Всероссийская научная конференция с международным участием
Параллельные вычислительные технологии (ПаВТ'2023)
Санкт-Петербург, 28–30 марта 2023 г.

Автоматизированный поиск аномалий временных рядов на графическом процессоре

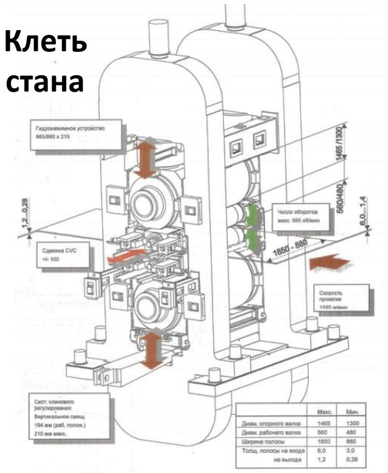
Я.А. Краева, М.Л. Цымблер

Южно-Уральский государственный университет (Челябинск)

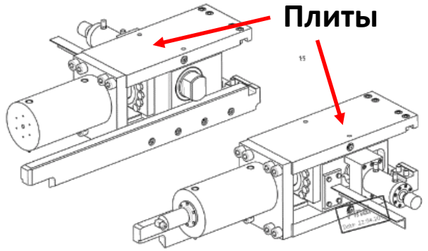
Работа выполнена при финансовой поддержке Российского научного фонда (грант № 23-21-00465).

Поиск аномалий во временных рядах из цифровой индустрии

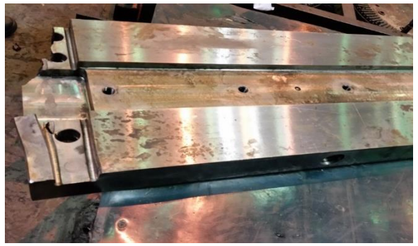
Клеть стана



Плиты

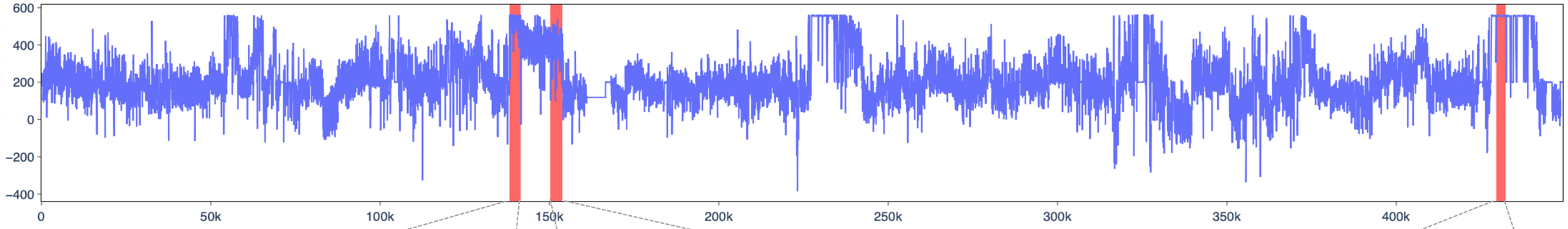


Система валков с непрерывно изменяемой кривизной

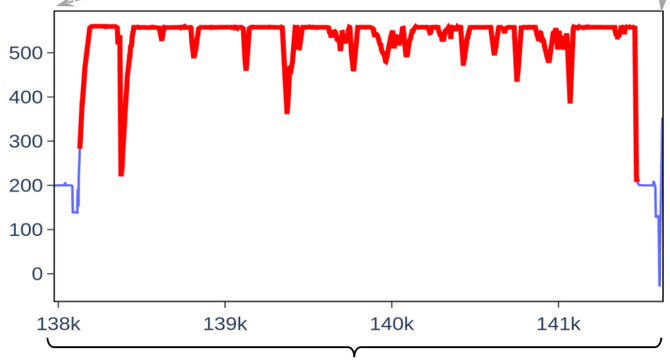


Разрушение плит

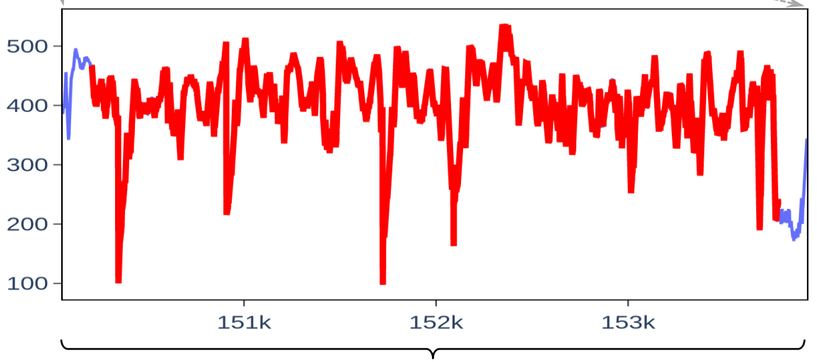
Фактическое изгибающее усилие прокатки



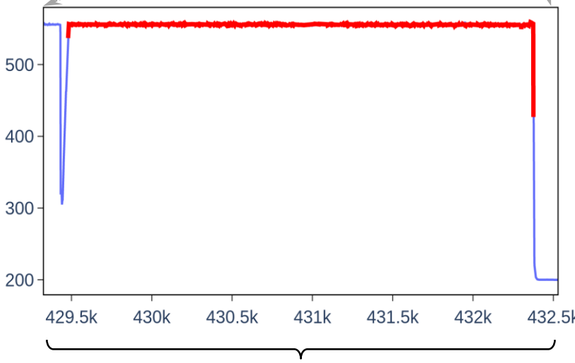
Аномальные напряжения плит



55 минут



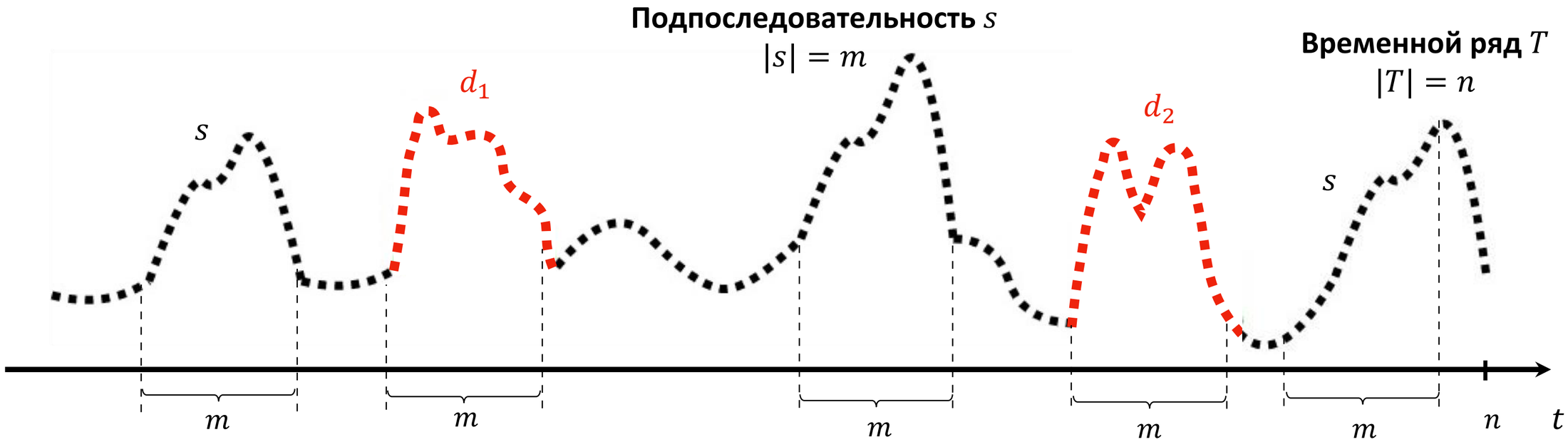
1 час



48 минут

Постановка задачи

- **Диссонанс**[1)] – подпоследовательность ряда, расстояние от которой до ближайшего соседа не ниже порога r
- **Дано:** временной ряд T, длина диссонанса m, порог r
- **Найти:** $D = \{d_1, d_2, \ldots\}$, $d_i \in D \Leftrightarrow \forall s \in T \min\limits_{s \cap d_i = \emptyset} \mathrm{dist}(d_i, s) \geq r$



Количество подпоследовательностей: $N = n - m + 1$

[1)] Yankov D., Keogh E.J., Rebbapragada U. Disk aware discord discovery: finding unusual time series in terabyte sized datasets. Knowl. Inf. Syst. 17(2): 241-262. 2008.

Основные работы по теме исследования

Алгоритм	Платформа	Критика
Последовательный алгоритм		
Nakamura T., *et al.* **MERLIN**: parameter-free discovery of arbitrary length anomalies in massive time series archives. IEEE ICDM 2020. pp. 1190-1195.	CPU	Квадратичная сложность от длины ряда
Параллельные алгоритмы		
DRAG: Yankov D., *et al*. Disk aware discord discovery: finding unusual time series in terabyte sized datasets. Knowl. Inf. Syst. 17(2): 241-262. 2008.	CPU	Симуляция MapReduce
PDD: Huang T., *et al*. Parallel discord discovery. PAKDD 2016. LNCS 9652. Springer, 2016. pp. 233-244.	Spark	Низкая производительность ввиду большого количества обменов между узлами
PhiDD: Zymbler M., *et al.* A Parallel Approach to Discords Discovery in Massive Time Series Data. Computers, Materials & Continua 66(2): 1867-1876. 2021.	Кластер Intel Xeon Phi	Квадратичная пространственная сложность от длины ряда
KBF_GPU: Thuy T.T.H., *et al*. A new discord definition and an efficient time series discord detection method using GPUs. ICSED 2021. pp. 63-70.	GPU	Полный перебор подпоследовательностей ряда
Zhu B., *et al*. A GPU Acceleration framework for motif and discord based pattern mining. IEEE Trans. on Parallel and Distr. Systems 32(8): 1987-2004. 2021.	GPU	Поиск одного (самого важного) диссонанса ряда
PD3: Zymbler M., Kraeva Ya. Parallel algorithm for time series discord discovery on a graphics processor. Pattern Recognition and Image Analysis 33(2). 2023.	GPU	Ручной подбор длины диссонанса и порога

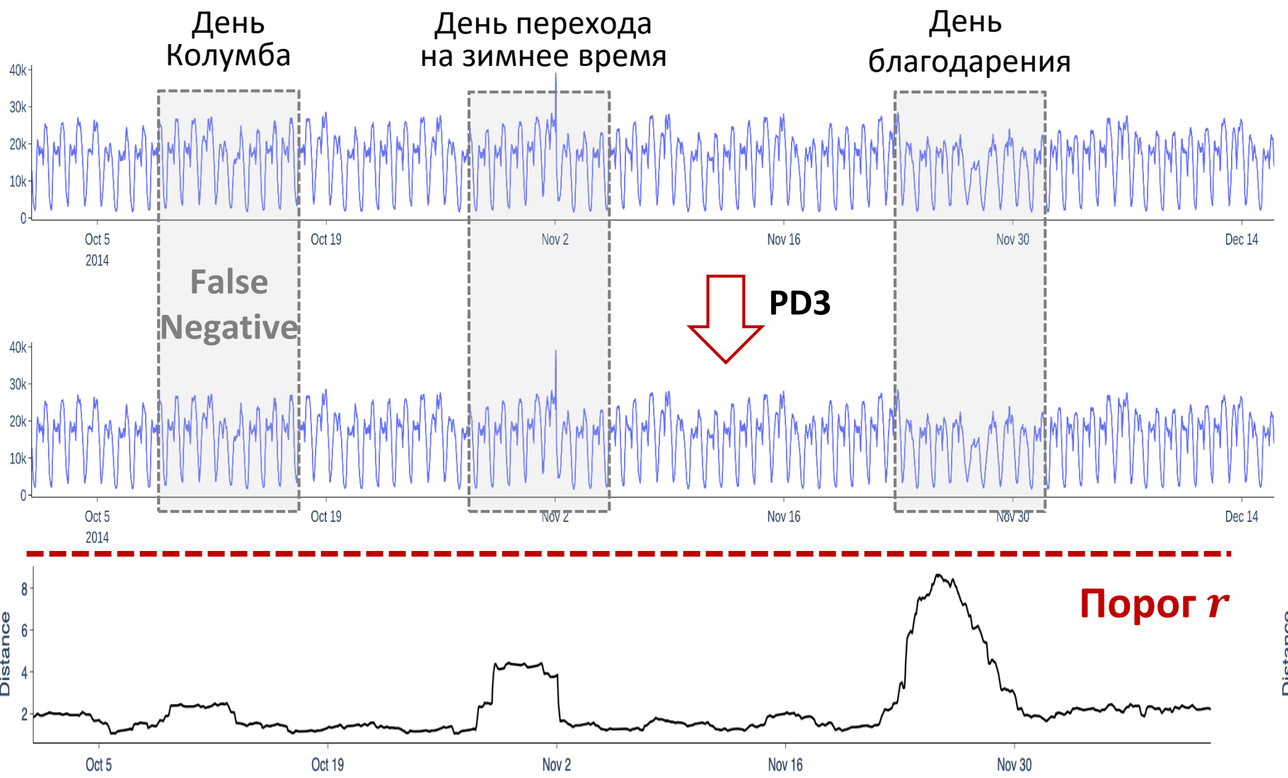
PD3 (Parallel DRAG-based Discord Discovery): Ручной подбор r

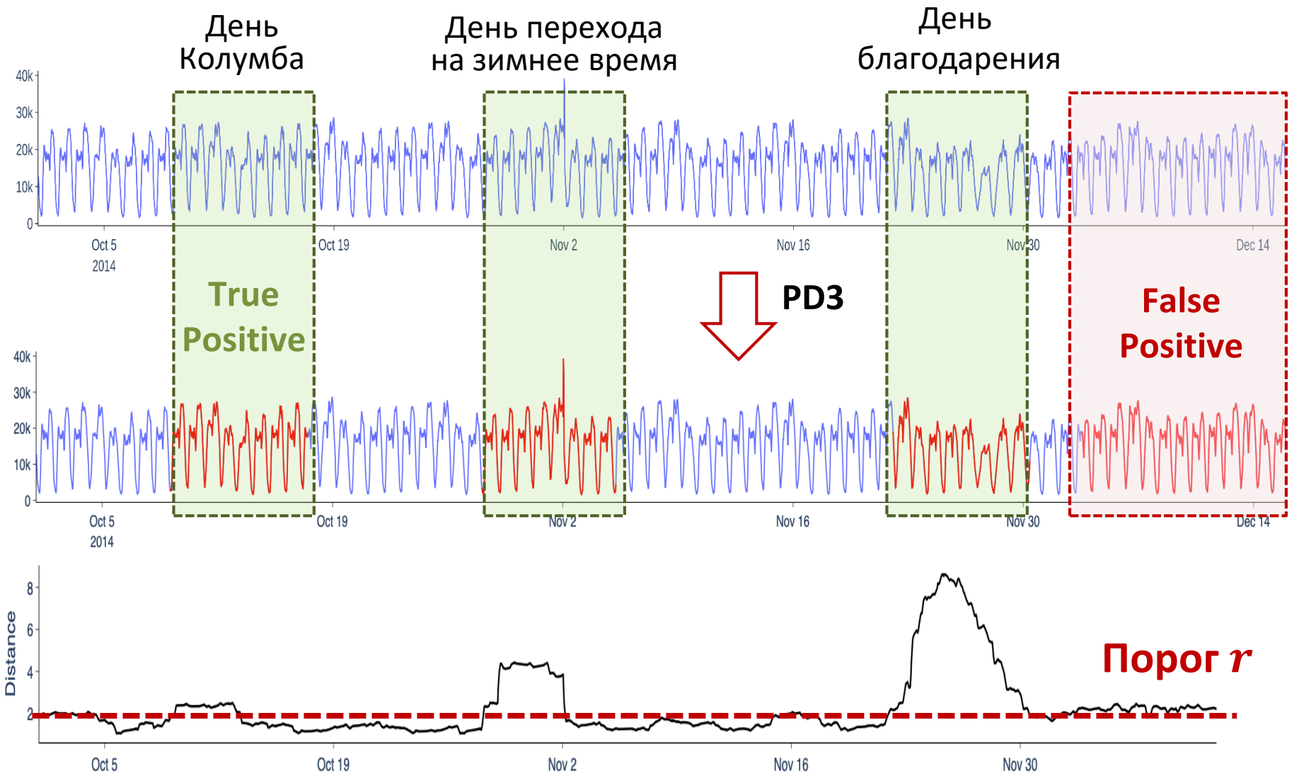
Порог r

Большой → **НЕТ** диссонансов

Маленький → Много **ЛОЖНЫХ** диссонансов

Среднее число пассажиров NY такси (осень 2014 г., каждые полчаса)



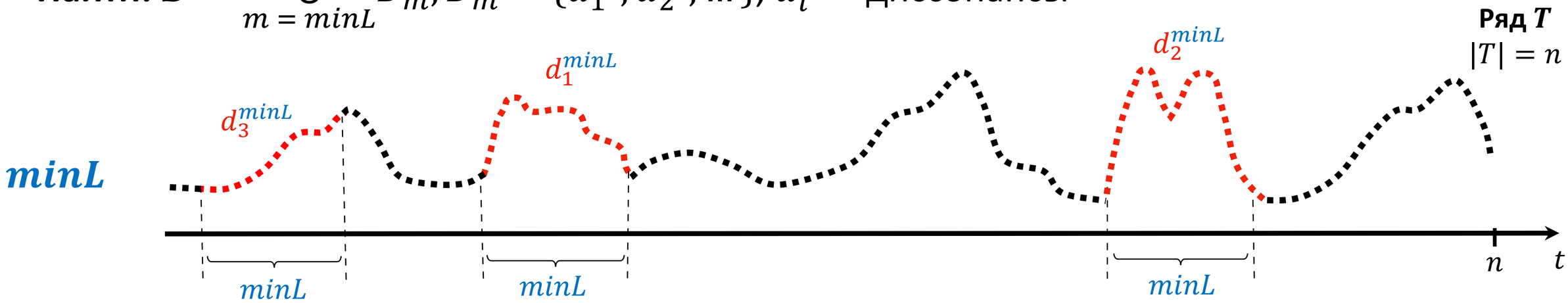


PD3 (Parallel DRAG-based Discord Discovery): Ручной подбор m

Накладные расходы на подбор параметров m и r

Постановка задачи автоматизированного поиска диссонансов[1)]

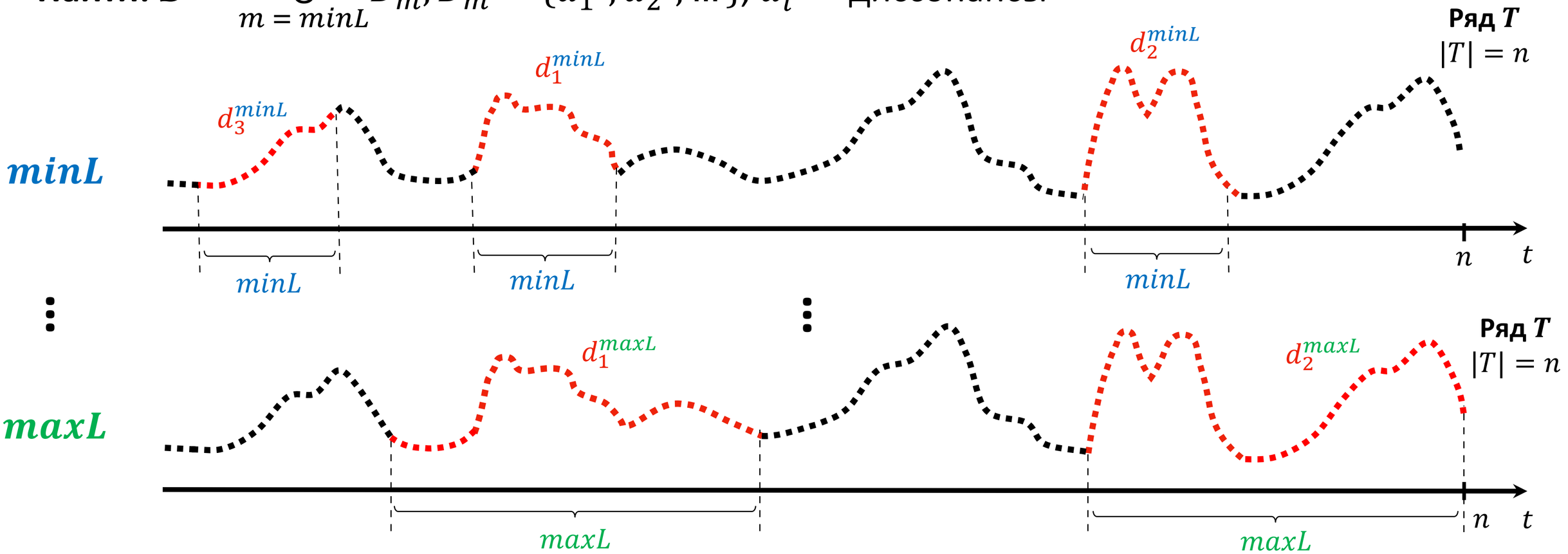
- **Дано:** временной ряд T, диапазон длин диссонансов $minL, \dots, maxL$
- **Найти:** $\mathcal{D} = \bigcup_{m=minL}^{maxL} D_m, D_m = \{d_1^m, d_2^m, \dots\}$, d_i^m – диссонансы



[1)] Nakamura T., *et al*. MERLIN: parameter-free discovery of arbitrary length anomalies in massive time series archives. IEEE ICDM 2020. pp. 1190-1195.

Постановка задачи автоматизированного поиска диссонансов[1)]

- **Дано:** временной ряд T, диапазон длин диссонансов $minL, \dots, maxL$
- **Найти:** $\mathcal{D} = \bigcup_{m=minL}^{maxL} D_m, D_m = \{d_1^m, d_2^m, \dots\}$, d_i^m – диссонансы



[1)] Nakamura T., *et al*. MERLIN: parameter-free discovery of arbitrary length anomalies in massive time series archives. IEEE ICDM 2020. pp. 1190-1195.

PALMAD: **P**arallel **A**rbitrary **L**ength **M**ERLIN-based **A**nomaly **D**iscovery

1. Применение $\mathrm{ED}^2_{\mathrm{norm}}$ в качестве функции расстояния[1)]

$$\mathrm{ED}^2_{\mathrm{norm}}(T_{i,m}, T_{j,m}) = 2m\left(1 - \frac{T_{i,m} \cdot T_{j,m} - m\mu_i\mu_j}{m\sigma_i\sigma_j}\right)$$

2. Сокращение избыточных вычислений μ и σ при вычислении $\mathrm{ED}^2_{\mathrm{norm}}$

Лемма. Пусть даны ряд T, $|T| = n$ и подпоследовательности $T_{i,m}$ и $T_{i,m+1}$. Тогда

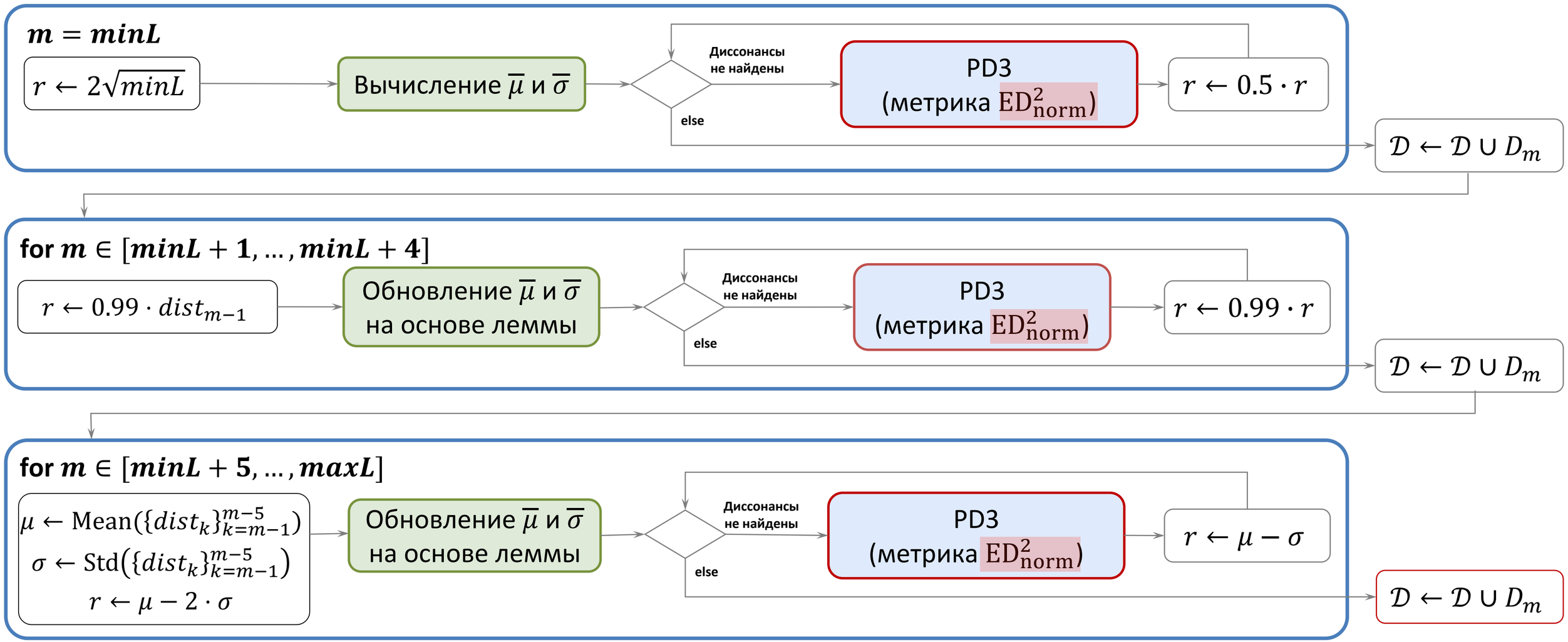
$$\mu_{T_{i,m+1}} = \frac{1}{m+1}\left(m\mu_{T_{i,m}} + t_{i+m}\right), \quad \sigma^2_{T_{i,m+1}} = \frac{m}{m+1}\left(\sigma^2_{T_{i,m}} + \frac{1}{m+1}\left(\mu_{T_{i,m}} - t_{i+m}\right)^2\right).$$

3. Автоматизированный подбор порога r

4. Тепловая карта диссонансов

[1)] Mueen A. et al. Fast approximate correlation for massive time-series data. SIGMOD 2010. pp. 171-182. ACM (2010). https://doi.org/10.1145/1807167.1807188

Схема PALMAD



Подбор порога r

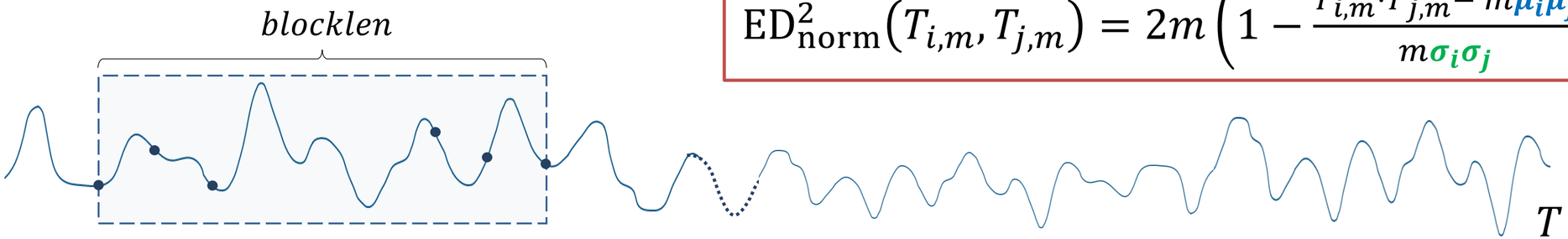
Предобработка (сокращение вычислений)

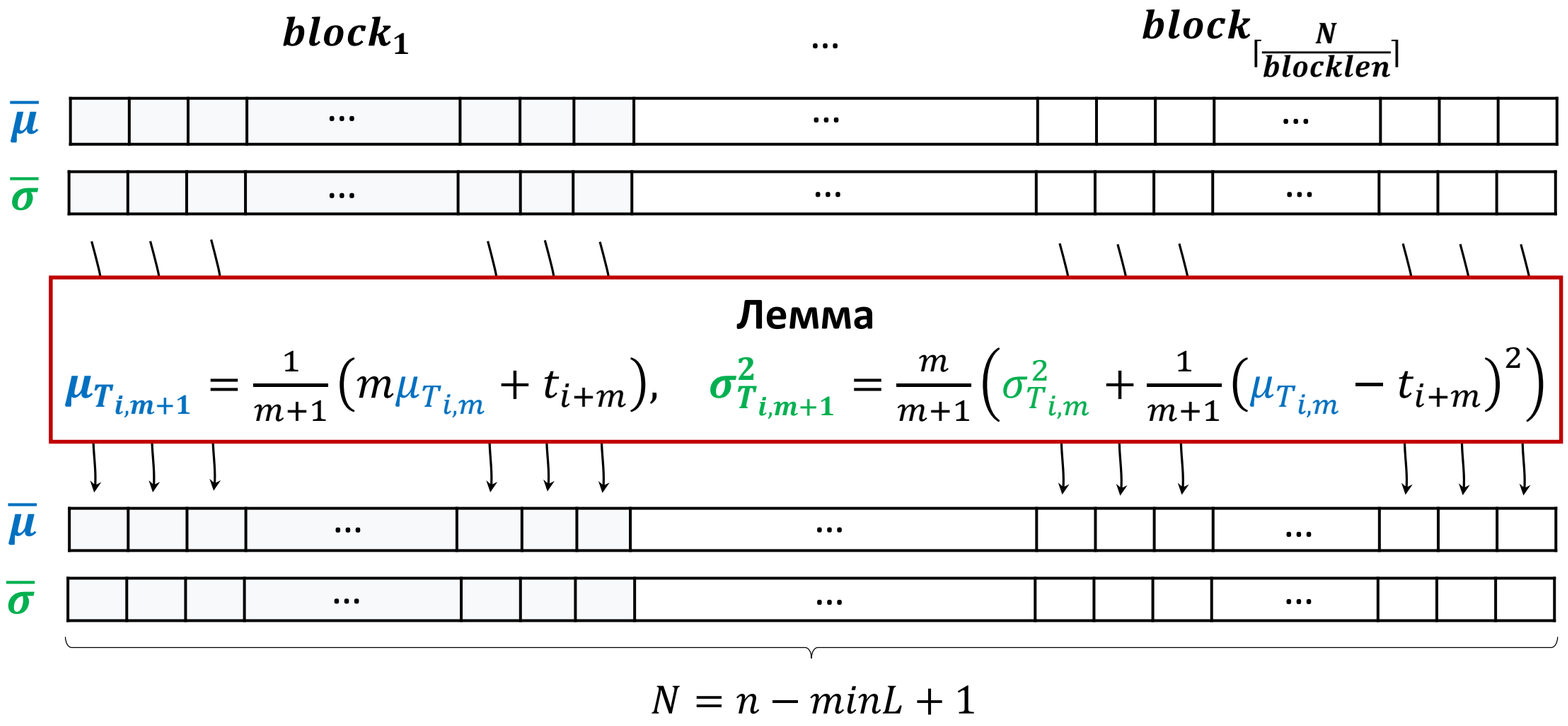
Поиск диссонансов (использование PD3 и $\mathrm{ED}^2_{\mathrm{norm}}$)

Сокращение избыточных вычислений $\overline{\mu}$ и $\overline{\sigma}$

$$\mathrm{ED}^2_{\mathrm{norm}}(T_{i,m}, T_{j,m}) = 2m\left(1 - \frac{T_{i,m} \cdot T_{j,m} - m\mu_i\mu_j}{m\sigma_i\sigma_j}\right)$$

Вычисленные $\overline{\mu}$ и $\overline{\sigma}$ для подп-ей ряда длины, меньшей на 1

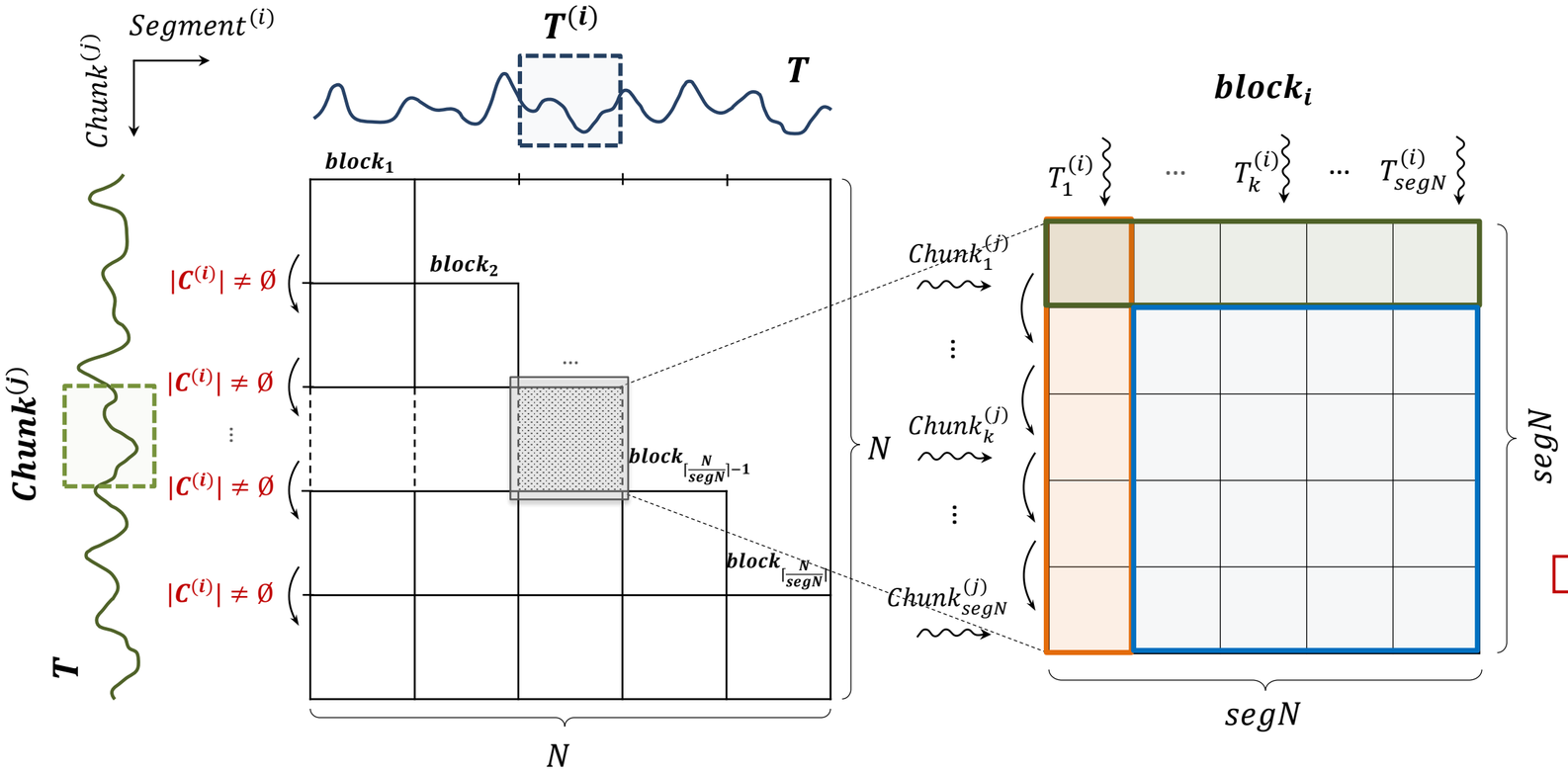




PD3: Parallel DRAG-based Discord Discovery

$$\mathrm{ED}^2_{\mathrm{norm}}\left(T_{i,m}, T_{j,m}\right) = 2m\left(1 - \frac{T_{i,m} \cdot T_{j,m} - m\mu_i\mu_j}{m\sigma_i\sigma_j}\right)$$

1. PD3 включает 2 фазы: отбор кандидатов в диссонансы и очистка кандидатов
2. Использование концепции параллелизма по данным
3. Эффективное вычисление скалярных произведений $T_{i,m} \cdot T_{j,m}$

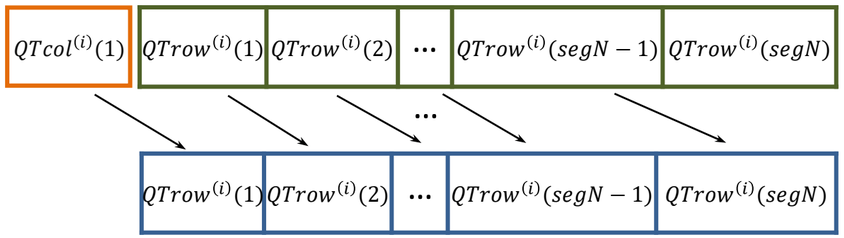


$$\boldsymbol{QTrow^{(i)}(tid)} = \sum_{k=1}^{m} T_{tid}^{(i)}(k) \cdot Chunk_1^{(j)}(k)$$

$$\boldsymbol{QTcol^{(i)}(tid)} = \sum_{k=1}^{m} T_1^{(i)}(k) \cdot Chunk_{tid}^{(j)}(k)$$

Выч. сложность $O(1)$ вместо $O(m)$!

$$\boldsymbol{QTrow^{(i)}(tid)} = QTrow^{(i)}(tid-1) - T_{tid-1}^{(i)}(1) \cdot Chunk_{tid-1}^{(j)}(1) + T_{tid}^{(i)}(m) \cdot Chunk_{tid}^{(j)}(m)$$



Эксперименты

- **Конкуренты** (поиск top-1 диссонанса на GPU)
  - **KBF_GPU**: Thuy T.T.H. et al. A new discord definition and an efficient time series discord detection method using GPUs. ICSED 2021. pp. 63–70. https://doi.org/10.1145/3507473.3507483.
  - **Zhu et al.** A GPU Acceleration framework for motif and discord based pattern mining. IEEE Transactions on Parallel and Distributed Systems 32(8): 1987–2004. 2021. https://doi.org/10.1109/TPDS.2021.3055765.
- **Данные**[1,2)

Временной ряд	Длина ряда, n	Длина диссонанса, $minL = maxL$
Space shuttle	5 000	150
ECG	45 000	200
ECG2	21 600	400
Koski-ECG	100 000	458
Power demand	33 220	750
Respiration	24 125	250
RandomWalk1M	$1 \cdot 10^7$	512
RandomWalk2M	$2 \cdot 10^7$	512

- **Аппаратные платформы**

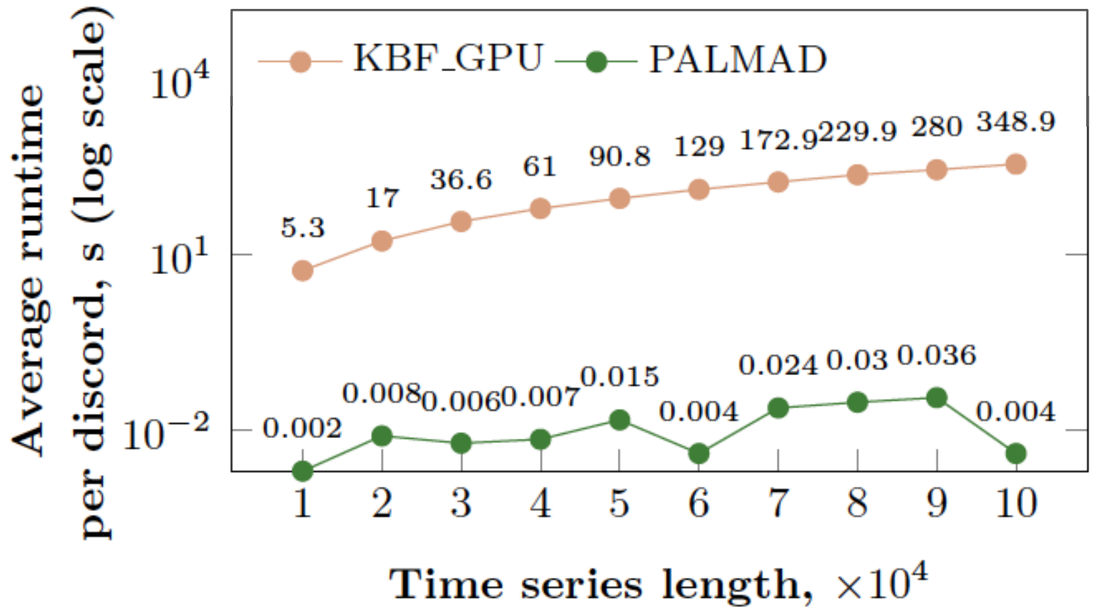
Характеристика	GPU-МГУ	GPU-ЮУрГУ
Производитель, семейство	NVIDIA Tesla	
Модель	P100	V100
# CUDA-ядер	3 584	5 120
Тактовая частота, GHz	1.19	1.3
Оперативная память, Gb	16	32
Пик. пр-ть (double), TFLOPS	4	7

[1)] Keogh E., Lin J., Fu A. HOT SAX: Finding the most unusual time series subsequence: Algorithms and applications. Proc. 5th IEEE Int. Conf. Data Mining 2004: 440–449. URL: http://www.cs.ucr.edu/~eamonn/discords/.

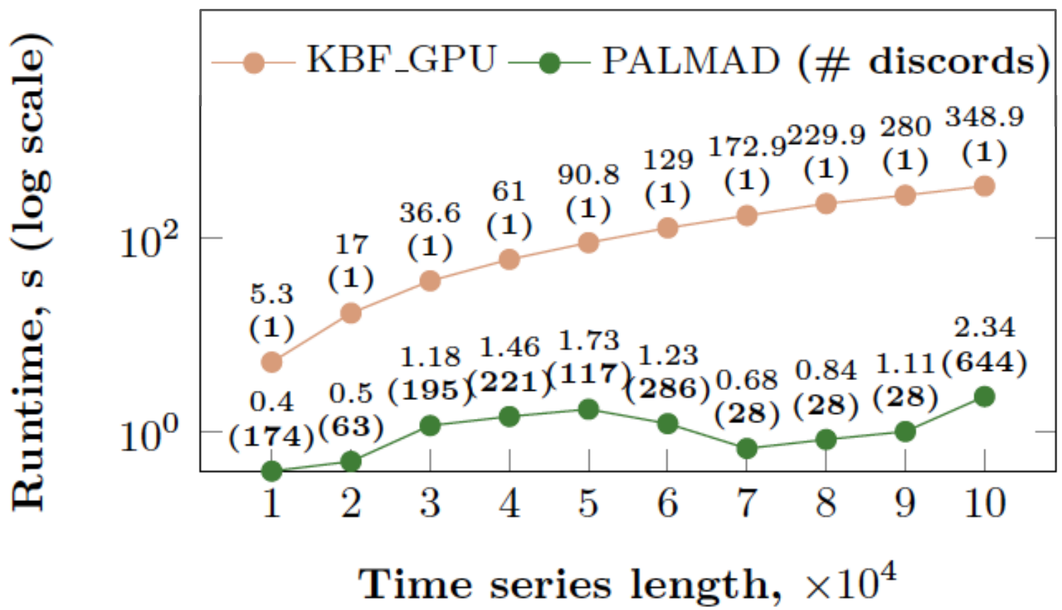
[2)] Pearson K. The problem of the random walk. Nature 72(394). https://doi.org/10.1038/072342a0.

Производительность: сравнение с KBF_GPU[1)]

Среднее время на поиск **одного** диссонанса



Время на поиск **всех** диссонансов

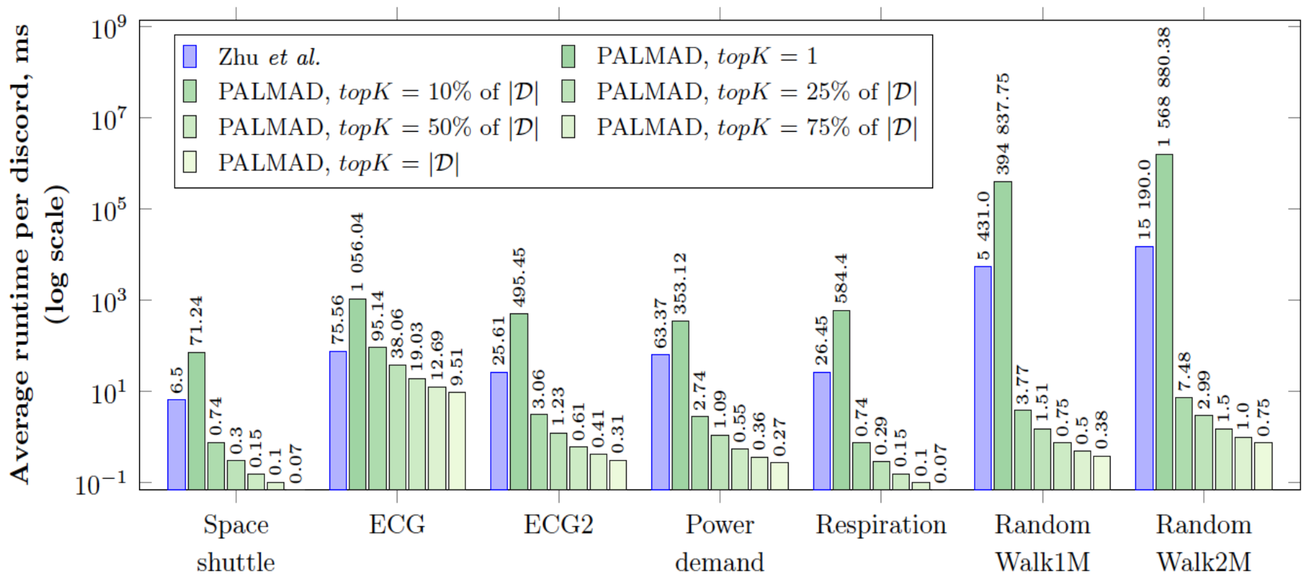


PALMAD опережает KBF_GPU как по общему времени работы,
так и по среднему времени на поиск одного диссонанса

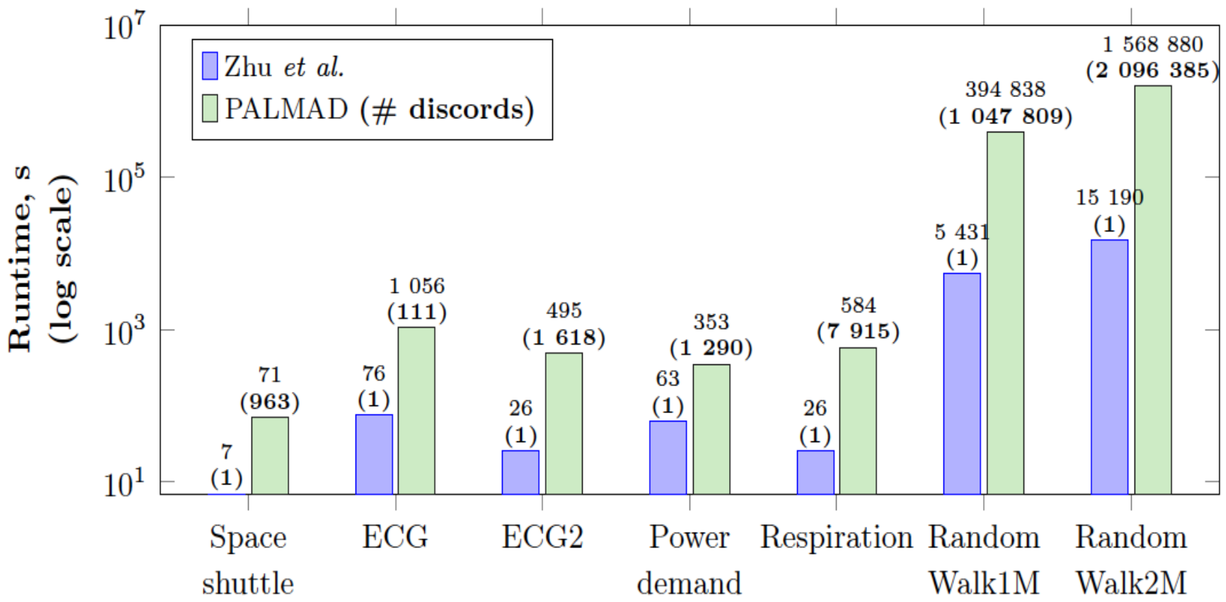
[1)] Thuy T.T.H. et al. A new discord definition and an efficient time series discord detection method using GPUs. ICSED 2021. pp. 63–70. https://doi.org/10.1145/3507473.3507483.

Производительность: сравнение с Zhu et al.[1)]

Среднее время на поиск **одного** диссонанса



Время на поиск **всех** диссонансов

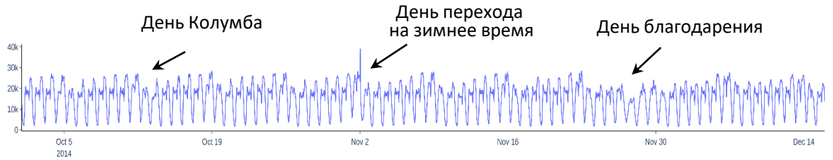


PALMAD значительно опережает алгоритм Zhu et al. по среднему времени на поиск одного диссонанса, начиная с поиска top-k диссонансов, где k=10% от фактического числа диссонансов

[1)] Zhu B. et al. A GPU Acceleration framework for motif and discord based pattern mining. IEEE Transactions on Parallel and Distributed Systems 32(8): 1987-2004. 2021. https://doi.org/10.1109/TPDS.2021.3055765.

Тепловая карта диссонансов

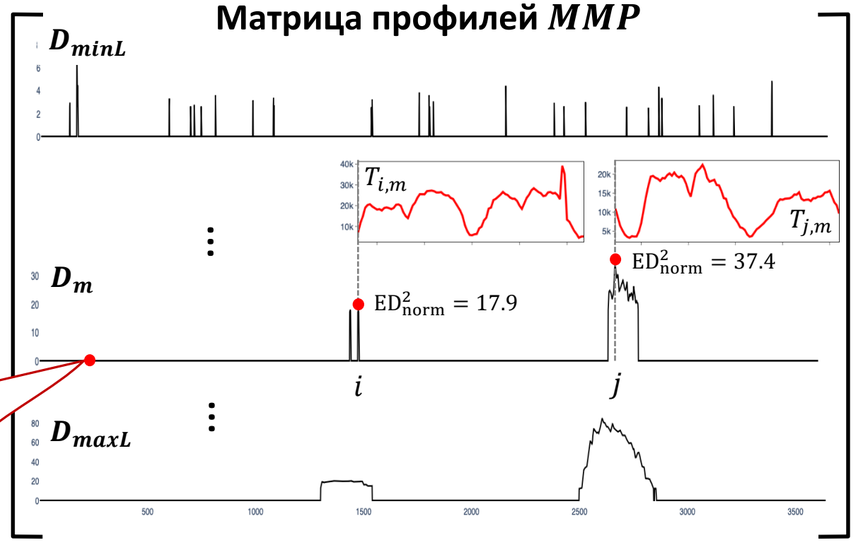
Исходный ряд



PALMAD

Автомат. поиск диссонансов

Матрица профилей MMP

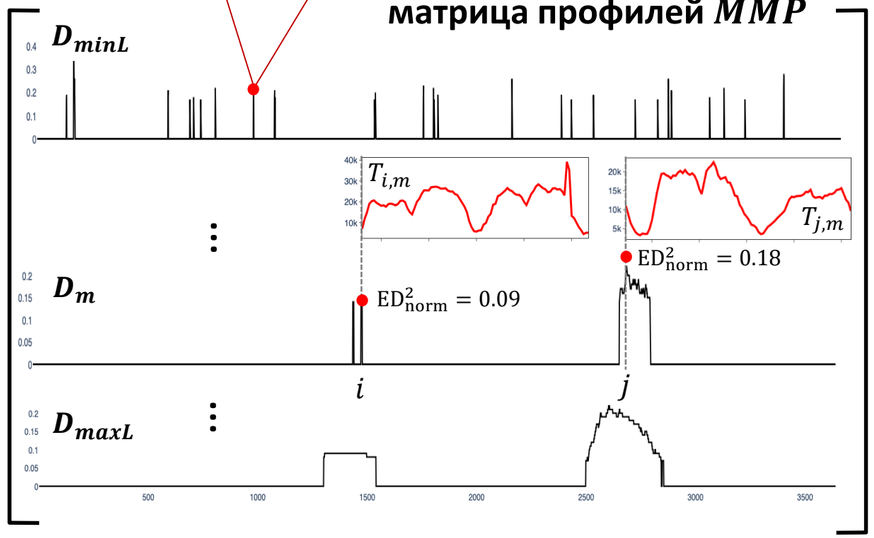


$$MMP(m,i) = \begin{cases} \mathrm{ED}^2_{\mathrm{norm}}(T_{i,m}, NN), T_{i,m} \in D_m \\ 0, otherwise \end{cases}$$

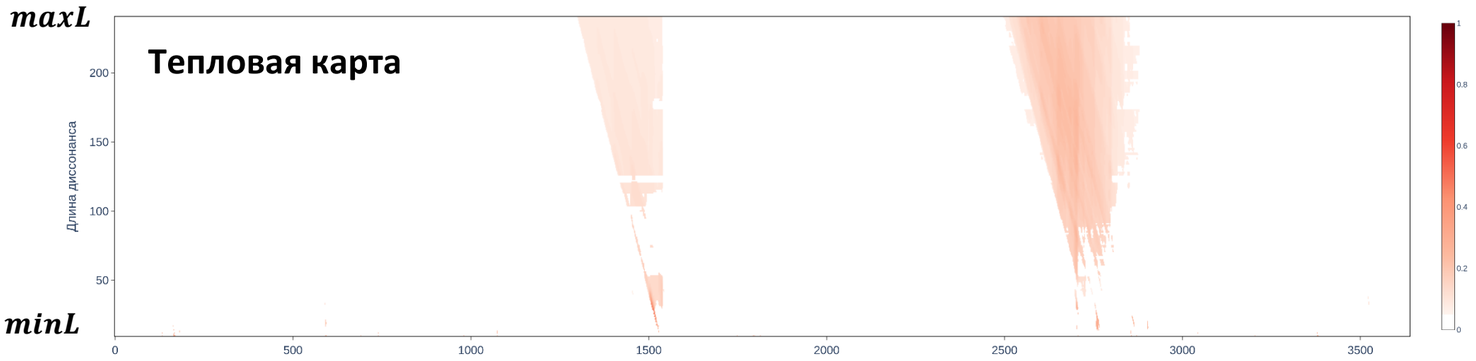
Норм ализа ция

$$MMP(m,i) = \frac{MMP(m,i)}{2m}$$

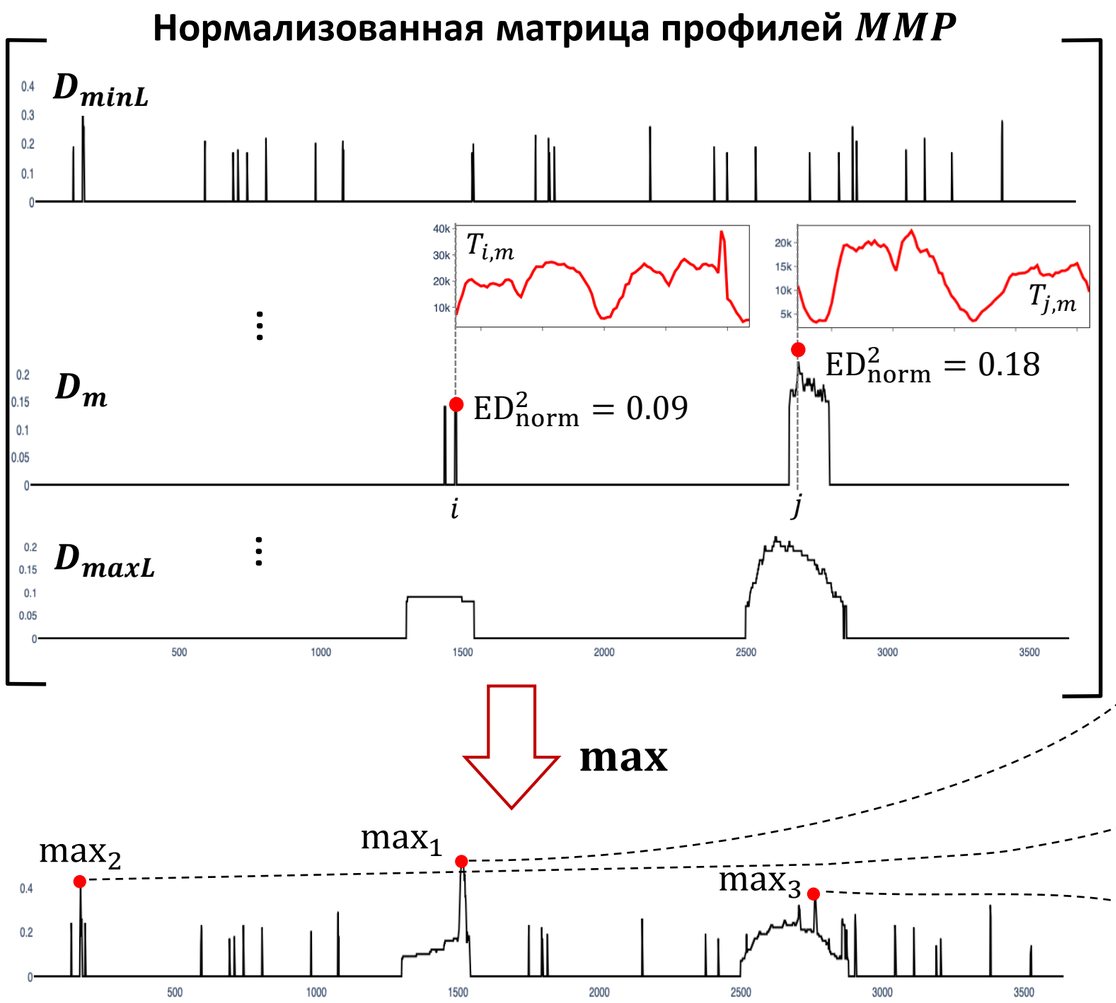
Нормализованная матрица профилей MMP

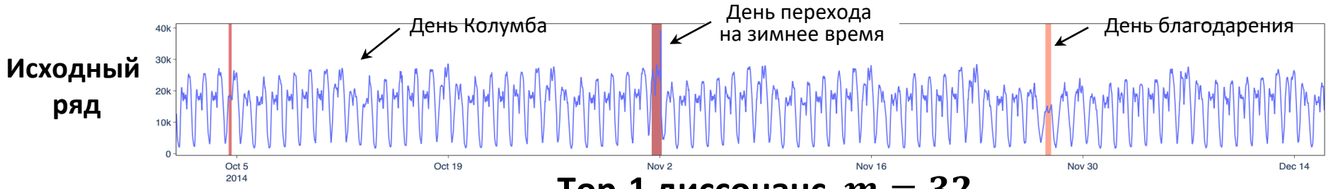


Визуализация (расстояние ⟶ яркость)



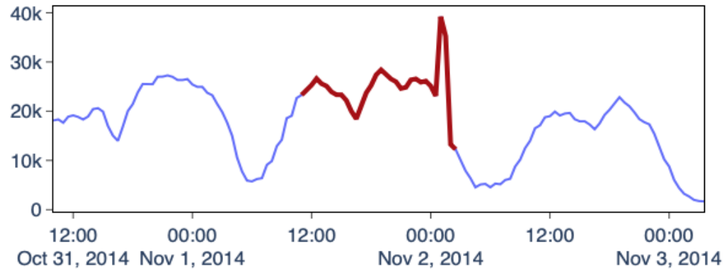
Ранжирование диссонансов





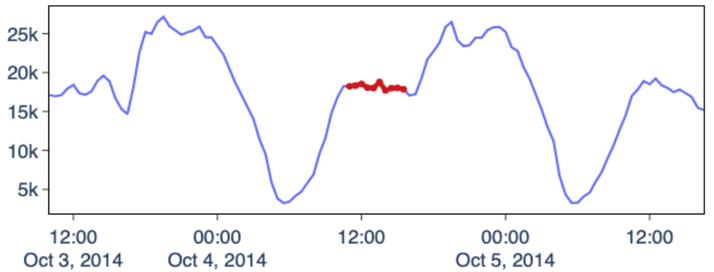
Top-1 диссонанс, $m = 32$

$\mathrm{ED}^2_{\mathrm{norm}} = 0.52$



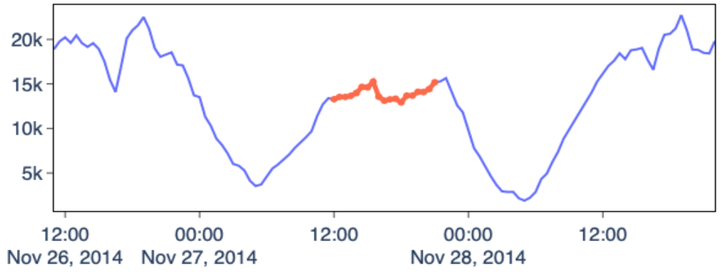
Top-2 диссонанс, $m = 10$

$\mathrm{ED}^2_{\mathrm{norm}} = 0.43$

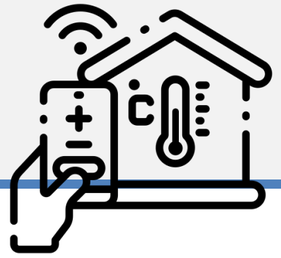


Top-3 диссонанс, $m = 19$

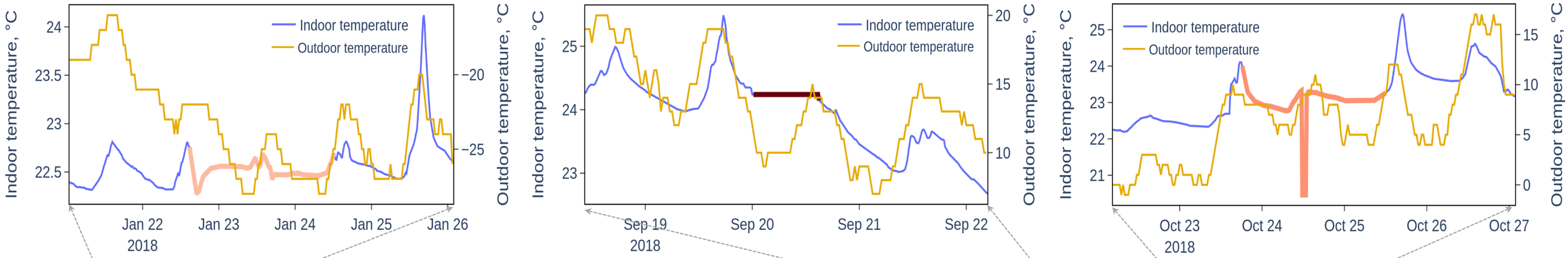
$\mathrm{ED}^2_{\mathrm{norm}} = 0.37$



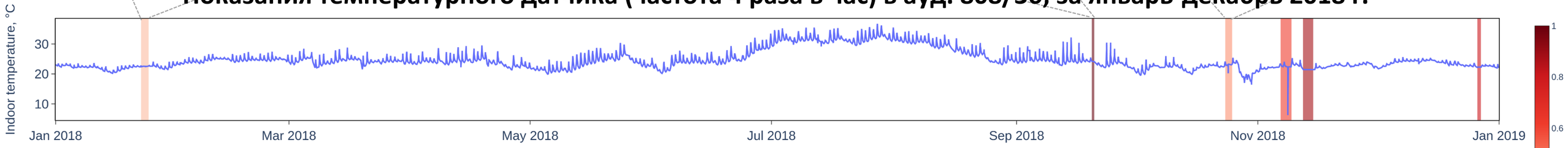
Выявление аномалий в системе отопления ЮУрГУ



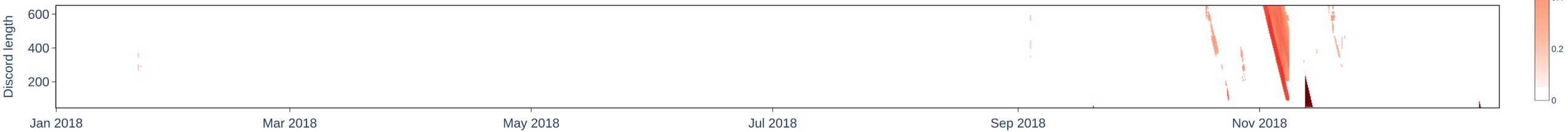
Примеры найденных аномалий длительностью 0.5-2 суток



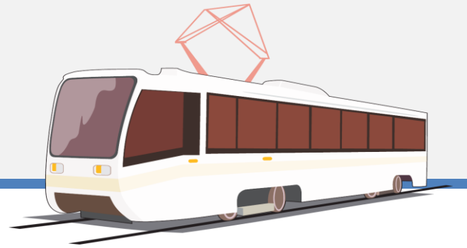
Показания температурного датчика (частота 4 раза в час) в ауд. 808/3б, за январь-декабрь 2018 г.



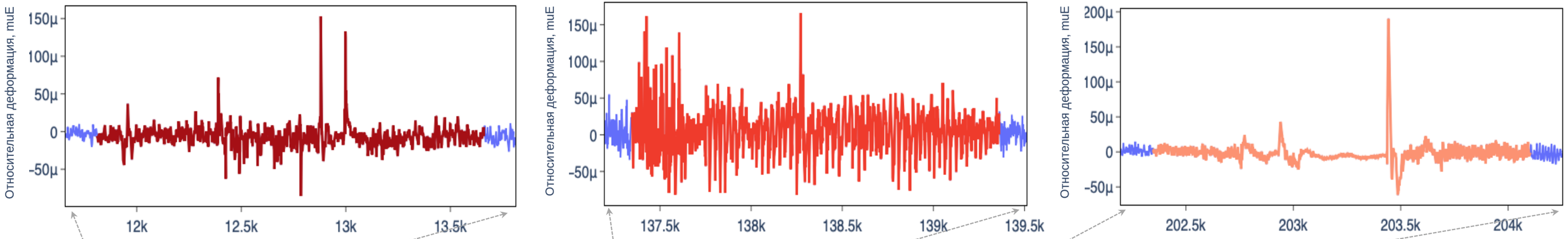
Тепловая карта найденных аномалий длительностью 0.5-2 суток



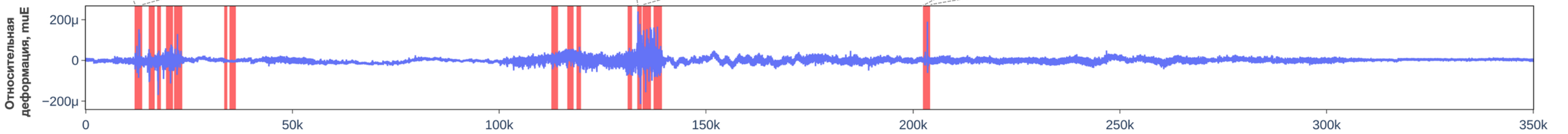
Выявление аномалий в машиностроении



Примеры найденных аномалий длительностью 0.25-0.5 секунд



Относительные деформации механизма стыковки вагонов трамвая (частота 4 096 раз в сек) за 1.5 минуты



Тепловая карта найденных аномалий длительностью 0.25-0.5 секунд

Заключение

- Предложен новый параллельный алгоритм поиска аномалий временного ряда PALMAD для GPU
- Будущие исследования:
  - Разработка версии PALMAD для кластера с GPU-узлами
  - Применение PALMAD в нейросетевой модели для поиска аномалий временного ряда в режиме реального времени

Спасибо за внимание! Вопросы?

Яна Александровна Краева
kraevaya@susu.ru