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
2019

Оглавление

Введение 3

Глава 1. Обнаружение разладки во временных рядах 4

1.1. Построение модели данных 4

1.2. Методы обнаружения разладки 6

1.2.1. Методы на основе аппроксимации 7

1.2.2. Методы на основе прогнозирования 9

1.3. Оценка качества 11

Глава 2. Численные эксперименты 13

2.1. Моделирование данных 13

2.2. Применение методов к моделированным данным и оценка качества этих методов 16

Заключение 19

Список литературы 20

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 отражаются в данных в виде аномальных значений, резких всплесков и внезапных изменений тренда. 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. Построение модели 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.

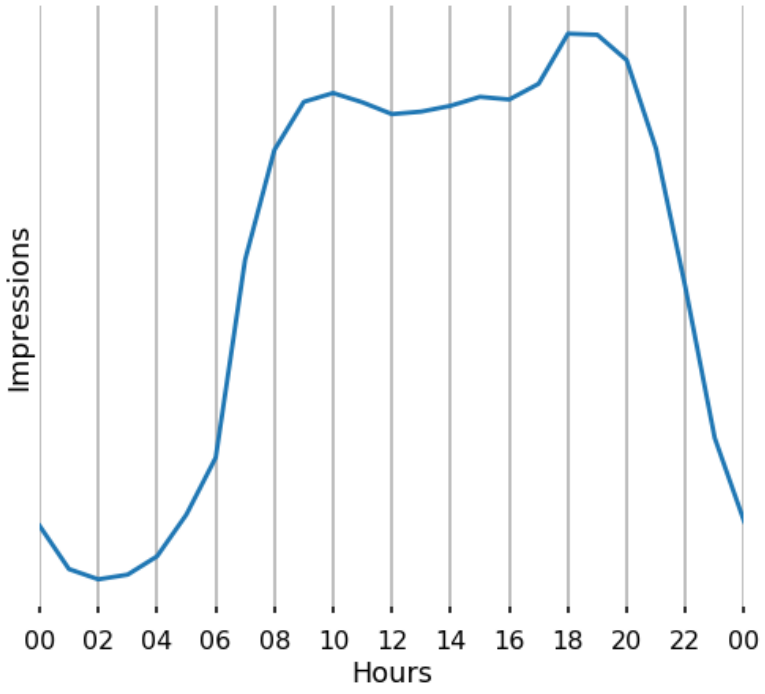

Рис. 1.1. Пример количества показов рекламы за сутки

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

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

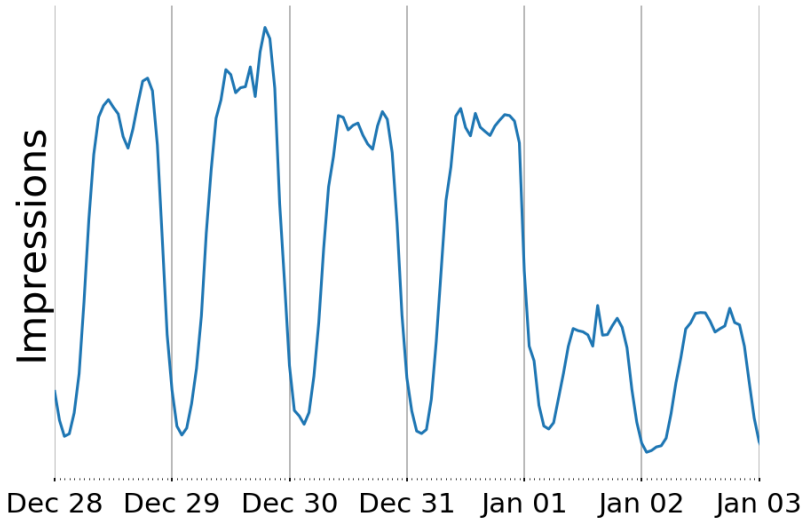

Рис. 1.2. Пример количества показов рекламы с разладкой за неделю

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. Для каждой из этих компонент требуется построить модель. 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.

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

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

Построим 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 обозначим $\tilde{T} = (\tilde{t_1}, \ldots, \tilde{t_n})$, 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. В данной работе значение разладки будет иметь нормальное распределение $\delta^* \sim N(\mu^{(\text{ср})}, \sigma^{2(\text{ср})})$, 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 может являться 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. Однако в данной работе, для упрощения оценки качества методов τ будет не случайной величиной, а фиксированным параметром, то есть $\tau = n_0$.

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 следующие параметры: $\{A\}_{j=1}^{J}, \{a\}_{j=1}^{J}, \{\phi\}_{j=1}^{J}, \mu, \sigma$ а модель разладки имеет еще четыре параметра: $\mu^{(\text{ср})}, \sigma^{(\text{ср})}, \rho, n_0$.

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 прогнозирования
- Методы на основе аппроксимации

1.2.1. Методы на основе аппроксимации

Пусть l — чётное 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): $W^{(\text{left})} = \{w_j^{(\text{left})}\} = \{(y_j, \ldots, y_{j+\frac{l}{2}-1})\}$ 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:

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

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:

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

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, $j = 1, \ldots, k$.

Отметим, что 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 :

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

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

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

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).

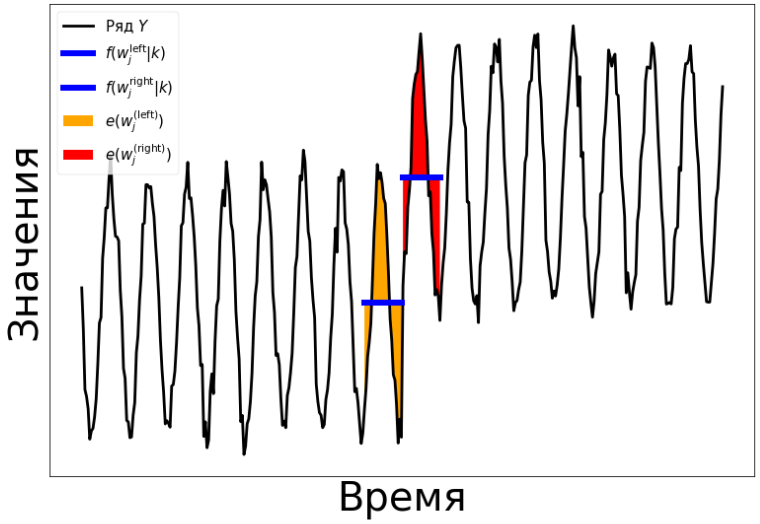

Рис. 1.3. Пример промежуточного расчета ошибки методом аппроксимации

Итого, взяв ряд Y, мы «скользим» по нему окном ширины 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, модель f и порог γ.

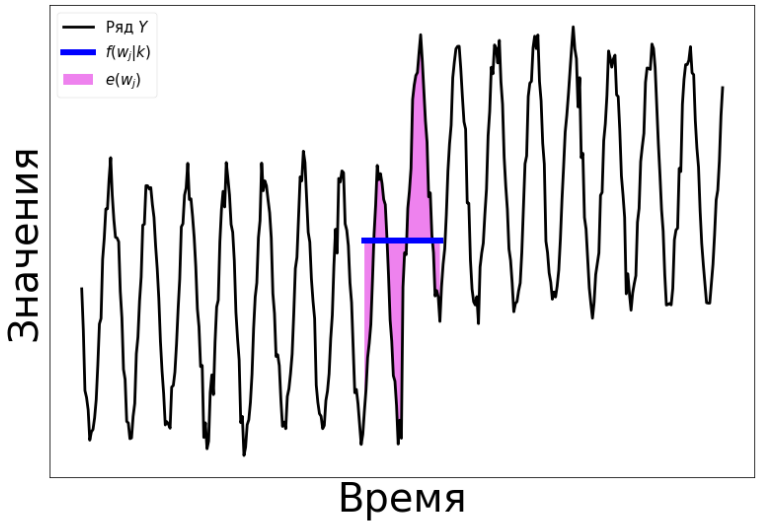

Рис. 1.4. Пример промежуточного расчета ошибки методом аппроксимации, продолжение.

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 $e(\cdot)$ на ряде $w_j^{(\text{right})}$. При этом функция разладки принимает следующий вид:

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

В остальном данные методы ничем не отличаются от методов на основе аппроксимации. Следовательно, синхронизированная функция разладки q_i синхронизируется с исходным рядом аналогичным образом.

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 $F(w_j^{(\text{right})})$ для всего ряда.

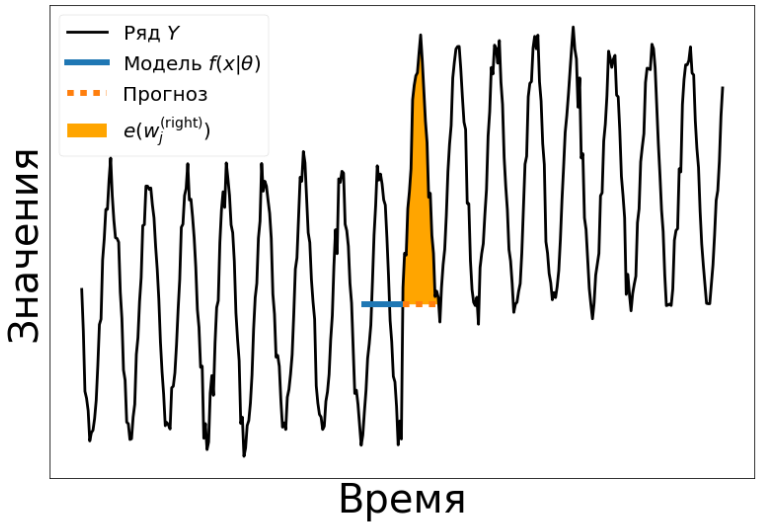

Рис. 1.5. Пример расчета ошибки методом прогнозирования

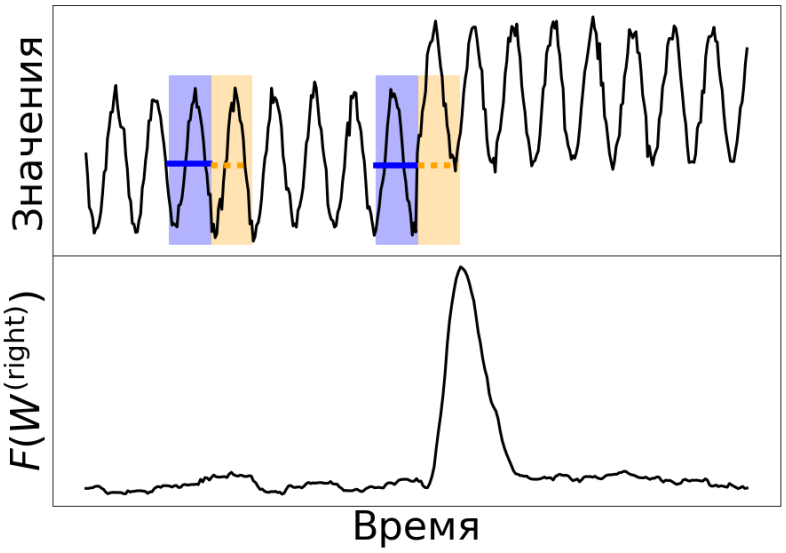

Рис. 1.6. Пример расчета функции разладки с помощью скользящего окна

В методах прогнозирования нужно задавать следующие параметры: ширину окна l, модель 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.

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 важно обнаружить разладку не позднее какого-то срока, иначе оповещение о разладке будет несвоевременным. Для простоты оценки качества методов мы фиксируем точку разладки τ параметром n_0, тем самым фиксируя приемлемую задержку обнаружения разладки на уровне $n - n_0$.

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.

Договорившись о таком способе оценки качества, 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. ROC-AUC удобно использовать, поскольку она удобно отражает качество метода одним числом. Но помимо оценки самого ROC-AUC нам бы хотелось оценить доверительный интервал в который попадает ROC-AUC с заданным уровнем значимости. В статье [1] предлагают способ оценки стандартного отклонения ROC-AUC. Оценка эта исходит из того, что для больших выборок значение ROC-AUC имеет нормальное распределение. Поэтому доверительный интервал с уровнем доверия $1-\alpha$

можно посчитать следующим образом:

$$AUC \pm z_{\frac{\alpha}{2}}\sigma(AUC),$$

где z — стандартизованная оценка, а $1-\alpha$ — уровень значимости. Способ оценки стандартного отклонения предложен авторами статьи:

$$\sigma(AUC) = \sqrt{\frac{AUC(1-AUC) + (N_p - 1)(Q_1 - AUC^2) + (N_n - 1)(Q_2 - AUC^2)}{N_p N_n}},$$

где N_p — количество положительных наблюдений в выборке (в нашем случае количество рядов с разладкой), N_n — количество негативных наблюдений в выборке (количество рядов без разладки),

$$Q_1 = \frac{AUC}{2 - AUC},$$
$$Q_2 = \frac{2AUC^2}{1 + AUC}.$$

Глава 2

Численные эксперименты

2.1. Моделирование данных

На рисунке 2.2 представлен пример реальных почасовых данных показов рекламы за пять недель. В рамках суток данные имеют два типа структуры — структуру буднего дня (1.1) и структуру выходного дня (2.3). Следовательно, моделировать будние и выходные дни лучше отдельно, в этой работе мы сфокусируемся на выходных днях, однако аналогичные действия применимы и к будним дням. Попробуем убрать из реальных данных будние дни и "склеить"выходные дни в один ряд. Получившийся ряд изображен на рисунке 2.4 Мы можем смоделировать данный ряд используя подход, описанный в разделе 1.1. Процедура оценки параметров SSA определила следующие параметры модели 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:

$$s_i = 1.05\cos(\frac{2\pi}{24}i+\frac{3\pi}{4})+0.82\cos(\frac{2\pi}{12}i+\frac{\pi}{12})+0.27\cos(\frac{2\pi}{8}i-\frac{2\pi}{3})+0.05\cos(\frac{2\pi}{6}i-\frac{\pi}{3}), \quad i = 1, \dots, n.$$

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. На рисунке 2.1 изображено сравнение реальных данных (из которых предварительно вычли среднее значение и отнормировали) и смоделированных по модели, описанной выше.

Таким образом, модель для генерации ряда имеет следующий вид:

$$y_i = s_i + N(0, 0.1).$$

Вероятность возникновения разладки выберем $\rho = 0.8$; величину разладки $\delta^* \sim N(\mu = 1, \sigma = 0.4)$; а место возникновения разладки зададим в самом конце ряда $n_0 = 396$, так, чтобы допустимая задержка в обнаружении разладки была равна четырем часам.

Пример сгенерированного ряда с разладкой показан на рисунке 2.5.

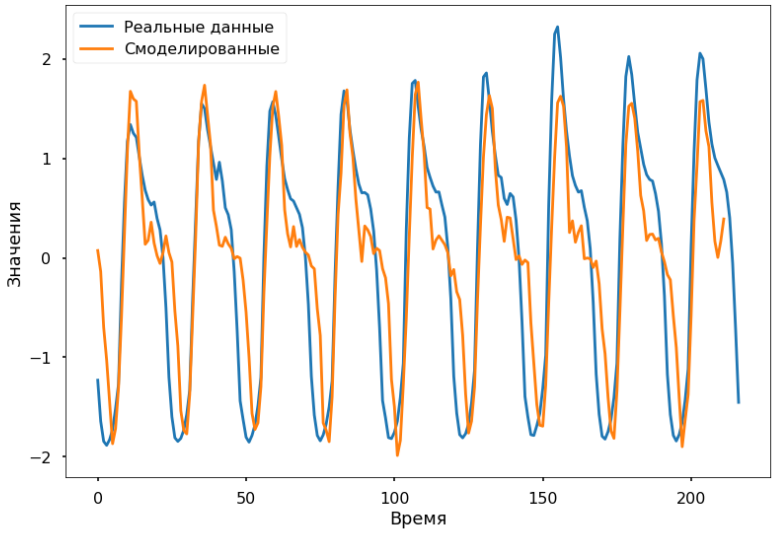

Рис. 2.1. Сравнение реальных отнормированных данных и смоделированных

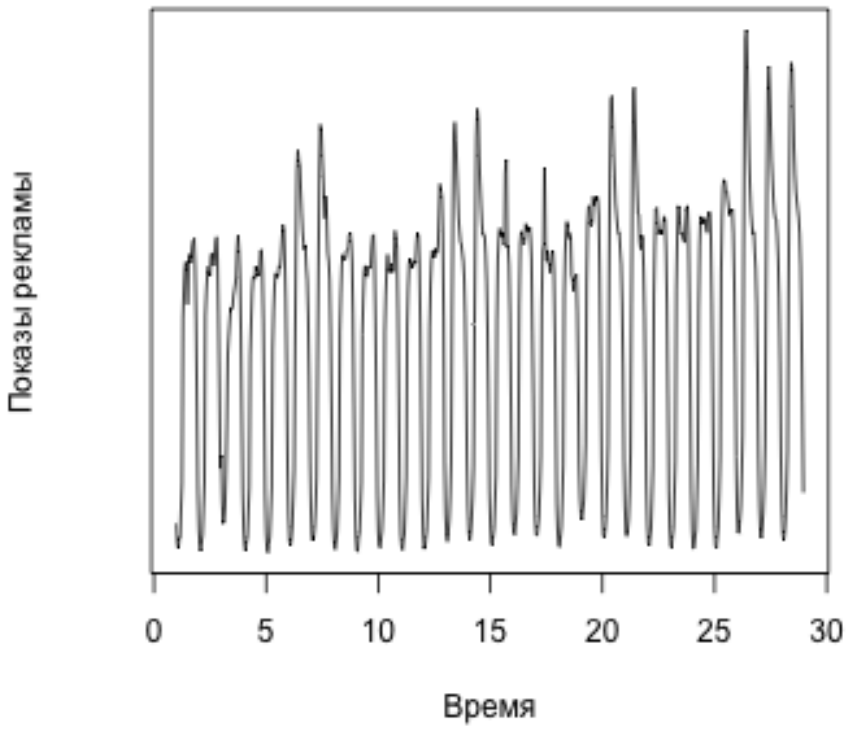

Рис. 2.2. Пример реальных данных показов рекламы

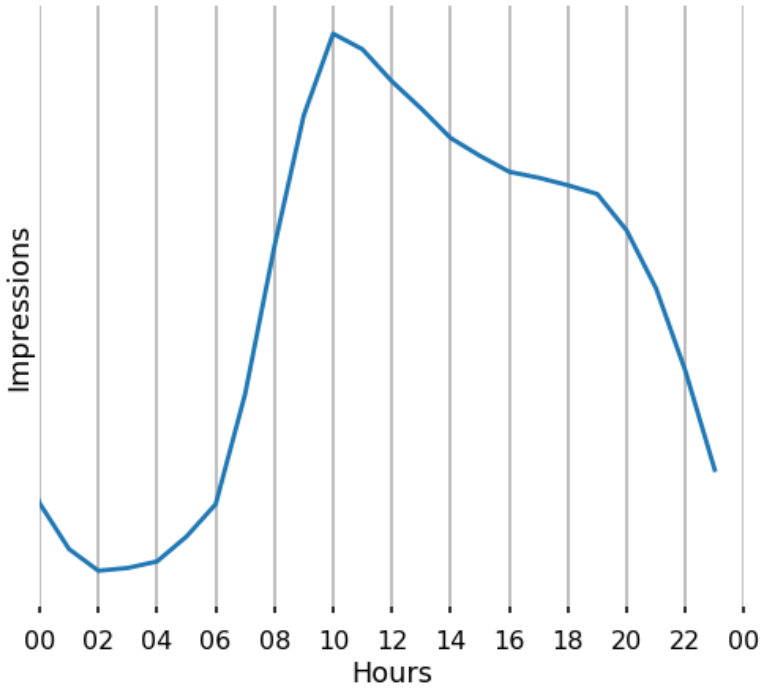

Рис. 2.3. Пример количества показов рекламы в выходной день

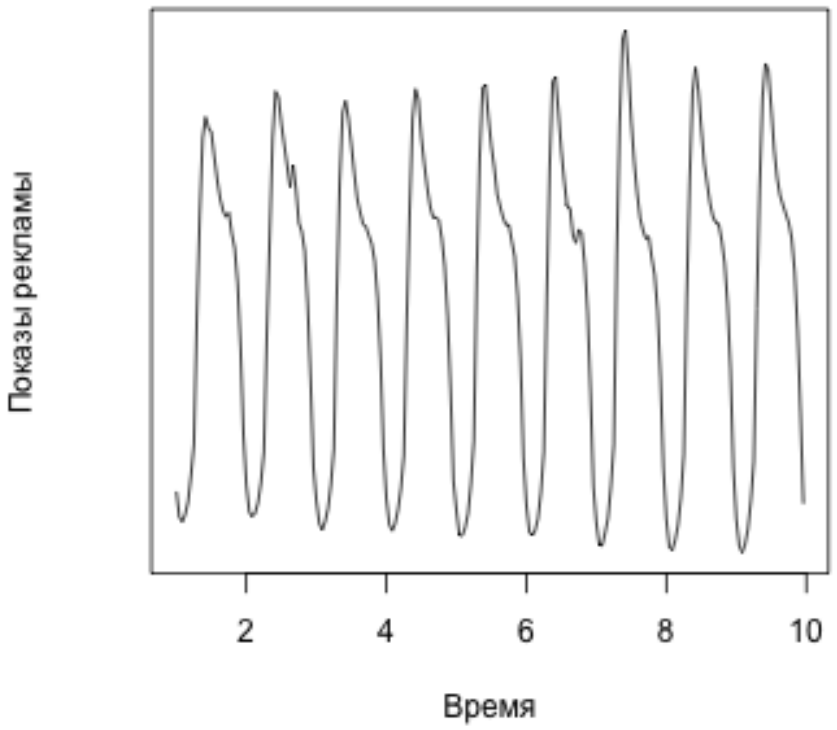

Рис. 2.4. Реальные данные показов рекламы без будних дней

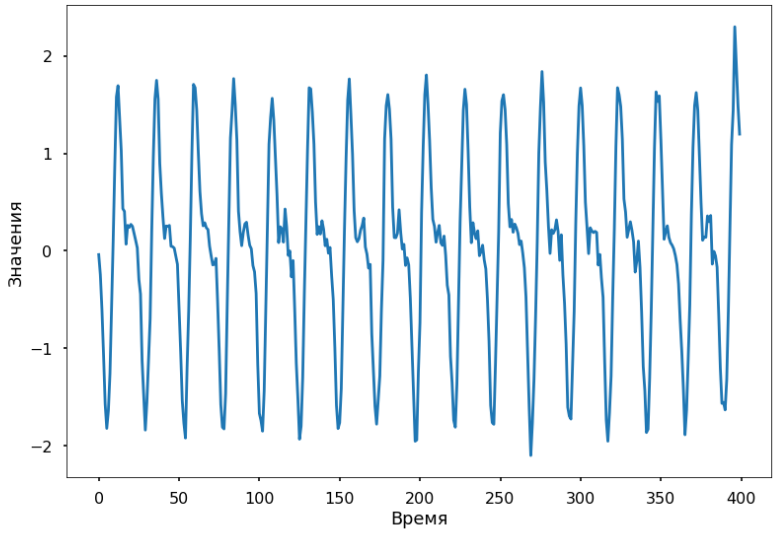

Рис. 2.5. Пример сгенерированного ряда с разладкой

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

Сгенерируем 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 $\theta = (b)$, 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 2.6. 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.

Для тысячи рядов с заданными параметрами метода обнаружения разладки и заданным диапазоном порога ROC-кривая получилась следующая (рисунок 2.7). ROC-кривая 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).

Попробуем применить метод обнаружения разладки на основе прогнозирования к такой же тысяче рядов. Параметры метода будем брать аналогичными параметрам аппроксимации (кроме порога). Функция прогнозирования константная $\theta = (b)$, ширина окна $l = 48$, индекс разделения окна $g = 24$, порог γ будем брать в диапазоне от 0 до 5 с шагом 0.05.

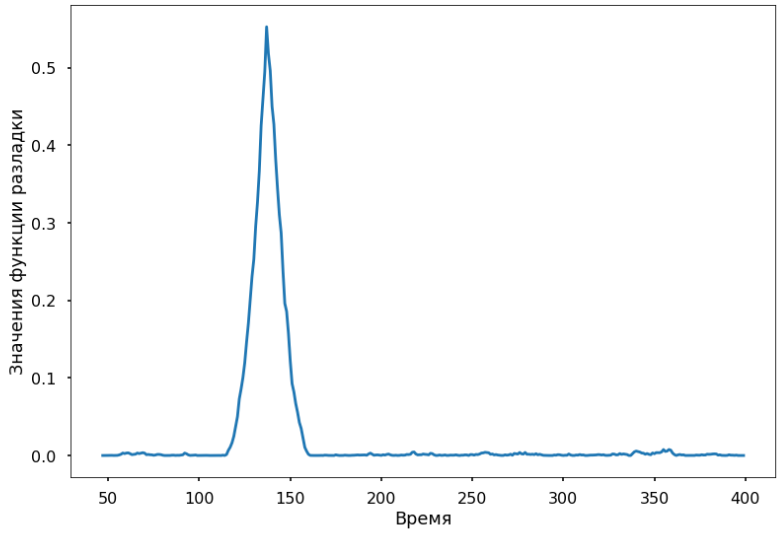

Рис. 2.6. Пример функции разладки для сгенерированного ряда с разладкой и без мультипликативности

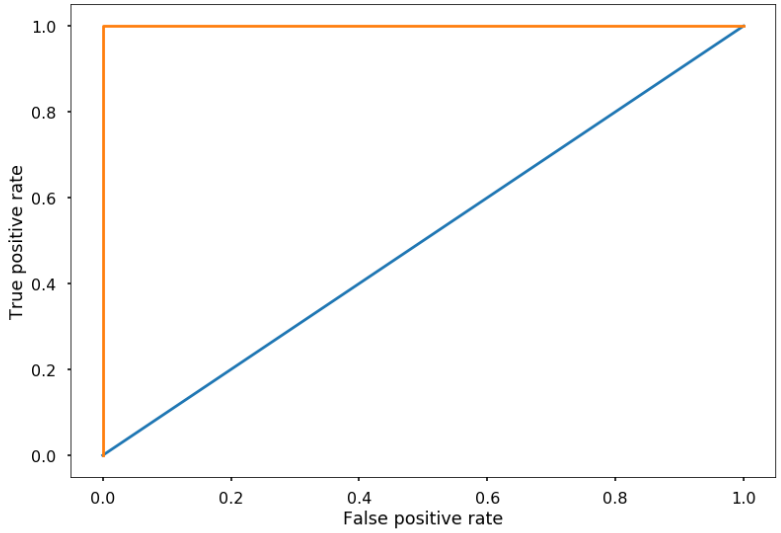

Рис. 2.7. ROC-кривая для 1000 смоделированных рядов и метода на основе аппроксимации

Получившийся результат представлен на рисунке 2.8. Как видно из ROC-кривых с данными параметрами метод на основе аппроксимации сработал лучше, нежели метод на основе прогнозирования. Безусловно, требуется настройка параметров, как для моделирования данных, так и для самих методов, но цель работы — создать стенд для проверки идей и гипотез по обнаружению разладок — достигнута, поскольку нам удалось сравнить качество двух разных методов в контролируемом эксперименте.

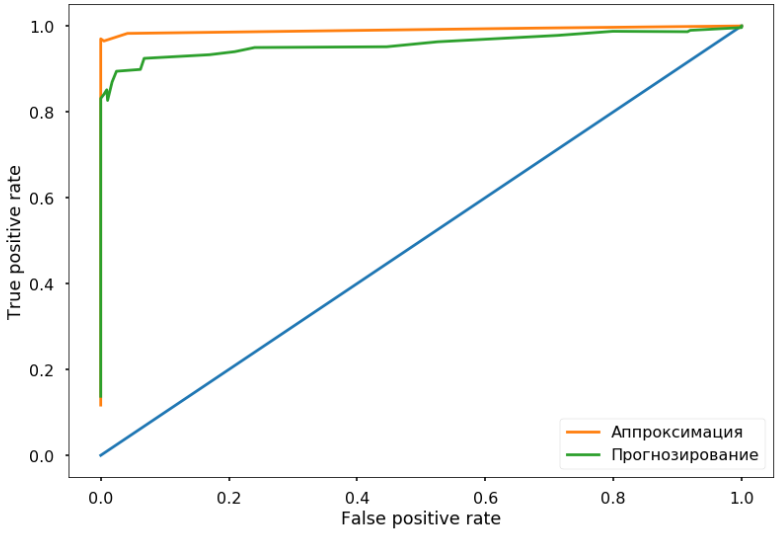

Рис. 2.8. ROC-кривые для 1000 смоделированных рядов

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. J.A. Hanley B.J. McNeil. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve // Radiology. 1982. April. T. 143, № 1. C. 29.
2. Athanasopoulos Rob J Hyndman; George. Forecasting: principles and practice. OTexts, 2013.
3. Nina Golyandina Vladimir Nekrutkin Anatoly A Zhigljavsky. Analysis of Time Series Structure - SSA and Related Techniques. Chapman and Hall/CRC, 2001.
4. Estela Bee Dagum Silvia Bianconcini. Seasonal Adjustment Methods and Real Time Trend-Cycle Estimation. Springer International Publishing, 2016.
5. Armstrong J. Scott. Principles of Forecasting: A Handbook for Researchers and Practitioners. Kluwer Academic Publishers, 2001.
6. Н.Э. Голяндина. Метод „Гусеница“-SSA: прогноз временных рядов. СПбГУ, 2003.