МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

ПРОГНОЗИРОВАНИЕ ВРЕМЕННЫХ РЯДОВ С ИСПОЛЬЗОВАНИЕМ ЯЗЫКА R

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 248 группы
направления 09.04.03 — ПРИКЛАДНАЯ ИНФОРМАТИКА

механико-математического факультета
Чуйкова Арсения Алексеевича

Научный руководитель
доцент, к. ф.-м. н., доцент _____ В. В. Новиков

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2020

Введение. При торговле товарами посредством сети Интернет постоянно возникает задача прогнозирования объёма продаж. Этот процесс нельзя осуществить без использования специальных знаний и приложений, которые позволяют уменьшить роль ошибок при принятии того или иного решения. Для повышения производительности труда, улучшения качества обслуживания и оптимизации процесса управления необходимо обратить внимание на планирование сбыта продукции.

Цель данной работы: спрогнозировать объём продаж интернет-магазина с использованием языка R и провести численный эксперимент по аппроксимации сезонной составляющей временного ряда с использованием непараметрической регрессионной модели на основе интерполяционных данных.

Задачами работы являются: анализ теоретических основ прогнозирования динамических рядов, применение полученных знаний на практике для реализации вычислительного эксперимента.

Работа состоит из трёх разделов.

Первый раздел посвящён теоретическим основам моделирования временных рядов. В нём рассматриваются основные факты и методы, относящиеся к анализу и прогнозированию временного ряда. В частности, обсуждаются такие понятия как автокорреляция, тренд, сезонность и особенности аддитивной модели.

Второй раздел посвящён использованию среды R для анализа временных рядов. Среди прочего рассмотрены средства языка R, с помощью которых происходит чтение данных временного ряда, его графическое представление, разложение, прогнозирование и экспоненциальное сглаживание. Анализируется работа интернет-магазина и производится прогнозирование объёма продаж на конкретном примере в среде R.

В третьем разделе был рассмотрен подход к моделированию сезонной компоненты с помощью непараметрической регрессионной модели, основанной на использовании частичных сумм Фурье-Лагранжа. Результаты работы опубликованы в статье.

Основное содержание работы. Примером временного ряда, который может быть описан с использованием аддитивной модели с тенденцией и сезонностью, является временный ряд объёма продаж интернет-магазина

Amazon. На официальном сайте www.amazon.com опубликованы годовые отчёты организации. Требуется, используя данные по объёму продаж за 2010-2019 год (1-40 кварталы), составить прогноз объема продаж Amazon на 2020 год (41-44 кварталы).

Будем предполагать что исходные данные находятся в файле "Amazon.txt"в текущем каталоге.

Чтение данных read.table в набор d:

```
d<- read.table("Amazon.txt",h=F)
```

Формирование временного ряда amazontimeseries из d:

```
amazontimeseries <- ts(d, frequency=4, start=c(2010,1))
amazontimeseries
```

Функция decompose() возвращает список объектов в качестве результата, где содержатся оценки периодической составляющей, тренда и нерегулярной компоненты, хранящиеся в именованых элементах этого списка объектов, называемых «seasonal», «trend» и «random» соответственно:

```
amazontimeseriescomponents <- decompose(amazontimeseries)
```

Вывод графика с элементами ряда на рисунке 1:

```
plot(amazontimeseriescomponents)
```

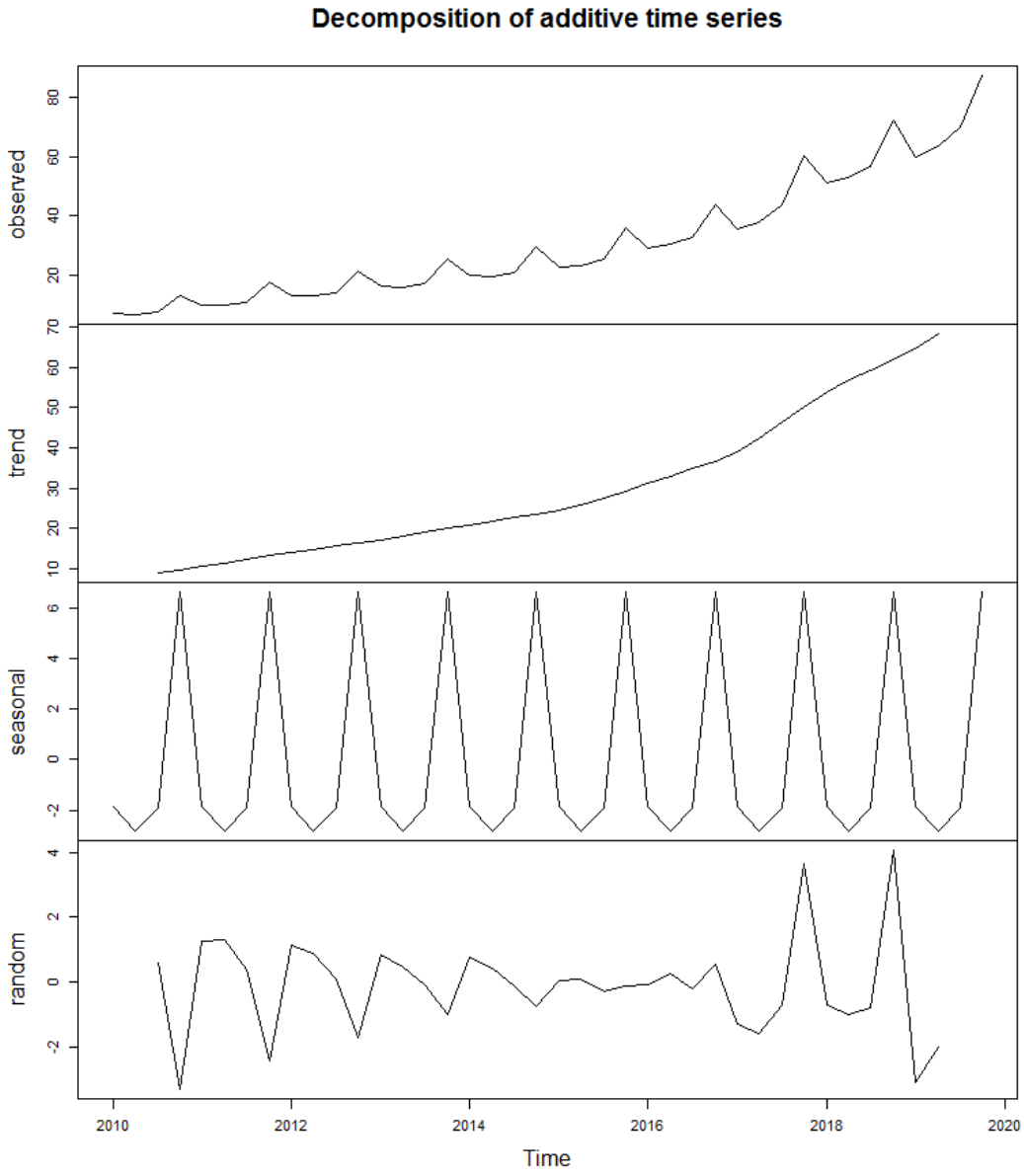

Рисунок 1 – Декомпозиция ряда Amazon

Построение модели HoltWinters. Результат находится в объекте mdHoltWinters:

```
mdHoltWinters<- HoltWinters(amazontimeseries, alpha = 0.2, beta = 0.3,
gamma = 0.8, seasonal = "additive")
```

Построение прогнозных значений (точечный и интервальный прогнозы) на 4 периода вперед с помощью модели mdHoltWinters. Результат находится в объекте dw:

```
dw <- predict(mdHoltWinters, 4, prediction.interval = TRUE)
```

Вывод результата прогнозирования из объекта dw, где значения точечного прогноза: fit ; интервальный прогноз: lwr - нижняя граница, upr - верхняя граница на рисунке 2:

	fit	upr	lwr
2020 Q1	73.02193	77.26224	68.78162
2020 Q2	76.72997	81.11126	72.34868
2020 Q3	82.76745	87.35405	78.18085
2020 Q4	99.45399	104.31539	94.59259

Рисунок 2 – Точечный и интервальный прогноз Amazon

На рисунке 3 представлены графики с фактическими, теоретическими и прогнозными значениями:

```
plot(mdHoltWinters,dw,main="Объём продаж интернет-магазина Amazon",
ylab="Факт / Прогноз",xlab="Годы")
```

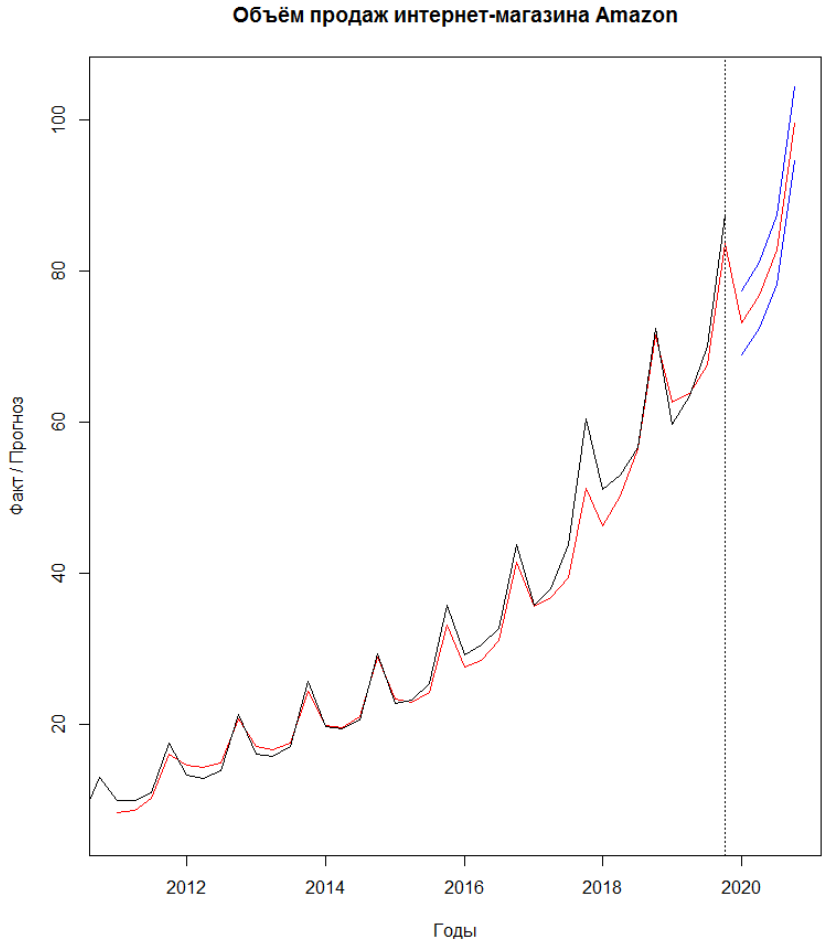

Рисунок 3 – Фактические, теоретические и прогнозные значения ряда Amazon

Рассмотрим непараметричесую регрессионную модель

$$Y_i = m(X_i) + \varepsilon_i\,,\ i = 1, ..., n,$$

где $m(x) = \mathsf{E}\,(Y|X = x)$ – неизвестная функция регрессии, подлежащая оцениванию на основе эмпирических данных $\{(X_i, Y_i)\}_{i=1}^n$, $\{\varepsilon_i\}_{i=1}^n$ – случайные ошибки. Обсудим достаточные условия состоятельности *оценок ортогонального разложения* $\hat{m}(x)$ основанных на представлении функции $m\,(x)$ рядом Фурье

$$m\,(x) = \sum_{j=0}^{\infty} \beta_j \varphi_j\,(x)\,, \tag{1}$$

по некоторой заранее выбранной ортонормированной системе $\{\,\varphi_j\,(x)\}_{j=0}^{\infty}$. При построении таких оценок бесконечный ряд (1) заменяется его частичной суммой подходящего порядка $N\,(n)$, а коэффициенты Фурье β_j — их оценками $\hat{\beta}_j$. Хотя методы оценивания непараметрической регрессии, связанные с ортогональными разложениями, уступают в популярности ядерным методам, они также представляют значительный интерес. Это связано, прежде всего, с простотой и естественностью конструкции таких оценок, а также с не слишком обременительными условиями, обеспечивающими их состоятельность.

Пусть система $\{\,\varphi_i\,(x)\}_{i=0}^{\infty}$ ортонормирована на $[-1,\,1]$ относительно скалярного произведения

$$(f, g) = \int_{-1}^{1} f\,(x)\,g\,(x)\,\rho(x)\,dx,$$

где $\rho(x)$ — весовая функция. Предположим, далее, что переменная X принимает равноотстоящие значения

$$X_i = -1 + \frac{2(i-1)}{n} + \theta_n,\ i = 1, \ldots, n,$$

где $\theta_n \in [0, 2/n]$ — постоянные числа и $\{A_i\}_{i=1}^n$ — разбиение отрезка $[-1,\,1]$ на неналегающие интервалы такие, что $X_i \in A_i$, $i = 1, \ldots, n$. Тогда

$$\beta_j = \sum_{i=1}^{n} \int_{A_i} m\,(x)\ \varphi_j\,(x)\,\rho(x)\,dx \approx \sum_{i=1}^{n} m\,(X_i) \int_{A_i} \varphi_j\,(x)\,\rho(x)\,dx.$$

Заменяя значение $m(X_i)$ на Y_i, получим оценку для коэффициента β_j:

$$\hat{\beta}_j = \sum_{i=1}^{n} Y_i \int_{A_i} \varphi_j(x)\,\rho(x)\,dx, \tag{2}$$

после чего, ограничиваясь конечным числом $N(n)$ членов разложения (1), получаем искомую оценку ортогонального разложения для функции регрессии

$$\hat{m}_{N(n)}(x) = \sum_{j=0}^{N(n)} \hat{\beta}_j \varphi_j(x).$$

Хорошо известным фактом является наличие равномерной сходимости как тригонометрического ряда Фурье, так и интерполяционного процесса Лагранжа $\{L_n(f,x)\}$ с равноотстоящими узлами для функций с «хорошими» структурными свойствами, например, для непрерывно дифференцируемых. Для узлов $x_{i,n} = 2\pi i/(2n+1), i = 0,...,2n$, коэффициенты Фурье-Лагранжа будут интегральными суммами для коэффициентов ряда Фурье, что и объясняет сходные аппроксимативные возможности этих двух типов операторов. Для других ортогональных систем связь между суммой ряда Фурье и соответствующим интерполяционным полиномом не столь явная, но также достаточно тесная. Это позволяет предположить, что вместо оценок коэффициентов (2), содержащих интегралы по частичным отрезкам, можно использовать коэффициенты Фурье-Лагранжа, содержащие только значения функции в точках наблюдения.

В случае тригонометрической системы и узлов $\{X_i = x_{i,n} = 2\pi i/(2n+1)\}$, $i = 0,...,2n$, соответствующие оценки будут иметь вид

$$\hat{a}_k = \frac{2}{2n+1}\sum_{i=0}^{2n} Y_i \cos kX_i,$$

$$\hat{b}_k = \frac{2}{2n+1}\sum_{i=0}^{2n} Y_i \sin kX_i,$$

$$\hat{m}_N(x) = \frac{\hat{a}_0}{2} + \sum_{k=0}^{N(n)} \hat{a}_k \cos jx + \hat{b}_k \sin jx. \tag{3}$$

Можно рассматривать аналогичные выражения и для других используемых в непараметрическом оценивании ортогональных систем.

В данной работе выполнен численный эксперимент по сглаживанию данных с помощью оценки (3). Его результаты показывают, что удовлетворительное качество аппроксимации может быть достигнуто при степени $N(n)$ полинома (3) существенно меньшей числа n.

В качестве тестовых наборов данных использовались значения функций:

$1) m(x) = \sqrt{1 + \cos x}$

$2) m(x) = 1 - |\cos(4 \arccos(\frac{x}{\pi} - 1))|$

$3) m(x) = \arccos(\cos(3x)) - \frac{\pi}{2}$

$4) m(x) = \cos(8 \arccos(\frac{x}{\pi} - 1))$

$5) m(x) = \frac{\sqrt{|x(x-\frac{5}{2})(x-2\pi)|}}{2},$

возмущённые нормальной случайной составляющей ε.

Код программы представлен в Приложении. Далее представлен графический результат численного эксперимента для тестовой функций 5) как наиболее характерной. Вывод графика исходной функции $m(x) = \frac{\sqrt{|x(x-\frac{5}{2})(x-2\pi)|}}{2}$ (синяя линия), возмущённой функции (зелёная линия) и аппроксимирующего полинома (красная линия) на рисунке 4:

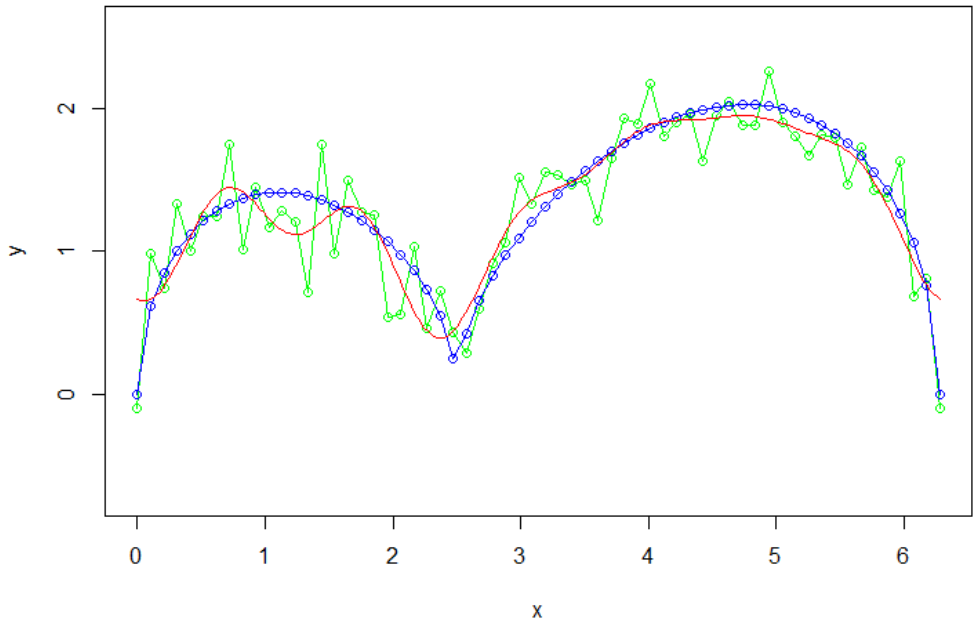

Рисунок 4 – $n = 30$ и $N(n) = 6$

В качестве реальных данных использовался ряд значений объёма продаж United Microelectronics Corporation. UMC — тайваньский производитель микроэлектроники (полупроводниковых изделий), образованный в 1980 из спонсируемого государством Industrial Technology Research Institute (ITRI). На официальном сайте www.umc.com опубликованы ежемесячные отчёты организации. Используем данные по объёму продаж за период 2009-2019 г.

Чтобы иметь "эталон"для оценки полученных результатов, выполним выделение сезонной компоненты с помощью функции decompose() языка R. Вывод графика с элементами ряда на рисунке 5:

```
UMCtimeseriescomponents <- decompose(UMCtimeseries)
plot(UMCtimeseriescomponents)
```

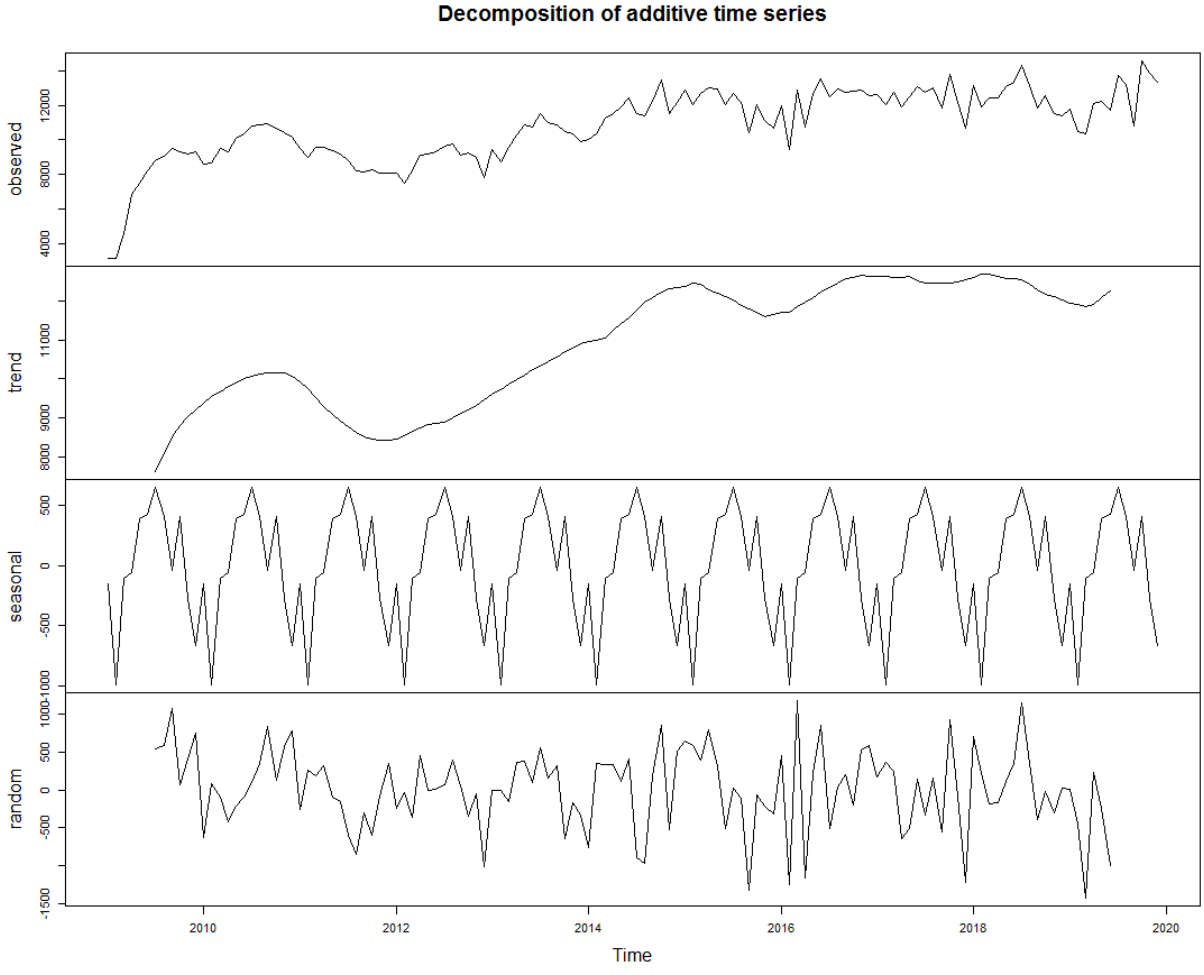

Рисунок 5 – Декомпозиция ряда UMC

Вычитаем тренд из наблюдаемых значений и строим график 6:

```
UMCtimeserieswithouttrend= UMCtimeseries- UMCtimeseriescomponents$trend
plot(UMCtimeserieswithouttrend)
```

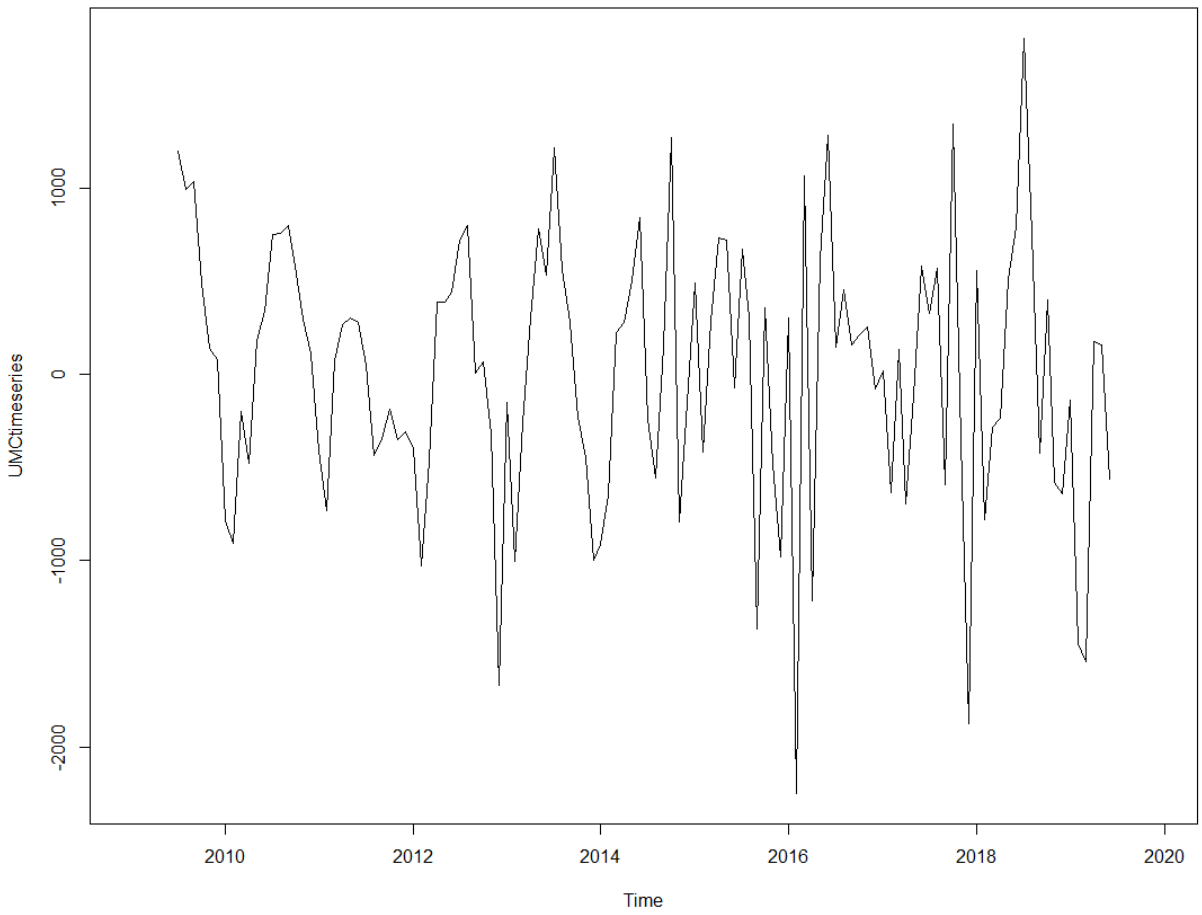

Рисунок 6 – Значения ряда UMC без тренда

Сглаживаем данные по каждому году с помощью оценки(3), где степень полинома $N(n) = 5$ и $n = 6$.

Находим усреднённые значения полинома для каждого месяца.

Предполагаем, что сезонная составляющая остаётся постоянной на всех интервалах. Вывод графика сезонности на рисунке 7:

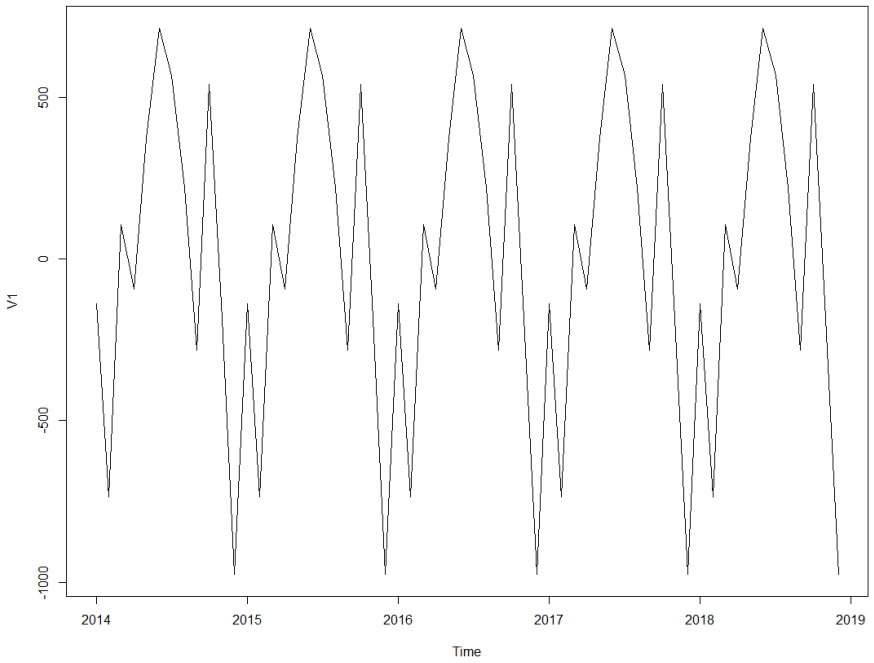

Рисунок 7 – Сезонная составляющая на 5 интервалах($N = 5$)

Вычитаем сезонность из исходных значений (без тренда). Вывод графика случайной компоненты на рисунке 8:

```
UMCrandom = UMCtimeserieswithouttrend[61:120]-fullN5seasonal
plot(UMCrandom, xlab = "years", ylab = "random")
```

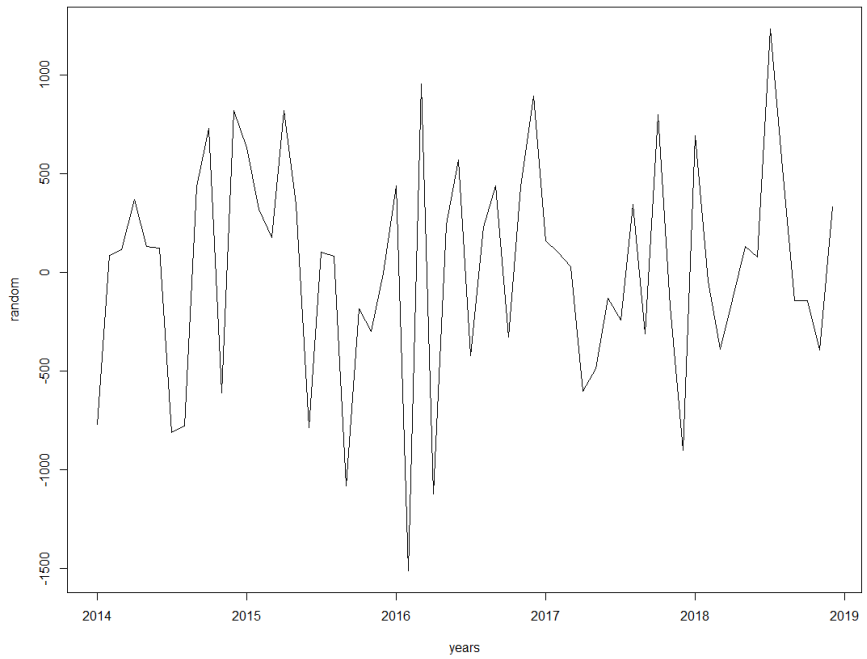

Рисунок 8 – Случайная компонента на 5 интервалах($N = 5$)

Проведя вычисления для $N(n) = 2$, $N(n) = 3$, $N(n) = 4$ и $N(n) = 5$ при $n = 6$, попарно сравним сезонные оставляющие за один период для различных $N(n)$ с сезонной составляющей seasonal, полученной алгоритмом decompose (см. рисунок 5). Проанализировав графики, можно сделать вывод, что результат выделения сезонной составляющей, наиболее близкий к результату работы функции decompose() получается при степени приближающего полинома $N(n) = 5$. Вывод графиков на рисунке 9

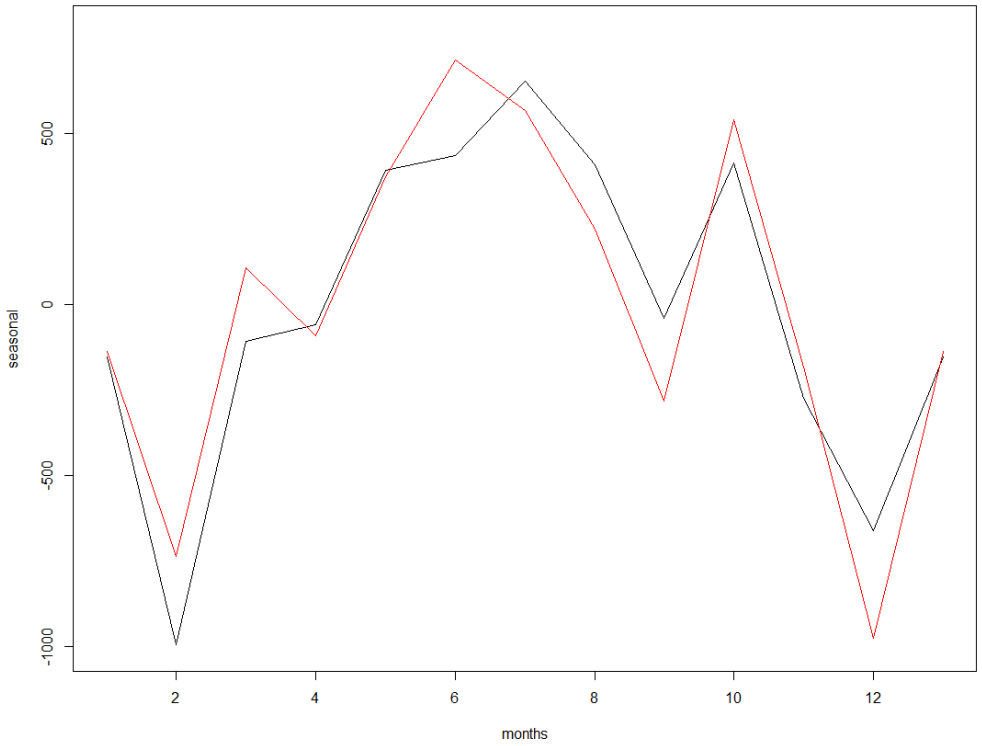

Рисунок 9 – Сезонная компонента, полученная алгоритмом decompose (чёрным) и сезонная компонента при $(N = 5, n = 6)$(красным)

Заключение. В данной работе был сделан обзор статистических методов анализа временных рядов, а также рассмотрены средства программной реализации этих методов в среде программирования R. В качестве практического приложения изученной теории было выполнено прогнозирование объёма продаж интернет-магазина средствами языка R и проведён численный эксперимент по сглаживанию тестовых данных с помощью непараметрической регрессионной модели, использующей частичные суммы Фурье-Лагранжа.

Результаты проведённого в третьей части работы численного моделирования показывают, что рассмотренный интерполяционный метод сглаживания, может быть использован для выделения сезонной составляющей временного ряда. При этом анализ тестовых примеров позволяет предположить, что при достаточном числе наблюдений ($2n+1 > 30$) и не слишком большой случайной компоненте удовлетворительное качество аппроксимации может быть достигнуто при порядке N приближающего полинома, существенно меньшем числа узлов $2n+1$. При работе с реальными данными, когда число узлов не слишком велико (ежемесячные наблюдения), удовлетворительный результат достигается при N, близком к n.