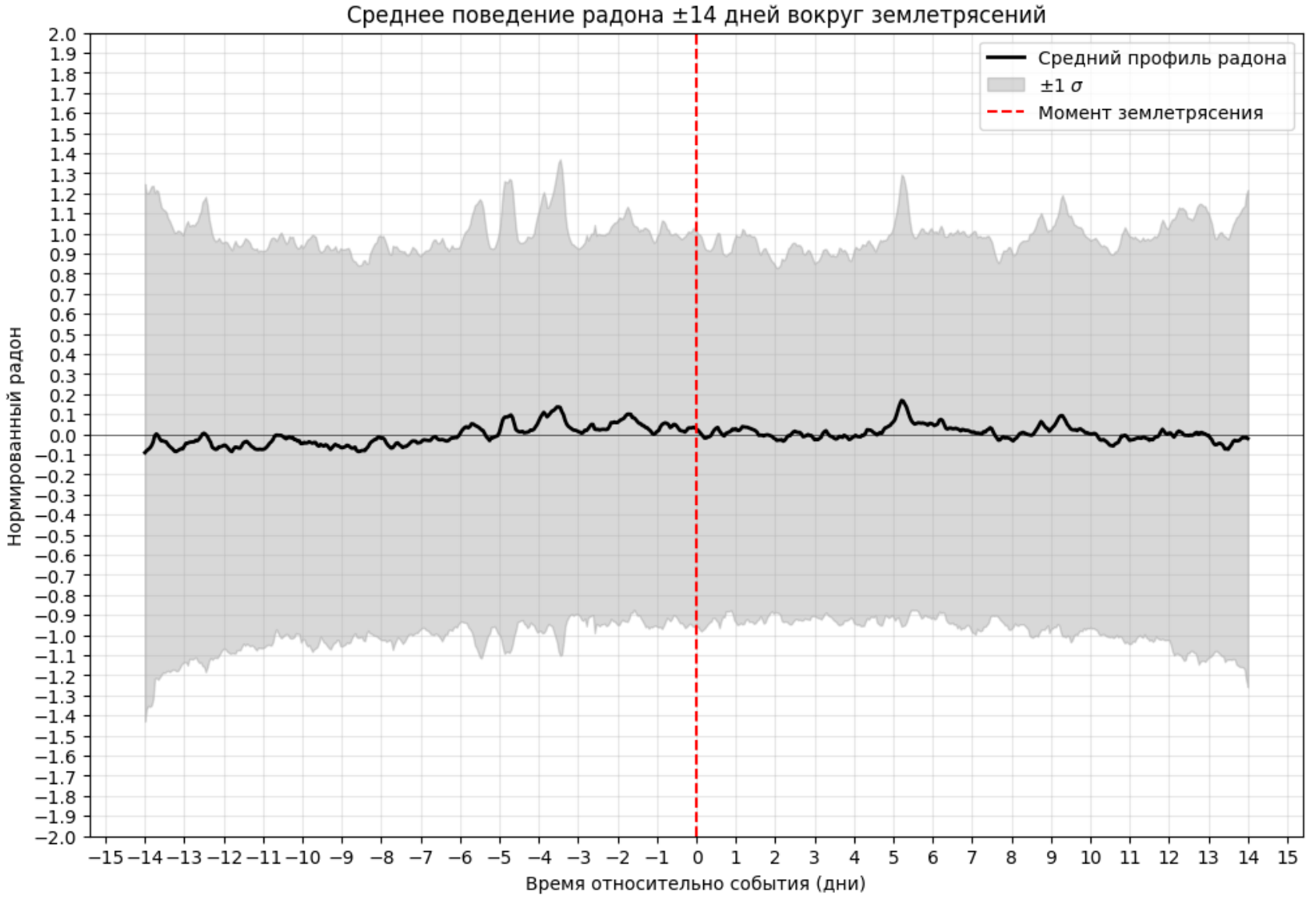
Для каждой даты проверялось наличие в ближайшие 24-144 часа после неё землетрясения, результат проверки принимал значение 0 или 1. Далее рассчитывалась корреляция при разных сдвигах (±14 дней) между датами двух векторов значений.



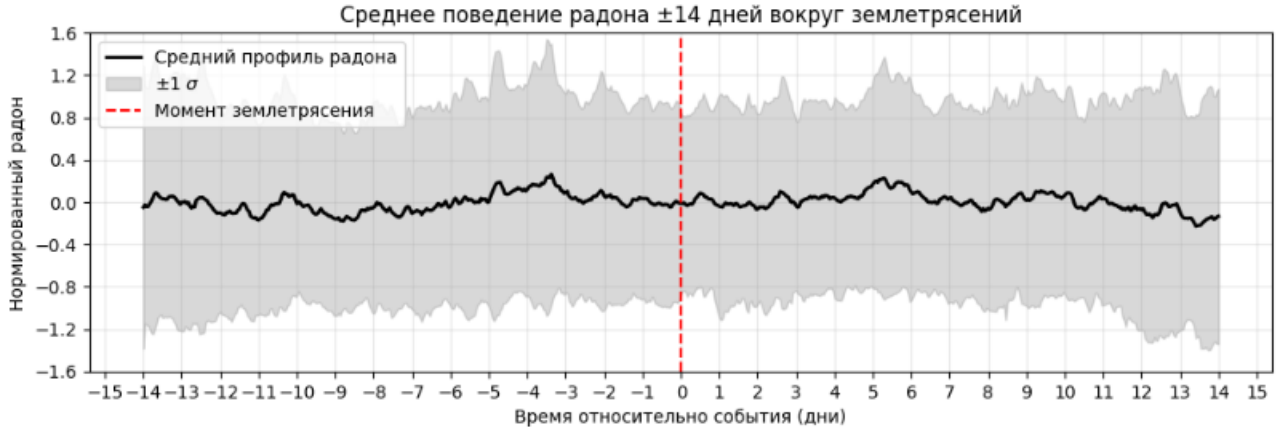
Поиск паттернов радона

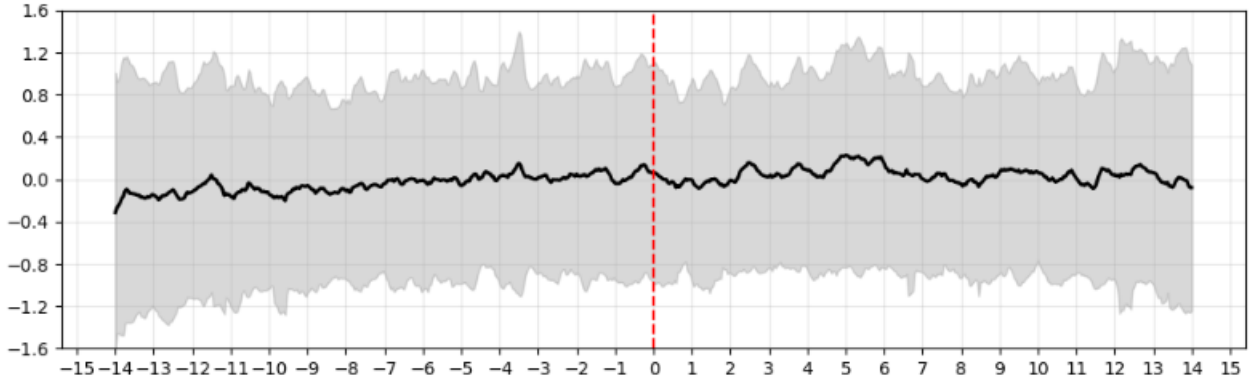
Для каждого землетрясения берётся окно ±14 дней вокруг даты события, из радонового ряда вырезается соответствующий фрагмент и нормируется относительно среднего в окне уровня радона. Далее все эти фрагменты усредняются между собой, получается «средняя форма сигнала» до и после землетрясения.

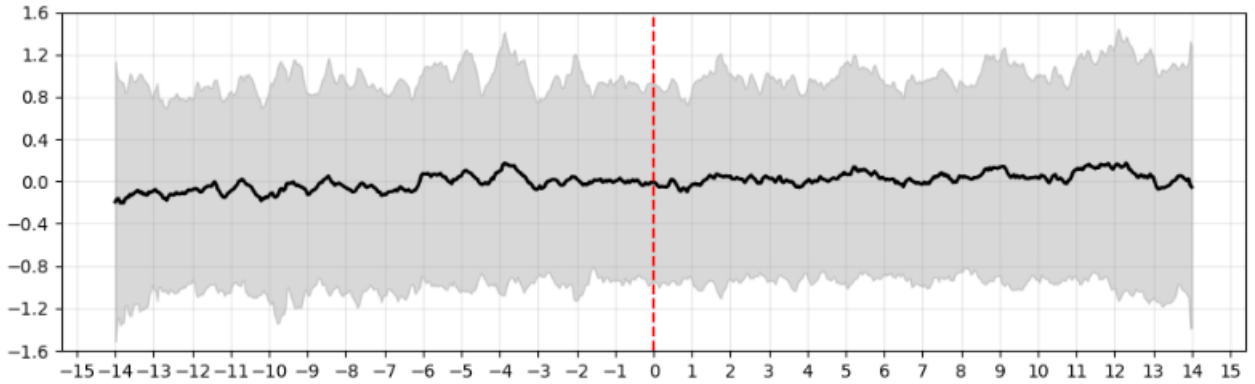
Характерные пики за 5.6, 4.8, 3.5 дня до события.



Проверка паттерна радона на устойчивость

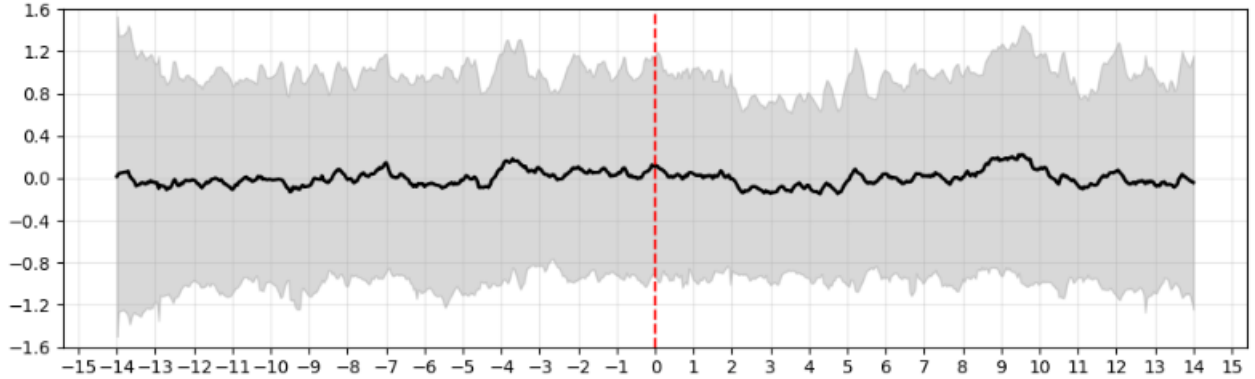


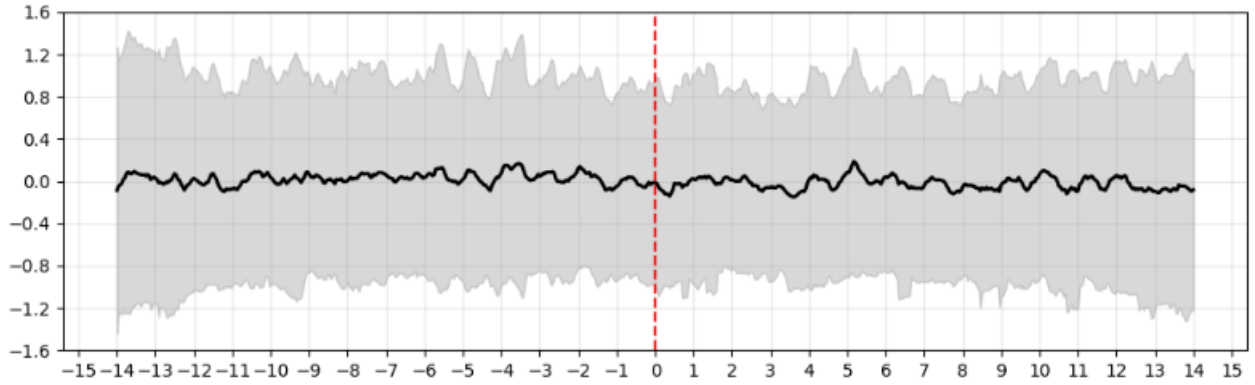




Выборки – случайные подмножества землетрясений в размере 10% от всего массива данных каждое.

- Паттерн наблюдается на выборках в 10%, но не всегда и не полностью
- Основной и самый стабильный пик – за 3.5 дня до события
- В ожидаемом диапазоне пиков среднее значение радона зачастую выше, чем в предыдущие дни

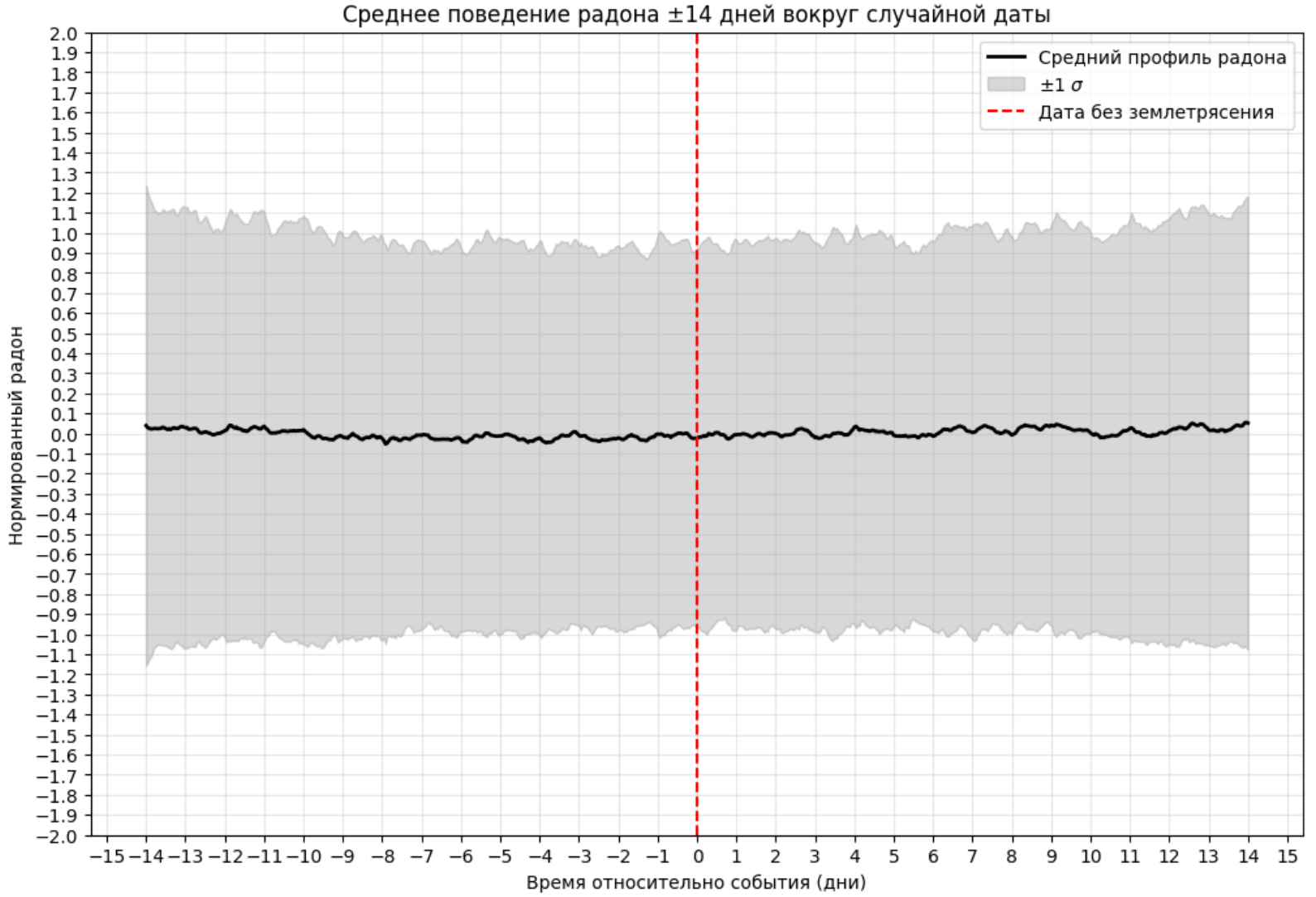




Проверка паттерна радона на значимость

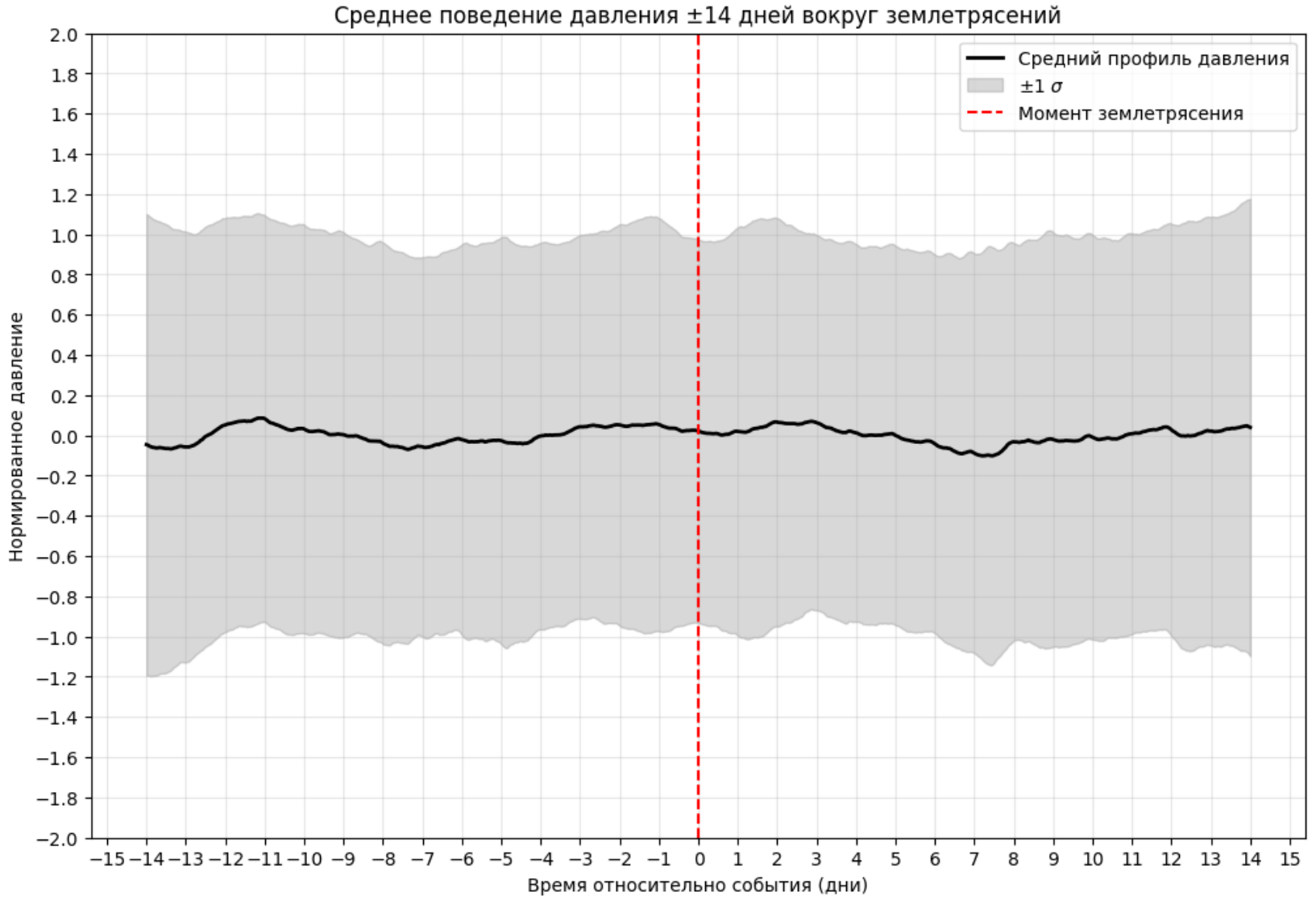
Выборка – случайное подмножество дат без землетрясений, размер равен всей выборки дат с землетрясениями.

Выделенный паттерн, как и какое-либо аномальное поведение, не наблюдается.

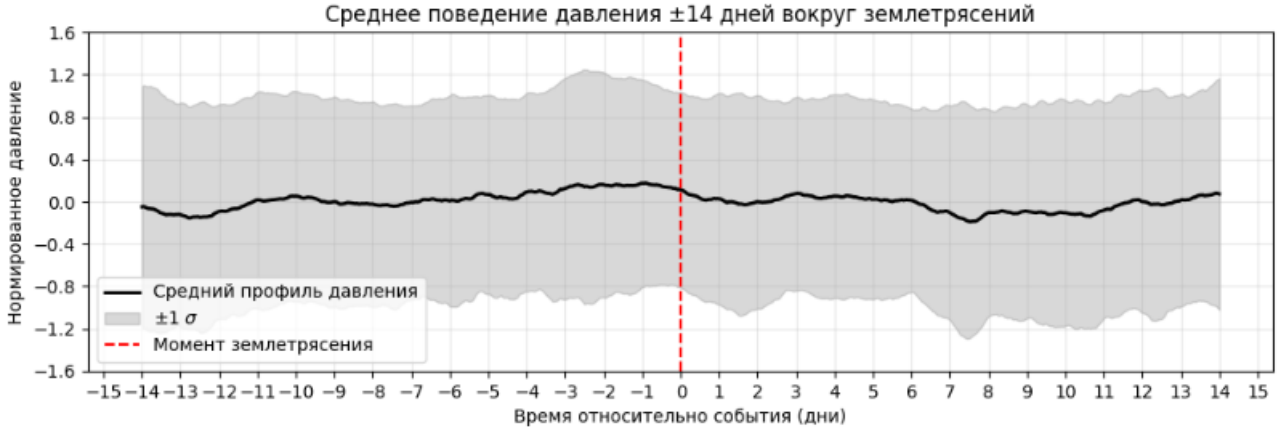


Поиск паттернов давления

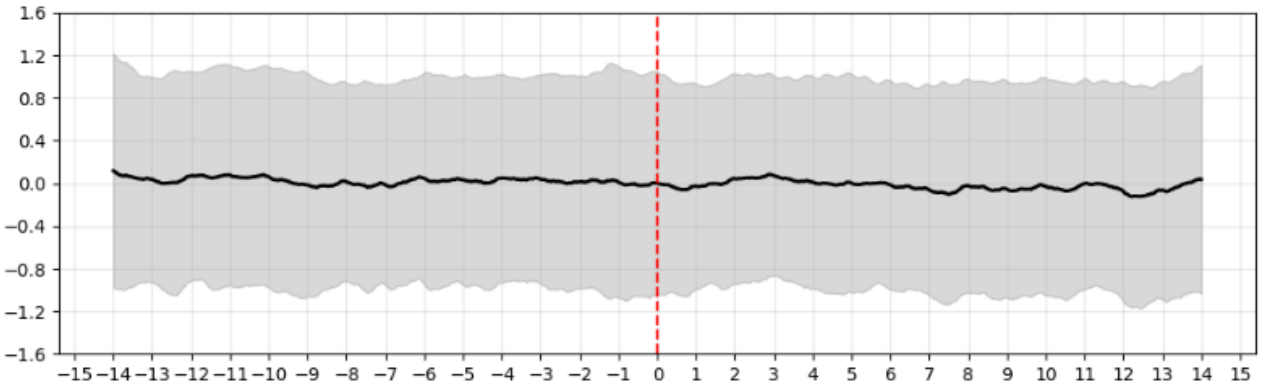
Гладкий пик за 1-2 дня до землетрясения в конце тренда на рост длительностью приблизительно 6 дней.

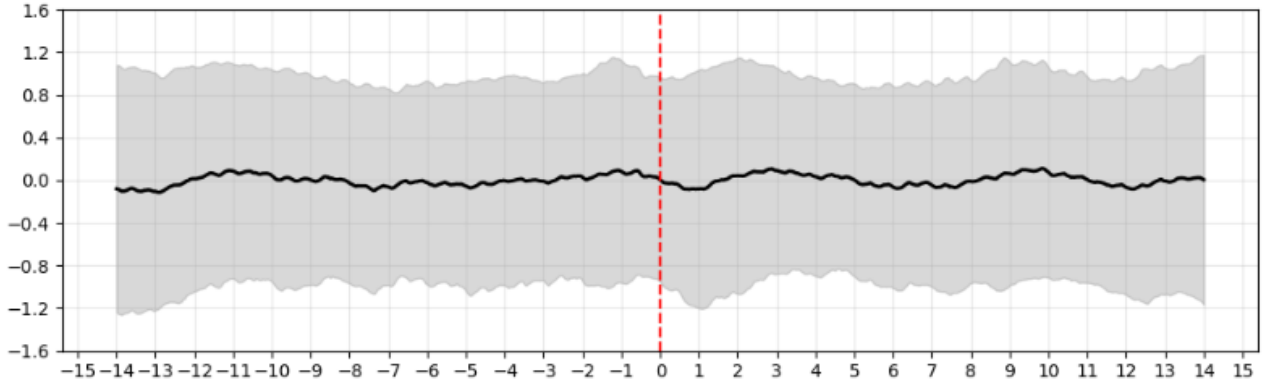


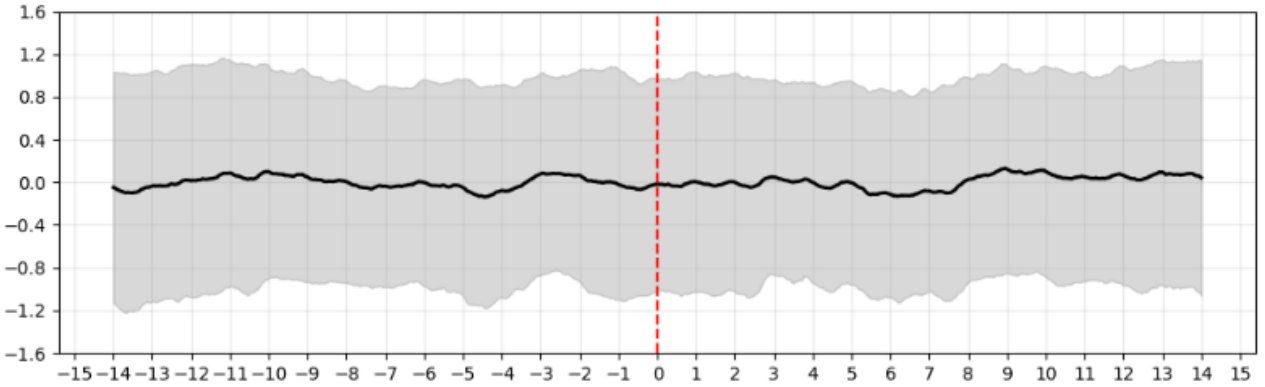
Проверка паттерна давления на устойчивость

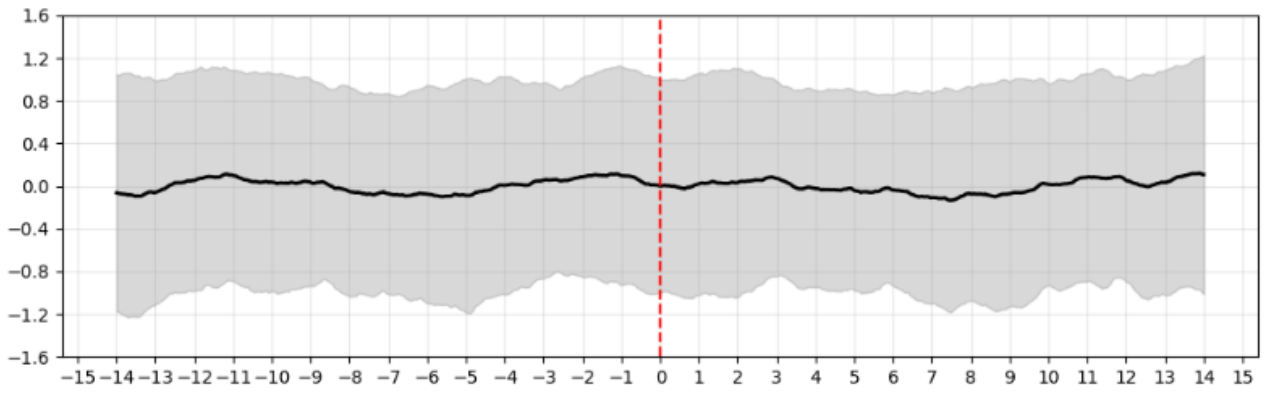


За 1-3 дня до события наблюдается небольшой пик.



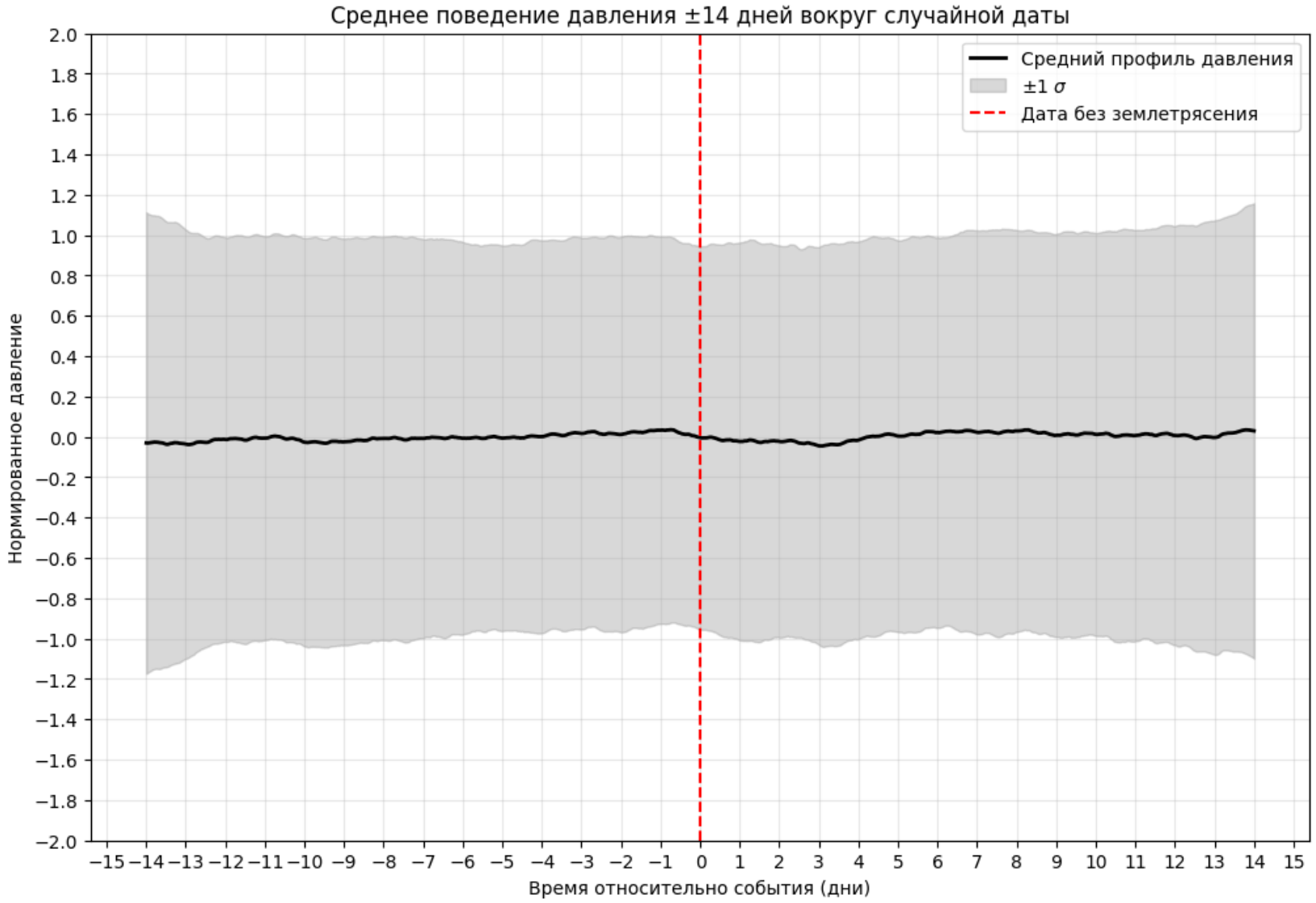






Проверка паттерна давления на значимость

Выделенный паттерн, как и какое-либо аномальное поведение, не наблюдается.



Формирование целевой переменной

Исходная постановка: «оценка вероятности возникновения землетрясения с $M \geq n$ в радиусе $R \leq m$ км в течение следующих k суток», где $n = M_{min}$ – минимальная магнитуда, $m = R_{max}$ – максимальное расстояние от LVD.

Фактически не прогнозируется вероятность землетрясения, а даётся оценка (score) состояния, в котором на данный момент находится система.

Задача бинарной классификации: грубо говоря, модель должна ответить на вопрос, будет в ближайшие k суток землетрясение или нет. Таким образом, формирование целевых меток происходит по следующему принципу:

$$y_t = 1, \text{если } \exists \text{ дата землетрясения } \in [t+1, t+k]$$

- k > 2 – дисбаланс в сторону целевого класса

- k = 2 – баланс классов
- k = 1 – дисбаланс в сторону нулевого класса, число меток равно числу землетрясений

Преобразование данных

	Count_rate_1	Count_rate_2	Count_rate_3	Pressure_1	Pressure_2	Pressure_3	Pressure_4
0	54.193	54.480	82.111	760.69	760.44	760.83	760.92
1	54.726	54.055	82.918	761.40	761.18	761.45	761.60
...	...	...	...	...	...	...	...
168961	37.814	40.825	61.653	760.17	759.91	760.14	760.18
168962	38.103	41.439	60.807	760.35	760.17	760.17	760.23

168963 rows × 7 columns

усреднение по башням

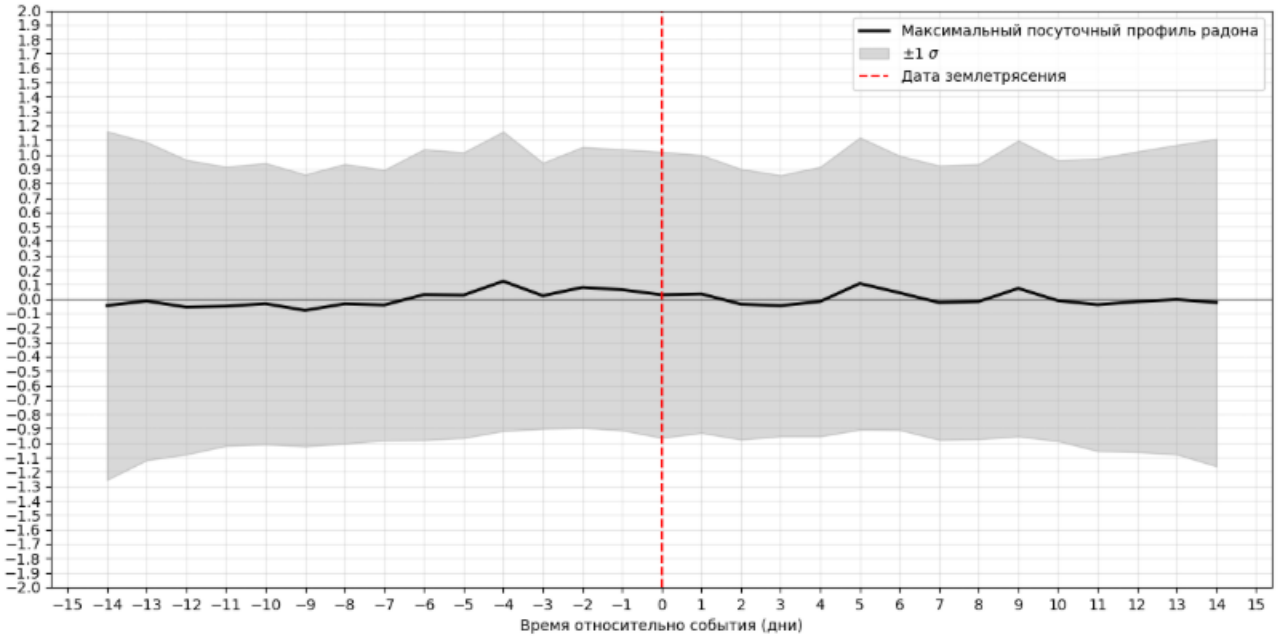
	Count_rate_avg	Pressure_avg
0	63.594667	760.653333
1	63.899667	761.343333
...	...	...
168961	46.764000	760.073333
168962	46.783000	760.230000

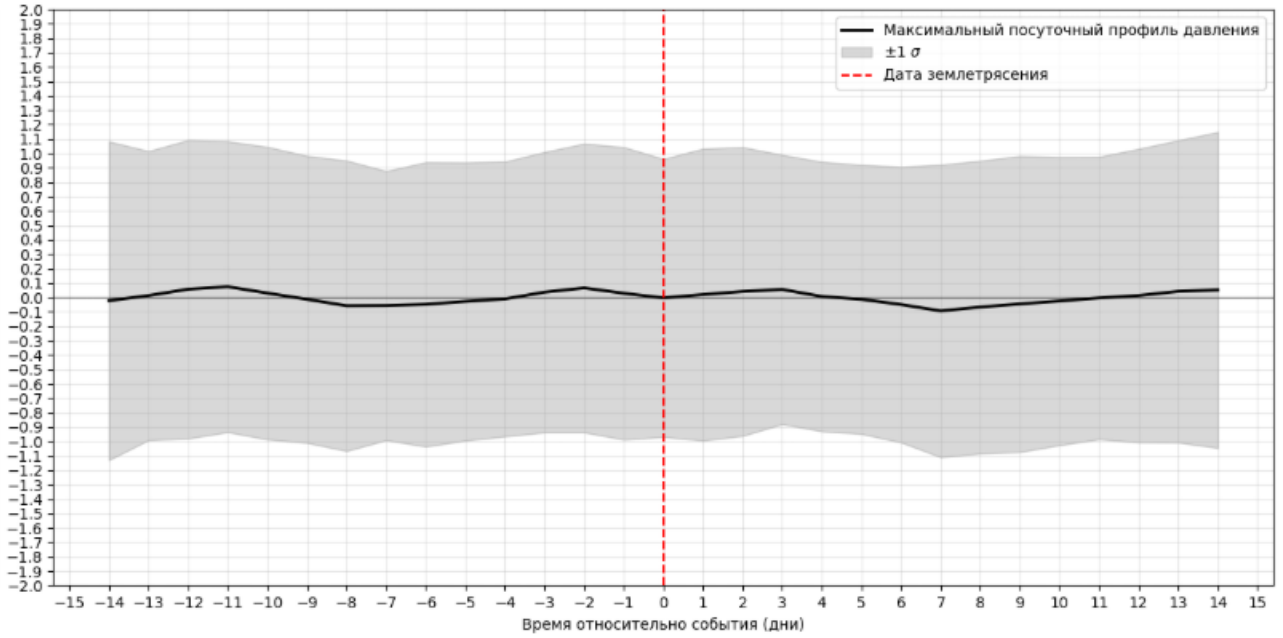
168963 rows × 2 columns

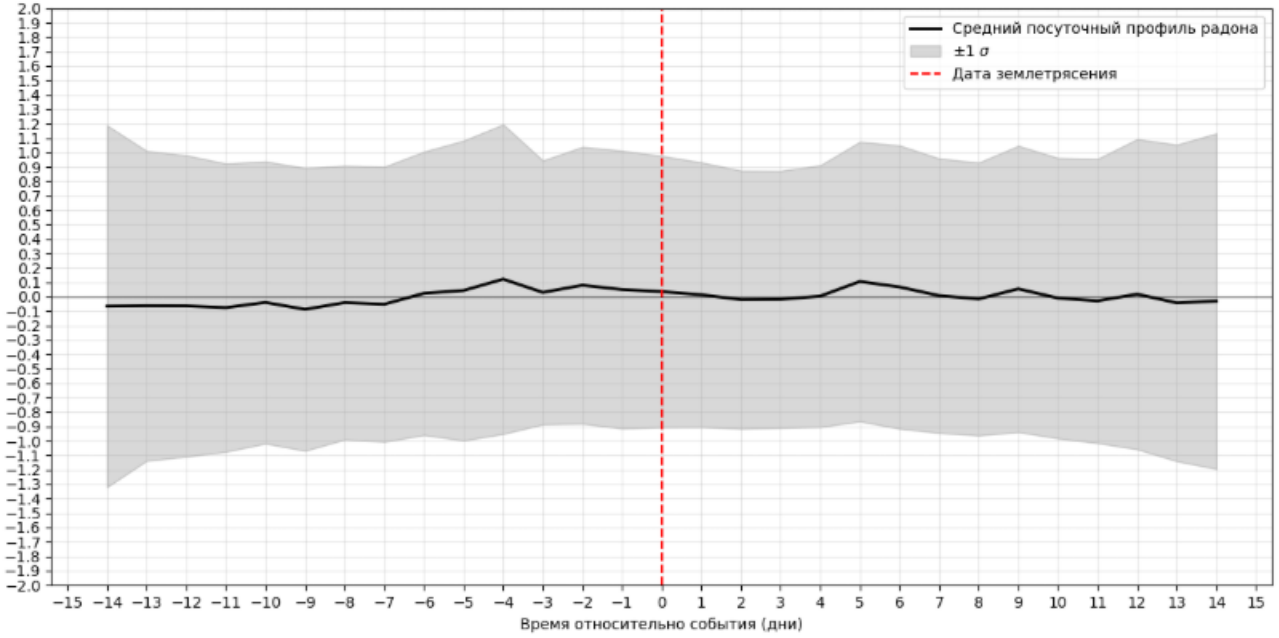
группировка и агрегация по суткам

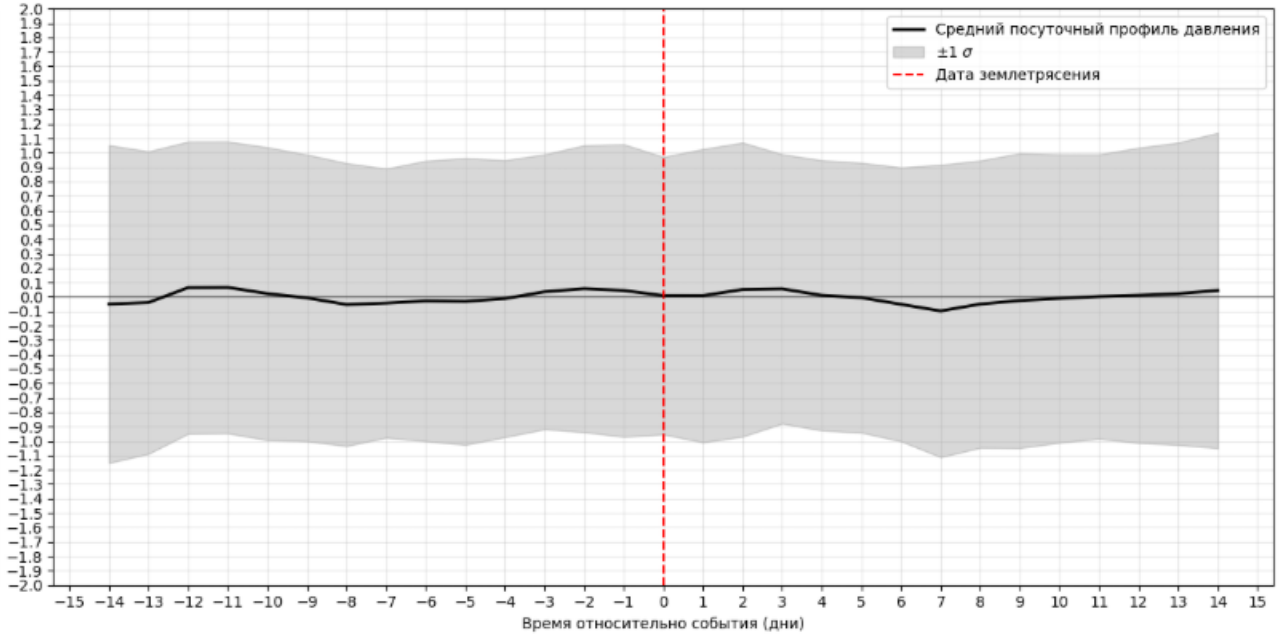
Date	Count_rate_avg_min	Count_rate_avg_max	Count_rate_avg_mean	Count_rate_avg_std	Count_rate_avg_var	Count_rate_avg_median
2004-01-10	62.094667	63.899667	63.042067	0.719298	0.517389	62.920333
2004-01-11	59.993667	61.943333	60.673222	0.533638	0.284769	60.558667
...	...	...	...	...	...	...
2023-04-19	46.742000	49.203000	47.628986	0.739322	0.546597	47.420333
2023-04-20	46.764000	50.035000	47.843182	0.931478	0.867651	47.775000

7041 rows × 12 columns









- Ранее выявленные явные пики радона стали выглядеть как заметный тренд на рост
- Картина для давления практически не изменилась

Формирование признаков на основе паттернов

k = 2 (baseline)

- Максимум радона
- Число шагов до максимума радона
- Максимум давления
- Число шагов до максимума давления
- Возрастающий тренд (линейная интерполяция) радона
- Убывающий тренд радона
- Тренд давления
- Отношение возрастающего тренда радона к тренду давления

8 признаков

k = 1

- Тренд радона
- Монотонность тренда радона
- Изменение тренда радона
- Находится ли максимум в окне 2-4 до текущей даты
- Выступ пика
- Спад после пика
- Скорость спада
- Интегральное превышение радона
- Волатильность радона

Аналогичные признаки для давления

Совместные:

- Совпадение трендов
- Пик радона раньше пика давления
- Отношение интегральных превышений радона и давления

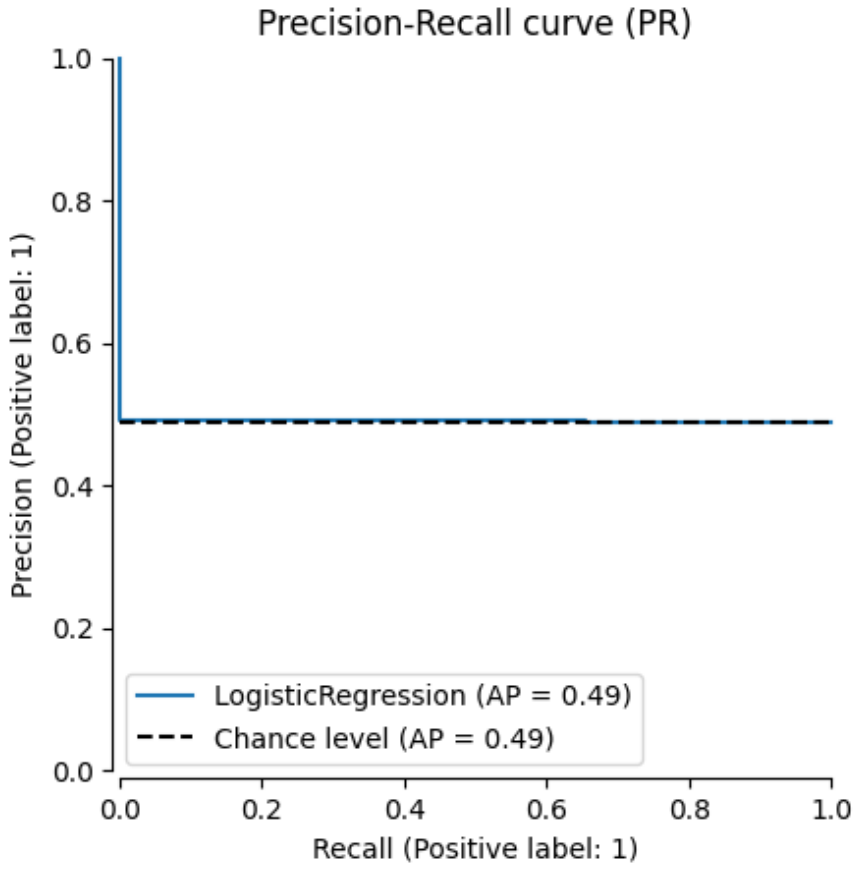
11 + 8 + 3 = 22 признака

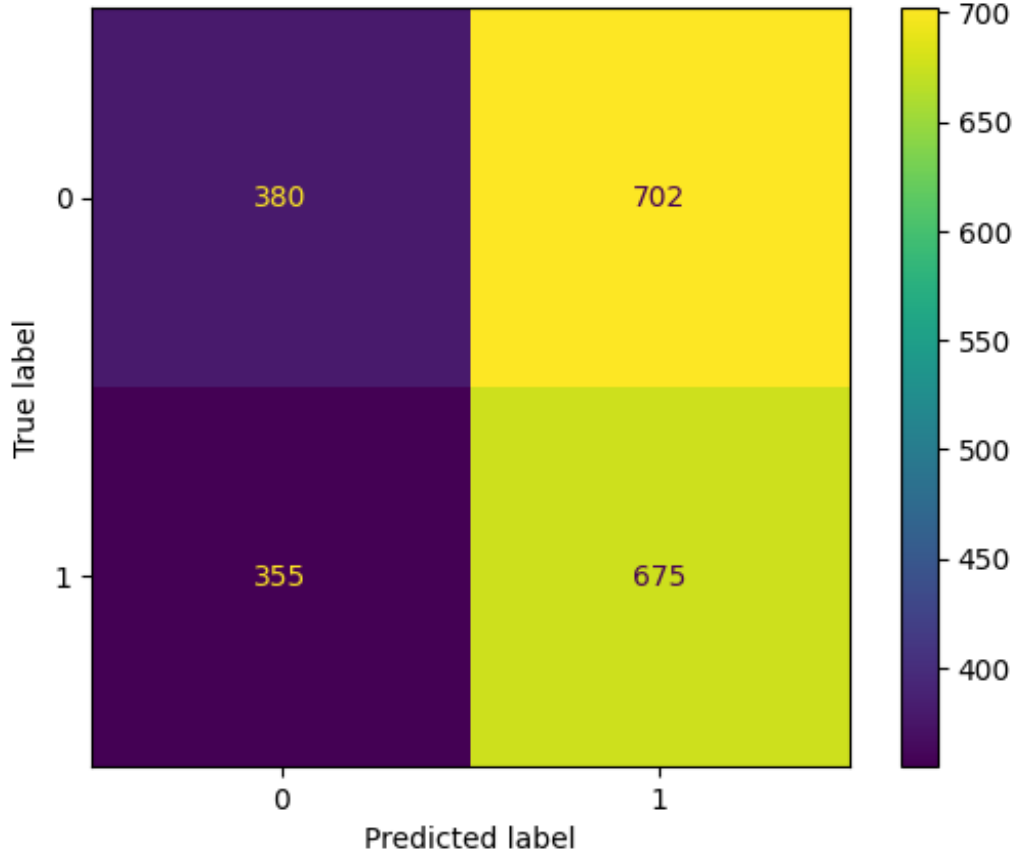
Результаты обучения

	Логистическая регрессия		Метод опорных векторов		Градиентный бустинг	
1	0.481	0.882	0.493	1.000	0.505	0.655
2	0.364	0.536	0.373	0.000	0.379	0.209
	Precision	|	Recall			

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$





Возможные проблемы и альтернативы

Возможные проблемы:

- Фундамент решения – агрегация исходных данных по суткам – сглаживание фона, но потеря информации
- Акцент на исходные данные в процессе анализа, но почти отсутствие анализа извлечённых признаков
- Расчёт практически всех признаков на основе только суточного среднего, игнорирование других результатов агрегации

Альтернативы:

- Агрегация по меньшим окнам: не 24, а 12 или 6 – это может поспособствовать сохранению ключевых значений временного ряда и при этом сократить шум
- Анализ извлечённых признаков: группировка и расчёт статистик, попарная визуализация признаков, метод главных компонент и др.
- Использовать разные исходные данные для разных признаков / обучать модель и с использованием результатов агрегации

В данной работе была попытка контекст учесть через формирование признаков на основе предыдущих значений. Можно также попробовать построить небольшую нейронную сеть с использованием рекуррентных блоков.

Заключение

- В ходе анализа были выявлены и подтверждены ключевые идеи, лежащие в основе исследования
- Были протестированы несколько из множества возможных подходов к решению задачи при помощи машинного обучения
- Несмотря на неудовлетворительный итоговый результат, остаётся большой простор для дальнейших экспериментов и работы с данными