Санкт-Петербургский государственный университет

Кафедра системного программирования

Байесовский подход к обнаружению разладки

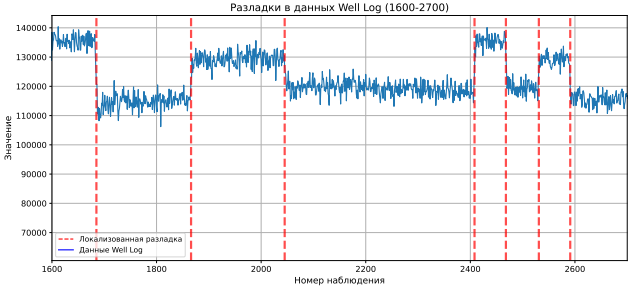
Алексей Дмитриевич Татьяненко

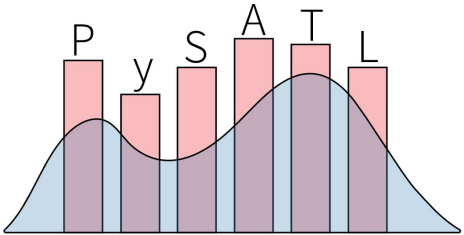
Научный руководитель: к.ф.-м.н. В. И. Гориховский, доцент кафедры системного программирования
Консультант: В. А. Кутуев, инженер-исследователь, Лаборатория технологий программирования инфраструктурных решений СПбГУ
Рецензент: к.ф.-м.н. М. С. Куля, старший инженер, ООО «Техкомпания Хуавэй»

Санкт-Петербург
2025

Введение

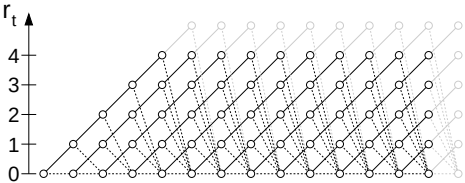
- Разладка (change point) — один из видов аномалий во временны́х рядах
- Задача её обнаружения возникает при обработке зашумлённых результатов бенчмаркинга [Matt Fleming, 2023], анализе поведения туристов [Zhan Zhao, 2018]
- Поскольку нет единых инструментов с широким набором алгоритмов, в проекте PySATL ведётся разработка пакета для обнаружения разладки
- Байесовский алгоритм является одним из наиболее точных и способен при некоторых модификациях решать задачу онлайн-обнаружения разладки



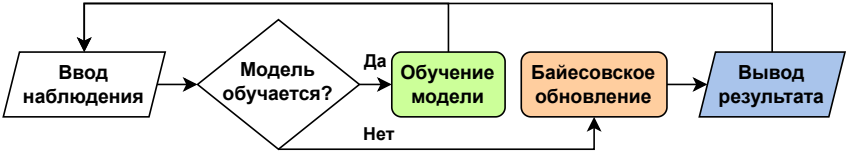


Байесовский подход

- Байесовский подход основан на оценке количества шагов от последней разладки
- Предложенный в оригинальной работе [Ryan Prescott Adams, 2007] алгоритм имеет недостатки:
  - Отсутствует критерий разладки
  - Не указаны конкретные предсказательные модели
  - Квадратичные затраты времени и памяти



Возможные длины пробега [Ryan Prescott Adams, 2007]



Постановка задач

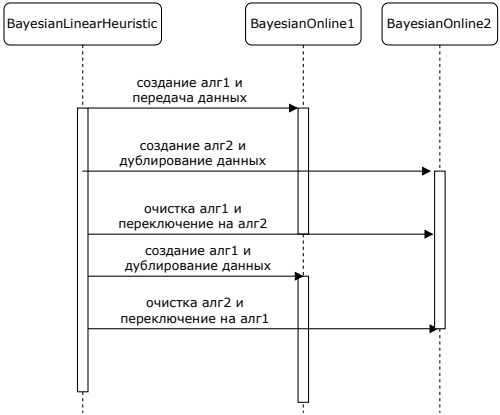
Цель: реализация онлайн-версии байесовского алгоритма, расширение списка предсказательных моделей (в том числе добавление адаптивной версии), настройка параметров алгоритма и оценка его эффективности.

Задачи:

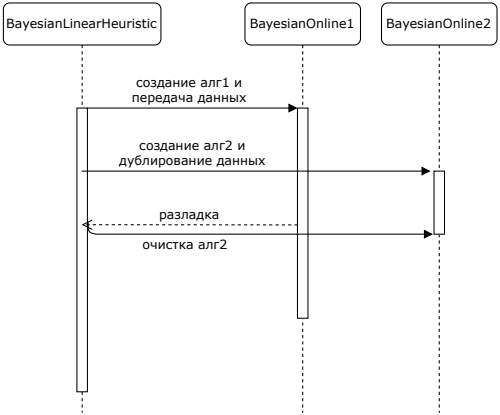
1. Адаптация алгоритма под задачу онлайн-обнаружения разладки
2. Реализация онлайн-версии байесовского алгоритма и его интеграция в PySATL-CPD
3. Интеграция альтернативных предсказательных моделей: экспоненциальной и адаптивной
4. Настройка параметров и оценка эффективности онлайн-алгоритма

Линейная эвристика

Случай без разладки

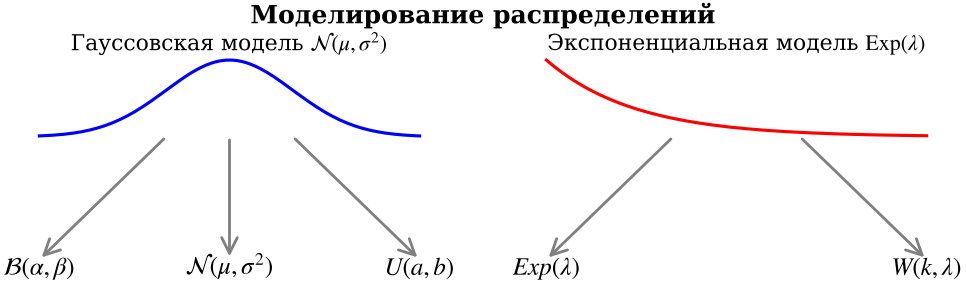


Случай с разладкой

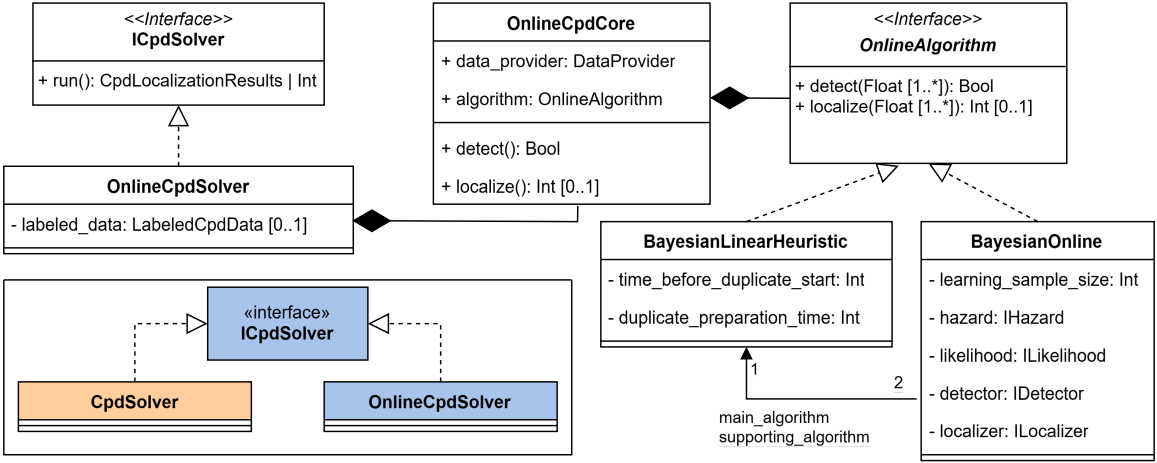


Предсказательные модели

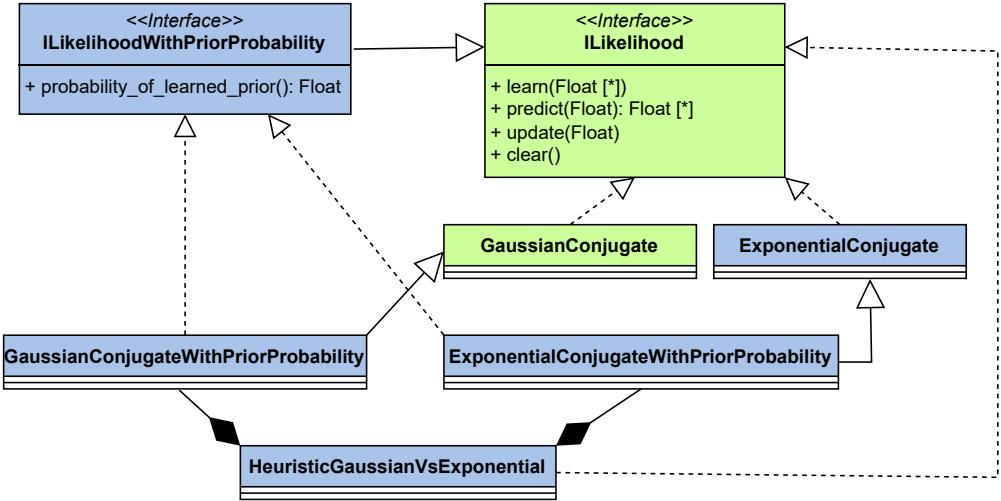
- Предсказательная модель оценивает правдоподобие принадлежности данных распределению с оценёнными параметрами
- В общем случае требует интегрирования и решения оптимизационной задачи, но это может быть заменено на явную оценку для некоторых семейств распределений (сопряжённые априорное/апостериорное)
- Позволяет по обучающей выборке оценивать начальные априорные параметры
- Рассматриваются гауссовская, экспоненциальная и эвристическая модели



Интеграция онлайн-алгоритма в PySATL-CPD



Реализация предсказательных моделей



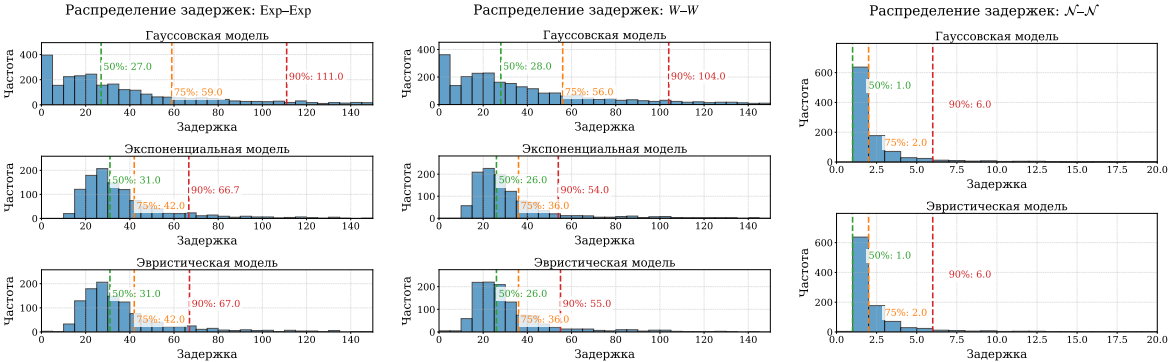
Настройка параметров алгоритма

- 25 конфигураций без разладки, 5 с разладкой, на каждую конфигурацию 1000 массивов данных
- Размер массива данных 500, разладка (в случае наличия) — 250, допустимое отклонение локализации — 25
- Эвристическая адаптивная модель, за исключением конфигураций $\mathcal{B}-\mathcal{B}$ и $W-U$, показала мощность не менее 0.79
- Среднее время обработки одного массива — 1.2 с., $\sigma = 0.4$ с.

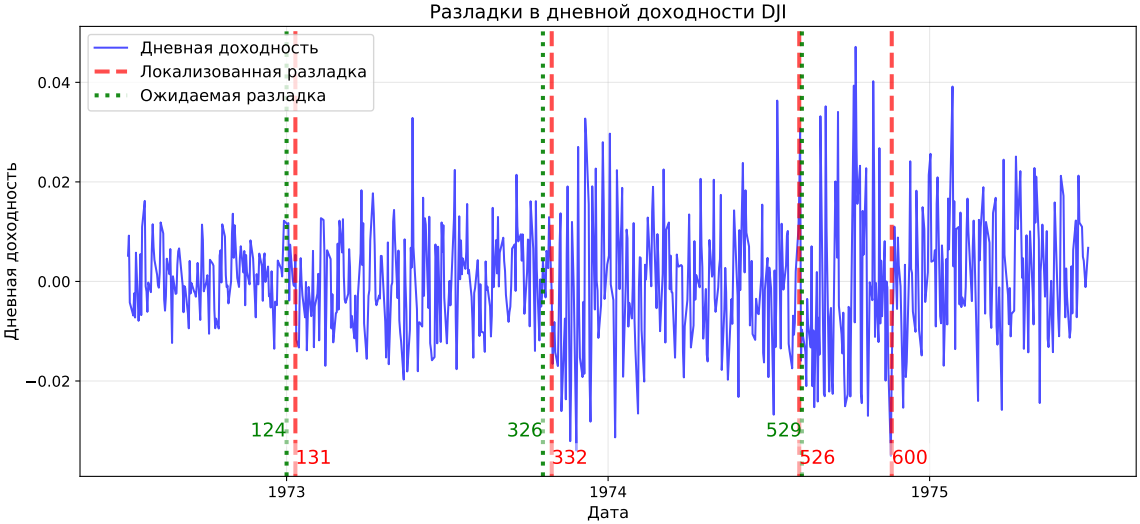
$\alpha_{\mathcal{N}}$	0.1	0.05	0.01	0.005
β\Порог	0.27	0.18	0.07	0.04
$\beta_{\mathcal{N}-\mathcal{B}}$	0.794	0.824	0.864	0.878
$\beta_{\mathcal{N}-Exp}$	0.779	0.805	0.840	0.863
$\beta_{\mathcal{N}-U}$	0.870	0.893	0.917	0.929
$\beta_{\mathcal{N}-W}$	0.783	0.811	0.853	0.863
$\beta_{\mathcal{N}-\mathcal{N}}$	0.968	0.975	0.981	0.981
$\beta_{\mathcal{B}-\mathcal{N}}$	0.997	0.998	1.0	1.0
$\beta_{Exp-\mathcal{N}}$	0.512	0.574	0.681	0.724
$\beta_{U-\mathcal{N}}$	0.998	1.0	1.0	1.0
$\beta_{W-\mathcal{N}}$	0.47	0.523	0.623	0.665

Таблица: Мощности для гауссовской модели

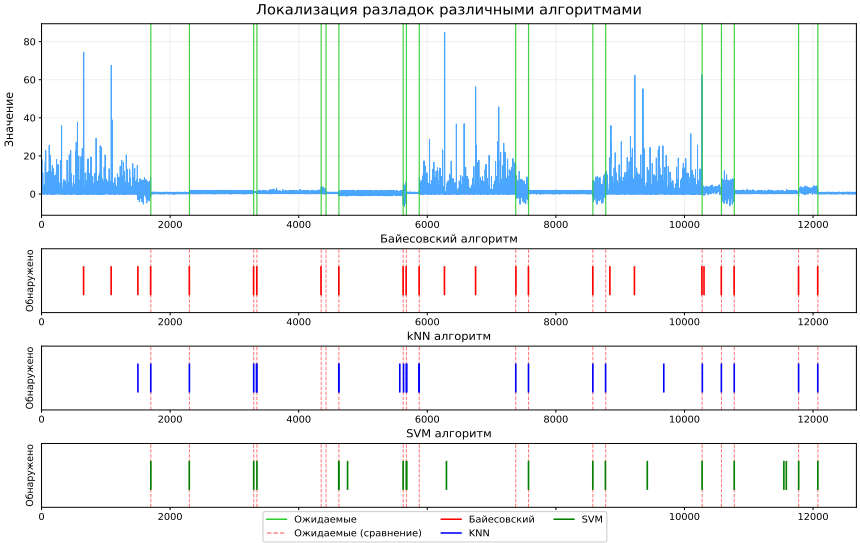
Задержки локализации



Данные доходности DJI



Сравнение алгоритмов на больших данных



Метрика	Байес	kNN	SVM
TP	15	15	13
FP	10	9	8
FN	4	5	7
Precision	0.60	0.62	0.62
Recall	0.79	0.75	0.65
F-score	0.68	0.68	0.63
Q_{50}	11	17	17
Q_{75}	80	23	23
Q_{90}	107	25	25

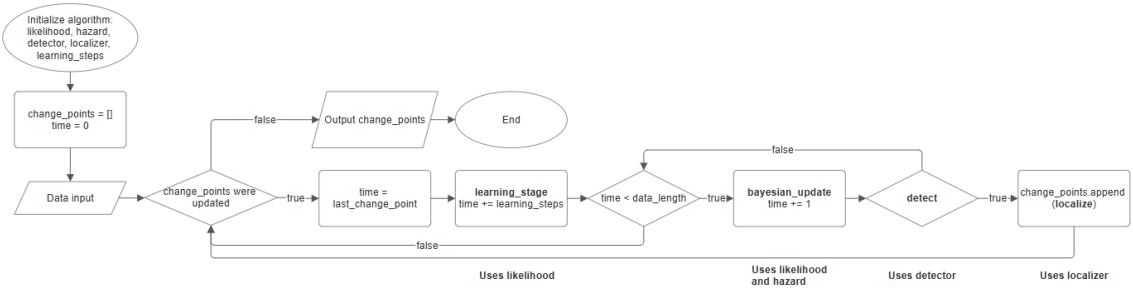
Заключение[1]

1. Проведена адаптация байесовского алгоритма под онлайн-задачу, добавлены обучение модели, детекция и локализация, предложена эвристика для получения линейной сложности
2. Адаптированный алгоритм реализован и интегрирован в PySATL-CPD
3. Были добавлены экспоненциальная и адаптивная (с эвристикой выбора функции правдоподобия) предсказательные модели
4. Проведена настройка параметров алгоритма, исследовано поведение алгоритма при решении онлайн-задачи обнаружения разладки, на его основе оценена применимость линейной эвристики

Репозиторий (nick: alexdtat): `https://github.com/PySATL/pysatl-cpd`

[1] Результаты работы также были представлены на конференции «МАТ-МЕХ. НАУКА 2025», Mendeleev AI seminar, поданы на DAMDID 2025

Базовая версия байесовского алгоритма



Обучение эвристической адаптивной модели

Algorithm Learning for Heuristic Adaptive Model

```
function Learn(array[float] LearningSample)
    GaussianConjugate.Learn(LearningSample)
    ExponentialConjugate.Learn(LearningSample)

    GaussianPrior ← GaussianConjugate.PriorProbability()
    ExponentialPrior ← ExponentialConjugate.PriorProbability()

    if GaussianPrior ≥ ExponentialPrior then
        SelectedModel ← GaussianConjugate
    else
        SelectedModel ← ExponentialConjugate
```

Гауссовская модель: мощности локализации

α	0.1	0.05	0.01	0.005
β\Порог	0.27	0.18	0.07	0.04
$\beta_{\mathcal{N}-\mathcal{B}}$	0.794	0.824	0.864	0.878
$\beta_{\mathcal{N}-Exp}$	0.779	0.805	0.840	0.863
$\beta_{\mathcal{N}-U}$	0.870	0.893	0.917	0.929
$\beta_{\mathcal{N}-W}$	0.783	0.811	0.853	0.863
$\beta_{\mathcal{N}-\mathcal{N}}$	0.968	0.975	0.981	0.981
$\beta_{\mathcal{B}-\mathcal{N}}$	0.997	0.998	1.0	1.0
$\beta_{Exp-\mathcal{N}}$	0.512	0.574	0.681	0.724
$\beta_{U-\mathcal{N}}$	0.998	1.0	1.0	1.0
$\beta_{W-\mathcal{N}}$	0.47	0.523	0.623	0.665

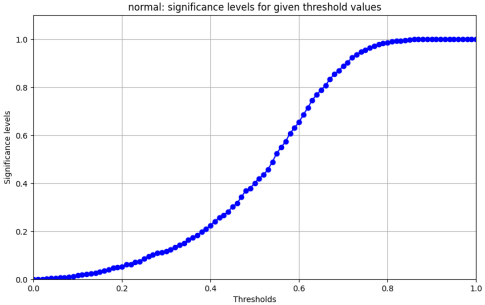
β\Порог	0.27	0.18	0.07	0.04
$\beta_{\mathcal{B}-\mathcal{B}}$	0.945	0.96	0.976	0.983
$\beta_{\mathcal{B}-Exp}$	0.818	0.839	0.869	0.881
$\beta_{\mathcal{B}-U}$	0.993	0.996	0.998	0.998
$\beta_{\mathcal{B}-W}$	0.813	0.832	0.862	0.876
$\beta_{Exp-\mathcal{B}}$	0.309	0.353	0.439	0.477
$\beta_{Exp-Exp}$	0.337	0.393	0.471	0.508
β_{Exp-U}	0.406	0.465	0.543	0.594
β_{Exp-W}	0.374	0.425	0.518	0.544
$\beta_{U-\mathcal{B}}$	0.892	0.917	0.946	0.955
β_{U-Exp}	0.829	0.851	0.875	0.888
β_{U-U}	0.987	0.991	0.996	0.997
β_{U-W}	0.851	0.877	0.904	0.914
$\beta_{W-\mathcal{B}}$	0.324	0.370	0.456	0.498
β_{W-Exp}	0.373	0.421	0.516	0.561
β_{W-U}	0.374	0.423	0.504	0.538
β_{W-W}	0.351	0.416	0.508	0.553

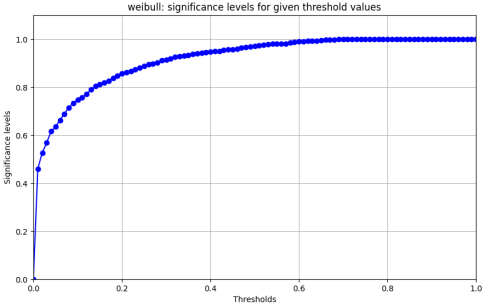
Гауссовская модель: мощности детекции

β\Порог	0.27	0.18	0.07	0.04
$\beta_{\mathcal{B}-\mathcal{B}}$	1.0	1.0	1.0	1.0
$\beta_{\mathcal{B}-Exp}$	1.0	1.0	1.0	1.0
$\beta_{\mathcal{B}-\mathcal{N}}$	1.0	1.0	1.0	1.0
$\beta_{\mathcal{B}-W}$	1.0	1.0	1.0	1.0
$\beta_{\mathcal{B}-U}$	1.0	1.0	1.0	1.0
$\beta_{Exp-\mathcal{B}}$	0.999	0.998	0.998	0.998
$\beta_{Exp-Exp}$	1.0	1.0	1.0	1.0
$\beta_{Exp-\mathcal{N}}$	1.0	1.0	1.0	1.0
β_{Exp-U}	1.0	1.0	1.0	1.0
β_{Exp-W}	1.0	1.0	1.0	1.0
$\beta_{\mathcal{N}-\mathcal{B}}$	1.0	1.0	1.0	1.0
$\beta_{\mathcal{N}-Exp}$	1.0	1.0	1.0	1.0
$\beta_{\mathcal{N}-\mathcal{N}}$	1.0	1.0	1.0	1.0

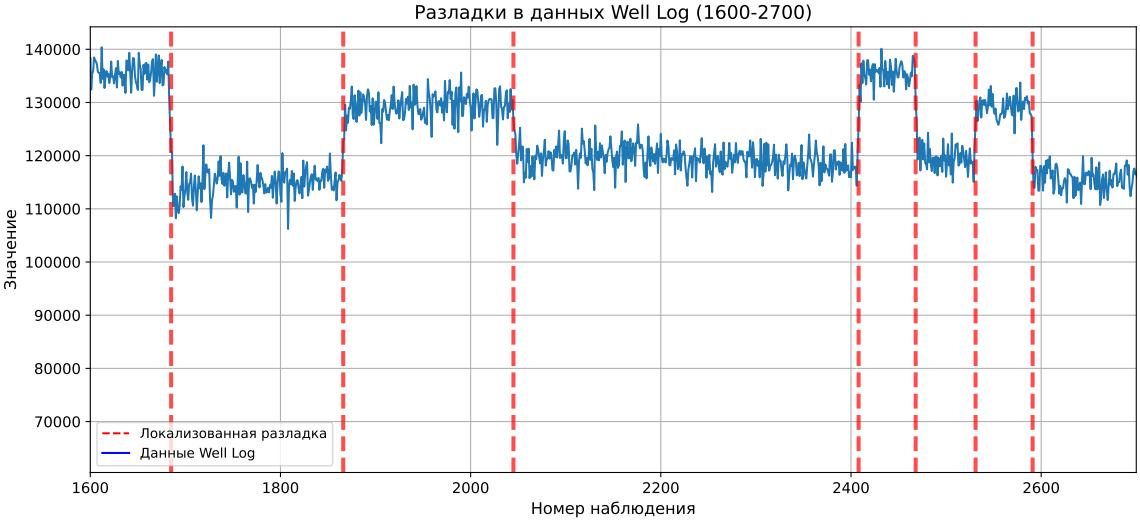
β\Порог	0.27	0.18	0.07	0.04
$\beta_{\mathcal{N}-U}$	1.0	1.0	1.0	1.0
$\beta_{\mathcal{N}-W}$	1.0	1.0	1.0	1.0
$\beta_{U-\mathcal{B}}$	1.0	1.0	1.0	1.0
β_{U-Exp}	1.0	1.0	1.0	1.0
$\beta_{U-\mathcal{N}}$	1.0	1.0	1.0	1.0
β_{U-U}	1.0	1.0	1.0	1.0
β_{U-W}	1.0	1.0	1.0	1.0
$\beta_{W-\mathcal{B}}$	0.995	0.994	0.993	0.992
β_{W-Exp}	1.0	1.0	1.0	1.0
$\beta_{W-\mathcal{N}}$	1.0	1.0	1.0	1.0
β_{W-U}	0.986	0.984	0.978	0.978
β_{W-W}	1.0	1.0	1.0	1.0

Уровень значимости и пороги (примеры)

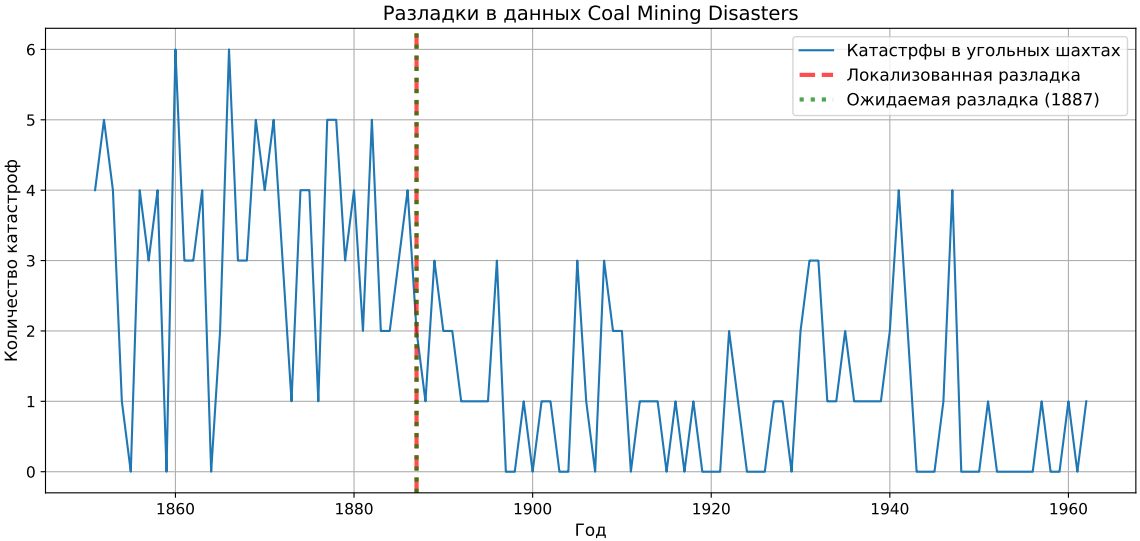




Well Log


Разладки в данных Well Log (1600-2700)
Значение
Номер наблюдения
140000
130000
120000
110000
100000
90000
80000
70000
1600
1800
2000
2200
2400
2600
Локализованная разладка
Данные Well Log

Coal Mining Disasters



Задачи детекции и локализации разладки

- Гипотеза разладки:

$$\{x_i\}, i = 1, \ldots, n$$
$$H_0 \colon x_i \sim F_0, i = 1, \ldots, n$$
$$H_a \colon \exists 1 \leq \tau < n, x_i \sim \begin{cases} F_1, i > \tau \\ F_0, \text{иначе} \end{cases}$$

- Гипотеза корректной локализации:

$$\{x_i\}, i = 1, \ldots, n$$
$$H_0 \colon x_i \sim F_0, i = 1, \ldots, n$$
$$H_a \colon 1 \leq \tau_1 < n \Rightarrow \exists \tau \in [\tau_1 - k; \tau_1 + k], x_i \sim \begin{cases} F_1, i > \tau \\ F_0, \text{иначе} \end{cases}$$

τ_1 — корректно распознанная с точностью k разладка

Оценка длины пробега

Имеем вероятность вида $P(r_t \mid \boldsymbol{x}_{1:t}) = P(r_t, \boldsymbol{x}_{1:t})/P(\boldsymbol{x}_{1:t})$, где при условии независимости наблюдений $P(r_t, \boldsymbol{x}_{1:t})$ раскладывается в сумму:

$$P(r_t, \boldsymbol{x}_{1:t}) = \sum_{r_{t-1}} P(r_t, r_{t-1}, \boldsymbol{x}_{1:t}) =$$

$$= \sum_{r_{t-1}} P(r_t, x_t \mid r_{t-1}, \boldsymbol{x}_{1:t-1}) P(r_{t-1}, \boldsymbol{x}_{1:t-1}) =$$

$$= \sum_{r_{t-1}} P(r_t \mid r_{t-1}) P(x_t \mid r_{t-1}, \boldsymbol{x}_t^{(r)}) P(r_{t-1}, \boldsymbol{x}_{1:t-1}).$$

Функция выживаемости

- **Функция выживаемости (hazard function)** определяет вероятность обнуления длины пробега в зависимости от прошлой длины пробега:

$$H(\tau) = \frac{P_{\text{gap}}(g = \tau)}{\sum_{t=\tau}^{\infty} P_{\text{gap}}(g = t)},$$

  где $P_{\text{gap}}(g)$ — дискретное априорное распределение на промежутке между разладками.
- Позволяет вычислять условную априорную вероятность на разладку:

$$P(r_t \mid r_{t-1}) = \begin{cases} H(r_{t-1}+1) & \text{если } r_t = 0 \\ 1 - H(r_{t-1}+1) & \text{если } r_t = r_{t-1} + 1 \\ 0 & \text{иначе} \end{cases}.$$

- Если $P_{\text{gap}}(g)$ — геометрическое, то функция правдоподобия — константа: $H(\tau) = 1/\lambda$

Гауссовская модель

Гауссовская функция правдоподобия, используется нормальное-обратное гамма сопряжённое:

- Обучение гиперпараметров:

$$\mu_0 = \bar{x}^{learn}; \qquad k_0 = n; \alpha_0 = n/2; \qquad \beta_0 = \sum_{i=1}^{n} \frac{(x_i^{learn} - \bar{x}^{learn})^2}{2}$$

- Обновление гиперпараметров:

$$\mu_{t+1} = \frac{k_t \mu_t + x_t^{obs}}{k_t + 1}; \quad k_{t+1} = k_t + 1; \alpha_{t+1} = \alpha_t + 1/2; \quad \beta_{t+1} = \beta_t + \frac{k_t}{k_t + 1} \frac{(x_t^{obs} - \mu_t)^2}{2}$$

- Предсказательная вероятность:

$$P\left(x_{t+1}^{obs} \mid r_t, \boldsymbol{x}_{t+1}^{(r)}\right) = t_{2\alpha_{t+1}}\left(x_{t+1}^{obs} \mid \mu_{t+1}, \frac{\beta_{t+1}(k_{t+1} + 1)}{k_{t+1}\alpha_{t+1}}\right)$$

Экспоненциальная модель

Экспоненциальная функция правдоподобия, используется гамма сопряжённое:

- Обучение гиперпараметров:

$$\alpha_0 = n; \qquad \beta_0 = \sum_{i=1}^{n} x_i^{learn}$$

- Обновление гиперпараметров:

$$\alpha_{t+1} = \alpha_t + 1 \qquad \beta_{t+1} = \beta_t + x_t^{obs};$$

- Предсказательная вероятность:

$$P(x_{t+1}^{obs} \mid r_t, \boldsymbol{x}_{t+1}^{(r)}) = \mathrm{Lomax}(x_{t+1}^{obs} \mid \beta_{t+1}, \alpha_{t+1})$$