Анализ временных рядов

Сусла Диана, БПМИ182

Что это такое?

Временной ряд — это последовательность значений признака, измеренных через постоянные временные интервалы.

$$Y = Y_1, Y_2, ..., Y_t, ...$$

где $\forall i \ Y_i \in \mathbb{R}$ — значение признака в момент времени *i*.

Пример

(Данные взяты с соревнования на Kaggle “Прогноз популярности статьи на Хабре”)

Зачем это нужно?

Прогнозирование временного ряда — это предсказание следующего значения признака или нескольких следующих значений признака в зависимости от уже имеющегося временного ряда.

Имеем:

$$Y_1, Y_2, ..., Y_t, ...$$

Хотим найти:

Функцию f_t, такую что:

$$\hat{Y}_{t+d}(w) = f_t(Y_1, ..., Y_t; w)$$

где w - вектор параметров модели, $d \in \{1, 2, ..., D\}$

D - горизонт прогнозирования

Отличия прогнозирования временных рядов от других задач машинного обучения

- Данные находятся не в произвольном порядке, а упорядочены по времени.
- Данные должны быть зависимы. Таким образом по значениям ряда в прошлом можно будет предугадать его поведение в будущем. Чем сильнее будущее зависит от прошлого, тем точнее можно сделать прогноз.

Где применяется?

- Прогнозирование объёмов продаж
- Анализ фондовых рынков
- Прогнозирование объемов потребления электроэнергии
- Прогнозирование объемов перевозок
- Прогнозирование пробок
- и т.д.

Основные свойства временных рядов

- ***Сезонность*** — циклическое изменение параметров ряда с постоянным периодом, связанное с сезонами и ритмами активности человека.
- ***Тренд*** — плавное изменение параметров временного ряда, проходящее в некотором определенном направлении, которое сохраняется в течение значительного промежутка времени.

- ***Стационарность*** — свойство, при котором не изменяется распределение вероятности — среднее значение, дисперсия и ковариация ряда не изменяются со временем.

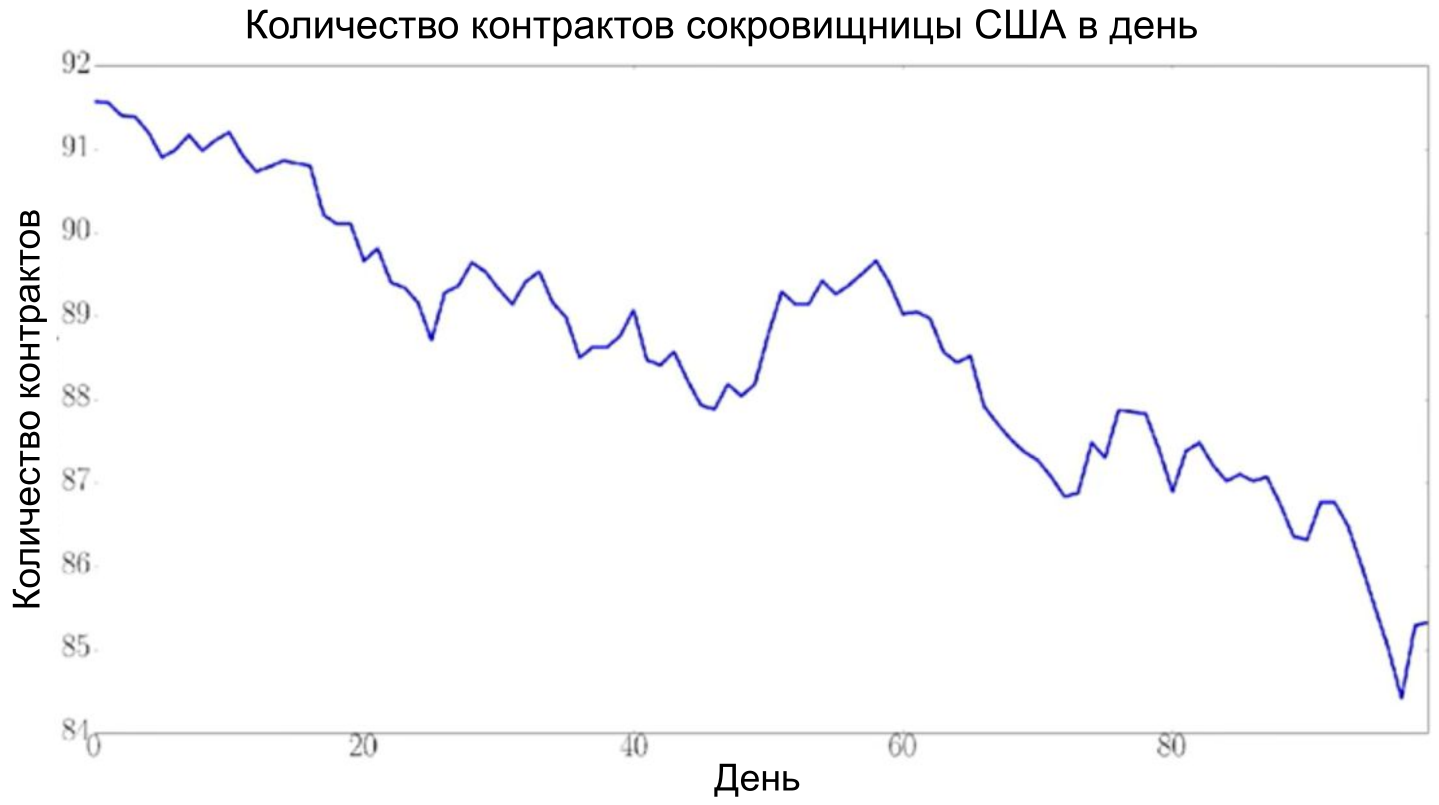
$$\mathbb{E}(Y_1) = \mathbb{E}(Y_2) = .. = \mathbb{E}(Y_t) = ... = const$$
$$\mathbb{D}(Y_1) = \mathbb{D}(Y_2) = ... = \mathbb{D}(Y_t) = ... = const$$
$$cov(Y_1, Y_2) = ... = cov(Y_{t-1}, Y_t) = ... = const$$

То есть:

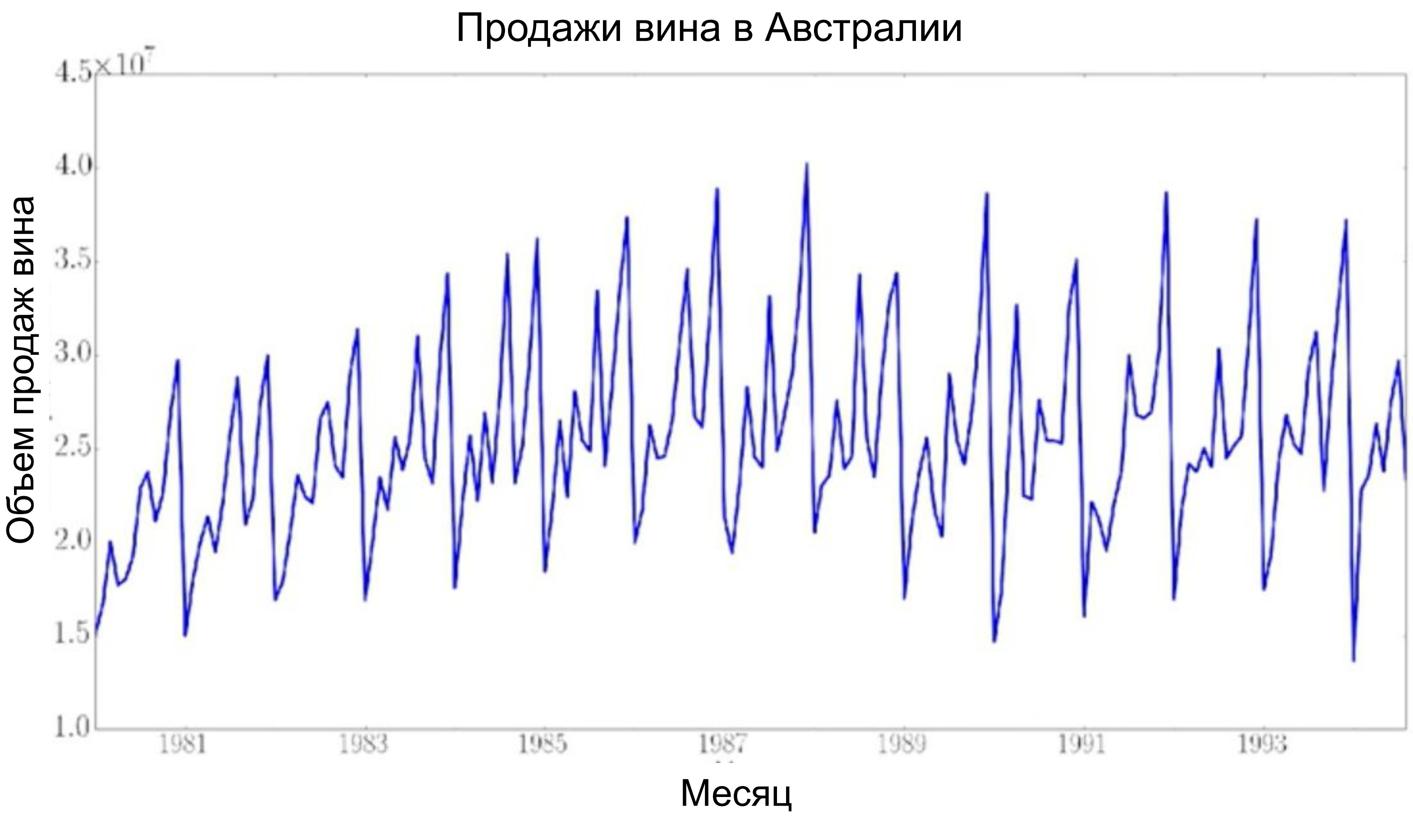
$$cov(Y_{t-k}, Y_t) = \gamma_k$$

Что значит, что ковариация между двумя показателями не зависит от их значений, а зависит только от разницы по времени между этими показателями. Функцию γ_k называют *автоковариационной функцией*.

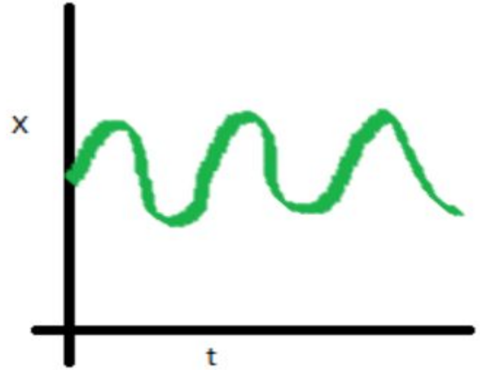
Тренд


Количество контрактов сокровищницы США в день
92
91
90
89
88
87
86
85
84
0
20
40
60
80
Количество контрактов
День

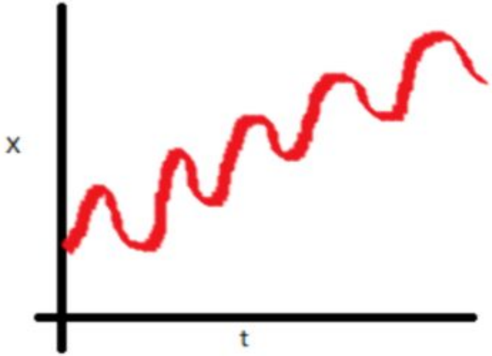
Тренд + сезонность



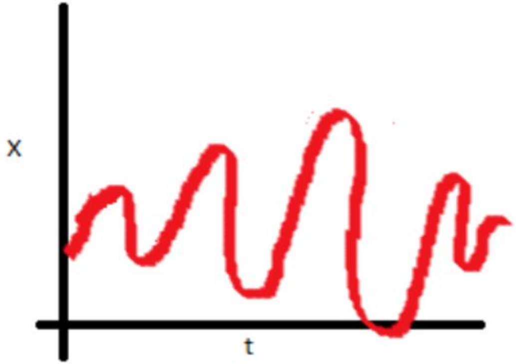
Стационарность



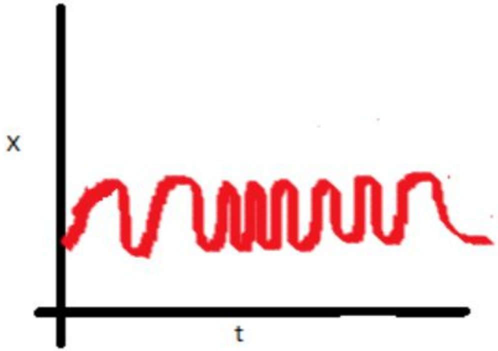
Stationary series



Non-Stationary series



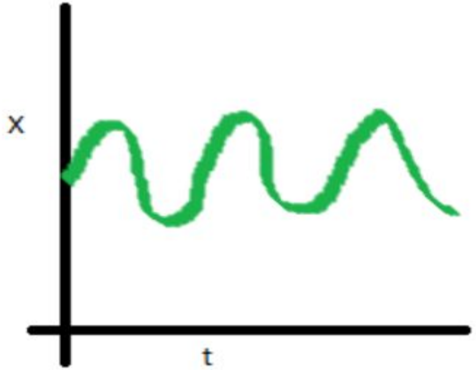
Non-Stationary series



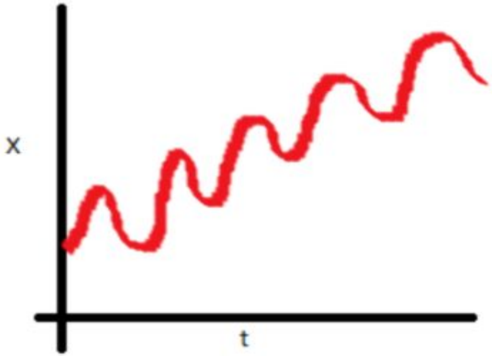
Non-Stationary series

Стационарность

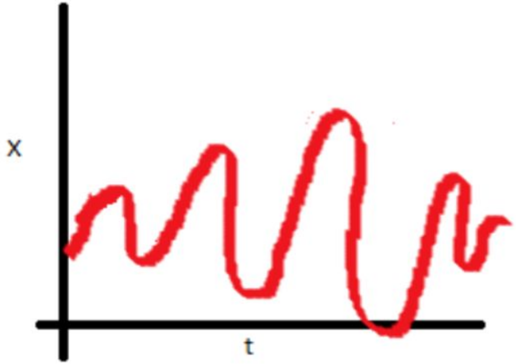
тут растет
матожидание



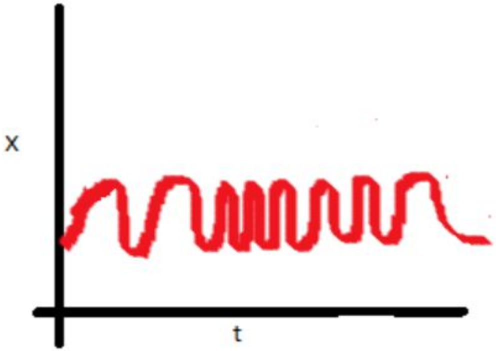
Stationary series



Non-Stationary series



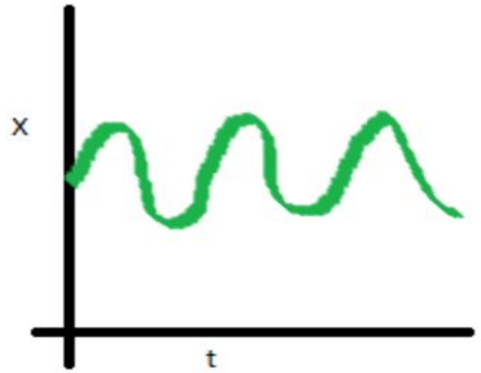
Non-Stationary series



Non-Stationary series

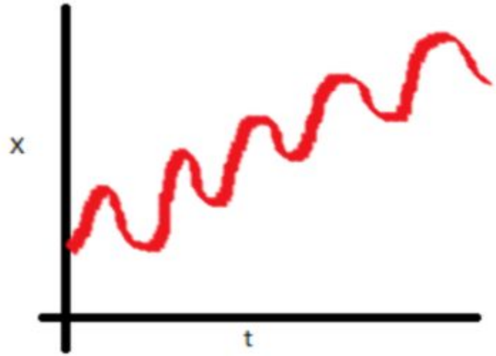
Стационарность

тут растет матожидание

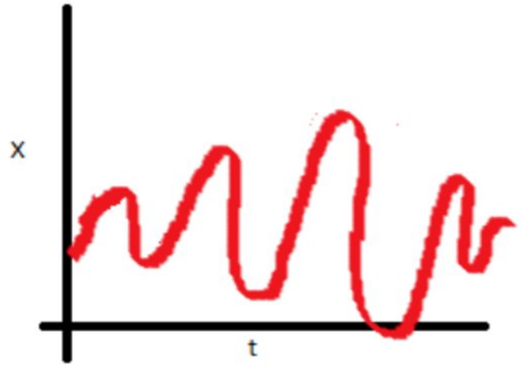


Stationary series

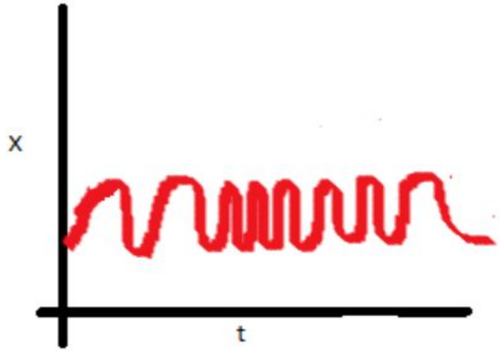
тут нестабильная дисперсия



Non-Stationary series



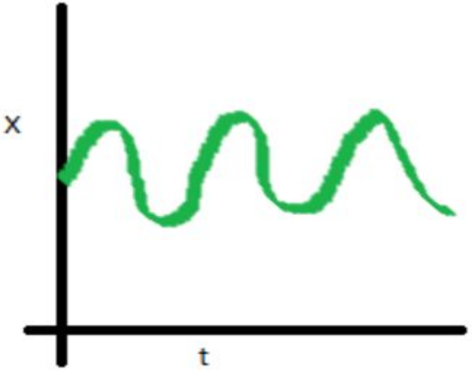
Non-Stationary series



Non-Stationary series

Стационарность

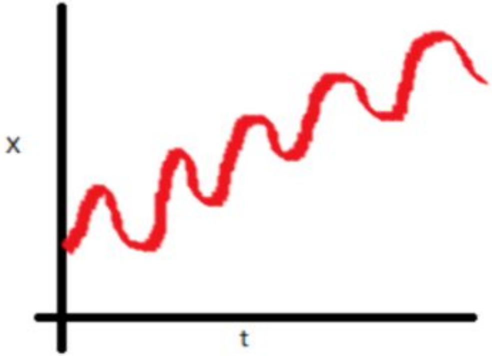
тут растет
матожидание



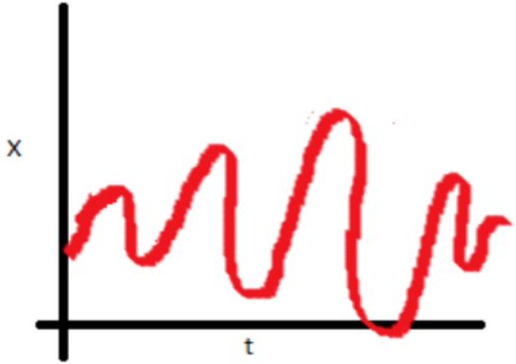
Stationary series

тут нестабильная
дисперсия

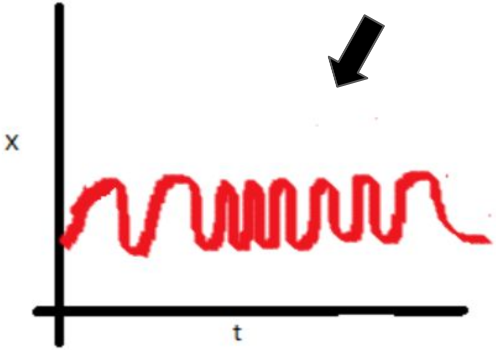
тут
непостоянная
ковариация



Non-Stationary series



Non-Stationary series



Non-Stationary series

Критерии проверки стационарности ряда

- Критерий KPSS

(Квятковского - Филлипса - Шмидта - Шина)

- Критерий Дики - Фуллера

(DF-тест)

Сведение нестационарного ряда к стационарному

- ***Дифференцирование*** ряда — это переход к попарным разностям его соседних значений. То есть:

$$Y_1, Y_2, ..., Y_t \to Z_2, Z_3, ..., Z_t,\ \ Z_i = Y_i - Y_{i-1}$$

При помощи дифференцирования ряда можно избавиться от тренда и сезонности, а также стабилизировать математическое ожидание.

- ***Логарифмирование*** ряда — применение логарифмирования к каждому члену ряда:

$$Y_1, Y_2, ..., Y_t \to \ln(Y_1), \ln(Y_2), ..., \ln(Y_t),\ \ Y_i > 0$$

Данная техника полезна для рядов с не постоянной дисперсией.

Преобразование Бокса-Кокса

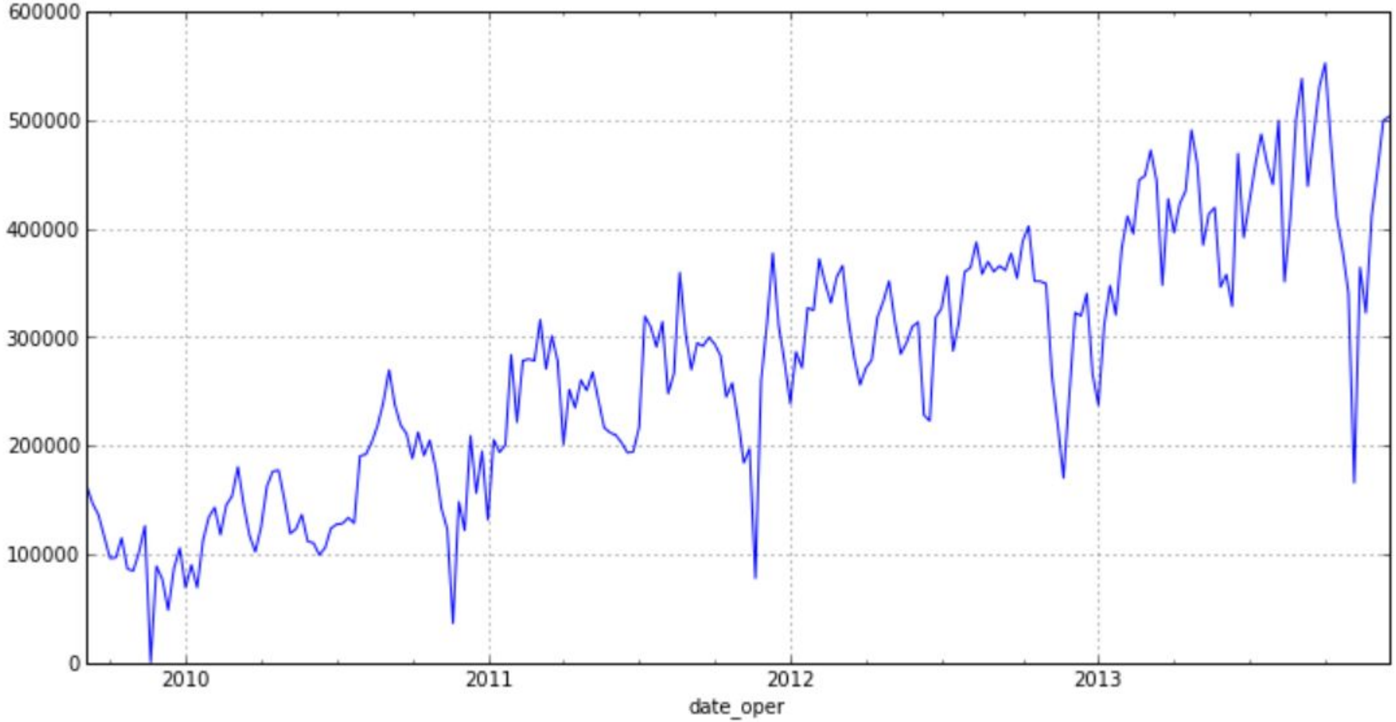
$$\forall i \; Y_i' = \begin{cases} \ln(Y_i), \; \lambda = 0 \\ \frac{{Y_i}^{\lambda} - 1}{\lambda}, \; \lambda \neq 0 \end{cases}$$

Параметр λ можно выбирать, максимизируя логарифм правдоподобия.

<u>Обратное преобразование:</u>

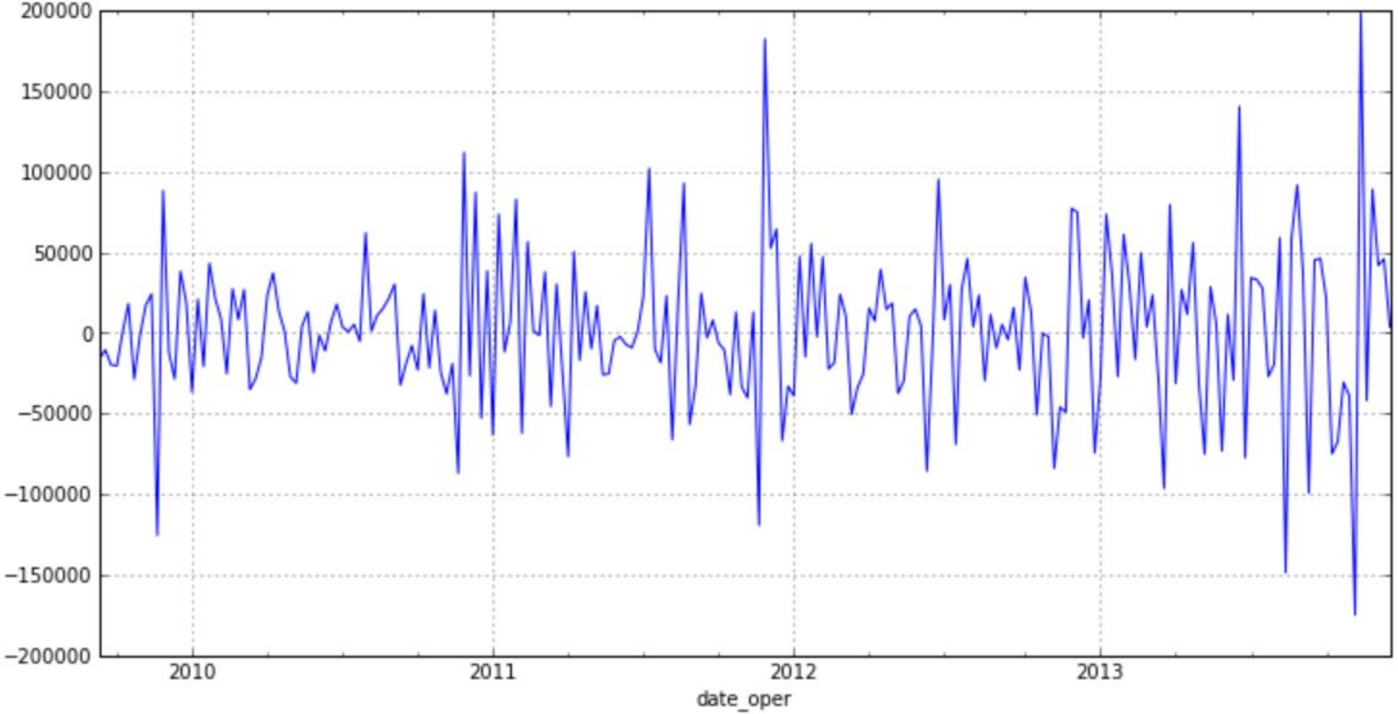
$$\forall i \; \hat{Y}_i = \begin{cases} \exp(\hat{Y}_i'), \; \lambda = 0 \\ (\lambda \hat{Y}' + 1)^{1/\lambda}, \; \lambda \neq 0 \end{cases}$$

Как это работает?

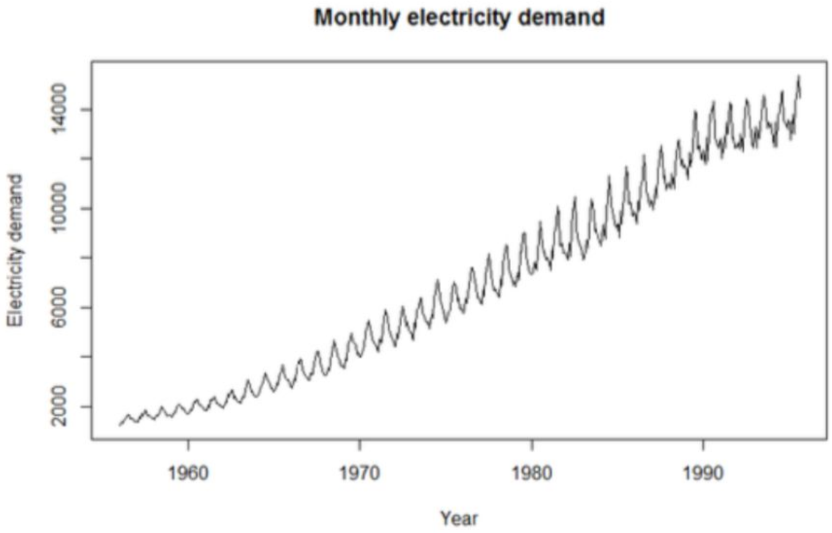


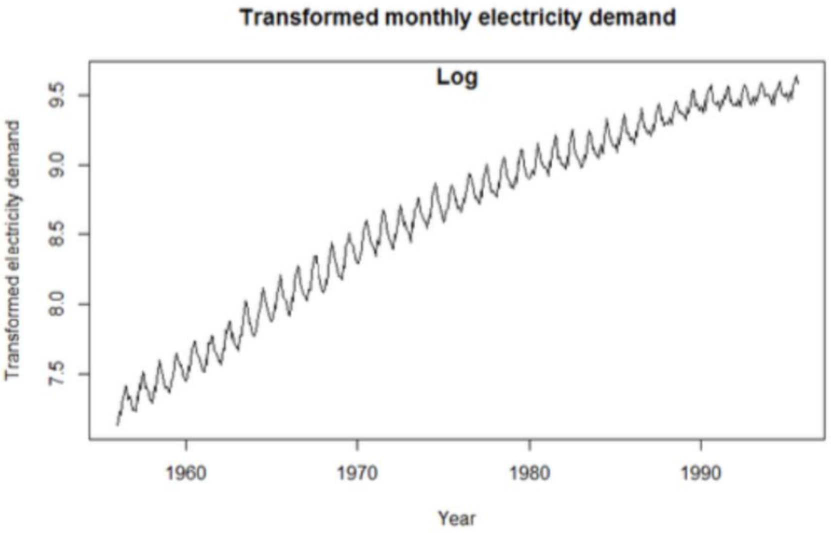
Данные по отгрузке товаров одного из складских комплексов Подмосковья.

Тот же ряд после дифференцирования



Как работает логарифмирование



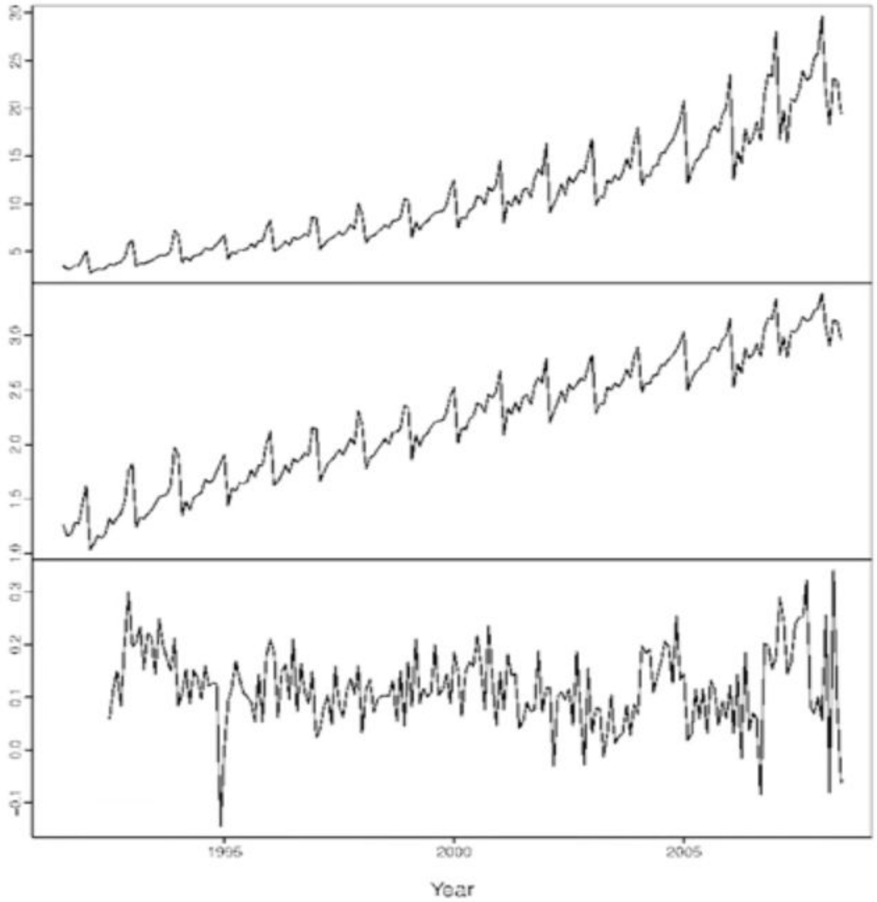


Исходный ряд

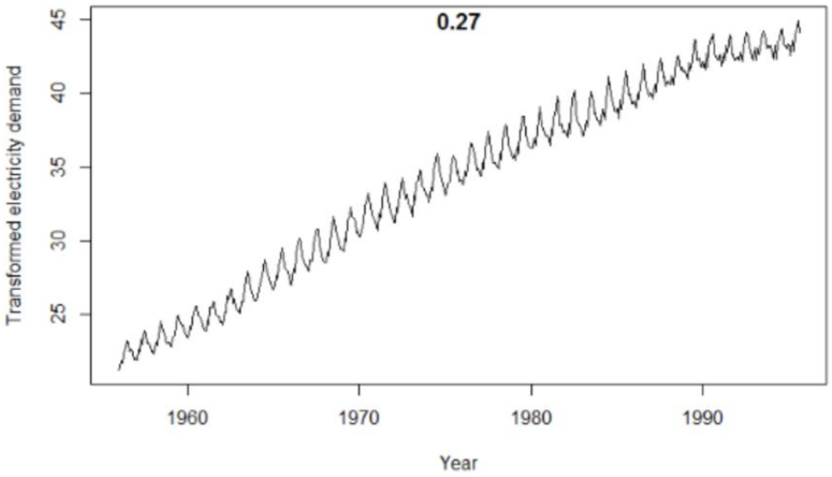
Ряд после логарифмирования

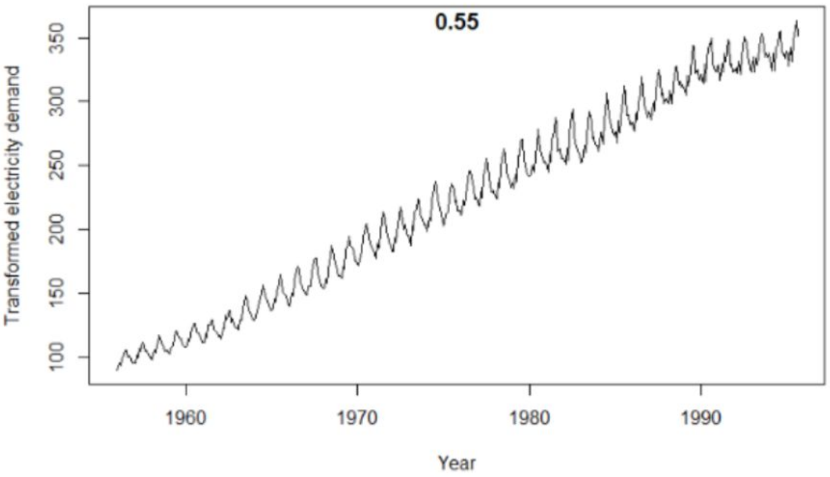
Ряд после логарифмирования и дифференцирования

Antidiabetic drug sales



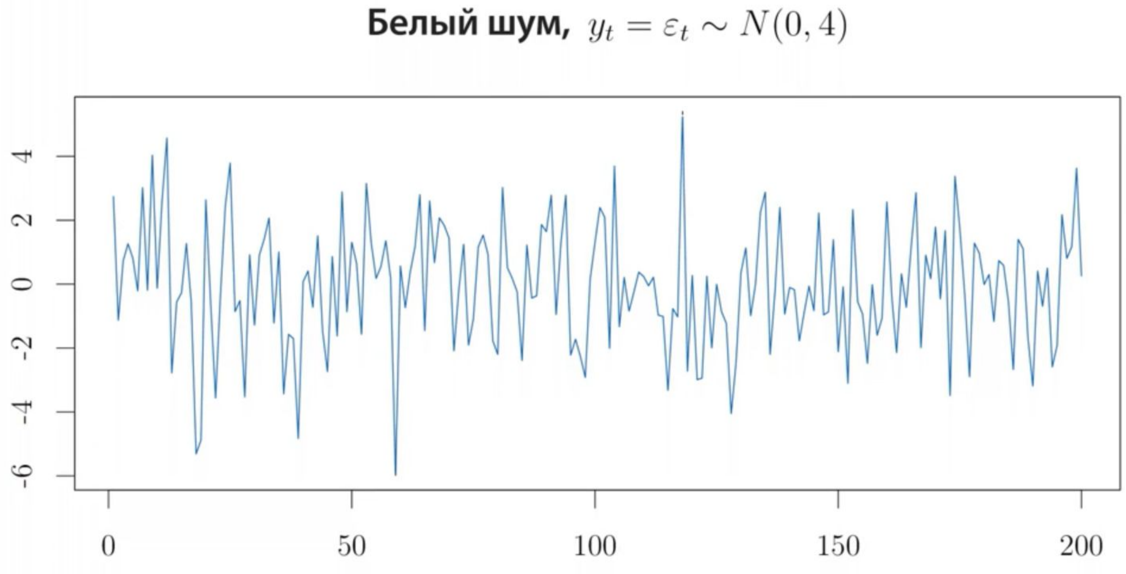
Работа преобразования Бокса-Кокса





Белый шум

Белый шум - процесс, имеющий постоянное математическое ожидание, постоянную дисперсию и нулевую автоковариационную функцию. Белый шум является одним из самых простых примеров стационарного ряда.



Модели прогнозирования временных рядов

Статистические

- Регрессия
- Авторегрессия
- Экспоненциальное сглаживание
- и др..

Структурные

- Нейронные сети
- Цепи Маркова
- Классификационные деревья
- и др.

AR
(Autoregression)

Авторегрессия - регрессия ряда на собственные значения в прошлом.

$$AR(p) : Y_t = \alpha + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \varepsilon_t$$

α - какая-то константа, ϕ_i- параметры модели, Y - стационарный ряд,
ε_t - гауссов белый шум с нулевым средним.

MA

(*Moving average*)

Скользящее среднее - авторегрессия, примененная к шуму.

$$MA(q): Y_t = \alpha + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + ... + \theta_q\varepsilon_{t-q}$$

α - какая-то константа, θ_i - параметры модели, Y - стационарный ряд,
ε_i - гауссов белый шум с нулевым средним.

ARMA
(Autoregressive moving average)

$$ARMA(p,q): Y_t = \alpha \ + \ \phi_1 Y_{t-1} \ + \ \phi_2 Y_{t-2} \ + \ ... \ + \ \phi_p Y_{t-p} \ +$$

$$+ \ \varepsilon_t \ + \ \theta_1 \varepsilon_{t-1} \ + \ \theta_2 \varepsilon_{t-2} \ + \ ... \ + \ \theta_q \varepsilon_{t-q}$$

α - какая-то константа, ϕ_i , θ_i - параметры модели, Y - стационарный ряд, ε_i - гауссов белый шум с нулевым средним. p - количество авторегрессионных компонент, а q - количество компонент скользящего среднего. p + q минимально возможна.

Теорема Вольда:

Любой стационарный ряд может быть описан моделью $ARMA(p,q)$ с любой наперёд заданной точностью.

ARIMA

(Autoregressive integrated moving average)

$$ARIMA(p,d,q) : \nabla^d Y_t = \alpha \; + \; \phi_1 Y_{t-1} \; + \; \phi_2 Y_{t-2} \; + \; ... \; + \; \phi_p Y_{t-p} \; +$$

$$+ \; \varepsilon_t \; + \; \theta_1 \varepsilon_{t-1} \; + \; \theta_2 \varepsilon_{t-2} \; + \; ... \; + \; \theta_q \varepsilon_{t-q}$$

α - какая-то константа, ϕ_i , θ_i - параметры модели, Y - стационарный ряд, ε_i - гауссов белый шум. p - количество авторегрессионных компонент, а q - количество компонент скользящего среднего.

Автокорреляция (ACF)

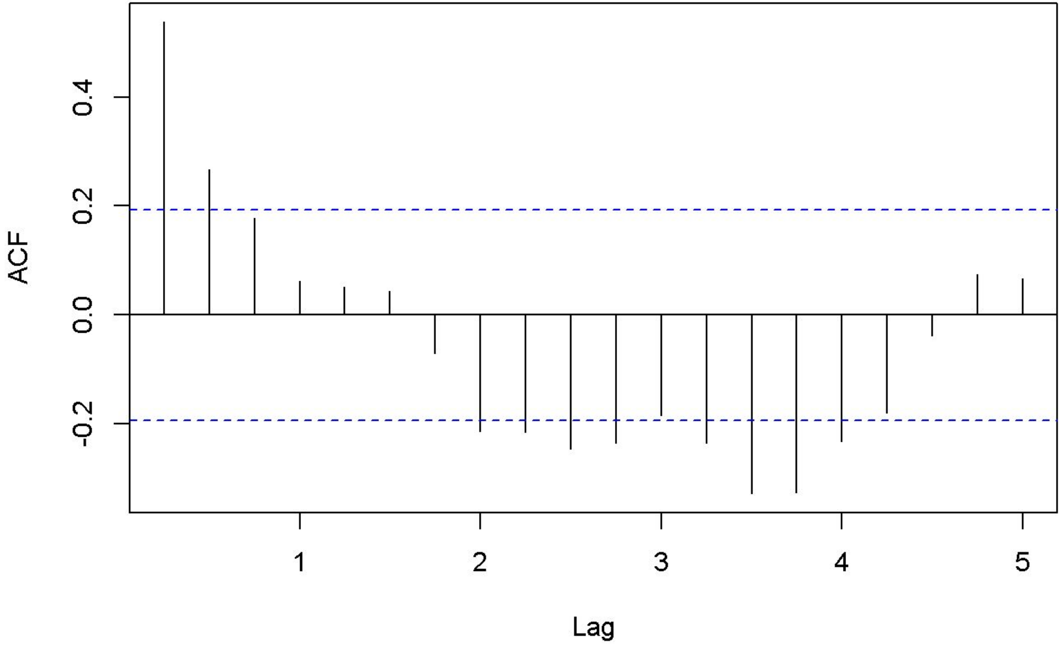
$$r_i = r_{Y_t, Y_{t-i}} = \frac{\sum\limits_{t=i+1}^{T} (Y_t - \overline{Y})(Y_{t-i} - \overline{Y})}{\sum\limits_{t=1}^{T} (Y_t - \overline{Y})^2}$$

Автокорреляция измеряет совокупный эффект воздействия.

i - лаг, сдвиг по времени, *T* - длина ряда

Коррелограмма

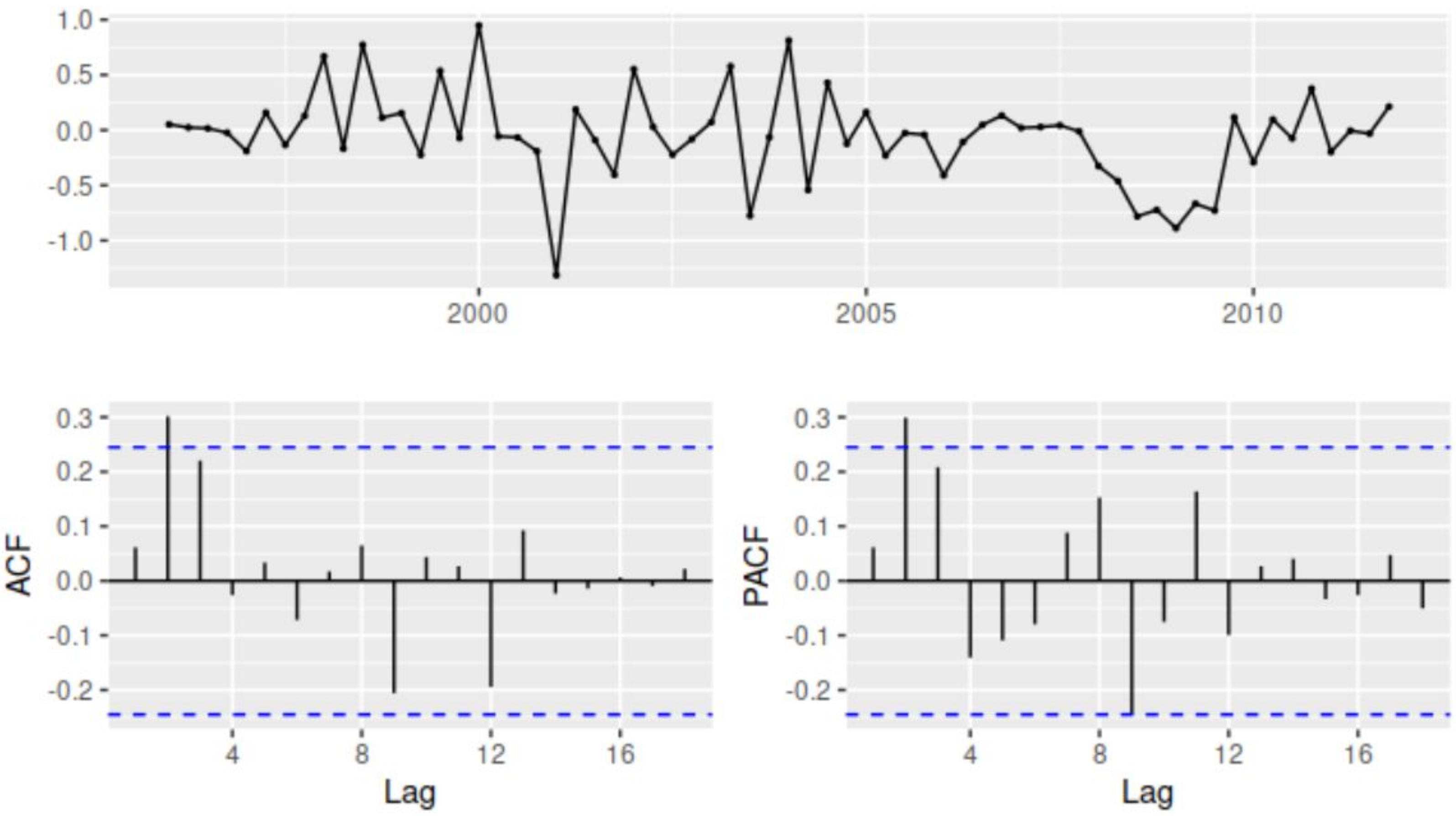
Коррелограмма — это график автокорреляций. Он помогает понять как значения ряда связаны со своими же значениями в прошлом. Лаг отражает степень временной задержки.



Частная автокорреляция (PACF)

$$\phi_i = \begin{cases} r_{Y_{t-1}, Y_t} & i = 1 \\ r_{\varepsilon_{t-i}, \varepsilon_t} & i > 1 \end{cases}$$

Частная автокорреляция измеряет прямой эффект воздействия



1.0
0.5
0.0
-0.5
-1.0
2000
2005
2010
ACF
0.3
0.2
0.1
0.0
-0.1
-0.2
4
8
12
16
Lag
PACF
0.3
0.2
0.1
0.0
-0.1
-0.2
4
8
12
16
Lag

Подбор параметров

- Если p , d , q фиксированы, то α ,ϕ_i, θ_i подбираются методом наименьших квадратов.
- Чтобы подобрать ε_i делают авторегрессию на ряд и считают остатки, затем подставляют вместо шума и применяют МНК.
- d подбирается так, чтобы ряд стал стационарным.
- p, q не можем выбрать по ММП, так чем больше p и q , тем больше параметров и тем больше ОМП - тем лучше модель обучается.
- p, q подбираем по кореллограмме. p - номер последнего лага при котором PACF значима, q - номер последнего лага при котором ACF значима.

Аддитивная модель временного ряда

$$Y = T + S + E$$

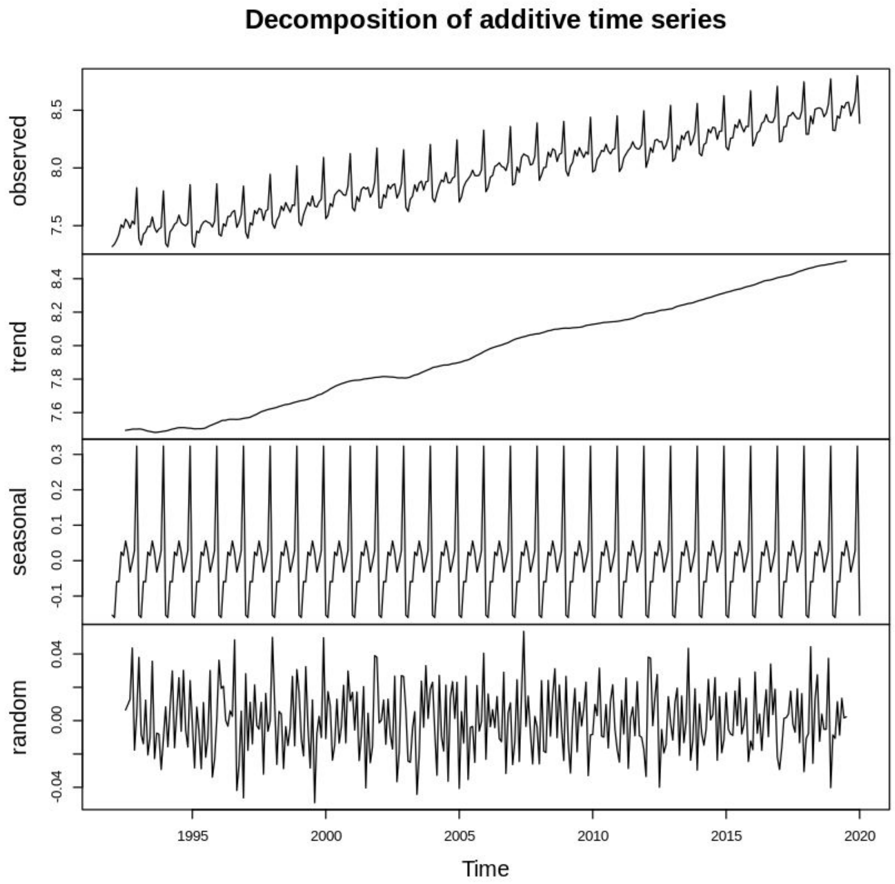
- T - трендовая составляющая
- S - сезонная составляющая
- E - случайная составляющая

Аддитивную модель строят если амплитуда сезонных колебаний относительно трендовой компоненты приблизительно постоянна.

Как работает?

1. Выравнивание исходного ряда скользящей средней
2. Оценка сезонной компоненты с учетом того, что для аддитивной модели сумма сезонных компонент за весь период равна нулю (Y - центрированная скользящая средняя)
3. Удаление сезонных компонент из исходных уровней ряда Y – S и получение T + E
4. Оценка параметров тренда по полученных по модели значений T + E (прогнозируем любой моделью)
5. Добавление к прогнозам сезонность последнего периода времени
6. Оценка качества полученной модели

Разложение ряда на компоненты



Fbprophet

(Facebook Prophet)

$$Y(t) = g(t) + s(t) + h(t) + \varepsilon_t$$

- $s(t)$ - сезонная составляющая
- $h(t)$ - тренд
- $g(t)$ - аномальные дни
- ε_t - случайная составляющая

Метрики качества прогнозирования

Имеем:

Y_t — фактическое значение временного ряда в момент времени t

Y_t' — прогнозируемое значение в момент времени t

n — *количество анализируемых значений*

Хотим:

Оценить качество прогнозирования, то есть понять, какая именно модель прогнозирования наиболее подходит к анализируемому ряду и дает значение, максимально близкое к реальному результату.

Основные метрики

- ***Средняя абсолютная ошибка*** *(Mean Absolute Error, MAE)*

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|Y_t - Y'_t|$$

- ***Средняя абсолютная процентная ошибка*** *(Mean Absolute Percentage Error, MAPE)*

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|Y_t - Y'_t|}{|Y_t|} \cdot 100\%$$

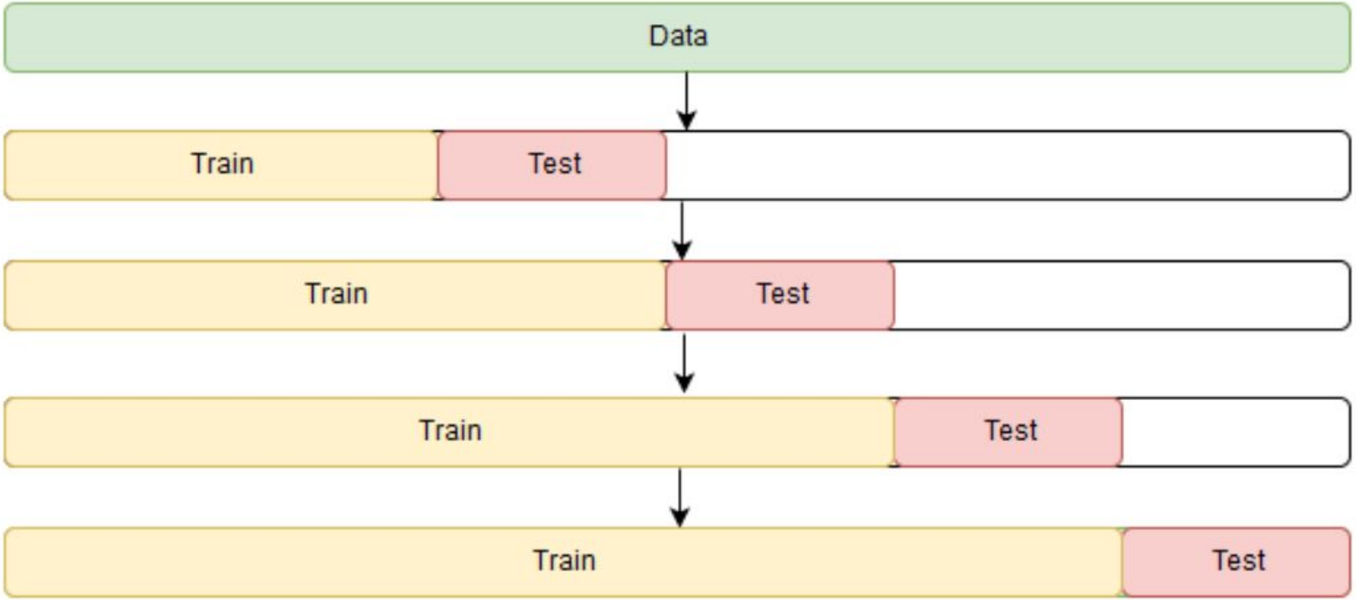
- ***Средний квадрат ошибок*** *(Mean Squared Error, MSE)*

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(Y_t - Y_t')^2$$

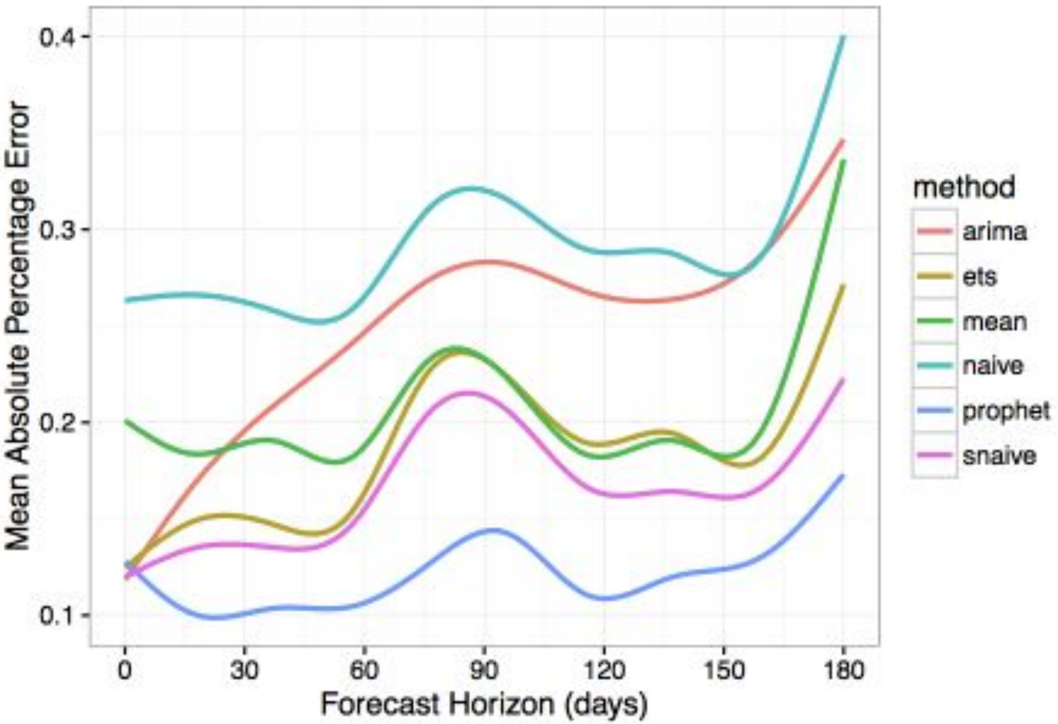
- ***Среднеквадратичная ошибка*** *(Root Mean Square Error, RMSE)*

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Y_t - Y_t')^2}$$

Кросс-валидация
(Cross-validation on a rolling basis)



Сравнение моделей ARIMA и fbprophet (метрика MAPE)



Сравнение моделей ARIMA и fbprophet

ARIMA(3, 1, 4): MAPE = 16.54%, MAE = 7.28
Fbprophet: MAPE = 26.79%, MAE = 8.49

Что нового мы узнали?

- Что такое временные ряды?
- Главная задача анализа временных рядов
- Применение
- Основные свойства
- Методы сведения нестационарного ряда к стационарному
- Аддитивные регрессионные модели
- Простейшие авторегрессионные модели
- Подбор параметров моделей
- Метрики качества прогнозирования
- Кросс-валидация по ряду

Список источников

- «Анализ временных рядов и прогнозирование», Сажин Ю.В., Катынь А.В., Сарайкин Ю.В, 2013
- «Прогнозирование и временные ряды», Кизбикенов К.О., Барнаул, ФГБОУ ВО, «АлтГПУ» 2017
- https://habr.com/ru/company/ods/blog/323730/
- Introduction to Time Series Analysis https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4.htm
- https://ranalytics.github.io/tsa-with-r/ch-intro-to-prophet.html
- https://www.youtube.com/watch?v=u433nrxdf5k
- https://habr.com/ru/company/ods/blog/327242/
- http://ainsnt.ru/file/out/863967
- https://facebook.github.io/prophet/static/prophet_paper_20170113.pdf
- http://www.agpu.net/fakult/ipimif/metodmater/ddv007_additivmodelvr3.pdf
- https://otexts.com/fpp2/
- https://www.youtube.com/playlist?list=PLu5flfwrnSD6wzkzgs4TocGL5GOXmEjZE

Вопросы

1. Какое преобразование стоит применить при сведении нестационарного ряда с не постоянной дисперсией к стационарному?
2. Запишите формулы моделей AR(p), MA(q). Как подбирать параметры данных моделей?
3. В чем преимущество модели ARIMA по сравнению с ARMA?
4. Из каких компонент состоит аддитивная модель fbprophet?