# Обнаружение разладки во временных рядах показов мобильной рекламы

К.В. Мерзляков, группа 622

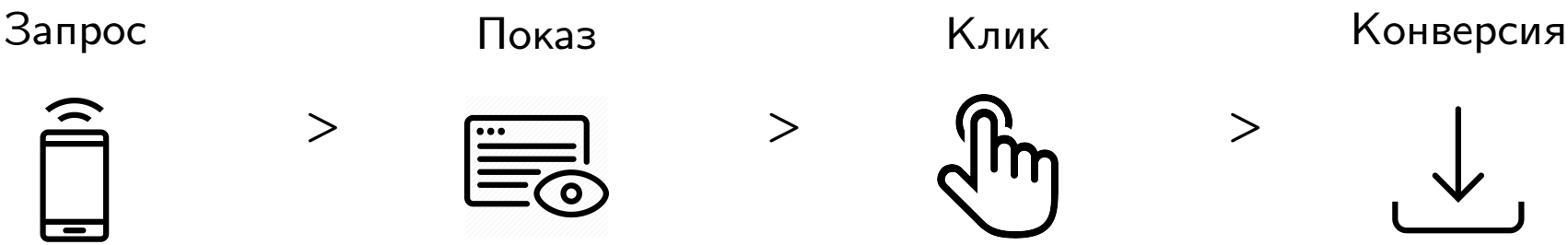
Санкт-Петербургский Государственный Университет
Кафедра статистического моделирования

18.05.2019

# Содержание

# Общие замечания

# Изменения в данных




Изменения на стороне пользователя

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

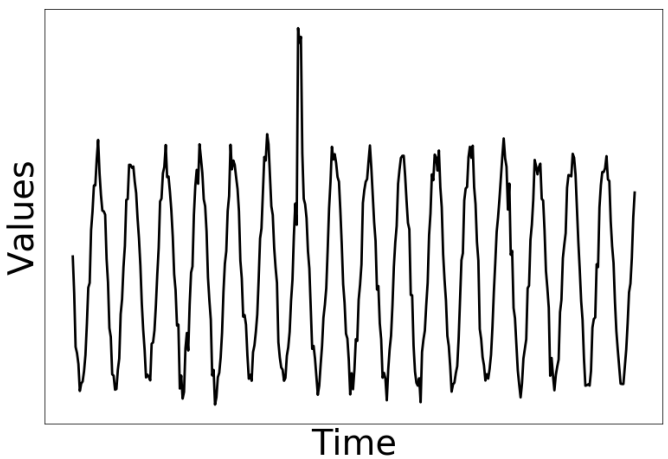
Изменения на стороне рекламной сети

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

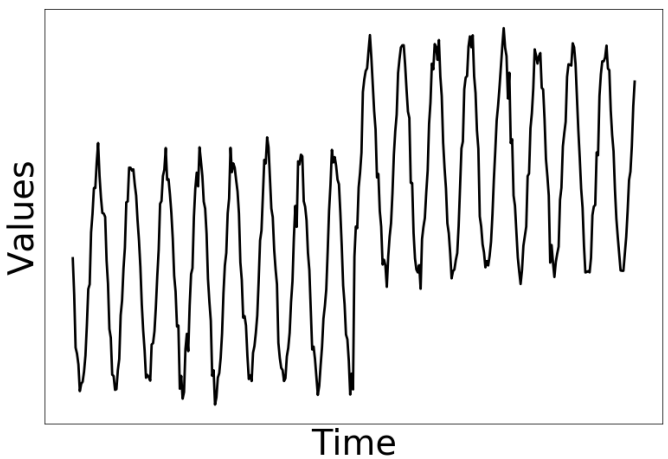
# Разладка во временных рядах

- 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**

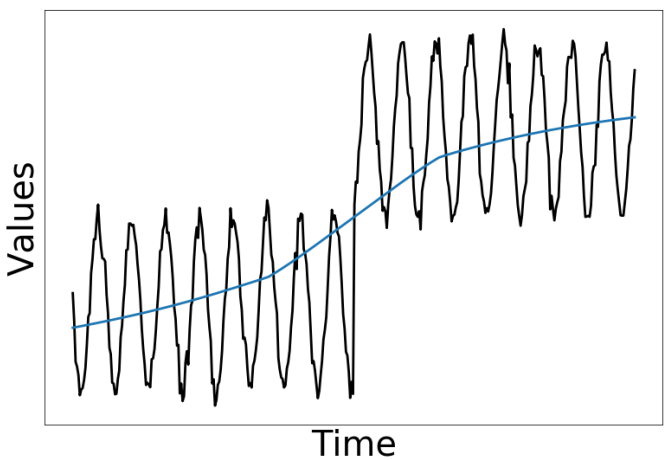


# Мотивация

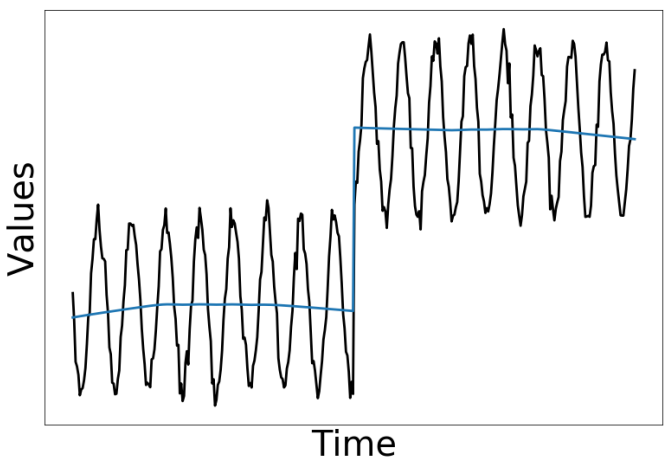
Практическая польза:

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

**Извлечение тренда без анализа разладок**

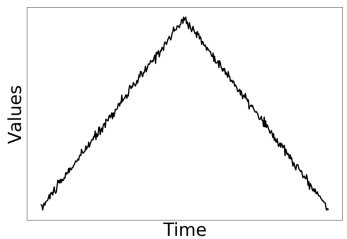



**Извлечение тренда с анализом разладок**

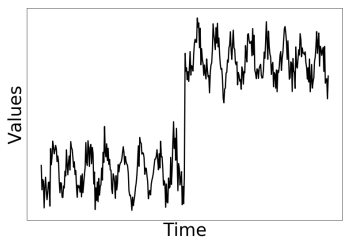


# Виды разладок

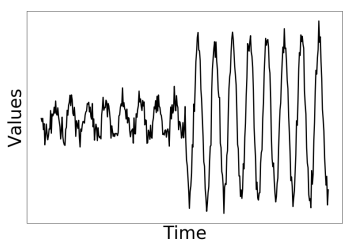
Изменение в тренде



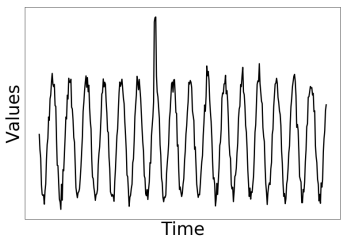

Изменение в среднем



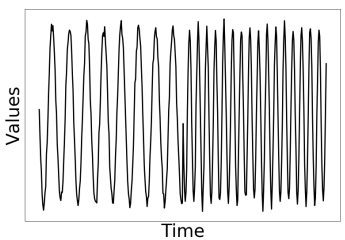

Изменение в амплитуде колебаний




Локальное изменение




Изменение в периодике



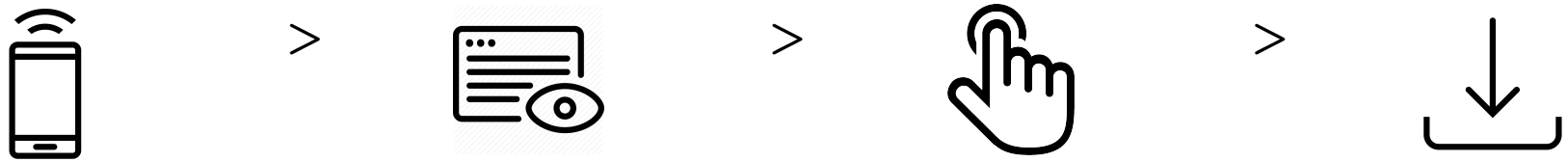
# Структура исследования

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

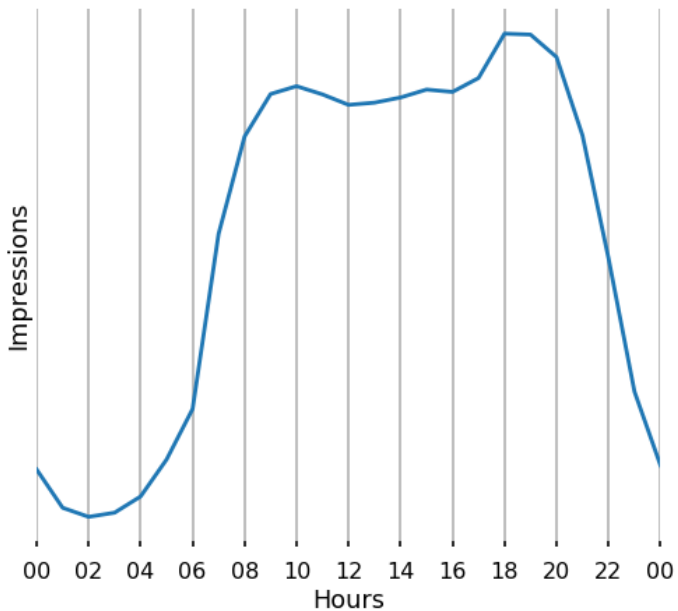
# Построение модели данных

# Данные мобильной рекламы

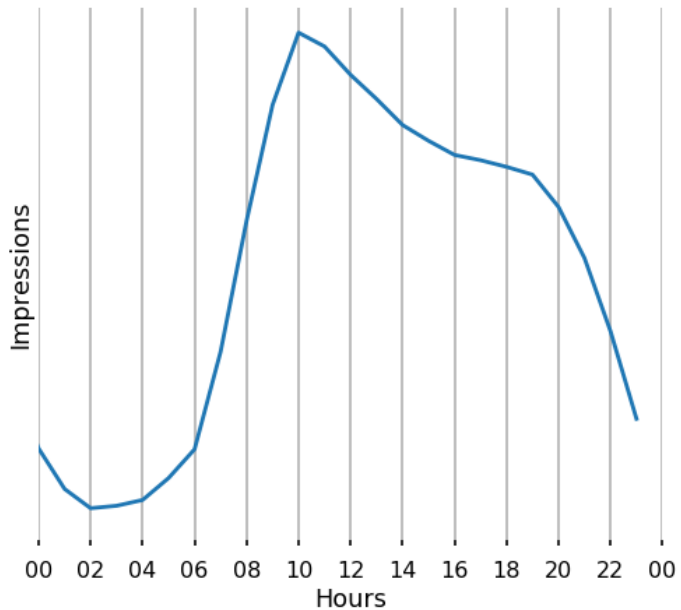
Запрос > Показ > Клик > Конверсия



**Типичный будний день**

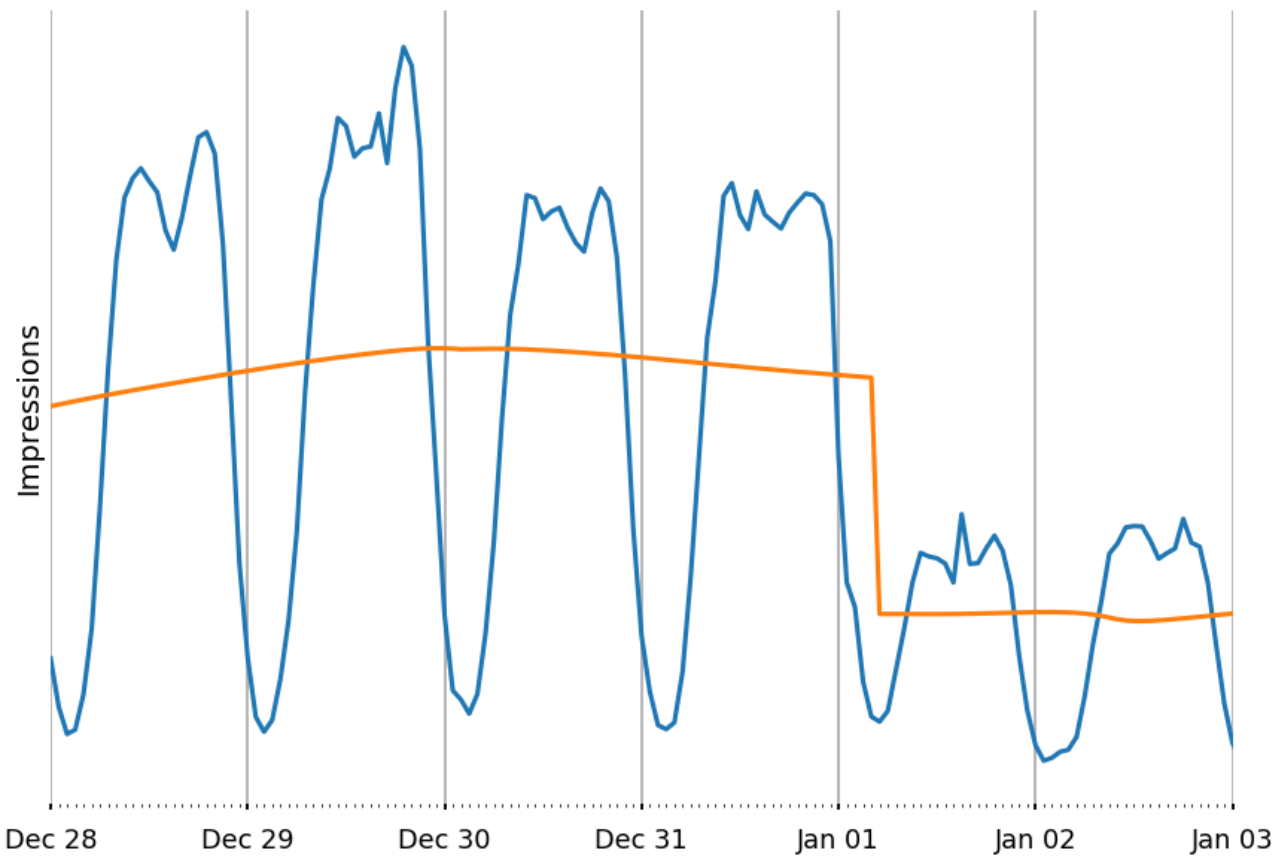



**Типичный выходной день**



# 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; $\phi_j$ — фаза $j$-го косинуса.

# Модель разладки. Изменение в среднем

Модель разладки 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 $\rho$.

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

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

$\delta^{(mean)}$ — значение разладки. Чтобы разладка была заметна, введем ещё минимальное допустимое значение разладки $\delta^{(mean)}_{min}$
Значение разладки является случайной величиной с некоторым распределением. $\delta^{(mean)*} \sim N(\mu^{(\text{ср})}, \sigma^{2(\text{ср})}); |\delta^{(mean)*}| \geqslant \delta^{(mean)}_{min}$:

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

# Модель разладки. Локальная

Отличие от предыдущего типа разладки в том, что в локальной разладки разладка влияет только на одну точку ряда.

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

Остальное остается идентичным предыдущему варианту.

# Модель разладки. Изменение в тренде

Модель разладки можно задать следующим образом:

- Разладка только в одной точке ряда;
- Разладка только в тренде и заключается в изменении коэффициента тренда;
- Разладка может произойти не всегда, а с некоторой вероятностью $\rho$.

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

$$\tilde{t_i} = \begin{cases} t_i, & i < \tau, \\ t_{\tau-1} + \delta^{(trend)}(i - \tau + 1), & i \geqslant \tau, \end{cases}$$

$\delta^{(trend)}$ — значение разладки. Чтобы разладка была заметна, введем ещё минимальное допустимое значение разладки $\delta^{(trend)}_{min}$
Таким образом, независимо от типа разладки, 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 $\theta$.
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 :

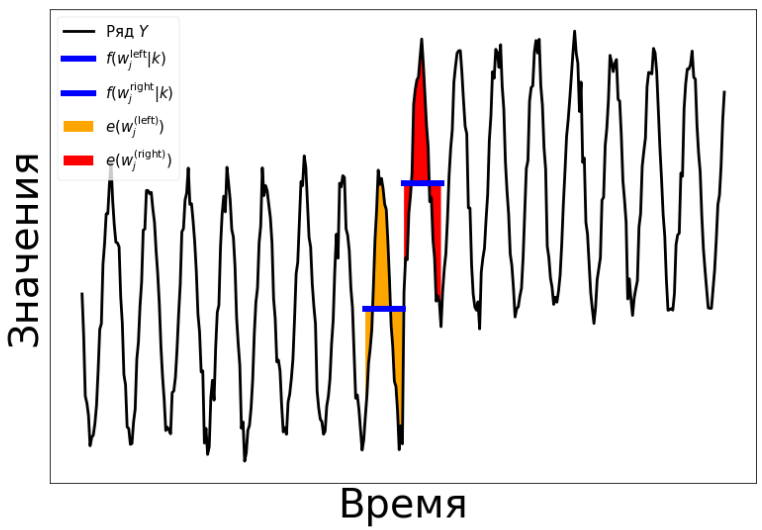
$$q_i = \begin{cases} f_{i-l+1}, & i \geq l, \\ 0, & i < l. \end{cases}$$

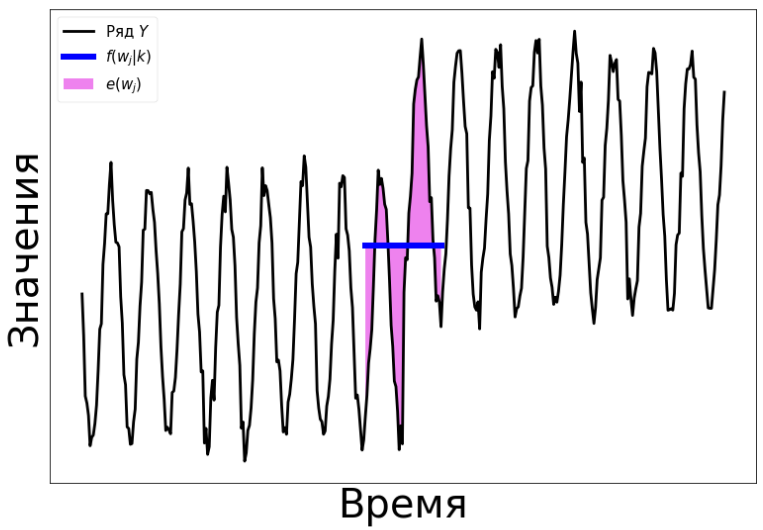
Нормирующую константу можно рассчитывать как ненормированное значение функции разладки на первом отрезке ряда (предполагая, что на этом отрезке не происходило разладок):

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

# Аппроксимация

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



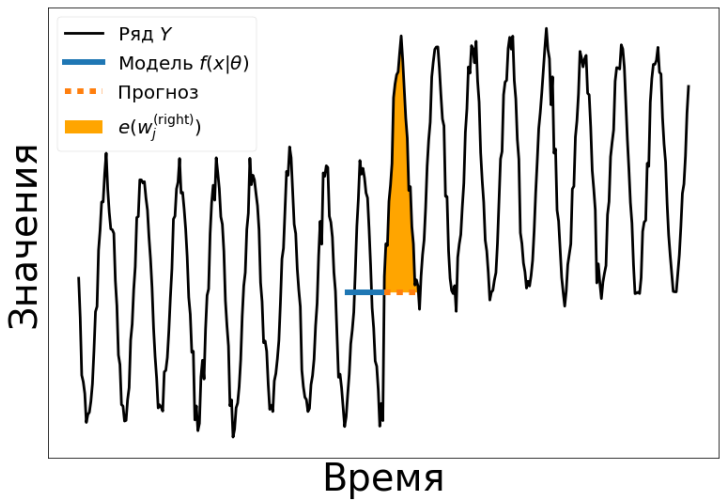
# Прогнозирование

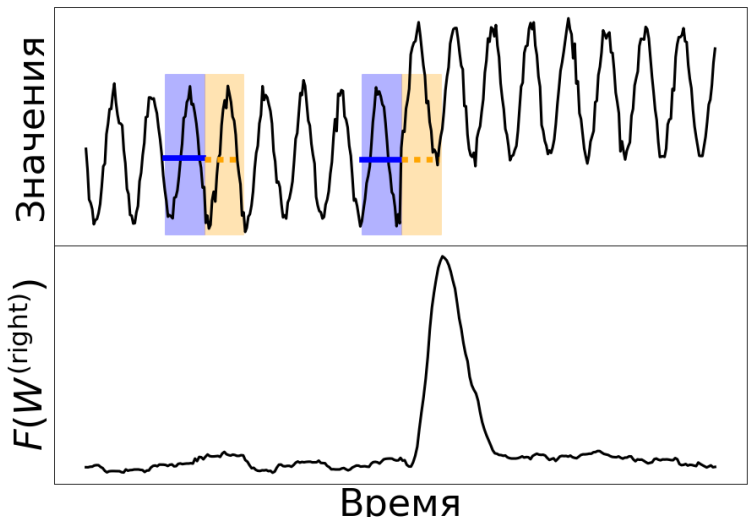
- Строим прогноз на несколько точек ряда вперед и считаем отклонение фактических значений от прогнозных
- В случае, если отклонение выше заданного порога, метод обнаруживает разладку
- Формально, оставаясь в тех же обозначениях, есть та же ширина окна $l$
- Есть последовательность подрядов $W = \{w_j\}_{j=1}^{k}$
- 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 $\theta$ 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 $\tau$ параметром $n_0$, тем самым фиксируя приемлемую задержку обнаружения разладки на уровне $n - n_0$

# Матрица сопряжённости

Таким образом, у нас имеется приемлемая задержка, в рамках которой нас интересует обнаружить разладку. При этом, за пределами приемлемой задержки нас не интересует что происходит с рядом. Исходя из этого возможны четыре варианта:

- Разладка произошла и метод обнаружил точку разладки **после** фактической точки $\tau$. 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

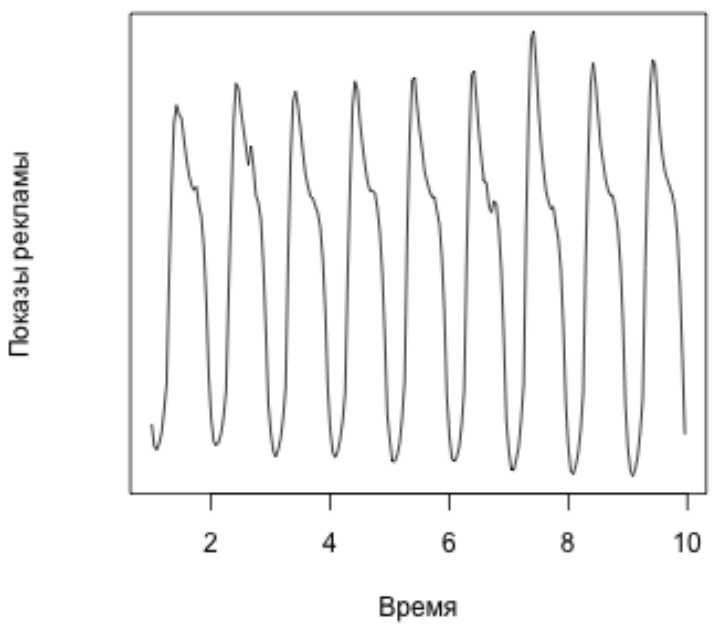
- Можно строить ROC-кривые (изменяя порог $\gamma$) для разных методов обнаружения разладки, сравнивая как работают те или иные методы в контролируемой среде эксперимента.
- ROC-кривая — график, позволяющий оценить качество бинарной классификации. Он отображает соотношение между долей верно-положительно классифицированных наблюдений от общего количества положительных классов, и долей ложно-отрицательно классифицированных наблюдений от общего количества отрицательных наблюдений при варьировании порога $\gamma$.
- Другими словами, ROC кривая это график, где по оси ординат откладывается TPR (англ. True Positive Rate), а по оси абсцисс откладывается FPR (англ. False Positive Rate). При этом каждая точка является значением TPR и FPR для какого-то конкретного значения порога.
- $TPR = \frac{\sum \text{Верно-положительные классификации}}{\sum \text{Все положительные наблюдения}}$, $FPR = \frac{\sum \text{Ложно-отрицательные классификации}}{\sum \text{Все отрицательные наблюдения}}$
- Для сравнения качества методов мы будем пользоваться метрикой ROC-AUC, которая является площадью под 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**

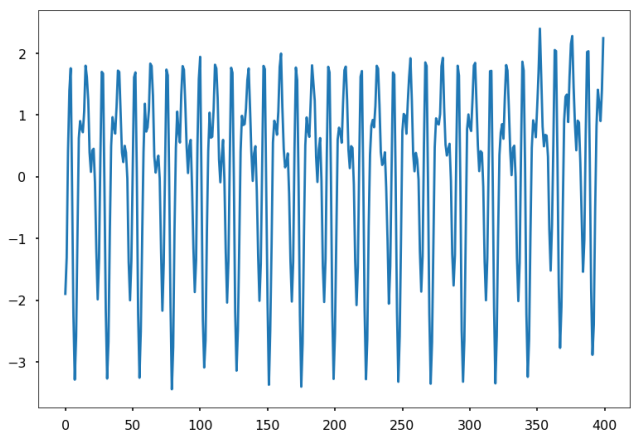


# Реальные данные

Длина ряда с предыдущего слайда 216 (то есть 9 суток). Применим к этому ряду метод SSA с окном 96. И оценим параметры периодичности по первым 10 компонентам (исключая тренд) используя .

| Периоды | Фазы | Амплитуды | Коэффициенты |
|---|---|---|---|
| 23.93 | 2.005 | 1.001 | 1 002 339 |
| 11.99 | -2.749 | 1.002 | 302 234 |
| 7.99 | -1.031 | 1.002 | 73 878 |
| 5.99 | 0.750 | 1.002 | 5 238 |

Получается 4 косинуса с периодами 24, 12, 8, 6 (это логично, так как у в ряде наблюдается суточная периодичность). Однако, если взять оцененные параметры фаз и амплитуд, то получится следующий график:



# Оценка парамтеров

К сожалению, график на предыдущем слайде не совсем похож на график исходных данных. Дело в коэффициентах амплитуды и в фазах.
Попробуем подобрать эти значения вручную.
В результате получились следующие параметры ряда:

- Ряд можно смоделировать четырьмя косинусами $J = 4$ с периодами $a_1 = 24, a_2 = 12, a_3 = 8, a_4 = 6$.
- Адекватные параметры амплитуд получились $A_1 = 1.05$, $A_2 = 0.82$, $A_3 = 0.27$, $A_4 = 0.05$
- 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:

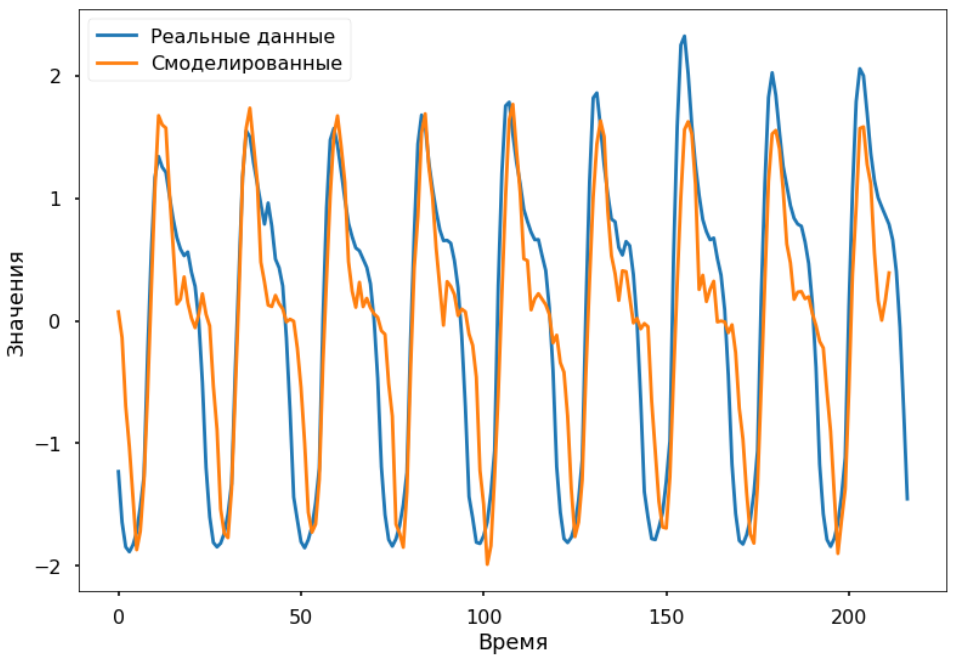
$$s_i = 1.05\cos(\frac{2\pi}{24}i+\frac{3\pi}{4})+0.82\cos(\frac{2\pi}{12}i+\frac{\pi}{12})+0.27\cos(\frac{2\pi}{8}i-\frac{2\pi}{3})+0.05\cos(\frac{2\pi}{6}i-\frac{\pi}{3}),$$

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

# 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 $\mu = 0, \sigma = 0.1$

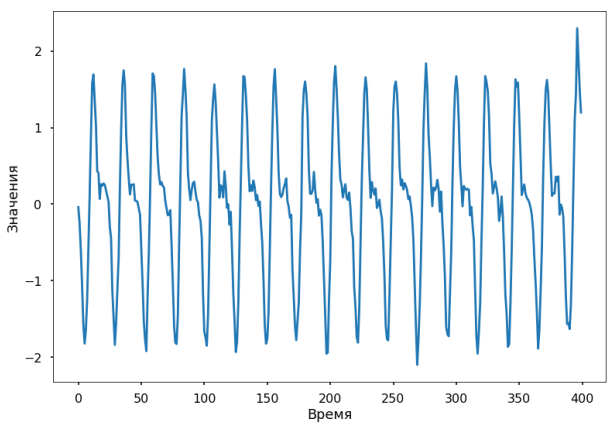
В результате, моделируемые ряды получились внешне достаточно похожими на реальные данные:



# Моделирование разладки

- Вероятность возникновения разладки выберем $\rho = 0.8$
- Величины разладки $\delta^{(mean)*} \sim N(\mu = 0, \sigma = 0.2)$, $\delta^{(local)*} \sim N(\mu = 0, \sigma = 1)$, $\delta^{(trend)*} \sim N(\mu = 0, \sigma = 0.01)$
- Минимальные допустимые значения разладок: $\delta^{(mean)*}_{min} = 0.3$, $\delta^{(local)*}_{min} = 0.5$, $\delta^{(trend)*}_{min} = 0.005$
- 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 (с разладкой и без) 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

И для подхода с аппроксимацией и для подхода с прогнозированием мы будем использовать следующие модели:

- Среднее $f(x|b) = b$
- Четыре косинуса с периодами из модели генерации ряда + тренд $f(x|P_i, p_i, \chi_i, b) = \sum_{i=1}^{4} P_i \cos(\frac{2\pi}{p_i}x + \chi_i) + bx$, где $p_1 = 24, p_2 = 12, p_3 = 8, p_4 = 6$
- Один косинус с периодом 24 + тренд $f(x|P, 24, \chi, b) = P\cos(\frac{2\pi}{24}x + \chi) + bx$
- Только тренд $f(x|b) = bx$

**Обратите внимание**

Следует отличать модель ряда, с помощью которого генерировался искусственный ряд и модель, используемая внутри метода обнаружения разладки.

# Методы

Всего будем сравнивать между собой 8 методов:

- Аппроксимация с выбранной моделью средним
- 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
- 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

# Замечания

Поскольку у нас есть 8 методов с одной стороны, и решетка параметров из 5 вариантов (длина окна $l$) с другой, то мы будем оценивать качество всех методов для комбинаций методов и параметров.
Однако не во всех случаях корректно применять методы, поэтому проговорим исключения, когда мы не будем считать качество:

- Методы, в которых лежит модель, отличная от 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 ROC-AUC для экспериментов 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 | 0,49 | 0,65 | 0,67 | 0,52 | 0,70 | 0,53 | 0,86 | 0,69 | 1,00 | 1,00 |
| approximation_sin_insight_trend | | | | 0,87 | 0,76 | | | | 0,85 | 0,71 | 0,61 | 0,96 |
| approximation_sin_trend | | | | 0,53 | 0,53 | | | | 0,52 | 0,55 | 0,58 | 0,83 |
| approximation_trend | | | | 0,63 | 0,71 | | | | 0,50 | 0,50 | 0,48 | 0,66 |
| prediction_mean | 0,70 | 0,66 | 0,57 | 0,62 | 0,48 | 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 | 0,58 | 0,99 |
| prediction_sin_trend | | | | 0,50 | 0,53 | | | | 0,48 | 0,48 | 0,58 | 0,85 |
| prediction_trend | | | | 0,51 | 0,54 | | | | 0,62 | 0,49 | 0,64 | 0,91 |

# Выводы

- Как мы видим, лучше всего сработал метод аппроксимации с моделью "Среднее". 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
- Методы прогнозирования с моделями "Среднее"и "4 синуса + тренд"сработали хорошо, но немного хуже чем методы аппроксимации
- Примечательно, что все методы хорошо определяют разладку в тренде, если выбрать окно $l = 96$. Вероятно имеет смысл понизить величину разладки при генерации ряда.
- Плохо сработали методы с моделью "Синус + тренд"и с моделью "Тренд"(и аппроксимация и прогнозирование).