Анализ временных рядов

часть 2

Алёна Ким

НИУ ВШЭ

9 ноября, 2018

Автокорреляционная функция (ACF)

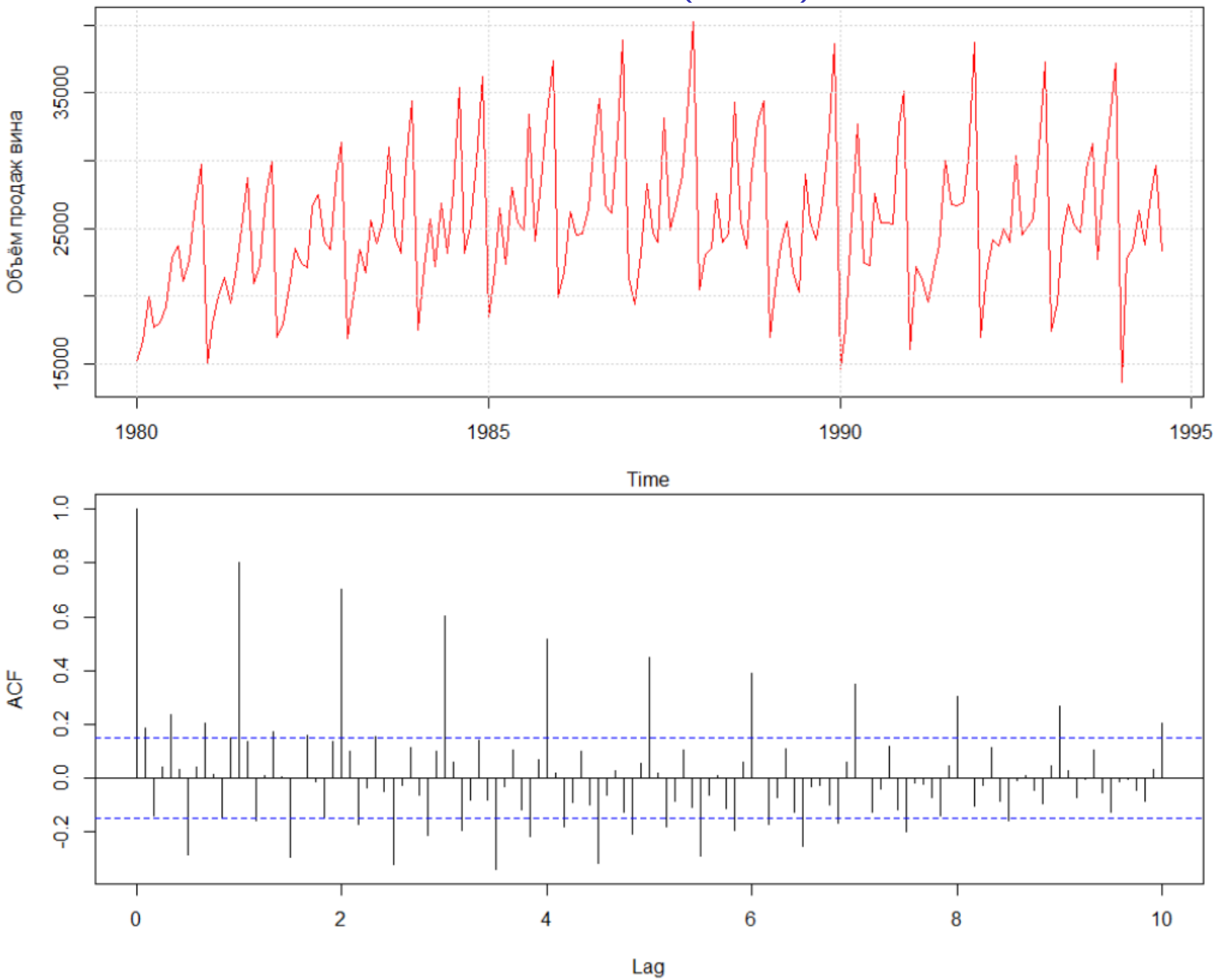
Временной ряд Y^T: $y_1, \ldots, y_T, \ldots, y_t \in \mathbb{R}$ - значения признака, измеренные через постоянные временные интервалы.
Автокорреляция:

$$r_\tau = r_{y_t t_{t+\tau}} = \frac{\sum_{t=1}^{T-\tau}(y_t - \vec{y})(y_{t+\tau} - \vec{y})}{\sum_{t=1}^{T}(y_t - \vec{y})^2}, \vec{y} = \frac{1}{T}\sum_{t=1}^{T} y_t.$$

$r_\tau \in [-1, 1], \tau$ – лаг автокорреляции

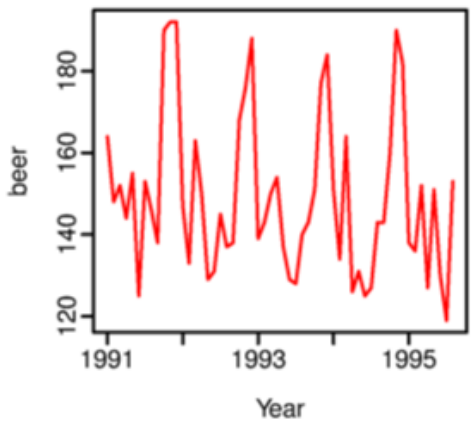
Нулевая гипотеза: $H_0 : r_\tau = 0$;
Альтернативная гипотеза: $H_1 : r_\tau \neq 0$;
Статистика: $T(Y^T) = \frac{r_\tau \sqrt{T-\tau-2}}{\sqrt{1-r_\tau^2}}$;

Автокорреляционная функция (ACF)



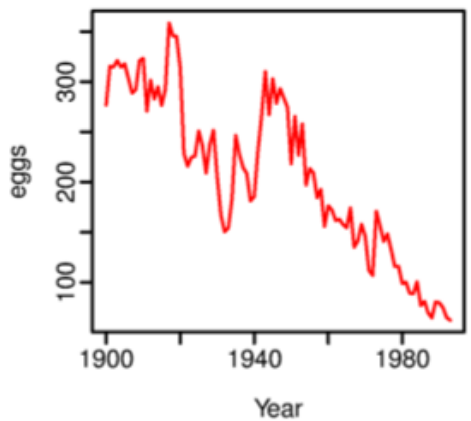
Нестационарность

- сезонность



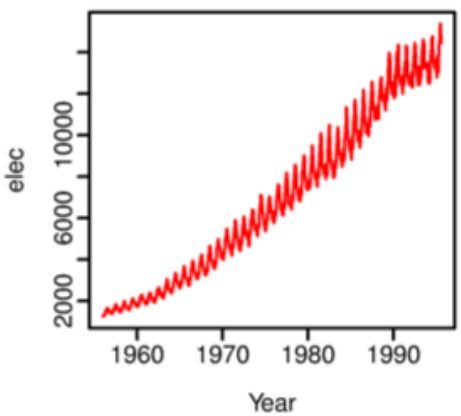
Нестационарность

- тренд



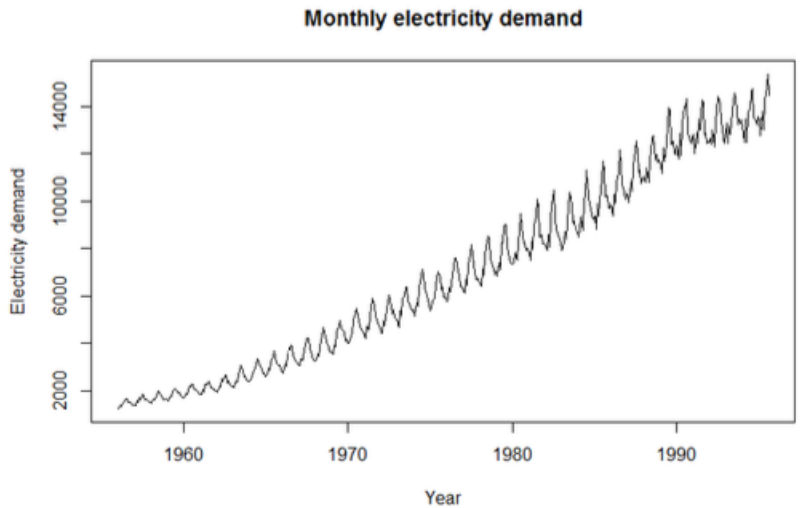
Нестационарность

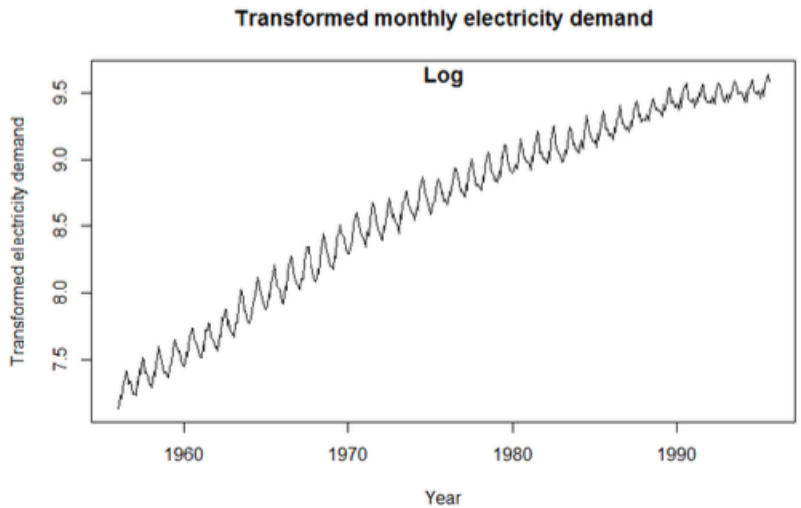
- меняющаяся дисперсия



Стабилизация дисперсии

Логарифмирование как стабилизирующее преобразование:

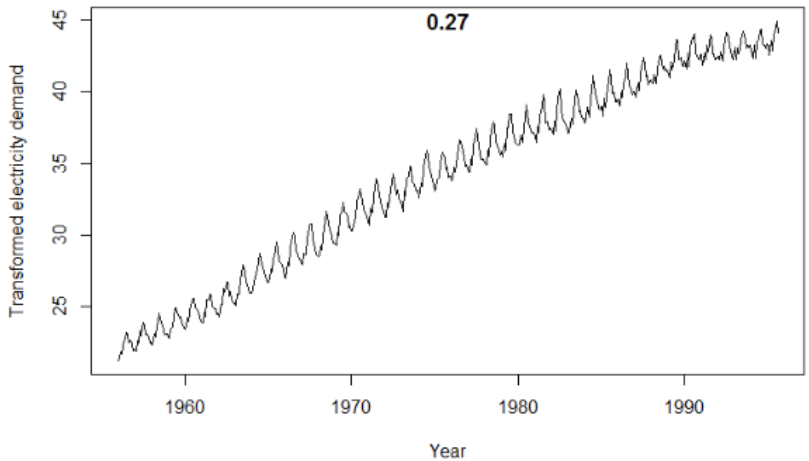


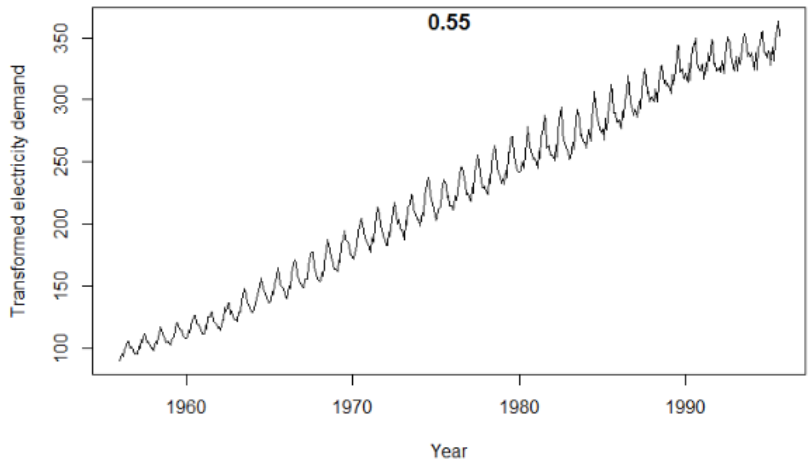


Преобразования Бокса-Кокса

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

Параметр λ выбирается так, чтобы минимизировать дисперсию или максимизировать правдоподобие





Преобразования Бокса-Кокса

$$\hat{y}_t = \begin{cases} exp(\hat{y}'_t), \lambda = 0, \\ (\lambda \hat{y}'_t + 1)^{\frac{1}{\lambda}}, \lambda \neq 0, \end{cases}$$

Если какие-то $y_t < 0$, преобразования Бокса-Кокса невозможны

Дифференцирование ряда

Временной ряд Y^T: $y_1, \ldots, y_T$
Дифференцирование:

$$y_1, \ldots, y_T \to y'_2, \ldots, y'_T$$

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

Зачем?

- стабилизирует значение ряда
- избавляет от сезонности/тренда

Неоднократное дифференцирование

$$y_1, \ldots, y_T \to y'_2, \ldots, y'_T \to y''_3, \ldots, y''_T$$

Сезонное дифференцирование ряда

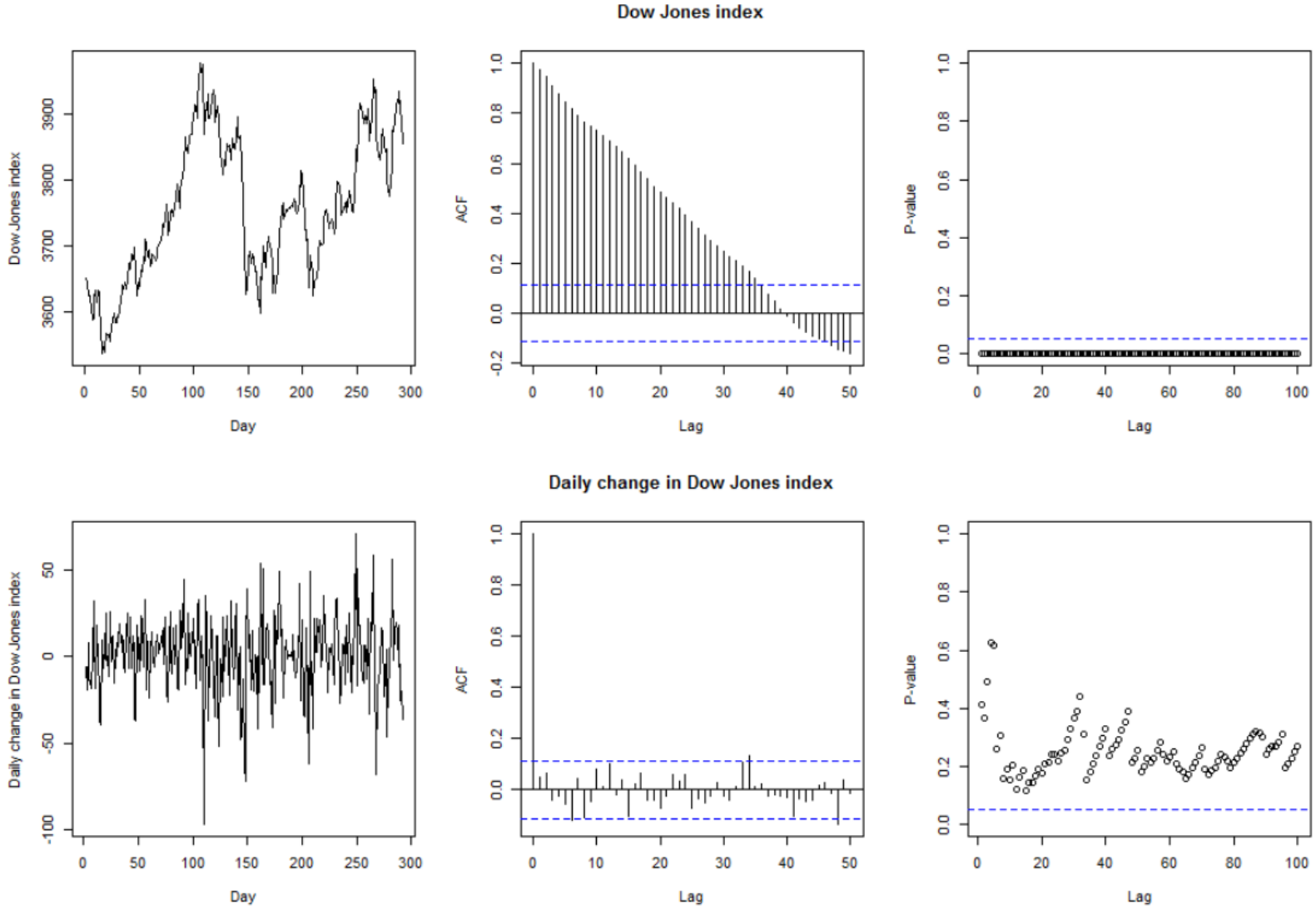
Переход к попарным разностям значений в соседних сезонах

$$y_1, \ldots, y_T \to y'_{s+1}, \ldots, y'_T$$

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

Обычное и сезонное дифференцирование можно выполнять в любом порядке.
Если ряд кажется сезонным, сначала лучше выполнить сезонное дифференцирование. После этого ряд может уже быть стационарным

Дифференцирование ряда



Критерий KPSS:

Для исходного ряда p-value < 0.01

Для ряда первых разностей p-value > 0.1

Критерий KPSS (Kwiatkowski-Philips-Schmidt-Shin)

ряд ошибок прогноза: $\epsilon^T = \epsilon_1, \ldots, \epsilon_T$;

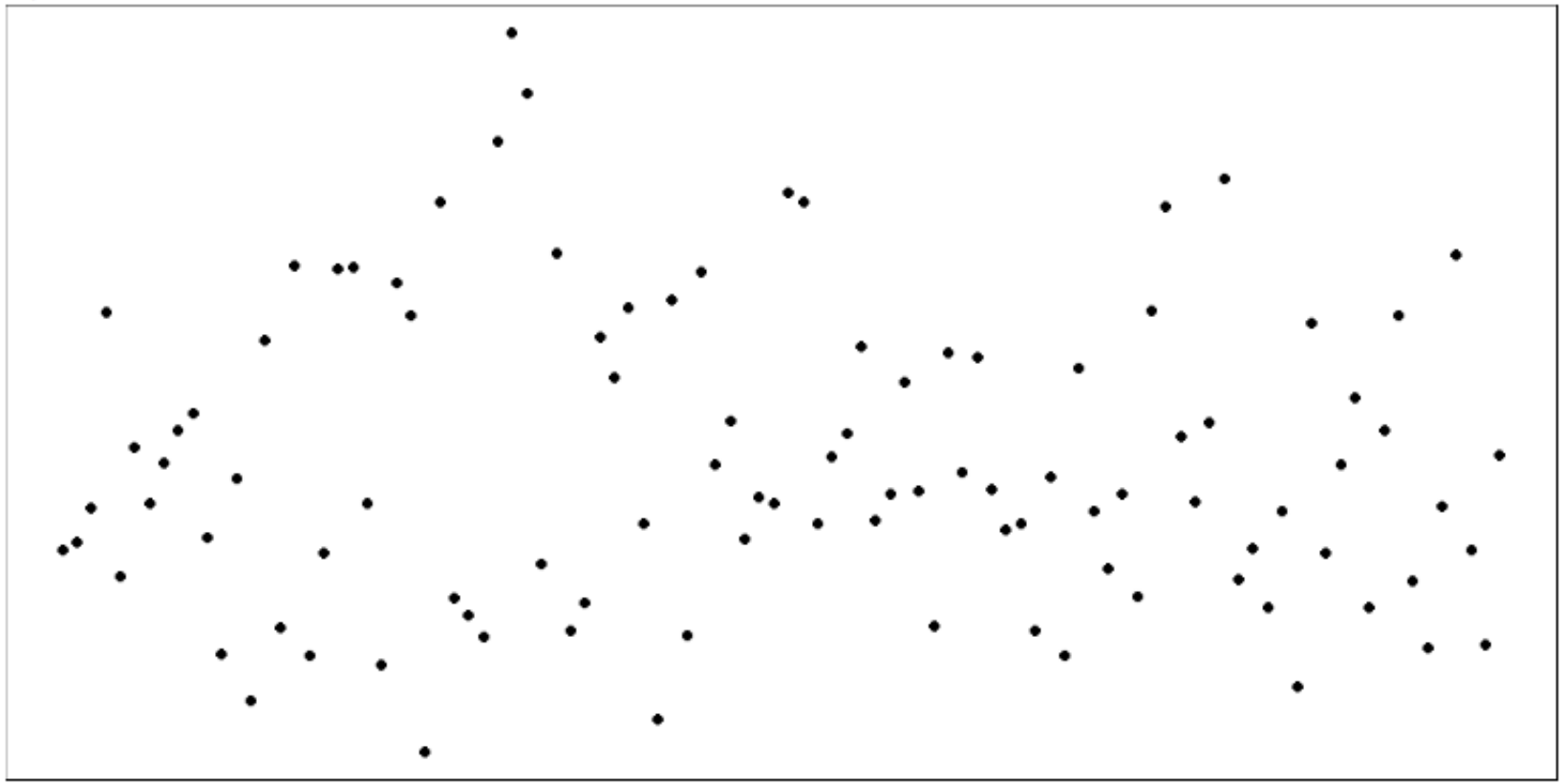
нулевая гипотеза: H_0 : ряд ϵ^T стационарен;

альтернативная гипотеза: H_1 : ряд ϵ^T описывается моделью вида $\epsilon_t = \alpha\epsilon_{t-1}$;

статистика: $KPSS(\epsilon^T) = \frac{1}{T^2\lambda^2}\sum_{i=1}^{T}(\sum_{t=1}^{i}\epsilon_t)^2$;

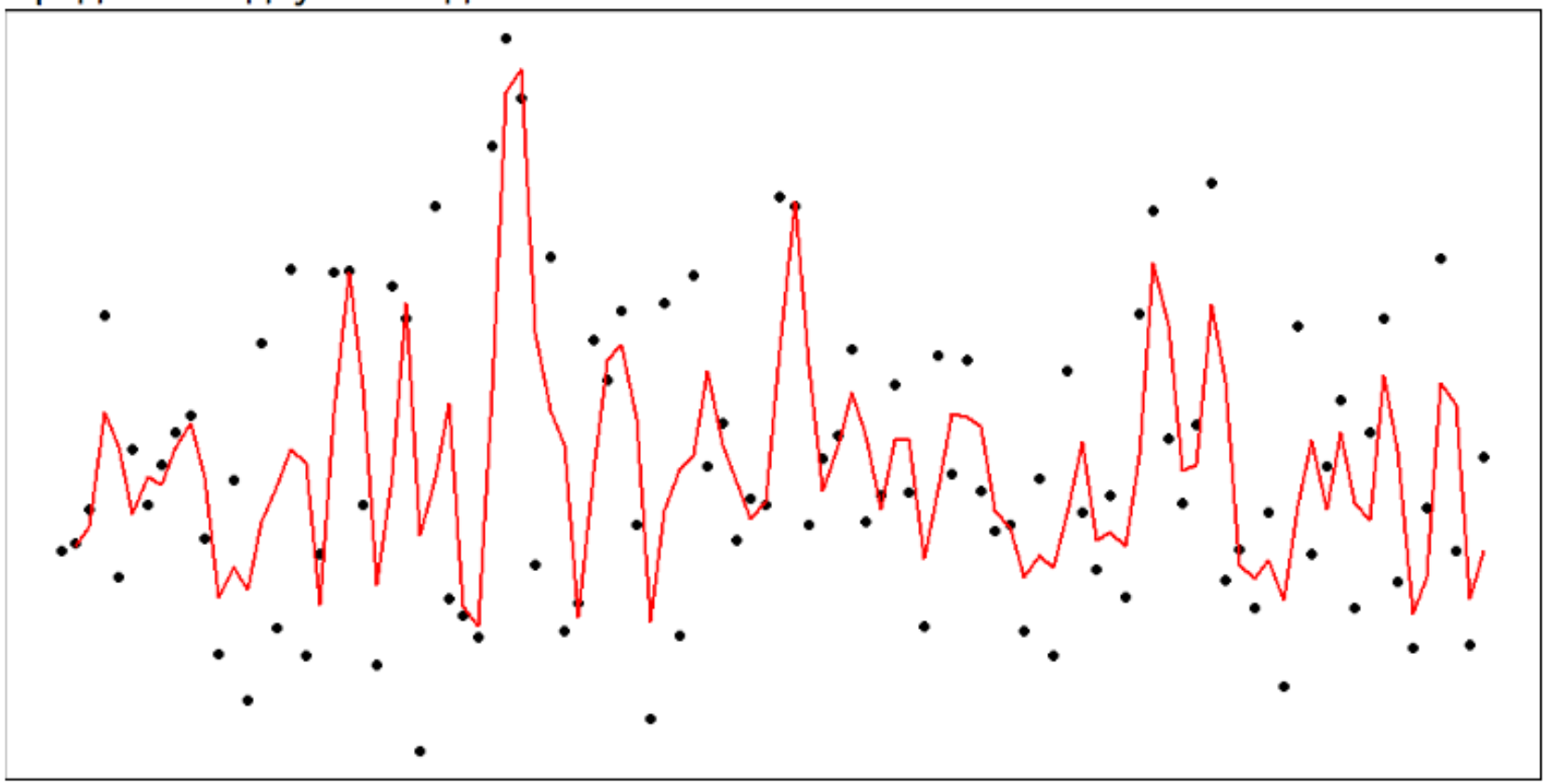
Скользящее среднее

Пусть у нас есть независимый одинаково распределённый во времени шум ε_t:



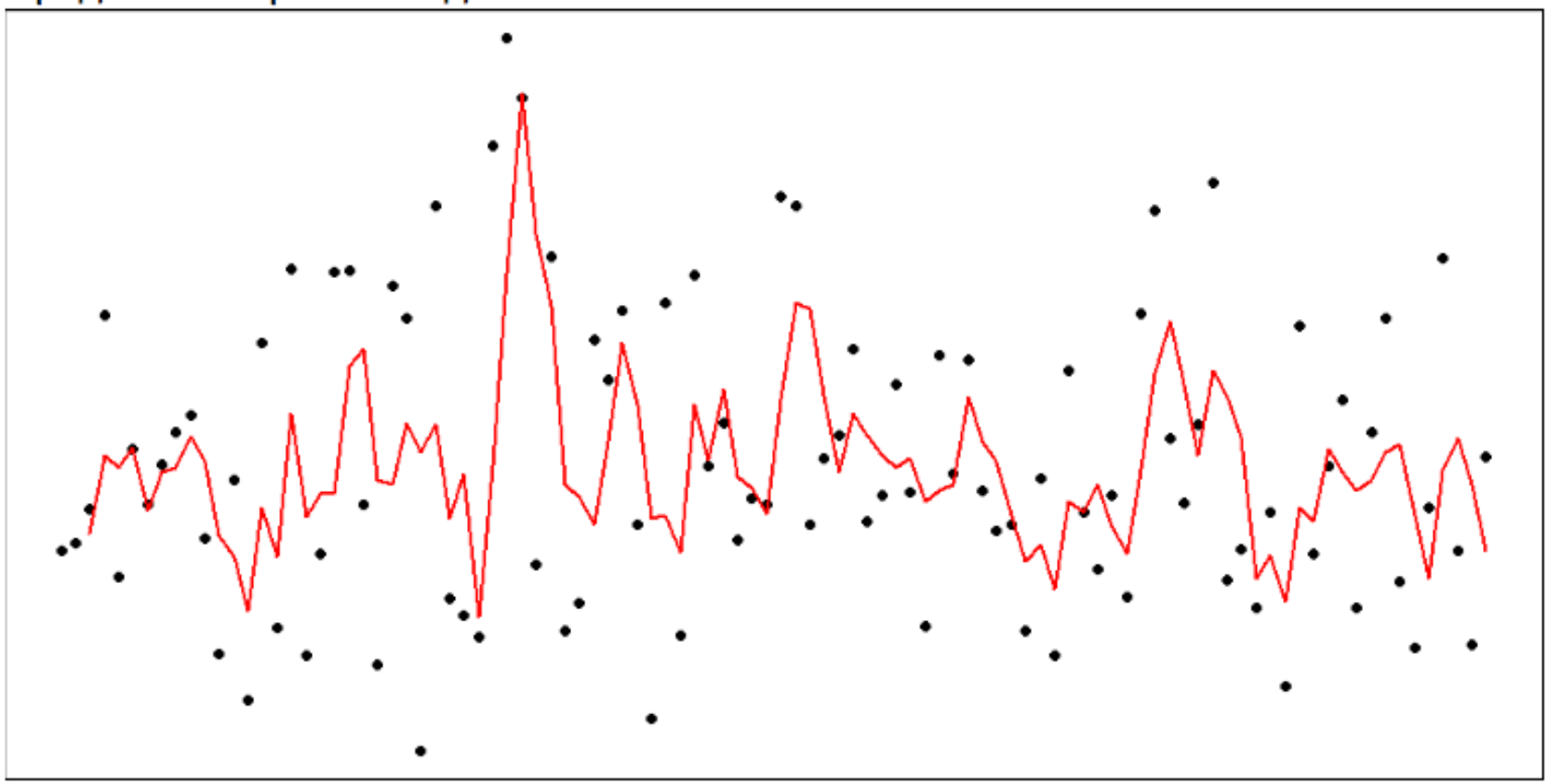
Скользящее среднее

Среднее по двум соседним точкам:



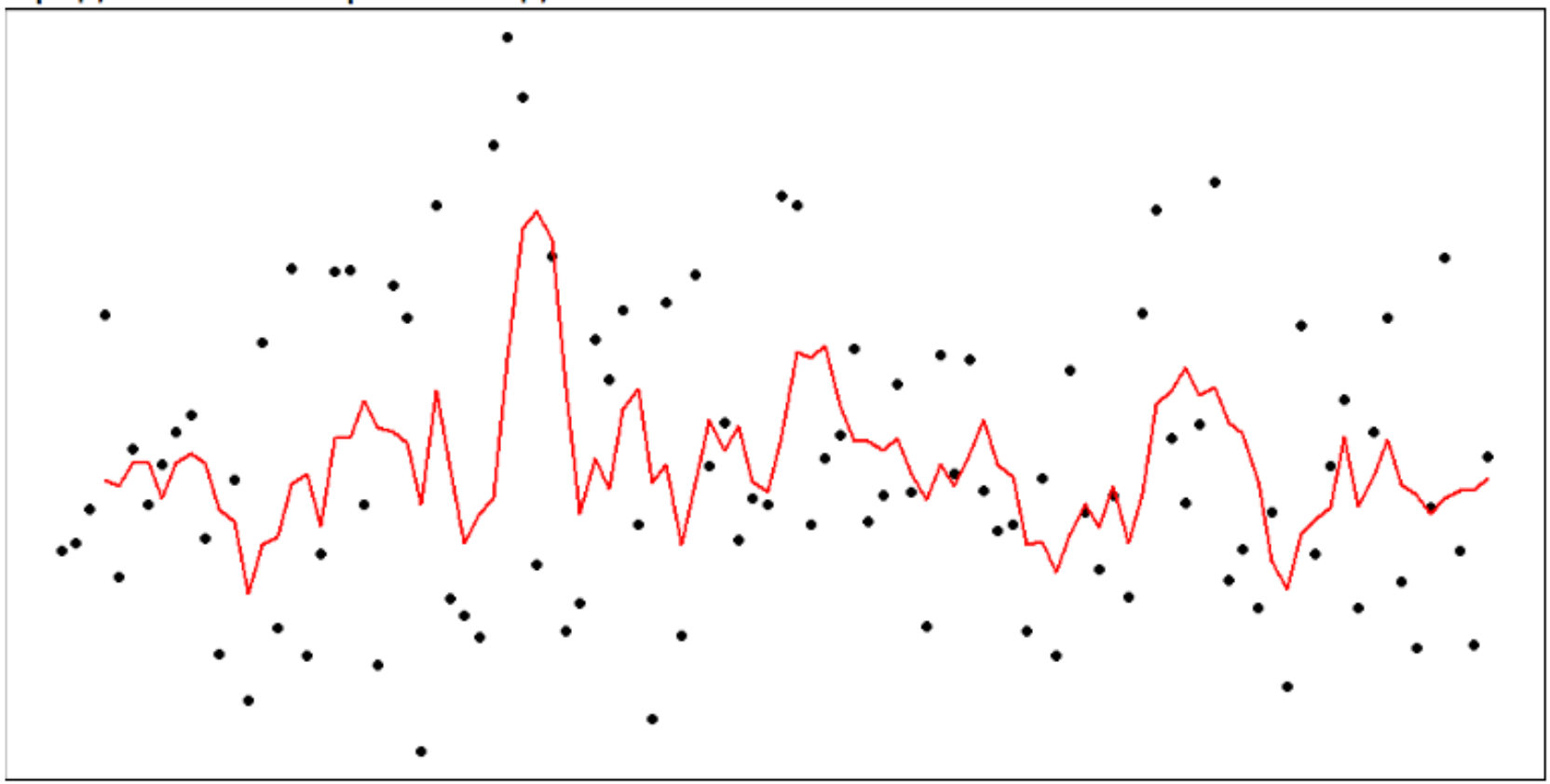
Скользящее среднее

Среднее по трём соседним точкам:



Скользящее среднее

Среднее по четырём соседним точкам:



AR(p), MA(q)

$$AR(p) : y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \epsilon_t,$$

$$MA(q) : y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q},$$

где y_t– стационарный ряд с нулевым средним, ϕ_t, θ_t – константы, ϵ_t – гауссов белый шум с нулевым средним и постоянной дисперсией σ_ϵ^2

Если среднее равно μ :

$$AR(p) : y_t = \alpha + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \epsilon_t,$$

$$\alpha = \mu(1 - \phi_1 - \ldots - \phi_p)$$

$$MA(q) : y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q},$$

ARMA (Autoregressive moving average)

$ARMA(p,q) : y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q}$, где y_t– стационарный ряд с нулевым средним, ϕ_t, θ_t – константы, ϵ_t – гауссов белый шум с нулевым средним и постоянной дисперсией σ_ϵ^2

Если среднее равно μ:

$$y_t = \alpha + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q},$$

$\alpha = \mu(1 - \phi_1 - \ldots - \phi_p)$
Другой способ записи:

$$\phi(B) y_t = \theta(B) \epsilon_t,$$

где B – разностный оператор $(By_t = y_{t-1})$

$$\phi(B) y_t = (1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p) y_t =$$

$$= \epsilon_t (1 + \theta_1 B + \ldots + \theta_q B^q) = \theta(B) \epsilon_t$$

Теорема Вольда

Любой стационарный ряд может быть аппроксимирован моделью $ARMA(p, q)$ с любой точностью

ARIMA (Autoregressive integrated moving average)

Ряд описывается моделью $ARIMA(p, d, q)$, если ряд его разностей

$$\nabla^d y_t = (1 - B)^d y_t$$

описывается моделью $ARMA(p, q)$

$$\phi(B)\nabla^d y_t = \theta(B)\epsilon_t$$

Подбор коэффициентов

- α, θ, ϵ
  – если все остальные параметры фиксированы, коэффициенты регрессии подбираются методом наименьших квадратов
  – чтобы найти коэффициенты θ, шумовая компонента предварительно оценивается с помощью остатков авторегрессии
  – если шум белый, то МНК даёт оценки максимального правдоподобия
- d
  – порядки дифференцирования подбираются так, чтобы ряд стал стационарным
  – чем меньше раз продифференцируем, тем меньше будет дисперсия итогового прогноза
- p, q
  – начальные приближения можно выбрать с помощью автокорреляций

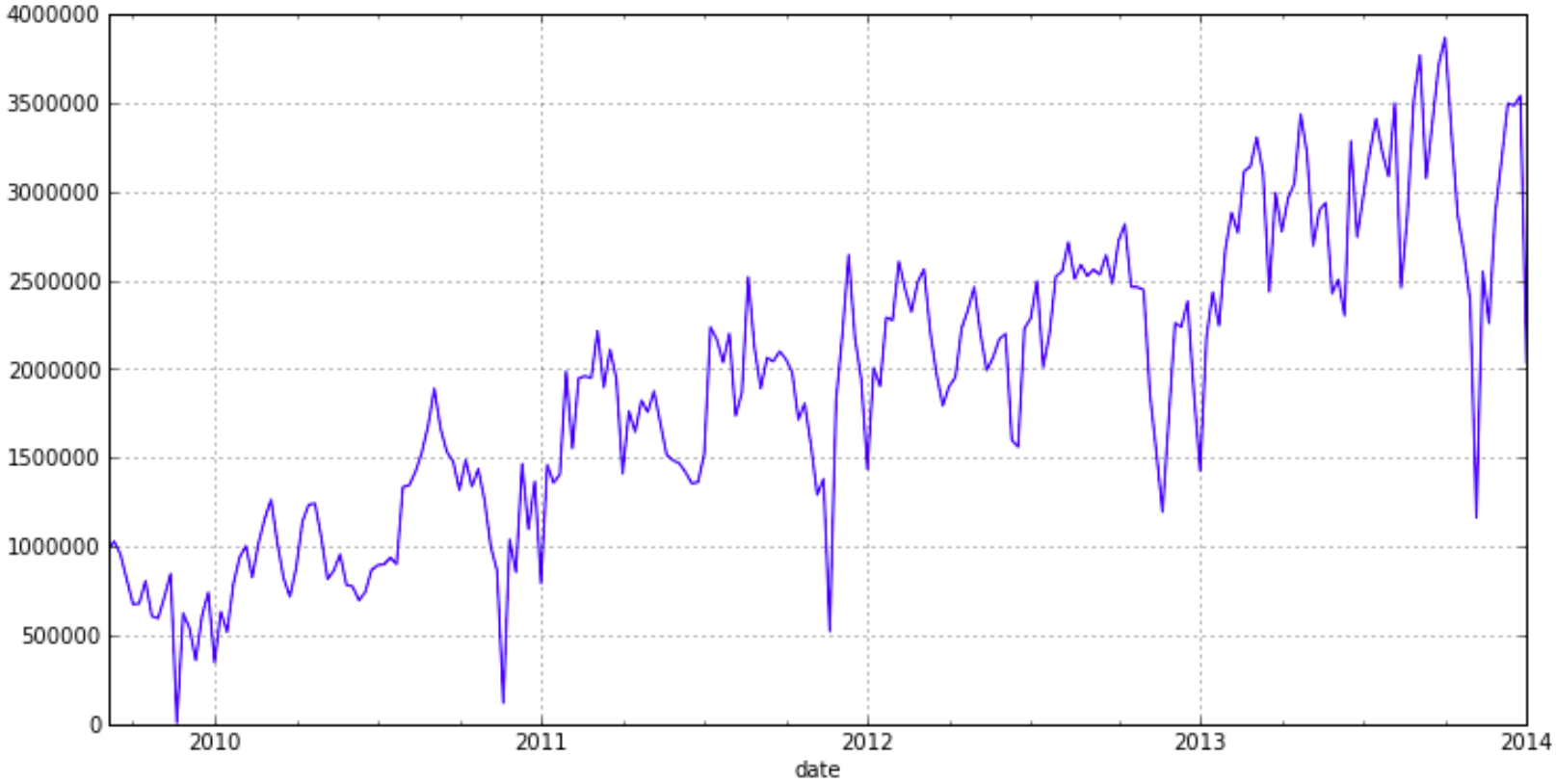
Частичная автокорреляционная функция (PACF)

Частичная автокорреляция: стационарного ряда y_t – автокорреляция остатков авторегрессии предыдущего порядка

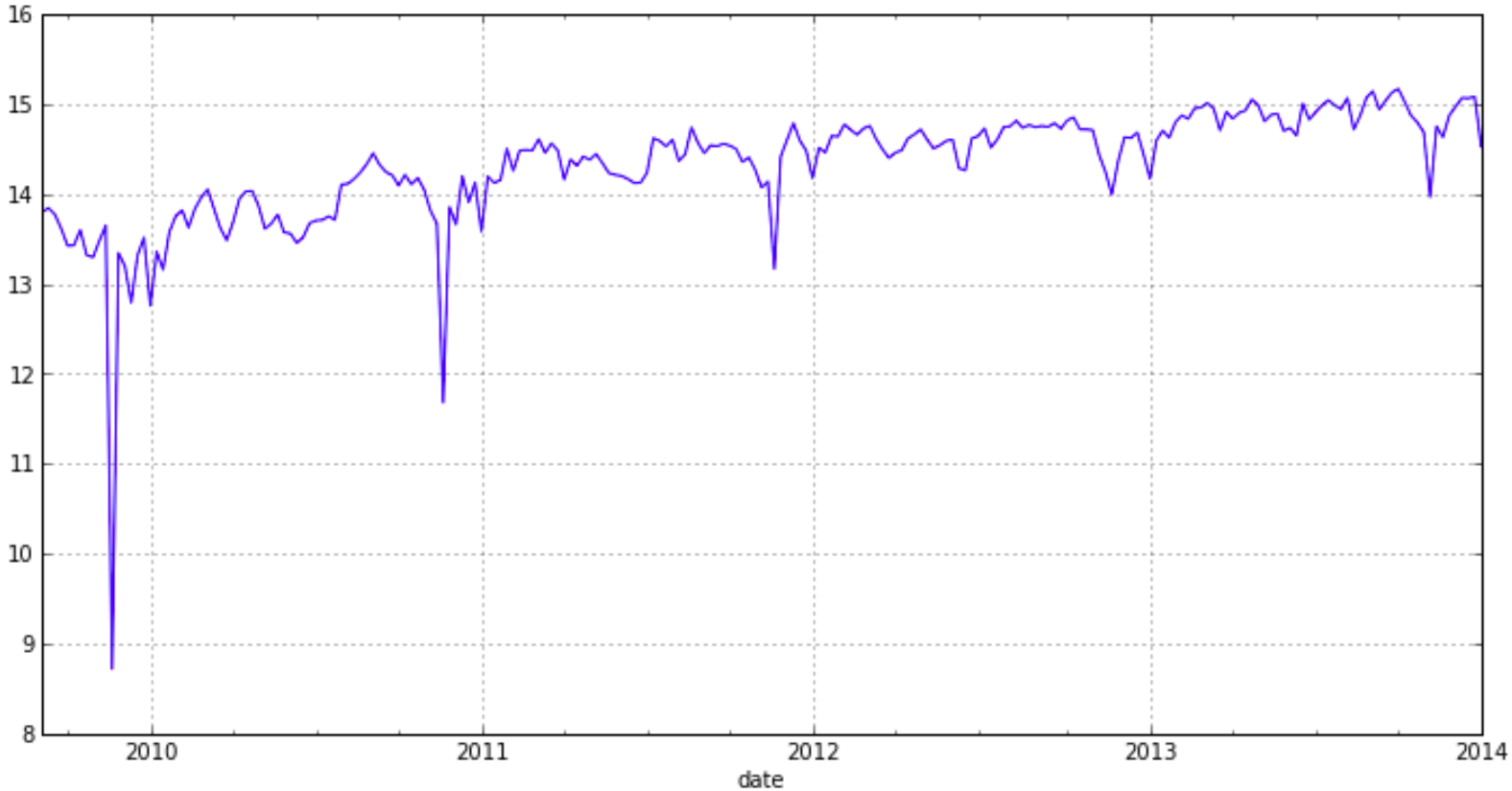
$$\phi_h h = \begin{cases} r(y_{t+1}, y_t), h = 1, \\ r(\epsilon_{t+h}, \epsilon_t), h > 1, \end{cases}$$

p – номер последнего лага при котором PACF значима
q – номер последнего лага при котором ACF значима

Построение ARIMA

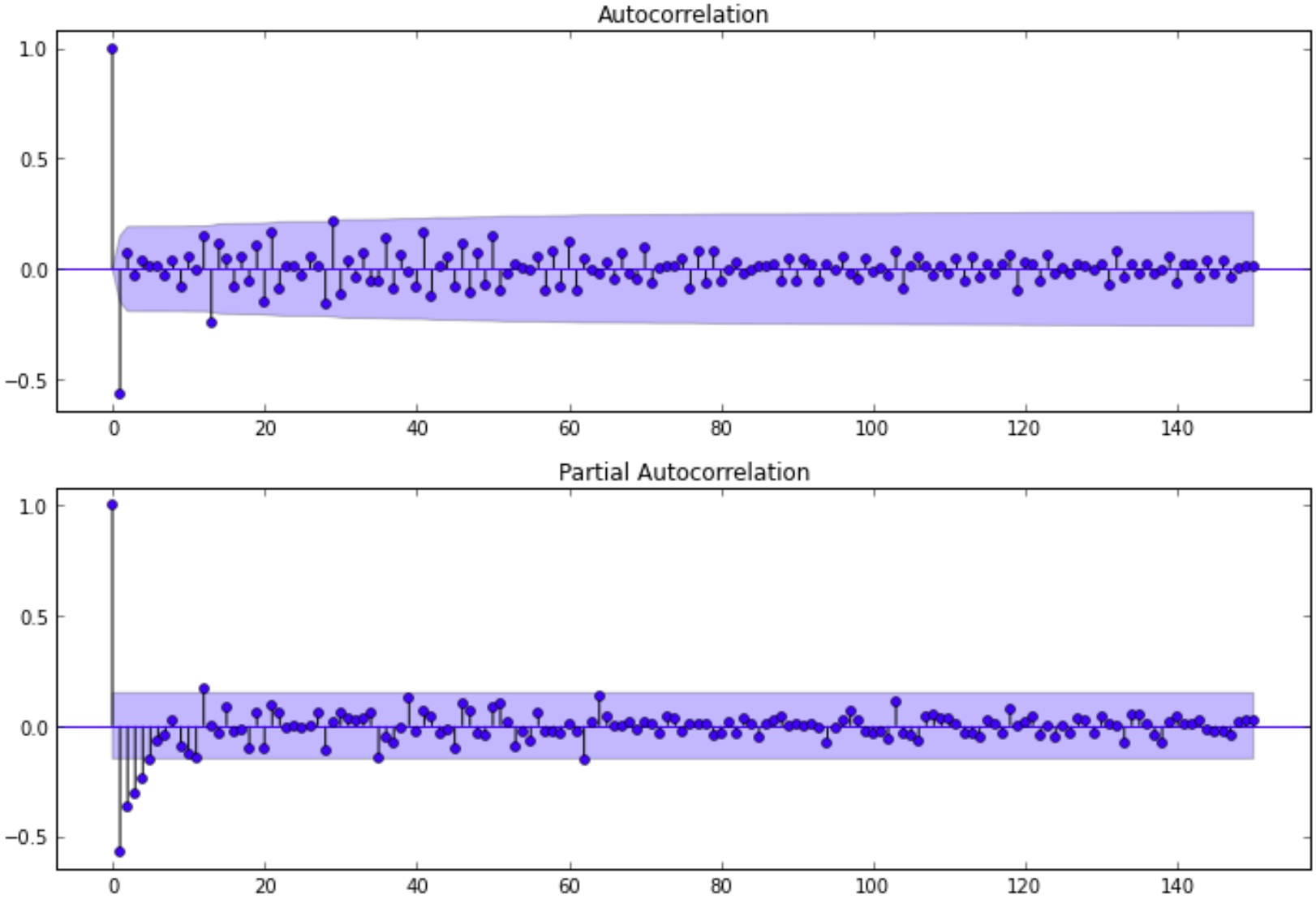


Построение ARIMA

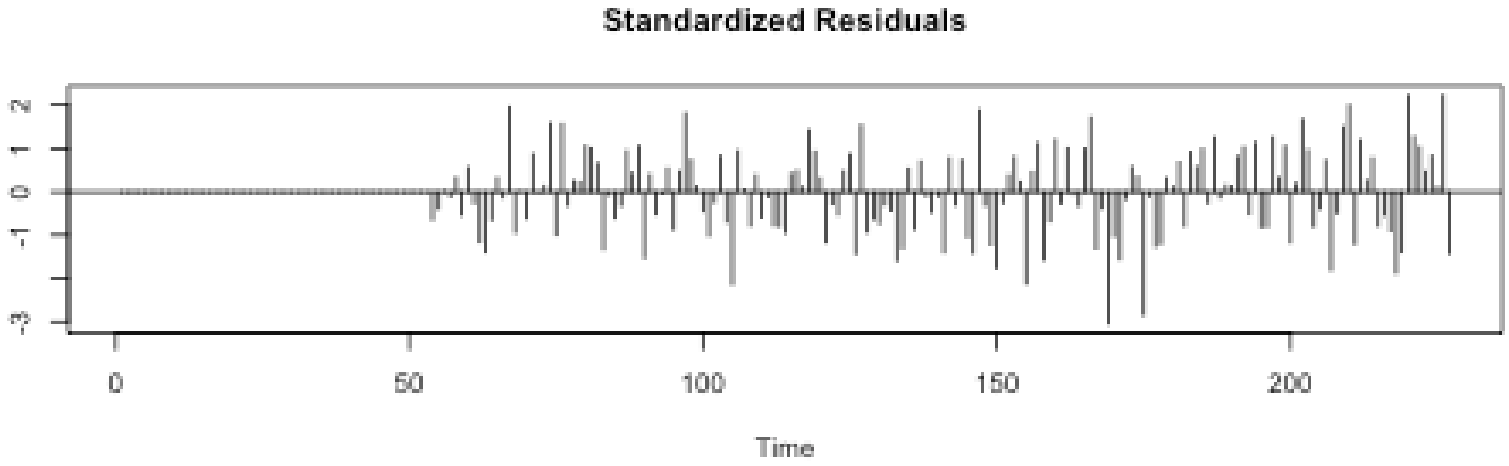


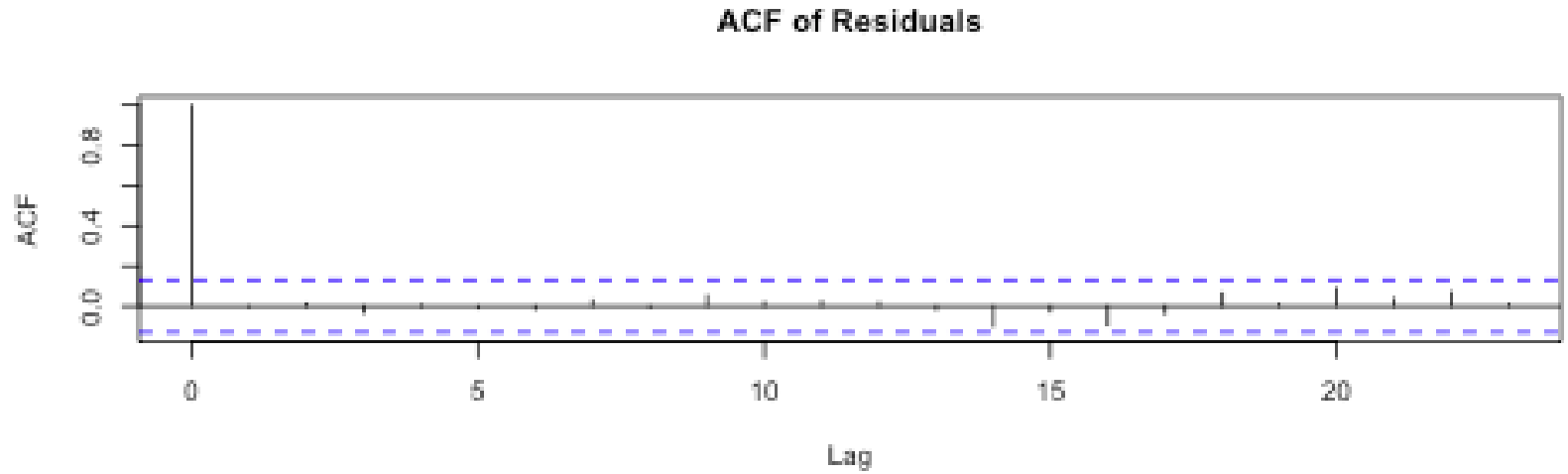
Построение ARIMA

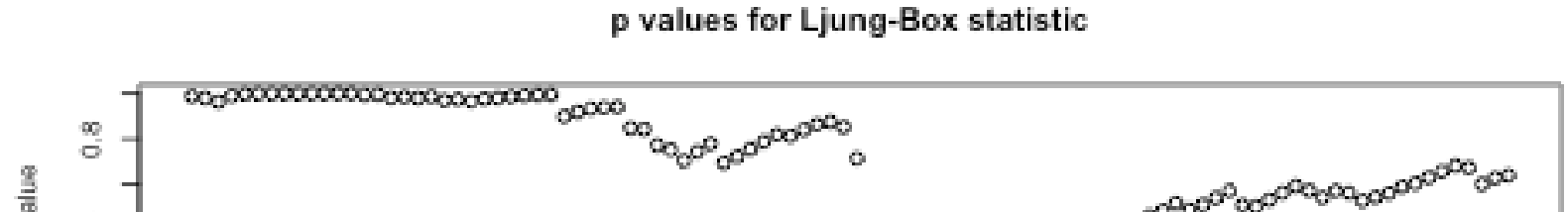
Построение ARIMA



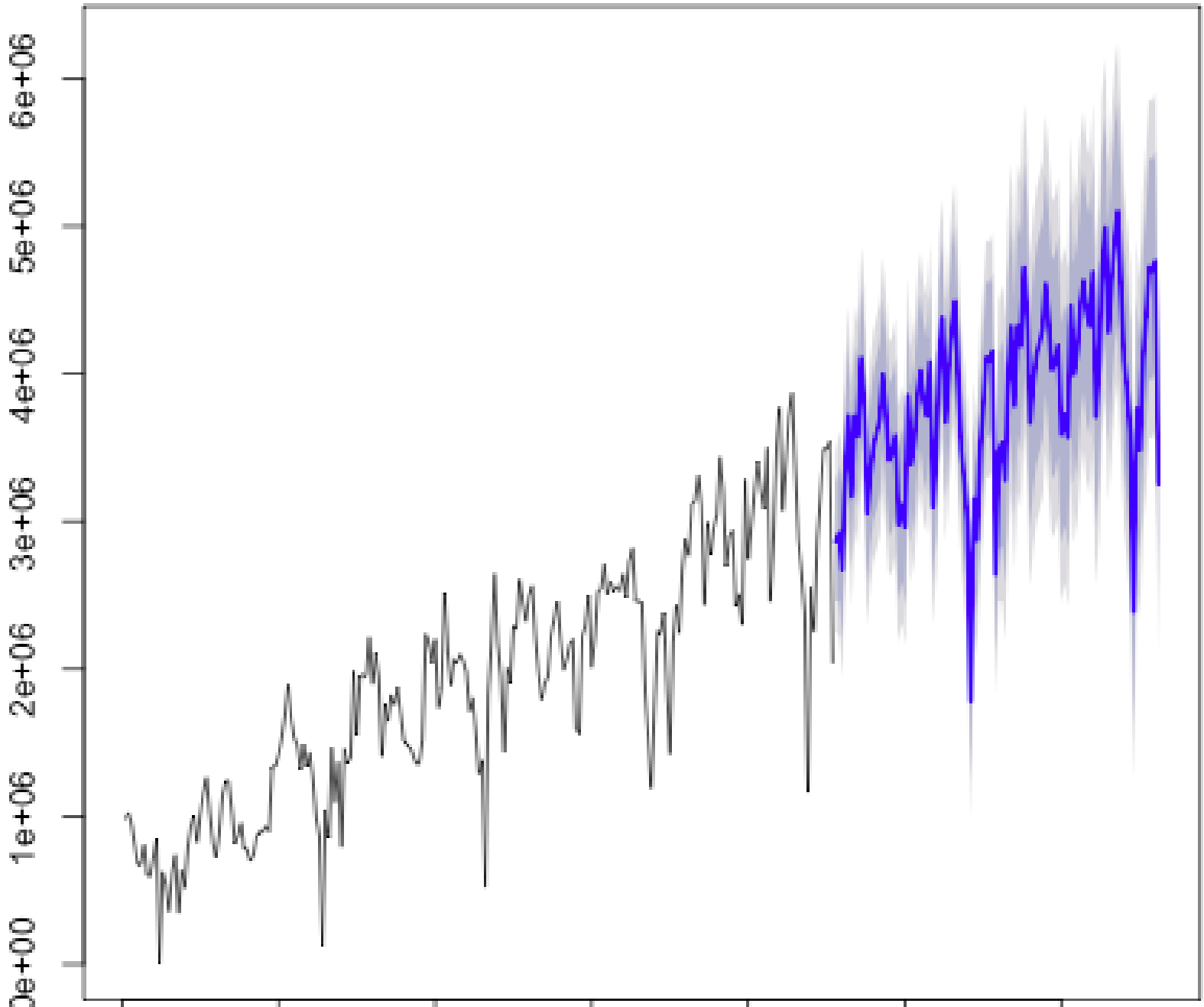
Построение ARIMA







Построение ARIMA(4,1,13)



Дискретное преобразование Фурье

DFT, Discrete Fourier Transform

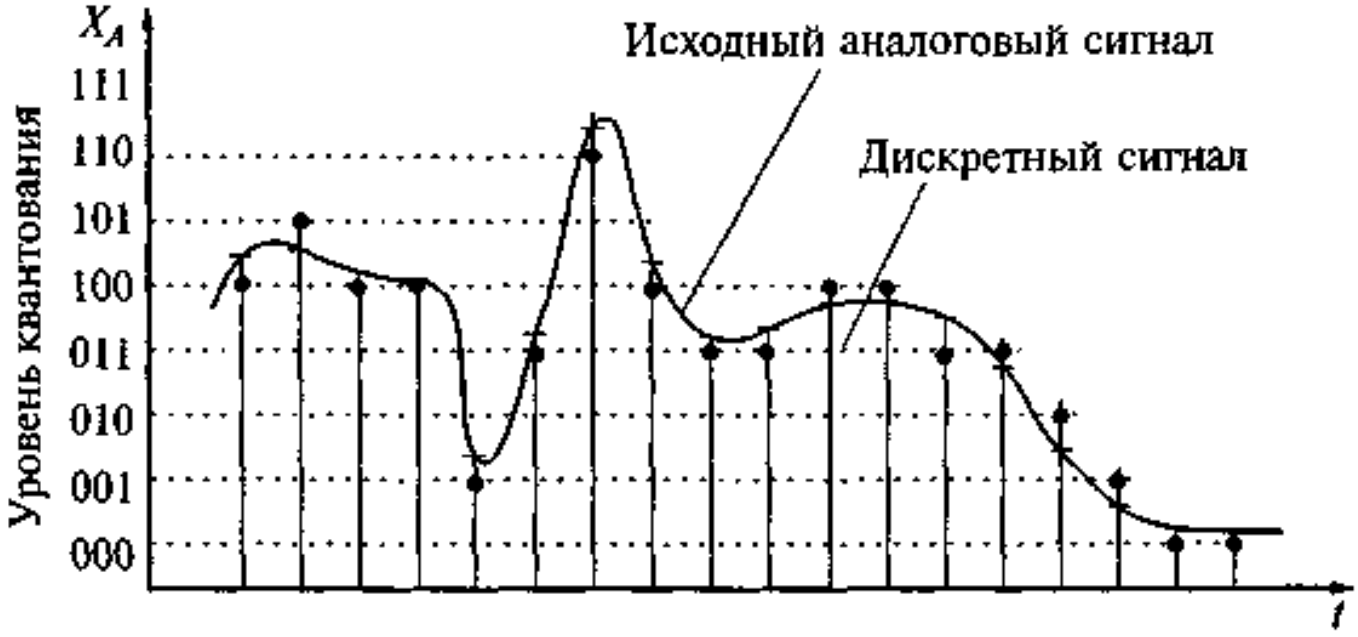
$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} =$$

$\sum_{n=0}^{N-1} x_n \cdot [cos(\frac{2\pi kn}{N}) - i \cdot sin(\frac{2\pi kn}{N})]$,

$$(k = 0, \ldots, N-1)$$

Ограничения:
При анализе с помощью преобразования Фурье, мы исходим из предположения, что он периодический на текущем временном интервале и состоит из элементарных синусоид.

Дискретное преобразование Фурье



- Получаем комплексные числа, содержащий информацию об амплитудном и фазовом спектрах анализируемого кадра. Причём спектры также являются дискретными с шагом (частота дискретизации)/(N отсчётов).
- Разложение в базис в некотором пространстве периодических функций. Коофициенты - веса этих функций.
- Выполняется за $O(Nlog(N))$ - БПФ.

Ссылки

[1]. http://www.machinelearning.ru/wiki/images/3/31/Psad$_t s_a rima_2$017.*pdf*
[2]. https://habr.com/post/210530/
[3]. https://ru.wikipedia.org