Обнаружение разладки во временных рядах показов мобильной рекламы

Мерзляков Климент Викторович, группа 622

Санкт-Петербургский Государственный Университет
Кафедра статистического моделирования

Научный руководитель — кандидат физико-математических наук, доцент Голяндина Нина Эдуардовна
Рецензент — кандидат физико-математических наук Пепёлышев Андрей Николаевич

2019г.

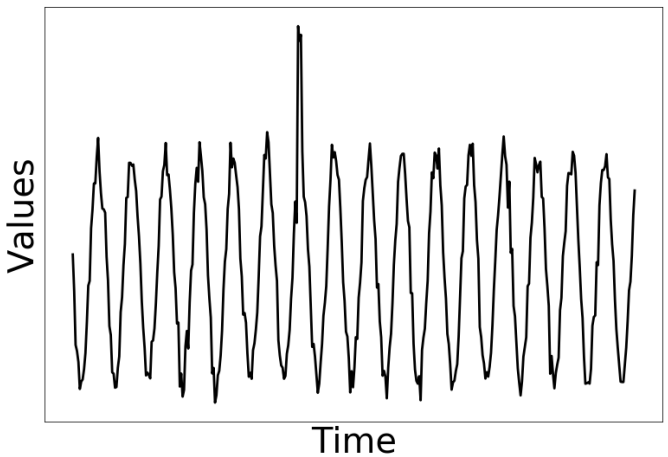
Содержание

- Общие замечания
- Моделирование данных
- Методы обнаружения разладки
- Оценка качества

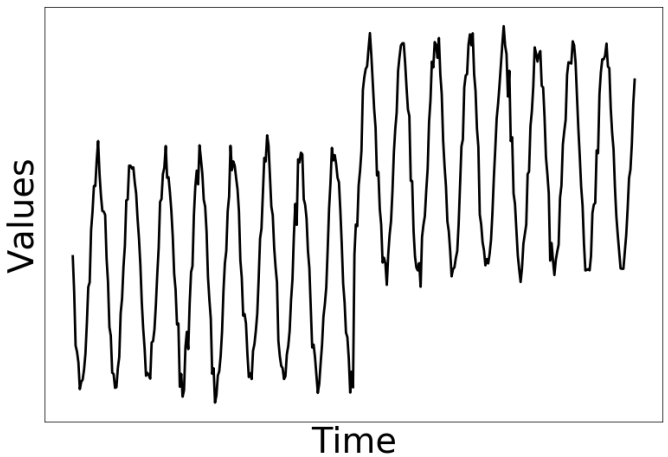
Общие замечания. Разладка во временных рядах

- Разладкой во временных рядах называют момент времени, в который произошло существенное изменение в структуре временного ряда
- Разладка может быть двух типов
  - Локальная — аномалия или выброс
  - Глобальная — изменение структуры ряда

Локальная разладка

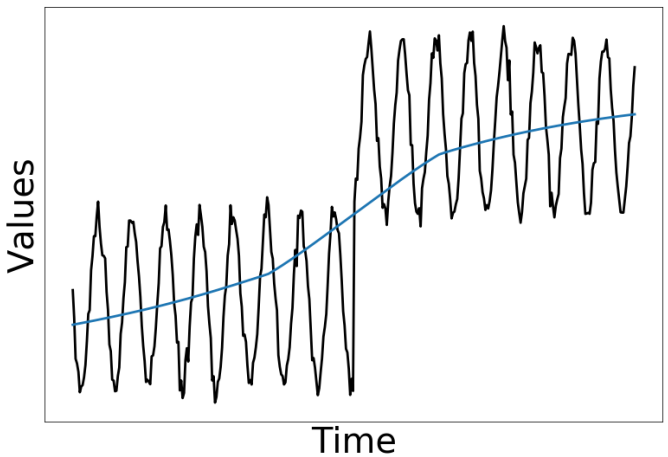


Глобальная разладка

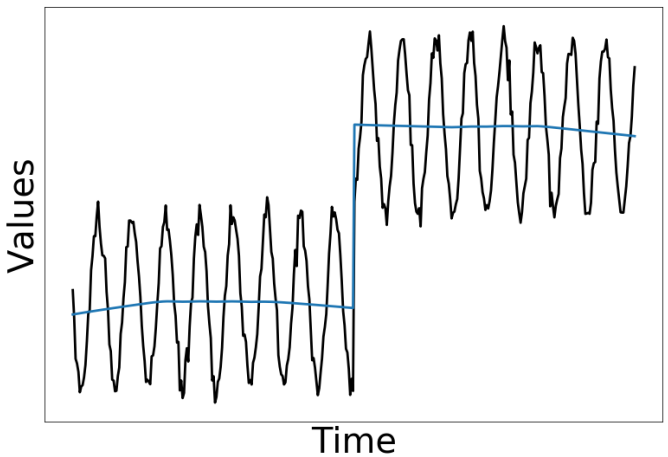


- Исторические данные
  - Прогнозирование
  - Извлечение тренда
  - Поиск проблем в исторических данных
- Текущие данные
  - Реакция на изменения своевременно

Извлечение тренда без анализа разладок

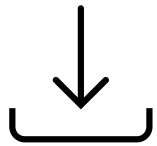


Извлечение тренда с анализом разладок

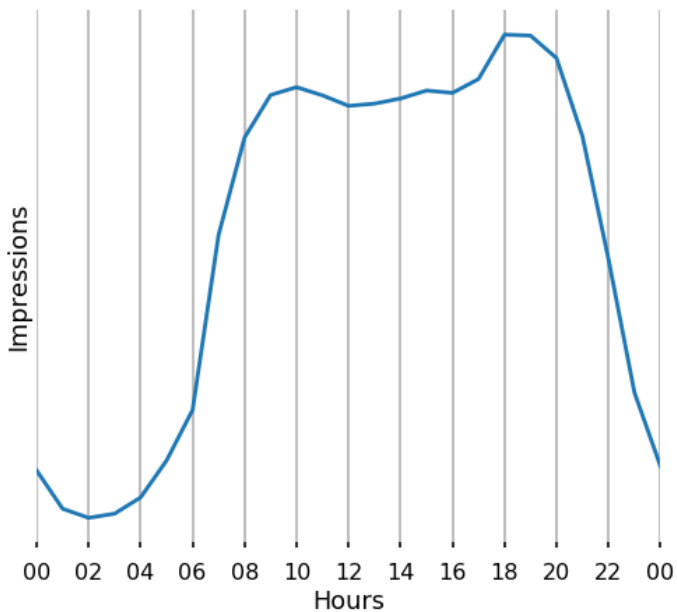


Запрос > Показ > Клик > Конверсия

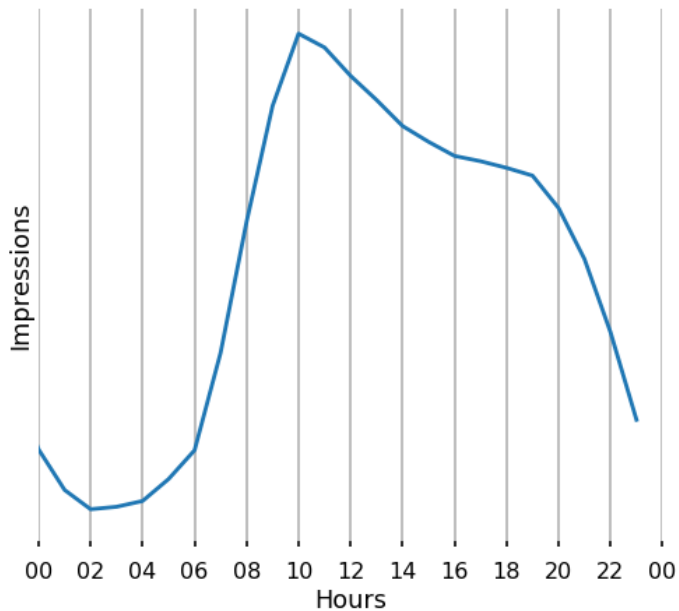


Типичный будний день



Типичный выходной день



Модель ряда $y_i = t_i + s_i + \epsilon_i$ можно задать следующим образом:

$$t_i = c, \quad i = 1, \ldots, n,$$

$$s_i = \sum_{j=1}^{J} A_j \cos\left(\frac{2\pi}{a_j} i + \phi_j\right), \quad i = 1, \ldots, n,$$

$$\epsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n,$$

где i индекс элемента ряда; j индекс косинуса в периодической компоненте; J — количество косинусов в периодической компоненте; c — константа; A_j — амплитуда j-го косинуса; a_j — период j-го косинуса; ϕ_j — фаза j-го косинуса.

τ — точка (индекс) разладки, тогда тренд с разладкой $\tilde{T} = (\tilde{t_1}, \ldots, \tilde{t_n})$.

Разладка — изменение в среднем

$$\tilde{t_i} = \begin{cases} t_i, & i < \tau, \\ t_i + \delta^{(mean)}, & i \geqslant \tau, \end{cases}$$

$$\delta^{(mean)} = \max(\delta^{(mean)*}, \delta^{(mean)}_{min}),$$

$$\delta^{(mean)*} \sim N(\mu^{(\text{cp\_mean})}, \sigma^{2(\text{cp\_mean})}).$$

Локальная разладка

$$\tilde{t_i} = \begin{cases} t_i, & i \neq \tau, \\ t_i + \delta^{(local)}, & i = \tau, \end{cases}$$

$$\delta^{(local)} = \max(\delta^{(local)*}, \delta^{(local)}_{min}),$$

$$\delta^{(local)*} \sim N(\mu^{(\text{cp\_local})}, \sigma^{2(\text{cp\_local})}).$$

Период исходного ряда — 24. Поэтому моделировать ряд будем как сумму тренда, периодики и шума. Применим к реальному ряду метод SSA (singular sprectrum analysis) с окном 96. И оценим параметры периодичности по 8 компонентам с номерами 3-10.

Период	Фаза	Амплитуда
$23.9 \approx 24$	$2.78 \approx 8\pi/9$	1.00
$11.9 \approx 12$	$1.55 \approx \pi/2$	0.39
$7.9 \approx 8$	$-1.56 \approx -\pi/2$	0.13
$5.9 \approx 6$	$-2.95 \approx -15\pi/16$	0.11

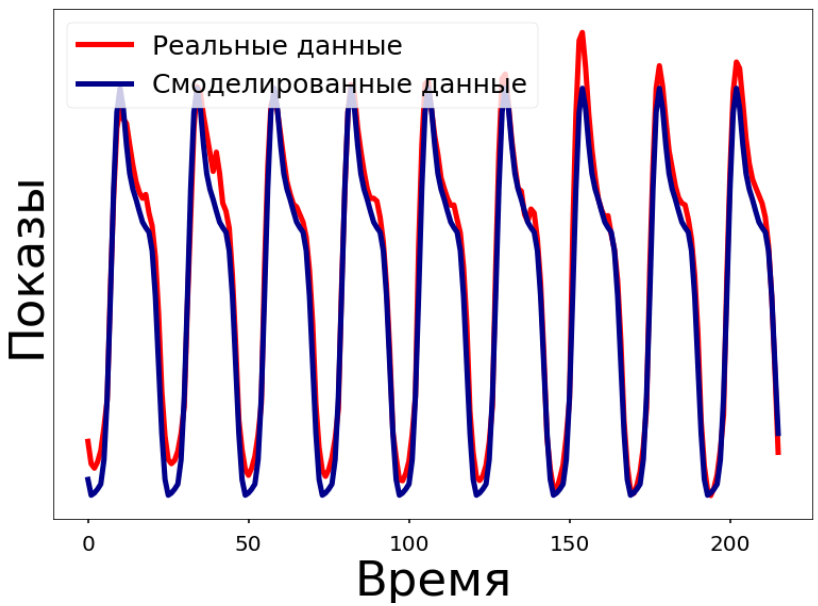
Таким образом, модель периодической составляющей s_i нашего ряда можно записать в следующем виде:

$$s_i = \cos(\frac{2\pi}{24}i + \frac{8\pi}{9}) + 0.39\cos(\frac{2\pi}{12}i + \frac{\pi}{2}) + 0.13\cos(\frac{2\pi}{8}i - \frac{\pi}{2}) + 0.11\cos(\frac{2\pi}{6}i - \frac{15\pi}{16}),$$

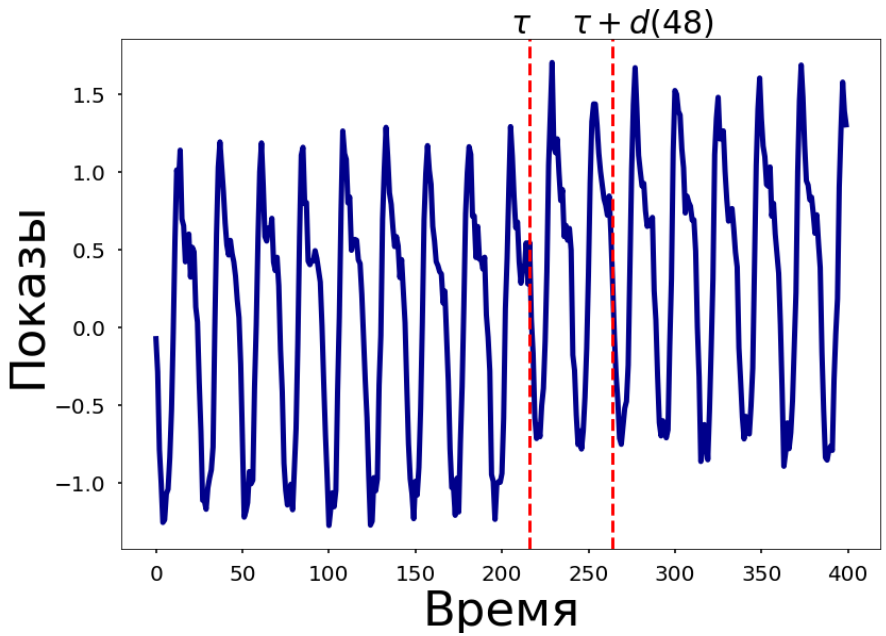
$$i = 1, \ldots, n.$$

- Длину ряда зафиксируем $n = 400$
- Значение тренда (до разладки) выберем нулевым: $c = 0$, то есть $t_i = 0, i = 1, \cdots, n$
- Параметры шума возьмем $\mu = 0, \sigma = 0.1$

Модель сигнала (без шума) в сравнении с реальным рядом



Сгенерированный ряд с шумом и разладкой



- Сигнал ряда может быть описан моделью
- Идея подхода: около точки разладки модель хуже описывает сигнал
- Используя меру ошибки можно измерить насколько плохо

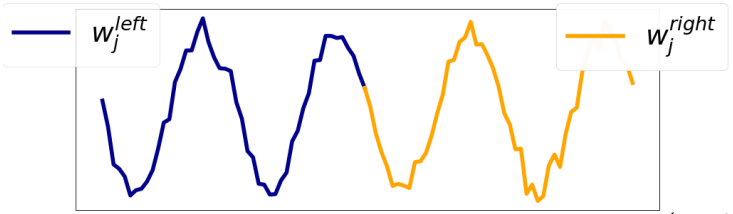
Исходная модель ряда одна:

$$y_i = t_i + s_i + \epsilon_i = c + \sum_{j=1}^{J} A_j \cos\left(\frac{2\pi}{a_j} i + \phi_j\right) + \epsilon_i, \quad i = 1, \ldots, n.$$

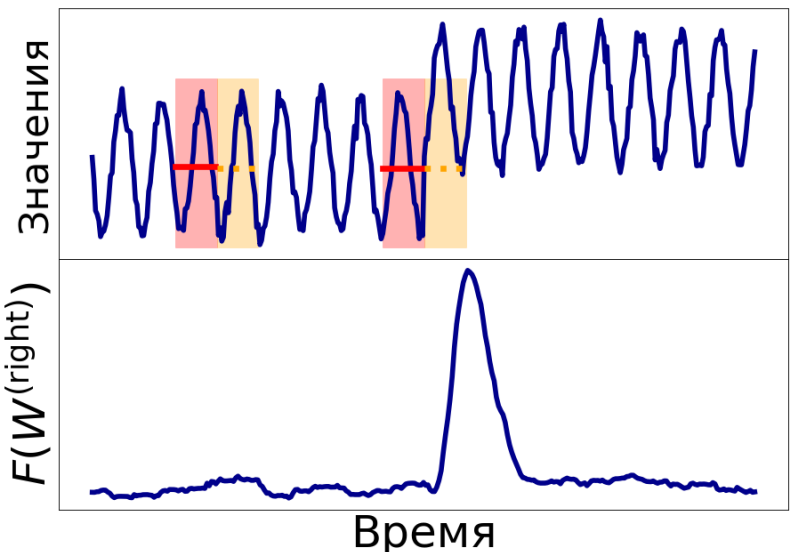
При этом моделей сигнала $f(x|\theta)$ в основе методов разладки много.

1. **„Среднее“**
$$f(x|b) = b,$$
2. **„Косинус“**
$$f(x|P, p, \chi, b) = P\cos(\frac{2\pi}{p}x + \chi) + b,$$
3. **„4 косинуса“**
$$f(x|\{P_j, p_j, \chi_j\}, b) = \sum_{j=1}^{J} P_j \cos(\frac{2\pi}{p_j}x + \chi_j) + b,$$

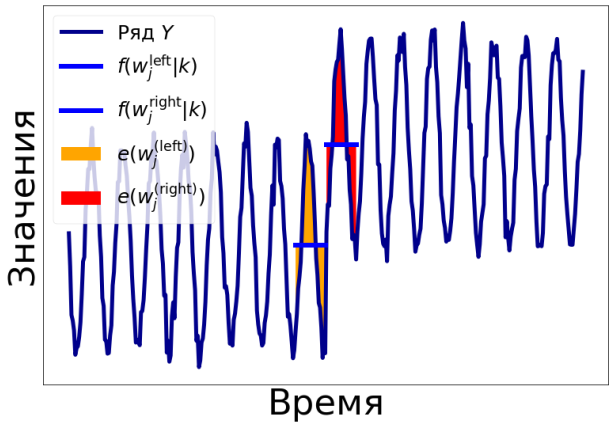
- Взяв ряд Y, мы «скользим» по нему окном ширины l
- Полученные отрезки делим на 2 части: $w_j^{(left)}$ и $w_j^{(right)}$

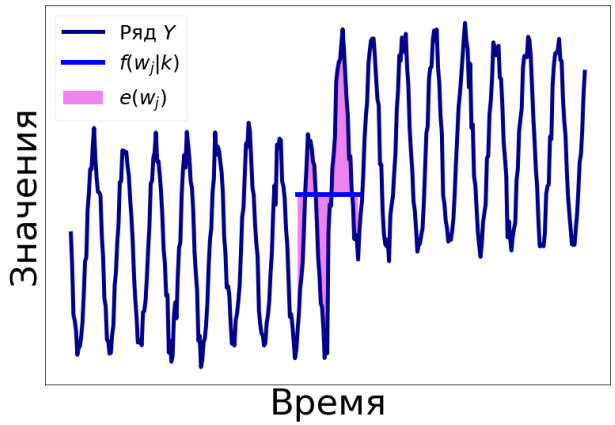


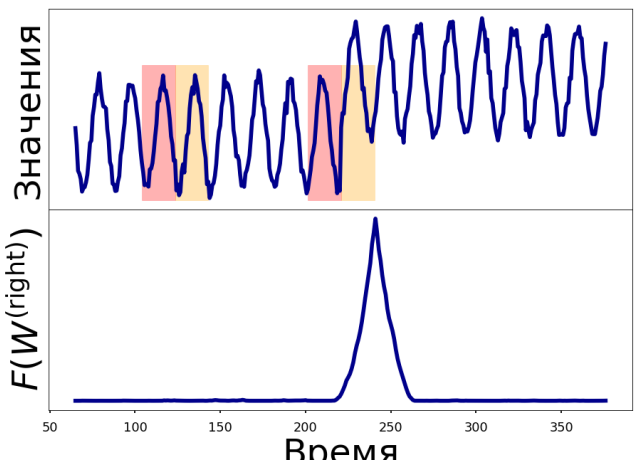
- Считаем параметры модели $\theta = \underset{\theta}{\operatorname{argmin}} \sum_{j=1}^{J} (w_j^{(left)} - f(w_j^{(left)}|\theta))^2$
- Делаем прогноз на правые отрезки и рассчитываем значения функции разладки $F(w_j^{(right)}) = e(w_j^{(right)}) = \sum_{j=1}^{J} (w_j^{(right)} - f(w_j^{(right)}|\theta))^2$,
- Функция разладки начинает расти в окрестности точки разладки τ,
- Разладка обнаружена при превышении $F()$ порога γ в точке $\hat{\tau}$



Ключевое отличие от методов прогнозирования: мы рассчитываем меры ошибки на тех же рядах на которых подбирались параметры модели θ. А функция разладки выглядит $F_j = e(w_j^{(all)}) - e(w_j^{(left)}) - e(w_j^{(right)})$



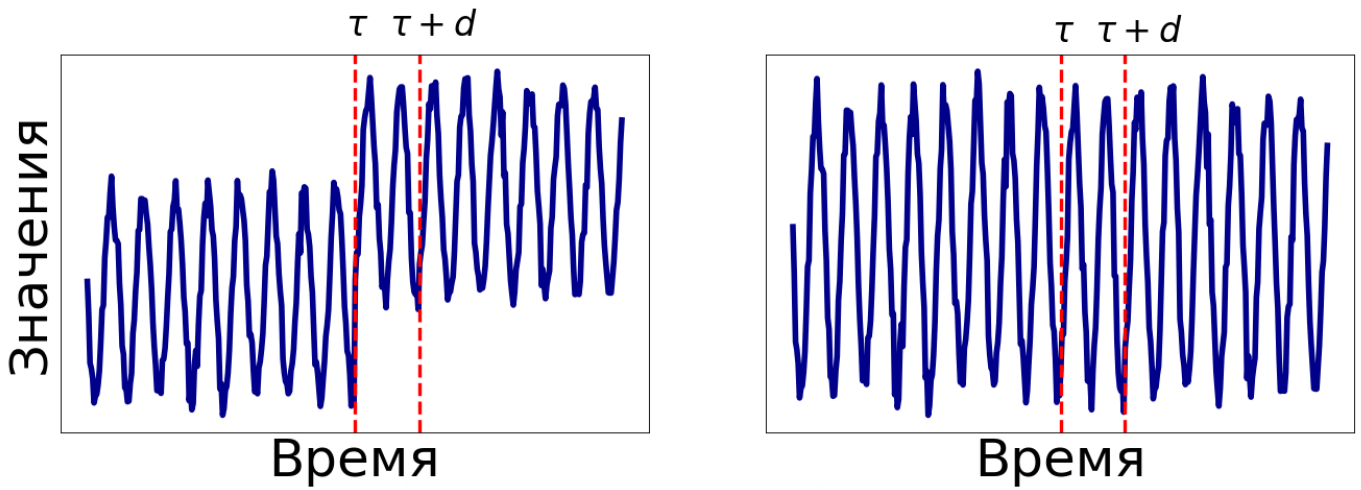




Оценка качества. Матрица ошибок классификации

Задан порог γ для функции разладки F_j. Фиксируем точку разладки τ и приемлемую задержку d. За пределами приемлемой задержки нас не интересует что происходит с рядом. Классифицирующее правило:

$$a(Y) = \begin{cases} 1, & \exists j : F_j \geq \gamma \text{ и } j \text{ лежит в отрезке } (\tau, \tau + d) \\ 0, & \text{иначе.} \end{cases}$$



Факт \ Метод	0	1
0	TN	FN
1	FP	TP

Результаты применения методов к смоделированным данным (ROC-AUC)

Тип разладки	локальная														
Задержка (d)	4					24					48				
Длина окна (l)	2	4	24	48	96	2	4	24	48	96	2	4	24	48	96
Место разладки	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216
Метод															
Аппроксимация. Среднее	0,98	0,76	0,50	0,77	0,67	0,95	0,64	0,56	0,83	0,71	0,92	0,67	0,63	0,86	0,71
Аппроксимация. Косинус				0,93	0,82				0,95	0,81				0,94	0,85
Аппроксимация. 4 косинуса				0,97	0,88				0,94	0,91				0,98	0,92
Прогноз. Среднее	0,95	0,80	0,59	0,69	0,70	0,93	0,66	0,57	0,82	0,71	0,93	0,67	0,61	0,75	0,73
Прогноз. Косинус				0,87	0,77				0,94	0,92				0,90	0,76
Прогноз. 4 косинуса				1,00	0,98				0,98	0,99				0,99	0,98

Тип разладки	в среднем														
Задержка (d)	4					24					48				
Длина окна (l)	2	4	24	48	96	2	4	24	48	96	2	4	24	48	96
Место разладки	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216
Метод															
Аппроксимация. Среднее	0,67	0,58	0,52	0,82	0,72	0,56	0,54	0,92	1,00	1,00	0,60	0,53	0,92	1,00	1,00
Аппроксимация. Косинус				0,95	0,85				1,00	1,00				1,00	1,00
Аппроксимация. 4 косинуса				0,99	0,86				1,00	1,00				1,00	1,00
Прогноз. Среднее	0,73	0,58	0,54	0,67	0,59	0,51	0,47	0,86	0,99	0,94	0,55	0,48	0,81	0,96	0,98
Прогноз. Косинус				0,70	0,65				1,00	0,99				1,00	1,00
Прогноз. 4 косинуса				0,99	0,96				1,00	1,00				1,00	1,00

Оценка качества. Выводы

- Оба подхода (аппроксимация и прогнозирование) работают примерно одинаково для условий нашей задачи;
- Чем более точно модель описывает данные, тем лучшее качество она показывает;
- Не кратные двум периодам окна работают хуже (потому что оценка происходит по неполной периодичности);
- Очень маленькие окна работают достаточно хорошо, но только для локальной разладки;
- Задержка для маленьких окон ухудшает качество метода при росте задержки. Но для больших окон эффект обратный — наблюдается улучшение при росте задержки;

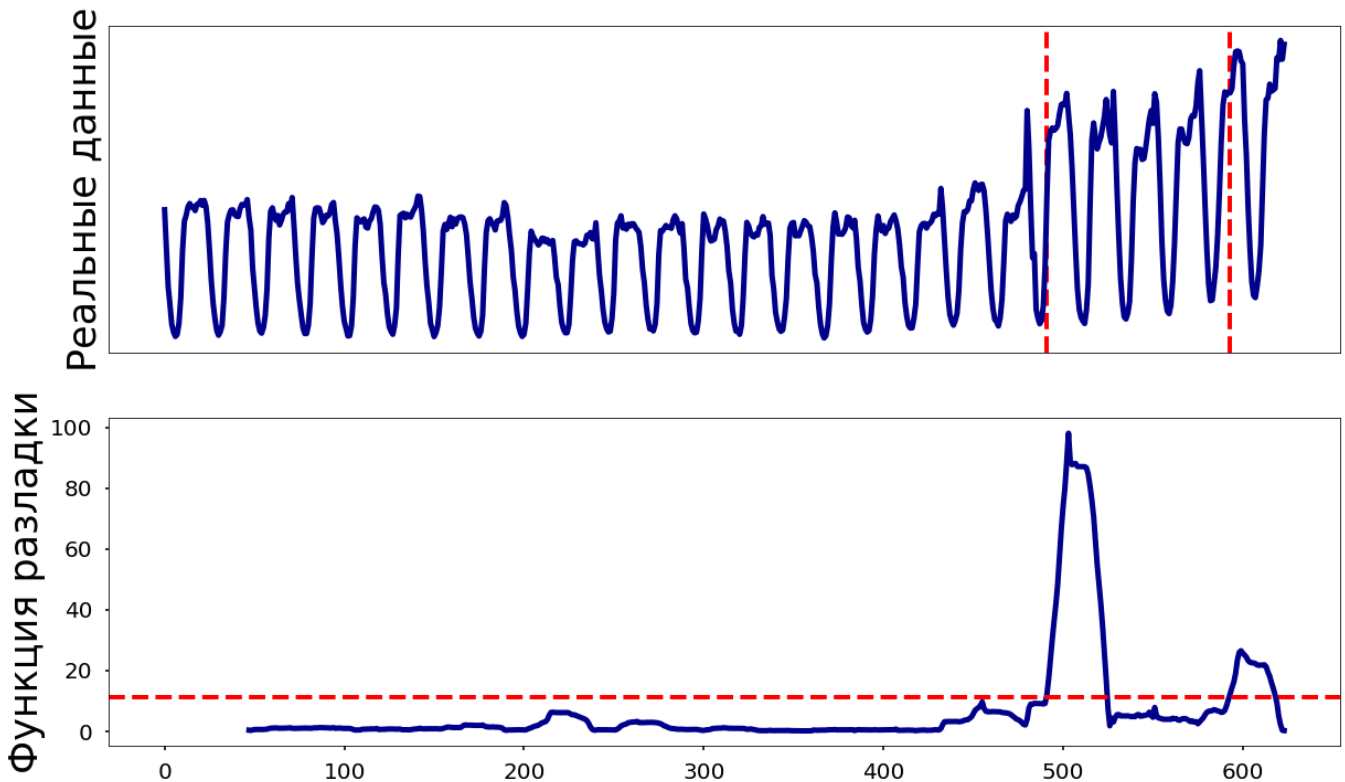
Оценка качества. Выбор порога γ

- Посчитать функцию разладки для начала ряда
- Построить выборку из максимумов скользящих отрезков длины d (d — величина допустимой задержки)
- В качестве порога γ взять q-квантиль из полученной выборки
- По построению, такой выбор порога будет фиксировать ошибку $FPR = 1 - q$

0.75 квантиль

		Классификатор 0	Классификатор 1	
Факт	0	366	134	FPR: 0,27
	1	31	469	TPR: 0,94

Реальные почасовые данные показов мобильной рекламы за 26 дней



Заключение

В рамках исследования мы:

- Создали модель, по которой можно создавать временные ряды близкие к реальным
- Сравнили разные подходы к обнаружению разладок
- Вывели рекомендации по выбору параметров, моделей и порога
- Применили наиболее подходящий метод к реальным данным

В качестве дальнейших шагов может быть предпринято следующее:

- Создание более сложной модели временного ряда, учитывающей отличия между будними в выходными днями
- Применение и оценка качества других методов обнаружения разладки