Санкт-Петербургский государственный университет
Прикладная математика и информатика
Статистическое моделирование

Мерзляков Климент Викторович

Анализ данных интернет-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 оплата, как правило, происходит за показ, поэтому показатель „количество показов“ является одним из ключевых в деятельности рекламной сети. Соответственно качественный прогноз данного показателя является актуальной задачей для любой рекламной сети.

Рекламные интернет-сети являются интересным объектом для исследования с точки зрения прогнозирования, 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 автономного прогноза 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 прогнозирования показов рекламы и апробация данной методики на реальных данных. В работе будут использоваться фактические, данные одной из работающих рекламных сетей.

Глава 1

Прогнозирование временных рядов

1.1. Первичная обработка временных рядов

Как и с любыми другими типами данных при работе с временными рядами важную роль занимает первичная обработка данных. В данных могут быть пропущенные, либо аномальные значения; в течение времени могут происходить структурные сдвиги; периодичность может меняться. Поэтому прежде чем строить модель прогнозирования, необходимо сначала тщательно проанализировать временной ряд. Временные ряды могут состоять из большого количества компонент. Принято выделять [3]:

- Тренд — медленно меняющаяся компонента
- Периодичность — цикличные изменения ряда с различными периодами
- Остатки — слабо предсказуемая случайная компонента, зависящая от большого количества факторов

При этом периодичная составляющая может состоять из нескольких компонент — сезонной (ежемесячной), еженедельной, ежедневной и не только. Традиционно компоненты ряда считаются независимыми друг от друга:

$$y_t = T_t + S_t + R_t.$$

где y_t обозначает временной ряд в момент времени t, T_t тренд, S_t - периодичность, R_t остатки, $t = 1, ..., n$. Такой ряд с независимыми компонентами называется аддитивным временным рядом. Зачастую компоненты ряда взаимосвязаны, тогда речь идет о мультипликативном характере ряда. В таком случае ряд можно записать следующим образом:

$$y_t = T_t \times S_t \times R_t.$$

При аддитивной модели ряда предполагается, что периодическая компонента не зависит от тренда и имеет стабильно одинаковую амплитуду. При мультипликативной модели амплитуда колебаний периодичной составляющей меняется при изменении тренда. Один из способов определить тип ряда описан в [2]. Он заключается в том, чтобы найти огибающую кривую ряда за вычетом тренда. Если огибающая кривая не имеет больших колебаний, то ряд аддитивный, иначе мультипликативный. Для того, чтобы найти огибающую кривую прежде всего

нужно извлечь тренд из исходного ряда. Тренд можно находить различными способами, но в контексте данной задачи можно воспользоваться простейшим методом скользящего среднего [6]. Модель скользящего среднего с окном k может быть записана следующим образом:

$$\hat{T}_t = \frac{1}{k+1}\sum_{j=0}^{k} y_{t-j}.$$

После того, как мы извлекли тренд оставшийся ряд можно обозначить следующим образом:

$$f_t = A(t)\cos(2\pi\omega t).$$

где $A(t)$ новая медленно меняющаяся компонента. Оказывается, если возвести в квадрат ряд f_t и умножить на 2, то компоненту $A(t)$ снова несложно будет извлечь (снова скользящим средним или каким-либо другим способом):

$$g_t = 2f_t^2 = A^2(t) + A^2(t)\cos(4\pi\omega t)$$

Далее останется только извлечь корень из компоненты $A^2(t)$, получившаяся $A(t)$ и будет огибающей кривой ряда f_t. В случае, если ряд оказался мультипликативным можно его трансформировать в аддитивный с помощью логарифма. Действительно, логарифм ряда $y_t = T_t \times S_t \times R_t$ становится аддитивным по определению $\log(y_t) = \log(T_t) + \log(S_t) + \log(R_t)$. Следовательно после логарифмирования с рядом можно работать как с аддитивным.

Стоит отметить, что для нахождения паттернов в периодичности полезно пользоваться графиками сезонности [5]. В отличие от обычного графика временного ряда, на графике сезонности изображаются отрезки исходного ряда соответствующие одному периоду.

1.2. Метод „Гусеница“ SSA

Рассмотрим базовый метод SSA, описанный в учебном пособии [1].

Будем рассматривать вещественнозначный временной ряд $F_N = (f_0, \ldots, f_{N-1})$ длины $N > 2$. Предполагаем, что ряд ненулевой.

Первый шаг — *вложение*. Процедура вложения переводит исходный временной ряд в последовательность многомерных векторов. Пусть L — некоторое целое число, которое называется *длиной окна*, $1 < L < N$. Процедура вложения образует
$K = N - L + 1$ *векторов вложения*

$$X = (f_{i-1}, \ldots, f_{i+L-2})^{\mathrm{T}}, \qquad 1 \le i \le K.$$

Траекторная матрица ряда F

$$\mathbf{X} = (x_{ij})_{i,j=1}^{L,K} = [X_1 : \ldots : X_K] \tag{1.1}$$

состоит из векторов вложения в качестве столбцов. Очевидно, что $x_{ij} = f_{i+j-2}$ и матрица $\mathbf{X}$ имеет одинаковые элементы на «диагоналях» $i + j = const$. Таким образом, траекторная матрица является *ганкелевой*.

Второй шаг — *сингулярное разложение* траекторной матрицы ряда. Определим $L \times L$ симметричную матрицу:

$$\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}}. \tag{1.2}$$

Матрица $\mathbf{S}$ имеет L линейно независимых собственных векторов и L собственных чисел. Обозначим $\lambda_1, \ldots, \lambda_L$ *собственные числа* матрицы $\mathbf{S}$, взятые в неубывающем порядке ($\lambda_1 \geqslant \ldots \geqslant \lambda_L \geq 0$) и $U_1, \ldots, U_L$ — ортонормированную систему *собственных векторов* матрицы S, соответствующую собственным числам.

Пусть $d = \max\{i : \lambda_i > 0\}$. Пусть $V_i = \mathbf{X}^T U_i / \sqrt{\lambda_i}, i = 1, \ldots, d$, тогда сингулярное разложение матрицы $\mathbf{X}$ может быть записано как

$$\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_d, \tag{1.3}$$

где $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$. Набор $(\sqrt{\lambda_i}, U_i, V_i)$ назовем *i-ой собственной тройкой* сингулярного разложения (1.3).

Третий шаг — *группировка собственных троек*. Процедура группировки делит все множество индексов $1, \ldots, d$, полученных в результате разложения (1.3), на m непересекающихся подмножеств $I_1, \ldots, I_m$.

Рассмотрим такое подмножество $I = i_1, \ldots, i_p$. Результирующая матрица $\mathbf{X}_I$, соответствующая I, определяется как

$$\mathbf{X}_I = \mathbf{X}_{i_1} + \ldots + \mathbf{X}_{i_p}.$$

Вычислив такие матрицы для всех групп $I = I_1, \ldots, I_m$, разложение (1.3) можно записать в сгруппированном виде

$$\mathbf{X} = \mathbf{X}_{I_1} + \ldots + \mathbf{X}_{I_m}. \tag{1.4}$$

Четвертый шаг — *диагональное усреднение*. На этом шаге алгоритма каждая матрица сгруппированного разложения (1.4) преобразуется в ряд длины N.

Пусть $\mathbf{X}_{I_k}$ — результирующая матрица из (1.4). Тогда ее преобразование соответствует усреднению элементов матрицы вдоль «диагоналей» $i + j = const$. Если $\mathbf{X}_{I_k}$ является траекторной матрицей некоторого ряда $(h_0, \ldots, h_{N-1})$, то h_i — есть элемент, лежащий на соответствующей «диагонали». Применяя диагональное усреднение к результирующим матрицам $\mathbf{X}_{I_k}$, получаем восстановленные ряды $\check{F}_N^{(k)} = (\check{f}_0^{(k)}, \ldots, \check{f}_{N-1}^{(k)})$. Исходный ряд $F_N =$

$(f_0, \ldots, f_{N-1})$ раскладывается в сумму m рядов

$$f_n = \sum_{k=1}^{m} \check{f}_n^k.$$

Отметим, что если рассматривать $L \times L$ матрицу

$$\mathbf{U} = [U_1 : \ldots : U_L], \tag{1.5}$$

состоящую из собственных векторов матрицы $\mathbf{S}$ в качестве столбцов, то можно определить матрицу

$$\mathbf{\Lambda} = \mathbf{U}^{\mathrm{T}}\mathbf{S}\mathbf{U}. \tag{1.6}$$

$\mathbf{\Lambda}$ — диагональная матрица, k-ым диагональным элементом которой является k-ое собственное число матрицы $\mathbf{S}$.

Глава 2

Прогнозирование данных интернет-рекламы

2.1. Описание и первичная обработка данных

В данной работе будут использоваться реальные данные компании, занимающейся рекламой в мобильных приложениях. На рисунке 2.1 приведены данные о показах онлайн-рекламы по часам за 5 месяцев.

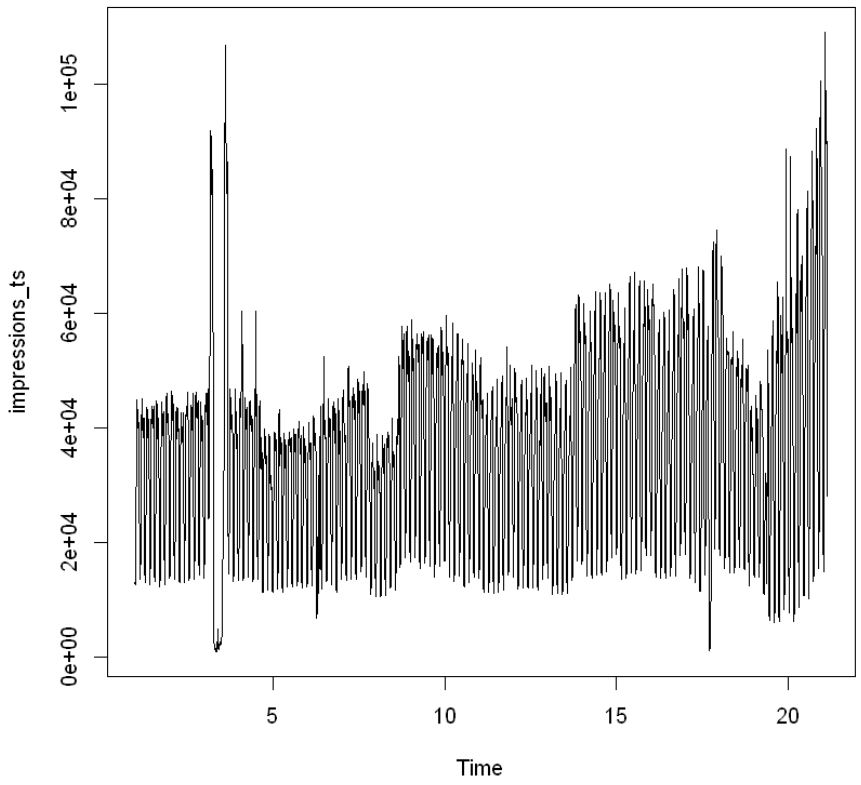

Рис. 2.1. Исходные данные показов рекламы.

Для прогнозирования, независимо от метода, тем выше будет качество прогноза, чем качественнее исходный ряд, поэтому постараемся локализовать и устранить часть проблем. Очевидно в исходном ряду есть несколько проблем:

- На графике видны явные выбросы — резкие скачки или провалы значений. Выбросы связаны, как правило, с техническими неполадками, поэтому заменим их на средние соседние значения.

- В первой половине временного ряда виден провал в данных в виде ступеньки и вскоре возврат к прежним значениям. Причиной этому являются технические неполадки, происходившие в этот период.

Суммарно в ряду были признаны аномальными 362 точки (что соответствует примерно 10%). На рисунке 2.2 изображен график с удаленными аномальными значениями. Заменим удаленные значения на среднее значение между аналогичным часом и днем недели прошлой и будущей недель (рисунок 2.3).

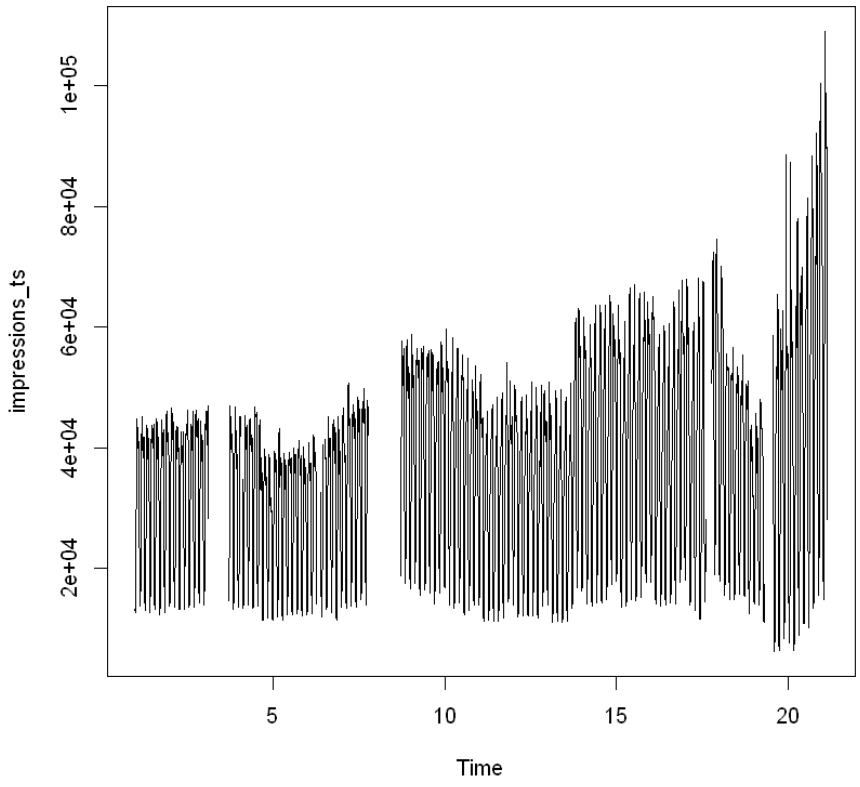

Рис. 2.2. Данные показов рекламы после удаления аномальных точек.

Далее необходимо произвести декомпозицию ряда на тренд, периодичную составляющую и остатки. Наибольшую сложность представляет выделение устойчивой периодичности, поэтому для этих целей необходимо изучить ряд внимательнее. Прежде всего, требуется определить какой характер имеет периодичность - аддитивный или мультипликативный (раздел 1.1). Для этого извлечем тренд из ряда с помощью SSA с окном равным одной неделе (168 часов). Получившийся график остатков (рисунок 2.4) не имеет сильно возрастающих колебаний, то есть явной мультипликативности в исходном ряде нет.

Однако, если провести аналогичную процедуру для логарифма исходного ряда, то ам-

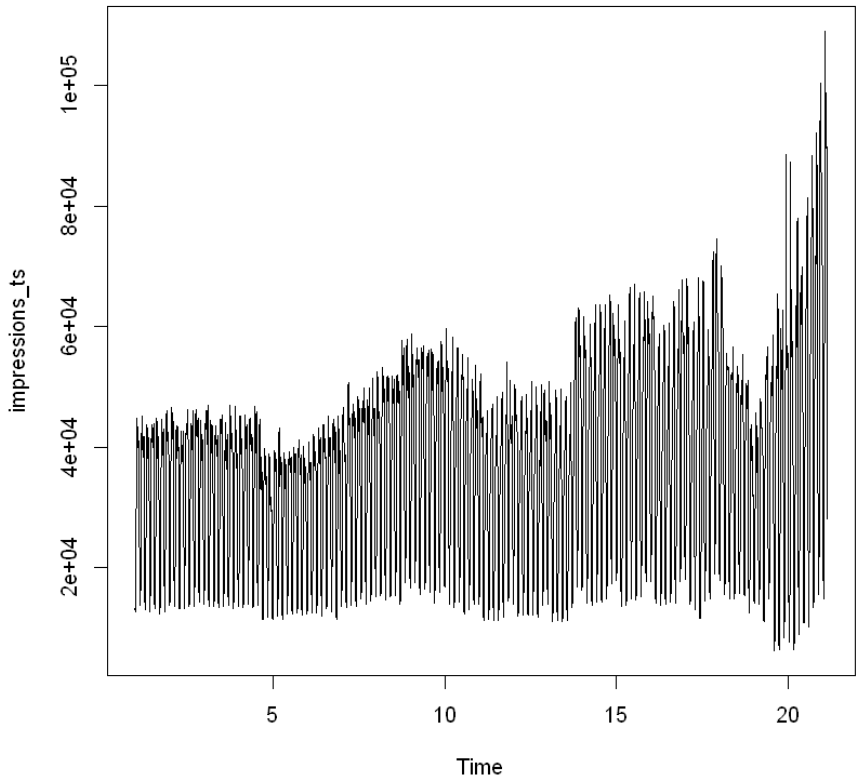

Рис. 2.3. Показы рекламы после замены аномальных значений.

плитуда колебаний графика остатков (после удаления тренда) больше похожа на одинаковую (рисунок 2.5).

Для подтверждения этой мысли построим огибающую кривую, описанную в первой главе (раздел 1.1) для первого и второго случаев. Из графиков 2.6 и 2.7 становится очевидным, что при логарифмировании амплитуда колебаний становится гораздо более ровной. Поэтому будем считать периодичность ряда мультипликативной, соответственно в дальнейшем работать с логарифмом исходного ряда.

Теперь изучим ряд ближе с точки зрения ежедневной периодичности. После детального изучения был сделан вывод о том, что ряд следует разделить на три части — первые 9 недель, последующие 8 недель и оставшиеся 3 недели. Для демонстрации этого вывода построим несколько графиков периодичности (раздел 1.1).

Дело в том, что в первой части временного ряда ежедневная периодичность мало чем отличается друг от друга — в этот период виден довольно устойчивый ежедневный паттерн (рисунок A.1).

Во второй части видно явное различие между будними (рисунок A.2) и выходными днями (рисунок A.3).

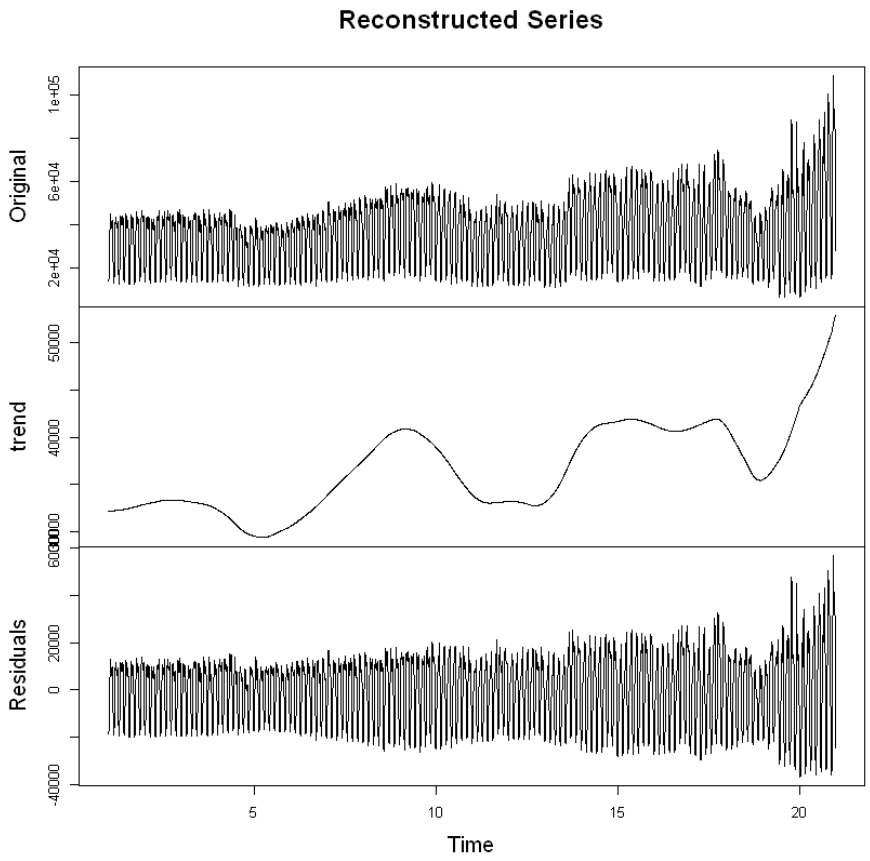

Рис. 2.4. Извлечение тренда из ряда с помощью SSA.

В последние 3 недели происходят зачастую хаотичные колебания в которых сложно уловить какой-либо паттерн (рисунок A.4). Далее вероятно потребуется рассматривать три данных отрезка ряда по отдельности.

Итого, после первичной обработки ряд стал гораздо более гладким и однородным, было выяснено, что ряд имеет мультипликативную периодичность, а также было выяснено что ряд имеет разные паттерны периодичности на разных отрезках времени.

2.2. Сравнение различных моделей для прогнозирования показов рекламы

Попробуем подобрать модель, которая будет наилучшим образом прогнозировать выбранные данные на одни сутки. Выбирать будем из следующих моделей:

- Базовая модель — прогноз равен последнему известному дню.
- SSA учитывающий все предыдущие значения с окном равным половине ряда

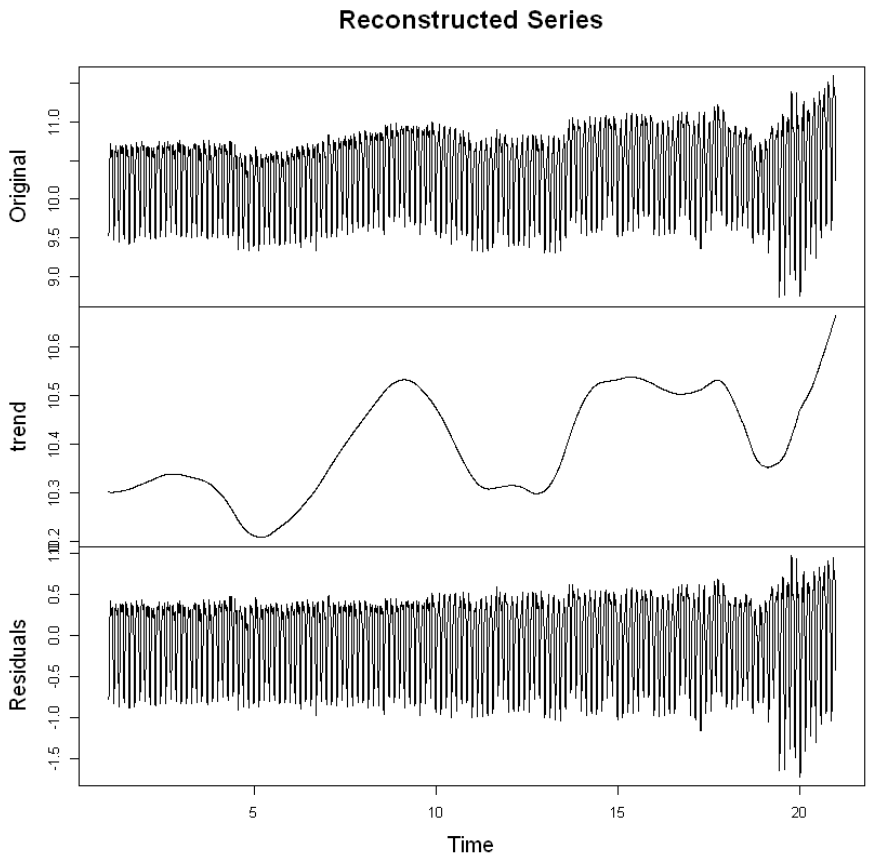

Рис. 2.5. Извлечение тренда из логарифма ряда с помощью SSA.

- Скользящий SSA учитывающий определенное количество последних значений с окном равным половине ряда

Для сравнения моделей прогноз будет строиться не только на последние сутки, а скользящим способом на все предыдущие сутки в ряду. Обозначим весь временной ряд как вектор $Y = y_1, ...y_T$; часть временного ряда, на котором будет строиться модель, как $\tilde{Y} = y_i, ..., y_t$, где $i \geq 1, t \leq T$; а прогноз, как $\hat{Y} = y_j, ...y_h$, где $h - j$ — период на который строится прогноз, при этом $h \leq t < T$. Таким образом, для каждого подхода мы будем строить m моделей и столько же прогнозов. Сравнивать результаты будем по показателю Average RMSE

$$AverageRMSE = \frac{\sum_{p=1}^{m}(\sqrt{\frac{\sum_{j=1}^{h}(y_j - \hat{y}_j)^2}{h}})}{m}$$

.

В базовой модели нет никаких параметров, поэтому мы просто посчитаем показатель ошибки этой модели, и результаты приведем ниже в сравнении с другими моделями. Попробуем применить SSA для прогноза, для этого нам нужно решить следующие задачи:

- Какое количество предыдущих значений учитывать при построении модели?

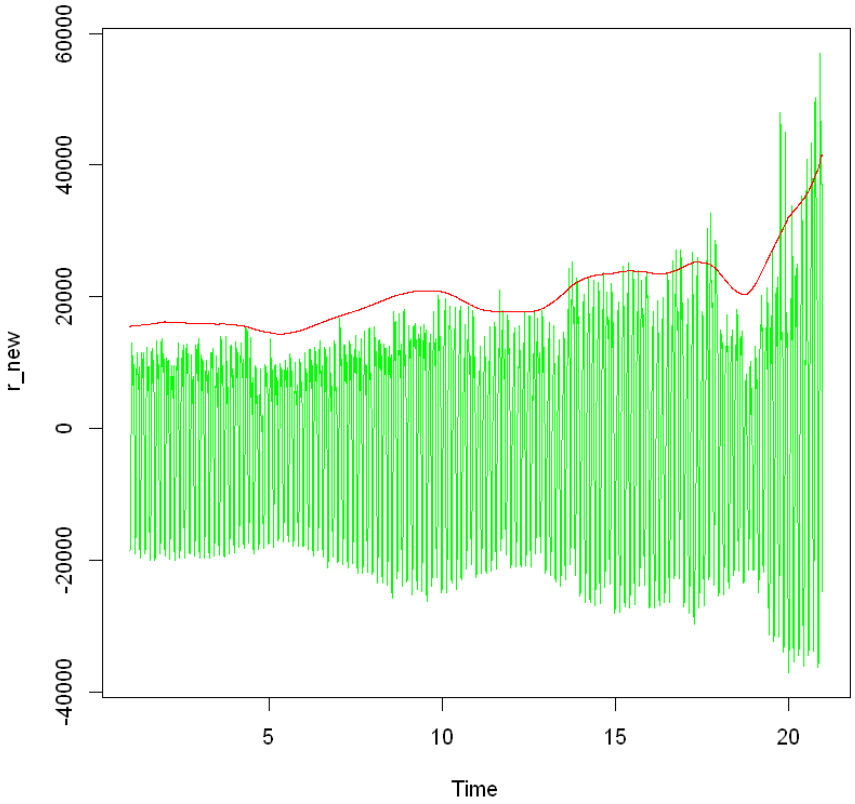

Рис. 2.6. Окаймляющая кривая после извлечение тренда из исходного ряда.

  - Все доступные
  - Некоторое определенное количество

- Какой размер окна выбирать на этапе разложения?
  - „Стандартный“ равный половине ряда
  - Фиксированный

- Какое количество троек выбрать на этапе разложения и как?
  - Фиксированное количество первых троек

- Учитывать ли специфику данного ряда, о которой писалось выше?
  - Не учитывать
  - Учитывать и рассматривать каждый из этих рядов по отдельности

Попробуем задать некоторое количество параметров и для каждой комбинации параметров посчитать среднюю ошибку. Путем перебора оптимальным вариантом оказался прогноз

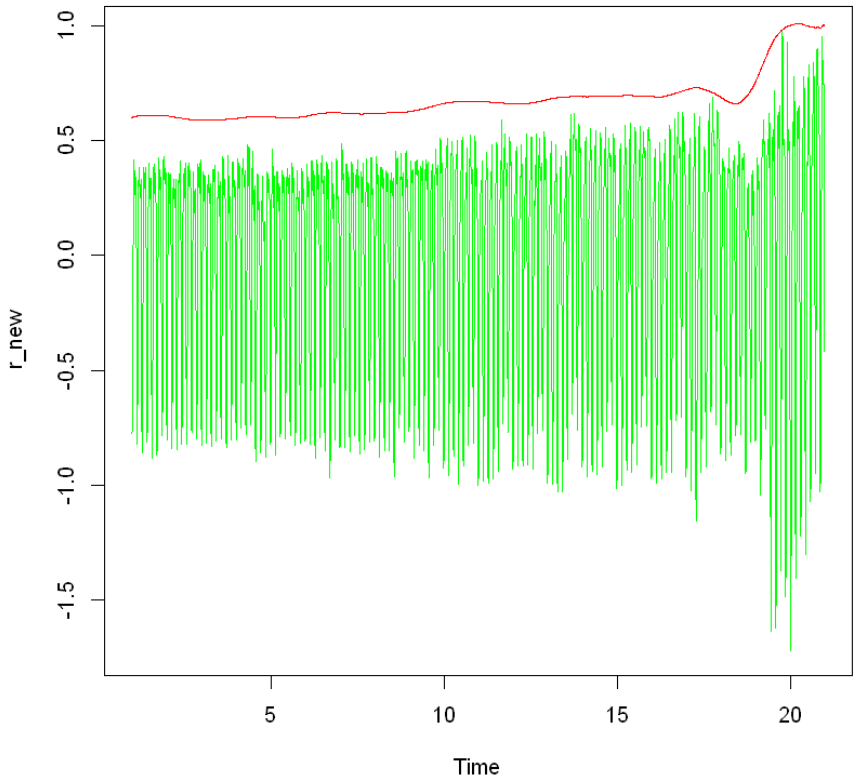

Рис. 2.7. Окаймляющая кривая после извлечение тренда из логарифма ряда.

с помощью модели с периодом (на котором строится модель) равным четырем неделям, длиной окна равной одной неделе и количеством собственных чисел равным пятидесяти. Сравнительные результаты приведены в таблице 2.1

Тип модели	Период	Длина окна	Количество собственных троек	Средний RMSE
SSA	672	168	50	2 754
Базовая	-	-	-	3 470
SSA	672	Стандартная	50	3 627
SSA	Весь доступный	168	50	3 996
SSA	Весь доступный	Стандартная	50	5 412
SSA	168	Стандартная	5	6 374

Таблица 2.1. Сравнительная характеристика моделей.

Примечательно, что прогноз с помощью такой модели оказался качественнее, чем прогноз с помощью базовой модели. Также стоит отметить, что прогноз с помощью SSA со стандартными параметрами показал результат значительно хуже.

Заключение

В данной работе были рассмотрены подходы к анализу временных рядов и метод анализа и прогнозирования SSA. Помимо этого описанные подходы были успешно применены к фактическим данным, были выявлены особенности и паттерны временного ряда, такие как мультипликативный характер ряда, ежедневная периодичность с отличиями в будние и праздничные дни. Более того, была проведена отдельная работа по поиску и замене аномальных значений. При этом методика прогнозирования, примененная в работе показала лучший результат в сравнении с базовым методом. В дальнейшем следует сравнить SSA с другими методами прогнозирования, продумать автономную методику первичного преобразования временного ряда, а так же опробовать полученные подходы к другим временным рядам.

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.

Приложение А

Графики периодичности

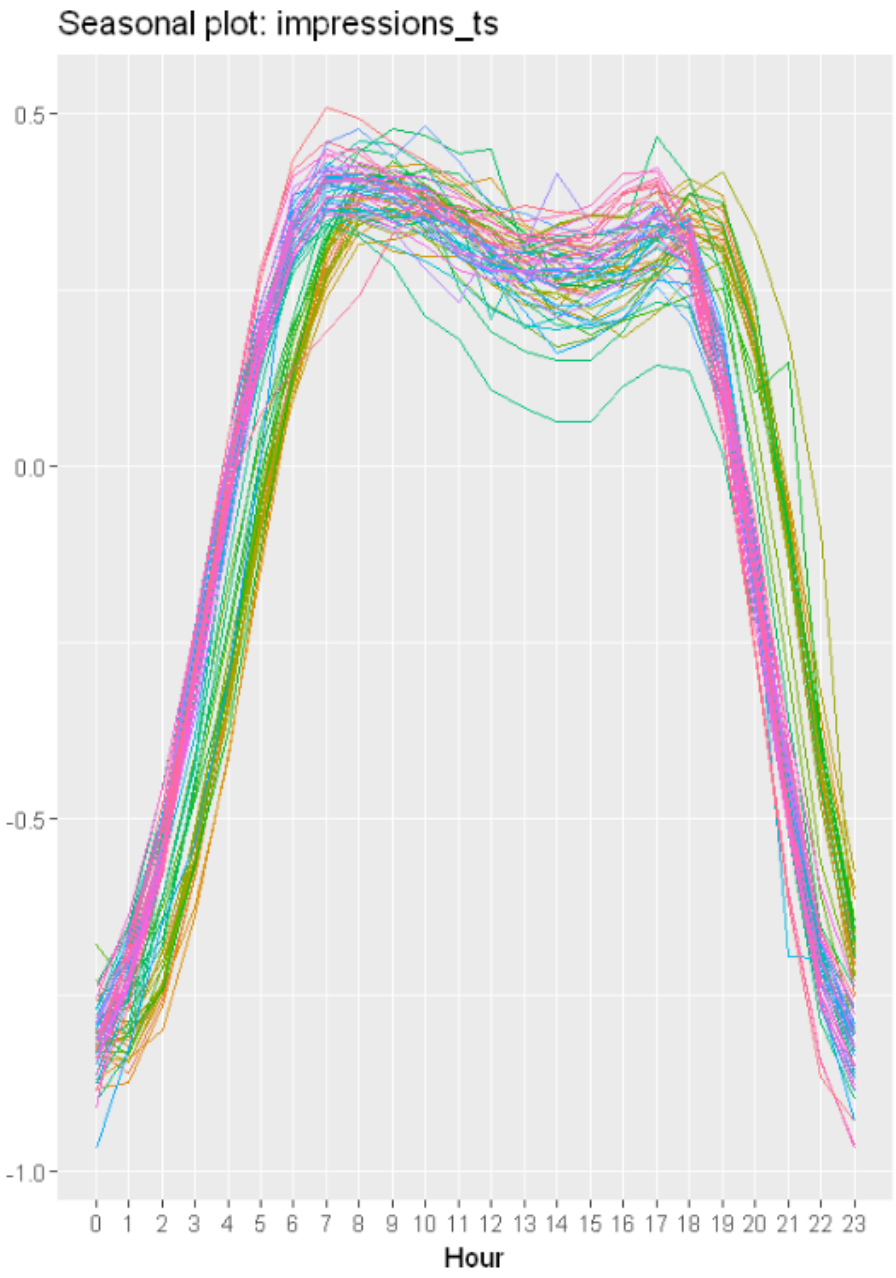

Рис. А.1. График периодичности для первых девяти недель логарифма ряда.

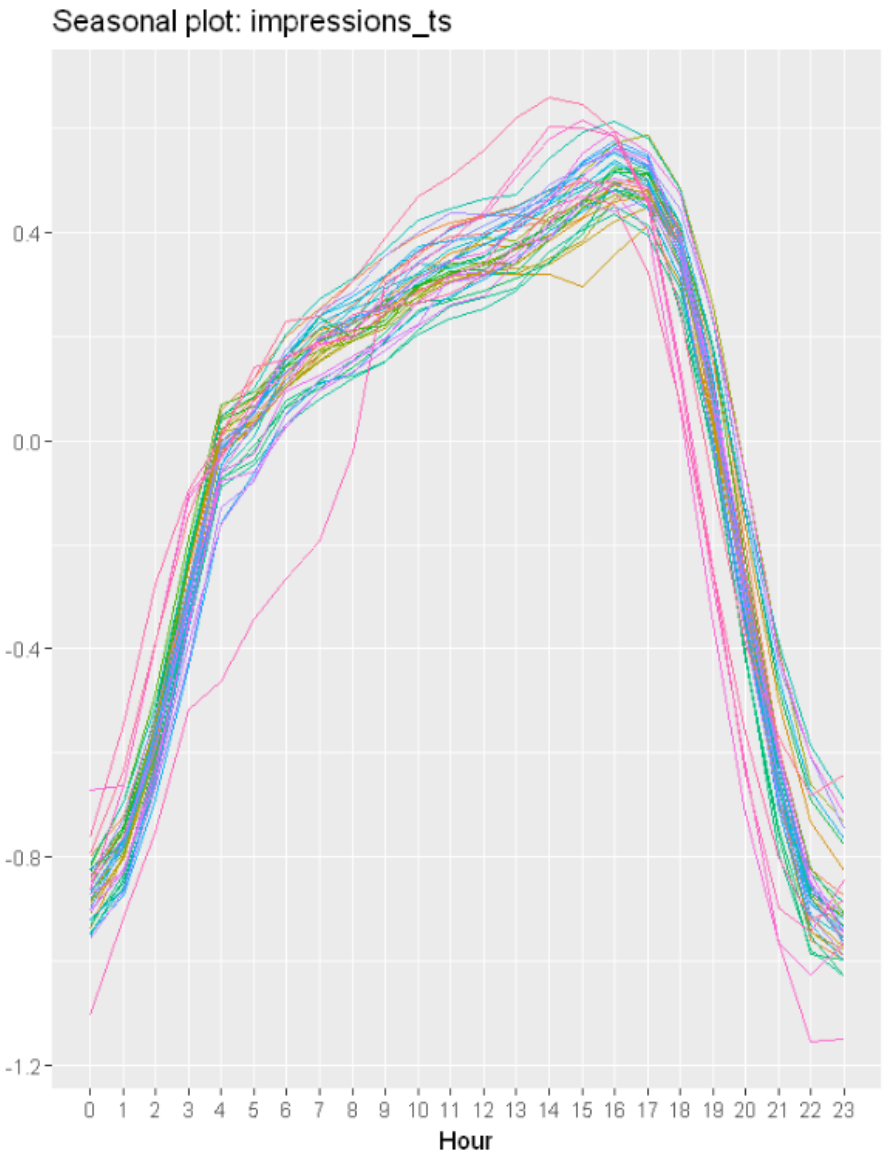

Рис. А.2. График периодичности будних дней для недель с 10 по 18.

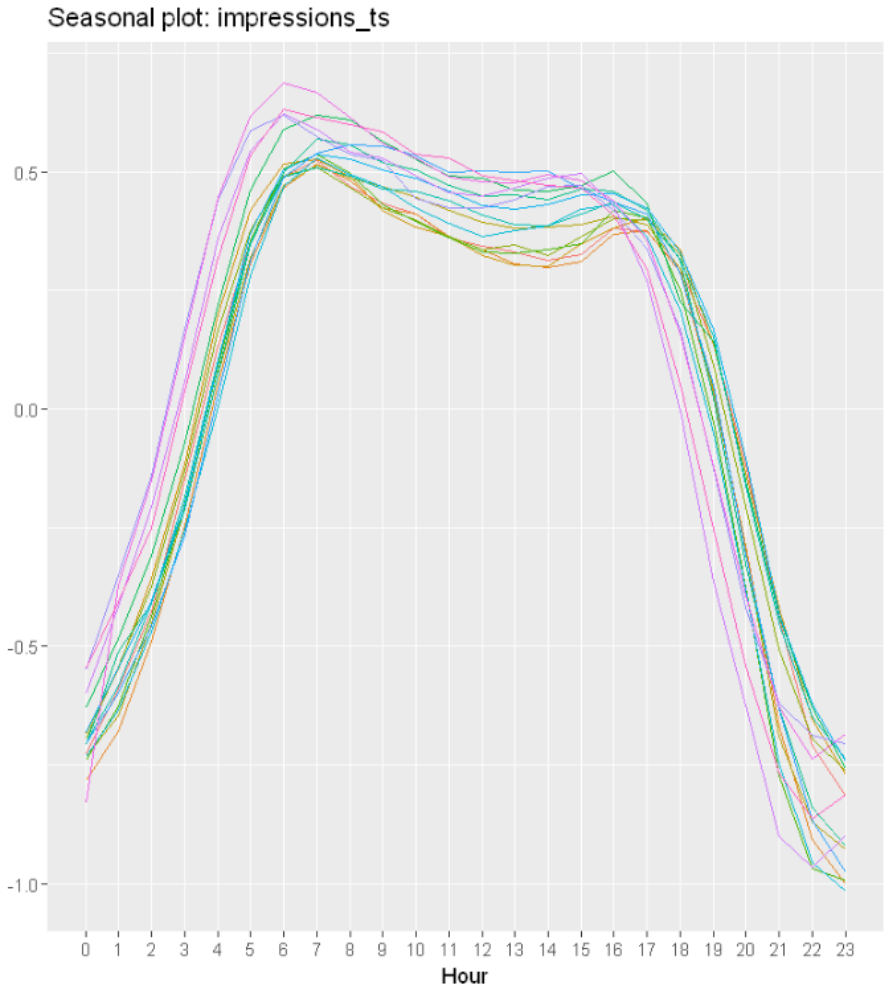

Рис. А.3. График периодичности выходных дней для недель с 10 по 18.

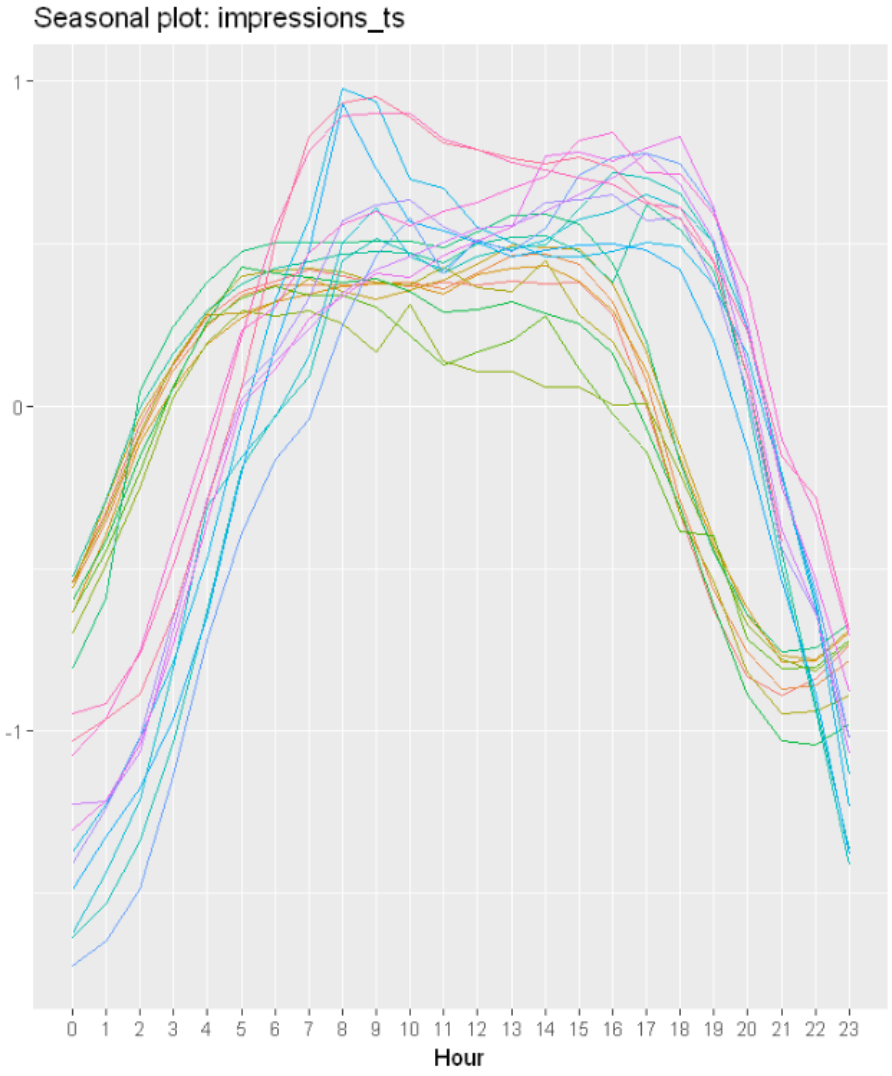

Рис. А.4. График периодичности последних 3 недель.