ИТОГОВЫЙ ПРОЕКТ ПО КУРСУ
Data Scientist

ИССЛЕДОВАНИЕ ЭКОНОМИЧЕСКИХ ВРЕМЕННЫХ РЯДОВ МЕТОДАМИ ЧАСТОТНОГО АНАЛИЗА С ПОМОЩЬЮ АМПЛИТУДНО-ФАЗОВЫХ ОПЕРАТОРОВ

Чкалова Дарья

ПЛАН

1. Постановка задач. Источники данных. Предобработка данных
2. Амплитудно-фазовый оператор (теоретические сведения)
3. Аппроксимация временного ряда с помощью АФО
4. Прогнозирование временного ряда методами регрессионного анализа
5. Прогнозирование ряда с помощью пакета prophet

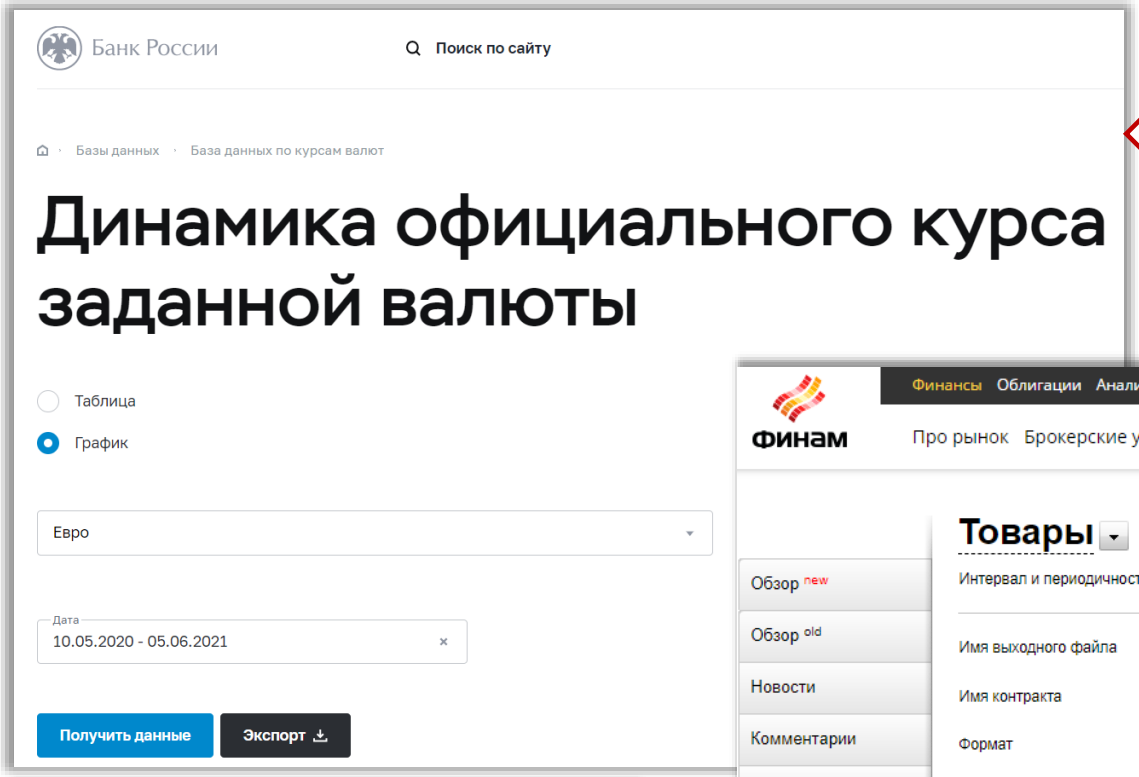
1. ПОСТАНОВКА ЗАДАЧ. ИСТОЧНИКИ ДАННЫХ. ПРЕДОБРАБОТКА ДАННЫХ

1.1 ПОСТАНОВКА ЗАДАЧ

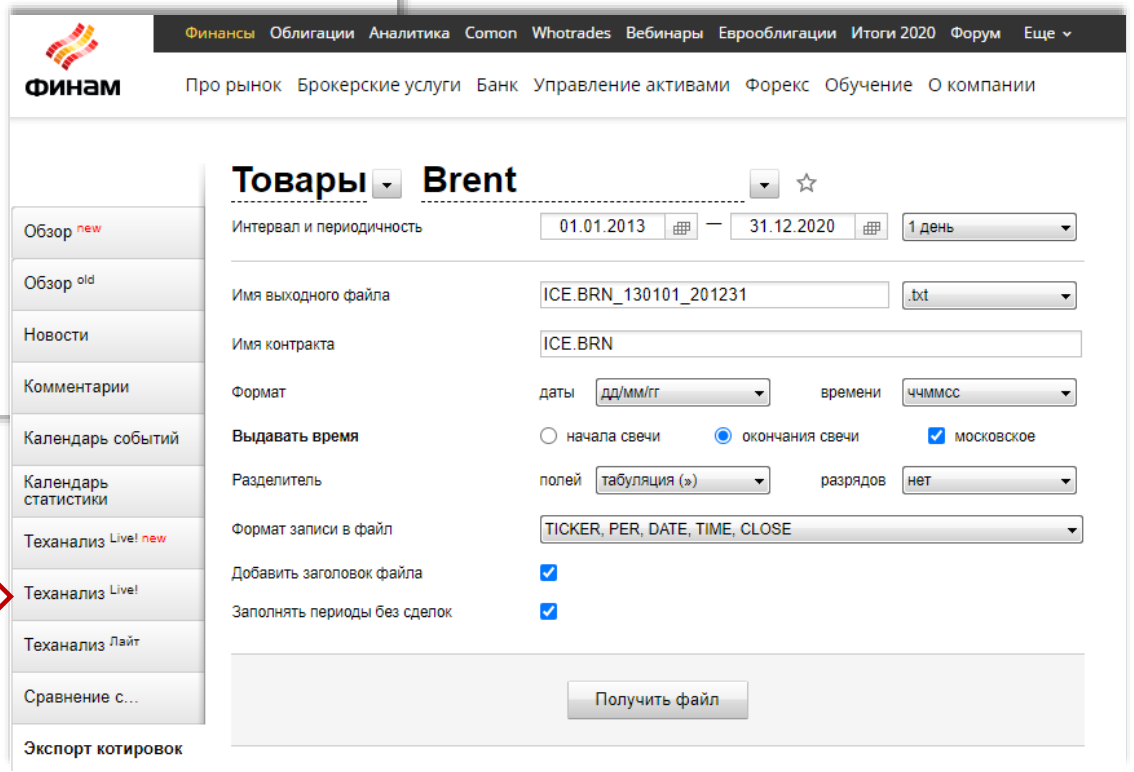
1. Реализовать действие амплитудно-фазового оператора на временной ряд с целью выделения заданных гармоник
2. Разработать и реализовать алгоритмы аппроксимации временных рядов
3. Исследовать структуру таких временных рядов, как: курсы валют, природных ископаемых и т.п.
4. Разработать и реализовать алгоритмы прогнозирования исследуемых временных рядов

1.2 ИСТОЧНИКИ ДАННЫХ

ДИНАМИКА КУРСА ВАЛЮТ
http://www.cbr.ru/



ПРИРОДНЫЕ ИСКОПАЕМЫЕ, МЕТАЛЛЫ, ГОСОБЛИГЦИИ, АКЦИИ, СТАТИСТИКА ПО ОТРАСЛЯМ ЭКОНОМИКИ
https://www.finam.ru/



1.3 ПРЕДОБРАБОТКА ДАННЫХ

Исходные данные представляют собой несортированные pandas.DataFrame из двух столбцов: **дата** (не обязательно последовательные значения) и **значение показателя**. Поэтому необходима предобработка:

- ✓ **Сортировка данных** в хронологическом порядке. Например, курсы валют с сайта ЦБ России экспортируются в обратном хронологическом порядке
- ✓ **Заполнение пропусков** в данных линейной интерполяцией по известным точкам. Большинство исследуемых временных рядов имеют пропуски по выходным и праздничным дням
- ✓ **Замена дат** на порядковые номера замеров временного ряда

2. АМПЛИТУДНО-ФАЗОВЫЙ ОПЕРАТОР (ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ)

2.1 АМПЛИТУДНО-ФАЗОВЫЙ ОПЕРАТОР

Амплитудно-фазовый оператор (АФО) – это сумма преобразований, осуществляющих изменение амплитуды и начального сдвига функции. В качестве базисной функции рассматривается многочлен

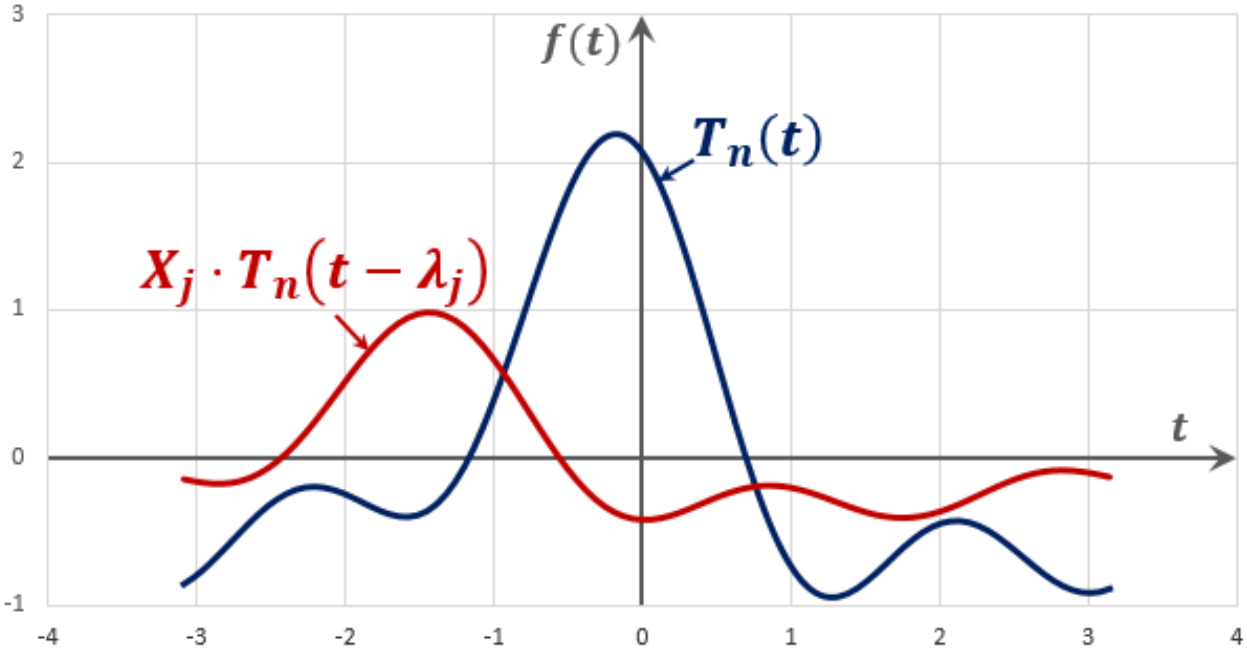
$$T_n(t) = \sum_{k=1}^{n} a_k \cos kt + b_k \sin kt\,.$$

АФО осуществляет преобразование по правилу

$$H_m\left(T_n, \{X_j\}, \{\lambda_j\}; t\right): \quad T_n(t) \to \sum_{j=1}^{m} X_j \cdot T_n(t - \lambda_j)$$

Параметры АФО – амплитуды X_j и сдвиги λ_j **вещественны**, порядок АФО $m \le n$.

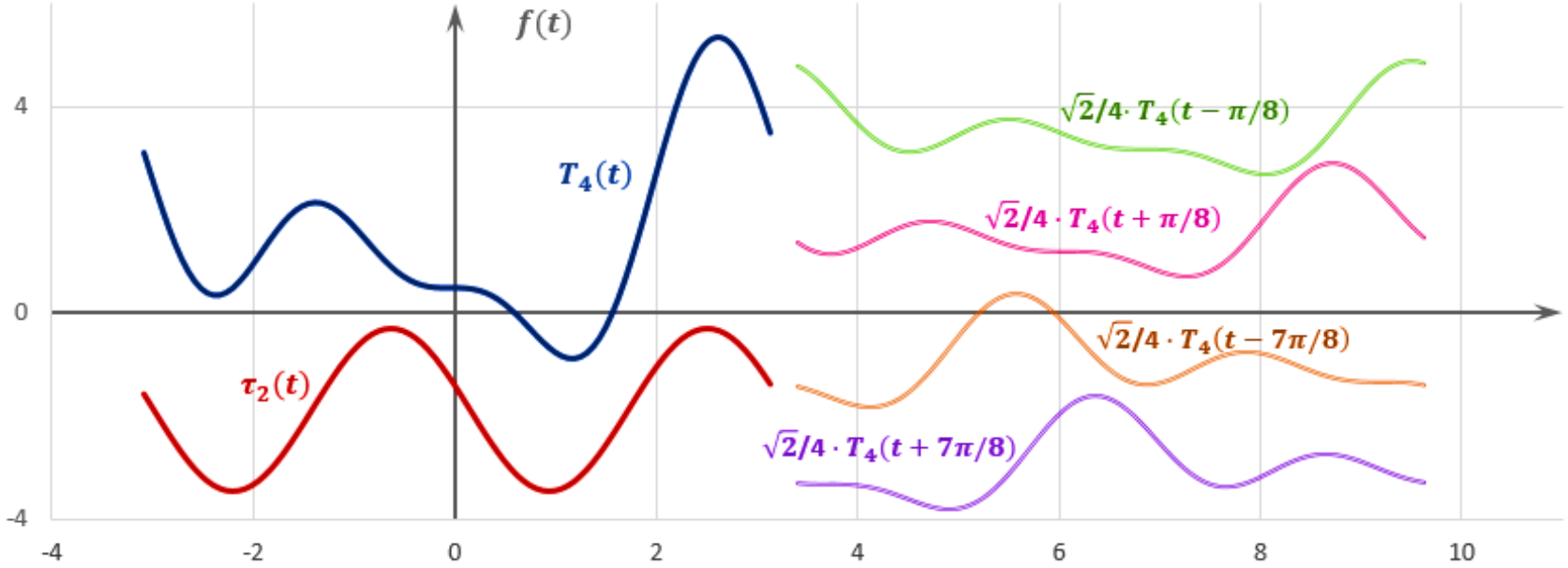
2.1 АМПЛИТУДНО-ФАЗОВЫЙ ОПЕРАТОР



2.1 АМПЛИТУДНО-ФАЗОВЫЙ ОПЕРАТОР

На данный момент аналитически решены задачи выделения с помощью АФО из многочлена $T_n(t)$ **одной заданной гармоники** и **суммы гармоник**:

$$H_m(T_n) - a_0 \sum_{j=1}^{n} X_j = \tau_k(t), \qquad H_m(T_n) - a_0 \sum_{j=1}^{n} X_j = \sum_{\mu_k \in \mathcal{M}} \tau_{\mu_k}(t)$$



2.2 НАУЧНОЕ ОБОСНОВАНИЕ МЕТОДА

Результаты были получены в рамках выполнения грантов Российского Фонда Фундаментальных исследований

- 16-31-00252 «Специальные интерполяционные формулы и их применение в численном анализе»
- 18-01-00744 «Экстремальные задачи для амплитудно-частотных операторов и рациональных функций»
- 20-31-90010 «Выделение гармоник и степеней из тригонометрических и алгебраических многочленов»

и опубликованы в ведущих математических журналах

- D.G. Vasilchenkova, V.I. Danchenko, *Extraction of Several Harmonics from Trigonometric Polynomials*. *Fejer-Type Inequalities*, Proceedings of the Steklov Institute of Mathematics, 2020, Vol. 308, 92-106. https://doi.org/10.1134/S0081543820010083
- D.G. Chkalova, *Fast Optical Signal Filtering by Means of Amplitude and Phase Operators*, 2020 Journal of Physics: Conference Series, Vol. 1679, 022092. https://iopscience.iop.org/article/10.1088/1742-6596/1679/2/022092
- D.G. Vasilchenkova, V.I. Danchenko, *Extraction of Harmonics From Trigonometric Polynomials by Phase-Amplitude Operators* // St. Petersburg Mathematical Journal, 2021, Vol. 32, 215-232. https://www.ams.org/journals/spmj/2021-32-02/S1061-0022-2021-01645-8/#Abstract
- V.I. Danchenko, D.G. Chkalova, *Algebraic Analogs of Fejer Inequalities*, Journal of Mathematical Sciences, 2021, Vol. 255, №. 5, 601-608. http://link.springer.com/article/10.1007/s10958-021-05397-0

2.3 ДИСКРЕТИЗАЦИЯ МЕТОДА

Если тригонометрический многочлен **задан на равномерной сетке узлов**:

$$T_n(\lambda_k), \qquad \lambda_k = \frac{2\pi(k-1)}{m}, \qquad k = 1, \dots, m$$

и при этом

$$m = (s+1)\mu, \qquad 1 \le \mu \le n, \qquad s = \min\{r:\ r\mu - 1 \ge n\}$$

$$\omega = -2\cos\varphi_\alpha, \qquad \varphi_\alpha = \frac{\pi\alpha}{s+1},$$

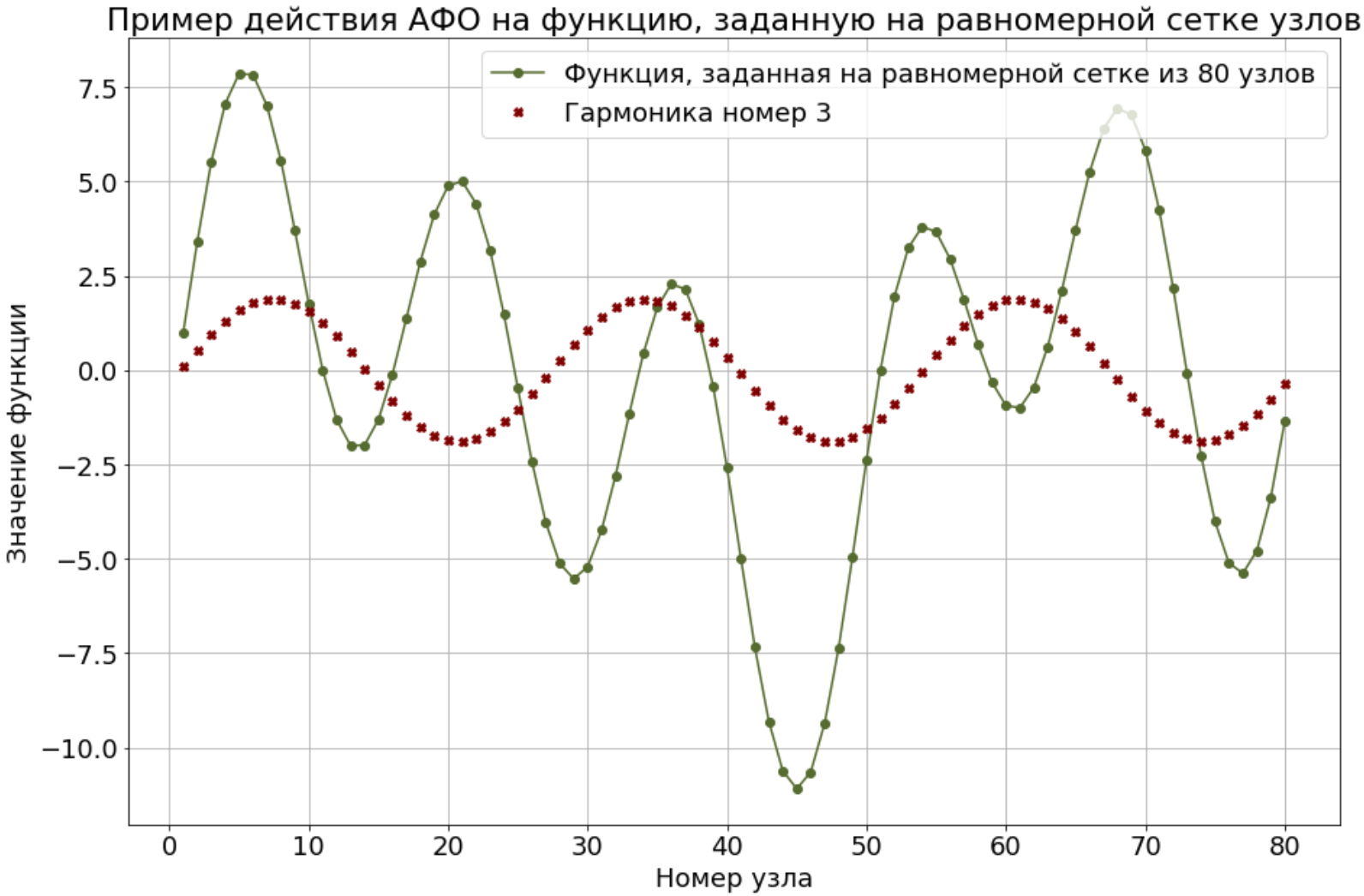
тогда точные значение гармоники $a_\mu \cos\mu t + b_\mu \sin\mu t = \tau_\mu(t)$ в узлах сетки вычисляются по формуле:

$$\tau_\mu(\lambda_k) + a_0\omega = \sum_{j=1}^{k} X_j \cdot T_n\left(\lambda_{k-j+1}\right) + \sum_{j=k+1}^{m} X_j \cdot T_n\left(\lambda_{m+k-j+1}\right),$$

$$X_j = \frac{\omega - 2\cos\mu\lambda_j}{(s+1)\mu}$$

ОСНОВНЫЕ ФОРМУЛЫ ДЛЯ РЕАЛИЗАЦИИ АФО

2.3 ДИСКРЕТИЗАЦИЯ МЕТОДА



2.4 ОСОБЕННОСТИ РЕАЛИЗАЦИИ АФО

При реализации АФО были учтены следующие моменты:

- Номер выделяемой гармоники **не должен превышать** $[N/2]$, где N– число замеров ряда (теорема Котельникова).
- Так как сетка узлов в данных фиксирована, а при реализации АФО она может меняться (незначительно) в зависимости от номера выделяемой гармоники, иногда приходилось **выбирать в качестве узлов неравномерно расположенные точки**, но это обстоятельство не вносит большую погрешность в вычисления, если ряд имеет высокую степень дискретизации (N порядка нескольких десятков замеров или больше).
- Для экономии памяти и ввиду проблемы, описанной в предыдущем пункте, было принято решение **запоминать не всю гармонику, а ее коэффициенты Фурье**, которые вычисляются по формулам:

$$a_\mu = H_m(0) - a_0\omega, \qquad b_\mu = H_m\left(\frac{\pi}{2\mu}\right) - a_0\omega$$

- Было учтено **несоответствие сеток узлов**: расчеты параметров АФО проводятся для сетки узлов, расположенной на $[0,2\pi)$, а замеры временного ряда – либо в натуральных точках, либо в точках формата datetime.

3. АППРОКСИМАЦИЯ ВРЕМЕННОГО РЯДА С ПОМОЩЬЮ АФО

3.1 АППРОКСИМАЦИЯ. ЛИНЕЙНЫЙ ТРЕНД

Для минимизации ошибки на границах интервала аппроксимации (из-за несовпадения значений ряда в граничных точках интервала) было принято решение **исключить из ряда линейный тренд**. После аппроксимации тренд прибавляется к результату

3.1 АППРОКСИМАЦИЯ. ЛИНЕЙНЫЙ ТРЕНД

АППРОКСИМАЦИЯ БЕЗ ЛИНЕЙНОГО ТРЕНДА

АППРОКСИМАЦИЯ С ЛИНЕЙНЫМ ТРЕНДОМ

3.2 АППРОКСИМАЦИЯ. ФУНКЦИИ

В проекте были реализованы следующие функции:

- **Выделение заданной гармоники** из ряда, вычисление ее периода, амплитуды и коэффициентов Фурье (т.е. амплитуд синуса и косинуса)
- **Сглаживание ряда** суммой из M первых гармоник, вычисление MAPE для полученной аппроксимации. Метрика MAPE выбрана, потому что: (1) абсолютная метрика позволяет более адекватно оценивать результат; (2) ряды, которые были рассмотрены в проекте, заведомо не содержат нулевых или близких к нулевым значений
- **Обратная задача** – построение суммы, аппроксимирующей ряд со значением MAPE, не превышающем заранее заданное MAPE_MAX. Учтена возможность некорректного значения MAPE_MAX. Поиск наилучшего приближения производится по всем суммам до порядка $[N/2]$ включительно. Получаем результат со значением

$$\mathrm{MAPE} = \max\{\mathrm{MAPE}_i\colon \mathrm{MAPE}_i \le \mathrm{MAPE\_MAX}\}$$

или, если MAPE_MAX слишком мало:

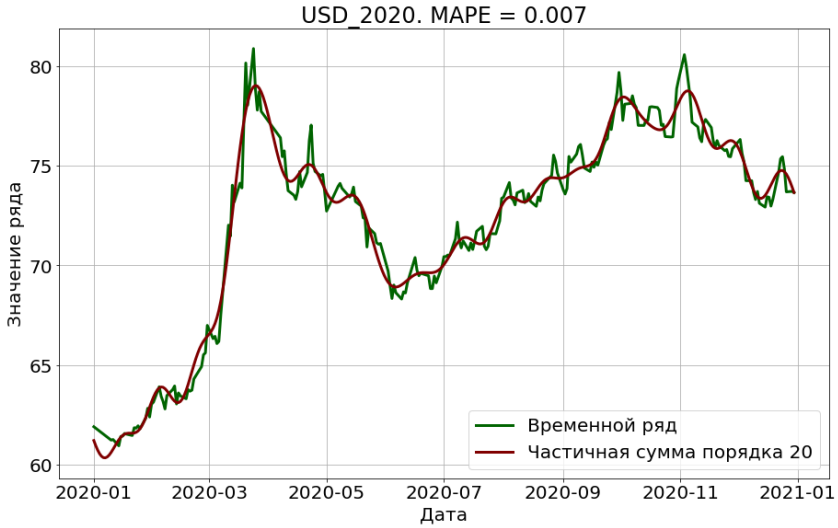
$$\mathrm{MAPE} = \min\{\mathrm{MAPE}_i\}$$

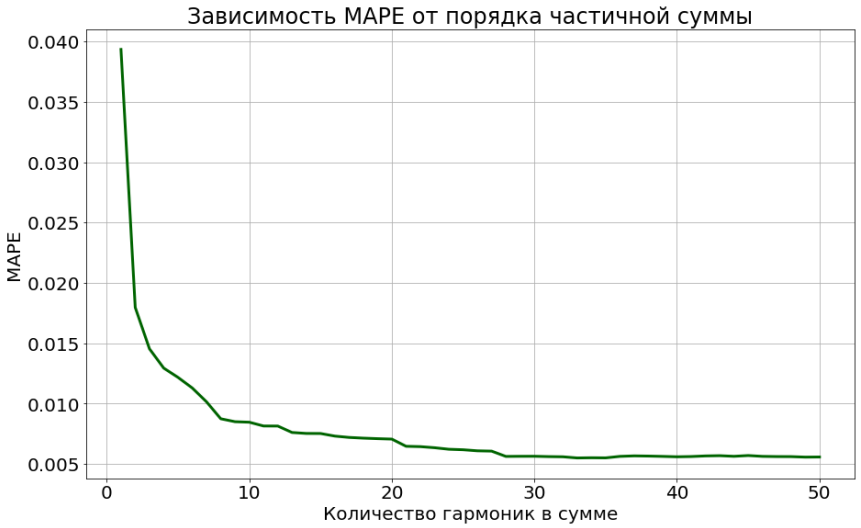
3.3 ВЫДЕЛЕНИЕ ЗАДАННОЙ ГАРМОНИКИ. ПРИМЕР 1

3.3 ВЫДЕЛЕНИЕ ЗАДАННОЙ ГАРМОНИКИ. ПРИМЕР 2

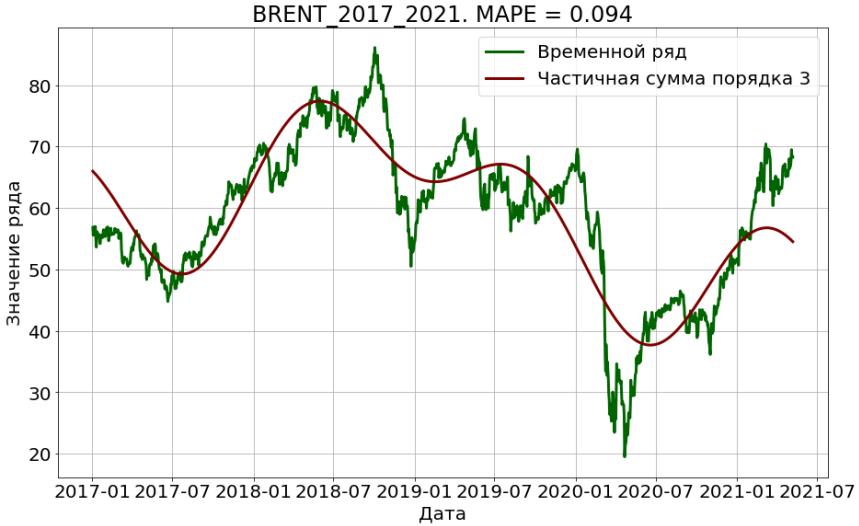
3.3 ВЫДЕЛЕНИЕ ЗАДАННОЙ ГАРМОНИКИ. ПРИМЕР 3

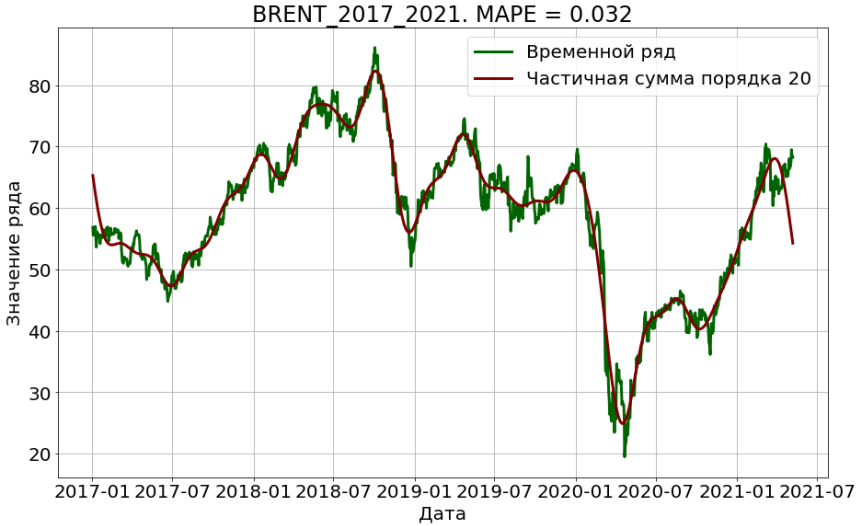
3.4 АППРОКСИМАЦИЯ. ПРИМЕР 1

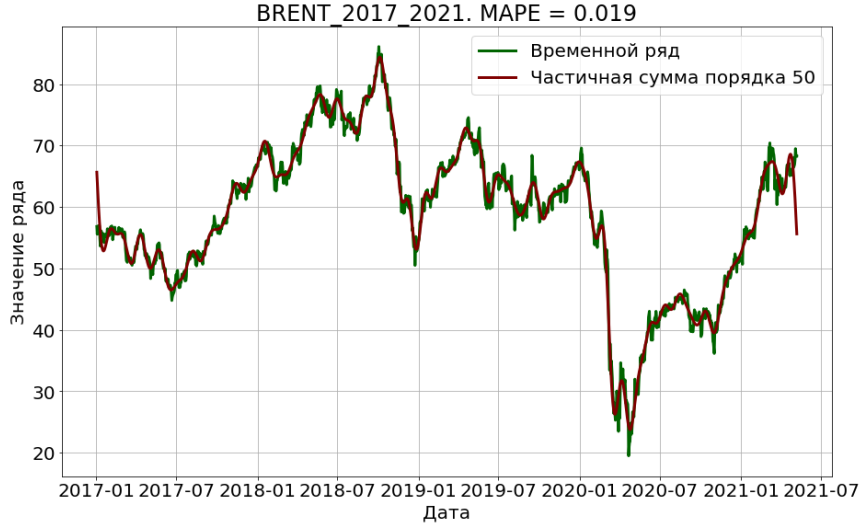


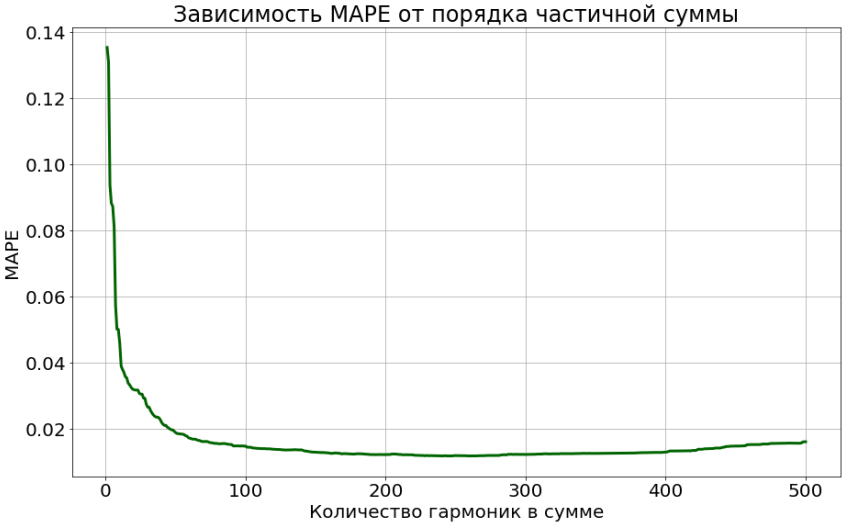


3.4 АППРОКСИМАЦИЯ. ПРИМЕР 2

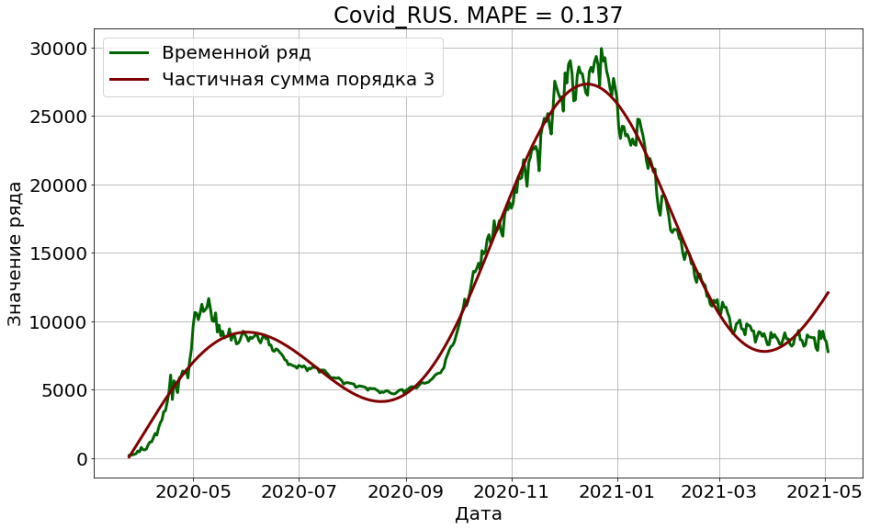


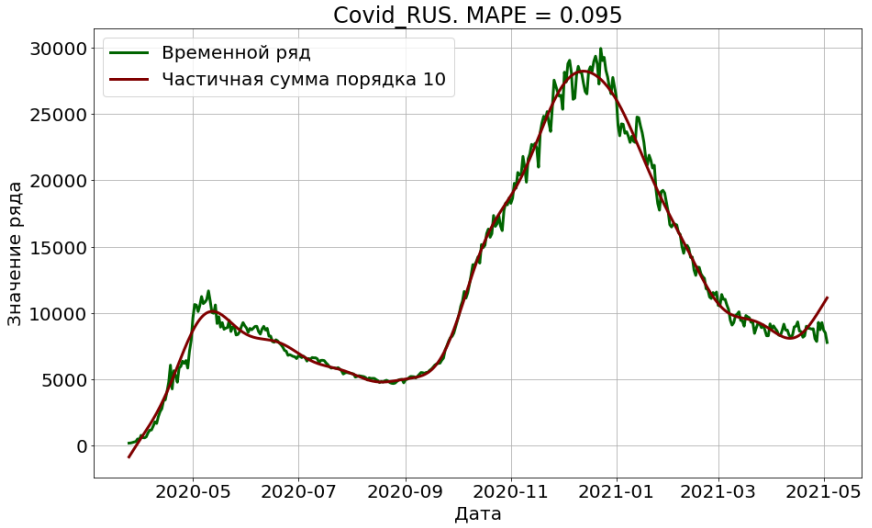


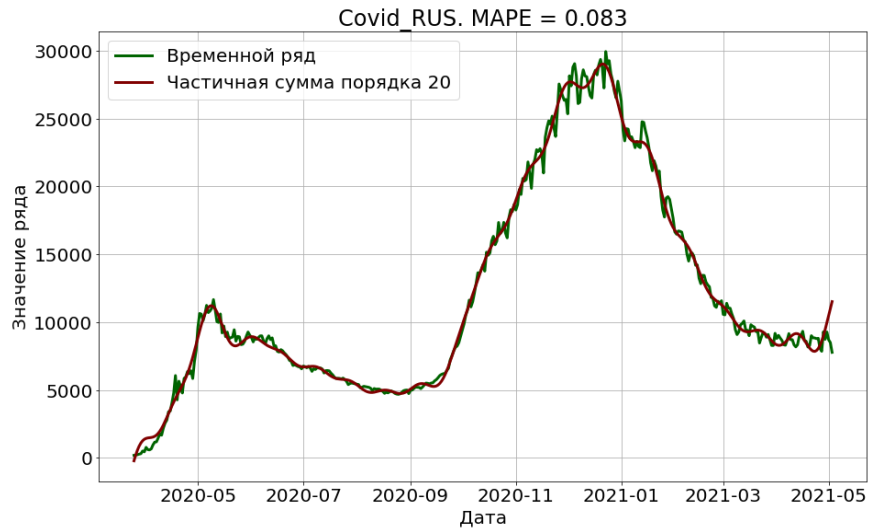


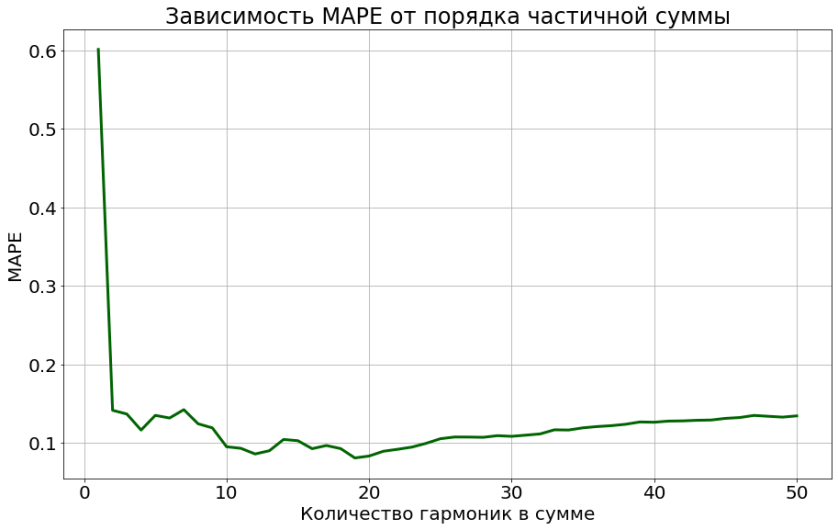


3.4 АППРОКСИМАЦИЯ. ПРИМЕР 3









3.5 ПОСТРОЕНИЕ СУММЫ ПО ЗАДАННОМУ ЗНАЧЕНИЮ MAPE. ПРИМЕР 1

ЗАДАННОЕ ПОРОГОВОЕ ЗНАЧЕНИЕ MAPE

ПОЛУЧЕННОЕ ЗНАЧЕНИЕ MAPE

3.5 ПОСТРОЕНИЕ СУММЫ ПО ЗАДАННОМУ ЗНАЧЕНИЮ MAPE. ПРИМЕР 2

3.5 ПОСТРОЕНИЕ СУММЫ ПО ЗАДАННОМУ ЗНАЧЕНИЮ MAPE. ПРИМЕР 3

Для очень маленького значения MAPE строится наилучшее приближение

4. ПРОГНОЗИРОВАНИЕ ВРЕМЕННОГО РЯДА

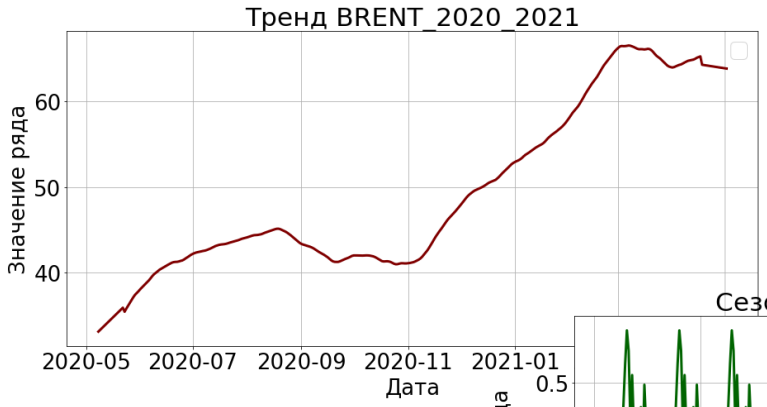
4. ПРОГНОЗИРОВАНИЕ

Экспериментально были получены следующие **оптимальные параметры**:

- Дискретизация данных: 1 день
- Размер периода: 30 дней (1 месяц)
- Количество периодов в тренировочной выборке: от 5 до 11 месяцев
- Количество периодов в тестовой выборке: 1 месяц
- Отсутствие валидационной выборки связано с тем, что качество предсказания при подборе макропараметров на валидационной выборке не улучшает, а только ухудшает результат

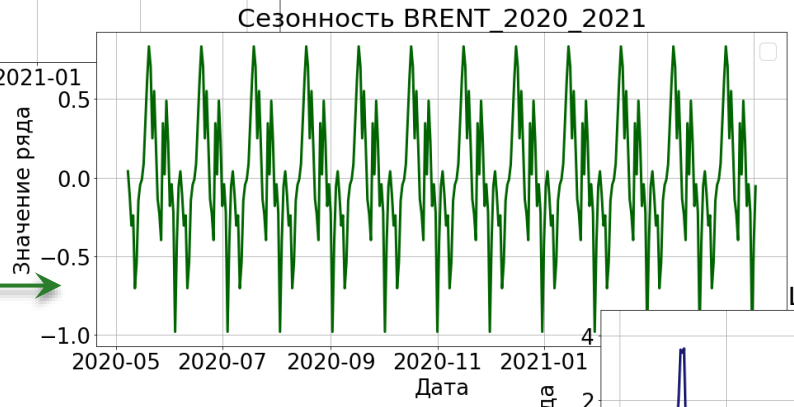
4.1 ПРОГНОЗИРОВАНИЕ. STL-РАЗЛОЖЕНИЕ

Типичная картина STL-разложения рассмотренных в проекте временных рядов:

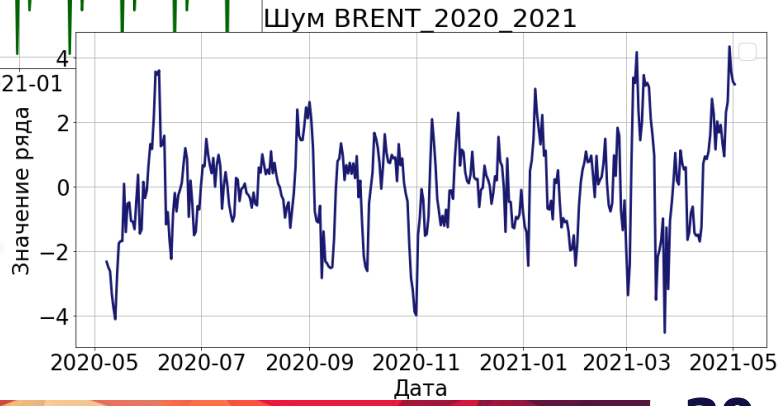


ТРЕНД ИМЕЕТ СТРУКТУРУ, БЛИЗКУЮ К КУСОЧНО-ЛИНЕЙНОЙ

СЕЗОННОСТЬ НЕЗНАЧИТЕЛЬНА (ПО МОДУЛЮ)

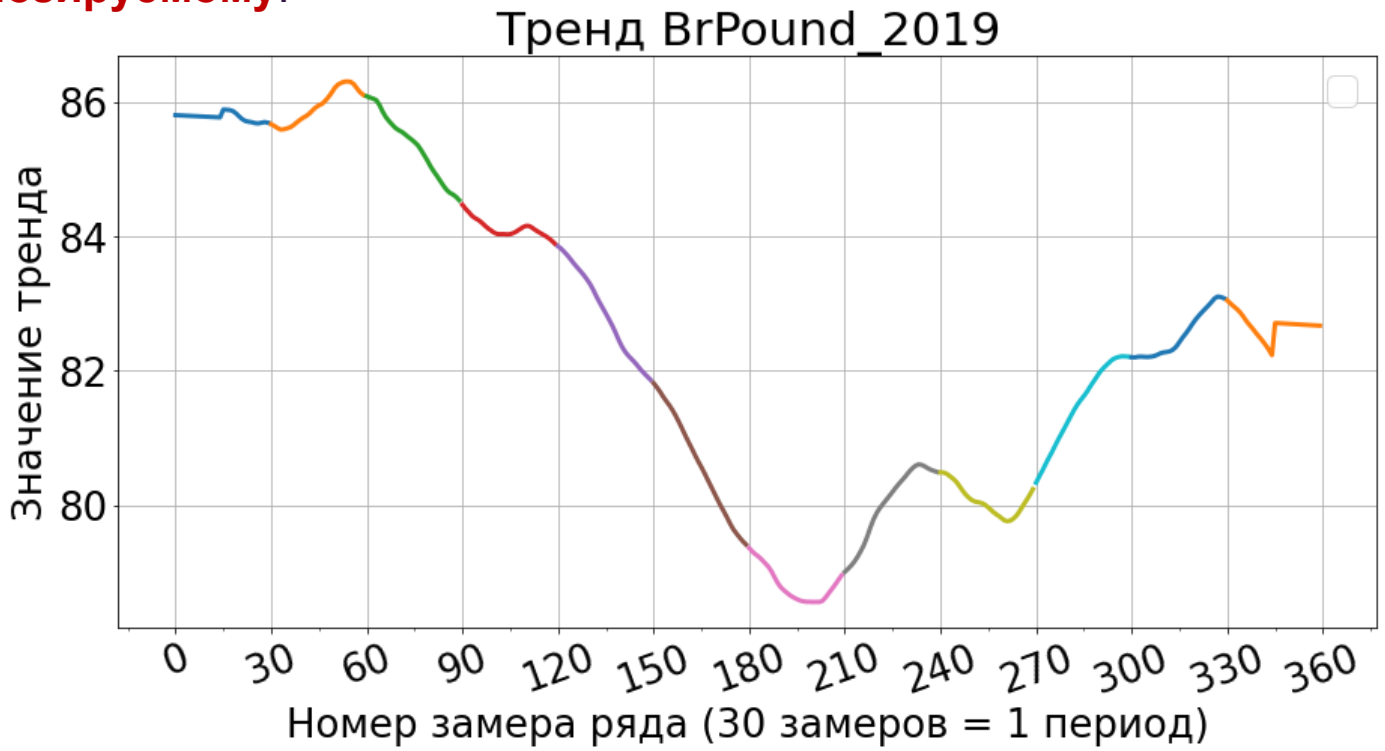


ЗНАЧИТЕЛЬНЫЙ ШУМ СПЕЦИФИЧЕСКОЙ СТРУКТУРЫ



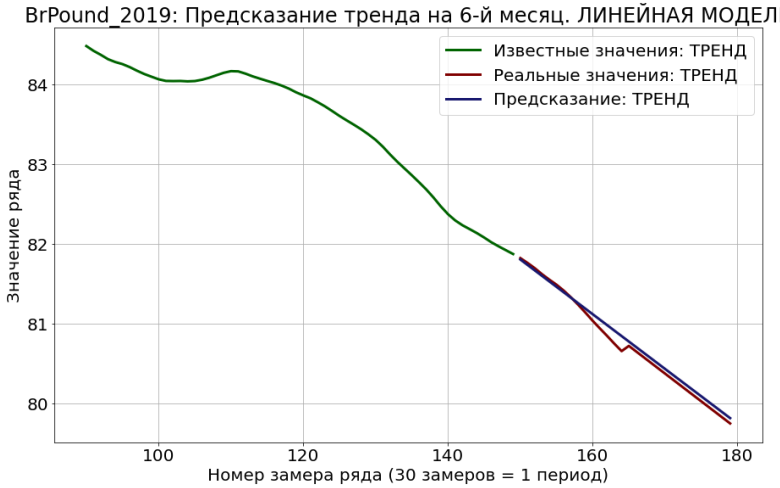
4.2 ПРОГНОЗИРОВАНИЕ. ТРЕНД. ВАРИАНТ 1

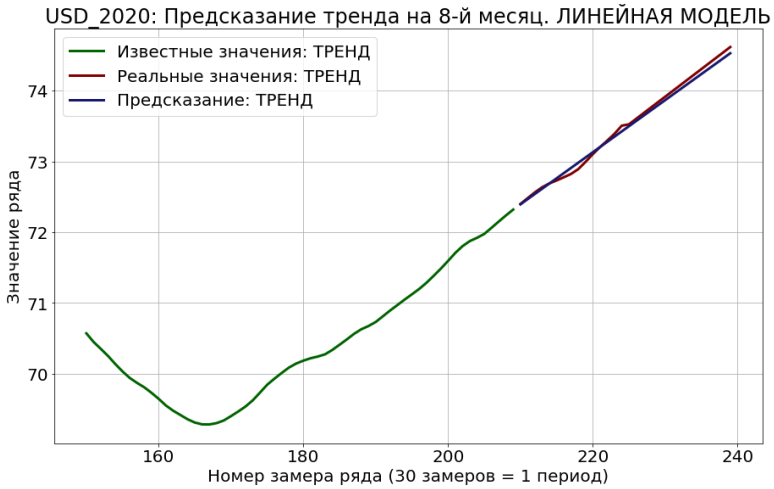
Ранее было отмечено, что в рассматриваемых временных рядах тренд имеет структуру, близкую к кусочно-линейной, причем линейность часто сохраняется на протяжении нескольких периодов. Это позволяет в ряде случаев делать качественный прогноз **линейной экстраполяцией периода, предшествующего прогнозируемому**.



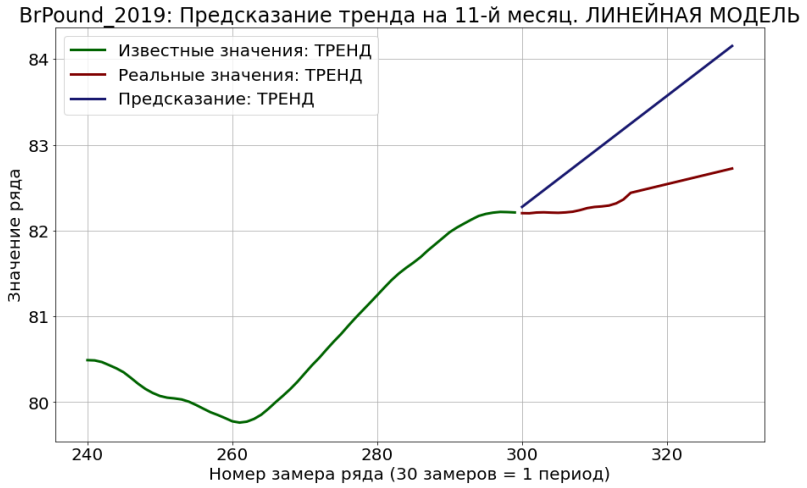
4.2 ПРОГНОЗИРОВАНИЕ. ТРЕНД. ВАРИАНТ 1

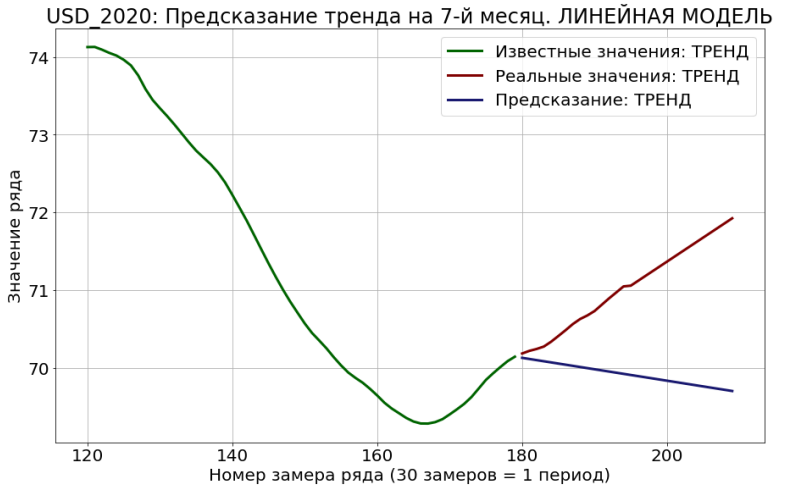
Положительные примеры





Отрицательные примеры





4.2 ПРОГНОЗИРОВАНИЕ. ТРЕНД. ВАРИАНТ 2

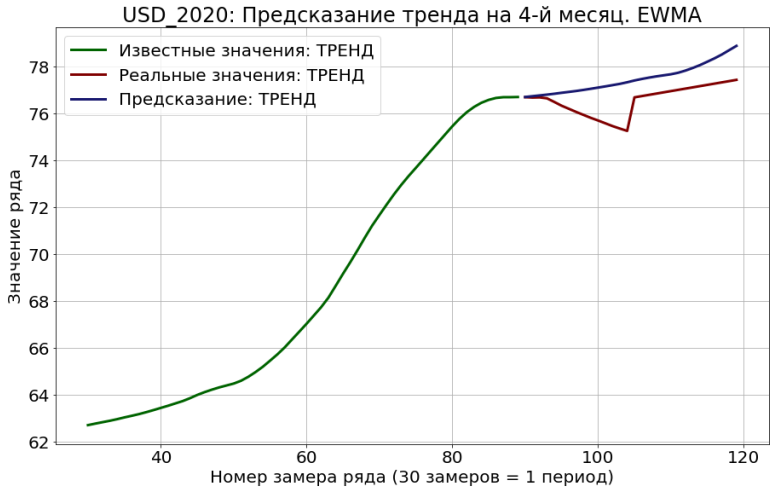
Альтернативный метод прогнозирования тренда – экстраполяция по известным значениям с помощью **экспоненциального скользящего среднего** (EWMA).

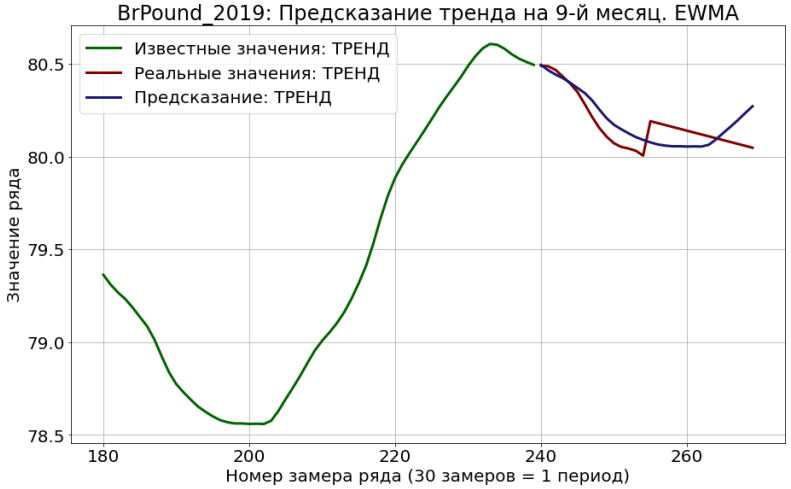
По известным W значениям тренда с помощью экспоненциальной скользящей средней (с параметром α) строится прогноз на одно наблюдение. Затем окно для EWMA сдвигается на единицу и аналогичным образом получаем следующее предсказание.

В примерах далее $W = 60, \alpha = 0.01$

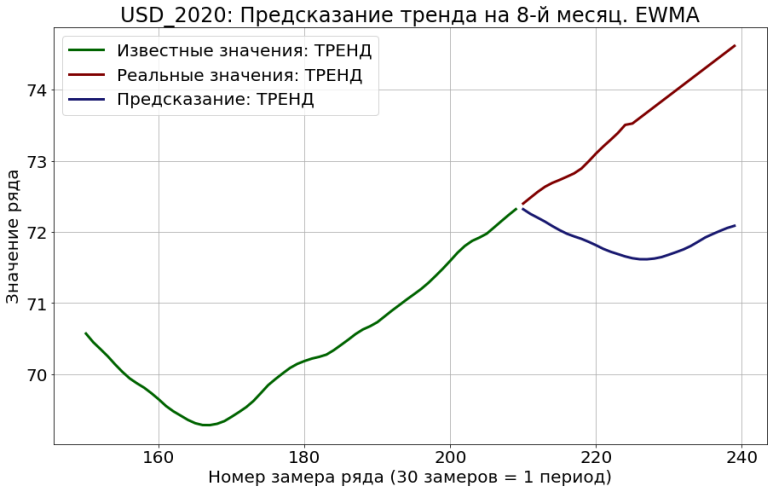
4.2 ПРОГНОЗИРОВАНИЕ. ТРЕНД. ВАРИАНТ 2

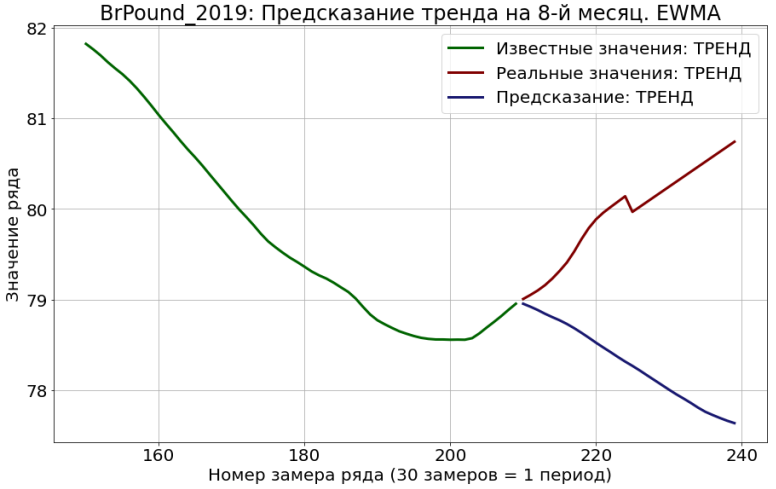
Положительные примеры





Отрицательные примеры





4.3 ПРОГНОЗИРОВАНИЕ. ШУМ

В процессе анализа шума было отмечено, что в нем присутствует **отчетливая структура**, а именно: некоторые гармоники (высокочастотные) имеют значительную амплитуду и хорошо сглаживают ряд. Если они доминируют на тренировочной выборке, то с большой долей вероятности будут и в будущем.

На графике приведен шум временного ряда GOLD_2013 (11 периодов по 30 замеров в каждом) и его аппроксимация гармониками с номерами [9, 13, 8, 18, 15], которые имеют амплитуды [15.25, 13.72, 11.15, 10.18, 9.28] соответственно.

4.3 ПРОГНОЗИРОВАНИЕ. ШУМ

Возникает вопрос: как найти нужные доминирующие гармоники. Было принято следующее решение.

Шаг 1. Находим процентиль $Q(\alpha)$ уровня α (близкого к единице) выборки, полученной из абсолютных значений шума. Другими словами, находим «типичную амплитуду» для ряда. Например, на приведенном графике это значение должно около 50. При $\alpha = 0.95$ получаем $Q(0.95) = 52.9$

Шаг 2. Находим номера гармоник ряда, которые имеют наибольшую амплитуду и их сумма максимально близка к $Q(\alpha)$

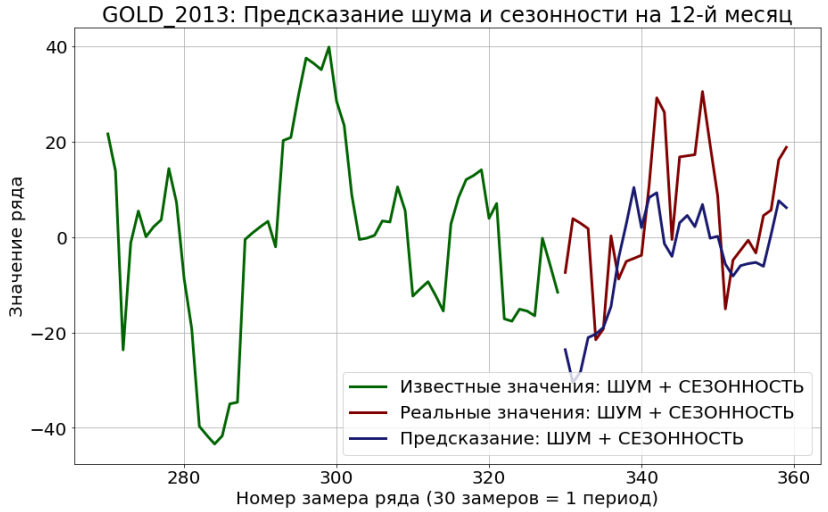
Гармоника	Амплитуда	Сумма
…	…	…
2	2,044709	120,35367
19	2,376922	118,30896
21	2,386834	115,93204
4	3,228641	113,54521
16	4,059446	110,31657
17	4,585459	106,25712
7	6,199014	101,67166
20	6,417968	95,472646
10	6,510667	89,054678
5	6,71408	82,544011
6	7,536077	75,829931
12	8,674401	68,293854
15	9,284523	59,619453
18	**10,189126**	**50,33493**
8	**11,159876**	40,145804
13	**13,72628**	28,985928
9	**15,259647**	15,259648

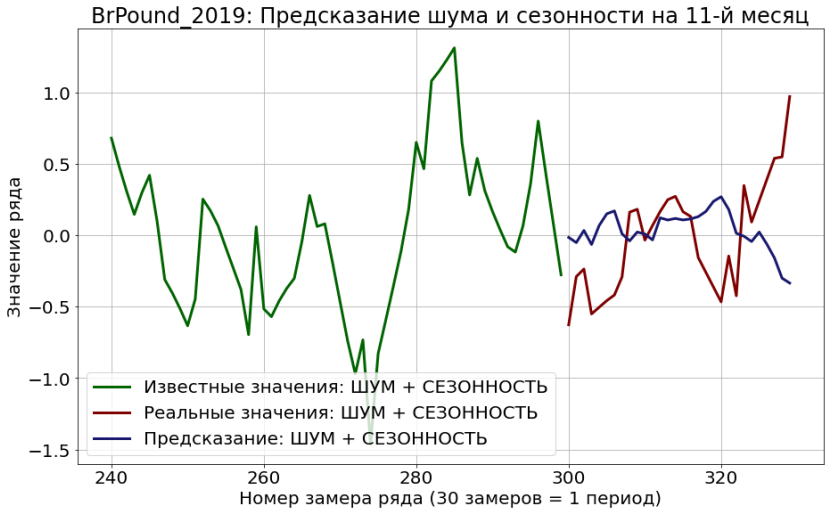
4.3 ПРОГНОЗИРОВАНИЕ. ШУМ

ПРИМЕРЫ РЕЗУЛЬТАТОВ

4.4 СЕЗОННОСТЬ + ШУМ

Приведем примеры объединения прогноза шума и полученной при STL-разложении сезонности





4.5 ТРЕНД + СЕЗОННОСТЬ + ШУМ

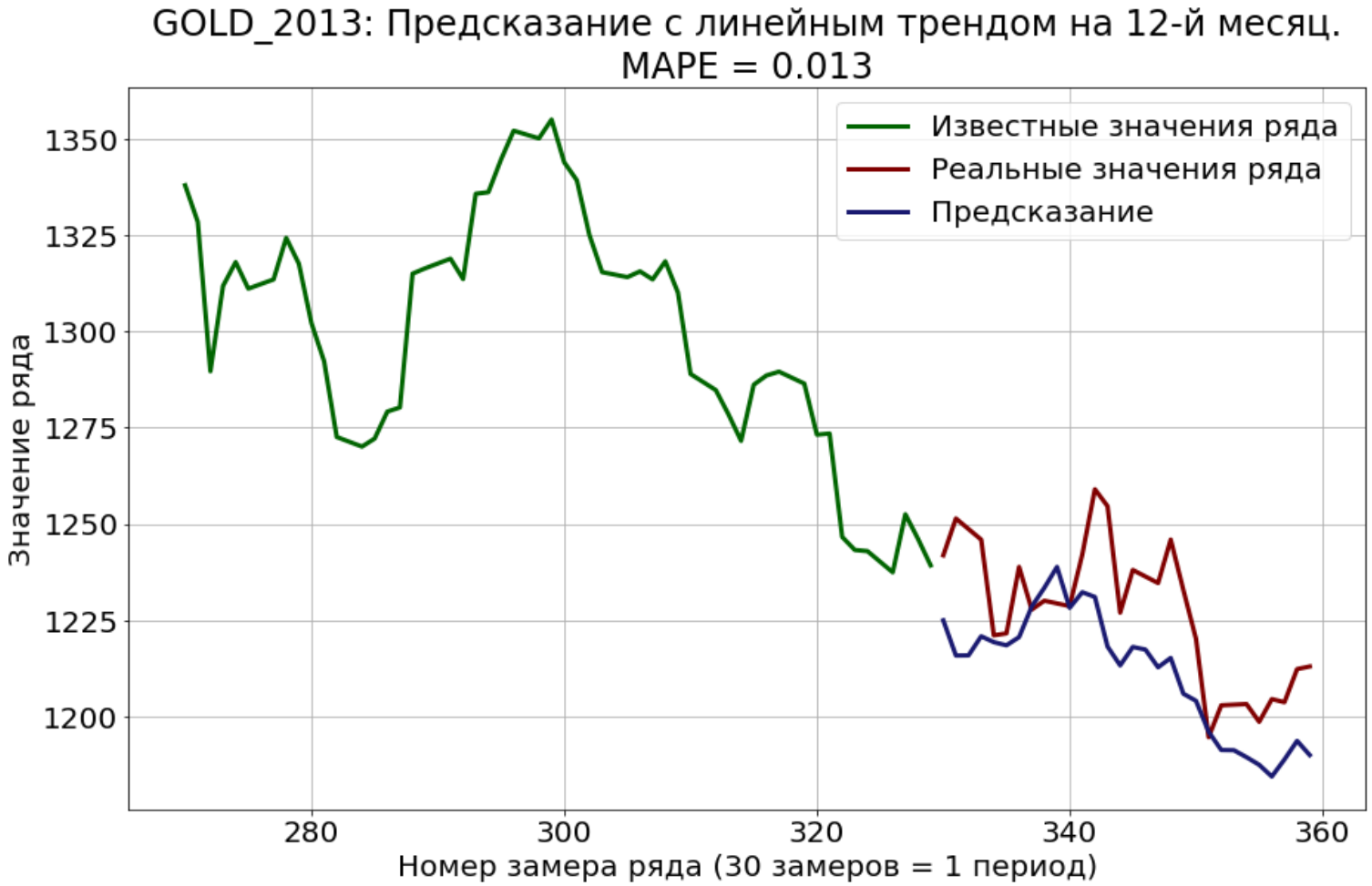
ПРИМЕР 1

4.5 ТРЕНД + СЕЗОННОСТЬ + ШУМ

ПРИМЕР 2

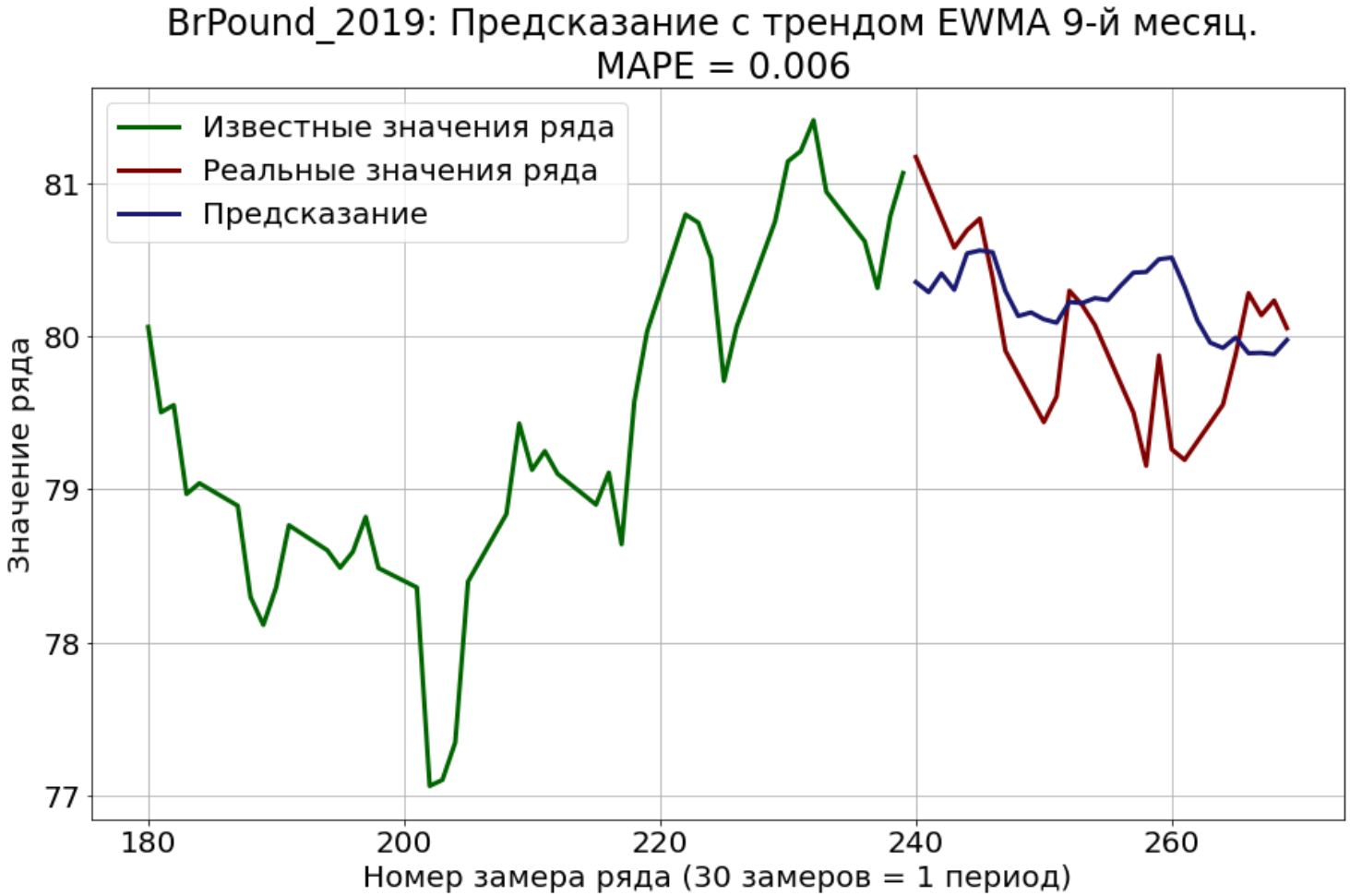
4.5 ТРЕНД + СЕЗОННОСТЬ + ШУМ

ПРИМЕР 3



4.5 ТРЕНД + СЕЗОННОСТЬ + ШУМ

ПРИМЕР 4



4.5 ТРЕНД + СЕЗОННОСТЬ + ШУМ

ПРИМЕР 5

5. Прогнозирование ряда с помощью пакета prophet

5.1 Facebook Prophet

В 2017 г. специалисты компании **Facebook** объявили о разработанном ими новом пакете для прогнозирования временных рядов – **Prophet**. Prophet позволяет в автоматическом режиме создавать модели, обладающие высокой точностью предсказаний.

В основе методологии этого пакета лежит процедура обучения **аддитивных регрессионных моделей** со следующими четырьмя основными компонентами:

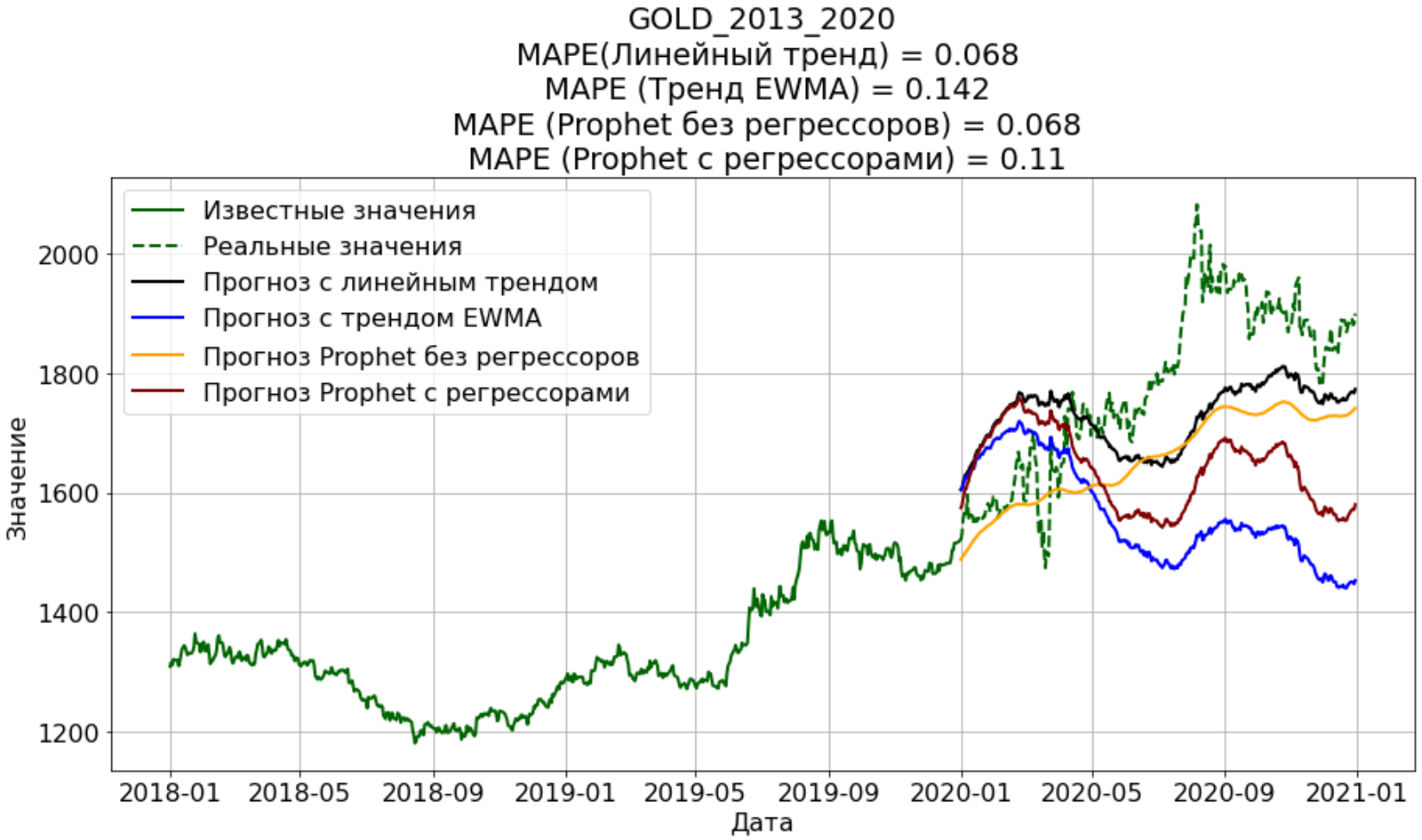
- тренд (моделируется с помощью кусочно-линейной регрессии или кусочной логистической кривой роста);
- годовая сезонность (моделируется как ряд Фурье);
- недельная сезонность (моделируется с использованием индикаторной переменной);
- праздники.

Оценивание параметров обучаемой модели выполняется с **использованием принципов байесовской статистики** (либо методом нахождения апостериорного максимума (MAP), либо путем полного байесовского вывода). Для этого применяется платформа вероятностного программирования Stan. Prophet представляет собой ни что иное, как удобный интерфейс для работы с этой платформой из среды R (имеется также аналогичная библиотека для Python).

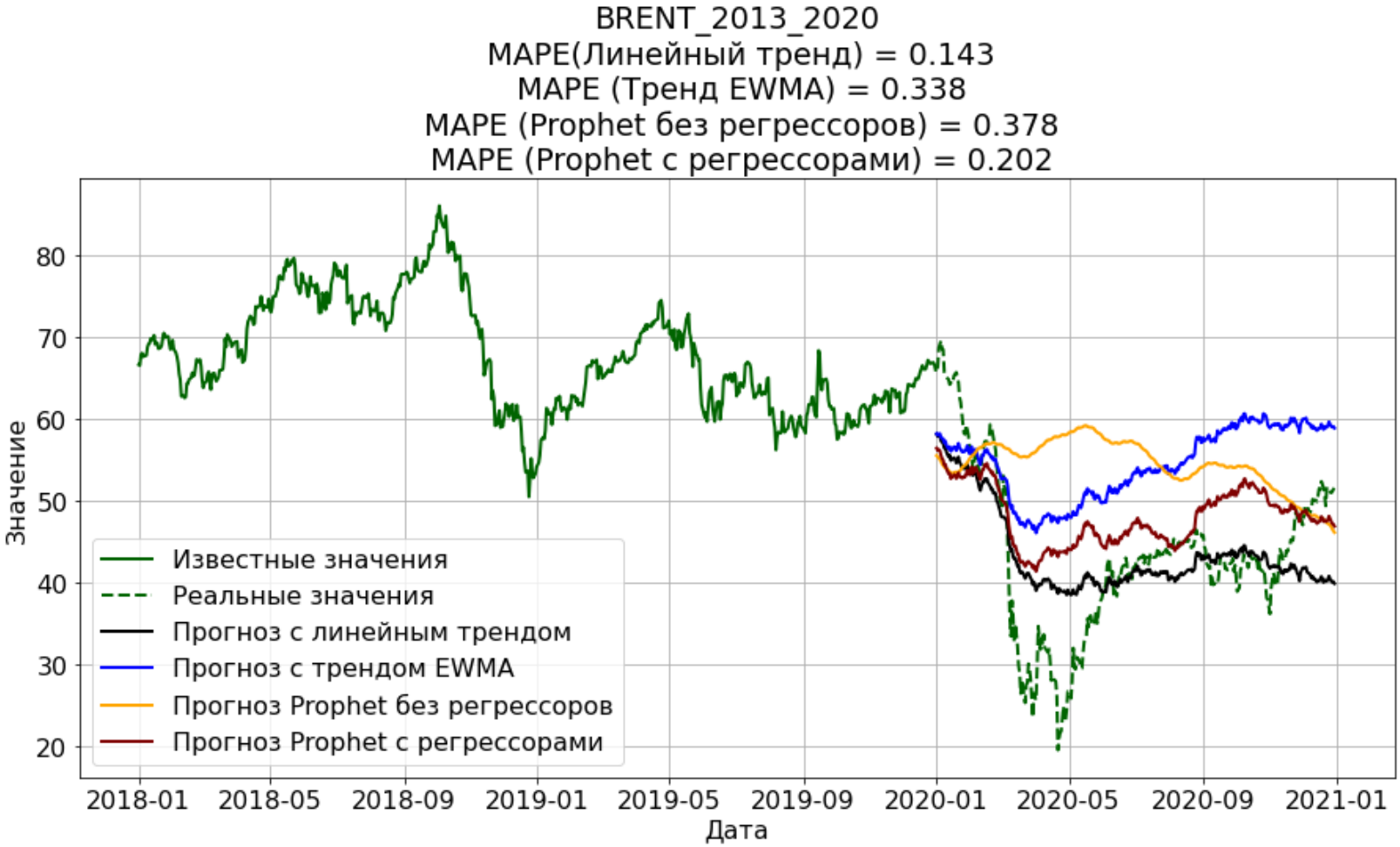
5.2 Построение прогнозов

- Строим прогноз с помощью пакета prophet по данным за 7 лет на 8-й год
- Строим прогноз с помощью пакета prophet, добавив два регрессора, которые в прошлом равны сглаженному гармониками ряду, а в будущем - прогнозам, построенным в пункте 4.

5.3 ПРИМЕР 1


GOLD_2013_2020
MAPE(Линейный тренд) = 0.068
MAPE (Тренд EWMA) = 0.142
MAPE (Prophet без регрессоров) = 0.068
MAPE (Prophet с регрессорами) = 0.11
Известные значения
Реальные значения
Прогноз с линейным трендом
Прогноз с трендом EWMA
Прогноз Prophet без регрессоров
Прогноз Prophet с регрессорами
Значение
2000
1800
1600
1400
1200
2018-01
2018-05
2018-09
2019-01
2019-05
2019-09
2020-01
2020-05
2020-09
2021-01
Дата

5.3 ПРИМЕР 2


BRENT_2013_2020
MAPE(Линейный тренд) = 0.143
MAPE (Тренд EWMA) = 0.338
MAPE (Prophet без регрессоров) = 0.378
MAPE (Prophet с регрессорами) = 0.202
Значение
Дата
Известные значения
Реальные значения
Прогноз с линейным трендом
Прогноз с трендом EWMA
Прогноз Prophet без регрессоров
Прогноз Prophet с регрессорами
80
70
60
50
40
30
20
2018-01
2018-05
2018-09
2019-01
2019-05
2019-09
2020-01
2020-05
2020-09
2021-01

СПАСИБО ЗА ВНИМАНИЕ