Анализ временных рядов

часть 1

Олег Дешеулин

НИУ ВШЭ

9 ноября, 2018

Что такое временной ряд?

Временной ряд: $y_1, \ldots, y_T, \ldots, y_t \in \mathbb{R}$ - значения признака, измеренные через постоянные временные интервалы.
Задача прогнозирования: найти функцию

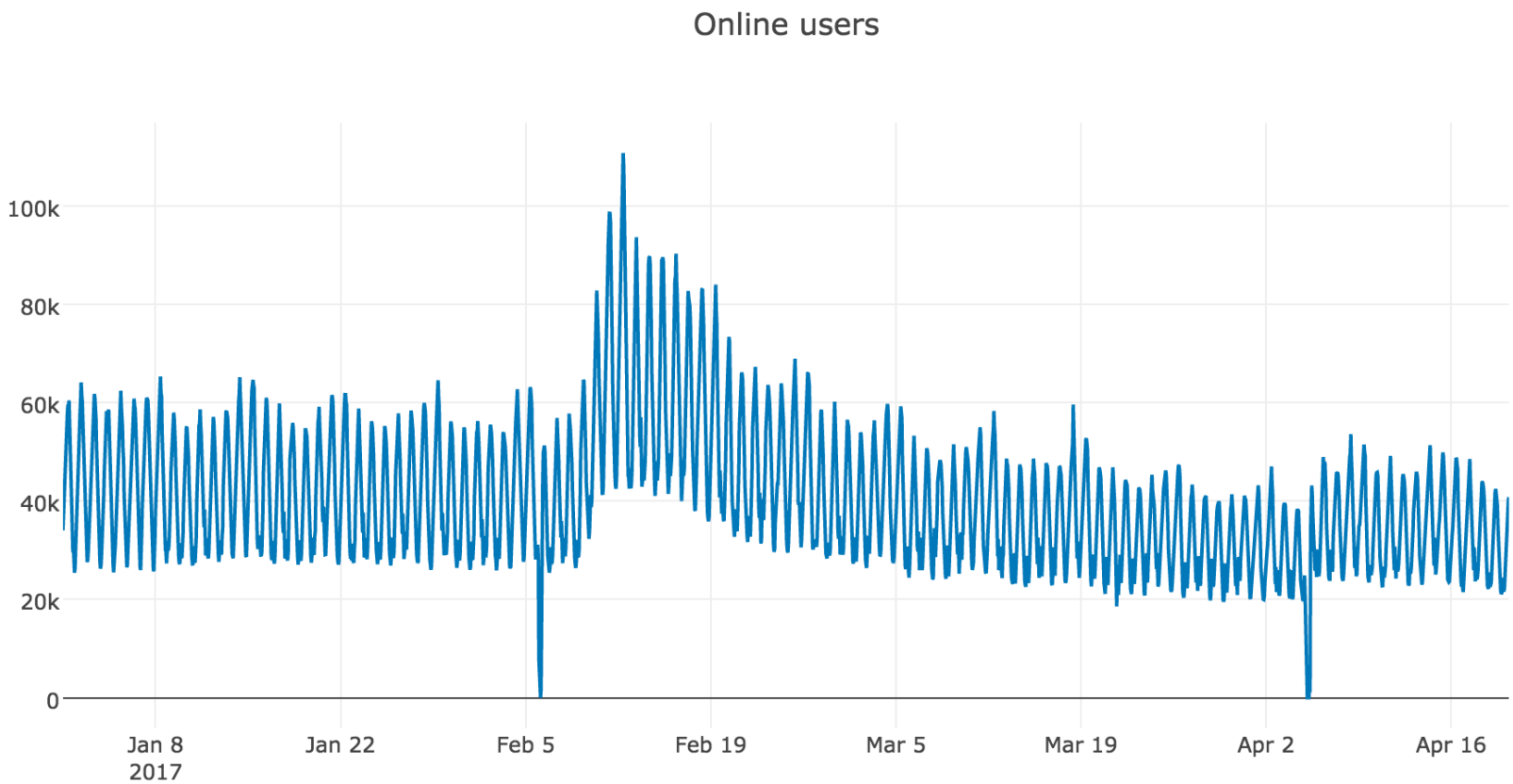
$$y_{T+d} \approx \mathrm{f}_T(y_T, \ldots, y_1, d) \equiv \hat{y}_{T+d|T},$$

где $d \in 1, \ldots, D$ - отсрочка прогноза, D - горизонт прогнозирования

Компоненты временных рядов

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

Что это значит?



Авторегрессия

Авторегрессионная (AR-(p)) модель –

$$\hat{y}_t = C + \sum_{i=1}^{p} w_i y_{t-i} + \epsilon_t$$

Например (AR-(1)):

$$\hat{y}_t = C + w_{t-1} y_{t-i} + \epsilon_t$$

Хотим найти w_i - параметры модели, ϵ_i - шум

Простейшие методы прогнозирования

- средним:

$$\hat{y}_{T+d} = \frac{1}{T}\sum_{t=1}^{T} y_t;$$

- средним за последние k отсчётов (скользящее среднее):

$$\hat{y}_{T+d} = \frac{1}{k}\sum_{t=T-k}^{T} y_t;$$

- наивный:

$$\hat{y}_{T+d} = y_T;$$

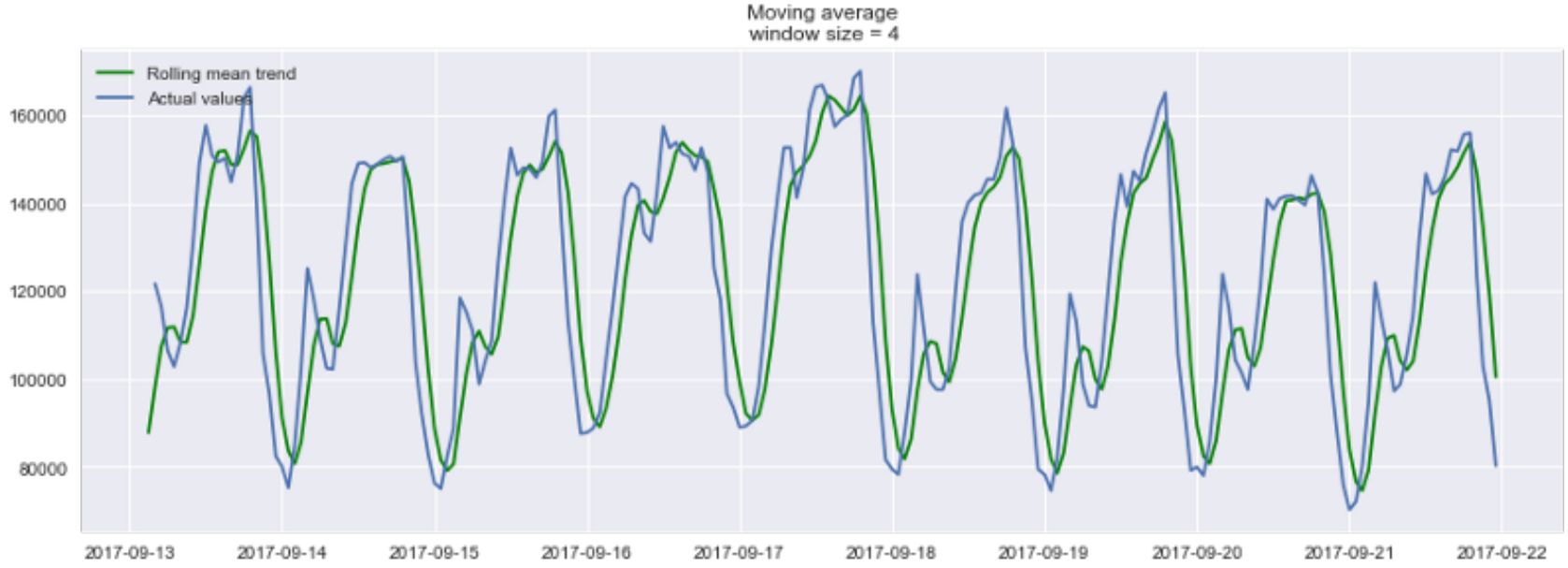
- наивный сезонный (s - период сезонности):

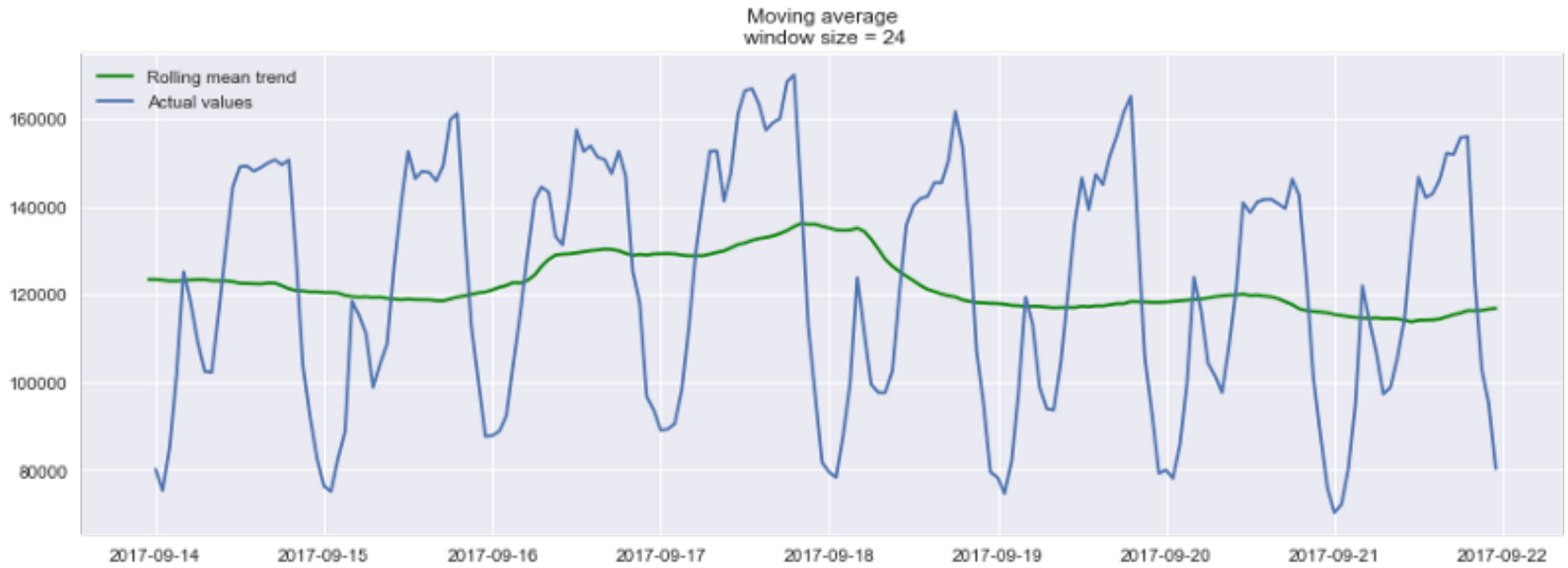
$$\hat{y}_{T+d} = y_{T+d-ks},$$
$$k = \lfloor \frac{d-1}{s} \rfloor + 1;$$

- экстраполяции тренда

$$\hat{y}_{T+d} = y_T + d\frac{y_t - y_1}{T-1}$$

Неужели это работает?





Да, скользящее среднее показывает дневной тренд

Взвешенное среднее

- Хотим учитывать разные элементы ряда с разными весами в прогнозе, как и было в авторегрессии:

$$\hat{y}_{T+d} = \frac{1}{k} \sum_{t=T-k}^{T} w_t y_t;$$

- Логично, например, большие веса отдавать наиболее свежим значениям - так приходит идея экспоненциального сглаживания

Простое экспоненциальное сглаживание

- Метод подходит для прогнозирования рядов без тренда и сезонности:

$$\hat{y}_{t+1|t} = l_t$$

$$l_t = \alpha y_t + (1-\alpha) l_{t-1} = \hat{y}_{t|t-1} + \alpha \cdot e_t$$

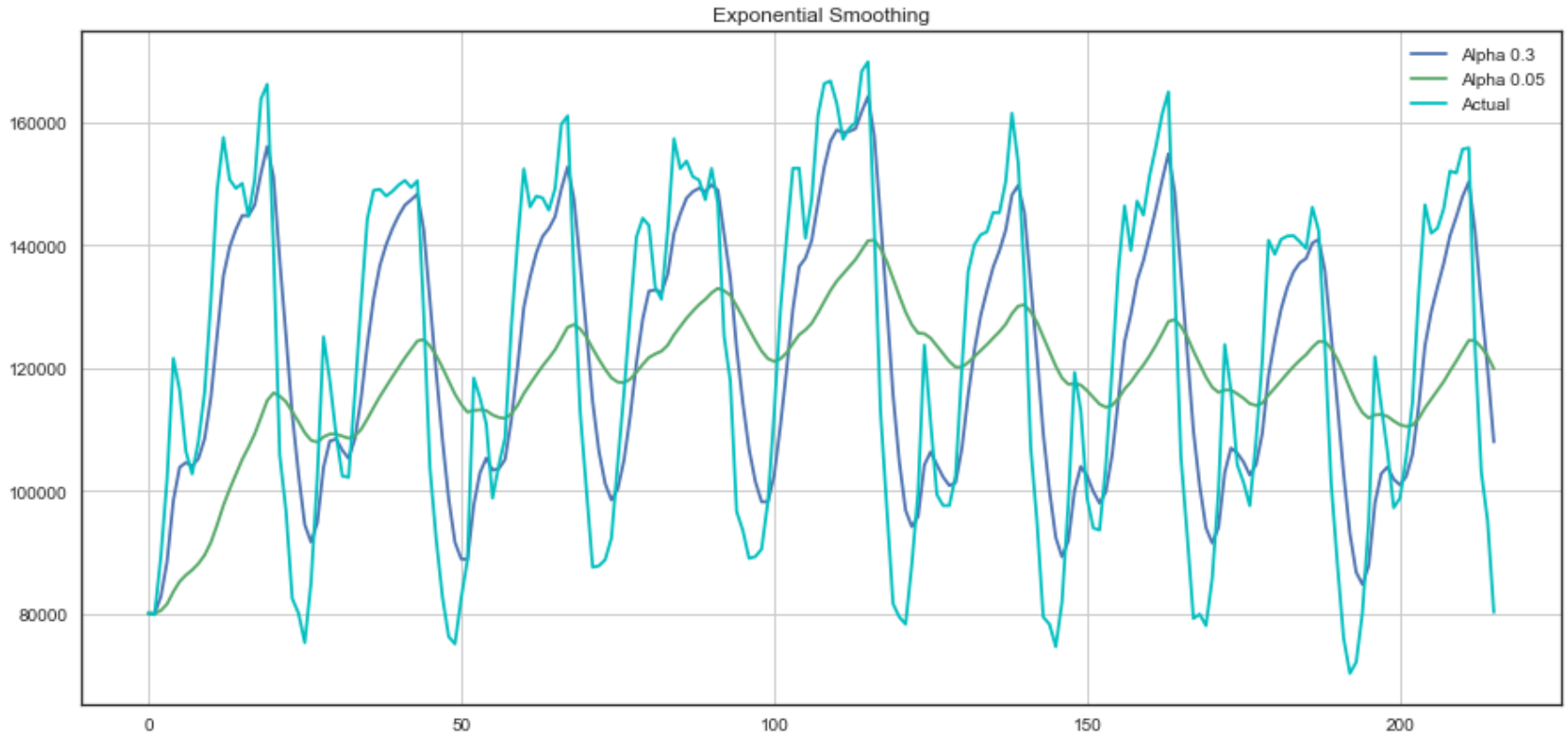
$e_t = y_t - \hat{y}_{t|t-1}$ - ошибка прогноза отсчёта времени t

- Прогноз зависит от l_0:

$$\hat{y}_{T+1|T} = \sum_{j=1}^{T-1} \alpha(1-\alpha)^j y_{T-1} + (1-\alpha)^T l_0$$

- Прогноз получается плоский, т.е. $\hat{y}_{t+d|t} = \hat{y}_{t+1|t}$

Простое экспоненциальное сглаживание



Методы, учитывающие тренд

Аддитивный линейный тренд (метод Хольта):

$$\hat{y}_{t+d|t} = l_t + db_t,$$

$$l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1}),$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1},$$

Мультипликативный линейный (экспоненциальный) тренд:

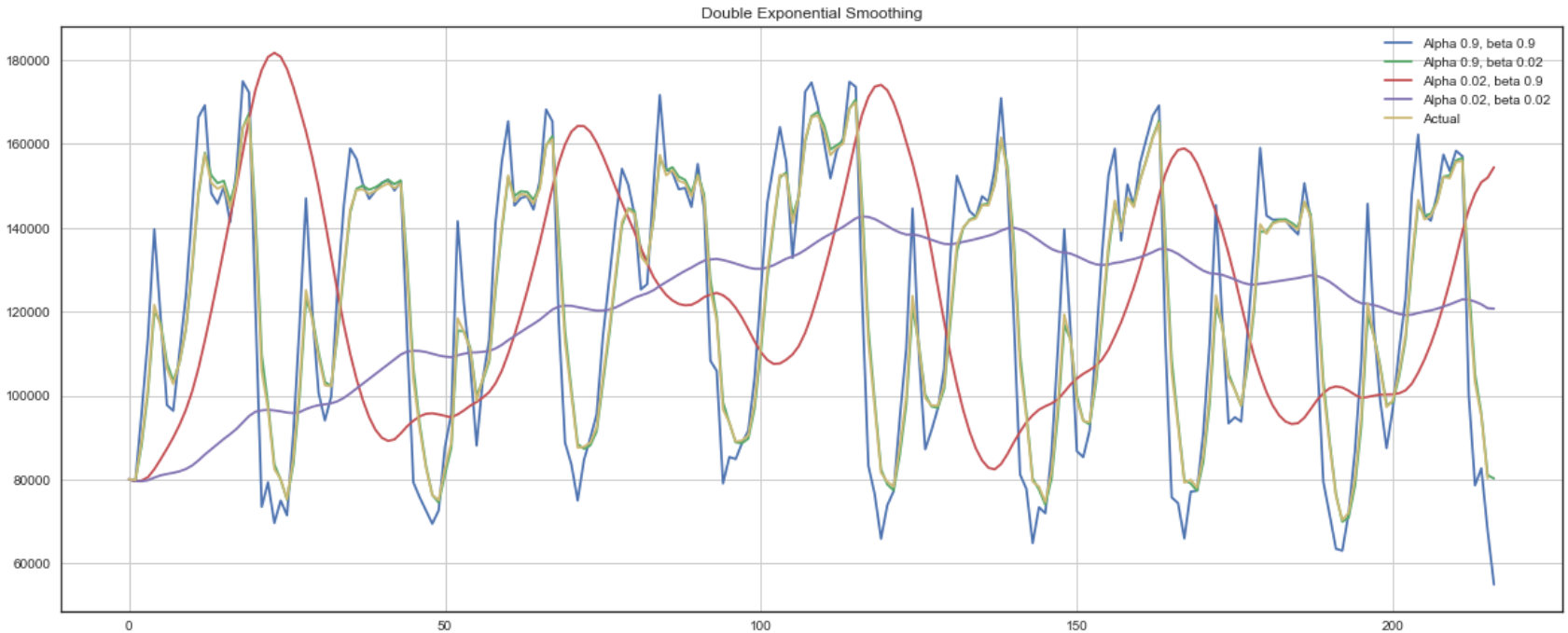
$$\hat{y}_{t+d|t} = l_t b_t^d,$$

$$l_t = \alpha y_t + (1-\alpha)l_{t-1}b_{t-1}),$$

$$b_t = \beta\frac{l_t}{l_{t-1}} + (1-\beta)b_{t-1},$$

$\alpha, \beta \in [0, 1]$.

Методы, учитывающие тренд



Методы, учитывающие сезонность

Мультипликативная сезонность (Хольта-Винтерса):

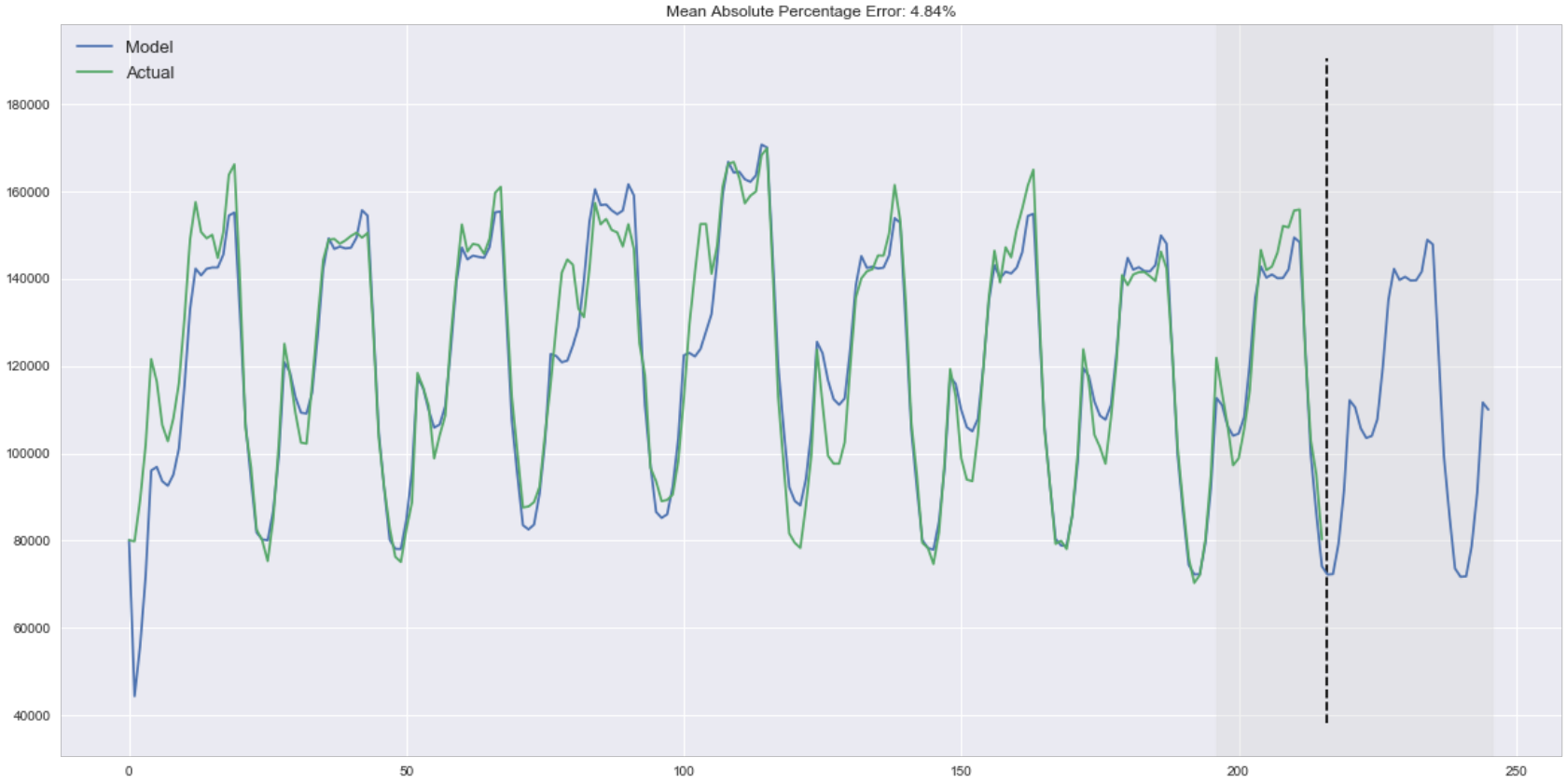
$$\hat{y}_{t+d|t} = (l_t + db_t)s_{t-m+(d \mod m)},$$

$$l_t = \alpha \frac{y_t}{s_{t-m}} + (1-\alpha)(l_{t-1} + b_{t-1}),$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1},$$

$$s_t = \gamma \frac{y_t}{l_{t-1} + bt - 1} + (1-\gamma)s_{t-m}.$$

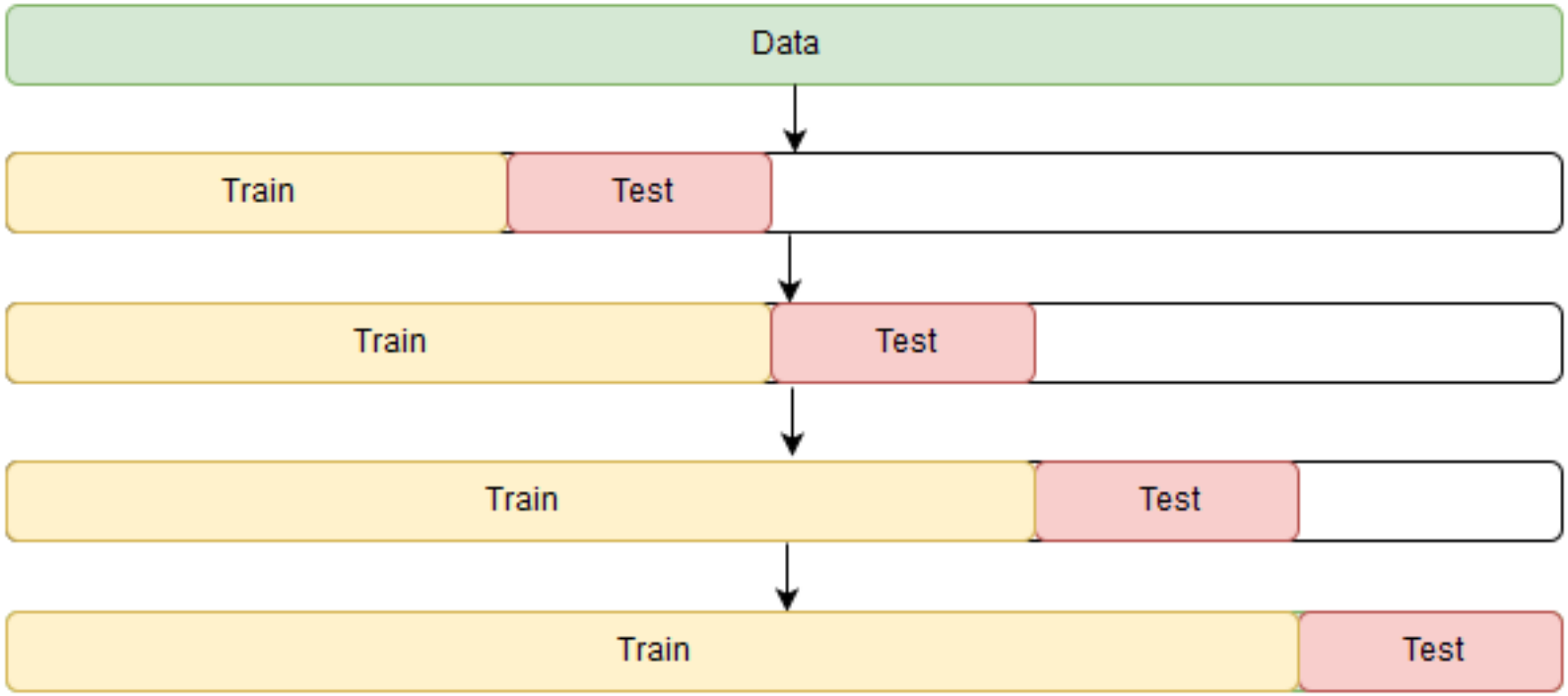
Методы, учитывающие сезонность



Аномалии


Mean Absolute Percentage Error: 4.84%
180000
160000
140000
120000
100000
80000
60000
40000
0
50
100
150
200
250
Model
Actual
Anomalies
Up/Low confidence

Кросс-валидация на временном ряду



Меры качества точечного прогноза

Mean Squared Error

$$MSE = \frac{1}{T - R + 1} \sum_{t=R}^{T} (\hat{y}_t - y_t)^2.$$

Mean Absolute Error

$$MAE = \frac{1}{T - R + 1} \sum_{t=R}^{T} |\hat{y}_t - y_t|.$$

То есть ровно такие же как для обычной регресии. На самом деле, всех этих метрик недостаточно, необходимо так же проверять остатки прогноза.

Стационарность

Ряд $y_1, \ldots, y_T$ **стационарен**, если $\forall s$ распределение $y_t, \ldots, y_{t+s}$ не зависит от t, т.е. его свойства не зависят от времени.

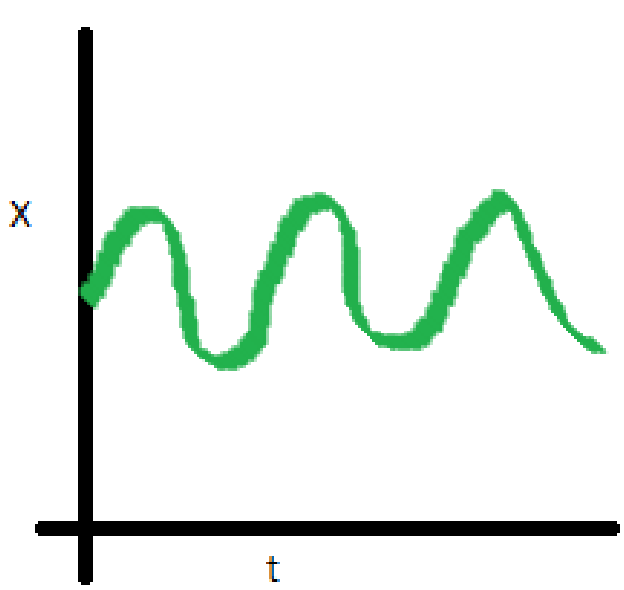
Стационарный ряд не меняет со временем свои характеристики, такие как матожидание, дисперсия и ковариации.

Ряды с трендом или сезонностью нестационарны.

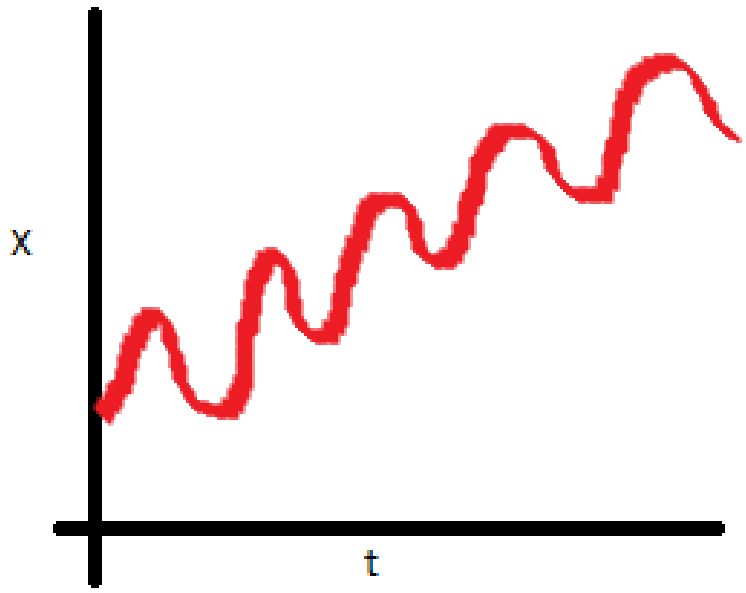
Ряды с непериодическими циклами стационарны, поскольку нельзя предсказать заранее, где будут находиться минимумы и максимумы.

Стационарность

Растет матожидание со временем:



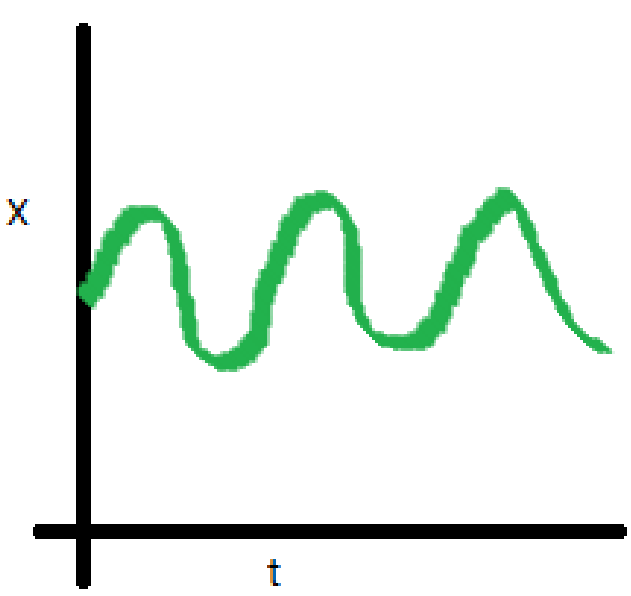
Stationary series



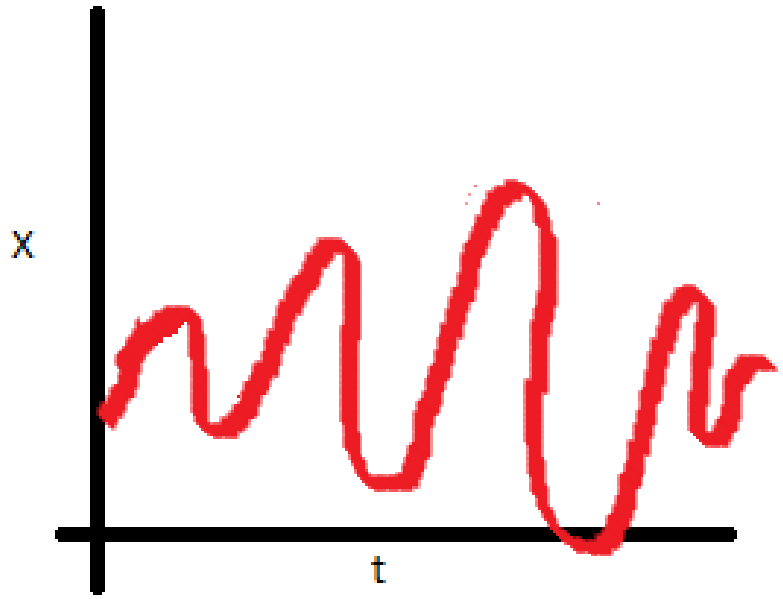
Non-Stationary series

Стационарность

Дисперсия зависит от периода:



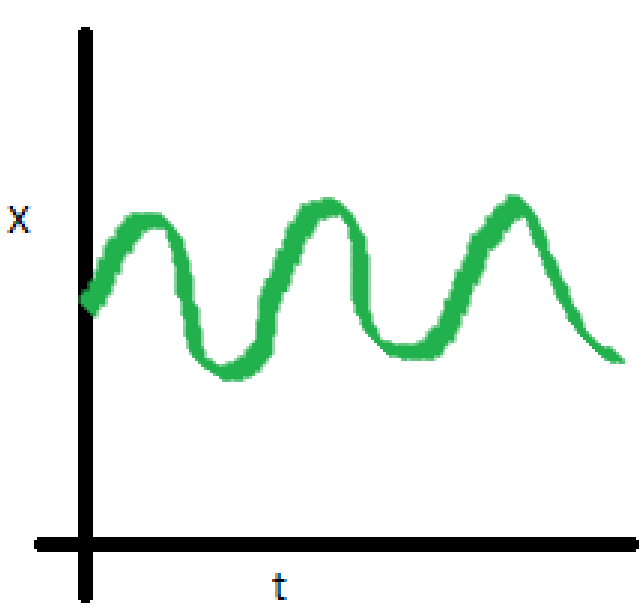
Stationary series



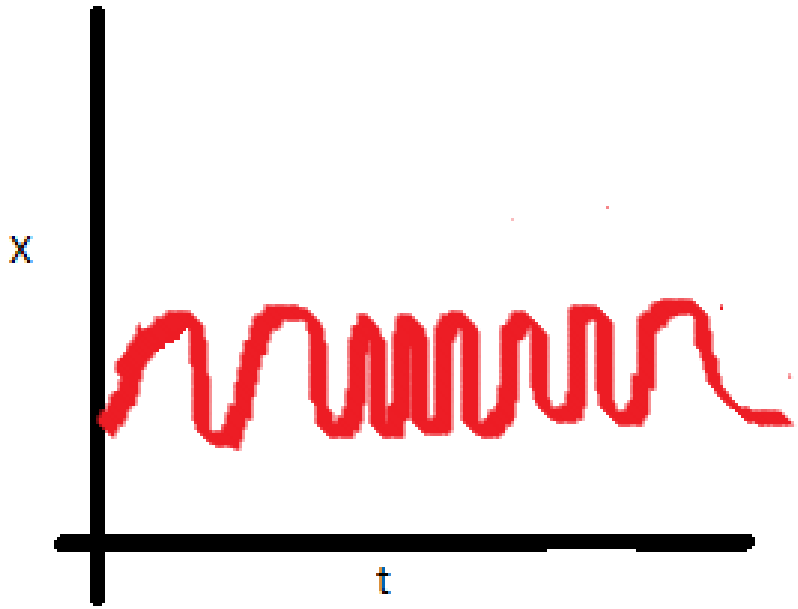
Non-Stationary series

Стационарность

Непостоянство ковариаций:



Stationary series



Non-Stationary series

Остатки

Остатки – разность между фактом и прогнозом:

$$\hat{\epsilon}_t = y_t - \hat{y}_t.$$

Прогнозы $\hat{y}_t$ могут быть построены с фиксированной отсрочкой:

$$\hat{y}_{R+d|R}, \dots, \hat{y}_{T|T-d},$$

или с фиксированным концом истории при разных отсрочках:

$$\hat{y}_{T-D+1|T-D}, \dots, \hat{y}_{T|T-d},$$

Необходимые свойства остатков прогноза

- Несмещённость – равенство среднего значения нулю: графички
- Неавтокореллированность – отсутствие неучтённнной зависимости от предыдущих наблюдений:
- Стационарность – отсутствие зависимости от времени
- Нормальность

Проверка на стационарность

Тест Дики-Фуллера (DF-тест):
Нулевая гипотеза H_0: $g = 0$ (существует единичный корень, ряд нестационарный)
Альтернативная гипотеза H_1: $g < 0$ (единичного корня нет, ряд стационарный)
Единичный корень:
Характеристическое уравнение (или характеристический полином) авторегрессионной модели временного ряда имеет корни, равные по модулю единице:

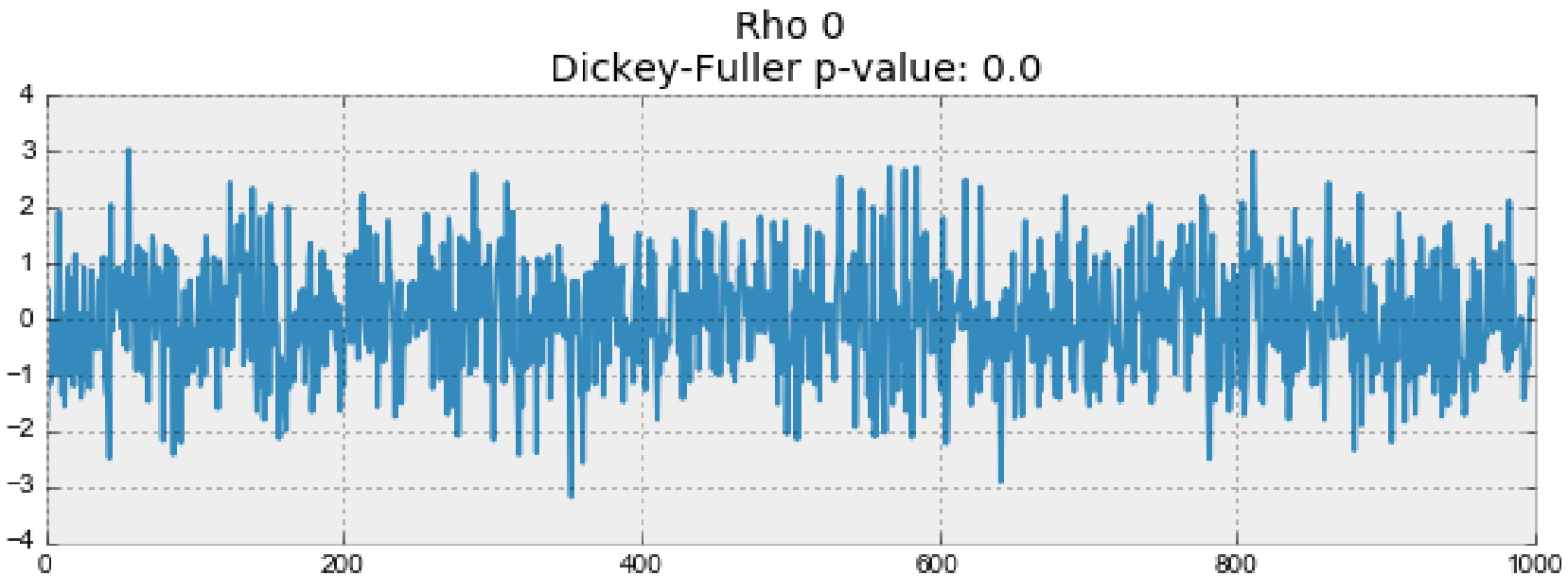
$$a(z) = 1 - \sum_{i=1}^{p} w_i z^i$$

Вообще говоря, если процесс стационарен корни находятся внутри единичного круга.

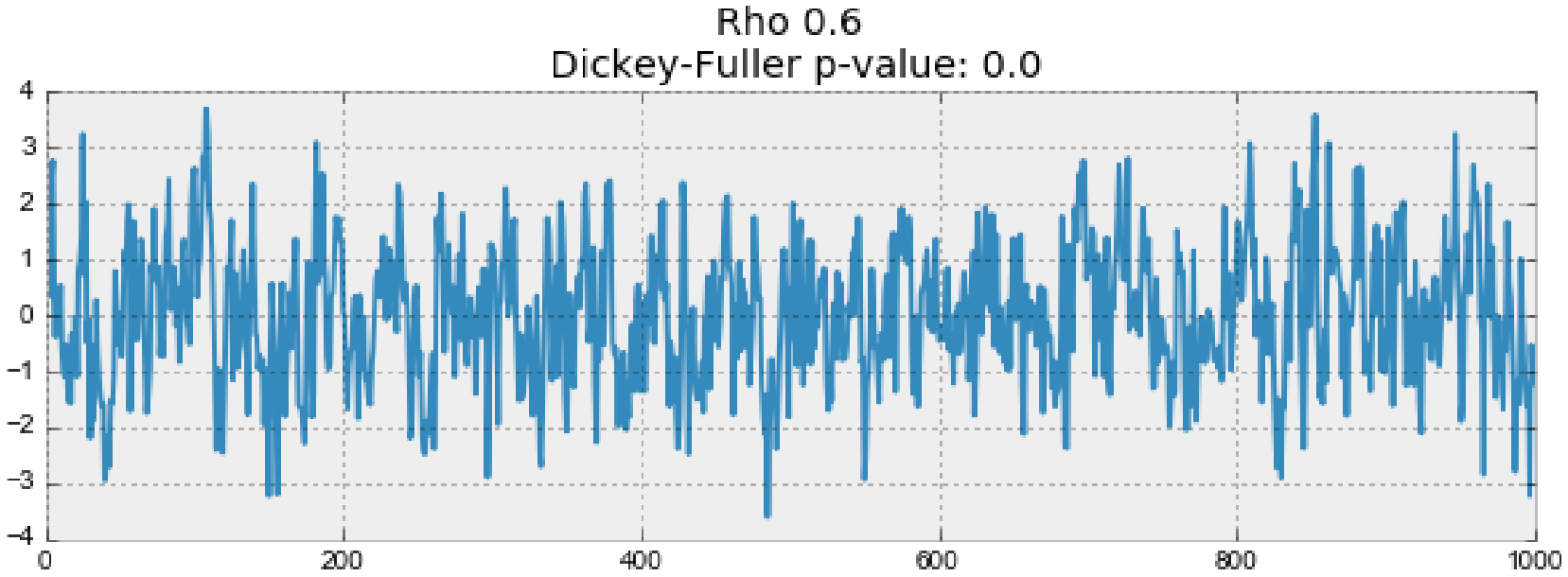
Что значит тест Дики-Фуллера?

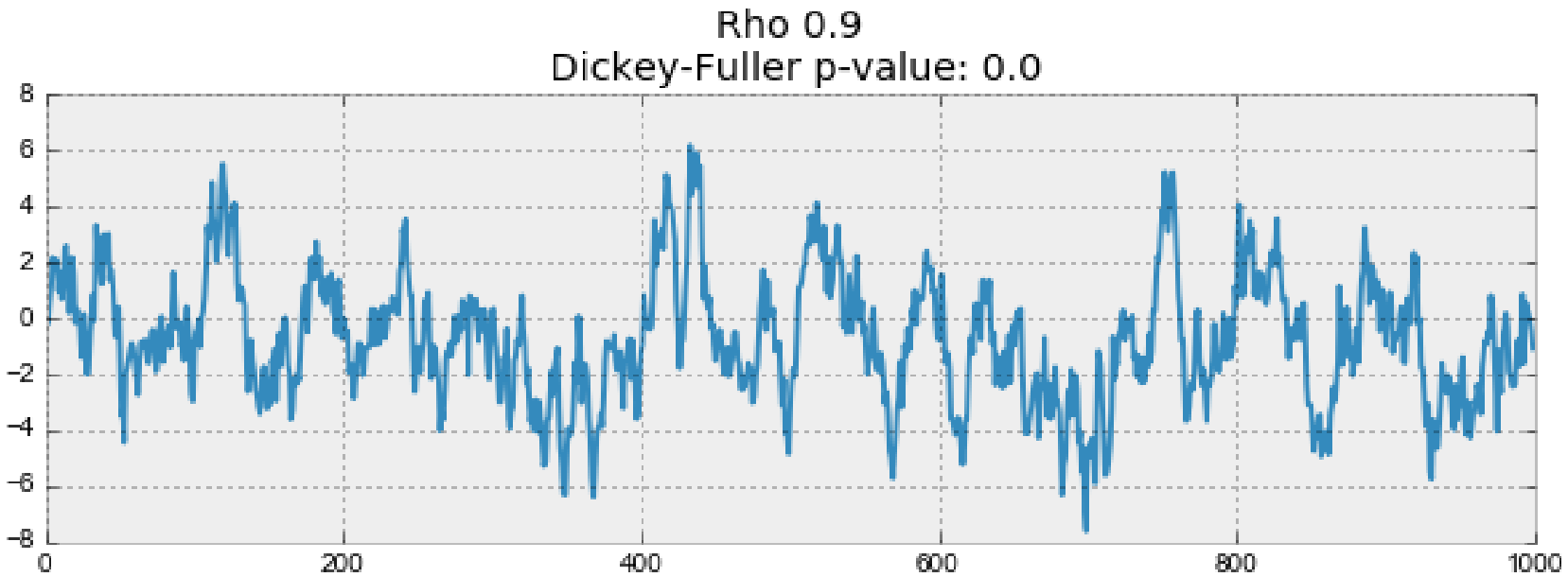
Генерируем стандарный нормальный шум. И порожаем им процесс, зависящий от ρ:

$$x_t = \rho x_{t-1} + e_t$$

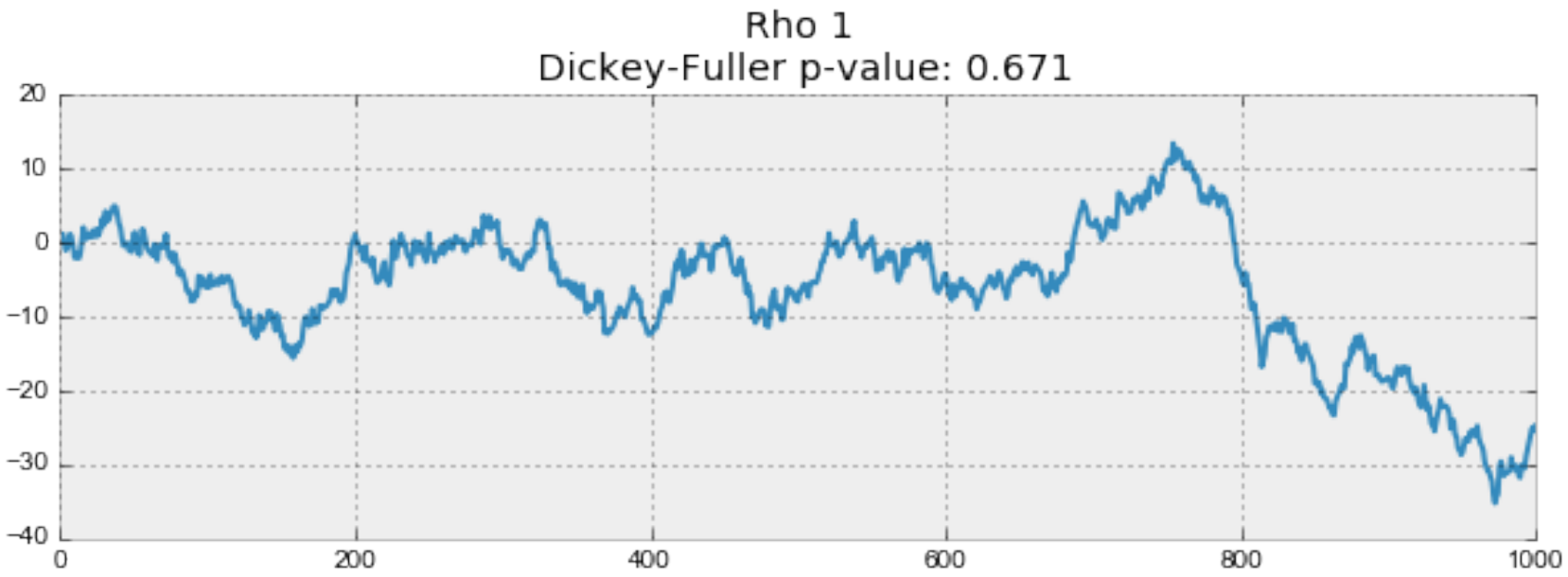


Что значит тест Дики-Фуллера?





Что значит тест Дики-Фуллера?



Если из нестационарного ряда первыми разностями удаётся получить стационарный, то он называется интегрированным первого порядка.

H_O отвергалась на первых трех графиках, и принялась на последнем.

Процесс может быть интегрированным с более высоким порядком тогда используют расширенный тест Дики-Фуллера.

Ссылки

[1]. http://www.machinelearning.ru/wiki/images/2/2d/Psad$_t s_e ts_2$017.*pdf*
[2]. https://habr.com/company/ods/blog/327242/
[3]. https://ru.wikipedia.org