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

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 главной задачей данной работы является разработка методики обнаружения разладки. В работе будут использоваться фактические данные одной из работающих рекламных сетей для построения модели данных.

Цель работы — сравнить методы обнаружения разладки на модельных данных, по виду и структуре похожих на реальные, и выработать рекомендации по использованию методов обнаружения разладки для реальных данных. В связи с этим, структура работы следующая: в первой главе описываются способы оценки параметров модели ряда, предлагается модель сигнала ряда, а также способ моделирования разладок разных видов. Помимо этого в первой

главе описываются методы обнаружения разладки и предлагается способ оценки качества этих методов. Во второй, практической, главе происходит подбор параметров модели сигнала в реальных данных, генерируются искусственные ряды, применяются методы описанные в первой главе к этим рядам, сравнивается качество этих методов с рекомендациями в выводах.

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 почасовой 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.

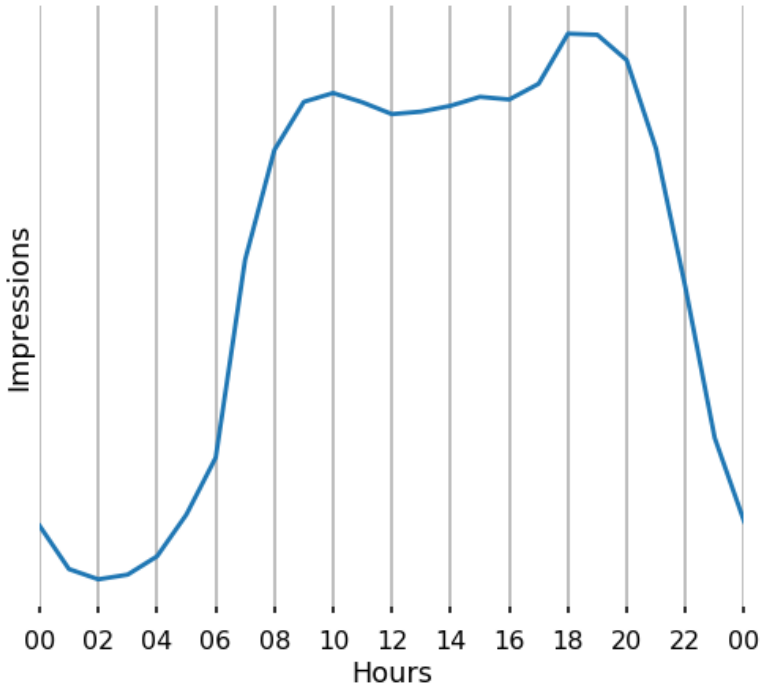

Рис. 1.1. Пример количества показов рекламы за сутки

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

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

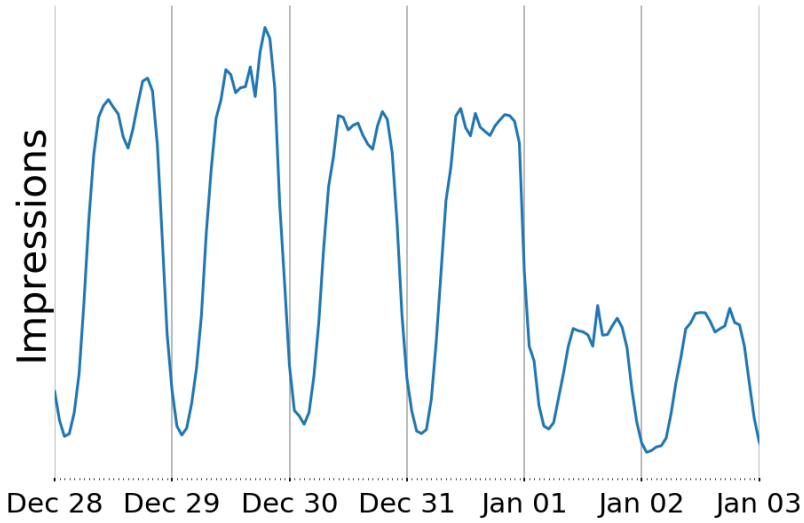

Рис. 1.2. Пример количества показов рекламы с разладкой за неделю

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:

$$t_i = c, \quad i = 1, \ldots, n,$$
$$s_i = \sum_{j=1}^{J} A_j \cos\left(\frac{2\pi}{a_j} i + \phi_j\right), \quad i = 1, \ldots, n,$$
$$\epsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n,$$

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

В данной работе мы будем проверять качество методов для двух наиболее часто встречающихся типов разладки — разладка в среднем и локальная разладка (в одной точке).

Построим модель разладки, исходя из следующего:

- Разладка только в одной точке ряда;
- Разладка заключается в сдвиге;

Формально это можно описать так: пусть τ — точка (индекс) разладки, тогда тренд с разладкой обозначим $\tilde{T} = (\tilde{t_1}, \ldots, \tilde{t_n})$. В случае разладки в среднем $\tilde{t_i}$:

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

где $\delta^{(mean)}$ — значение разладки. Чтобы разладка была заметна, введем ещё минимальное допустимое значение разладки $\delta_{min}^{(mean)}$.

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

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

В случае локальной разладки модель разладки будет выглядеть следующим образом:

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

Со схожим ограничением на минимальное значение разладки:

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

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

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

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

В результате модель временного ряда имеет следующие параметры: $\{A\}_{j=1}^{J}$, $\{a\}_{j=1}^{J}$, $\{\phi\}_{j=1}^{J}$, σ, c, а модель разладки имеет еще три параметра: $\mu^{(\text{cp})}, \sigma^{(\text{cp})}, \delta_{min}$.

1.2. Оценка параметров модели

В [**?**] описан подход к определению параметров модели временного ряда $\{A\}_{j=1}^{J}$, $\{a\}_{j=1}^{J}$, $\{\phi\}_{j=1}^{J}$, μ, σ, c основанный на траекторной матрице ряда. Предполагая, что периодическая компонента ряда управляется линейной рекуррентной формулой, можно записать её в виде $s_i = \sum_{j=1}^{r} c_j \mu_j^n$.

Оценить сигнальные параметры μ_j позволяет метод, который называется ESPRIT [**?**].

Обозначим $\{U_1, \ldots, U_r\}$ ортонормированный базис оцениваемого подпространства, интересующей компоненты. Такой базис можно оценить, например, с помощью метода SSA, где r будет параметром, обозначающий каким количеством компонент мы оцениваем сигнал временного ряда.

Обозначим $\mathbf{P}_r = [U_1 : \cdots : U_r]$, а $\underline{\mathbf{P}_r}$ матрица без последней строки, а $\overline{\mathbf{P}_r}$ матрица без первой строки. Тогда μ_i может быть оценено с помощью собственных значений матрицы $\underline{\mathbf{P}_r}^{\dagger}\overline{\mathbf{P}_r}$, где $\dagger$ обозначает псевдо-инверсию. Соответственно, оцениваемые частоты и являются аргументами μ_i.

Для упрощения модели, в работе мы исправляем корни μ_j, заменяя их на корни с модулем 1 и точными периодами. Получив оценки частот, можно взять сумму косинусов и синусов с этими частотами как модель и с помощью МНК подобрать параметры амплитуд. Далее, каждую пару синусов и косинусов с одинаковыми частотами можно сгруппировать в один косинус следующим образом:

$$a_1\cos(\omega t) + a_2\sin(\omega t) = a_1\cos(\omega t) + a_2\cos(\omega t - \frac{\pi}{2}) = \sqrt{a_1^2 + a_2^2}\cos(\omega t - \operatorname{arctg}(\frac{a_1}{a_2})).$$

1.3. 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

Следует обратить внимание, что исходная модель ряда одна:

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

При этом оба типа методов в своей основе будут иметь ту или иную модель ряда $f(x|\theta)$ сигнала для обнаружения разладки, где θ — параметры модели. Этих моделей ряда может быть много. Модель может быть константной ($\theta = (b)$):

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

либо другой подходящей под наш ряд функцией, один косинус:

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

либо более сложной функцией — суммой косинусов:

$$f(x|\{P_j, p_j, \chi_j\}, b) = \sum_{j=1}^{J} P_j \cos(\frac{2\pi}{p_j}x + \chi_j) + b,$$

и для отдельной проверки модель с трендом:

$$f(x|c, b) = cx + b.$$

1.3.1. 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 отрезков $W = \{w_j\}_{j=1}^{k}$, где $k = n - l + 1$ — количество таких отрезков; а $w_j = (y_j, \ldots, y_{j+l-1})$ — j-ый отрезок. Каждый отрезок w_j в свою очередь делится на два отрезка 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.

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 :

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

Нормирующая константа h нам необходима для того, чтобы приводить значения функции разладки к значениям близким к диапазону от 0 до 1, поскольку без нормировки значения функции разладки могут принимать сколь угодно большие значения. Подход к расчету 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):

$$h = e(w_1) - e(w_1^{(\text{left})}) - e(w_1^{(\text{right})}) + 1.$$

Либо, можно проходить скользящим окном начало ряда и рассчитывать ненормированное значение функции разладки для каждого окна и затем усреднять полученные значения. Такой подход помогает избежать ситуаций с выбранным малым окном, когда мы не охватываем всю периодичность одним окном.

$$h = \frac{\sum_{j=1}^{v} e(w_j) - e(w_j^{(\text{left})}) - e(w_j^{(\text{right})})}{v} + 1,$$

где v — сколько отрезков в начале ряда мы используем для расчета h.

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). А на рисунке 1.5 изображен пример расчета функции разладки с помощью скользящего окна.

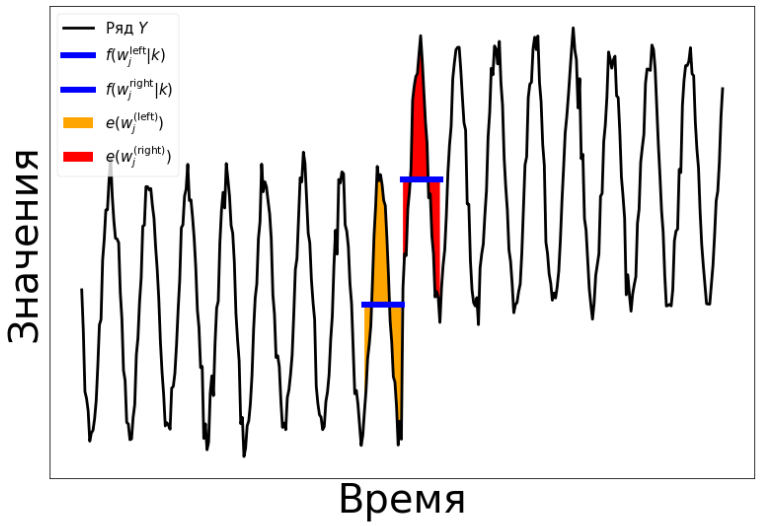

Рис. 1.3. Пример промежуточного расчета ошибки методом аппроксимации

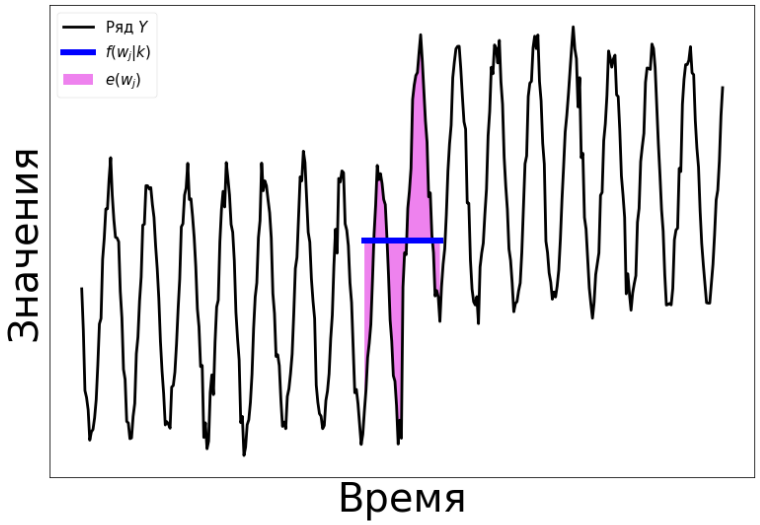

Рис. 1.4. Пример промежуточного расчета ошибки методом аппроксимации, продолжение.

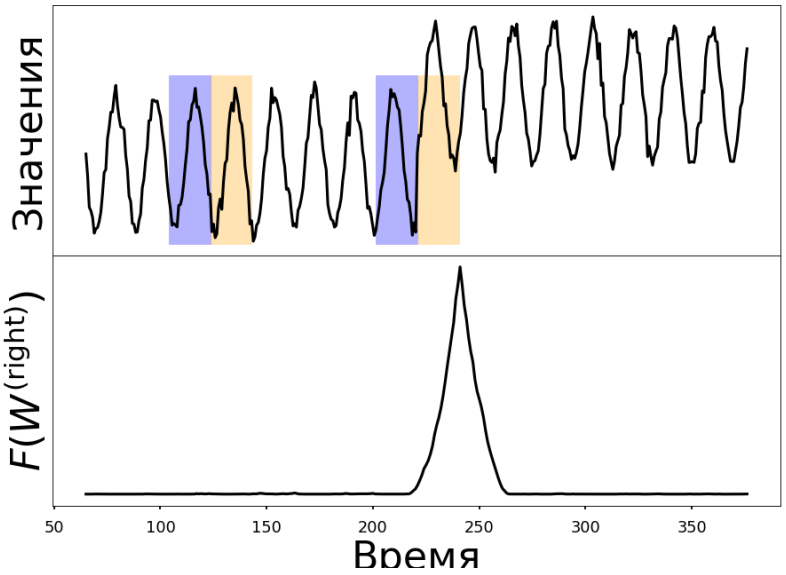

Рис. 1.5. Пример расчета функции разладки с помощью скользящего окна

Итого, взяв ряд Y, мы «скользим» по нему окном ширины l и рассчитываем значения функции разладки $F()$ для каждого из получаемых отрезков $W^{(\text{all})}$. 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.

1.3.2. 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 отрезков $W = \{w_j\}_{j=1}^{k}$. Каждый отрезок 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: $W^{(\text{left})} = \{w_j^{(\text{left})}\} = (y_j, \ldots, y_{j+g-1})$ 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 мер ошибок на тех же рядах, на которых подбирались параметры моделей (параметры модели и ошибка для левого отрезка оценивались на левом отрезке и так для всех трех моделей $w_j^{(all)}$), 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 1.6 приведен пример расчета ошибки с помощью метода прогнозирования на одном ряде $w_j^{(\text{right})}$. А на рисунке 1.7 показан пример расчета функции разладки $F(w_j^{(\text{right})})$ для всего ряда.

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.

1.4. 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

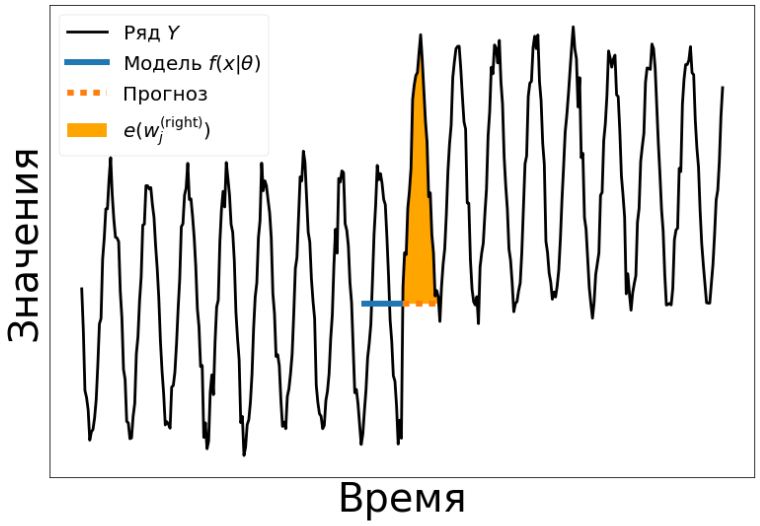

Рис. 1.6. Пример расчета ошибки методом прогнозирования

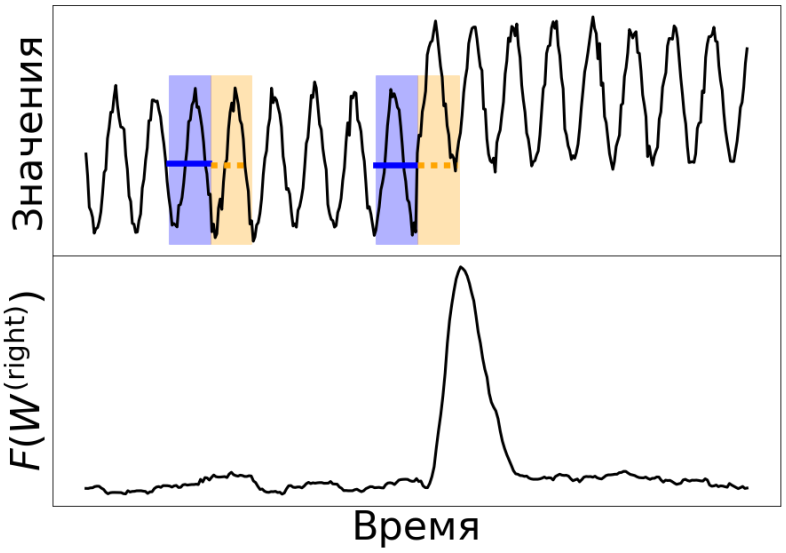

Рис. 1.7. Пример расчета функции разладки с помощью скользящего окна

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 ошибок классификации и считать метрики качества. При этом важно обнаружить разладку не позднее какого-то срока, иначе оповещение о разладке будет несвоевременным. Поэтому мы фиксируем точку разладки τ и вводим параметр допустимой задержки d.

Следовательно, в рамках данной работы мы будем строить классифицирующее правило:

$$a(Y) = \begin{cases} 1, & f_j \geq \gamma \text{ и } \tau - l \leq j \leq \tau - l + d - 1, \\ 0, & \text{иначе,} \end{cases}$$

где 0 означает, что разладки нет (negative), 1 означает, что разладка есть (positive). Таким образом, метод будет определять, либо не определять разладку в заданном диапазоне в зависимости от задаваемого порога γ.

Исходя из этого возможны четыре варианта:

- Разладка произошла и метод обнаружил точку разладки в диапазоне $(\tau, \cdots, \tau + d - 1)$. Такая ситуация попадает под категорию True positive.
- Разладка произошла и метод не обнаружил точку разладки в диапазоне $(\tau, \cdots, \tau + d - 1)$. Это случай False negative.
- Метод обнаружил разладку в диапазоне $(\tau, \cdots, \tau + d - 1)$ в ряде без разладки. Это ситуация False positive.
- Разладки не было и метод не обнаружил разладку в диапазоне $(\tau, \cdots, \tau + d - 1)$. 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.

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 γ. 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.

$$TPR(\gamma) = \frac{\text{Верно-положительные классификации}}{\text{Все положительные наблюдения}},$$

$$FPR(\gamma) = \frac{\text{Ложно-отрицательные классификации}}{\text{Все отрицательные наблюдения}}.$$

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 [**?**] 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 (рисунок 1.1) и структуру выходного дня (рисунок 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 $a_1 = 23.93 \approx 24$, $a_2 = 11.99 \approx 12$, $a_3 = 7.99 \approx 8$, $a_4 = 5.99 \approx 6$ (это логично, поскольку наши данные имеют суточные колебания). Оценка амплитуд данным методом получилась $A_1 = 1.00$, $A_2 = 0.39$, $A_3 = 0.13$, $A_4 = 0.11$. А фазы косинусов получились $\phi_1 = 2.78 \approx \frac{8\pi}{9}$, $\phi_2 = 1.55 \approx \frac{\pi}{2}$, $\phi_3 = -1.56 \approx -\frac{\pi}{2}$, $\phi_4 = -2.95 \approx -\frac{15\pi}{16}$. Таким образом, модель периодической составляющей s_i нашего ряда можно записать в следующем виде:

$$s_i = 1.00\cos(\frac{2\pi}{24}i+\frac{8\pi}{9})+0.39\cos(\frac{2\pi}{12}i+\frac{\pi}{2})+0.13\cos(\frac{2\pi}{8}i-\frac{\pi}{3})+0.11\cos(\frac{2\pi}{6}i-\frac{15\pi}{16}), \quad i = 1, \dots, n.$$

Длину ряда зафиксируем $n = 400$. Значение тренда 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:

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

Моделировать будем 100 рядов (50 с разладкой и 50 без). Величины разладки $\delta^{(mean)*} \sim N(\mu = 0, \sigma = 0.2)$, $\delta^{(local)*} \sim N(\mu = 0, \sigma = 1)$. Минимальные допустимые значения разладок выберем $\delta^{(mean)*}_{min} = 0.3$, $\delta^{(local)*}_{min} = 0.5$. Место возникновения разладки зададим в середине ряда $\tau = 216$, а допустимых задержек выберем несколько $d = (4, 24, 48)$.

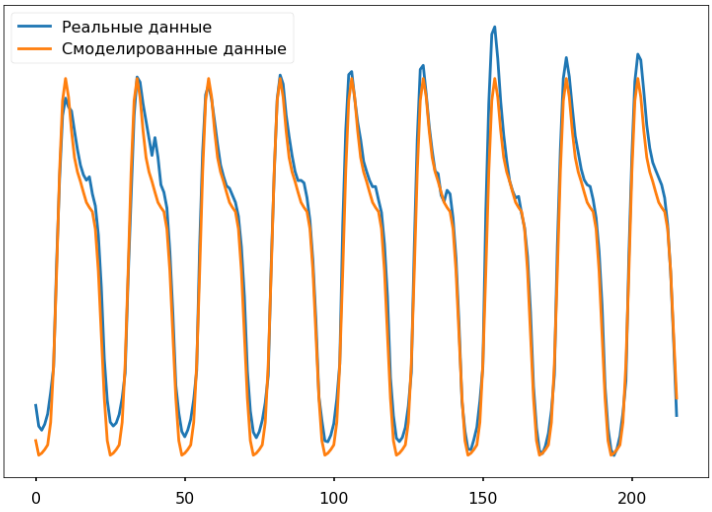

Рис. 2.1. Сравнение реальных отнормированных данных и смоделированных

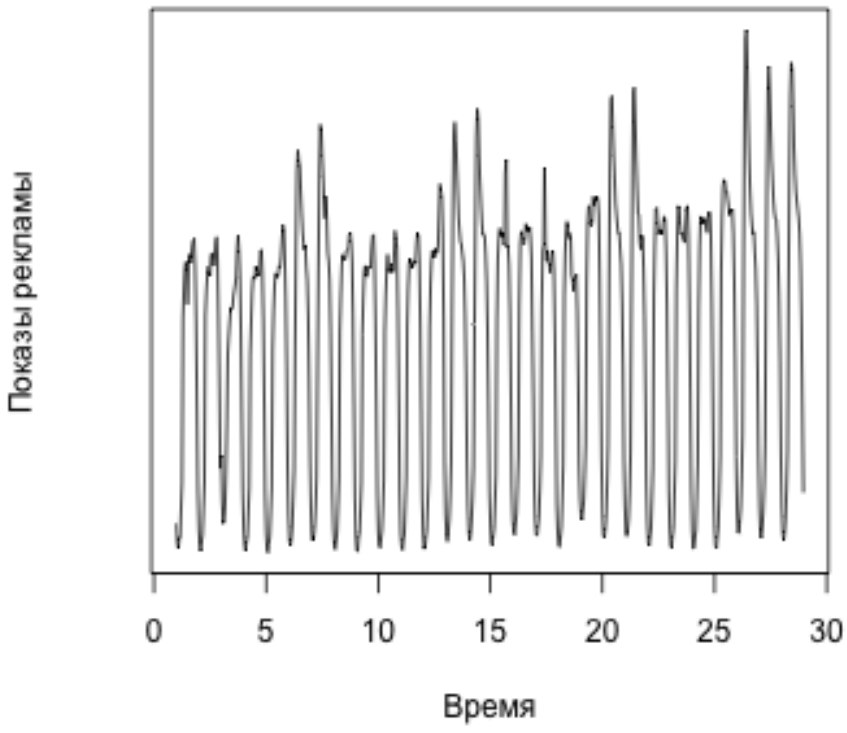

Рис. 2.2. Пример реальных данных показов рекламы

Пример сгенерированного ряда с разладкой показан на рисунке 2.5.

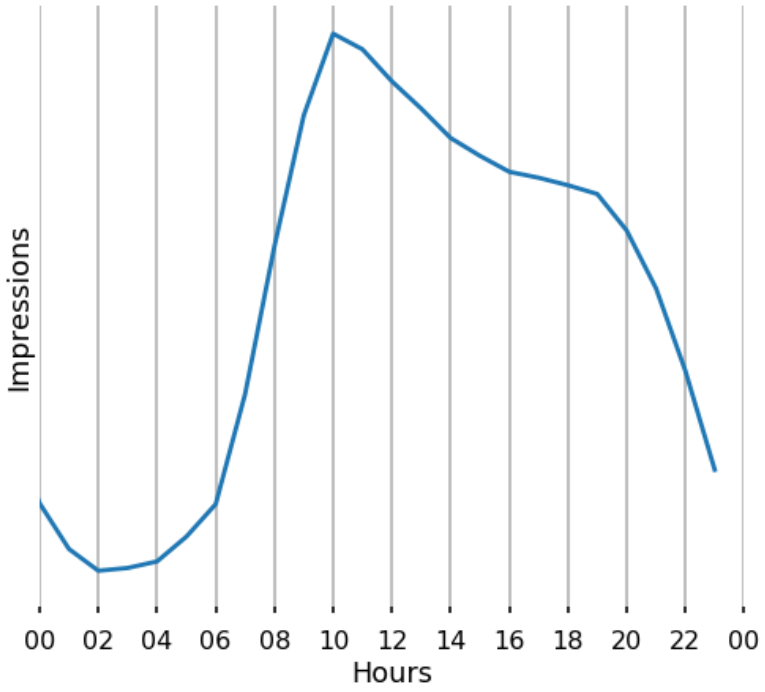

Рис. 2.3. Пример количества показов рекламы в выходной день

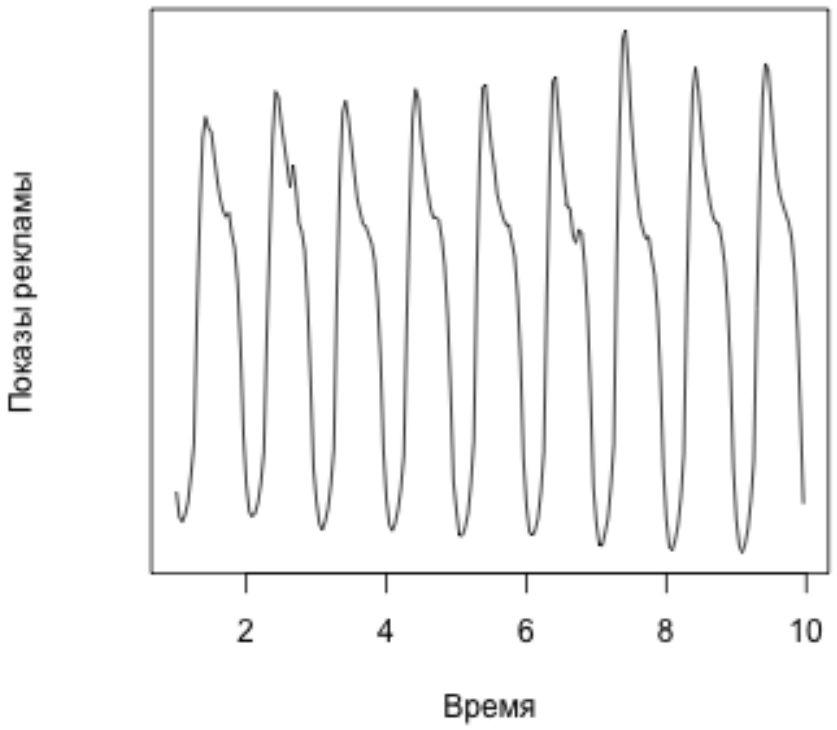

Рис. 2.4. Реальные данные показов рекламы без будних дней

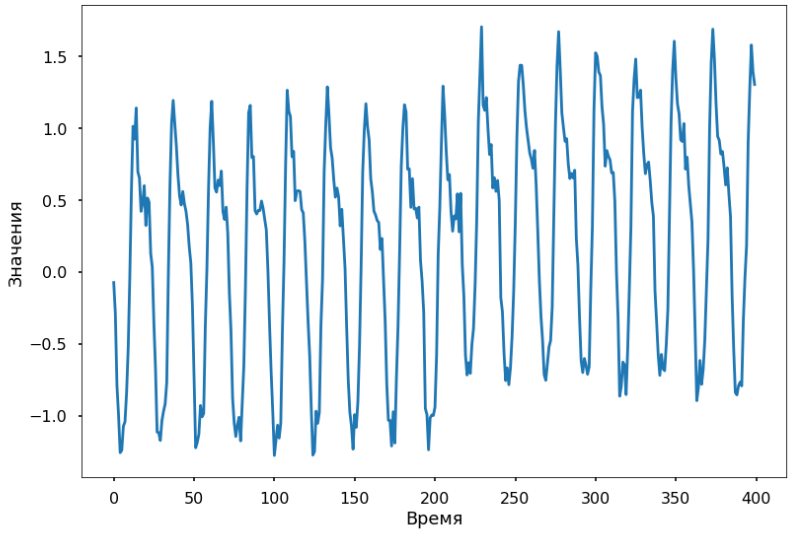

Рис. 2.5. Пример сгенерированного ряда с разладкой

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

Сгенерируем 50 рядов с локальной разладкой, 50 рядов с разладкой в среднем и 100 рядов без разладки. К каждому из этих рядов применим каждый из 7 методов (аппроксимация с моделями „среднее“, „косинус“, „4 косинуса“, „тренд“ и прогнозирование с моделями „среднее“, „косинус“, „4 косинуса“) 15 раз (ко всем комбинациям из задержек $d = (4, 24, 48)$ и длин окон $l = (2, 4, 24, 48, 96)$. Результат будем оценивать по метрике ROC-AUC с 95%-м доверительным интервалом. На рисунке 2.6 изображены результаты ROC-AUC для всей решетки параметров и методов.

Тип разладки	локальная															в среднем														
Задержка (d)	4					24					48					4					24					48				
Длина окна (l)	2	4	24	48	96	2	4	24	48	96	2	4	24	48	96	2	4	24	48	96	2	4	24	48	96	2	4	24	48	96
Место разладки	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216
Метод																														
approximation_mean	0,98	0,76	0,50	0,77	0,67	0,95	0,64	0,56	0,83	0,71	0,92	0,67	0,63	0,86	0,71	0,67	0,58	0,52	0,82	0,72	0,56	0,54	0,92	1,00	1,00	0,60	0,53	0,92	1,00	1,00
approximation_sin				0,93	0,82				0,95	0,81				0,94	0,85				0,95	0,85				1,00	1,00				1,00	1,00
approximation_sin_insight				0,97	0,88				0,94	0,91				0,98	0,92				0,99	0,86				1,00	1,00				1,00	1,00
approximation_trend				0,58	0,50				0,53	0,56				0,56	0,52				0,50	0,54				0,55	0,52				0,77	0,52
prediction_mean	0,95	0,80	0,59	0,69	0,70	0,93	0,66	0,57	0,82	0,71	0,93	0,67	0,61	0,75	0,73	0,73	0,58	0,54	0,67	0,59	0,51	0,47	0,86	0,99	0,94	0,55	0,48	0,81	0,96	0,98
prediction_sin				0,87	0,77				0,94	0,92				0,90	0,76				0,70	0,65				1,00	0,99				1,00	1,00
prediction_sin_insight				1,00	0,98				0,98	0,99				0,99	0,98				0,99	0,96				1,00	1,00				1,00	1,00

Рис. 2.6. ROC AUC для разных методов

Как мы видим метод аппроксимация с моделью „тренд“ сработал очень плохо для любых комбинаций параметров. Отсюда мы делаем первый вывод: если добавить в модель метода лишний, несвязанный с исходными данными компонент, то он испортит весь метод. Далее модель с трендом не будем рассматривать за ненадобностью.

Теперь сравним какой подход работает лучше: аппроксимация или прогнозирования. На рисунках 2.7, 2.8 изображены сравнительные графики ROC-AUC с 95%-ми доверительными интервалами. Как мы видим, для любых комбинаций моделей и параметров доверительные интервалы аппроксимации и прогнозирования пересекаются. Это означает, что оба подхода работает примерно одинаково для условий нашей задачи. Поэтому далее будем рассматривать только один из этих подходов, например, аппроксимацию.

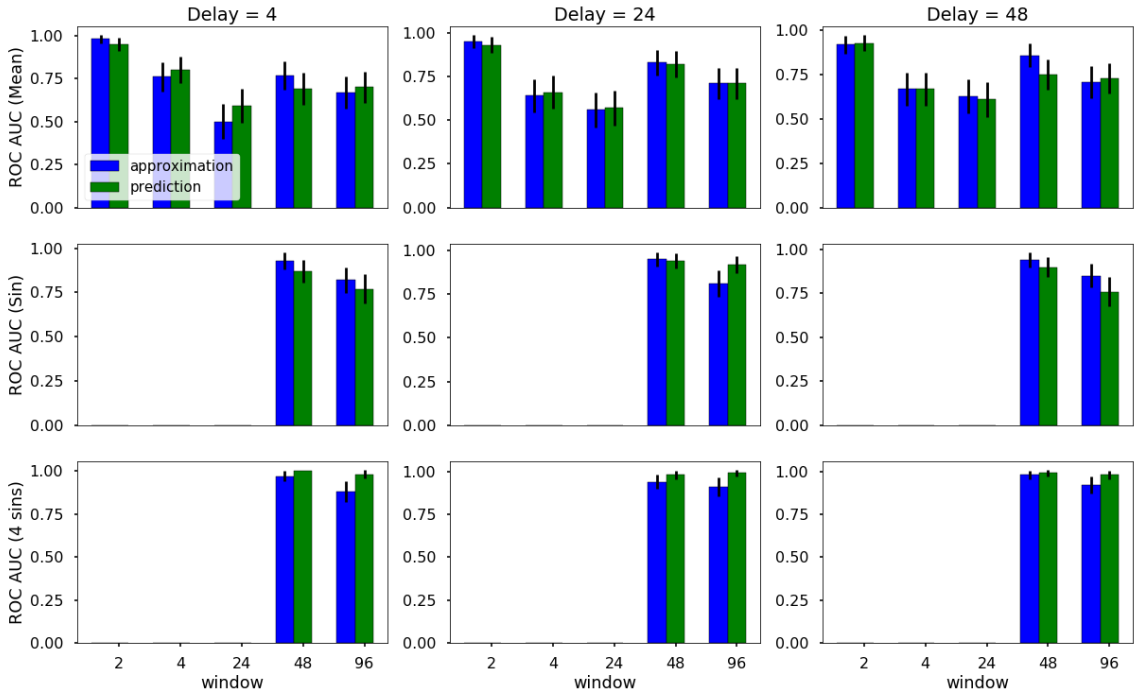

Рис. 2.7. Сравнение ROC-AUC для локальной разладки

Теперь сравним результаты друг с другом по сложности модели $f(x|\theta)$ в методе обнаружения разладки. Как мы видим из рисунка 2.6, чем более точно модель описывает данные, тем лучше. При этом надо иметь в виду что лишние компоненты модели портят результат, поэтому следует усложнять модель только тогда, когда вы уверены что усложнение корректно.

Далее сравним результаты по длине окна l. Снова из рисунка 2.6 видно, что не кратные двум периодам окна работают хуже (потому что оценка происходит по неполной периодичности). При этом, очень маленькие окна работают достаточно хорошо.

С точки зрения задержки d, задержка для маленьких окон ухудшает качество метода при росте задержки. Но для больших окон эффект обратный — наблюдается улучшение при росте задержки.

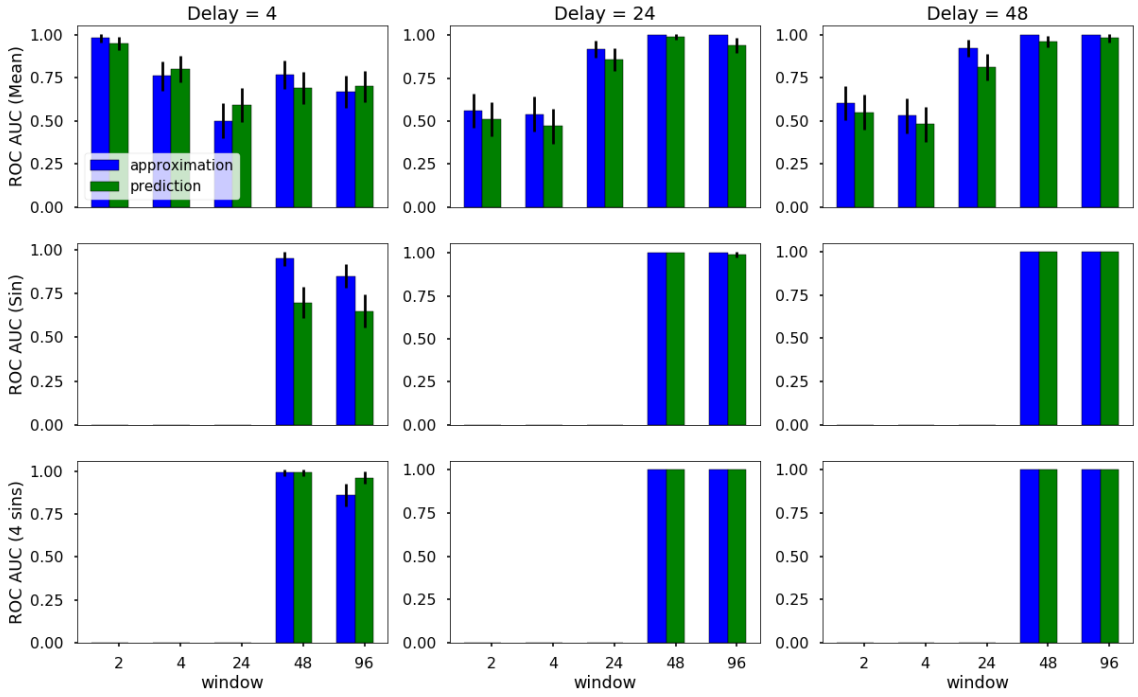

Рис. 2.8. Сравнение ROC-AUC для разладки в среднем

Если сравнивать между собой качество на локальной разладке и на разладке в среднем, то видно следующее. Общая модель „среднее“ хорошо работает для обнаружения локальной разладки, но плохо работает для обнаружения разладки в среднем. При этом, более сложные модели хорошо работают для любого типа разладки. Следовательно, с разумно подобранной моделью можно обнаруживать оба типа разладки.

Заключение

В данной работе мы установили подход к моделированию временных рядов показов мобильной 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. В результате были получены рекомендации по использованию методов обнаружения разладки и выборе параметров этих методов.