Временные ряды - 1

Елена Кантонистова

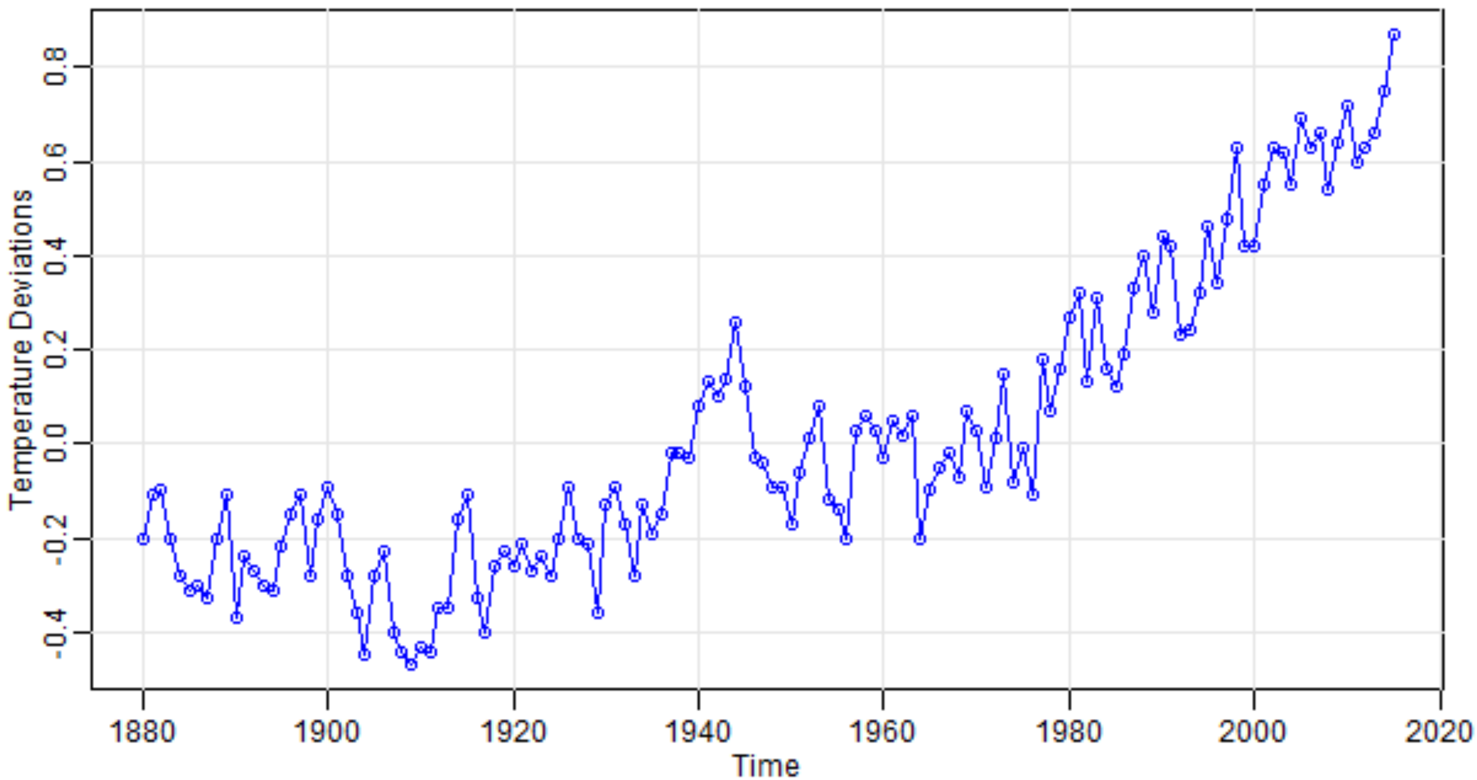
ВШЭ

Занятие 1

Статистические модели прогнозирования

ВРЕМЕННОЙ РЯД

Временной ряд **–** это последовательность значений, описывающих протекающий во времени процесс, измеренных в последовательные моменты времени, обычно через равные промежутки.



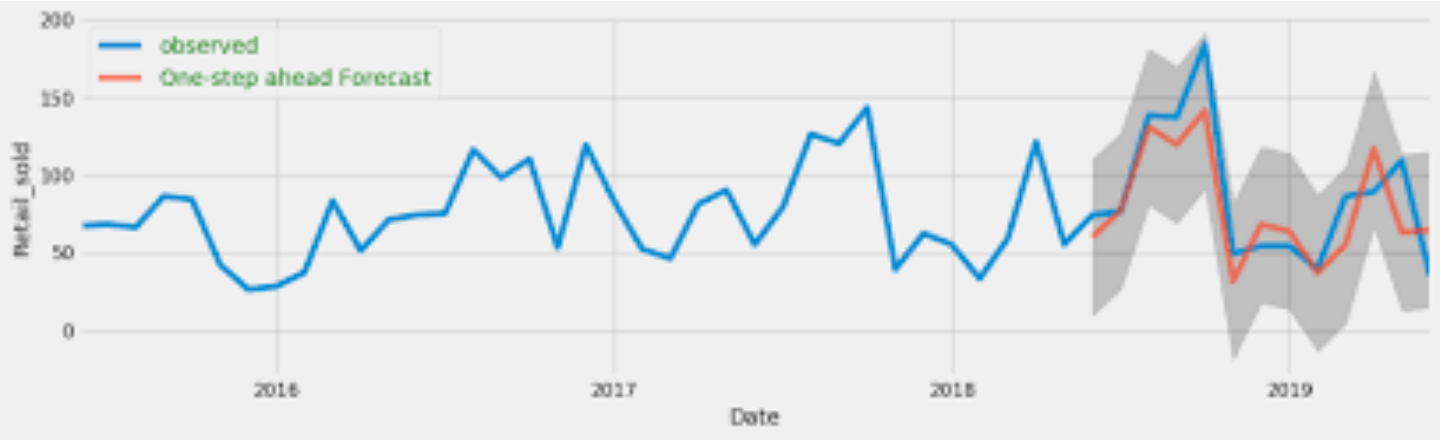
ЗАДАЧА ПРОГНОЗИРОВАНИЯ

$y_0, y_1, \dots, y_t, \dots$ - временной ряд, $y_i \in \mathbb{R}$.

<u>Задача:</u> построить функцию

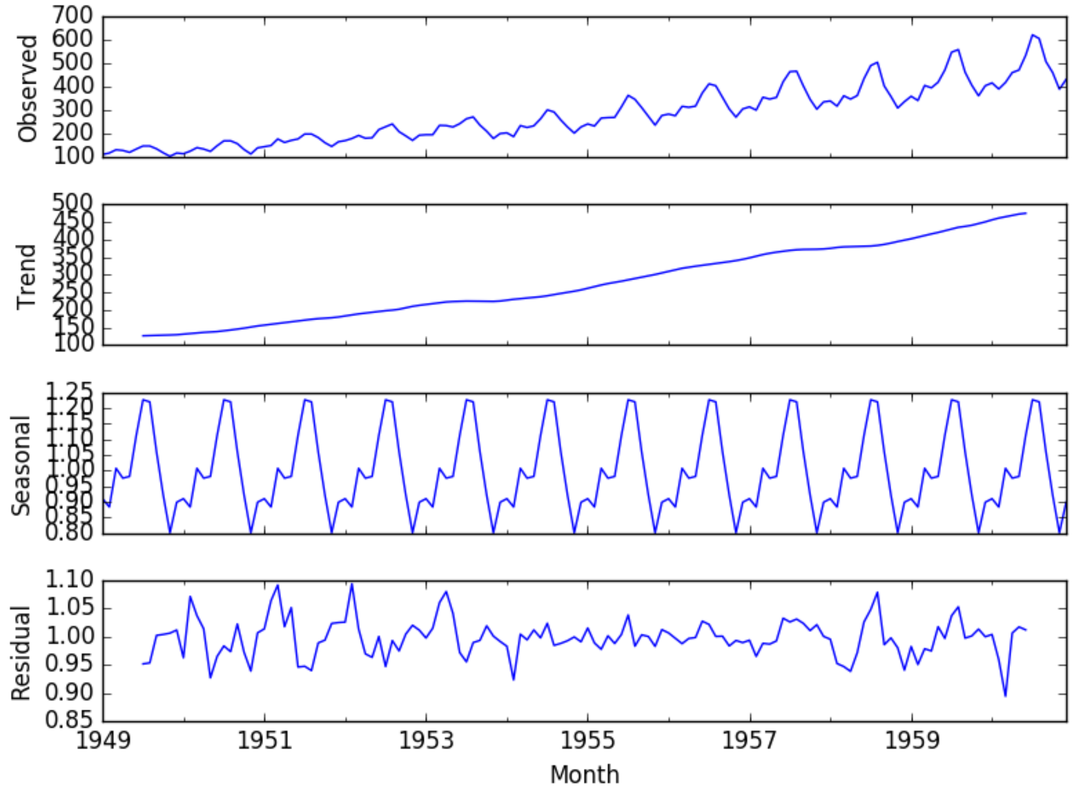
$$\hat{y}_{t+d}(w) = a_{t,d}(y_1, \dots, y_t; w)$$

- $d = 1, \dots, D$, где D – горизонт прогнозирования
- w – вектор параметров модели



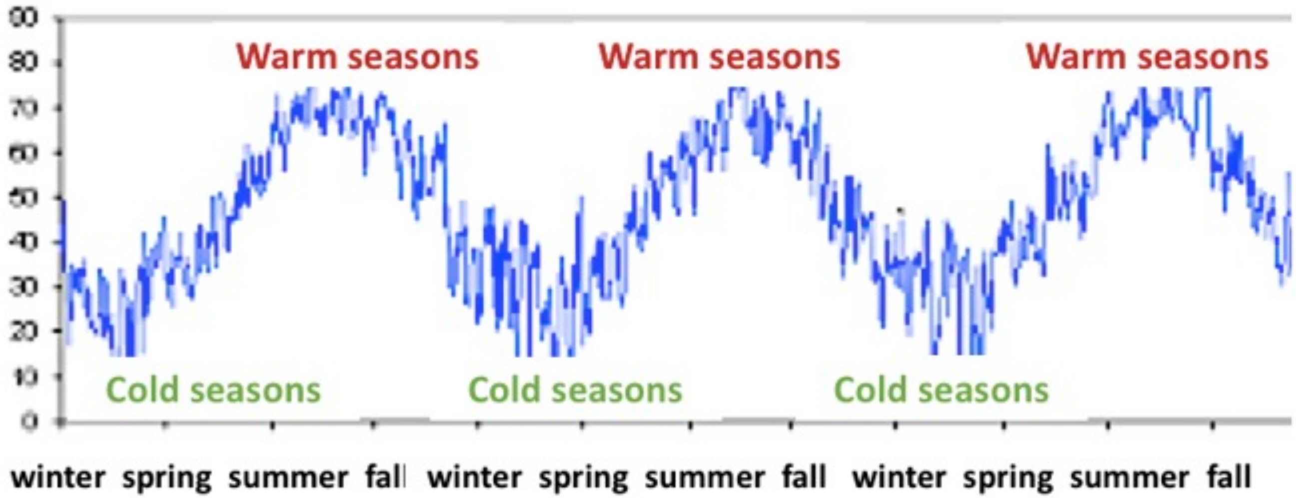
КОМПОНЕНТЫ ВРЕМЕННОГО РЯДА

- *Тренд* – плавное долгосрочное изменение уровня ряда
- *Сезонность* – циклические изменения уровня ряда с постоянным периодом
- *Циклы* – изменения уровня ряда с переменным периодом (цикл жизни товара, экономические волны, периоды солнечной активности)
- *Ошибка (шум)* – непрогнозируемая случайная компонента ряда

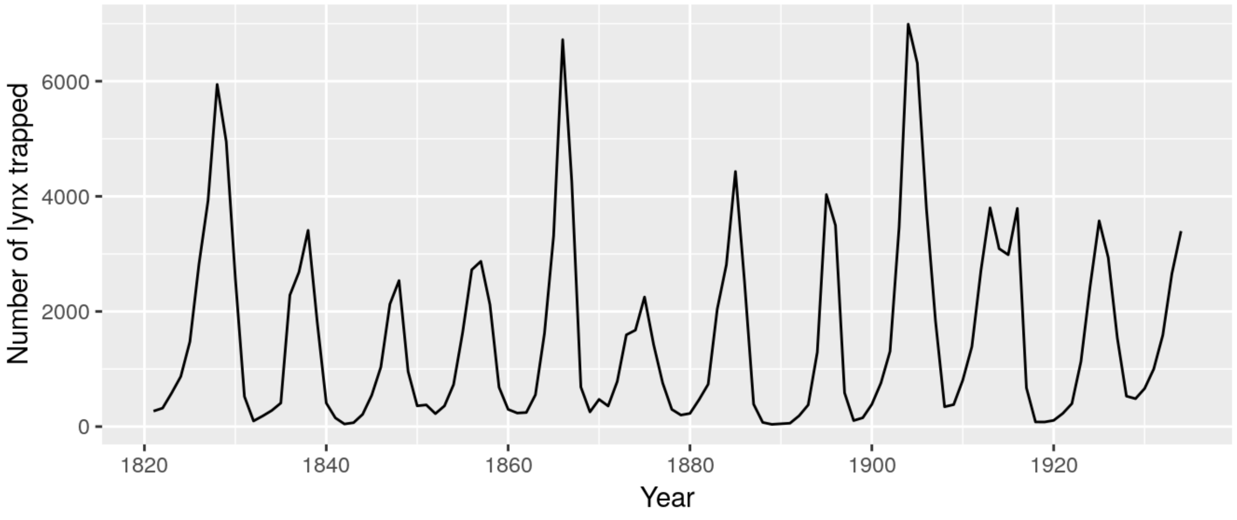


ЦИКЛЫ И СЕЗОННОСТЬ

Сезонность:



Цикл:



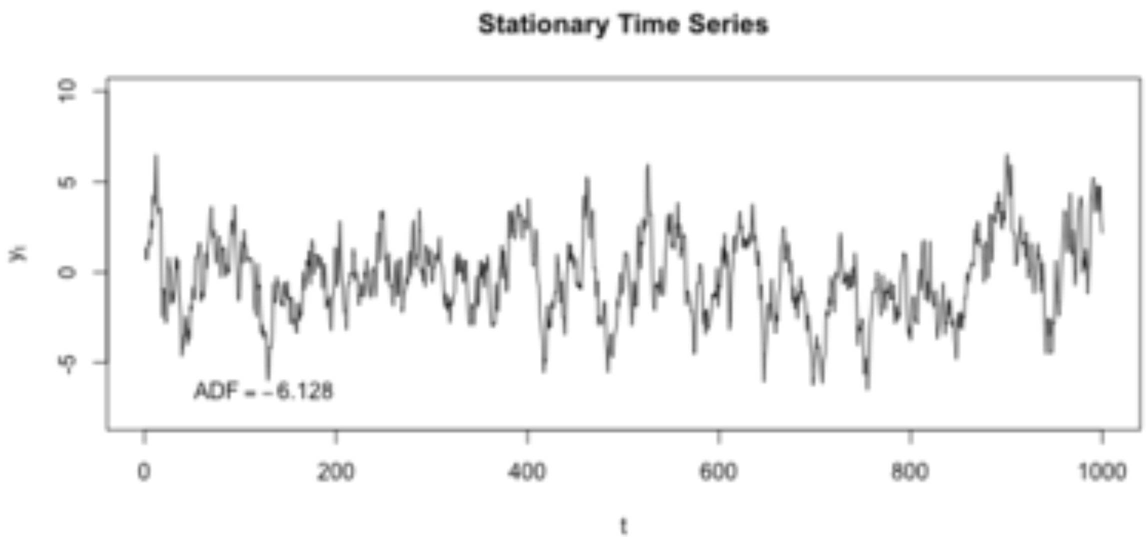
СТАЦИОНАРНОСТЬ

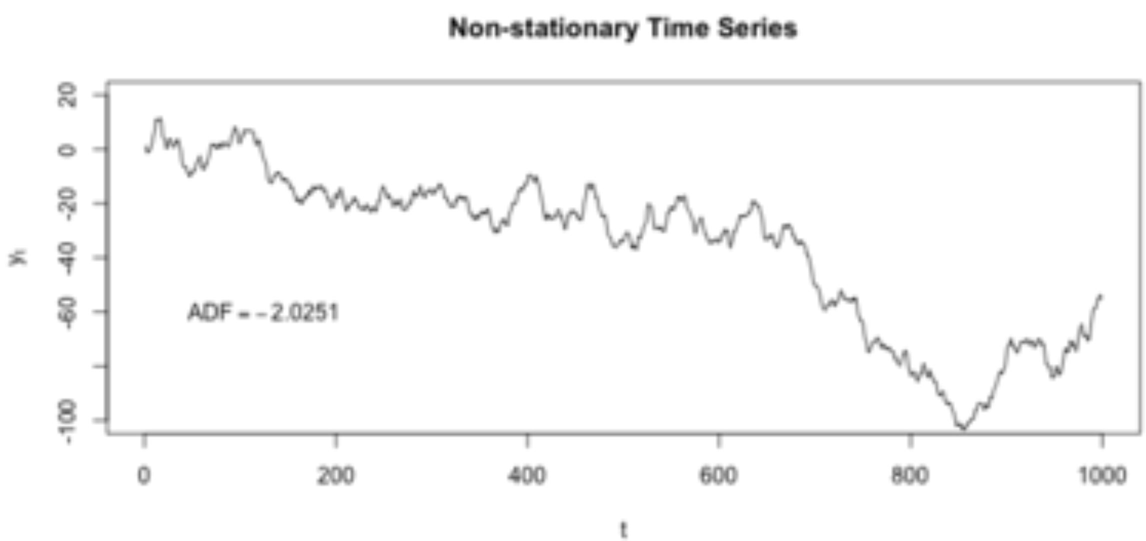
Стационарный временной ряд **-** это временной ряд, у которого статистические свойства не меняются со временем.

Формально, стационарный временной ряд должен удовлетворять следующим условиям:

- *Постоянное среднее значение*: среднее значение ряда должно быть постоянным и не зависеть от времени. Это означает, что в разные моменты времени ряд не должен демонстрировать систематический тренд вверх или вниз.
- *Постоянная вариация*: дисперсия ряда должна быть постоянной и не зависеть от времени. Это означает, что в разные моменты времени амплитуда колебаний ряда не должна изменяться.
- *Некоррелированность*: корреляция между значениями ряда на разных временных отрезках должна быть незначительной или отсутствовать вовсе. Это означает, что отсутствуют систематические зависимости или паттерны, которые повторяются во времени.

СТАЦИОНАРНОСТЬ





СТАЦИОНАРНОСТЬ

Ряд $y_1, \ldots, y_T$ ***стационарен***, если для любого s распределение $y_t, \ldots, y_{t+s}$ не зависит от t, то есть его свойства не зависят от времени.

- тренд ⇒ нестационарность
- сезонность ⇒ нестационарность
- цикл – заранее неизвестно

По стационарному ряду просто построить прогноз, так как мы полагаем, что его будущие статистические характеристики не будут отличаться от наблюдаемых текущих.

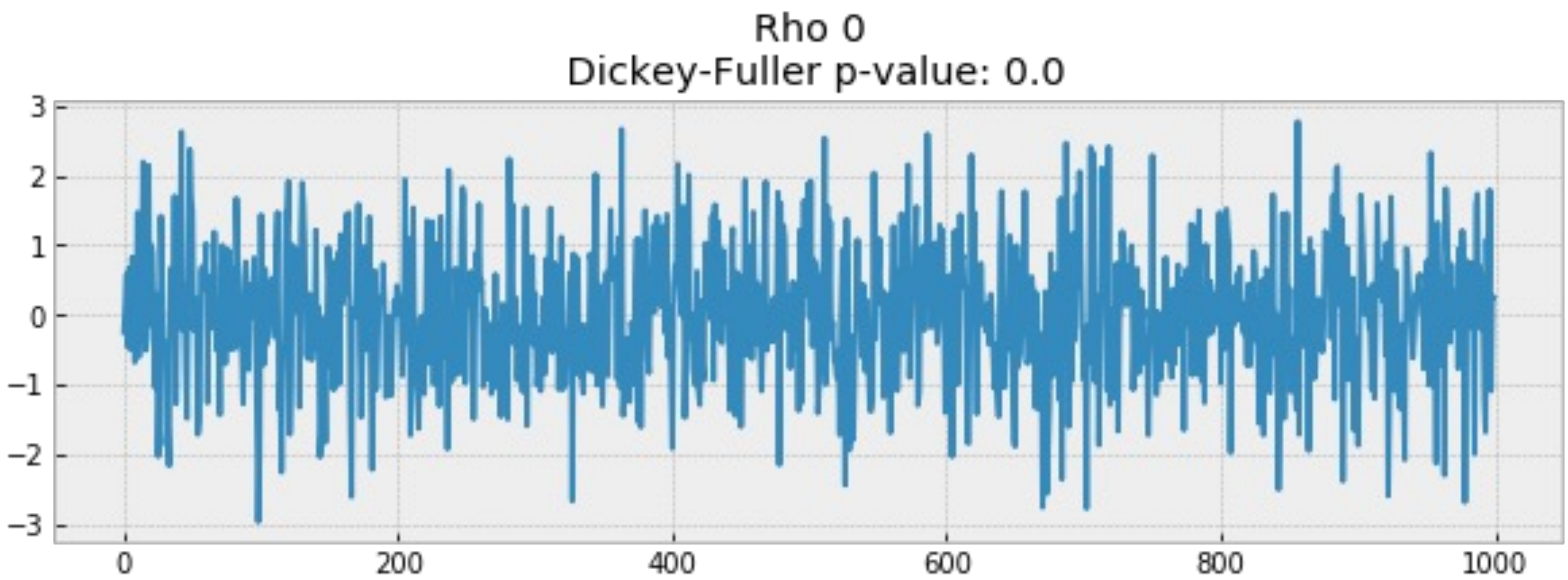
Для проверки стационарности ряда можно использовать критерий Дики-Фуллера.

ЕДИНИЧНЫЙ КОРЕНЬ

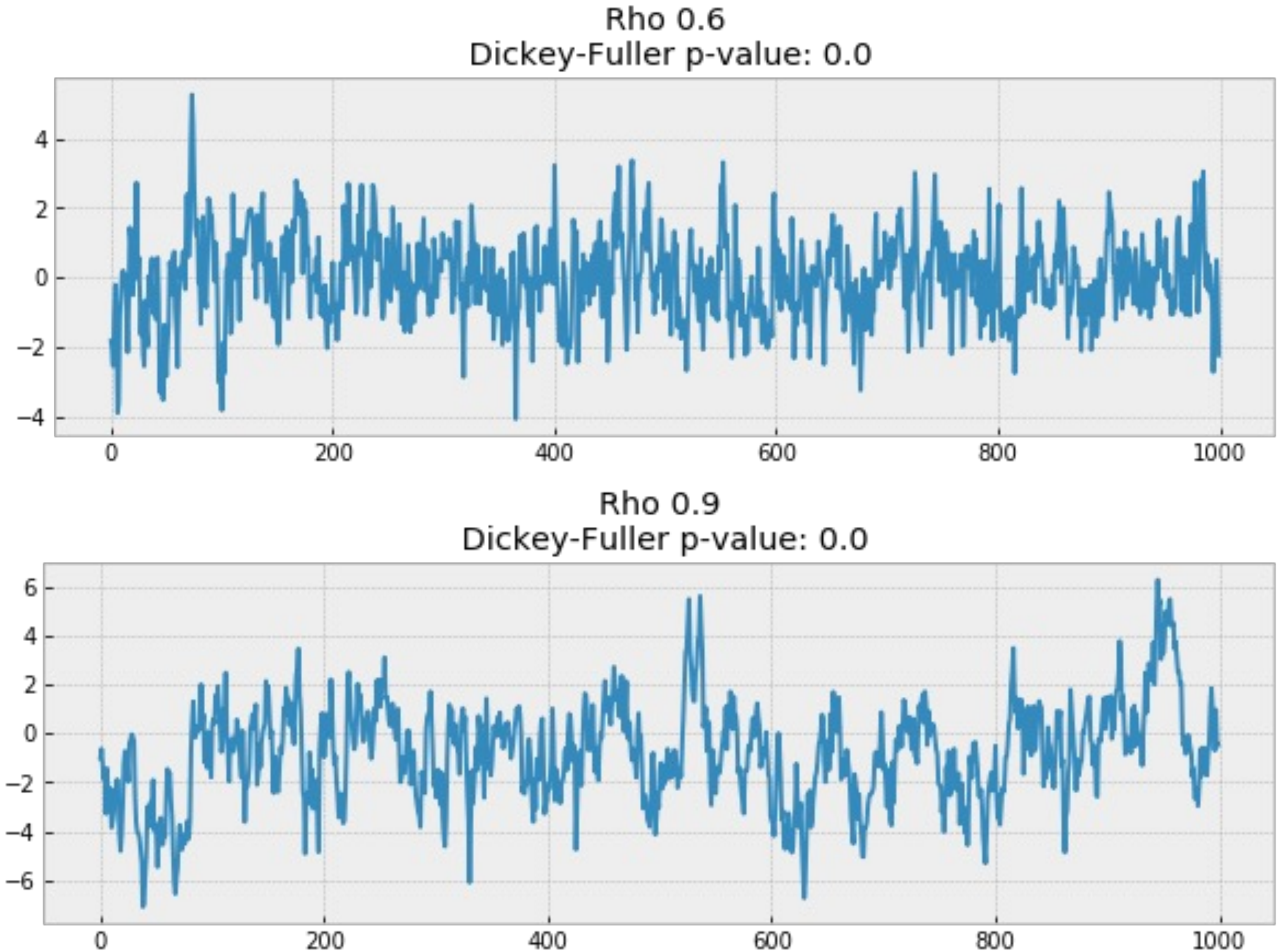
Рассмотрим модель временного ряда $\boldsymbol{X_t = \rho \cdot X_{t-1} + \varepsilon_t}$,

где ε_t - ошибка, не зависящая от значений временного ряда.

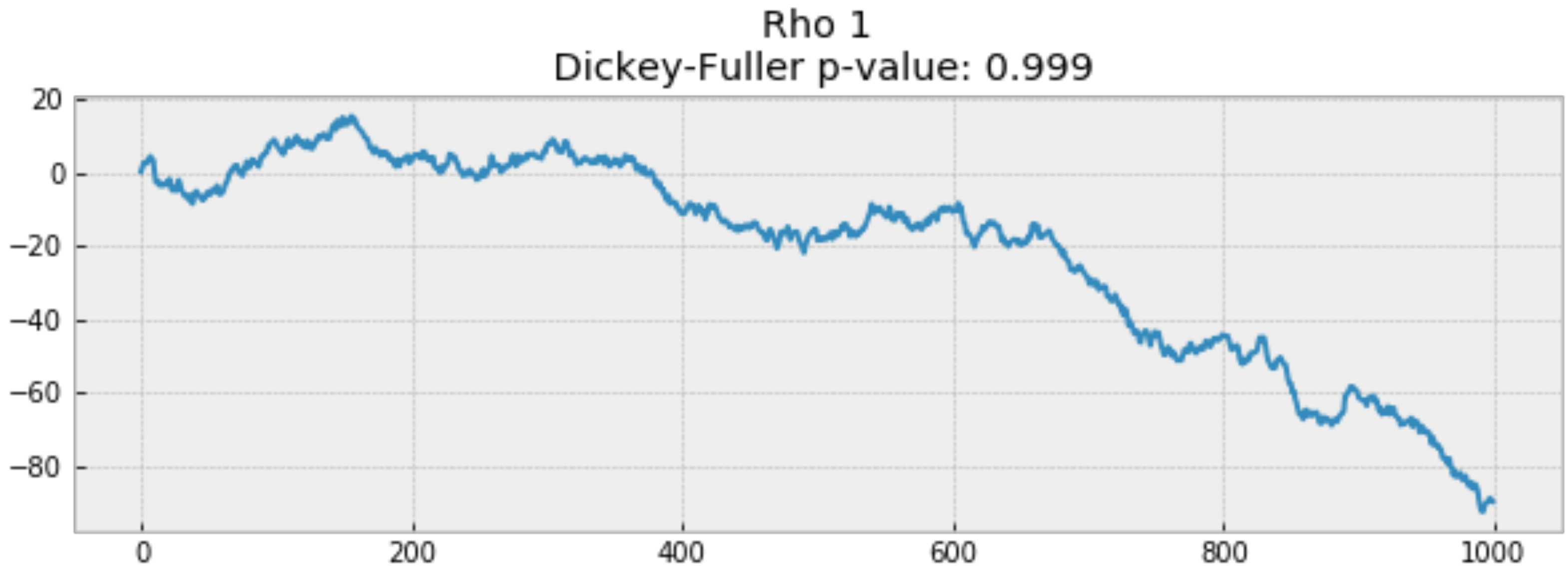
Определение. Если $\rho = 1$, то говорят, что ряд имеет единичный корень.



ЕДИНИЧНЫЙ КОРЕНЬ



ЕДИНИЧНЫЙ КОРЕНЬ



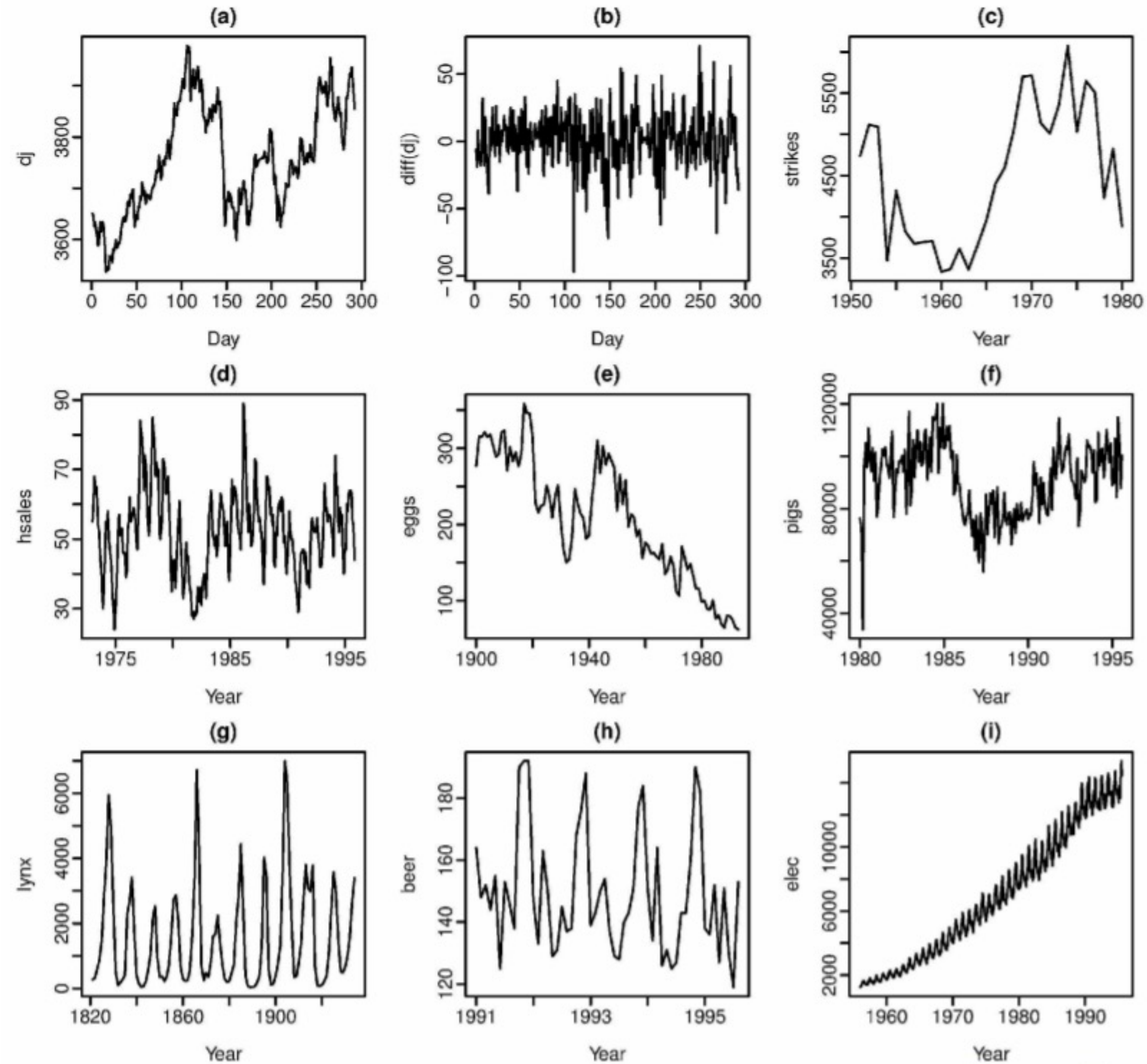
Видно, что при $\rho = 1$ процесс не возвращается к своему среднему, а значит, не является стационарным.

ПРОВЕРКА СТАЦИОНАРНОСТИ РЯДА

Проверку стационарности ряда можно осуществлять с помощью критерия Дики-Фуллера.

- Критерий Дики-Фуллера проверяет гипотезу $\rho = 1$.

ПРИМЕРЫ ВРЕМЕННЫХ РЯДОВ

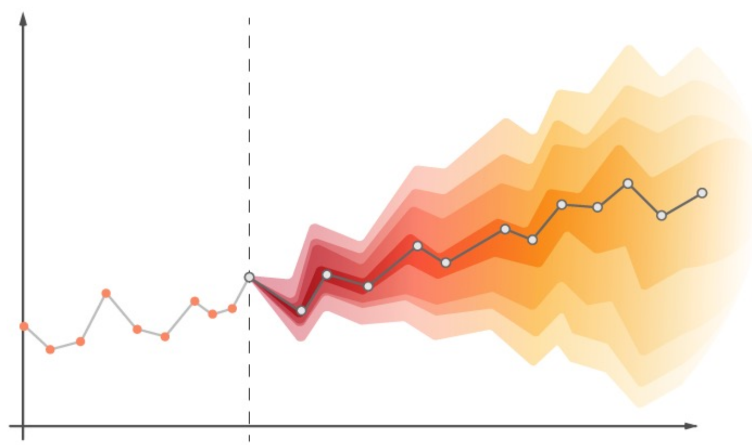


ПРАКТИКА-1

- Time series intro
- Time series intro (bonus)

СТАТИСТИЧЕСКИЕ МЕТОДЫ ПРОГНОЗА ВРЕМЕННЫХ РЯДОВ

- Этот подход основан на том, что стационарный временной ряд прогнозировать несложно, поэтому общая идея такая:
- Приводим ряд к стационарному
- С помощью линейной регрессии прогнозируем стационарный временной ряд
- Применяем обратные преобразования к прогнозу



МЕТОДЫ ИЗБАВЛЕНИЯ ОТ НЕСТАЦИОНАРНОСТИ

1. Стабилизация дисперсии

- для рядов с монотонно меняющейся дисперсией можно использовать стабилизирующее ***преобразование Бокса-Кокса*** (λ – параметр метода):

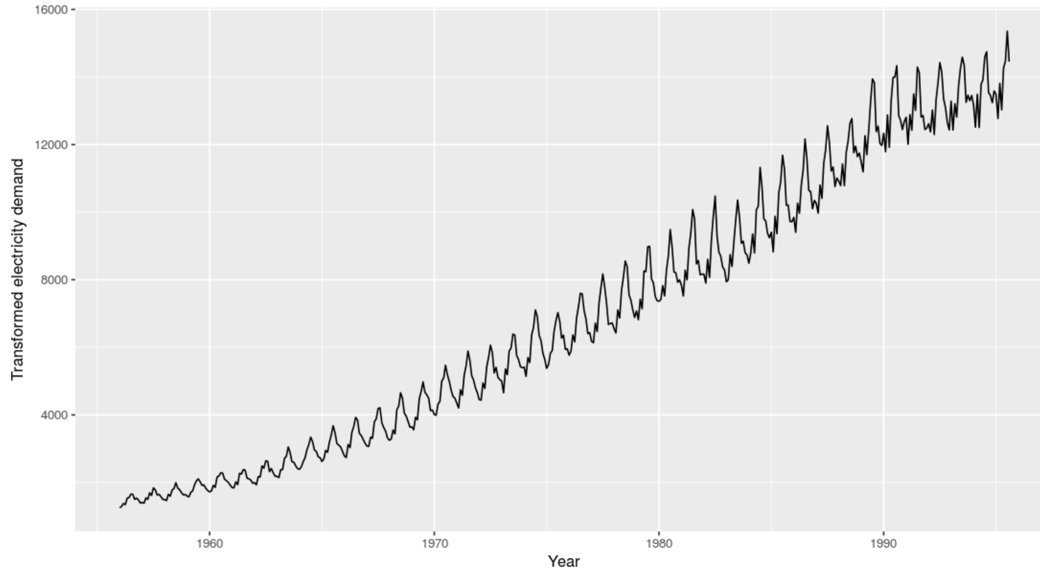
$$y_t' = \begin{cases} \ln y_t, \lambda = 0 \\ \dfrac{y_t^{\lambda} - 1}{\lambda}, \lambda \neq 0 \end{cases}$$

- логарифмирование – частный случай

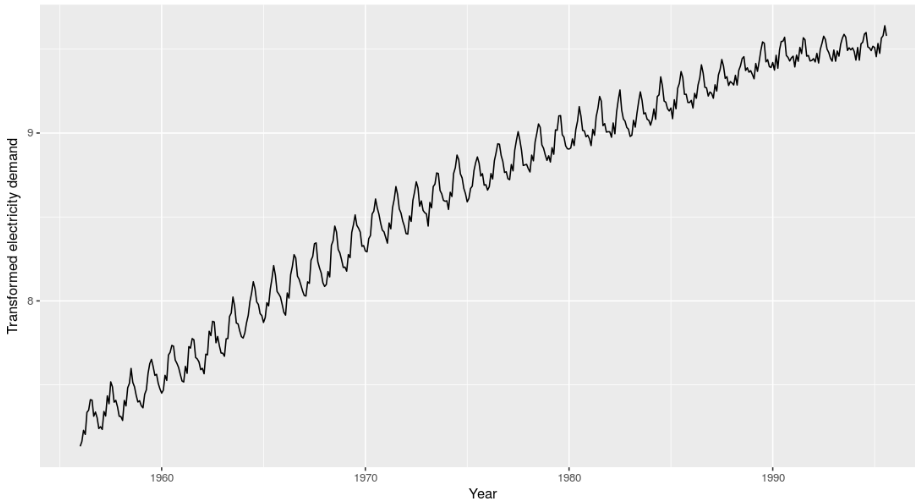
Параметр λ подбирается так, чтобы сделать дисперсию как можно более однородной.

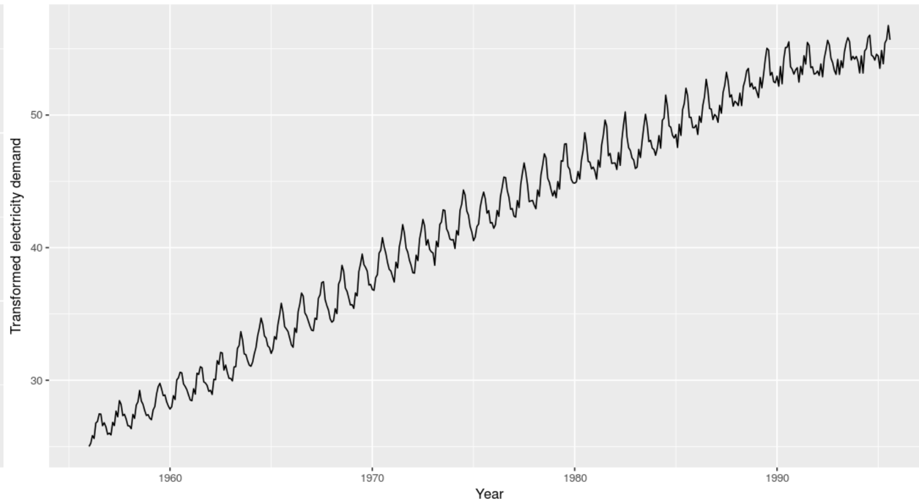
СТАБИЛИЗАЦИЯ ДИСПЕРСИИ (ПРИМЕР)

Исходный ряд:



Преобразование Бокса-Кокса с $\lambda = 0$ (слева) и $\lambda = 0.3$ (справа):





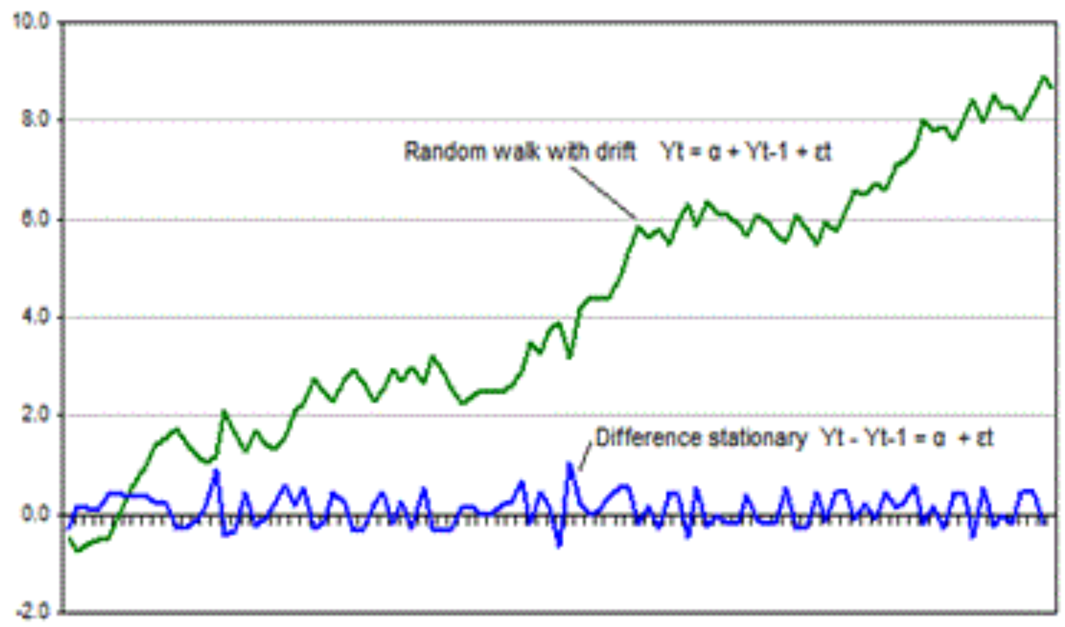
МЕТОДЫ ИЗБАВЛЕНИЯ ОТ НЕСТАЦИОНАРНОСТИ

2. Дифференцирование – переход к попарным разностям для соседних значений ряда

$$y_t' = y_t - y_{t-1}$$

- стабилизирует среднее значение ряда, позволяет избавиться от тренда
- можно применять

неоднократно

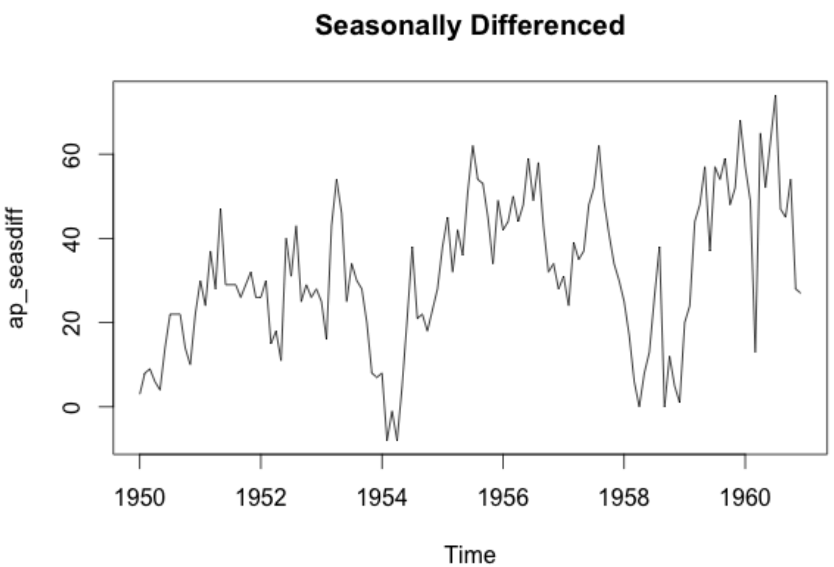


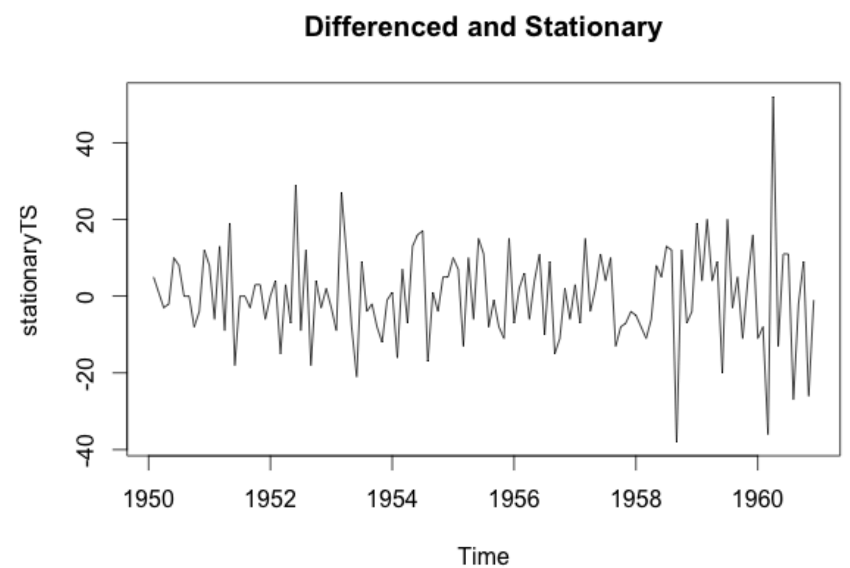
МЕТОДЫ ИЗБАВЛЕНИЯ ОТ НЕСТАЦИОНАРНОСТИ

3. <u>Сезонное дифференцирование</u> **–** переход к попарным разностям значений в соседних сезонах

$$y'_t = y_t - y_{t-s}$$

- убирает сезонность





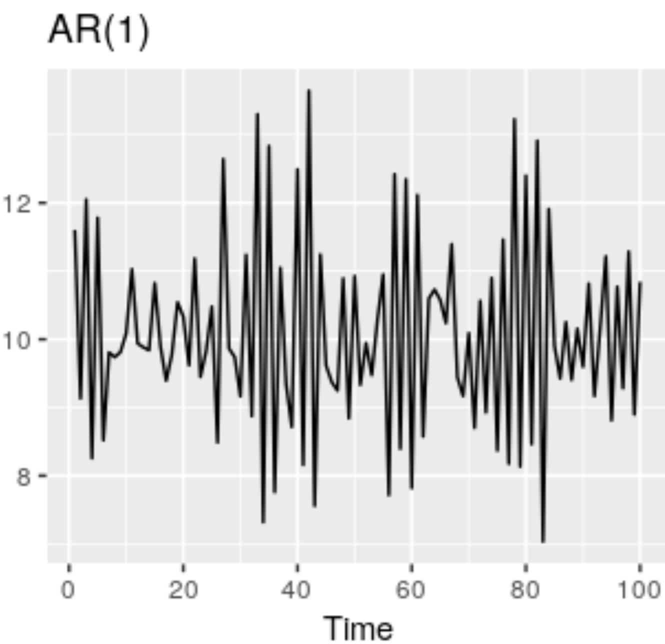
Сезонное дифференцирование лучше применять в начале – возможно, после него ряд уже станет стационарным.

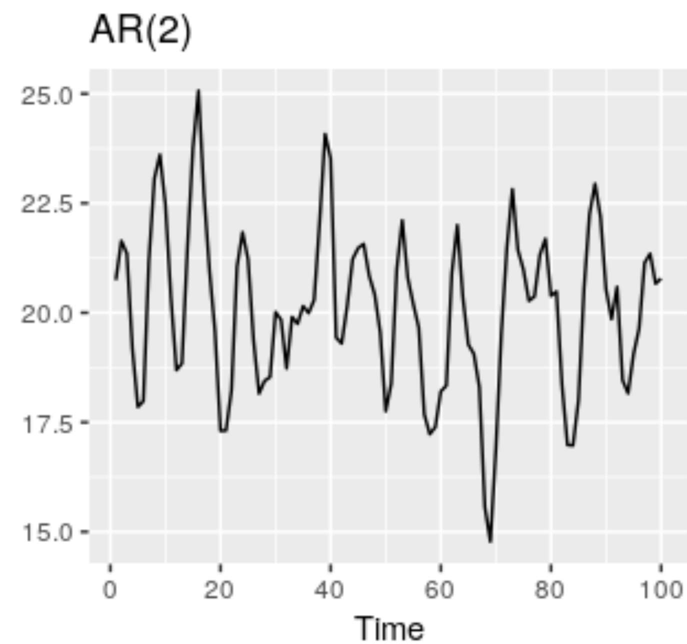
МОДЕЛИ ВРЕМЕННЫХ РЯДОВ – ЭКОНОМЕТРИЧЕСКИЙ ПОДХОД

- Модель авторегрессии $AR(p)$:

$$\boldsymbol{y_t = a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_p y_{t-p} + \varepsilon_t},$$

$a_p \neq 0, \varepsilon_t$ - процесс белого шума, $E\varepsilon_t = 0, D\varepsilon_t = \sigma_\varepsilon^2, cov(\varepsilon_t, \varepsilon_s) = 0$, некоррелируемый с y_t.



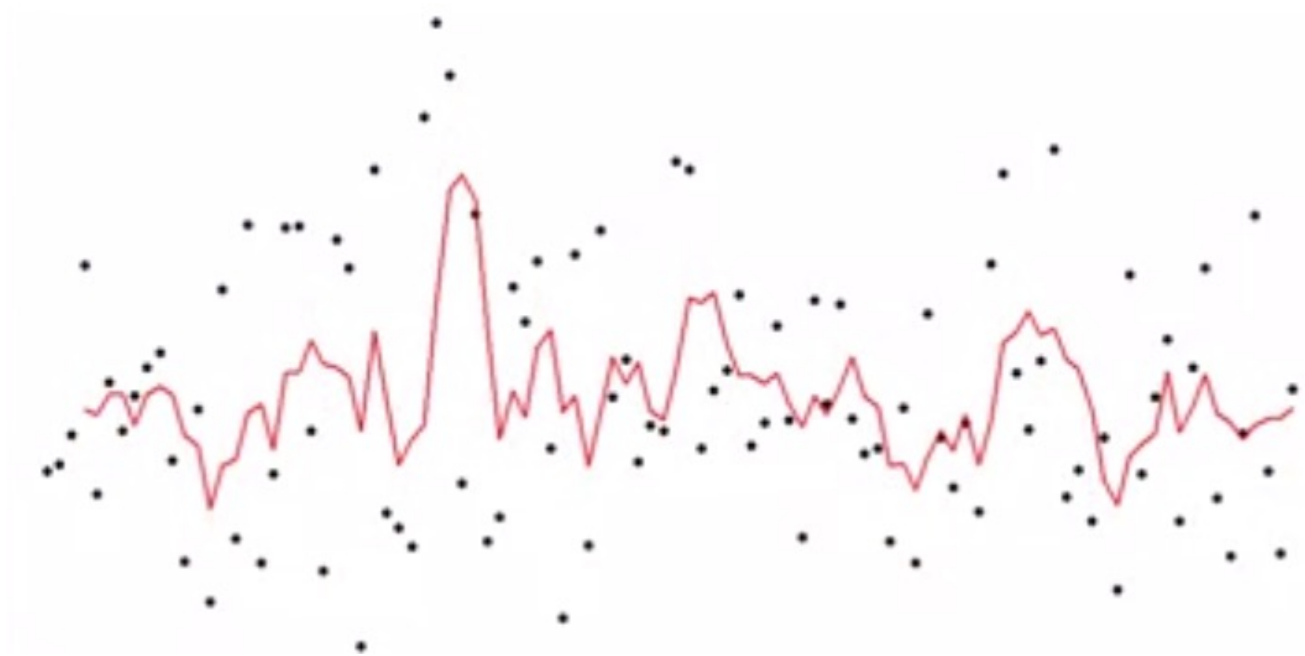


МОДЕЛИ ВРЕМЕННЫХ РЯДОВ – ЭКОНОМЕТРИЧЕСКИЙ ПОДХОД

- <u>Процесс скользящего среднего порядка q $\left(MA(q)\right)$:</u>

$$\boldsymbol{y_t = \alpha + \varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + b_q\varepsilon_{t-q}},$$

$b_q \neq 0, \varepsilon_t$ - процесс белого шума, $E\varepsilon_t = 0, D\varepsilon_t = \sigma_\varepsilon^2, cov(\varepsilon_t, \varepsilon_s) = 0$, некоррелируемый с y_t. Такой процесс всегда стационарен.



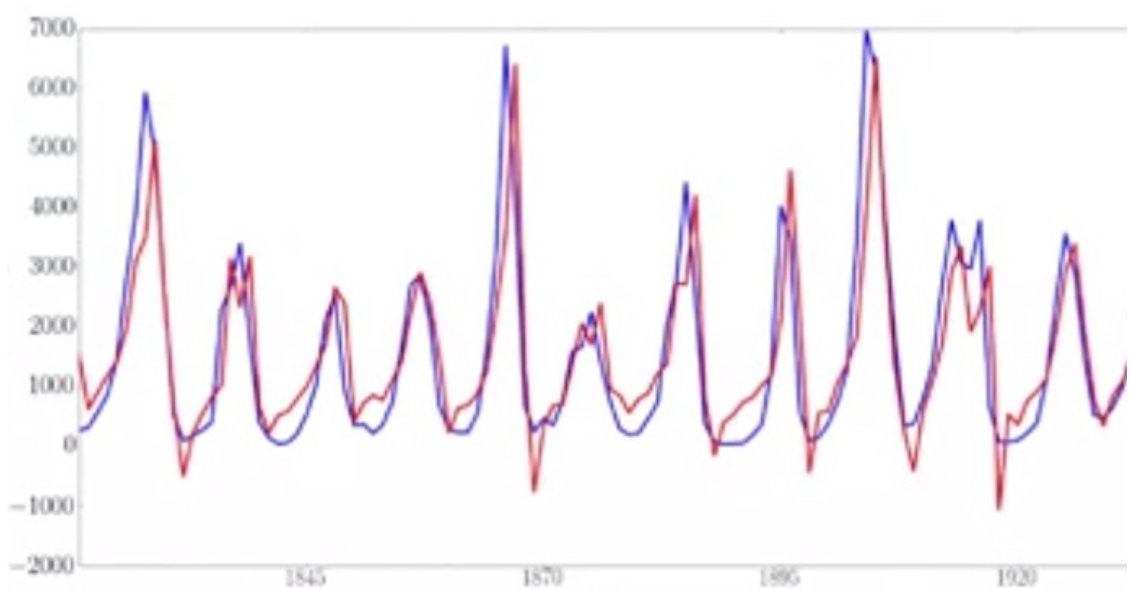
МОДЕЛИ ВРЕМЕННЫХ РЯДОВ – ЭКОНОМЕТРИЧЕСКИЙ ПОДХОД

- Смешанный процесс авторегрессии $ARMA(p, q)$:

$$\boldsymbol{y_t = \alpha + a_1 y_{t-1} + \cdots + a_p y_{t-p} + \varepsilon_t + b_1 \varepsilon_{t-1} + \cdots + b_q \varepsilon_{t-q}},$$

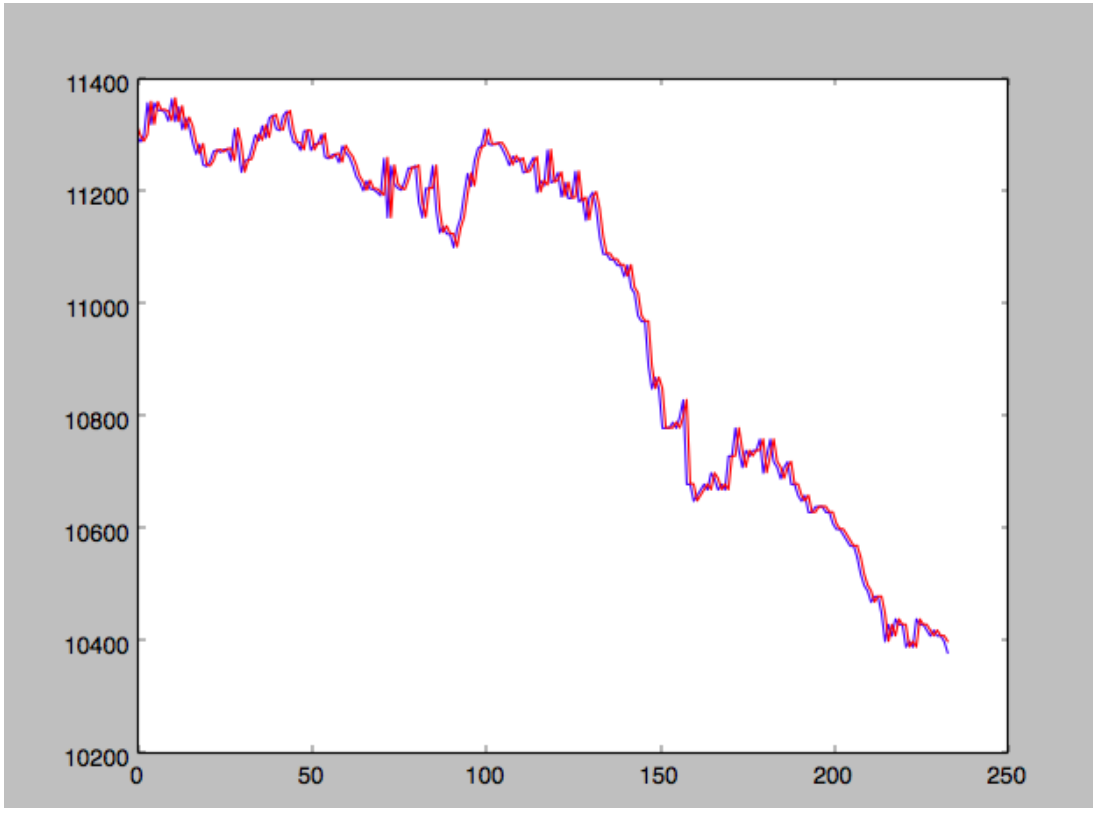
$a_p, b_q \neq 0.$

Теорема Вольда. Любой стационарный ряд можно приблизить моделью $ARMA(p, q)$ сколь угодно точно.



МОДЕЛИ ВРЕМЕННЫХ РЯДОВ – ЭКОНОМЕТРИЧЕСКИЙ ПОДХОД

- Модель $ARIMA(p,d,q)$ - модель $ARMA(p,q)$ для d раз продифференцированного ряда.



- *Модель SARIMA* – модель ARMA с учетом наличия тренда и сезонности

АВТОКОРЕЛЛЯЦИЯ

Для анализа временного ряда очень полезно изучить автокорреляционную функцию.

- Автокорреляция - это просто корреляция временного ряда с собой же, но сдвинутым на несколько моментов времени.
- То есть, например, мы можем посчитать корреляцию между рядом

$$y_t, y_{t-1}, y_{t-2}, \dots$$

и им же, но, например, два момента времени назад:

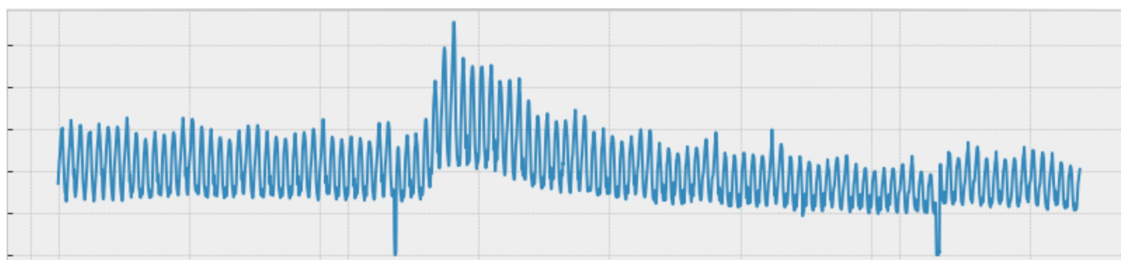
$$y_{t-2}, y_{t-3}, y_{t-4}, \dots$$

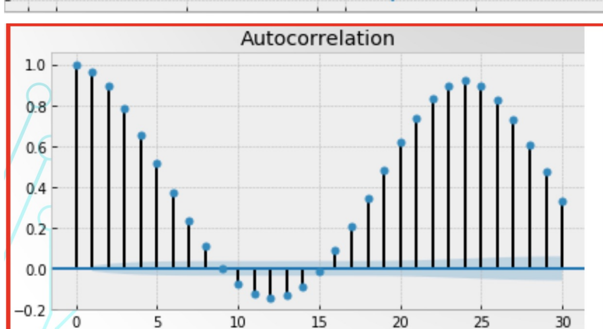
Большое значение коэффициента корреляции будет сигнализировать о том, что на значение ряда сегодня сильно влияет значение ряда два момента времени назад.

АВТОКОРЕЛЛЯЦИОННАЯ ФУНКЦИЯ (ACF)

Теперь, зная что такое автокорреляция, определим автокорреляционную функцию (AutoCorrelation Function, ACF):

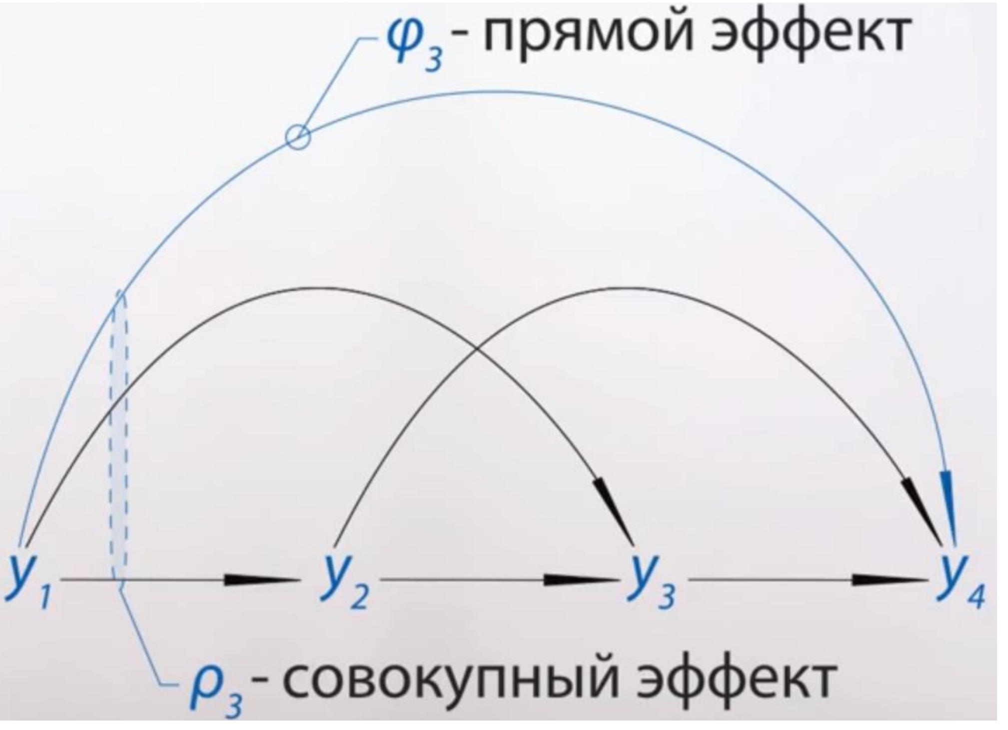
- для каждого значения $k = 0, 1, 2, \dots$ посчитаем корреляцию ряда в текущий момент времени и k моментов времени назад
- по оси x отложим k, по оси y - полученные значения корреляции





ЧАСТНАЯ АВТОКОРРЕЛЯЦИЯ (PACF)

Когда мы говорим об автокорреляции между рядом в моментами времени t и $t-k$, мы считаем не чистое влияние y_{t-k} на y_t, а на самом деле совокупный эффект, с которым ряд в моменты времени $t-k, t-k+1, \ldots, t-1$ влияет на ряд в момент времени t.



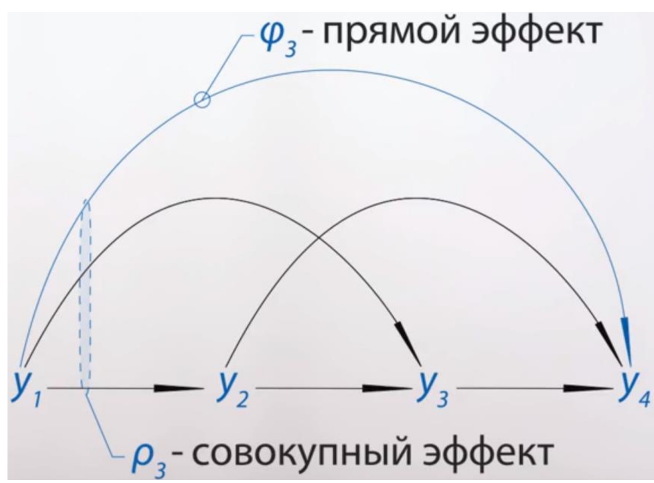
ЧАСТНАЯ АВТОКОРРЕЛЯЦИЯ (PACF)

Это не совсем то, что мы бы хотели видеть. Поэтому определяют частную (частичную) автокорреляцию **–** это часть корреляции между моментами времени t и $t - k$, которая не объясняется промежуточными корреляциями.

Если говорить точнее, то пусть

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ \dots + \varphi_k y_{t-k} + \varepsilon_t$$

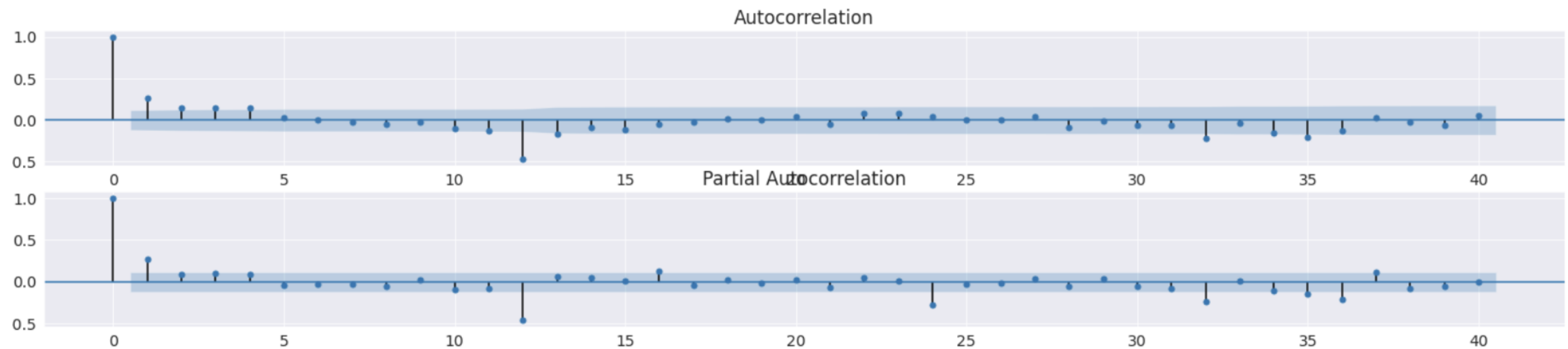
Тогда частная корреляция между рядом в момент времени t и $t - k$ равна φ_k.



ACF И PACF ДЛЯ ПОДБОРА ГИПЕРПАРАМЕТРОВ МОДЕЛИ ARMA

Правило:

p и q - число значимых лагов за период (если период/сезон есть) - то есть столбиков, выходящих за закрашенные области вокруг оси x.



Из графика можно определить кандидаты для оптимальных p и q:

- p=1 (по графику PACF)
- q=1,2,3,4 (по графику ACF).

ПРАКТИКА-2

- Statsmodels intro
- Statsmodels ARIMA

SARIMA

- В случае, если в ряде есть и тренд, и сезонность, то можно использовать модель ***SARIMA***, которая учитывает в себе и тренд, и сезонные эффекты.
- У модели SARIMA два набора гиперпараметров:
- *order*= (*p*,*d*,*q*) - те же гиперпараметры, что у ARIMA
- *seasonal_order*= (*P*,*D*,*Q*,*S*) - сезонные гиперпараметры

С увеличением числа гиперпараметров растет сложность их подбора. Подбирать гиперпараметры можно двумя способами:

- Анализируя графики функций ACF и PACF
- Перебором, минимизируя ошибку модели

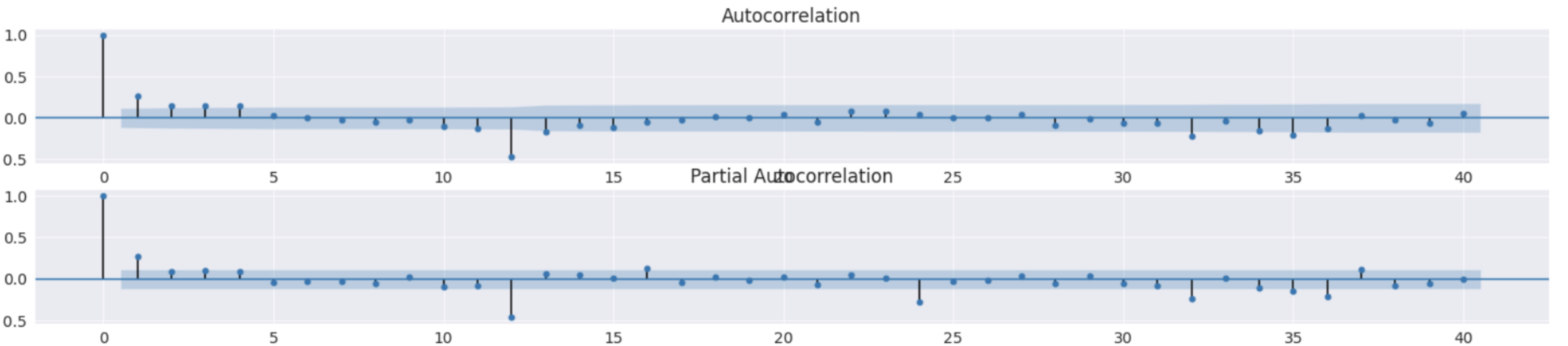
Хорошо работает комбинация этих подходов.

SARIMAX

- Иногда, кроме непосредственно прогнозируемого временного ряда, в данных есть другие факторы, также зависящие от времени - они называются ***экзогенными факторами***. Для улучшения качества прогноза кроме самого ряда полезно учитывать экзогенные факторы.
- Модель, которая позволяет учесть эти факторы называется ***SARIMAX*** (Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors model). В модели появляются дополнительные гиперпараметры:
- endog - прогнозируемый ряд
- exog - столбцы с экзогенными факторами.

ПОДБОР ГИПЕРПАРАМЕТРОВ (P,D,Q,S) ПО ГРАФИКАМ ACF И PACF

- Построим графики ACF и PACF для ряда уже без тренда
- Сезон S определяется по ACF визуально - это период, через который корреляции начнут себя примерно повторять



- По графику ACF видим, что сезон S=12
- P - число значимых пиков PACF в сезон (при сезоне S=12 анализируем столбцы с номерами 12,24,36…): $P = 1,2,3$
- Q - число значимых пиков ACF в сезон: $Q = 1$

ПОДБОР ГИПЕРПАРАМЕТРОВ ПЕРЕБОРОМ

Гиперпараметры (p,d,q) и (P,D,Q,S) можно найти перебором - при переборе мы минимизируем некоторую ошибку модели. Это может быть классическая ***RMSE***.

- Но чаще в моделях SARIMA(X) используют ***AIC***/***BIC*** критерии для поиска оптимальных гиперпараметров.
- Простыми словами, AIC и BIC - это модификации ошибки модели с учетом ее сложности.

ПОДБОР ГИПЕРПАРАМЕТРОВ ПЕРЕБОРОМ

- Формула для критерия Акаике (AIC):

$$\boldsymbol{AIC = -2ln\Pi + 2k}$$

- Формула для Байесовского информационного критерия (BIC):

$$\boldsymbol{BIC = -2ln\Pi + 2k \cdot lnn}$$

Π - правдоподобие (в классических моделях $ln\Pi$ часто совпадает с известными функциями потерь: *MSE* (регрессия) или *log-loss* (классификация))

k - число признаков (весов), используемых в модели (= сложность модели)

n - число объектов (константа)

СТАТИСТИЧЕСКИЕ МЕТОДЫ (ИТОГ)

Подведем итог по статистическому подходу. Он состоит в следующем:

- проверяем ряд на стационарность и в случае нестационарности приводим его к стационарному (либо самостоятельно, либо внутри готовых моделей ARIMA/SARIMA/SARIMAX)
- делаем прогноз для стационарного ряда с помощью линейной регрессии (авторегрессии)
- делаем обратные преобразования для получения прогноза для исходного ряда (либо самостоятельно, либо внутри готовых моделей ARIMA/SARIMA/SARIMAX)
- Подход дает хорошие результаты, однако требует понимание математики методов.

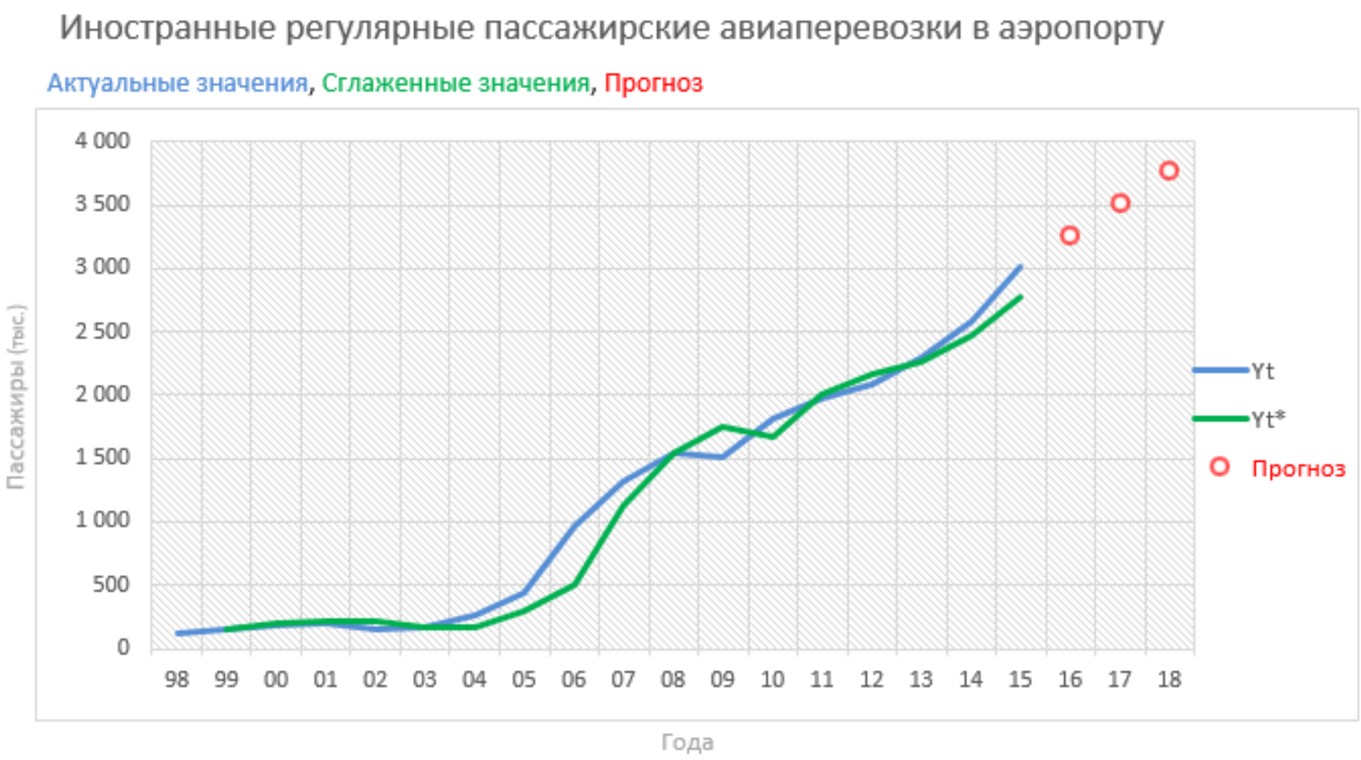
ПРАКТИКА-3

- SARIMAX intro
- SARIMAX

АДАПТИВНЫЕ МОДЕЛИ

АДАПТИВНЫЕ МЕТОДЫ

Адаптивные методы прогнозирования временных рядов представляют из себя методы, цель которых заключается в построении самокорректирующихся моделей, которые способны отражать изменяющееся во времени поведение ряда.



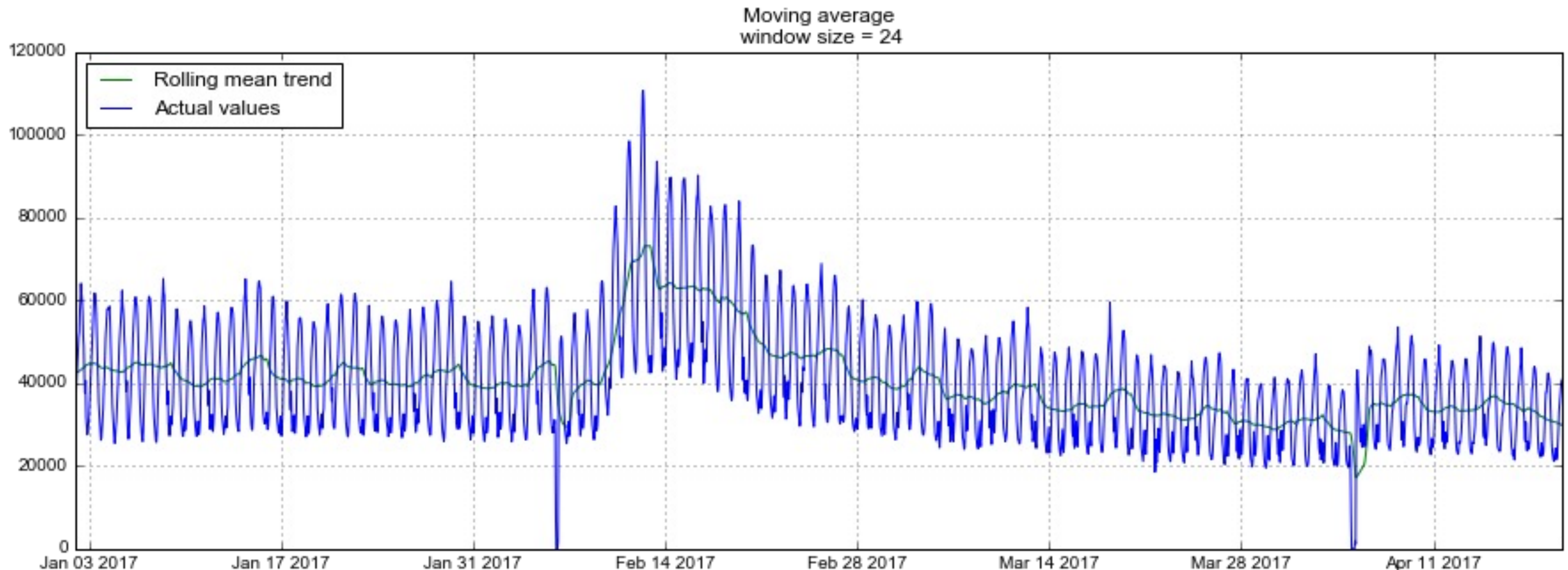
СКОЛЬЗЯЩЕЕ СРЕДНЕЕ

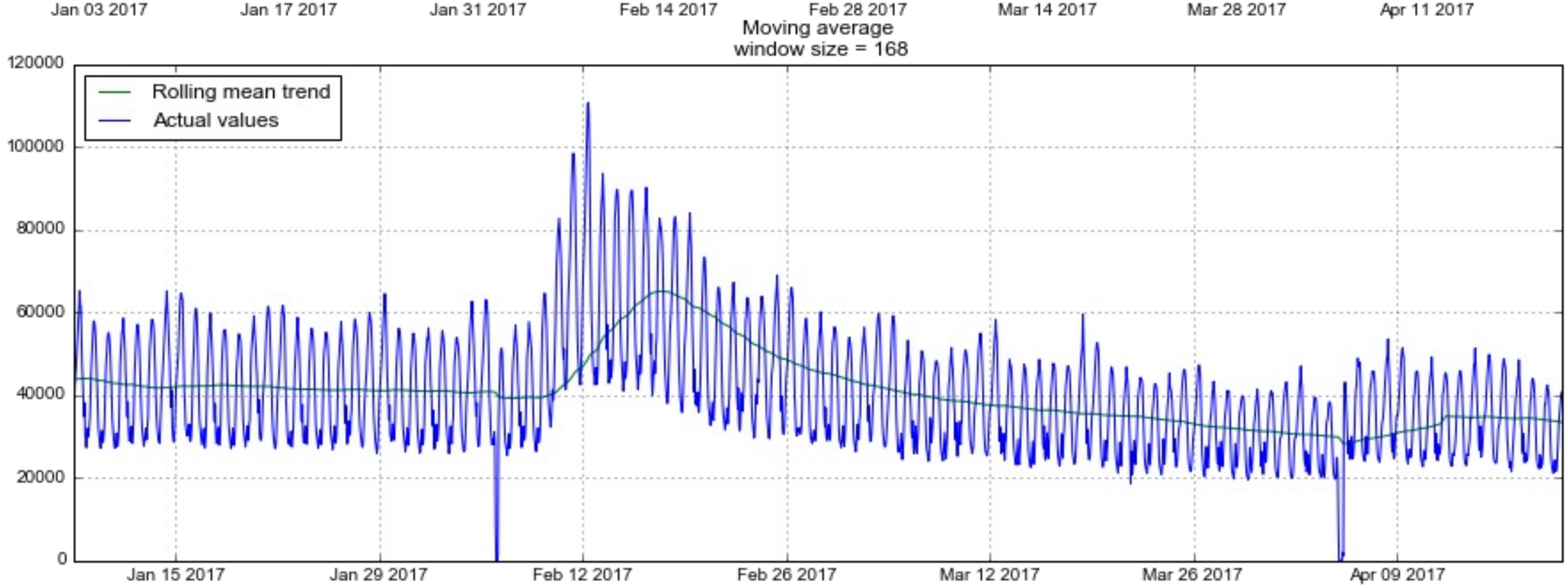
$$\hat{y}_{t+1} = \frac{1}{k}\sum_{i=0}^{k-1} y_{t-i}$$

- *чтобы сделать прогноз на следующий период времени, надо знать значение на текущий период (т.е. долгосрочный прогноз невозможен)*

\+ *сглаживает данные*

СКОЛЬЗЯЩЕЕ СРЕДНЕЕ





ЭКСПОНЕНЦИАЛЬНОЕ СКОЛЬЗЯЩЕЕ СРЕДНЕЕ

Идея: на значение ряда в данный момент времени больше всего влияет значение в предыдущий момент времени, затем – значение в предпредыдущий момент времени и т.д (то есть более поздние данные – более важные).

Пример:

$$EMA(t) = \frac{1}{2}p_t + \frac{1}{4}p_{t-1} + \frac{1}{8}p_{t-2} + \cdots$$

ЭКСПОНЕНЦИАЛЬНОЕ СКОЛЬЗЯЩЕЕ СРЕДНЕЕ

Пусть ряд ведет себя следующим образом:

$$y_t = l_t + \varepsilon_t,$$

- l_t - некоторое медленно меняющееся во времени значение (level)
- ε_t - шумовая компонента (со средним значением ноль).

Тогда прогнозировать мы должны компоненту l_t.

ЭКСПОНЕНЦИАЛЬНОЕ СКОЛЬЗЯЩЕЕ СРЕДНЕЕ

- Простейшая регрессионная модель – константа:

$$\boldsymbol{\hat{y}_{t+d} = \hat{l}_t = l}$$

Минимизируем квадратичную ошибку с весами β^i, убывающими в прошлое:

$$\sum_{i=0}^{t} \beta^i (y_{t-i} - l)^2 \to \min_c$$

<u>Аналитическое решение (формула Надарая-Ватсона):</u>

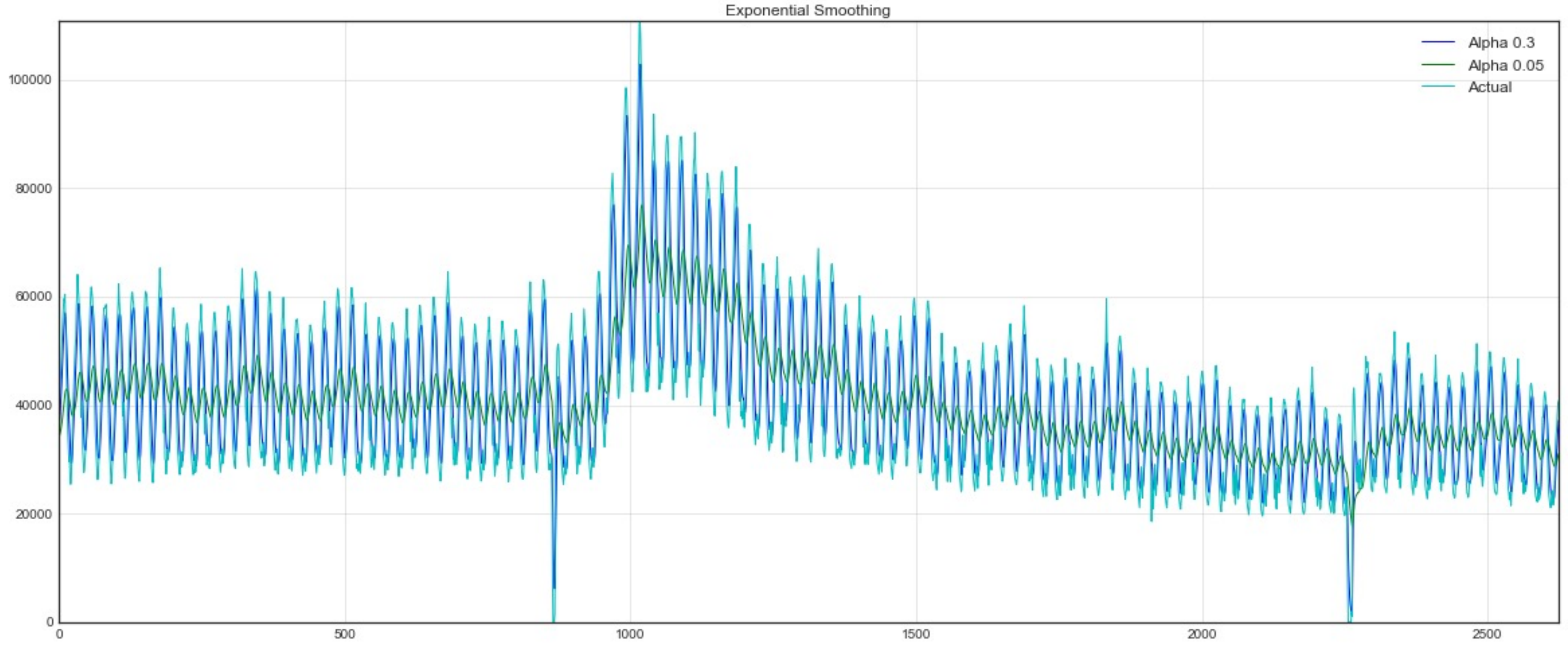
$$\hat{y}_{t+1} = l = \frac{\sum_{i=0}^{t} \beta^i y_{t-i}}{\sum_{i=0}^{t} \beta^i}$$

ЭКСПОНЕНЦИАЛЬНОЕ СКОЛЬЗЯЩЕЕ СРЕДНЕЕ (ЭСС)

Утверждение. Модель ЭСС можно записать в виде

$$\boldsymbol{\hat{y}_{t+1} = \alpha y_t + (1-\alpha)\hat{y}_t, \alpha \in (0,1)}$$

- чем больше α, тем больше вес последних точек
- чем меньше α, тем сильнее сглаживание



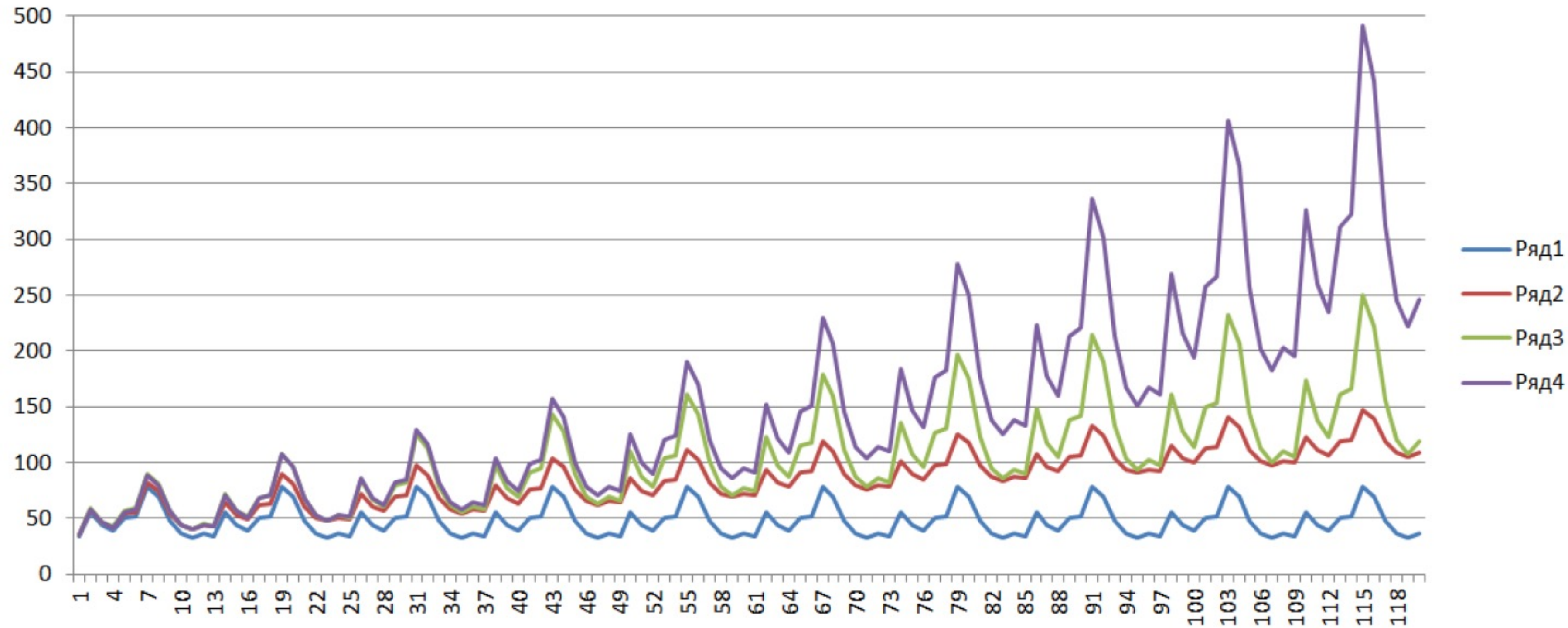
ЭКСПОНЕНЦИАЛЬНОЕ СКОЛЬЗЯЩЕЕ СРЕДНЕЕ (ЭСС)

Оптимальное значение α подбираем по скользящему контролю:

$$Q(\alpha) = \sum_{t=t_0}^{T} (\hat{y}_t(\alpha) - y_t)^2 \to \min_{\alpha}$$

- при $\alpha \in (0,0.3)$ ряд стационарен, модель ЭСС работает
- при $\alpha \in (0.3,1)$ ряд нестационарен, нужна модель тренда

МОДЕЛИ С ТРЕНДОМ И СЕЗОННОСТЬЮ



- Ряд 1 - сезонность без тренда
- Ряд 2 - линейный тренд, аддитивная сезонность
- Ряд 3 – линейный тренд, мультипликативная сезонность
- Ряд 4 – экспоненциальный тренд, мультипликативная сезонность

МОДЕЛИ С ТРЕНДОМ И СЕЗОННОСТЬЮ

Модель Хольта

- модель линейного тренда

$$\boldsymbol{\hat{y}_{t+d} = a_t + b_t d},$$

где a_t, b_t - адаптивные компоненты линейного тренда.

- формулы для a_t, b_t:

$$a_t = \alpha_1 y_t + (1 - \alpha_1)(a_{t-1} + b_{t-1})$$
$$b_t = \alpha_2(a_t - a_{t-1}) + (1 - \alpha_2)b_{t-1},$$

где α_1, α_2 - параметры сглаживания.

МОДЕЛИ С ТРЕНДОМ И СЕЗОННОСТЬЮ

Модель Винтерса (с аддитивной/мультипликативной сезонностью)

- модель мультипликативной сезонности периода s

$$\widehat{\boldsymbol{y}}_{\boldsymbol{t}+\boldsymbol{d}} = \boldsymbol{a}_{\boldsymbol{t}} \cdot \boldsymbol{\theta}_{\boldsymbol{t}+(\boldsymbol{d} \bmod \boldsymbol{s})-\boldsymbol{s}},$$

$\theta_0, \dots, \theta_{s-1}$ - сезонный профиль периода s без тренда.

формулы для a_t, b_t, θ_t:

$$a_t = \alpha_1(y_t/\theta_{t-s}) + (1-\alpha_1)\, a_{t-1}$$
$$\theta_t = \alpha_2(y_t/a_t) + (1-\alpha_2)\theta_{t-s},$$

где α_1, α_2 - параметры сглаживания.

МОДЕЛИ С ТРЕНДОМ И СЕЗОННОСТЬЮ

<u>Модель Хольта-Винтерса</u>

- модель линейного тренда с аддитивной сезонностью

$$\boldsymbol{\hat{y}_{t+d} = (a_t + b_t d) + \theta_{t+(d \bmod s)-s}},$$

$a_t + b_t d$ – тренд, очищенный от сезонных колебаний,

$\theta_0, \dots, \theta_{s-1}$ - сезонный профиль периода s без тренда.

формулы для a_t, b_t, θ_t:

$$a_t = \alpha_1(y_t - \theta_{t-s}) + (1 - \alpha_1)(a_{t-1} + b_{t-1})$$
$$b_t = \alpha_2(a_t - a_{t-1}) + (1 - \alpha_2)b_{t-1}$$
$$\theta_t = \alpha_3(y_t - a_t) + (1 - \alpha_3)\theta_{t-s},$$

где $\alpha_1, \alpha_2, \alpha_3$ - параметры сглаживания.

МОДЕЛИ С ТРЕНДОМ И СЕЗОННОСТЬЮ

Модель Хольта-Винтерса с аддитивным трендом и мультипликативной сезонностью

- модель мультипликативной сезонности периода s с линейным трендом

$$\boldsymbol{\hat{y}_{t+d} = (a_t + b_t d) \cdot \theta_{t+(d \bmod s)-s}},$$

$a_t + b_t d$ – тренд, очищенный от сезонных колебаний,

$\theta_0, \dots, \theta_{s-1}$ - сезонный профиль периода s без тренда.

формулы для a_t, b_t, θ_t:

$$a_t = \alpha_1 (y_t / \theta_{t-s}) + (1 - \alpha_1)(a_{t-1} + b_{t-1})$$
$$b_t = \alpha_2 (a_t - a_{t-1}) + (1 - \alpha_2) b_{t-1}$$
$$\theta_t = \alpha_3 (y_t / a_t) + (1 - \alpha_3) \theta_{t-s},$$

где $\alpha_1, \alpha_2, \alpha_3$ - параметры сглаживания.

МОДЕЛИ С ТРЕНДОМ И СЕЗОННОСТЬЮ

Модель Хольта-Винтерса с экспоненциальным трендом и мультипликативной сезонностью

- модель мультипликативной сезонности периода s с экспоненциальным трендом

$$\boldsymbol{\hat{y}_{t+d} = a_t(r_t)^d \cdot \theta_{t+(d \bmod s)-s}},$$

$a_t(r_t)^d$ – тренд, очищенный от сезонных колебаний,

$\theta_0, \dots, \theta_{s-1}$ - сезонный профиль периода s без тренда.

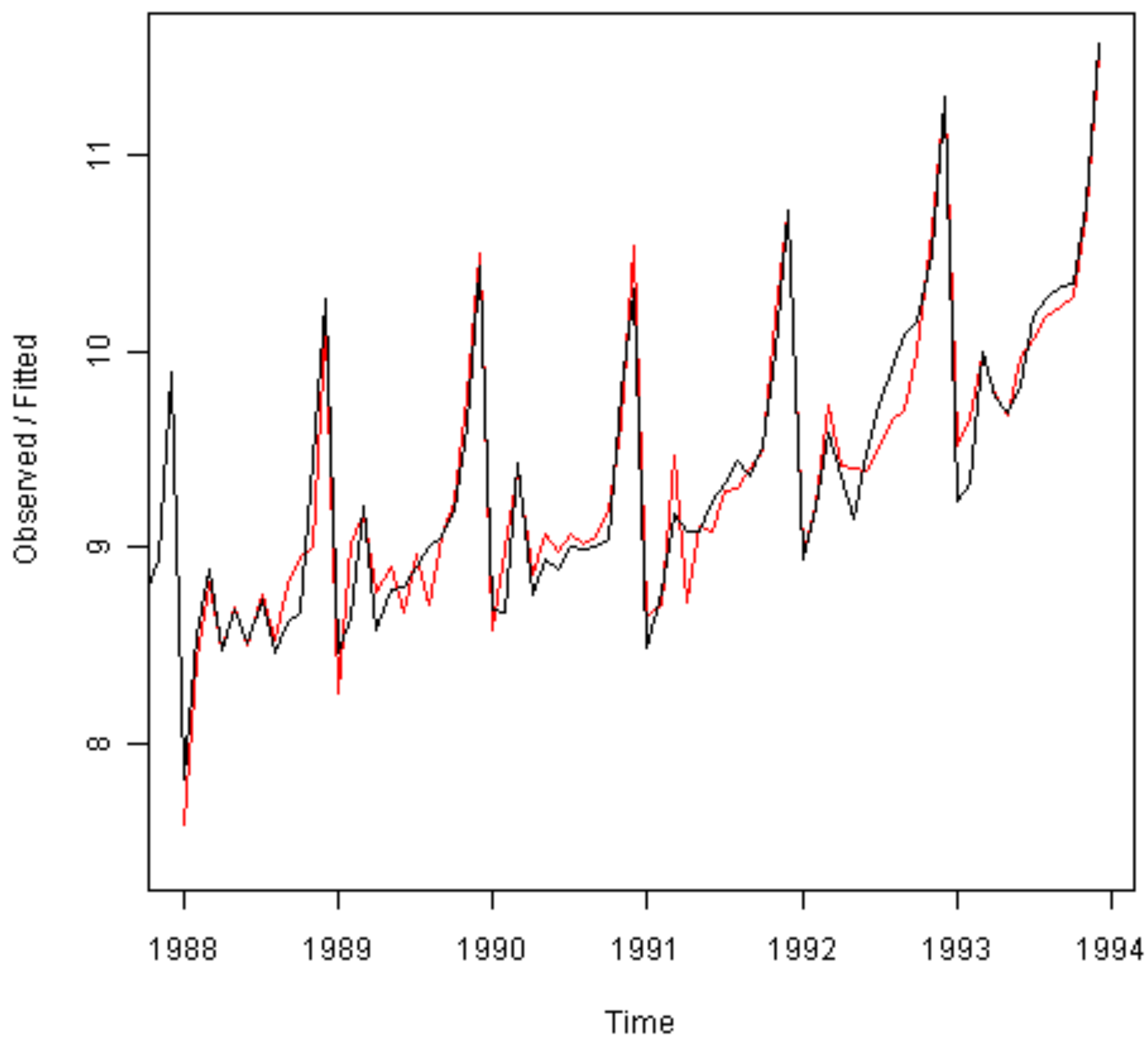
формулы для a_t, b_t, θ_t:

$$a_t = \alpha_1(y_t/\theta_{t-s}) + (1-\alpha_1)a_{t-1}r_{t-1}$$
$$r_t = \alpha_2(a_t/a_{t-1}) + (1-\alpha_2)r_{t-1}$$
$$\theta_t = \alpha_3(y_t/a_t) + (1-\alpha_3)\theta_{t-s},$$

где $\alpha_1, \alpha_2, \alpha_3$ - параметры сглаживания.

МОДЕЛИ С ТРЕНДОМ И СЕЗОННОСТЬЮ

Модель Уинтерса с линейным трендом (модель Хольта-Уинтерса)



ПРАКТИКА-4

- Простое экспоненциальное сглаживание
- Модели Хольта-Винтерса