Санкт-Петербургский государственный университет
Прикладная математика и информатика
Статистическое моделирование

Мерзляков Климент Викторович

Обнаружение разладки во временных рядах показов мобильной рекламы

Отчет о научно-исследовательской работе

Научный руководитель:
к. ф.-м. н., доцент Н. Э. Голяндина

Санкт-Петербург
2018

Оглавление

Введение

Рекламной сетью называют некоторую площадку или систему, которая является посредником между рекламодателями и собственниками рекламных мест — владельцев сайтов, мобильных приложений и каких-либо других пространств, где можно размещать рекламу.

В интернет-рекламе взаимодействие рекламной сети с пользователем можно описать следующей последовательностью событий. При выполнении некоторых условий (например, пользователь открыл мобильное приложение) с устройства пользователя отправляется запрос на показ рекламы. Если запрос удовлетворяется, то происходит событие „показ“, то есть пользователь непосредственно видит рекламу. После этого может произойти событие „клик“ и далее какое-либо целевое действие. В мобильной интернет-рекламе „показ“ является одним из ключевых событий, поскольку он отражает количество рекламы доставленное до конечного пользователя.

Рекламные интернет-сети являются интересным объектом для исследования с точки зрения обнаружения разладки, поскольку все показатели отслеживаются с точностью до секунды, происходит большое количество событий, а так как рекламные сети, как правило, работают на международном рынке, то существует возможность тестировать гипотезы на большом количестве различных временных рядов.

Одной из текущих проблем, стоящих перед рекламными сетями — это низкая скорость реагирования на любые резкие изменения текущего состояния. Такие изменения безусловно отражаются в данных в виде аномальных значений, резких всплесков и внезапных изменений тренда. Однако проблема заключается в том, что показателей требующих отслеживания могут быть десятки, при этом на каждый показатель может влиять большое количество факторов. Поэтому зачастую, чтобы локализовать и устранить проблему требуется просмотреть сотни графиков. Отсюда следует, что наличие качественного метода обнаружения разладки каждого показателя по каждому измерению позволило бы не только существенно сэкономить ресурсы, но и в целом повысить эффективность бизнеса. Поэтому целью данной работы является разработка методики обнаружения разладки. В работе будут использоваться фактические, данные одной из работающих рекламных сетей.

Глава 1

Обнаружение разладки во временных рядах

1.1. Моделирование данных

Реальные данные интернет-рекламы имеют стабильную дневную периодичность (на рисунке 1.1 приведен пример типичной динамики в рамках дня). По более длинному ряду изображенному на рисунке 1.2 видно, что в данных время от времени возникают разладки разных видов, при этом сам ряд имеет мультипликативный характер (с изменением среднего уровня ряда пропорционально меняется и амплитуда колебаний). В реальных временных рядах достаточно сложно разметить наличие разладок — зачастую сложно отделить разладку от шума. Поэтому вместо разметки реальных рядов мы будем моделировать искусственные ряды похожие на ряды данных интернет-рекламы с определенным шумом и разладками в известных местах.

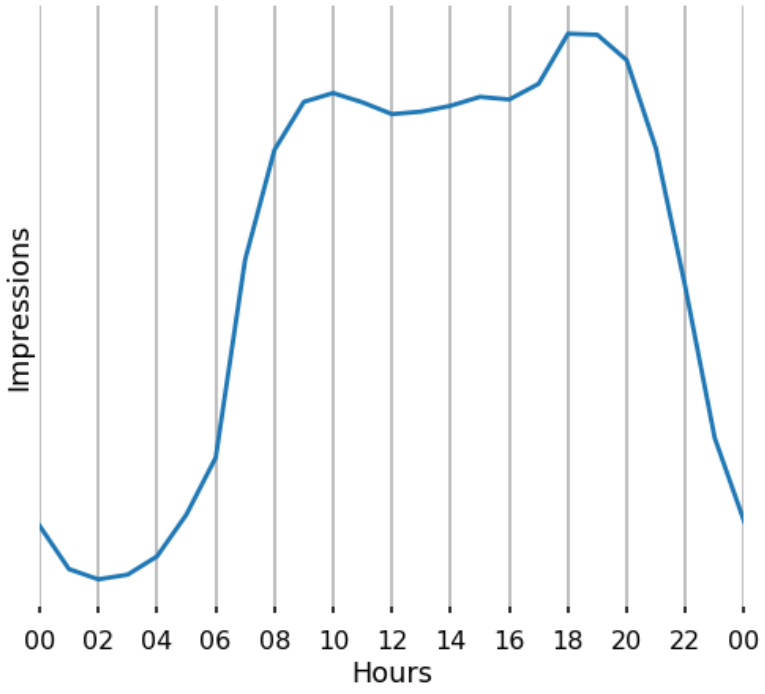

Рис. 1.1. Пример показов рекламы за сутки

Обозначим временной ряд $Y = (y_1, \ldots, y_n)$. Наблюдаемые значения ряда можно представить в виде суммы компонент:

$$Y = T + S + E,$$

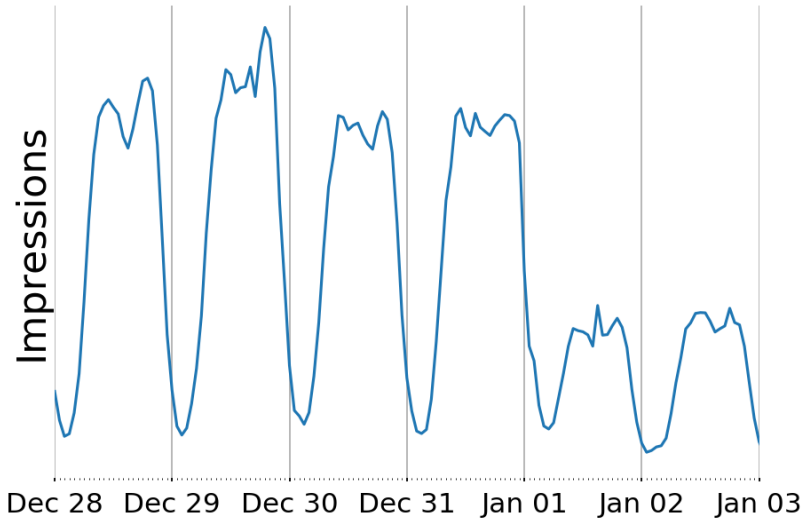

Рис. 1.2. Пример показов рекламы с разладкой за неделю

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. По каждой из этих компонент требуется построить модель. Это можно сделать, например, следующим образом:

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

$$s_i = A\cos(\frac{2\pi}{a}i + \phi), \quad i = 1, \ldots, n,$$

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

где i индекс элемента ряда; c — константа, A — амплитуда; a — период; ϕ — фаза.

Ряды, модель которых мы хотим построить имеют мультипликативность (амплитуда колебаний меняется пропорционально изменению тренда). Такого эффекта можно достичь, взяв экспоненту от исходной модели ряда:

$$Y^{(\text{mult})} = e^Y.$$

Далее нам нужно построить модель разладки, исходя из следующего:

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

Формально это можно описать так: пусть τ — точка (индекс) разладки, тогда тренд с разладкой разумно будет записать $\tilde{T} = (\tilde{t_1}, \ldots, \tilde{t_n})$. И элементы тренда с разладкой будут иметь вид:

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

где δ — значение разладки.

Значение разладки является случайной величиной с некоторым распределением. В данной работе значение разладки будет иметь нормальное распределение $\delta^* \sim N(\mu^{(\text{cp})}, \sigma^{2(\text{cp})})$, с некоторой вероятностью возникновения ρ:

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

Таким образом, δ является случайной величиной с распределением-смесью. При этом точка разладки τ тоже является случайной величиной с равномерным распределением на $[n_0, \ldots, n - n_0]$, где n_0 — самая первая возможная точка разладки, которая задается параметром. n_0 введена намеренно, чтобы разладка при моделировании не возникала в первых и последних точках ряда.

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

$$\tilde{Y}^{(\text{mult})} = e^{\tilde{T}+S+E}.$$

В результате модель временного ряда имеет пять параметров: $A, a, \phi, \mu, \sigma, c$, а модель разладки имеет еще пять параметров: $\delta, \mu^{(\text{cp})}, \sigma^{(\text{cp})}, \rho, n_0$.

1.2. Методы обнаружения разладки

Опишем один из подходов к обнаружению разладки. Данный подход не является единственным, хотя заключает в себе широкое разнообразие методов. Как правило, у временного ряда есть некоторая структура (сигнал), которая может быть описана той или иной моделью. Идея подхода заключается в том, что около точки разладки модель плохо описывает временной ряд. Используя некоторую меру ошибки мы можем измерять то, насколько хорошо или плохо 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 предсказания

- Методы на основе аппроксимации

1.2.1. Методы на основе аппроксимации

Пусть l четное вещественное число, называемое шириной окна. При этом $1 < l < n$. 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 (это возможно поскольку l четное по условию): $W^{(\text{left})} = \{w_j^{(\text{left})}\} = (y_j, \ldots, y_{\frac{j+l+1}{2}})$ и $W^{(\text{right})} = \{w_j^{(\text{right})}\} = (y_{\frac{j+l+1}{2}}, \ldots, y_{j+l+1})$.

Таким образом, для каждого ряда W можно сформировать тройки рядов:

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

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

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

где $X = (x_1, \ldots, x_m)$ некоторый вещественный временной ряд длины m, а $f(x|\theta)$ некоторая модель этого временного ряда с параметрами θ.

Функция $f(x|\theta)$ может быть константной ($\theta = (k)$):

$$f(x|k) = k.$$

Либо другой подходящей под наш ряд функцией, например:

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

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

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

где h — значение нормировки.

Расчет нормирующей константы является открытой проблемой, поскольку имеются разные варианты её расчета со своими плюсами и минусами. Например, можно рассчитывать её как значение функции ошибки на первом отрезке ряда (предполагая, что на этом отрезке не происходило разладок):

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

Для наглядности, на рисунке 1.3 приведен пример расчета ошибки на одном левом ряде $w_j^{(\text{left})}$ и на одном правом ряде $w_j^{(\text{right})}$. А на рисунке 1.4 показан пример расчет ошибки на одном общем ряде (в который входит и левая и правая части).

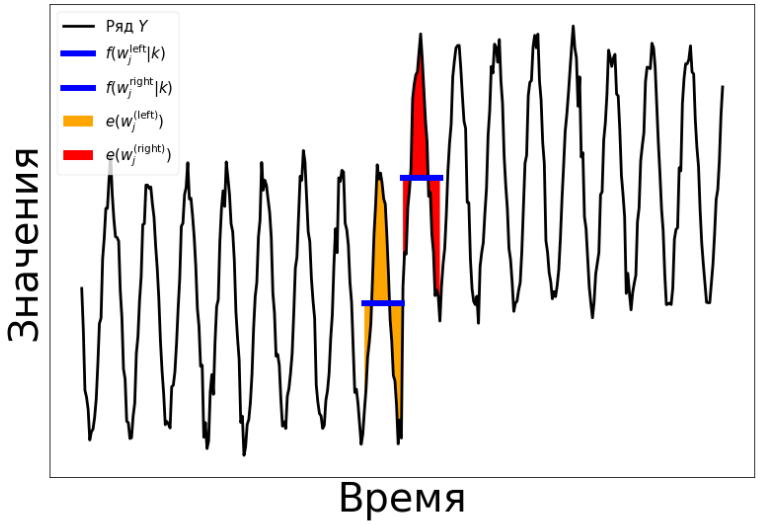

Рис. 1.3. Пример промежуточного расчета ошибки методом аппроксимации

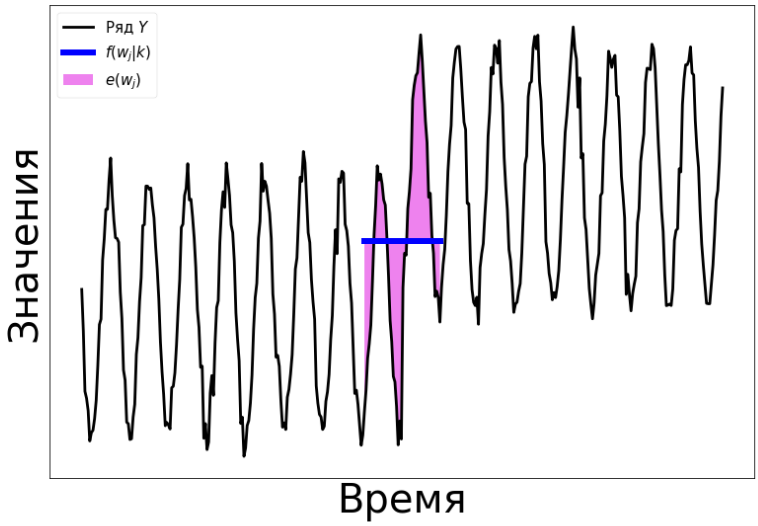

Рис. 1.4. Пример промежуточного расчета ошибки методом аппроксимации, продолжение.

Итого, взяв ряд Y мы «скользим» по нему окном длины l и 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 некий порог γ, такой что при превышении функции разладки этого порога в какой-то точке $\hat{\tau}$ будем считать,

что разладка обнаружена в этой точке.

В результате, в данных методах нужно задавать следующие параметры: ширину окна l, функцию аппроксимации f и порог γ.

1.2.2. Методы на основе прогнозирования

Методы на основе прогнозирования очень похожи на методы с использованием аппроксимации. Суть их заключается в том, что мы 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 всё та же ширина окна l (однако l в данном случае может быть нечетным) и 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. Таким образом, формируется набор из пар рядов: $W^{(\text{left})} = \{w_j^{(\text{left})}\} = (y_j, \ldots, y_{j+g})$ и $W^{(\text{right})} = \{w_j^{(\text{right})}\} = (y_{j+g}, \ldots, y_{j+l})$. Ключевое отличие от методов аппроксимации заключается в том, что 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 $c(\cdot)$ на ряде $w_j^{(\text{right})}$. При этом функция разладки принимает следующий вид:

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

В остальном данные методы ничем не отличаются от методов на основе аппроксимации.

Для наглядности, на рисунке 1.5 приведен пример расчета ошибки с помощью метода прогнозирования на одном ряде $w_j^{(\text{right})}$. А на рисунке 1.6 показан пример расчета функции разладки $F(W^{(\text{right})})$ для всего ряда.

В методах прогнозирования нужно задавать следующие параметры: ширину окна l, функцию прогнозирования f, индекс разделения окна (по сути с помощью него определяется на основе какого количества точек подбираются параметры модели, а на сколько точек происходит прогноз) g и порог γ.

Есть смысл в дальнейшем переписать аппроксимацию (l не обязательно четное, а функции ошибок делятся на ширину. Тогда получится универсальная запись для прогнозирования и для аппроксимации)

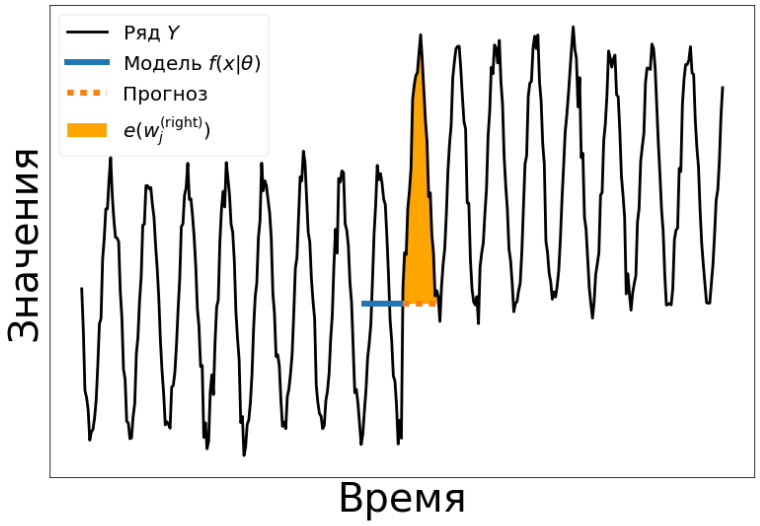

Рис. 1.5. Пример расчета ошибки методом прогнозирования

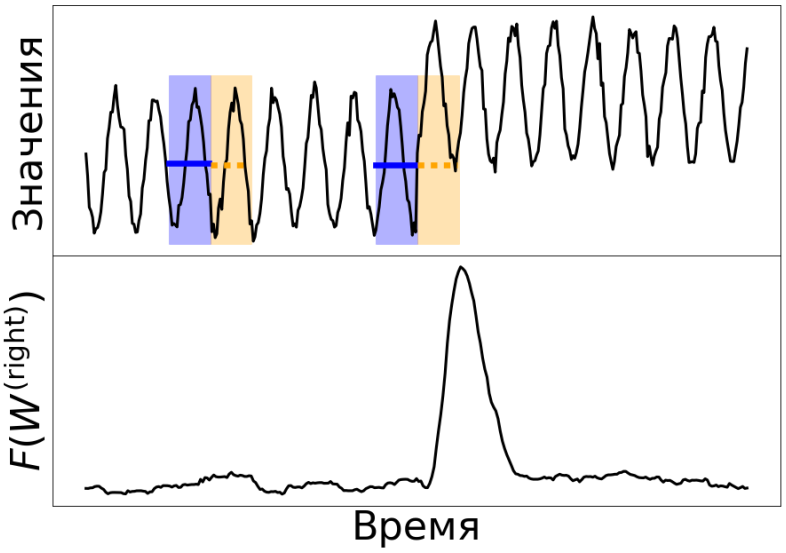

Рис. 1.6. Пример расчета функции разладки с помощью скользящего окна

1.3. Оценка качества

В рамках данной работы мы разрабатываем систему своевременного оповещения о разладках во временных рядах. При такой постановке задачи важны две характеристики: точность обнаружения разладки и скорость обнаружения разладки. Поскольку мы используем моделированные данные, то мы точно знаем в каких из смоделированных нами рядов произошла разладка, а в каких разладки не было. Более того, мы точно знаем момент разладки. Благодаря этому, мы можем строить матрицы сопряжённости и считать метрики качества.

При этом надо понимать, что в такой постановке задачи нам важно обнаружить разладку не позднее какого-то срока, иначе оповещение о разладке будет несвоевременным. Также, чтобы получать разумные результаты оценки методов, мы будем исходить из того, что метод может обнаруживать сколь угодно много разладок, в то время как фактическая разладка будет происходить только в одном месте (либо не происходить вовсе). Такая поправка введена намерено, в частности во избежания случаев, когда метод с низким значением порога γ будет останавливаться практически сразу, из-за чего матрицы сопряженности становится проблематично интерпретировать.

Таким образом, после каждого применения метода обнаружения разладки мы получаем вектор точек разладок $\hat{\mathrm{T}} = (\hat{\tau}_1, \cdots, \hat{\tau}_q)$, где q — количество разладок, обнаруженное методом. Стоит отметить, что q может быть равно нулю.

Исходя из этого возможны четыре варианта:

- Разладка произошла и метод обнаружил точку разладки **после** фактической точки τ, но не слишком поздно, то есть не позднее, чем $\tau + u$, где u параметр. Параметром u мы задаем приемлемую для нас задержку в обнаружении разладки. Такая ситуация попадает под категорию True positive.

- Разладка произошла и метод не обнаружил точку разладки в диапазоне $(\tau, \cdots, \tau + u)$. Это случай False negative.

- Метод обнаружил разладку там, где ее не было. То есть либо за пределами $(\tau, \cdots, \tau+u)$, либо когда разладки вообще не было в ряде. Это ситуация False positive.

- Разладки не было и метод не обнаружил разладку. Это случай True negative.

Договорившись о таком способе оценки качества, можно строить ROC-кривые (изменяя порог γ) для разных методов обнаружения разладки и таким образом, сравнивая как работают те или иные методы в контролируемой среде эксперимента.

Глава 2

Применение методов обнаружения разладки к смоделированным рядам

2.1. Моделирование данных

На рисунке 1.2 представлен пример реальных данных за семь дней. Существенно упрощая, мы можем смоделировать данный ряд используя подход, описанный в разделе 1.1. Для простоты дальнейшей работы зафиксируем длину ряда $n = 400$. Значение тренда пока что выберем нулевым: $t_i = 0, i = 1, \cdots, n$; амплитуду колебаний $A = 1$; период $a = 24$ (поскольку реальные данные имеют очевидную суточную периодичность); фазу $\phi = \frac{\pi}{2}$; параметры шума $\mu = 0, \sigma^2 = 1$.

Таким образом, модель для генерации ряда имеет следующий вид:

$$y_i = \cos(\frac{2\pi}{24}i + \frac{\pi}{2}) + N(0, 1).$$

Пример сгенерированного ряда показан на рисунке 2.1.

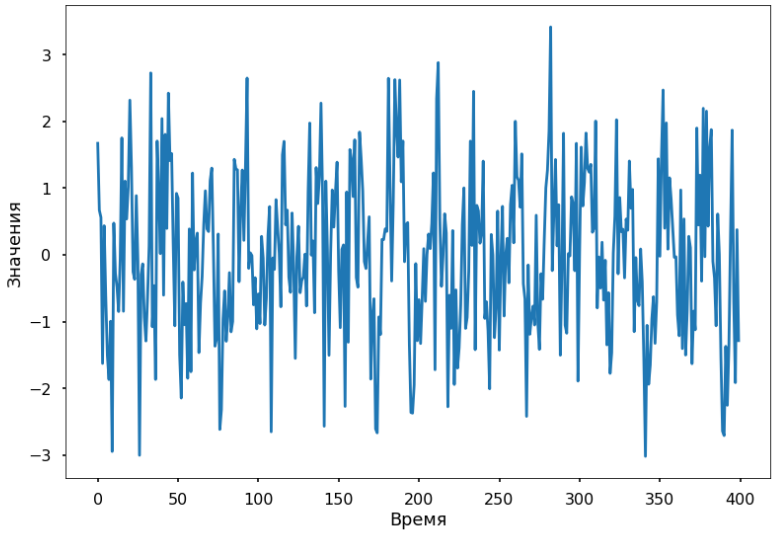

Рис. 2.1. Пример сгенерированного ряда без разладки

Вероятность возникновения разладки выберем $\rho = 0.8$; величину разладки $\delta^* \sim N(\mu = 0, \sigma^2 = 3)$; а отступ с обеих сторон ряда, где разладка не может возникать $n_0 = 48$.

Пример сгенерированного ряда с разладкой показан на рисунке 2.2.

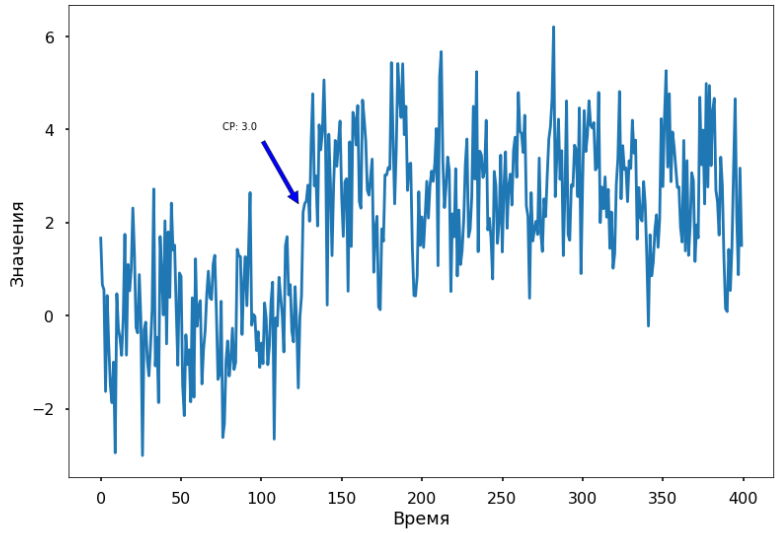

Рис. 2.2. Пример сгенерированного ряда с разладкой

Поскольку реальные данные имеют мультипликативный характер, то мы будем брать экспоненту от сгенерированных с разладкой рядов. На рисунке 2.3 показан пример того же ряда, что и на рисунке 2.2, но со взятой экспонентой от ряда.

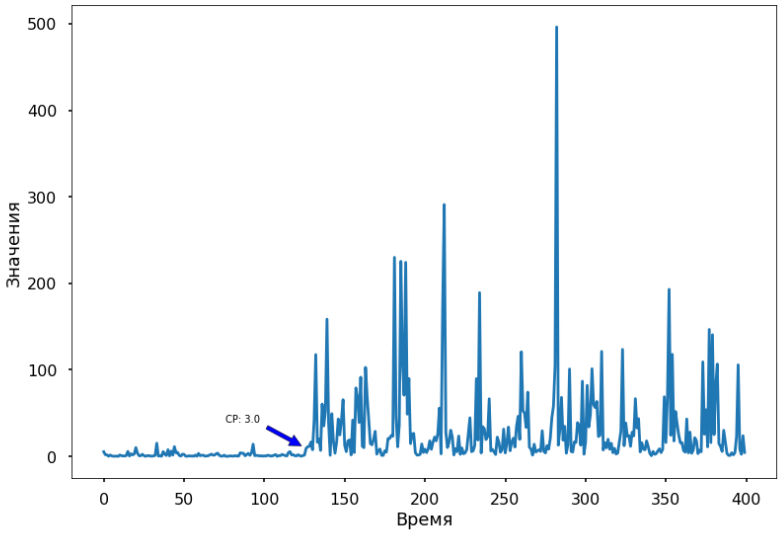

Рис. 2.3. Пример сгенерированного ряда с разладкой и мультипликативностью

Как мы видим, за счет мультипликативности, в некоторых случаях ряд может слишком сильно изменяться — вероятно нужно будет поменять некоторые параметры моделирования ряда, чтобы он стал больше похож на реальный ряд. Поэтому пока начнем с рядов без мультипликативности. В целом, результат похож на реальные ряды, поэтому пока будем

двигаться дальше с такими параметрами моделирования.

В качестве идеи на полях: можно ввести какую-то метрику схожести сгенерированных рядов и реальных.

2.2. Применение методов к моделированным данным и оценка качества этих методов

Попробуем сгенерировать 1000 рядов со случайным местом возникновения разладки τ, случайной величиной разладки δ и случайным шумом e и применить к каждому из этих рядов метод обнаружения разладки на основе аппроксимации. Функцию аппроксимации возьмем константную $\theta = (k)$, ширину окна $l = 48$, а порог γ будем брать в диапазоне от 0 до 1 с шагом 0.01. Пример функции разладки для данного метода с такими параметрами изображен на рисунке 2.4. При этом, в качестве приемлемой задержке возьмем $u = 48$. Это означает, нотификация с запаздыванием в двое суток является для нас приемлемой, что, разумеется, не всегда верно в реальных задачах.

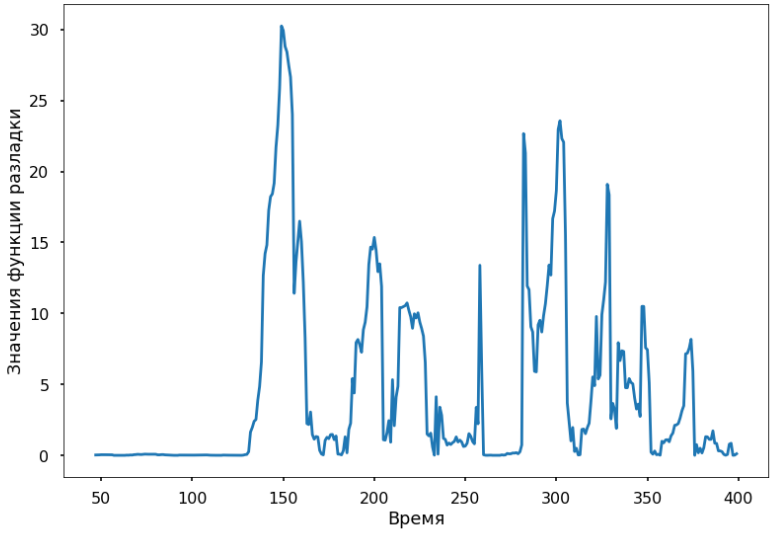

Рис. 2.4. Пример сгенерированного ряда с разладкой и мультипликативностью

Для тысячи рядов с заданными параметрами метода обнаружения разладки и заданным диапазоном порога ROC-кривая получилась следующая (рисунок 2.5). Уже сейчас можно сделать две наблюдения. Во-первых, она проходит довольно далеко от базовой линии, что говорит о том, что в целом метод с такими параметрами работает уже достаточно хорошо (разумеется, надо понимать, что пока что мы взяли приемлемую задержку равную 48 часам,

что является очень мягкими условиями эксперимента, в сравнении с реальной жизнью). Во-вторых, мы видим нетипичное для ROC-кривой поведение в верхнем правом углу: кривая уходит под базовую линию. Дело всё в том, что это происходит только в случае очень низких порогов и связано с моментом обнаружения разладки. Каждый раз после обнаружения разладки в ряде, метод начинает искать следующую разладку начиная с точки последней разладки (как бы отрезая и не учитывая данные, которые были до последней разладки). Это приводит к тому, что если разладка случайно обнаружилась (а при низком пороге это типичная ситуация) незадолго до фактической разладки, то скорее всего сама разладка не будет обнаружена. Этим объясняется почему, при очень низком пороге может быть True positive rate существенно ниже 1.

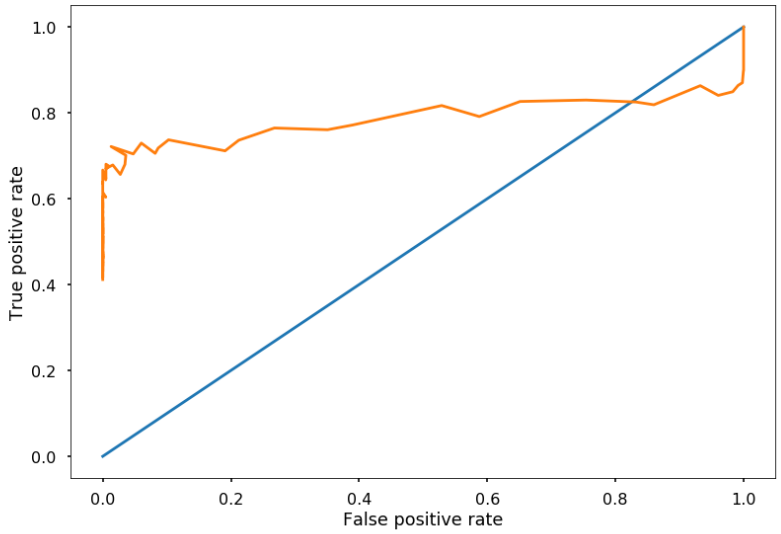

Рис. 2.5. ROC-кривая для 1000 смоделированных рядов и метода на основе аппроксимации

Заключение

В данной работе мы установили подход к моделированию временных рядов данных интернет-рекламы, формально описали некоторые методы обнаружения разладки и описали способ оценки качества методов обнаружения разладки во временных рядах. Во второй части работы, мы применили описанные методы к сгенерированным данным и сравнили качество методов. Данная работа может служить платформой для дальнейших экспериментов как с точки зрения моделирования данных более похожих на реальные, так и с точки зрения оценки качества различных методов обнаружения разладки.

Список литературы

1. Голяндина Н.Э. Метод „Гусеница“-SSA: прогноз временных рядов. Учебное пособие. СПб., 2003.
2. Armstrong. Principles of Forecasting: A Handbook for Researchers and Practitioners. Kluwer Academic Publishers, 2001.
3. Dagum, Estela, Bianconcini. Seasonal Adjustment Methods and Real Time Trend-Cycle Estimation, 2016.
4. N. Golyandina, V. Nekrutkin, A. Zhigljavsky. Analysis of Time Series Structure - SSA and Related Techniques, 2001.
5. R. Hyndman, G. Athanasopoulos. Forecasting: Principles and Practice. 2013.
6. R. Hyndman. Moving averages. 2009.