Обнаружение разладки во временных рядах показов мобильной рекламы

К.В. Мерзляков, группа 622

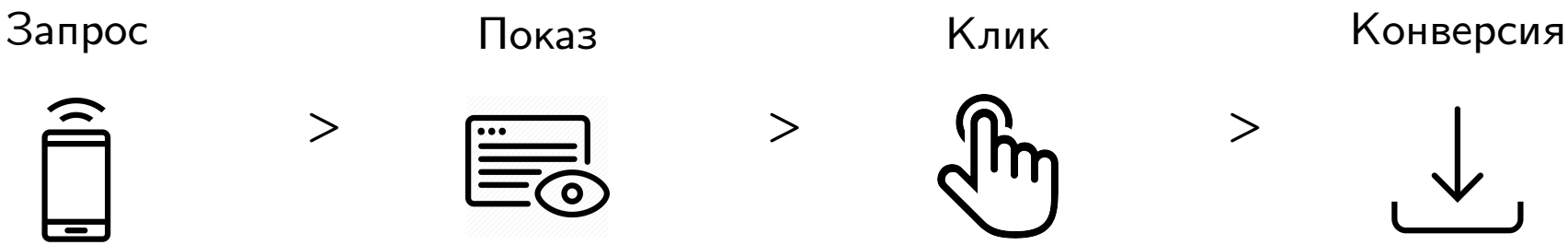
Санкт-Петербургский Государственный Университет
Кафедра статистического моделирования

18.05.2019

Содержание

Общие замечания

Изменения в данных



Изменения на стороне пользователя

- Популярность приложения
- Конкуренция
- Маркетинговая активность приложения
- ...

Изменения на стороне рекламной сети

- Релиз новых функций
- Партнерство с новыми рекламодателями
- Новые способы таргетинга
- ...

Временные ряды

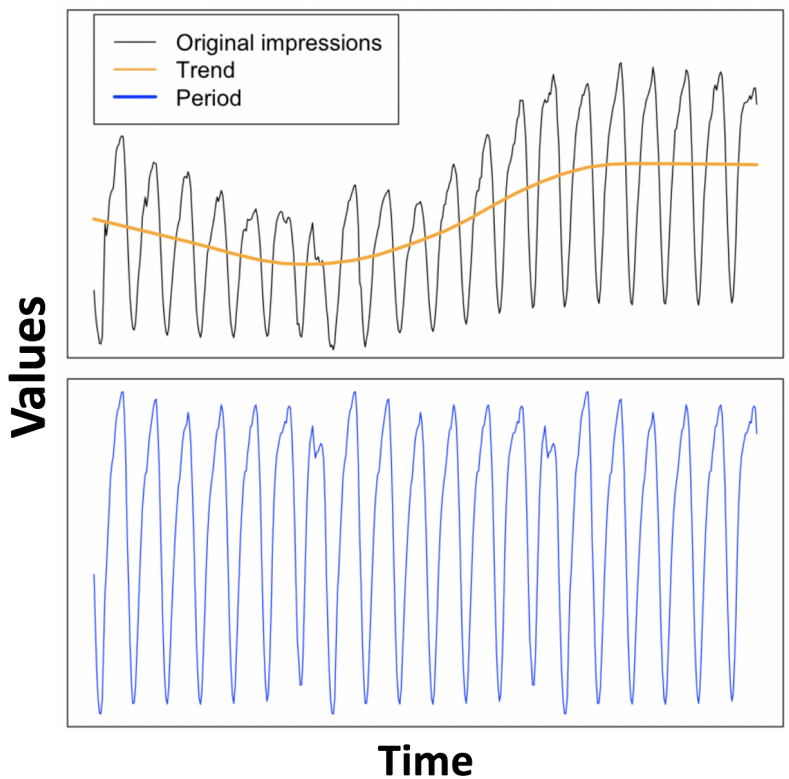
Временной ряд это ряд, состоящий из числовых значений упорядоченных по времени. Как правило, с равными промежутками времени между значениями (минута, час, день, неделя и т.д.)

Time	Data
18-Май-2019 19:00	435 098
18-Май-2019 20:00	431 248
18-Май-2019 21:00	420 329
...	...

Формальное обозначение:
$X = (x_1, x_2, ..., x_{N-1}, x_N)$

Обычно, временной ряд может быть представлен в виде суммы его компонент $X = T + S + E$

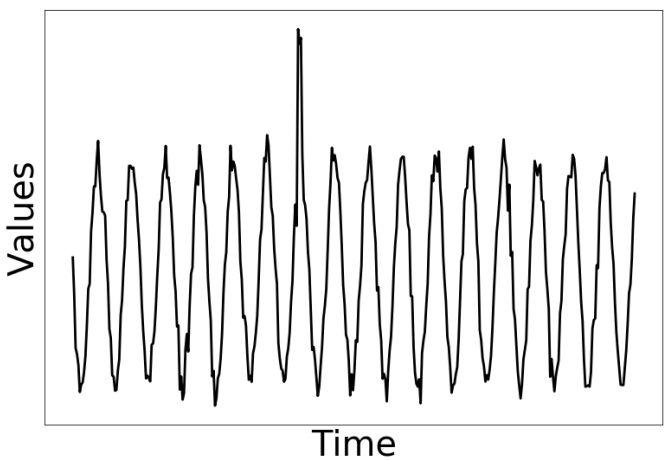
Временной ряд. Пример разложения ряда на компоненты



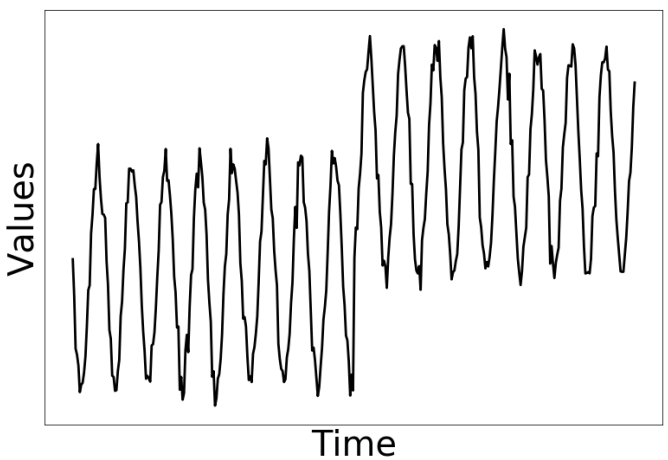
Разладка во временных рядах

- Разладкой во временных рядах называют момент времени, в который произошло существенное изменение в структуре временного ряда
- Методы обнаружения разладки — это группа методов, с помощью которых можно находить такие точки разладки
- Разладка может быть двух типов
  - Локальная — аномалия или выброс
  - Глобальная — изменение структуры ряда

Локальная разладка



Глобальная разладка

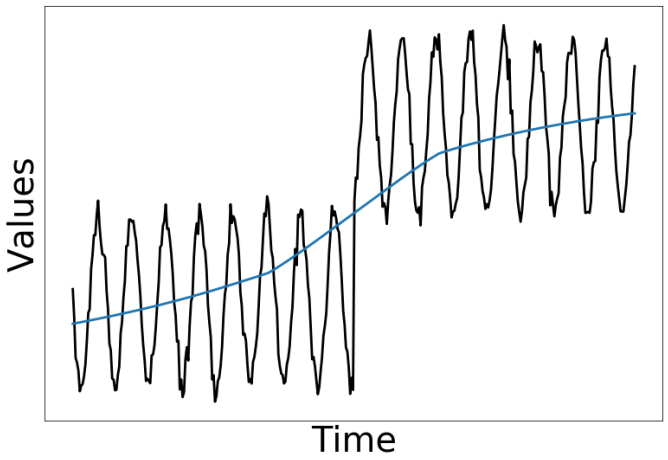


Мотивация

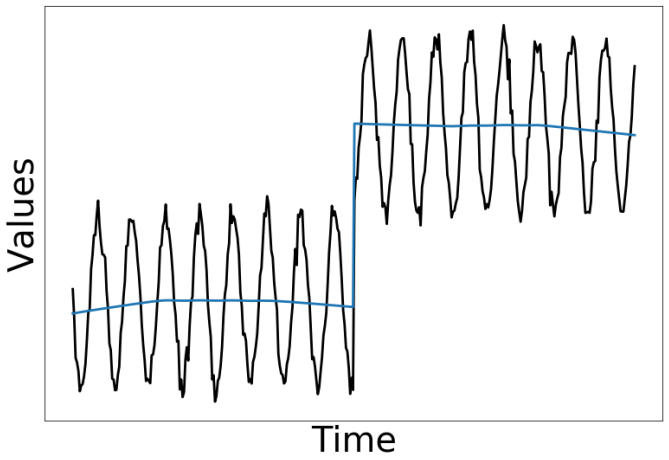
Практическая польза:

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

Извлечение тренда без анализа разладок



Извлечение тренда с анализом разладок

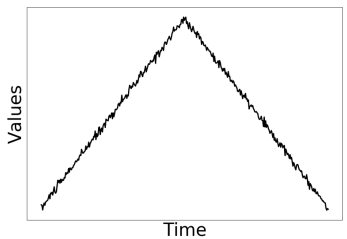


Что можно сделать после обнаружения разладки:

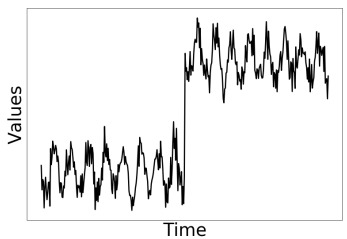
- Убрать/изменить выбросы

Виды разладок

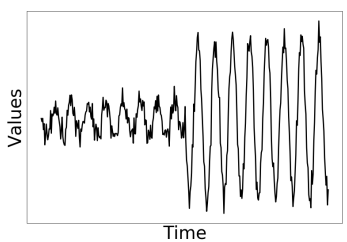
Изменение в тренде



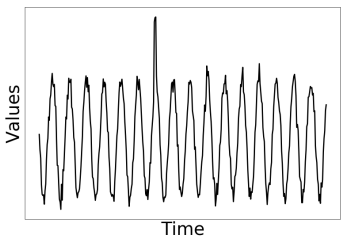
Изменение в среднем



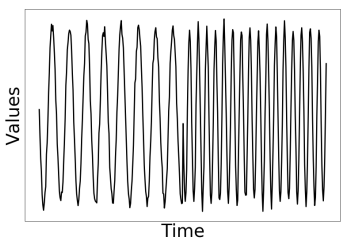
Изменение в амплитуде колебаний



Локальное изменение



Изменение в периодике



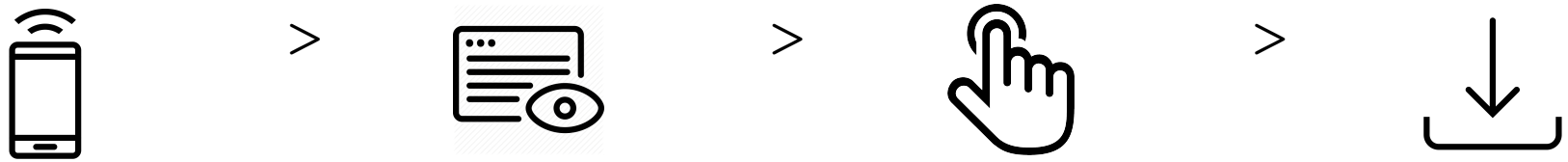
Структура исследования

- Смоделировать данные, близкие к реальным
- Применить к смоделированным данным набор методов
- Оценить и сравнить качество примененных методов

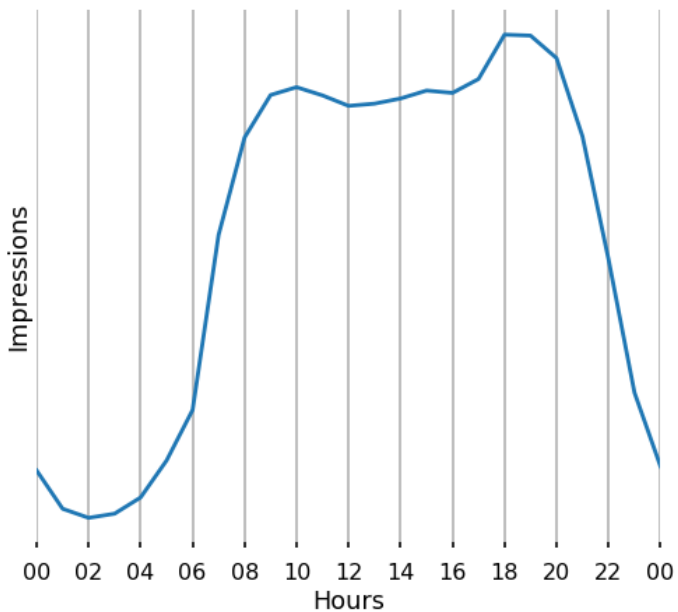
Построение модели данных

Данные мобильной рекламы

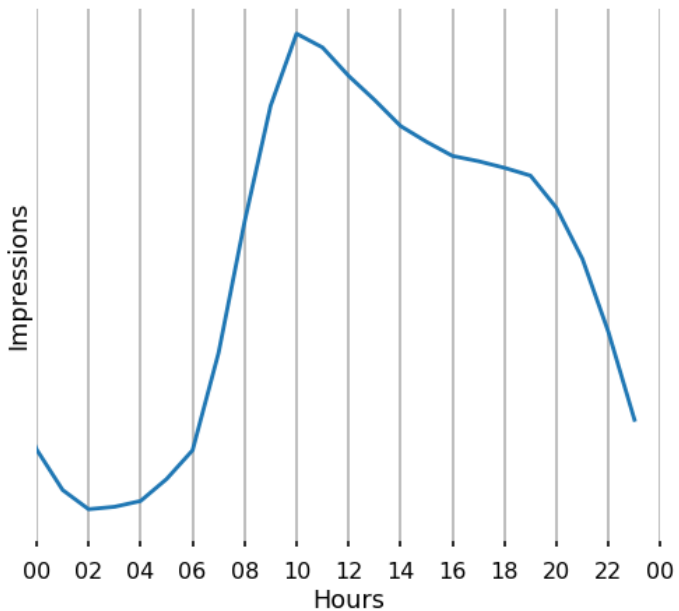
Запрос > Показ > Клик > Конверсия



Типичный будний день

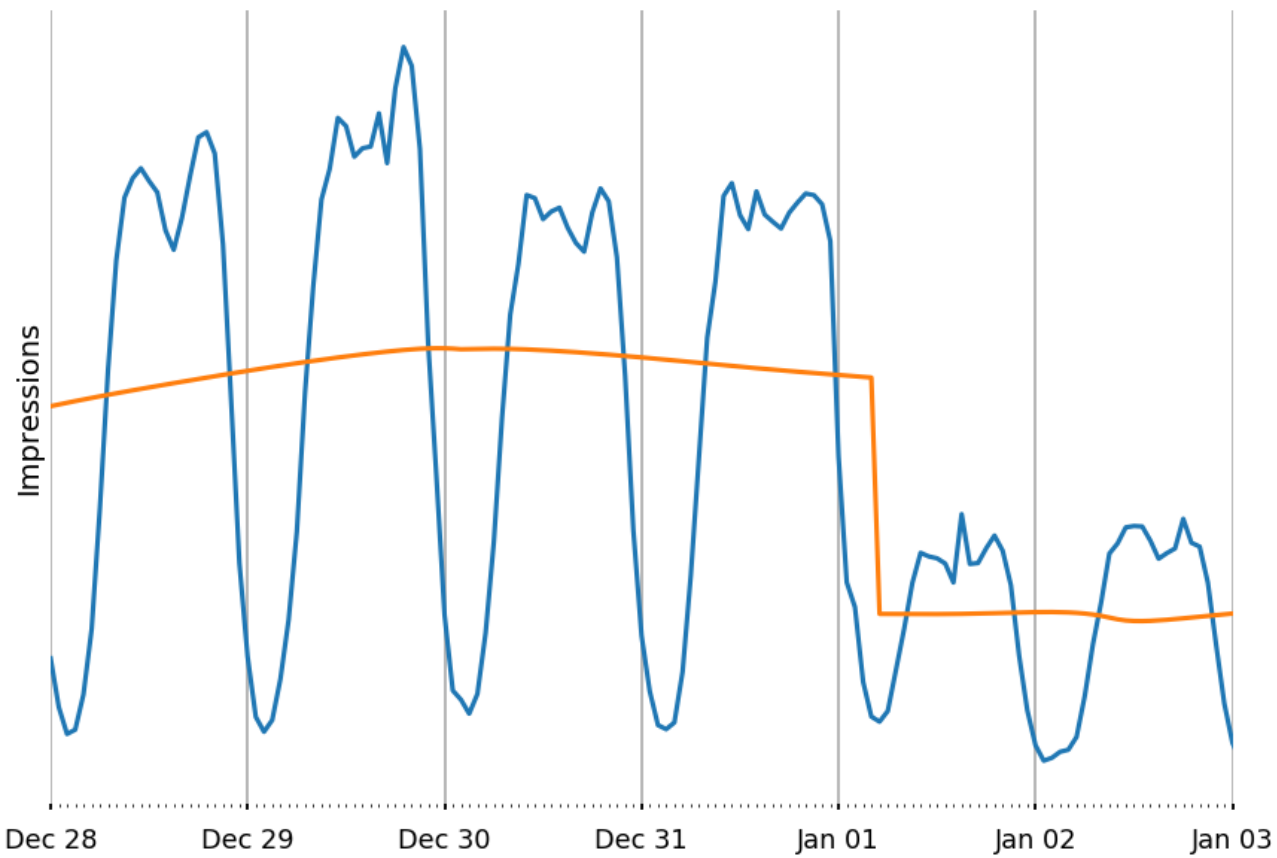


Типичный выходной день



Примеры разладки в реальных данных

Изменение в среднем



Построение модели ряда

- Обозначим временной ряд $Y = (y_1, \ldots, y_n)$
- Наблюдаемые значения можно представить в виде $Y = T + S + E$, где $T = (t_1, \ldots, t_n)$ компонента-тренд, $S = (s_1, \ldots, s_n)$ периодическая компонента, $E = (\epsilon_1, \ldots, \epsilon_n)$ остатки или шум
- Для каждой из этих компонент требуется построить модель

Построение модели ряда

Модель можно задать следующим образом:

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

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

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

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.

Построение модели ряда

Модель можно задать следующим образом:

- Разладка только в одной точке ряда;
- Разладка только в тренде и заключается в сдвиге;
- Разладка может произойти не всегда, а с некоторой вероятностью ρ.

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

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

δ — значение разладки.

Значение разладки является случайной величиной с некоторым распределением. $\delta^* \sim N(\mu^{(\text{ср})}, \sigma^{2(\text{ср})})$:

$$\delta = \begin{cases} \delta^*, & \text{с вероятностью } \rho, \\ 0, & \text{с вероятностью } 1 - \rho. \end{cases}$$

Таким образом, моделируемый ряд с разладкой будет иметь следующий вид:

$$\tilde{Y} = e^{\tilde{T}+S+E}.$$

Методы обнаружения разладки

Общая канва

- У временного ряда есть некоторая структура (сигнал)
- Сигнал может быть описан моделью
- Идея подхода: около точки разладки модель плохо описывает временной ряд
- Используя меру ошибки мы можем измерять насколько хорошо описывает выбранная модель реальные данные
- Как только ошибка (отклонение модели от реальных данных) превышает заданный порог, метод сигнализирует о разладке

Можно выделить два типа методов в данном подходе:

- Методы на основе прогнозирования
- Методы на основе аппроксимации

Аппроксимация

Пусть l — ширина окна. При этом $1 < l < n$, l чётное. С помощью ширины окна из исходного ряда образуется последовательность подрядов $W = \{w_j\}_{j=1}^{k}$, где $k = n - l + 1$ — количество таких подрядов; а $w_j = (y_j, \ldots, y_{j+l-1})$ — j-ый подряд. Каждый подряд w_j в свою очередь делится на два подряда одинаковой длины:
$W^{(\text{left})} = \{w_j^{(\text{left})}\} = \{(y_j, \ldots, y_{j+\frac{l}{2}-1})\}$ и
$W^{(\text{right})} = \{w_j^{(\text{right})}\} = \{(y_{j+\frac{l}{2}}, \ldots, y_{j+l-1})\}$.
Таким образом, для каждого ряда W можно сформировать тройки рядов:

$$W^{(\text{all})} = \{w_j^{(\text{all})}\}_{j=1}^{k} = \{(w_j; w_j^{(\text{left})}; w_j^{(\text{right})})\}_{j=1}^{k}.$$

Аппроксимация

Пусть есть функция ошибки $e(\cdot)$, такая что:

$$e(X) = \min_{\theta} \sum_{p=1}^{m} (x_p - f(x_p|\theta))^2,$$

где $X = (x_1, \ldots, x_m)$ — вещественный временной ряд длины m, а $f(x|\theta)$ — модель сигнала этого временного ряда с параметрами θ.
Функция $f(x|\theta)$ может быть константной ($\theta = (b)$):

$$f(x|b) = b,$$

либо другой подходящей под наш ряд функцией, например:

$$f(x|P, p, \chi) = P \cos(\frac{2\pi}{p} x + \chi) + b.$$

Аппроксимация

Мера ошибки позволяет нам рассчитать, насколько хорошо аппроксимируется отрезок ряда с помощью выбранной модели. Однако, для обнаружения самой разладки необходимо еще ввести функцию разладки:

$$f_j = F(w_j^{(\text{all})}) = \frac{e(w_j) - e(w_j^{(\text{left})}) - e(w_j^{(\text{right})})}{h},$$

где h — значение нормировки, $j = 1, \ldots, k$.
Значения функции разладки синхронизируются с исходным рядом по последнему индексу окна. То есть f_1 соответствует y_l, а f_k соответсвует y_n. Введем синхронизированную функцию разладки :

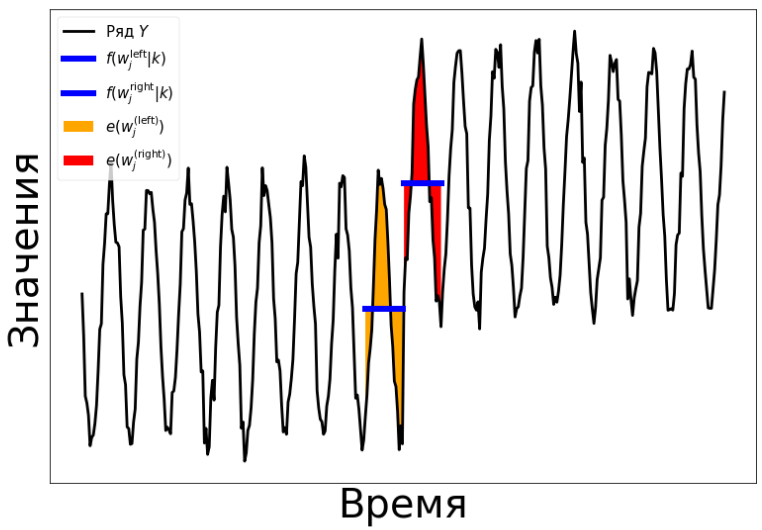
$$q_i = \begin{cases} f_{i-l+1}, & i \geq l, \\ 0, & i < l. \end{cases}$$

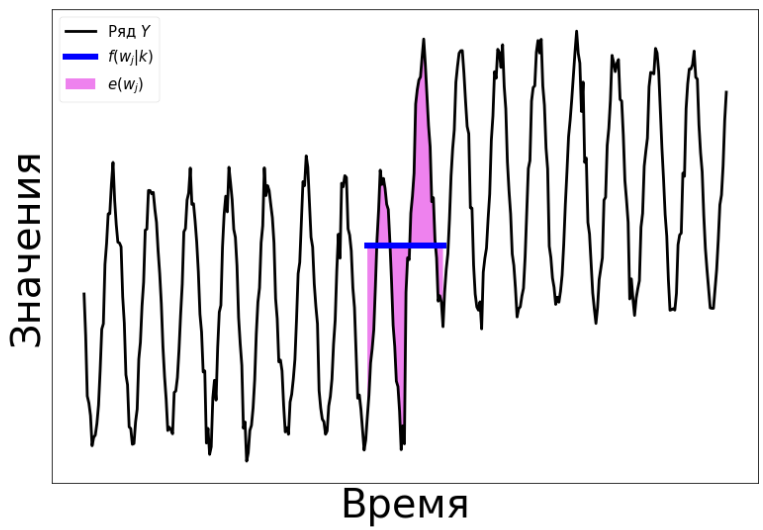
Нормирующую константу можно рассчитывать как ненормированное значение функции разладки на первом отрезке ряда (предполагая, что на этом отрезке не происходило разладок):

$$h = e(w_1) - e(w_1^{(\text{left})}) - e(w_1^{(\text{right})}).$$

Аппроксимация

- Итого, взяв ряд Y, мы «скользим» по нему окном ширины l
- Рассчитываем значения функции разладки $F()$ для каждого из получаемых подрядов $W^{(\text{all})}$
- Функция разладки начинает расти в окрестности точки разладки τ,
- Следовательно можно задать порог γ, такой что при превышении функции разладки этого порога в какой-то точке $\hat{\tau}$, разладка будет обнаружена





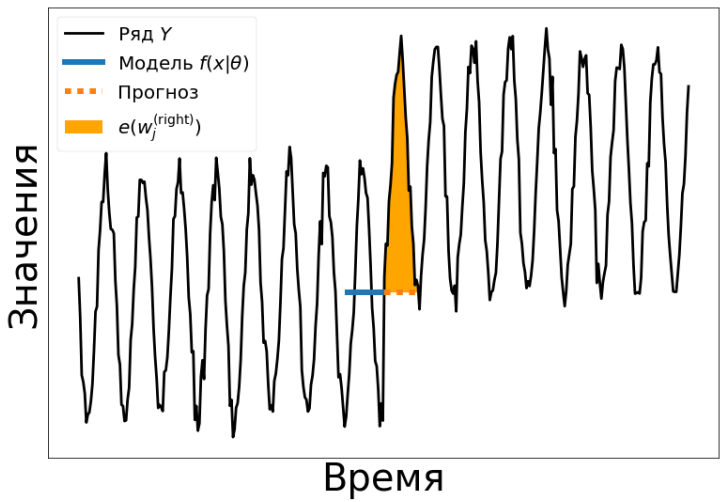
Прогнозирование

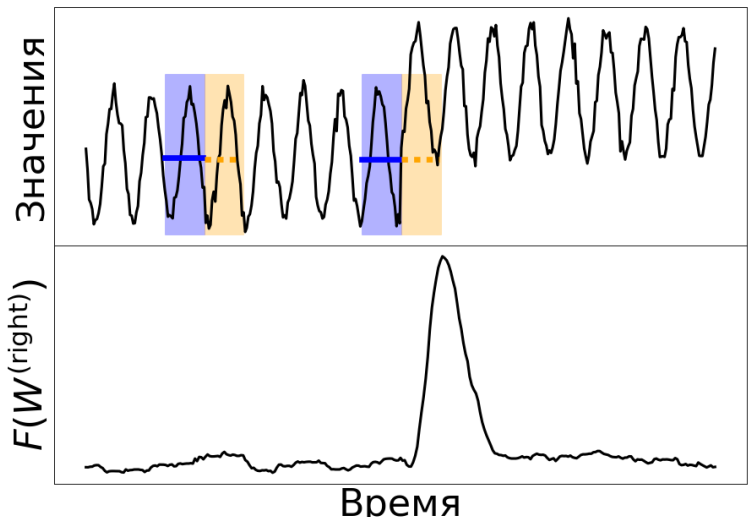
- Строим прогноз на несколько точек ряда вперед и считаем отклонение фактических значений от прогнозных
- В случае, если отклонение выше заданного порога, метод обнаруживает разладку
- Формально, оставаясь в тех же обозначениях, есть та же ширина окна l
- Есть последовательность подрядов $W = \{w_j\}_{j=1}^{k}$
- Каждый подряд w_j делится в этом методе на два ряда не обязательно одинаковой длины
- Введем индекс g, который будет указывать в какой точке ряда w_j он будет разделен на два
- формируется набор из пар рядов: $W^{(\text{left})} = \{w_j^{(\text{left})}\} = (y_j, \ldots, y_{j+g})$ и $W^{(\text{right})} = \{w_j^{(\text{right})}\} = (y_{j+g}, \ldots, y_{j+l})$

Прогнозирование

Ключевое отличие от методов аппроксимации: вместо расчета меры ошибки на том же ряду на котором подбирались параметры модели, мы оцениваем параметры θ модели $f(x|\theta)$ на ряде $w_j^{(\text{left})}$, делаем прогноз на $l - g$ точек и рассчитываем функцию ошибки $e(\cdot)$ на ряде $w_j^{(\text{right})}$.
Функция разладки принимает следующий вид:

$$f_j = F(w_j^{(\text{right})}) = \frac{e(w_j^{(\text{right})})}{h}.$$





Оценка качества

Допущения

В рамках данной работы мы разрабатываем систему своевременного оповещения о разладках во временных рядах.

- Нам важны две характеристики: точность и скорость обнаружения разладки
- Нам точно известны ряды с разладками и без
- Можем строить матрицы сопряжённости и считать метрики качества
- Для простоты оценки качества методов мы фиксируем точку разладки τ параметром n_0, тем самым фиксируя приемлемую задержку обнаружения разладки на уровне $n - n_0$

Матрица сопряжённости

Исходя из этого возможны четыре варианта:

- Разладка произошла и метод обнаружил точку разладки **после** фактической точки τ. Такая ситуация попадает под категорию True positive.
- Разладка произошла и метод не обнаружил точку разладки в диапазоне $(\tau, \cdots, n)$. Это случай False negative.
- Метод обнаружил разладку в диапазоне $(\tau, \cdots, n)$ в ряде без разладки. Это ситуация False positive.
- Разладки не было и метод не обнаружил разладку в диапазоне $(\tau, \cdots, n)$. Это случай True negative.

ROC-кривая

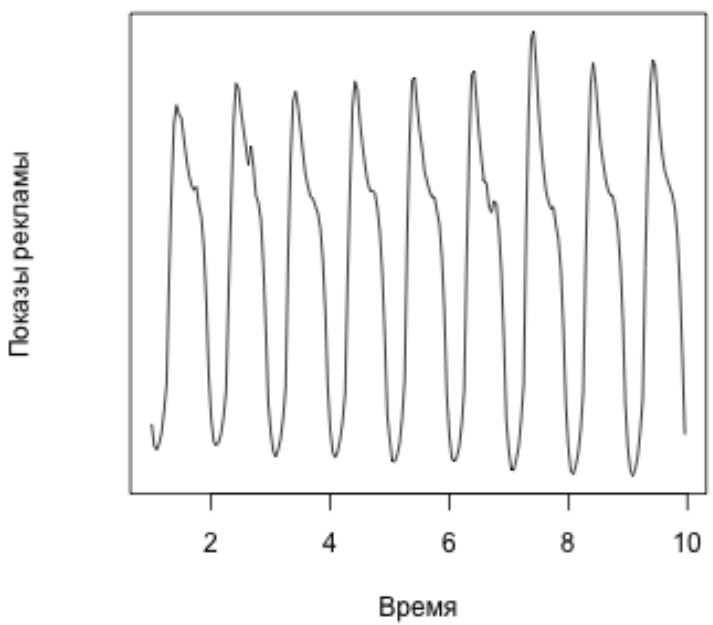
Можно строить ROC-кривые (изменяя порог γ) для разных методов обнаружения разладки, сравнивая как работают те или иные методы в контролируемой среде эксперимента.
Для сравнения качества методов мы будем пользоваться метрикой ROC-AUC, которая является ничем иным как площадью по ROC-кривой.

Моделирование данных

Реальные данные

Моделировать ряд будем как сумму тренда, периодики и шума. Тренд будем брать за константу, а периодику зададим как сумму косинусов с определенными периодичностями, амплитудами и фазами.

Пример реального ряда



Оценка парамтеров

С помощью процедуры ESPRIT нам удалось оценить следующие параметры периодической компоненты ряда:

- Ряд можно смоделировать четырьмя косинусами $J = 4$ с периодами $a_1 = 24, a_2 = 12, a_3 = 8, a_4 = 6$ (это логично, поскольку наши данные имеют суточные колебания).
- Оценка амплитуд данным методом получилась $A_1 = 1.05$, $A_2 = 0.82$, $A_3 = 0.27$, $A_4 = 0.05$
- А фазы косинусов возьмем $\phi_1 = \frac{3\pi}{4}$, $\phi_2 = \frac{\pi}{12}$, $\phi_3 = -\frac{2\pi}{3}$, $\phi_4 = -\frac{\pi}{3}$

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

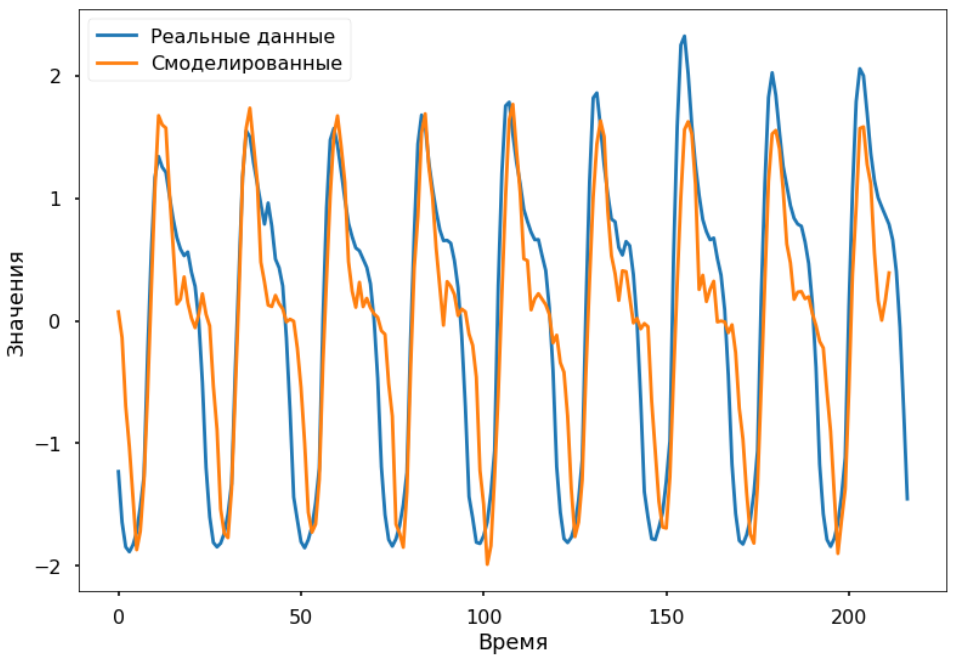
$$s_i = 1.05\cos(\frac{2\pi}{24}i+\frac{3\pi}{4})+0.82\cos(\frac{2\pi}{12}i+\frac{\pi}{12})+0.27\cos(\frac{2\pi}{8}i-\frac{2\pi}{3})+0.05\cos(\frac{2\pi}{6}i-\frac{\pi}{3}),$$

$i = 1, \ldots, n.$

Прочие параметры модели

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

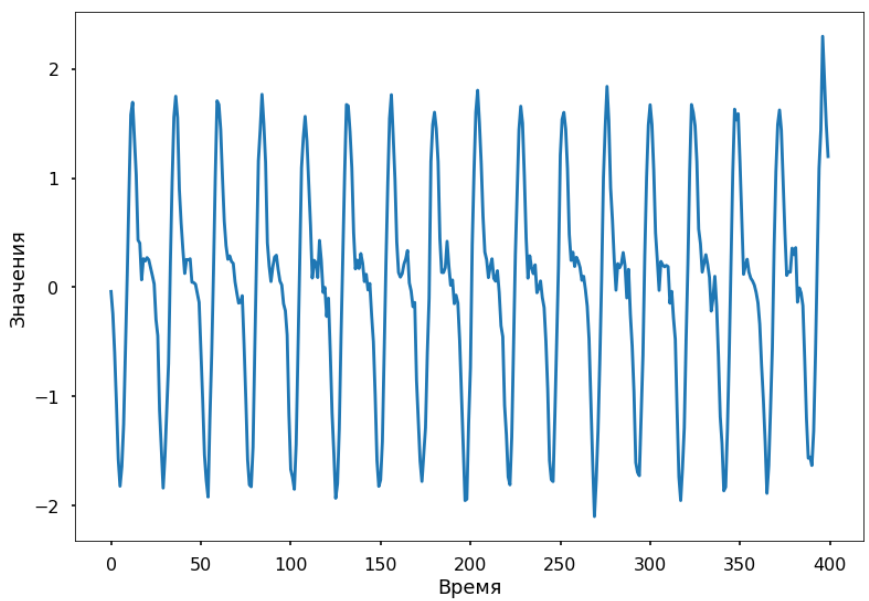
В результате, моделируемые ряды получились внешне достаточно похожими на реальные данные:



Моделирование разладки

- Вероятность возникновения разладки выберем $\rho = 0.8$
- Величину разладки $\delta^* \sim N(\mu = 1, \sigma = 0.4)$
- Место возникновения разладки зададим в самом конце ряда $n_0 = 396$ для изменения в среднем и локальной разладки
- Для изменения в тренде место возникновения разладки зададим с большей задержкой $n_0 = 328$

Пример сгенерированного ряда с разладкой



Применение методов

Моделирование рядов

Попробуем применить, описанные выше модели к смоделированным данным.

- Смоделируем 50 рядов
- У каждого ряда начало периодической компоненты выбирается случайно (то есть первый ряд может начинаться с нулевого часа, второй с пятого и т.п.). Это сделано, чтобы невелировать влияние периодичности на оценку качества метода.
- Параметры методов выбраны следующие. Длина окна l принимает значения 2, 4, 24, 48, 96. Разладка возникает трех типов: локальная, разладка в среднем, разладка в тренде
- Список значений порогов выбирается следующим образом. Моделируются 50 отдельных рядов и на них запускается расчет значений функции разладки при заданном методе и заданных параметрах. Далее берется 95 квантиль из полученных значений. После чего берётся 100 значений в диапазоне от нуля до 95 квантили с равными промежутками.

Методы

Всего будем сравнивать между собой 9 методов:

- Аппроксимация с выбранной моделью средним (или константной моделью)
- Аппроксимация с моделью из четырёх синусов с периодичностью 24, 12, 8, 6 и трендовой составляющей
- Аппроксимация с моделью из одного синуса с периодичностью 24 и тренда
- Аппроксимация с моделью только из тренда
- Прогнозирование с выбранной моделью средним (или константной моделью)
- Прогнозирование с моделью из четырёх синусов с периодичностью 24, 12, 8, 6 и трендовой составляющей
- Прогнозирование с моделью из одного синуса с периодичностью 24 и тренда
- Прогнозирование с моделью только из тренда
- Базовая простоя модель, в которой функция разладки это прирост текущих значений к значениям аналогичных часов сутки назад.

Замечания

Прежде, чем переходить к результатам следует оговорить следующие нюансы

- Методы, в которых лежит хоть какая-то сложная модель (то есть отличные от модели среднего) бессмысленно применять для окон l менее 48. Поскольку невозможно оценить какие либо параметры синуса, если длина ряда менее одного периода.
- В случае с разладкой в тренде бессмысленно применять методы с окном l менее 48 по тем же причинам

Результаты

В таблице приведены сводные результаты для экспериментах на 50 временных рядах.

Результаты применения методов к смоделированным данным

Тип разладки	local					mean					trend	
Окно	2	4	24	48	96	2	4	24	48	96	48	96
Точка разладки	396	396	396	396	396	396	396	396	396	396	328	328
Метод												
approximation_mean	0,88	0,65	0,50	0,65	0,67	0,51	0,70	0,53	0,86	0,69	1,00	1,00
approximation_sin_insight_trend				0,87	0,76				0,85	0,71	0,52	0,96
approximation_sin_trend				0,51	0,53				0,52	0,52	0,50	0,53
approximation_trend				0,63	0,71				0,54	0,51	0,50	0,52
naive	0,50	0,50	0,51	0,75	0,58	0,50	0,50	0,51	0,76	0,52	0,51	0,51
prediction_mean	0,70	0,66	0,57	0,62	0,50	0,54	0,57	0,54	0,54	0,62	0,95	0,96
prediction_sin_insight_trend				0,79	0,90				0,65	0,75	0,50	0,99
prediction_sin_trend				0,50	0,53				0,52	0,52	0,51	0,51
prediction_trend				0,53	0,52				0,62	0,50	0,50	0,51

Выводы

- Как мы видим, лучше всего сработал метод аппроксимации со средней моделью. Причем для разных типов разладки. Однако для каждого типа разладки у этого метода своя оптимальная длина окна
- Также, хорошо сработал метод аппроксимации с моделью из четырех синусов
- Базовый простой метод (в таблице это "naive") показал себя плохо. Причина тут в том, что наш ряд колеблется около нуля. Соответственно любые отклонения от нуля могут иметь очень большой прирост за счет эффекта низкой базы. Из-за этого базовый метод вероятно сработал бы неплохо в случае ряда с ненулевым средним, но в нашем случае он сработал плохо.
- В целом, результаты методов с использованием аппроксимации не столь существенно отличаются от результатов методов с использованием прогнозирования.