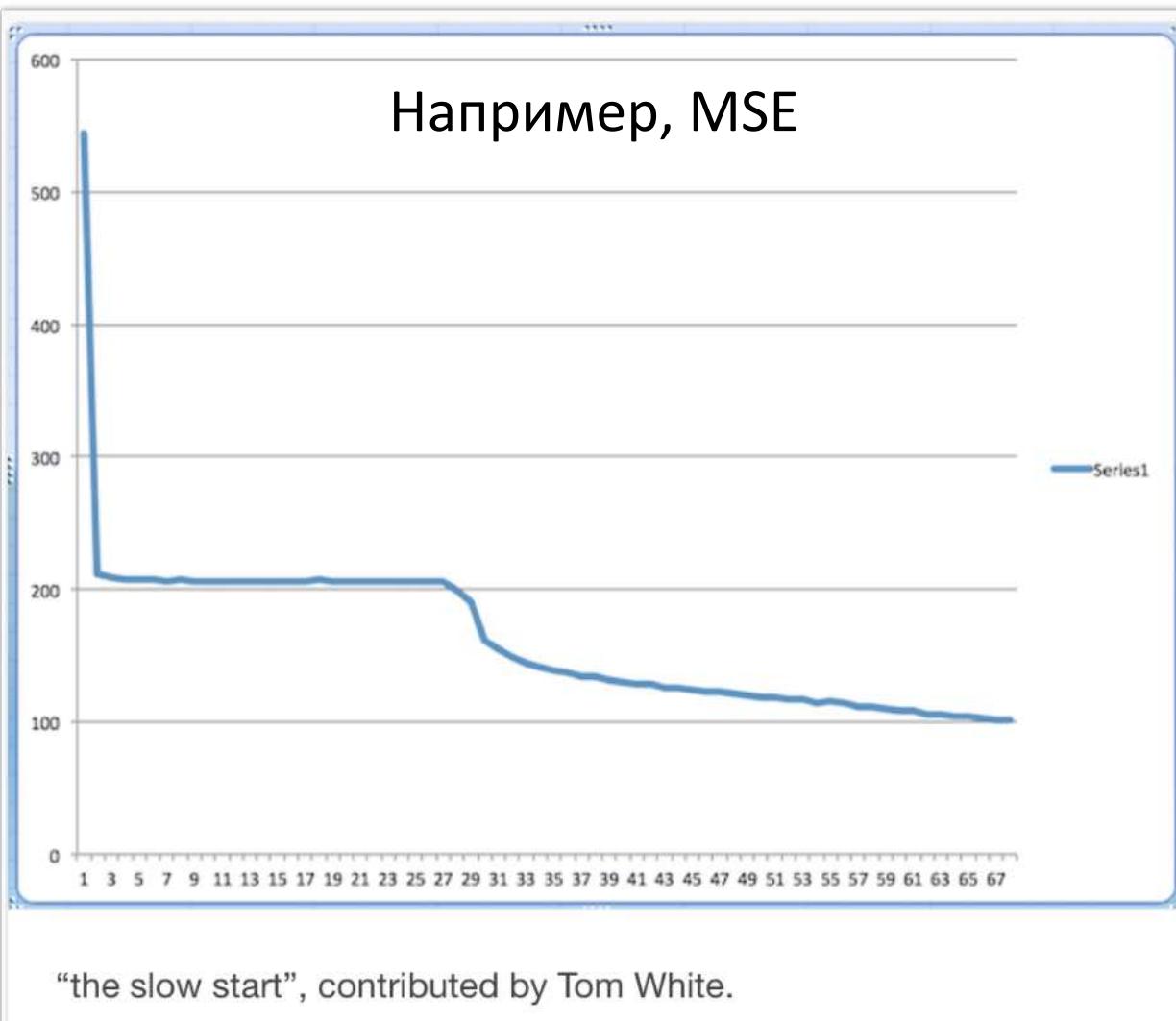


MML minor #10

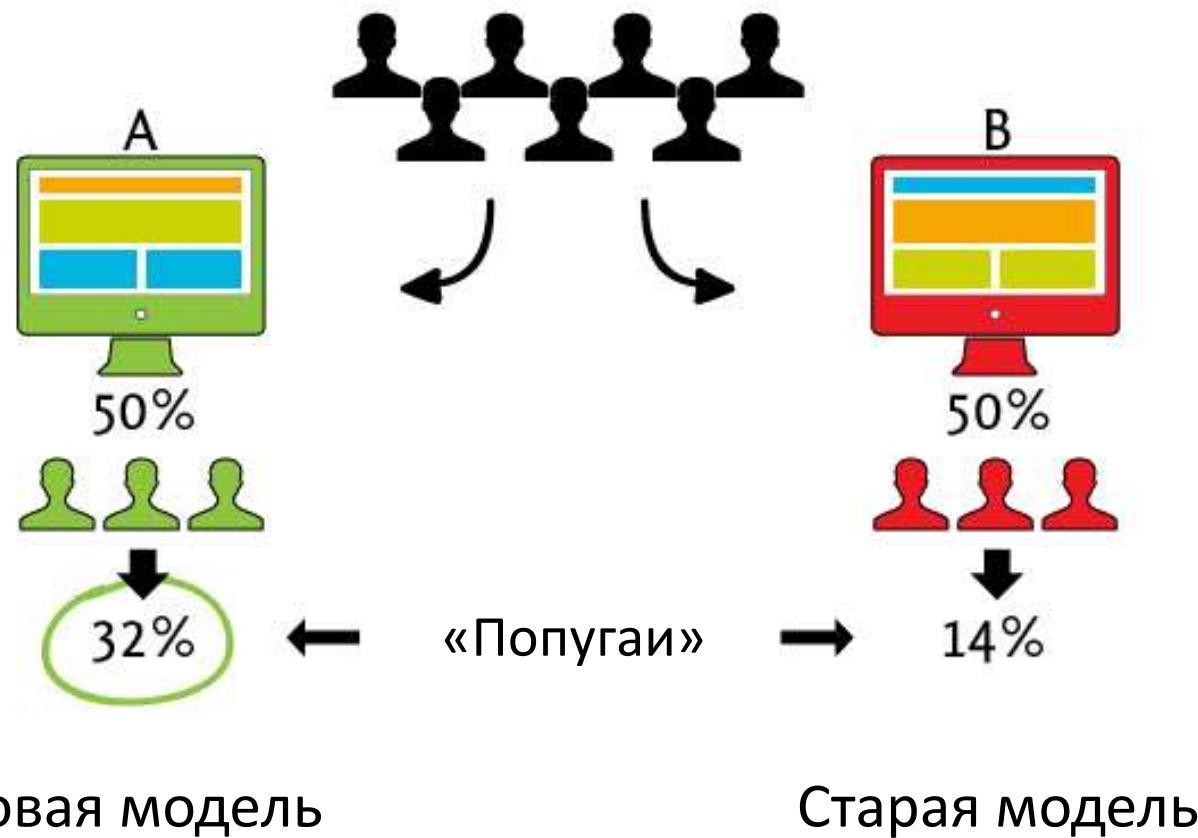
Не RNN единственным: ARIMA

Минимизация ошибки – это только начало!



А/В-тестирование

Делим пользователей случайно



Попугай:

1. Конверсия
2. Деньги
3. Время в сервисе

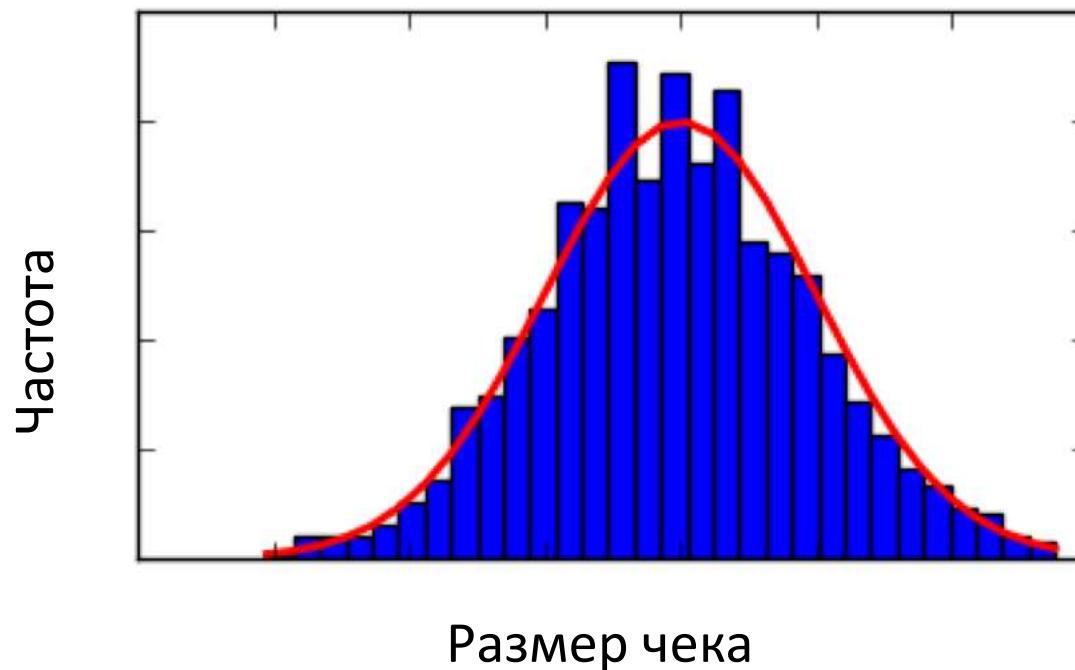


Статистическая проверка гипотез

- **Клиенты из группы А:** средний чек 77 р.
- **Клиенты из группы Б:** средний чек 78 р.
- Значимо ли различие или это шум?

Проверка гипотез

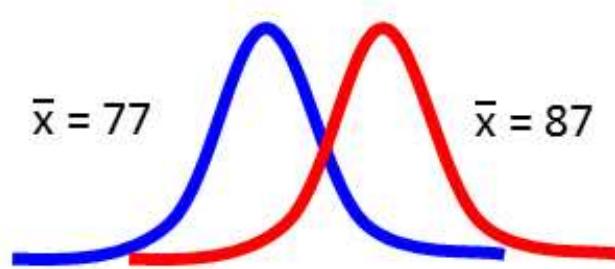
- Клиенты из группы А: средний чек 77 р.
- Клиенты из группы Б: средний чек 78 р.
- Значимо ли различие или это шум?



Построим для
каждой группы!

Проверка гипотез

- Клиенты из группы А: средний чек 77 р.
- Клиенты из группы Б: средний чек 78 р.
- Значимо ли различие или это шум?



Меньше доверия

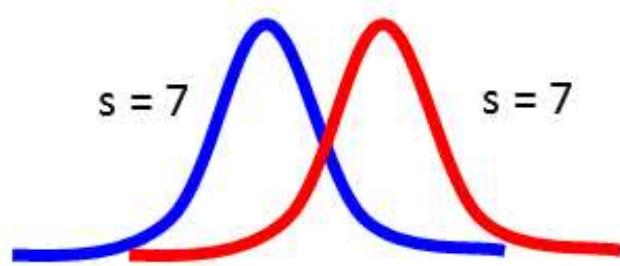


Больше доверия

Средние \bar{x} далеко

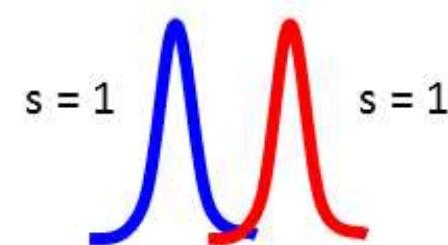
Проверка гипотез

- Клиенты из группы А: средний чек 77 р.
- Клиенты из группы Б: средний чек 78 р.
- Значимо ли различие или это шум?



Меньше доверия

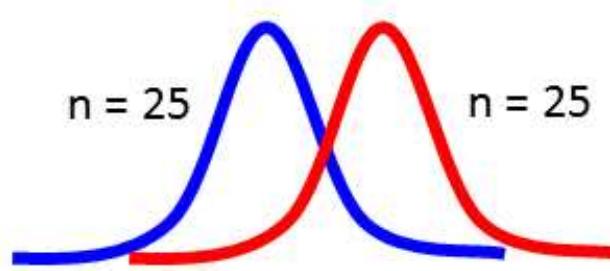
Меньше разброс σ



Больше доверия

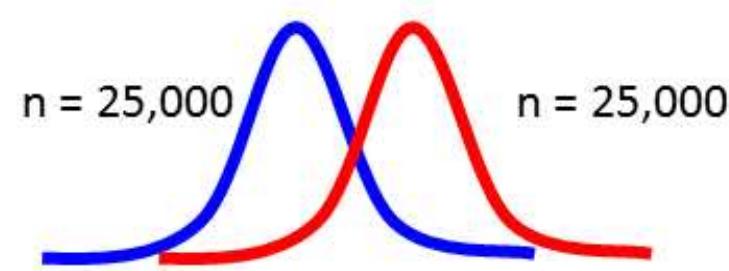
Проверка гипотез

- Клиенты из группы А: средний чек 77 р.
- Клиенты из группы Б: средний чек 78 р.
- Значимо ли различие или это шум?



Меньше доверия

Больше наблюдений n



Больше доверия

Взболтать, но не смешивать

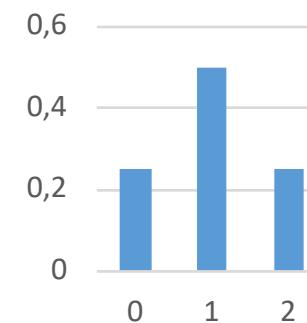
- Слепой тест: предложим Джеймсу Бонду n раз пару напитков и спросим, какой из двух он предпочитает.
- Выборка: n чисел (предпочёл взболтанный — 1, смешанный — 0)
- Нулевая гипотеза H_0 : Джеймс Бонд выбирает наугад (все исходы равновероятны)
- Статистика: число единиц в выборке

Исходы для $n=2$:	0	0
	0	1
	1	0
	1	1

Статистика:

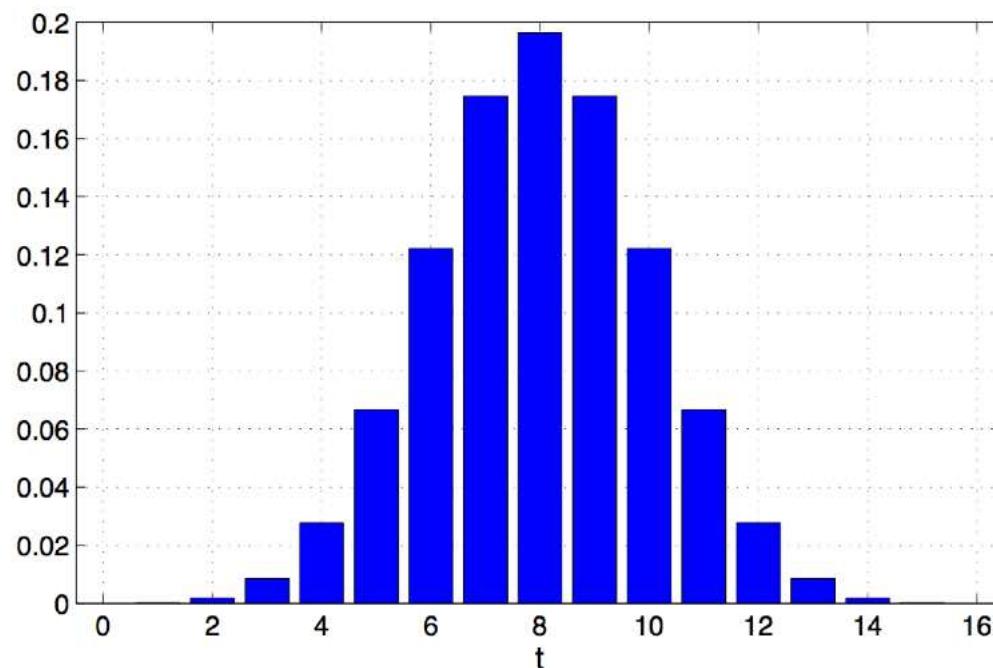
0
1
2

Распределение
статистики при H_0



Взболтать, но не смешивать

- Пусть $n = 16$ — тогда существует 65535 различных бинарных векторов
- Распределение количества единиц в векторе:

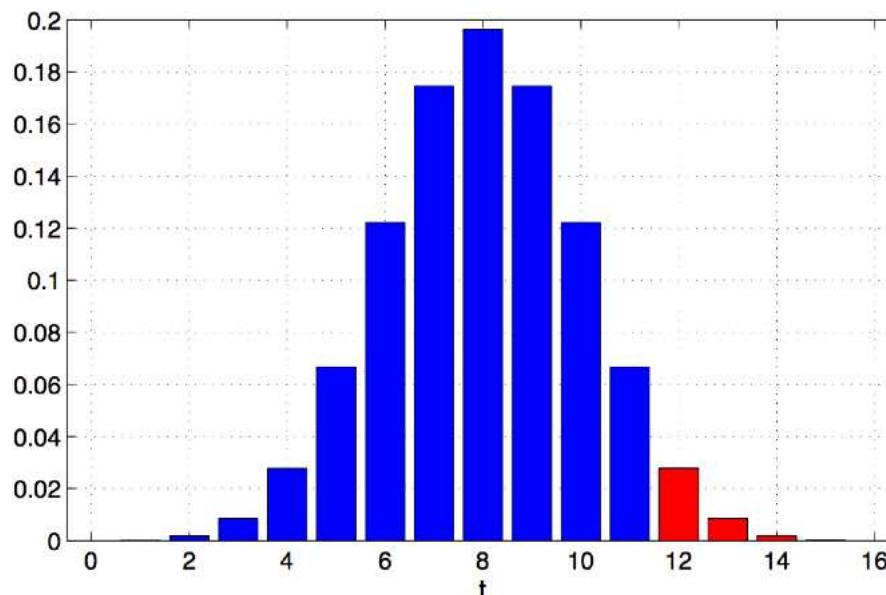


Распределение статистики при H_0

Взболтать, но не смешивать

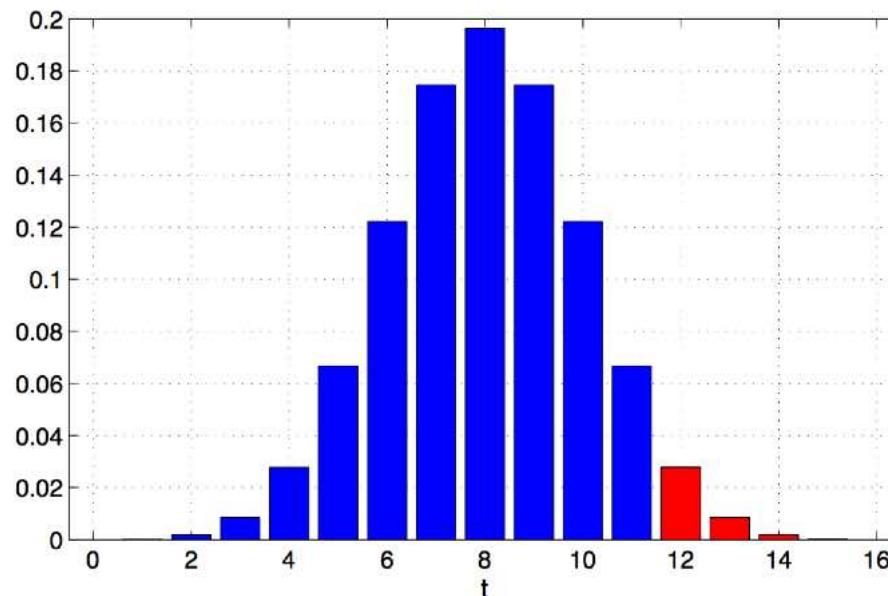
- Эксперимент: Джеймс Бонд выбрал взболтанный мартини в 12 из 16 раз!
- Вероятность того, что он выберет взболтанный мартини 12 раз или больше при условии, что выбирает наугад: $\frac{2512}{65536} \approx 0.0384 \leftarrow p\text{-value}$
- Отклоняем гипотезу о том, что Джеймс Бонд не разбирается в мартини ☺
(если $p\text{-value} < 0.05$)

Что это значит?



Взболтать, но не смешивать

- Эксперимент: Джеймс Бонд выбрал взболтанный мартини в 12 из 16 раз!
- Вероятность того, что он выберет взболтанный мартини 12 раз или больше при условии, что выбирает наугад: $\frac{2512}{65536} \approx 0.0384 \leftarrow \text{p-value}$
- Отклоняем гипотезу о том, что Джеймс Бонд не разбирается в мартини 😊
(если p-value < 0.05)



В 5% случаев отвергнем
верную гипотезу H_0

Пример: рассылка писем

	A (старая)	B (новая)
Кликов из писем	100	110
Отправлено писем	1000	1000
Конверсия	$100 / 1000 = 0.1$	$110 / 1000 = 0.11$

Значим ли этот результат?



Применяем Approximate Z-test (H_0 : пропорции в выборках равны)



Результат не значим (вероятно про наблюдать при верной H_0)



Собираем больше данных или останавливаем

Пример: рассылка писем

	A (старая)	B (новая)
Кликов из писем	1000	1100
Отправлено писем	10000	10000
Конверсия	$1000 / 10000 = 0.1$	$1100 / 10000 = 0.11$

Значим ли этот результат?



Применяем Approximate Z-test (H_0 : пропорции в выборках равны)



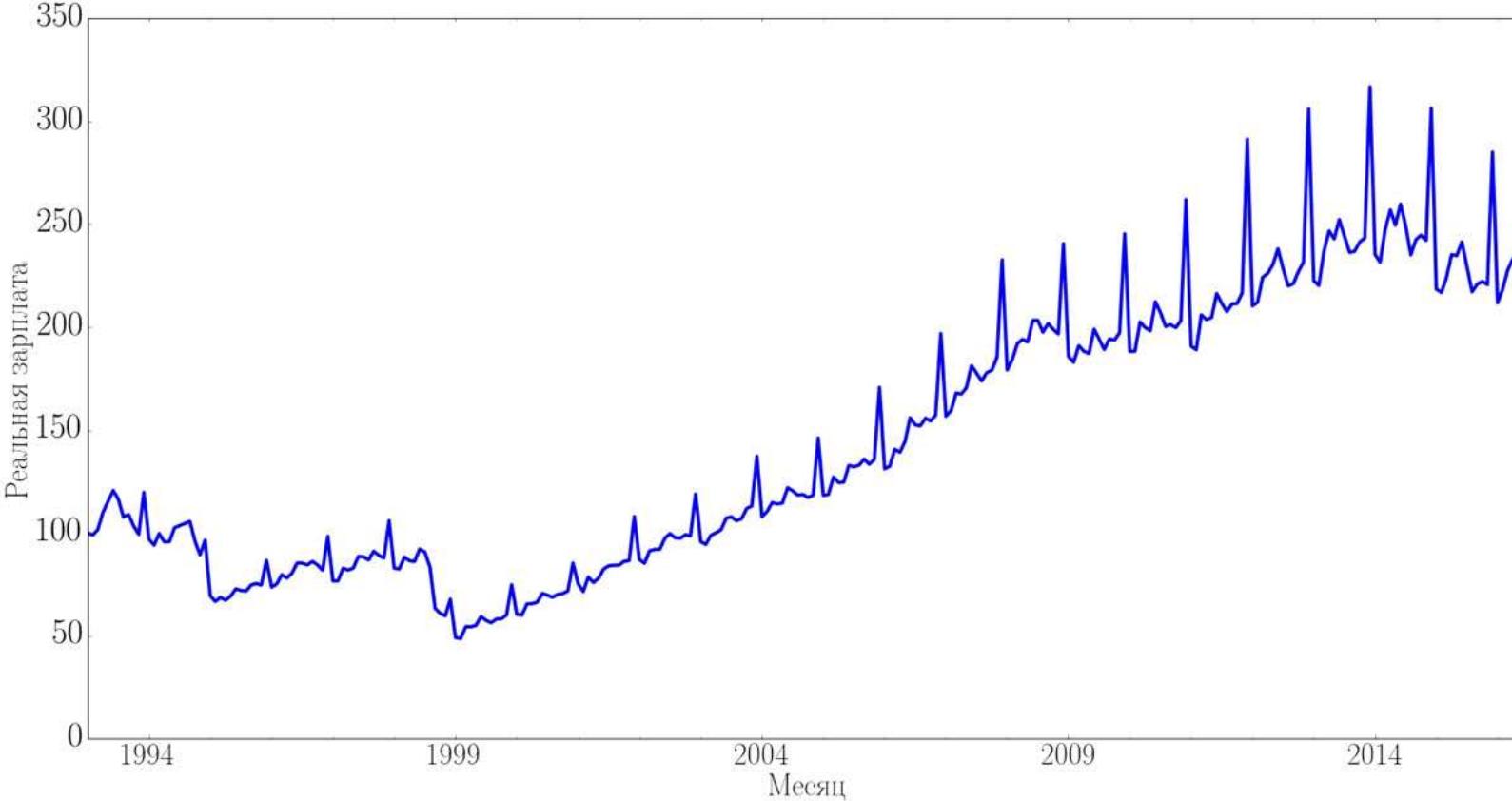
Результат значим (маловероятно про наблюдать при верной H_0)



Останавливаем эксперимент, победа!

Прогнозирование временного ряда

Временной ряд: $y_1, \dots, y_T, \dots, y_t \in \mathbb{R}$, —признак, измеренный через постоянные временные интервалы.



Задача прогнозирования — найти функцию f_T :

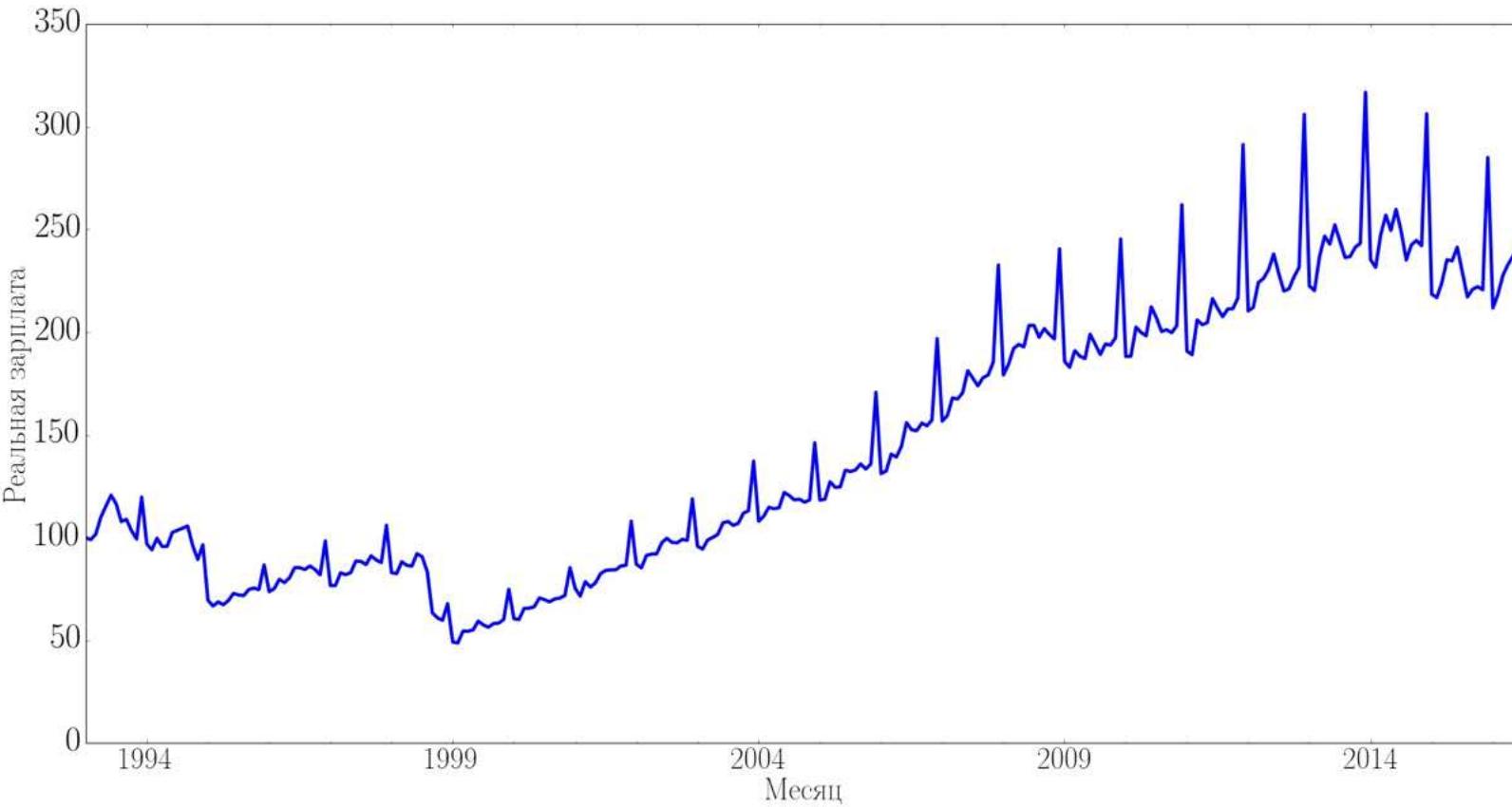
$$y_{T+d} \approx f_T(y_T, \dots, y_1, d) \equiv \hat{y}_{T+d|T},$$

$d \in \{1, \dots, D\}$, D — горизонт прогнозирования.

Главная особенность временных рядов

- В классических задачах анализа данных предполагается независимость наблюдений
 - При прогнозировании временных рядов, наоборот, мы надеемся, что значения ряда в прошлом содержат информацию о его поведении в будущем

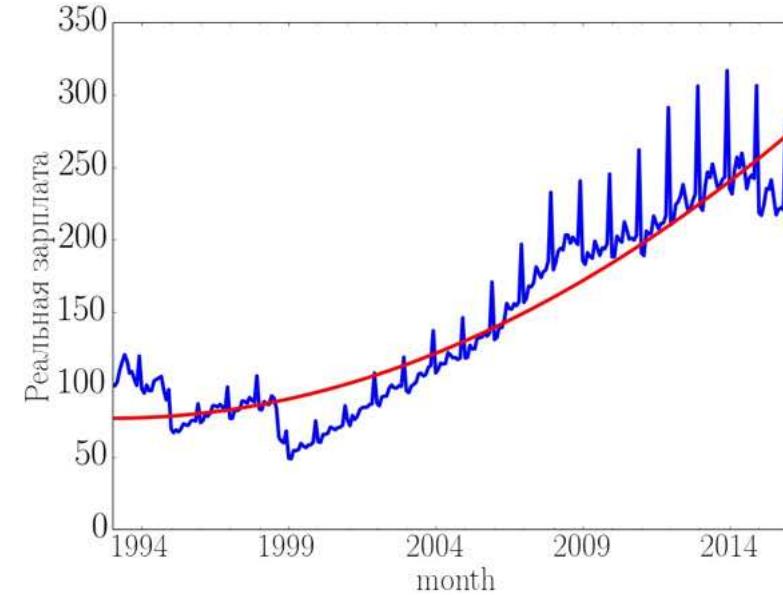
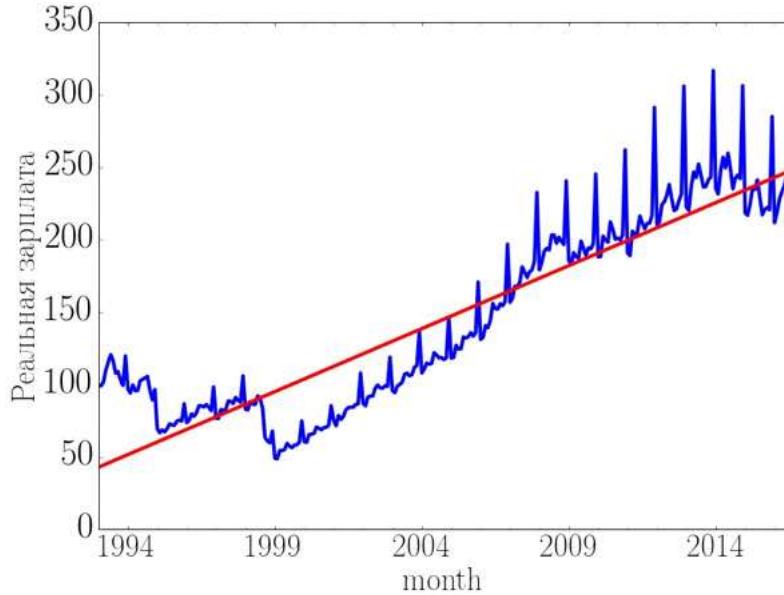
He i.i.d.



Это явно не случайная выборка!

Регрессия?

Простейшая идея: сделать регрессию на время.

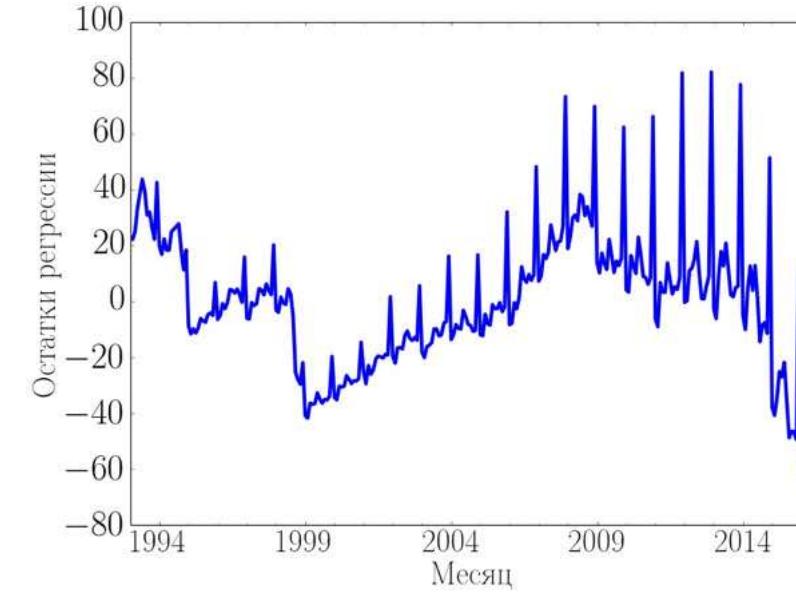
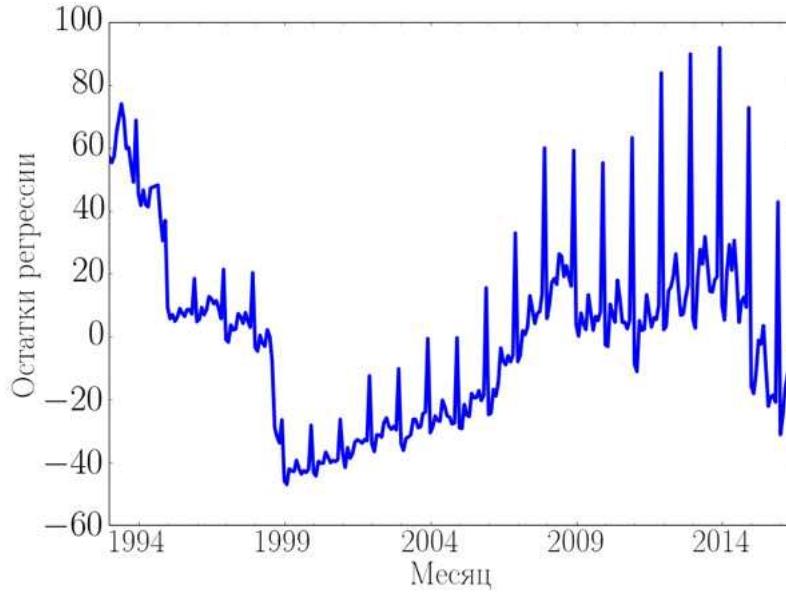


Предположение регрессии:

$$y | x \sim N(x\beta, \sigma^2)$$

Регрессия?

Остатки не выглядят как шум:



Компоненты временных рядов

Тренд —плавное долгосрочное изменение уровня ряда.

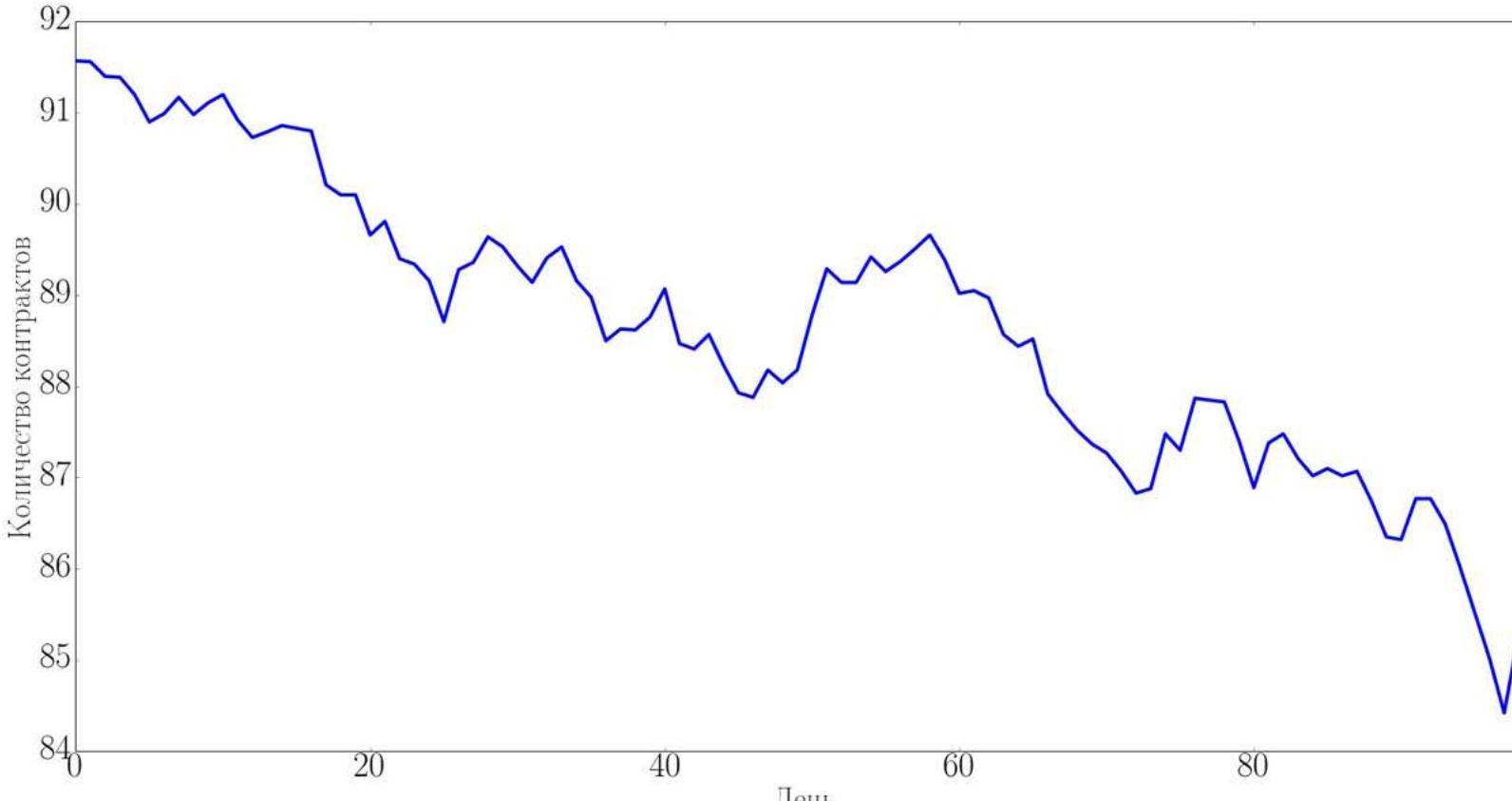
Сезонность —циклические изменения уровня ряда с постоянным периодом.

Цикл —изменения уровня ряда с переменным периодом (экономические циклы, периоды солнечной активности).

Ошибка —непрогнозируемая случайная компонента ряда.

Компоненты временных рядов

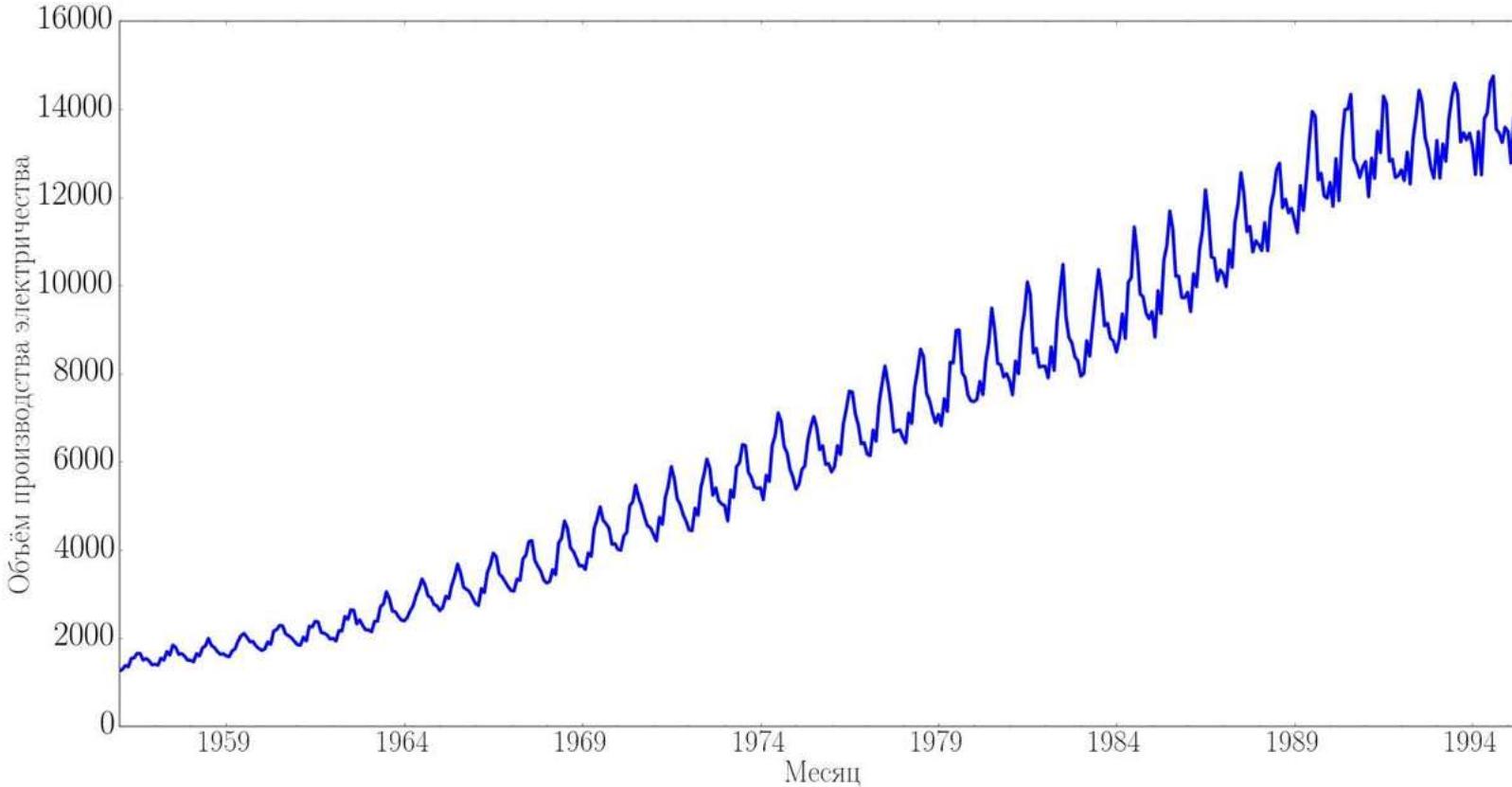
Количество контрактов сокровищницы США в день:



Тренд

Компоненты временных рядов

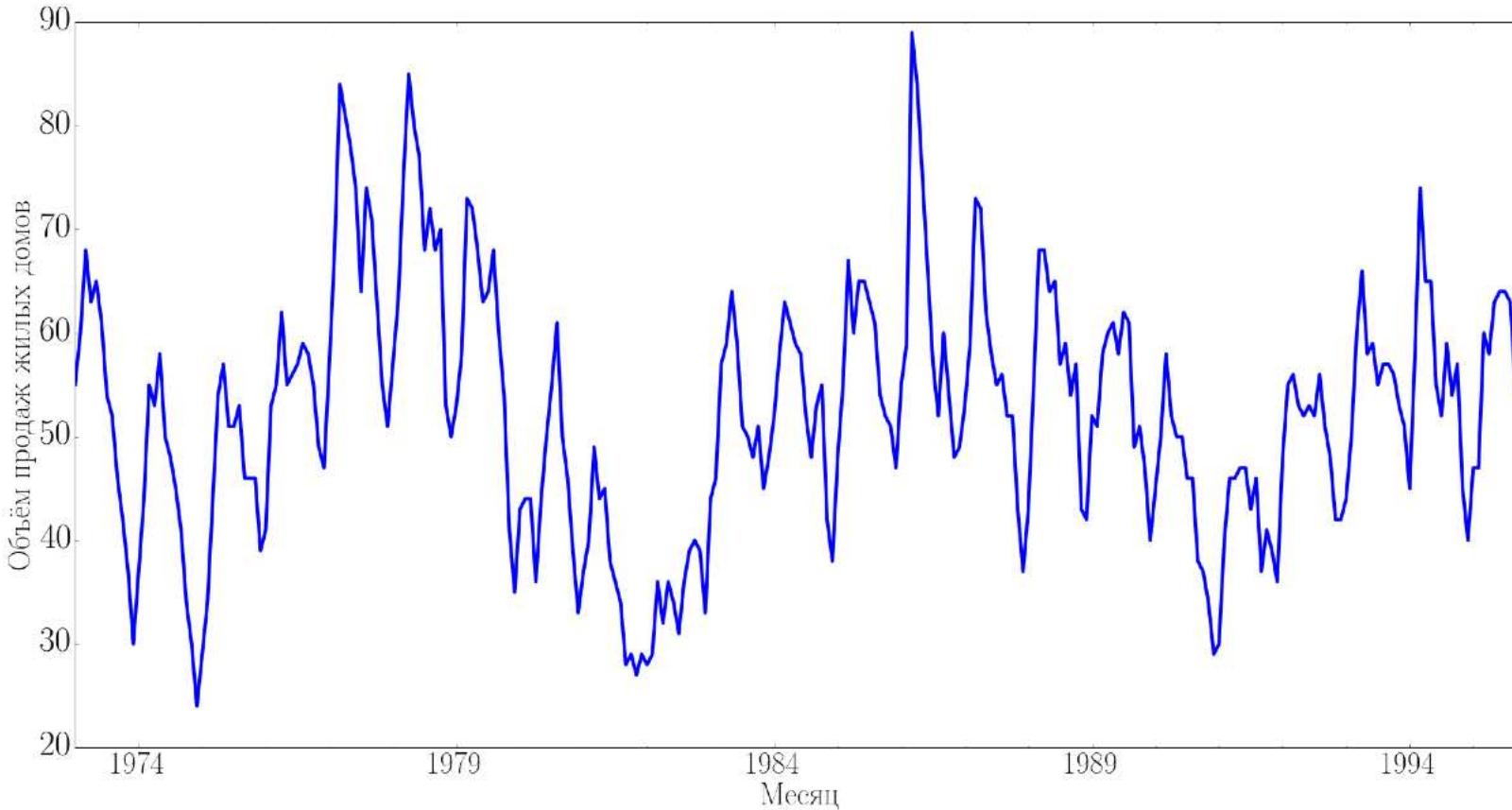
Объём производства электричества в Австралии:



Тренд, годовая сезонность

Компоненты временных рядов

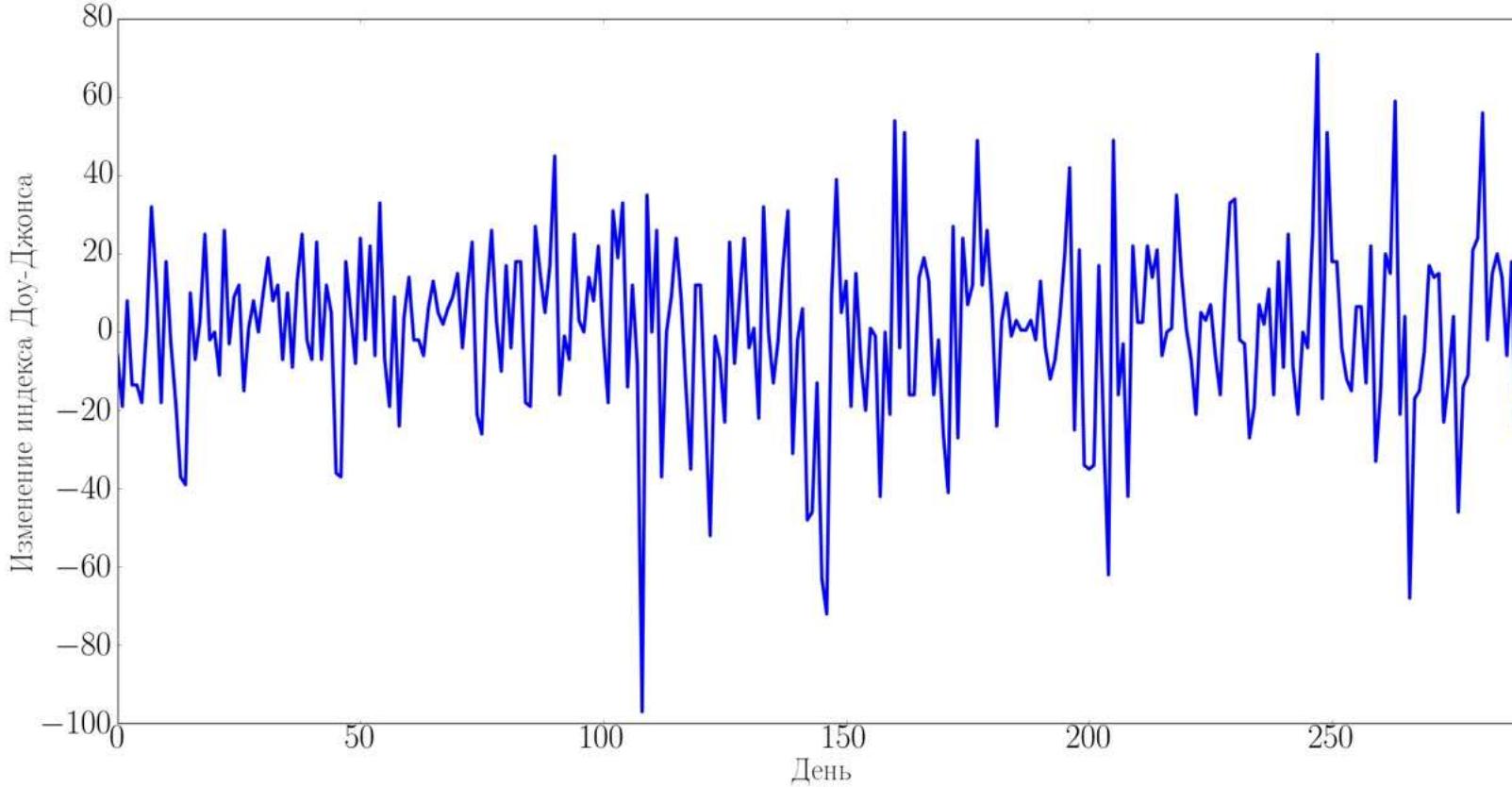
Объём продаж жилых домов:



Годовая сезонность, экономические циклы

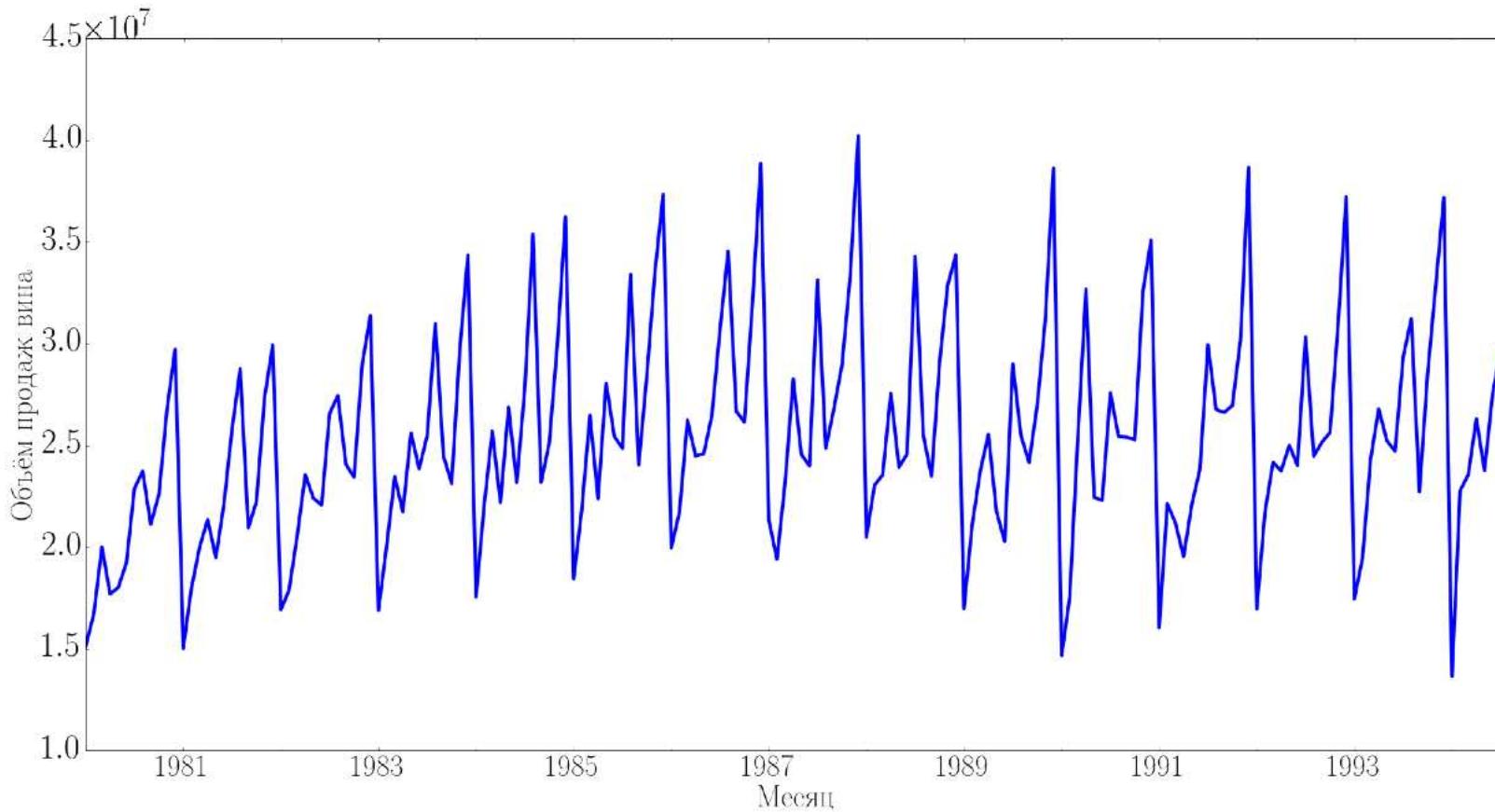
Компоненты временных рядов

Ежедневные изменения индекса Доу-Джонса:



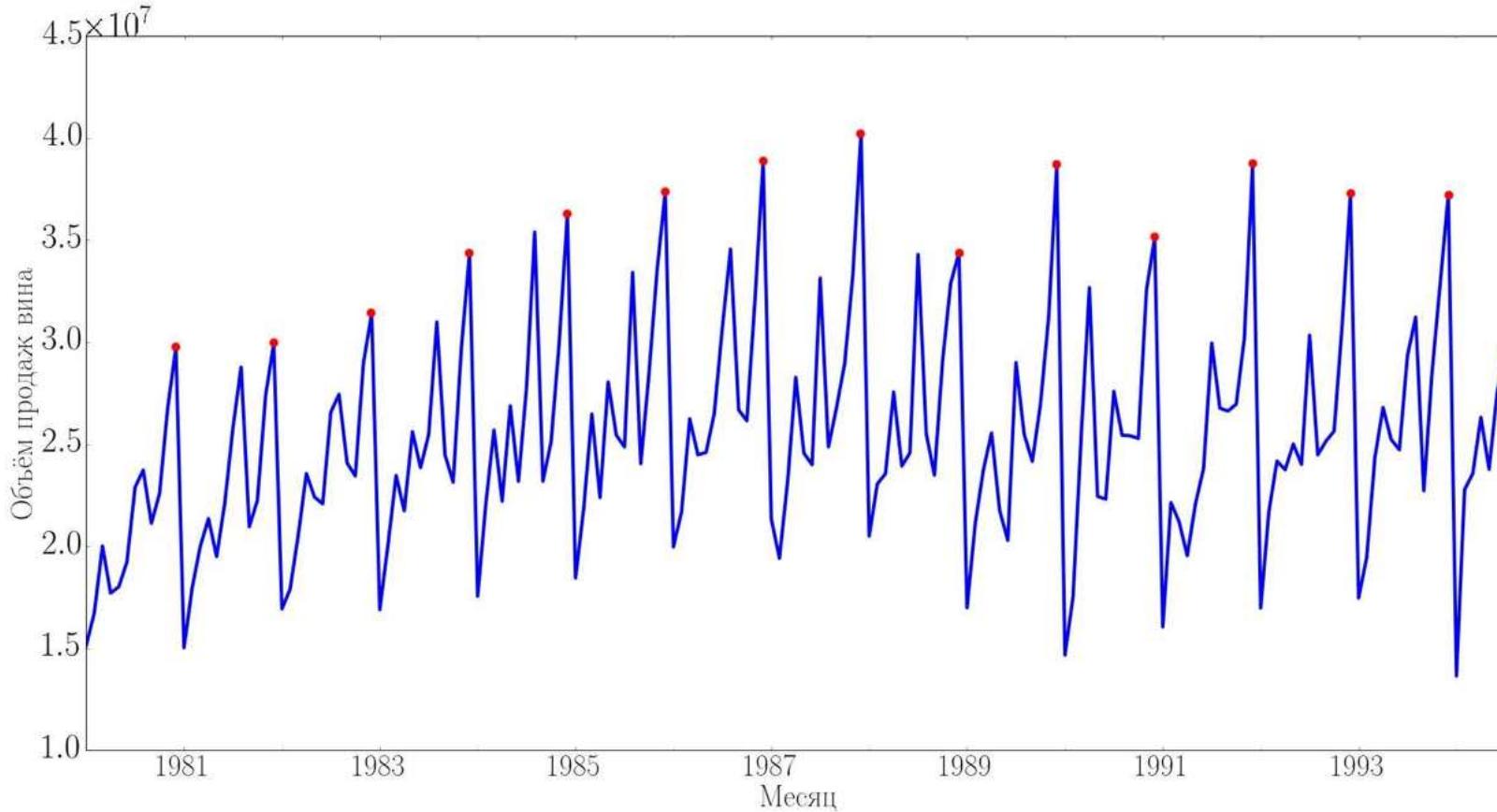
Ничего

Продажи вина в Австралии



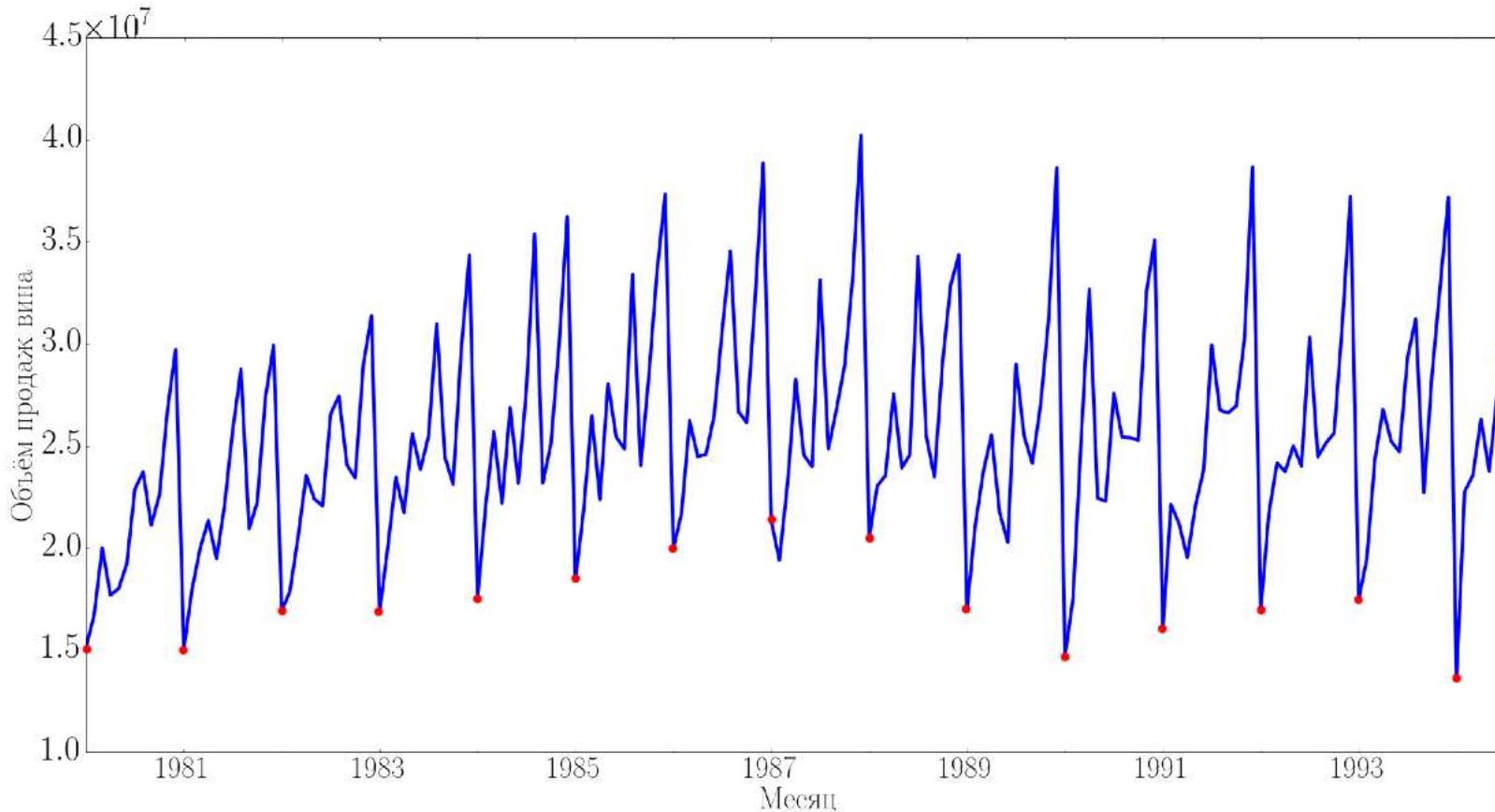
Продажи вина в Австралии

Каждый декабрь продажи большие:

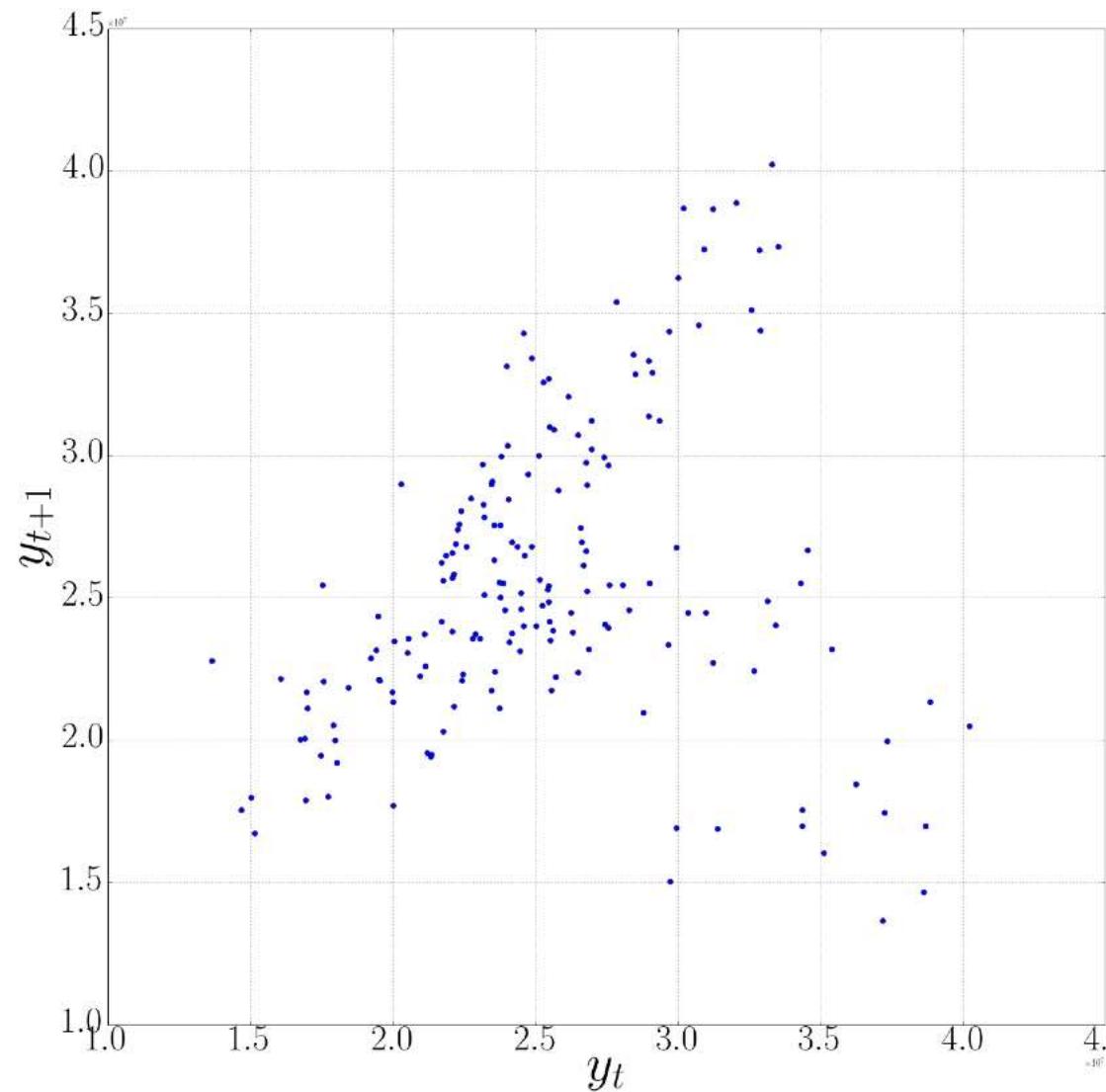


Продажи вина в Австралии

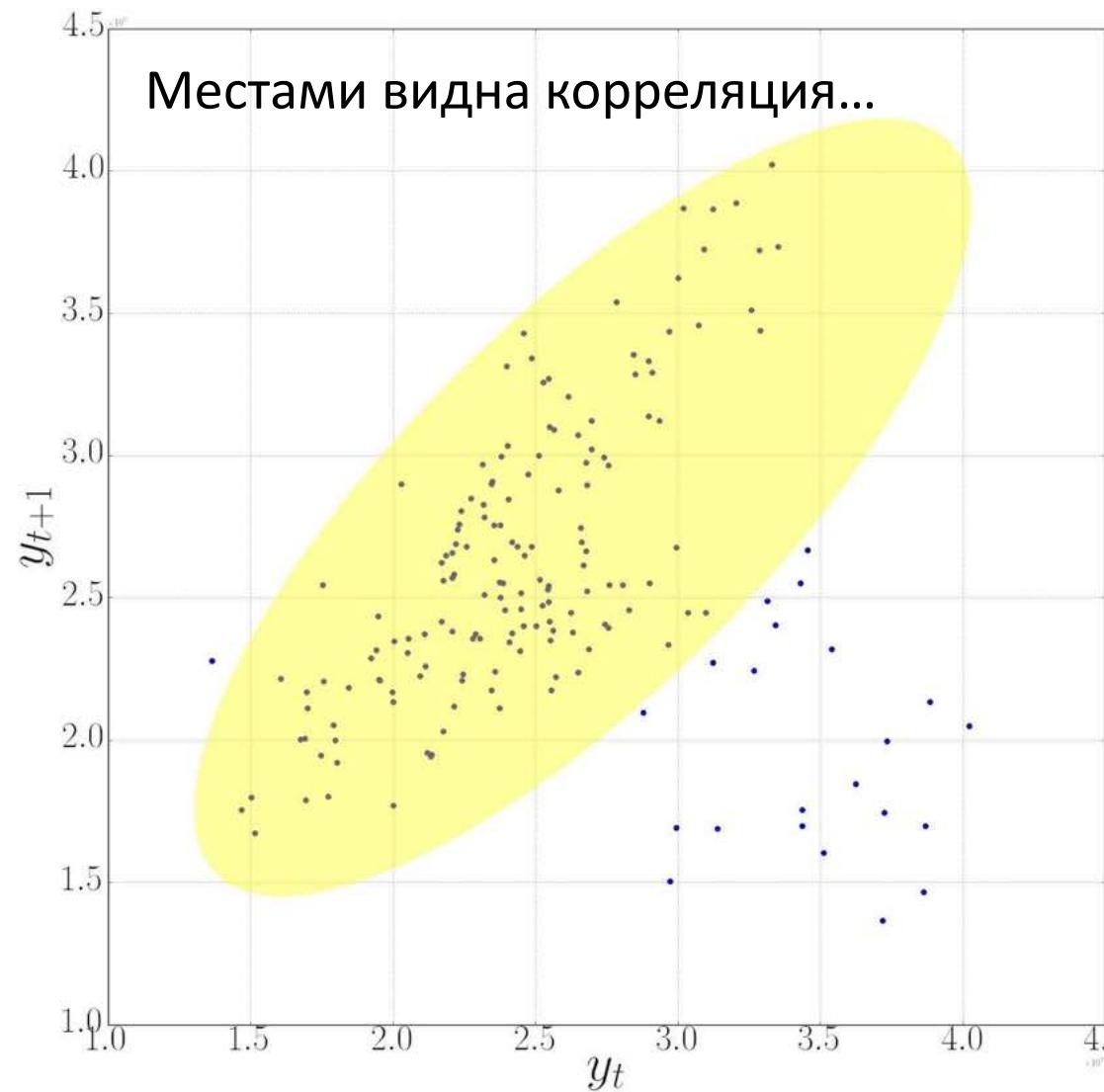
Каждый январь продажи падают:



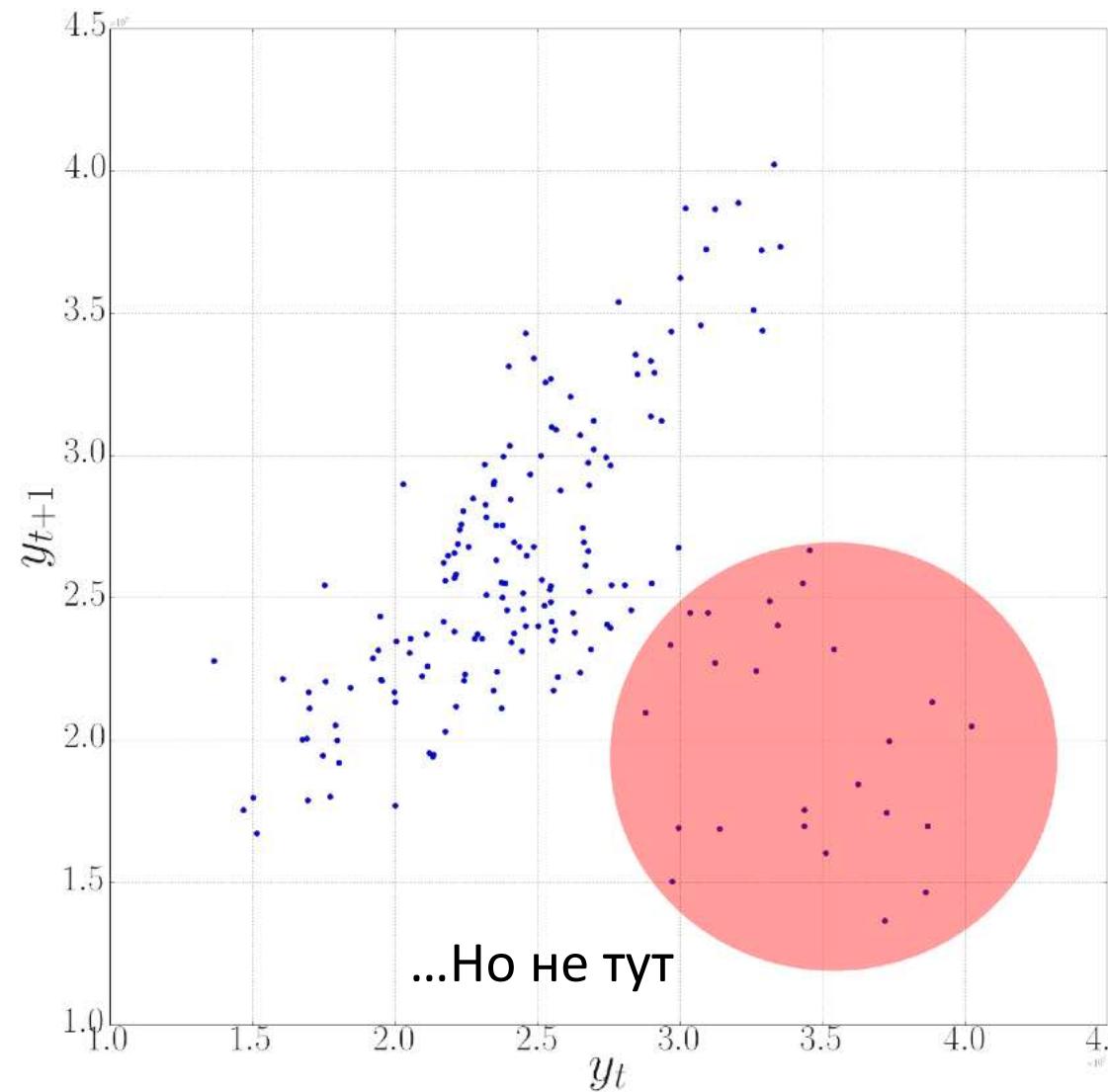
Продажи в соседние месяцы



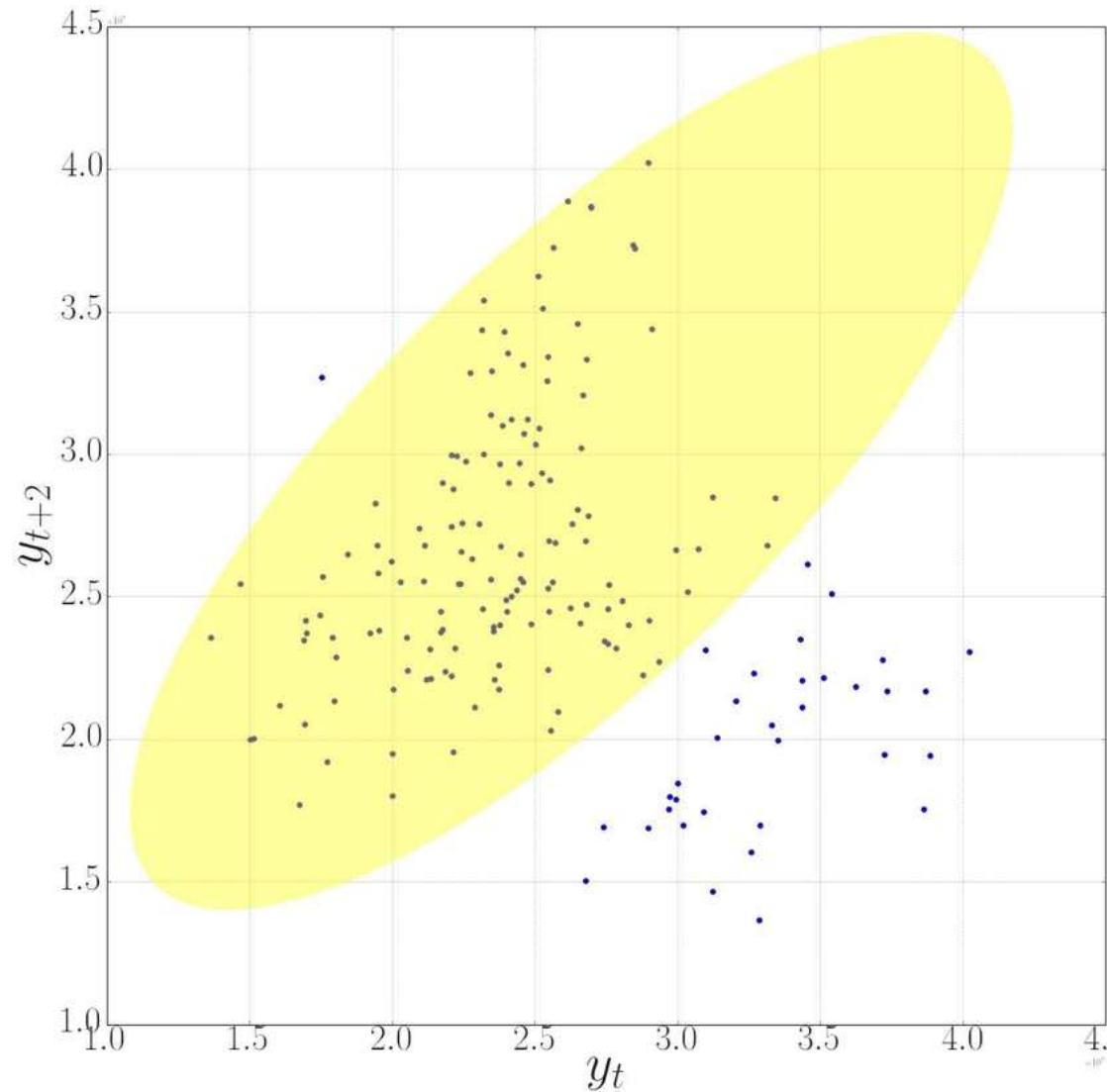
Продажи в соседние месяцы



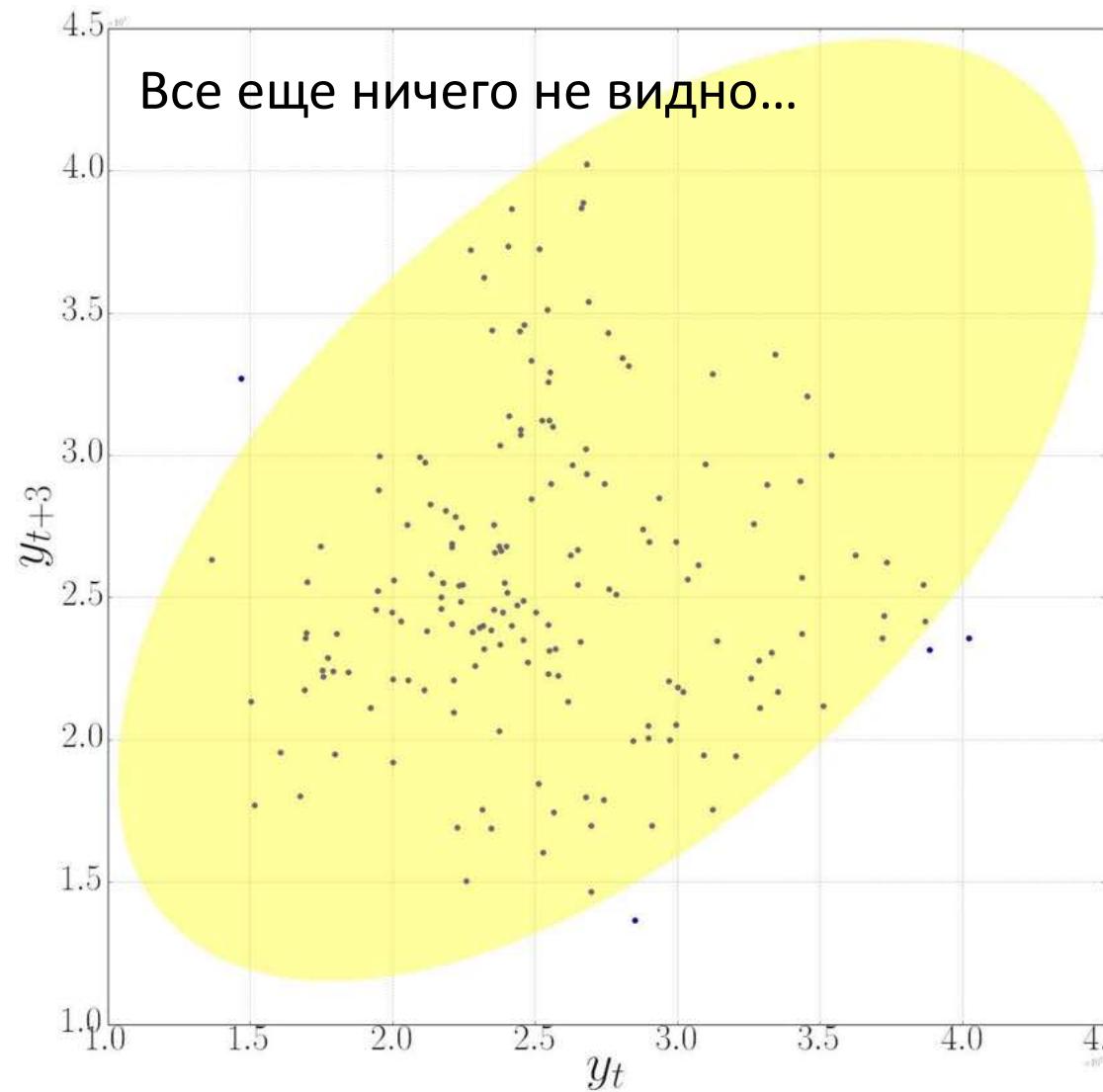
Продажи в соседние месяцы



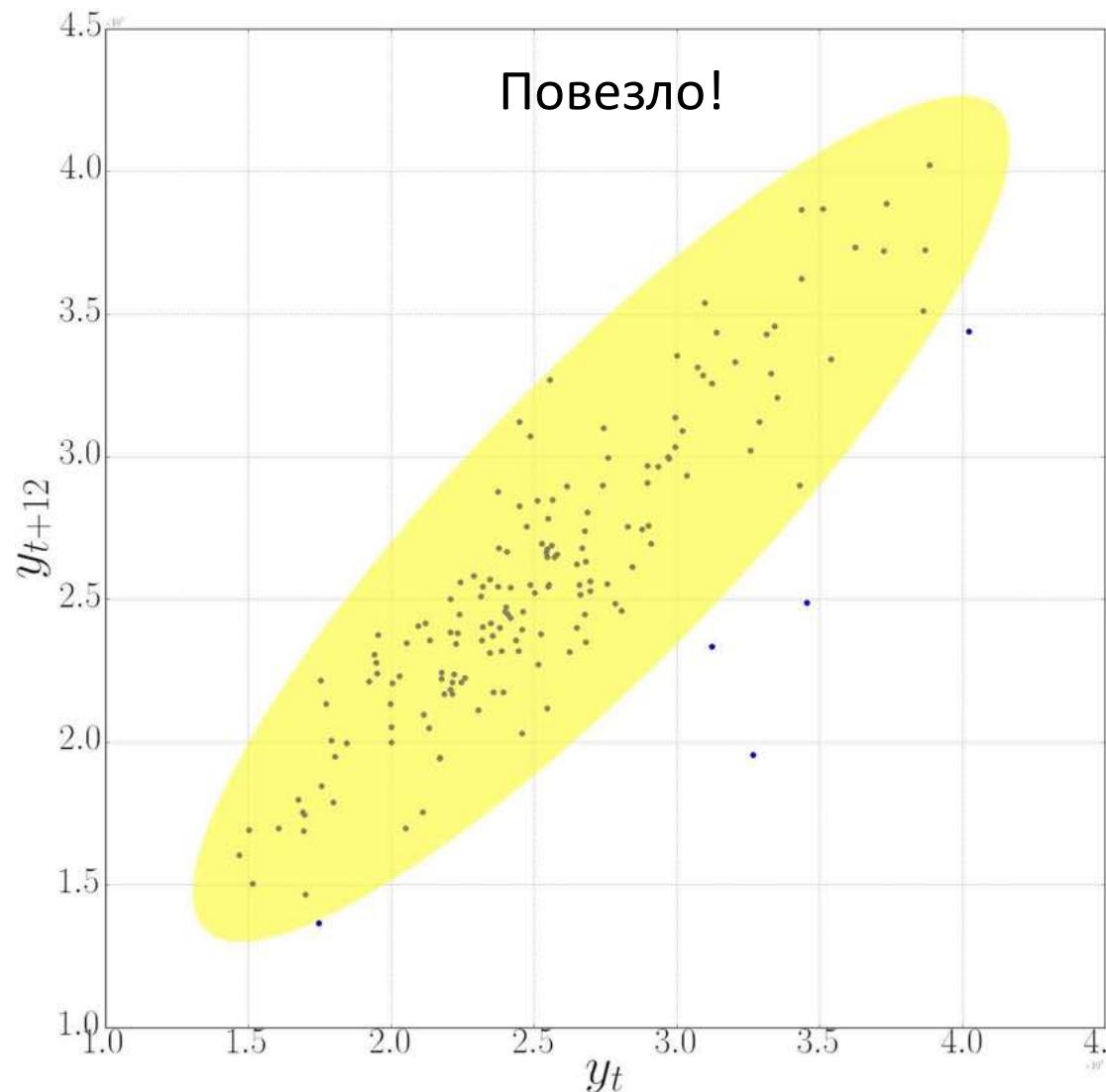
Продажи через 1 месяц



Продажи через 2 месяца



Продажи через год



Автокорреляция

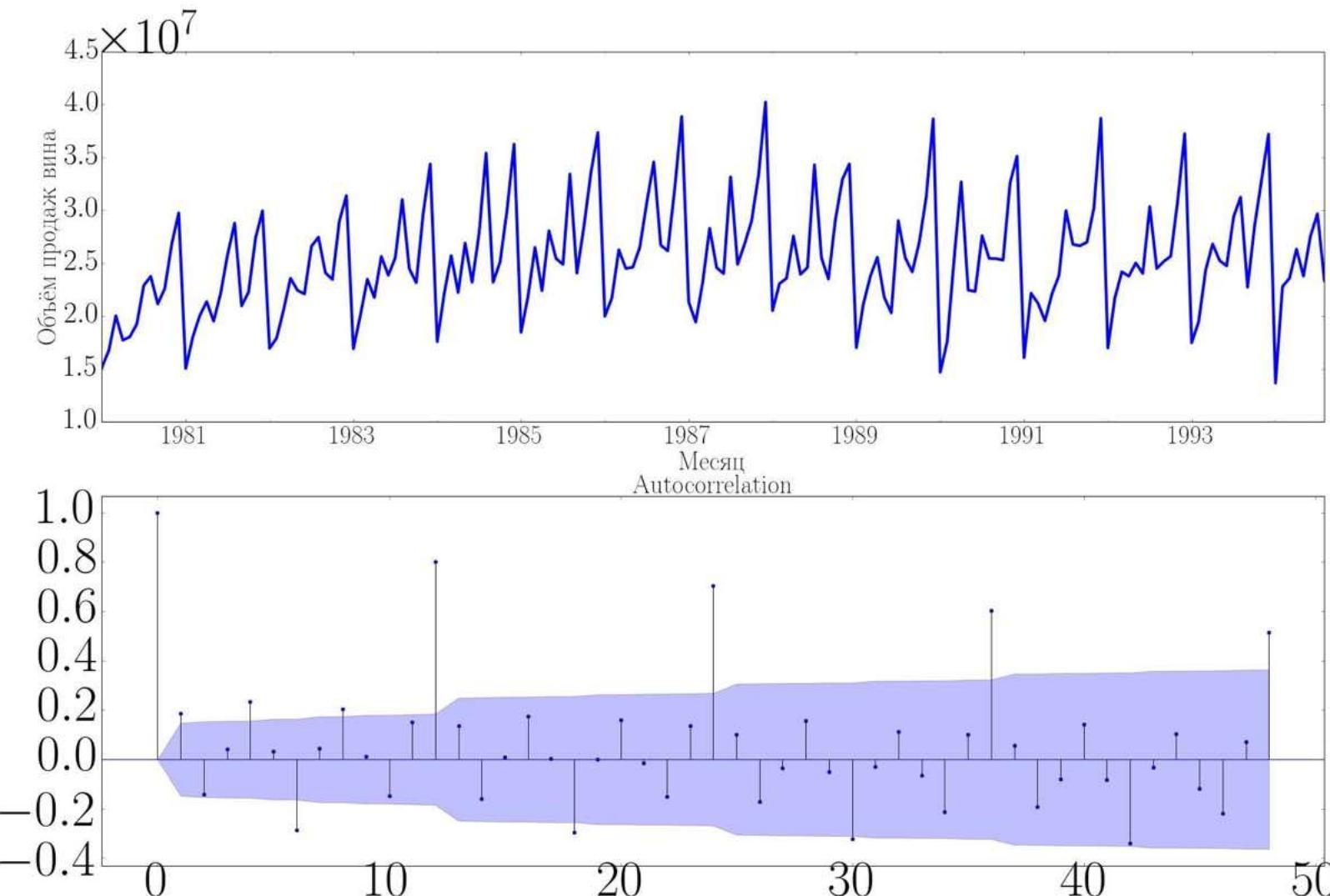
$$r_\tau = \frac{\mathbb{E}((y_t - \mathbb{E}y)(y_{t+\tau} - \mathbb{E}y))}{\mathbb{D}y}.$$

$r_\tau \in [-1, 1]$, τ — лаг автокорреляции.

Выборочная автокорреляция:

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

Коррелограммы



Значимость автокорреляции

временной ряд: $y^T = y_1, \dots, y_T$;

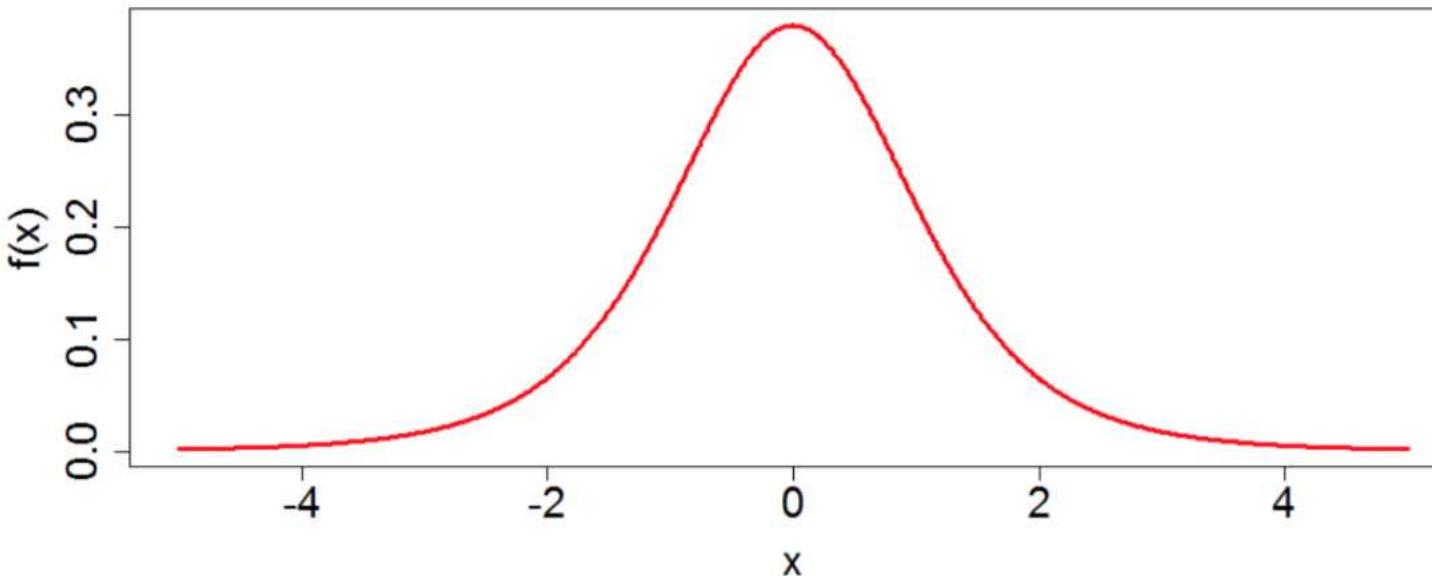
нулевая гипотеза: $H_0: r_\tau = 0$;

альтернатива: $H_1: r_\tau \neq 0$;

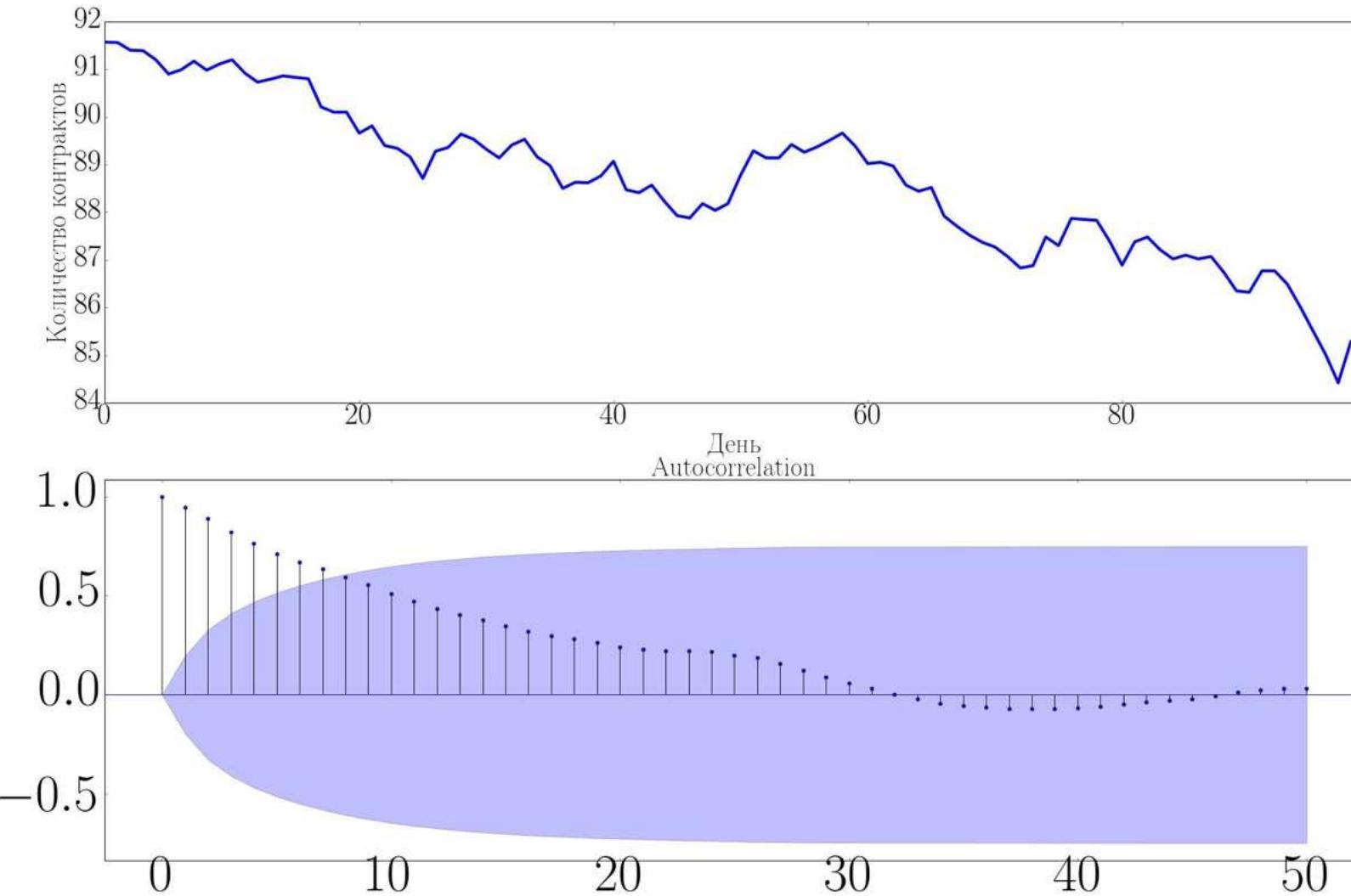
статистика: $T(y^T) = \frac{r_\tau \sqrt{T-\tau-2}}{\sqrt{1-r_\tau^2}}$;

нулевое распределение: $T(y^T) \sim St(T - \tau - 2)$ при H_0 .

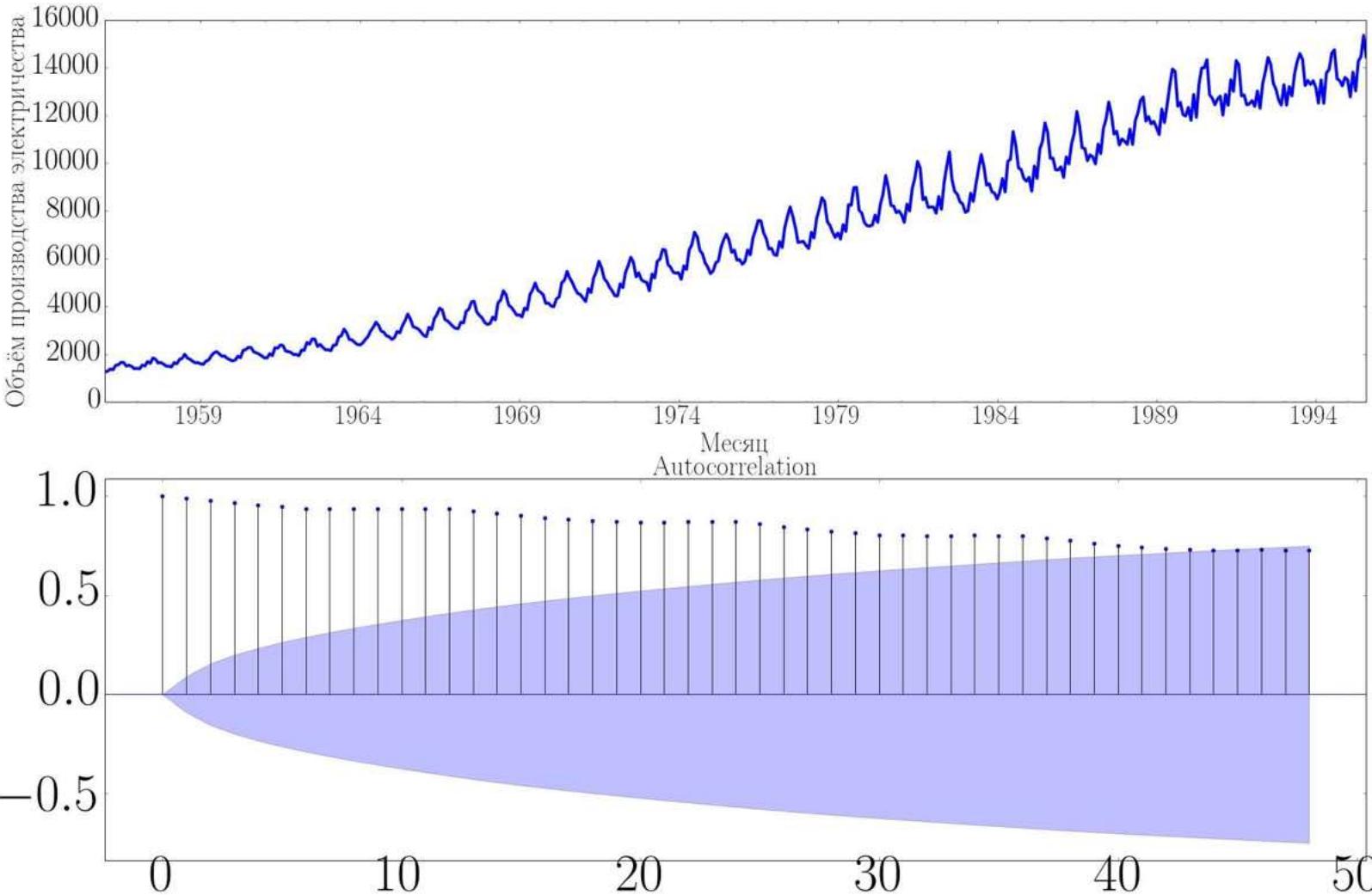
Эти формулы
не надо учить 😊



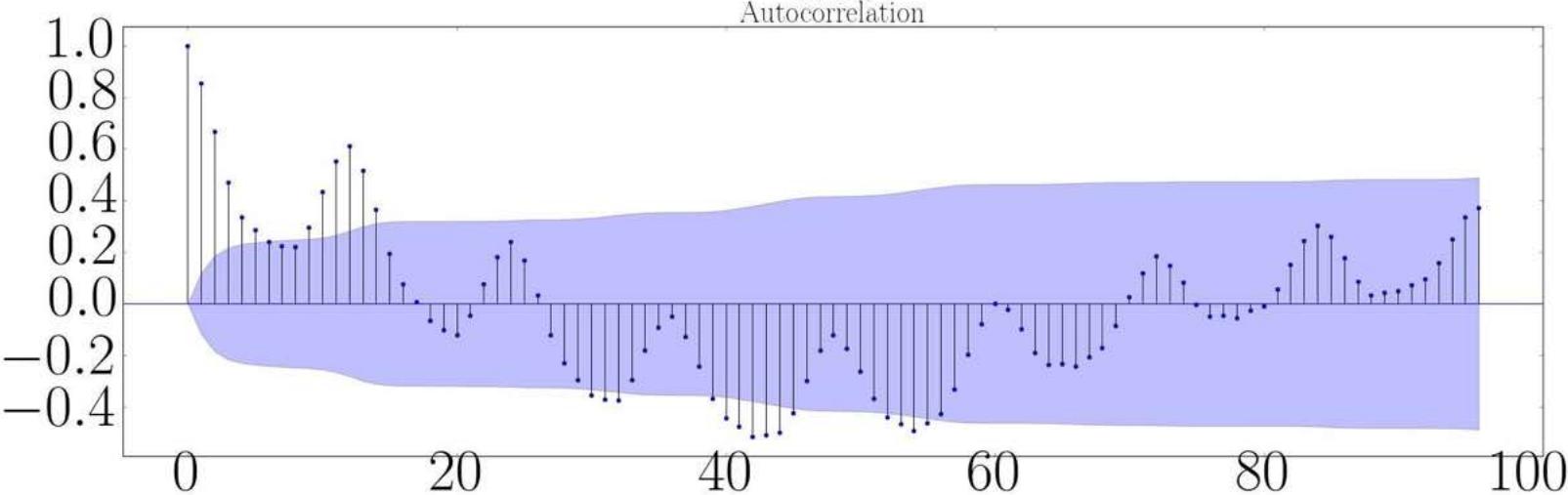
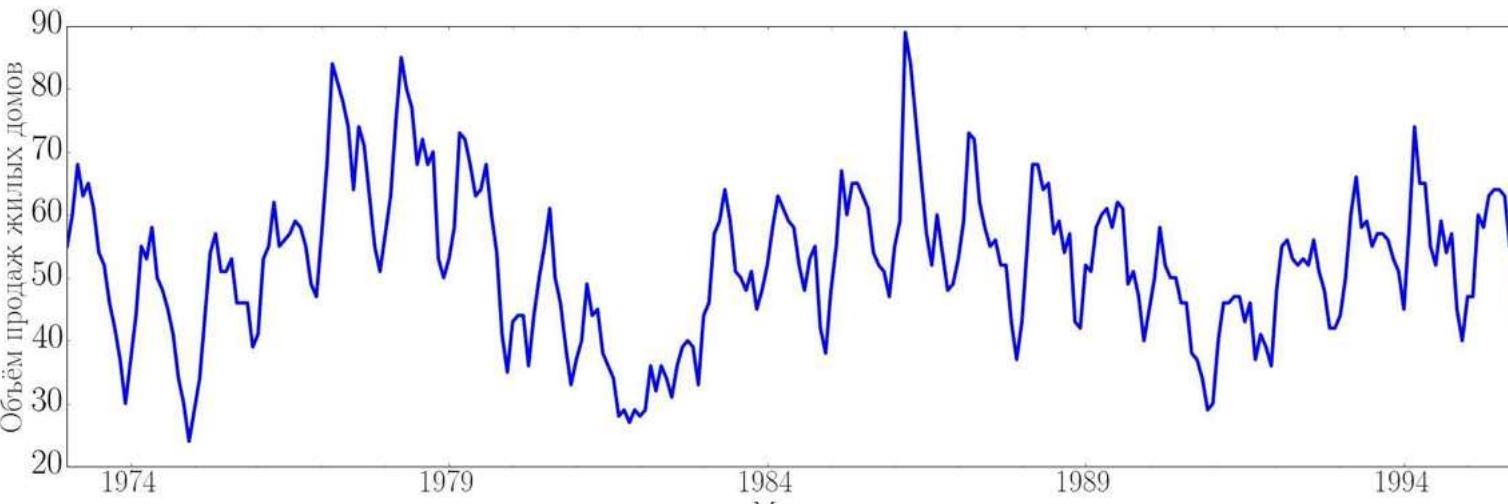
Коррелограммы



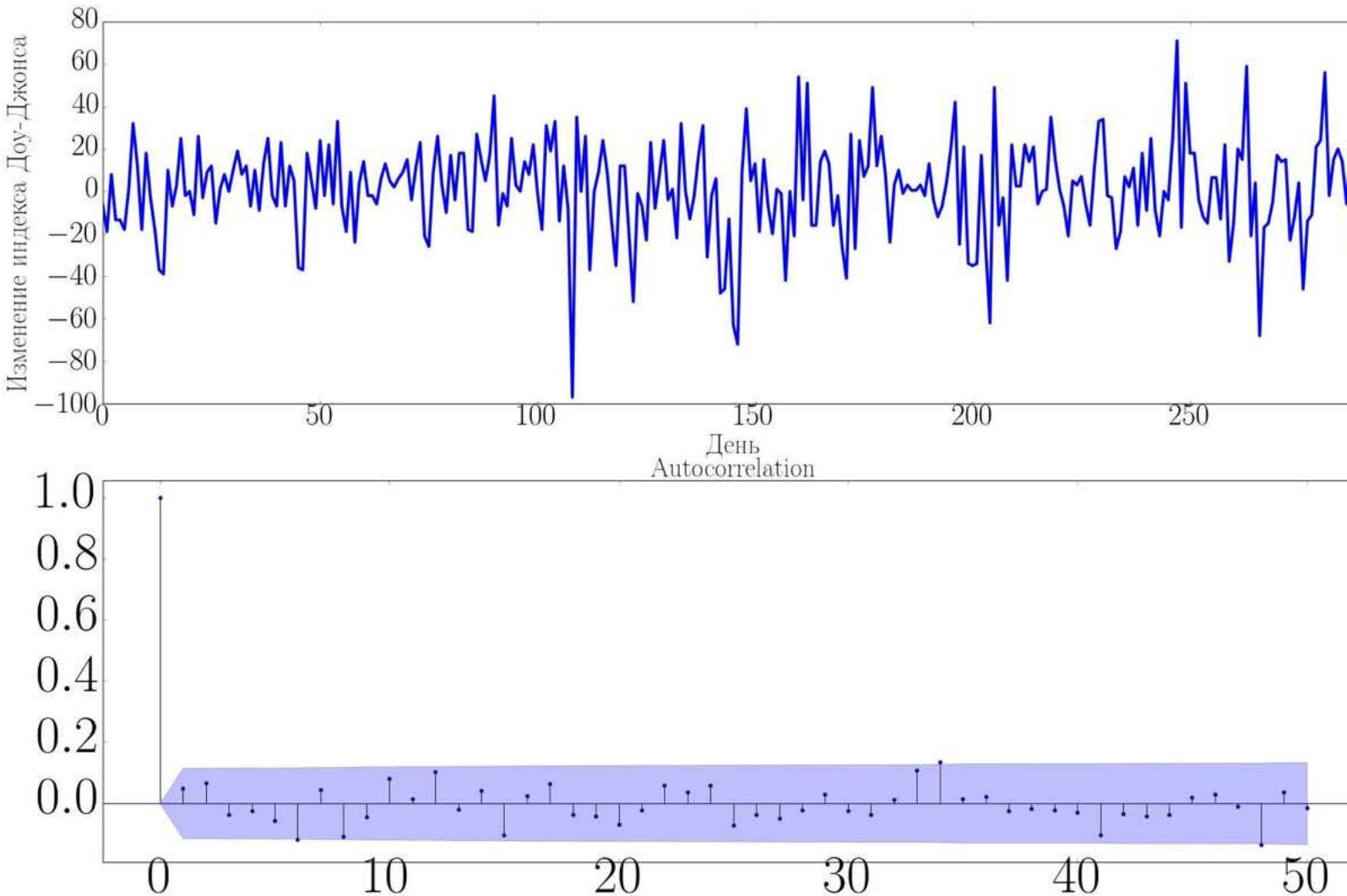
Коррелограммы



Коррелограммы



Коррелограммы

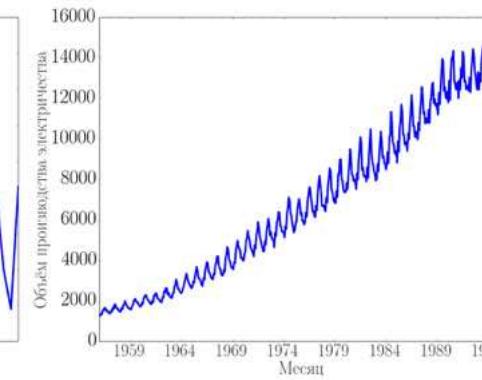
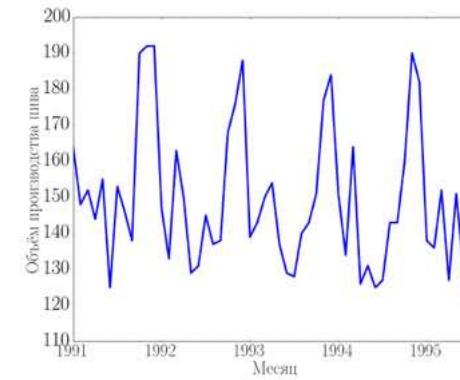
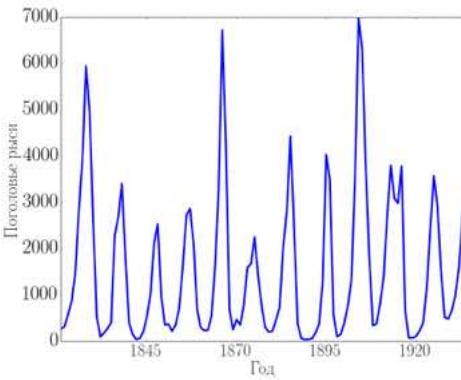
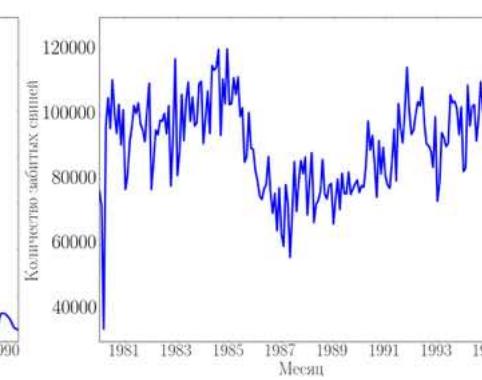
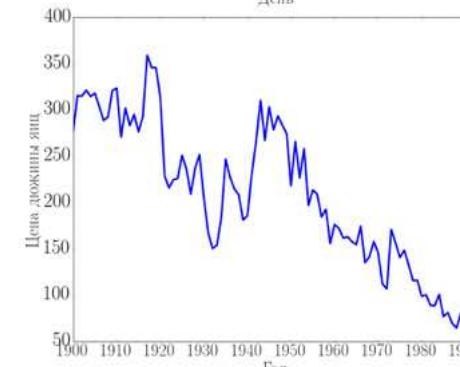
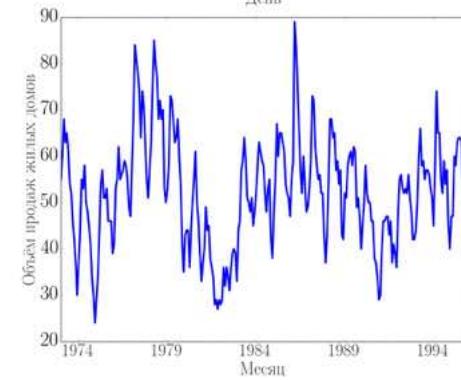
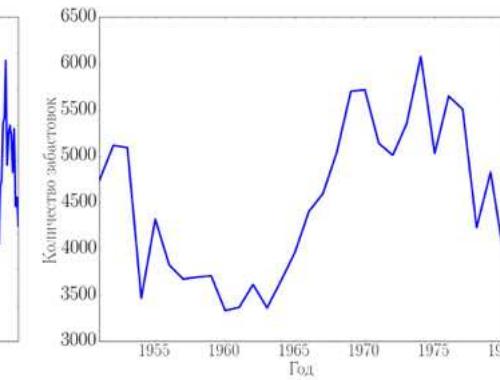
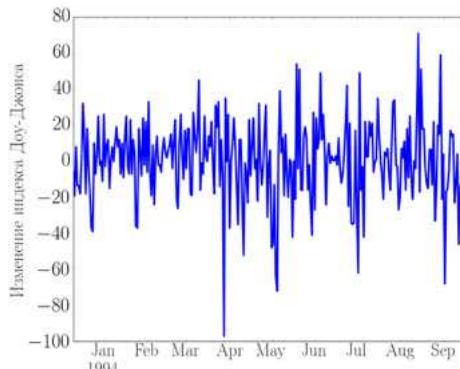
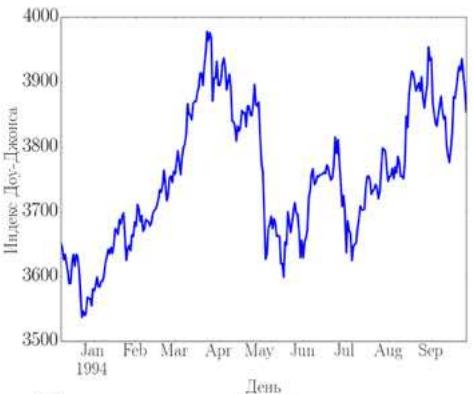


Стационарность

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

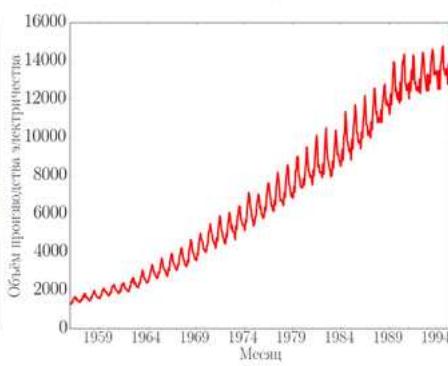
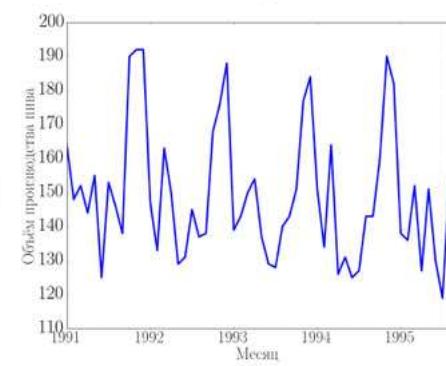
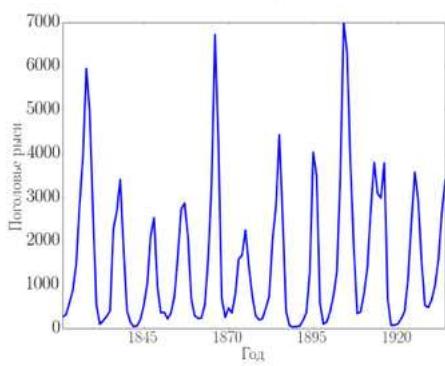
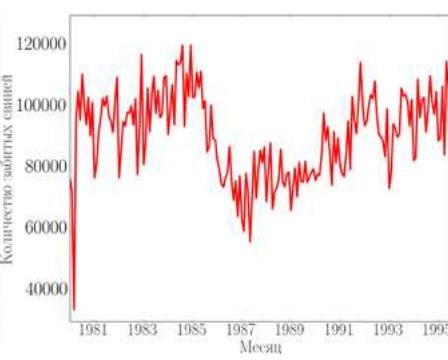
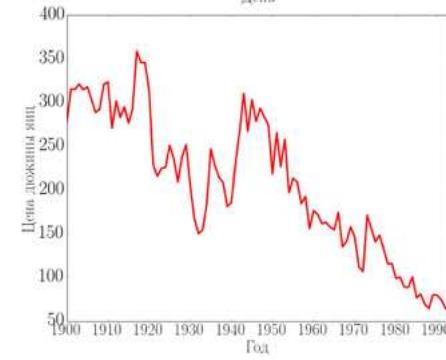
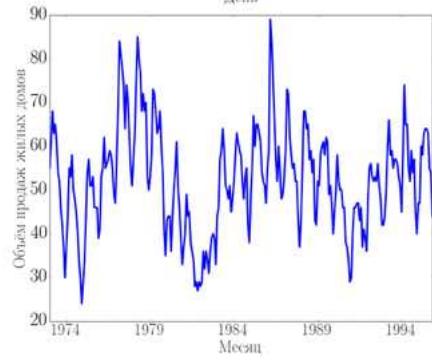
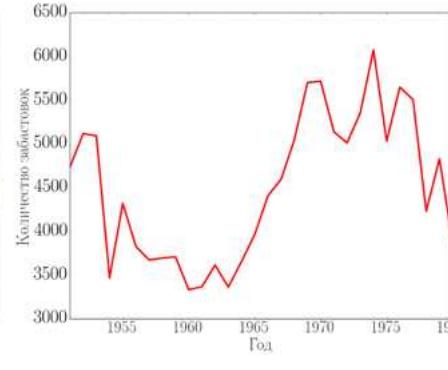
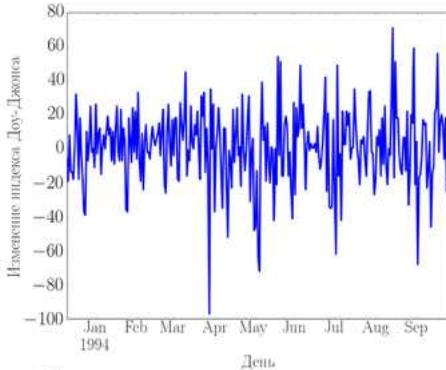
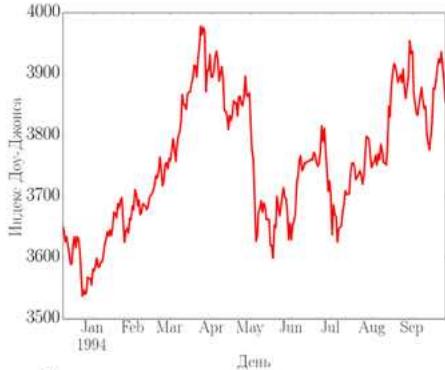
- тренд \Rightarrow нестационарность
- сезонность \Rightarrow нестационарность
- цикл $\not\Rightarrow$ нестационарность (нельзя предсказать заранее, где будут находиться максимумы и минимумы)

Примеры



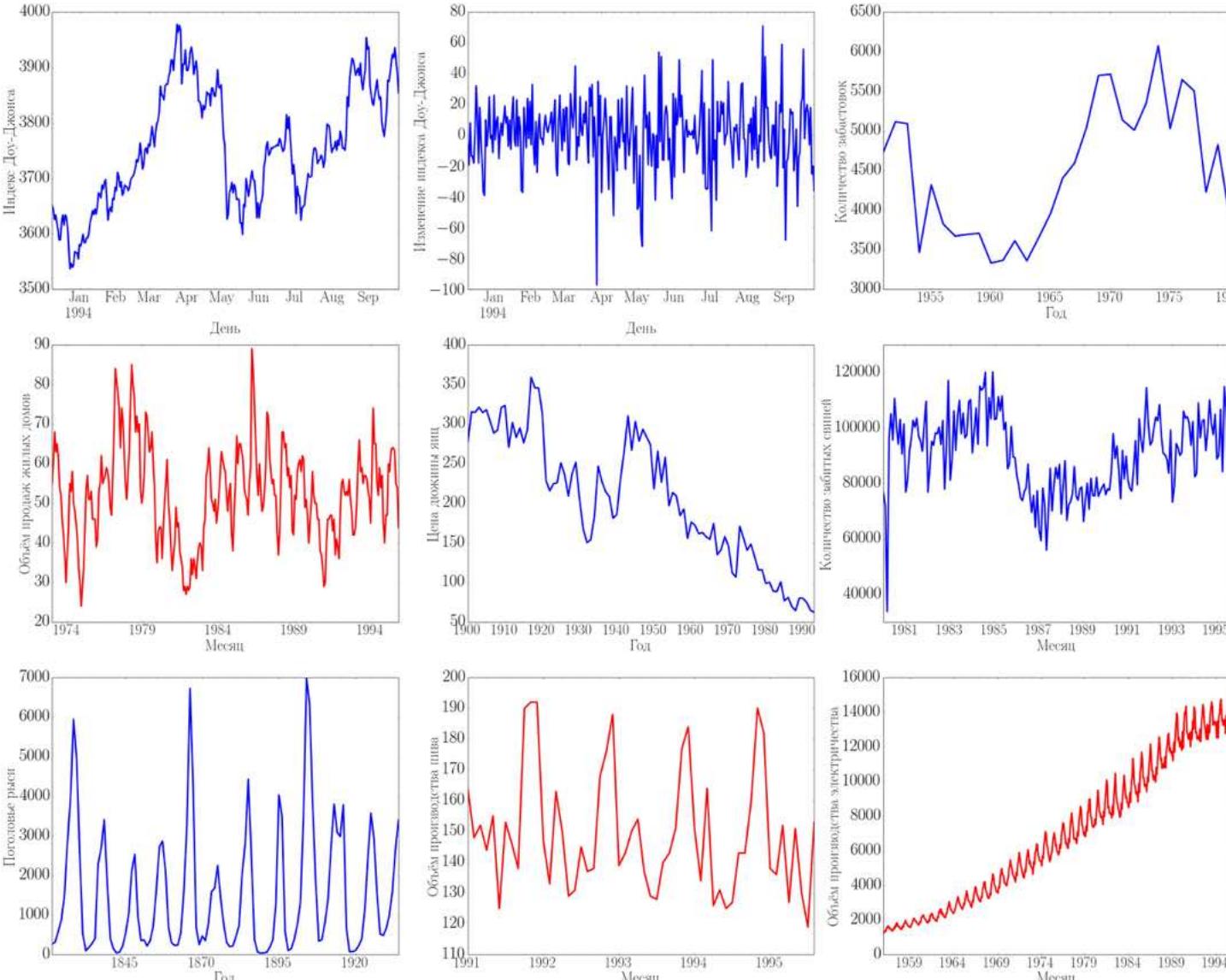
Примеры

Нестационарны из-за тренда:



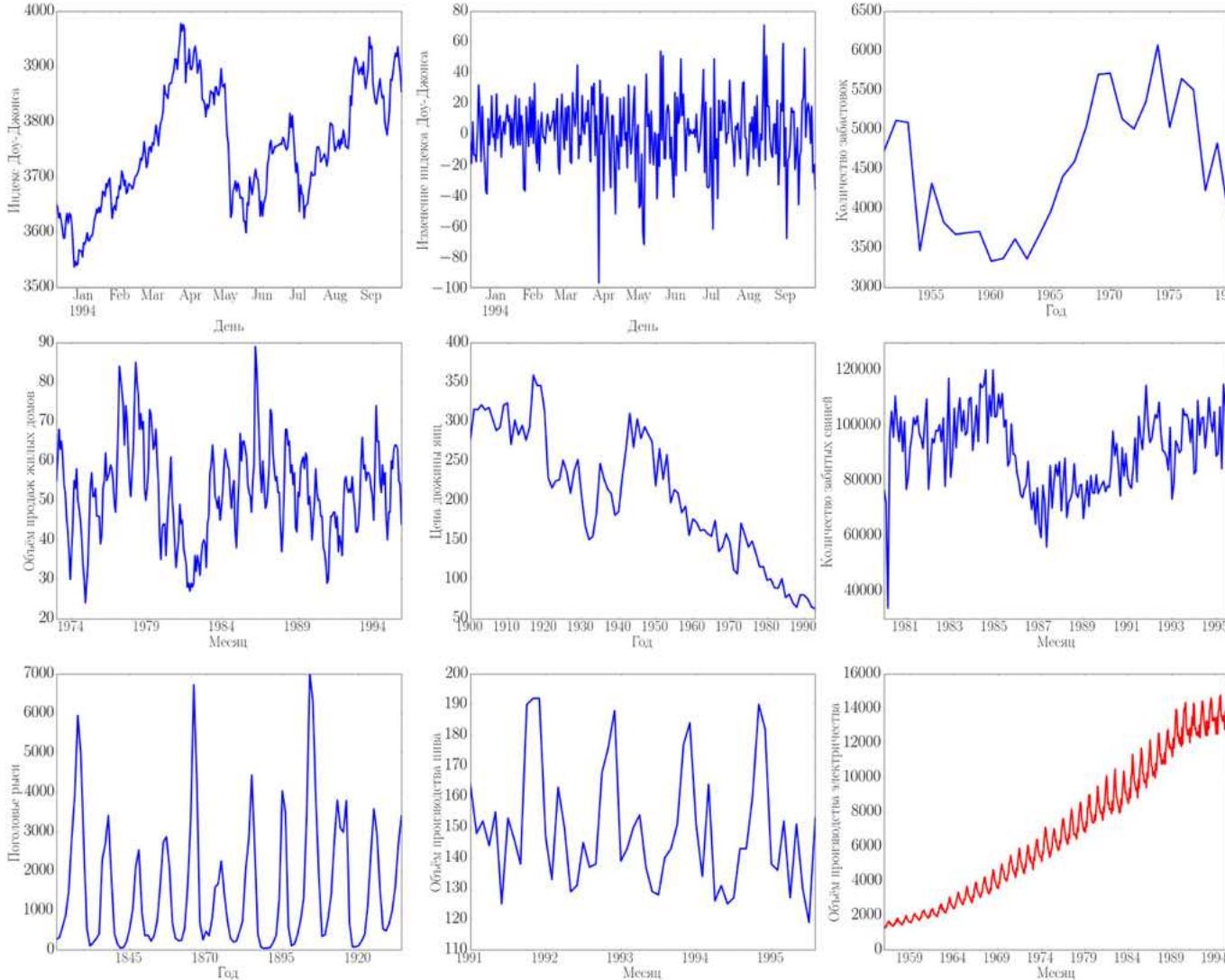
Примеры

Нестационарны из-за сезонности:



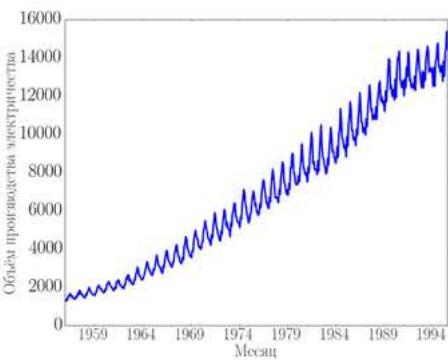
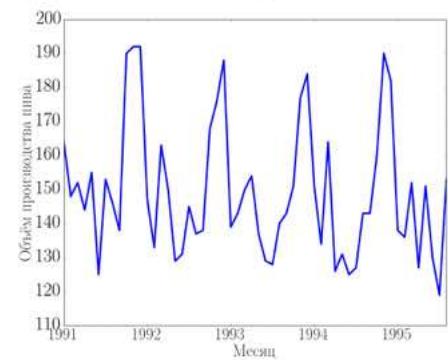
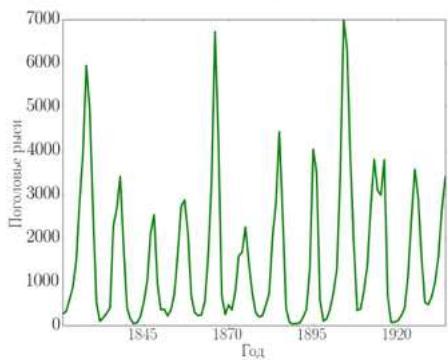
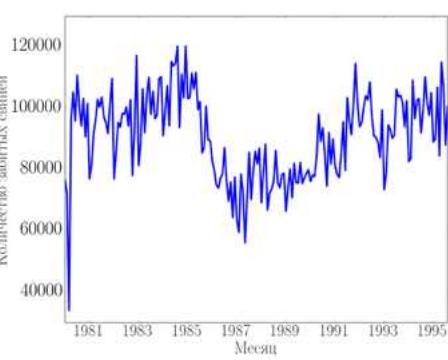
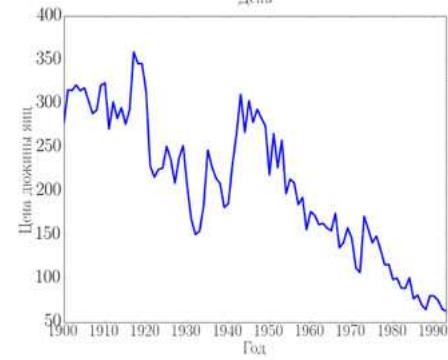
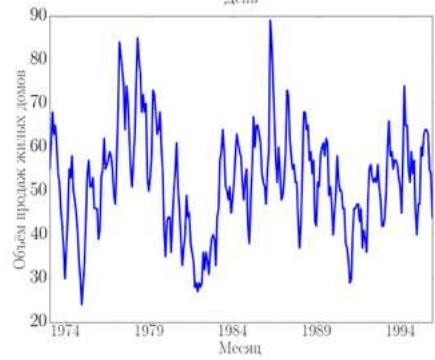
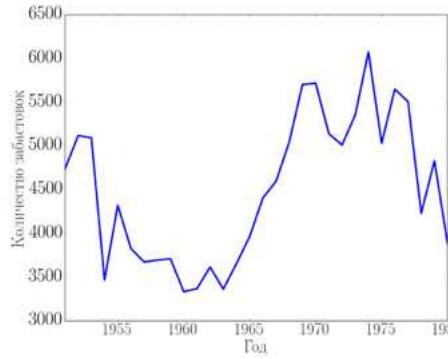
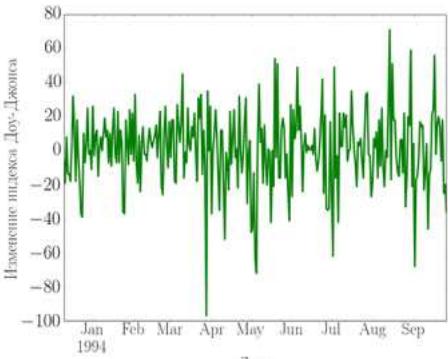
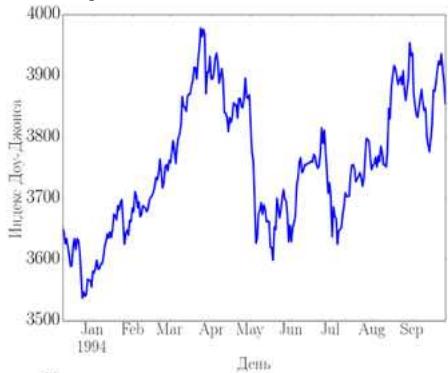
Примеры

Нестационарны из-за меняющейся дисперсии:



Примеры

Стационарны:



Критерий Дики-Фуллера

временной ряд: $y^T = y_1, \dots, y_T$;

нулевая гипотеза: H_0 : ряд нестационарен;

альтернатива: H_1 : ряд стационарен;

статистика: неважно;

нулевое распределение: табличное.

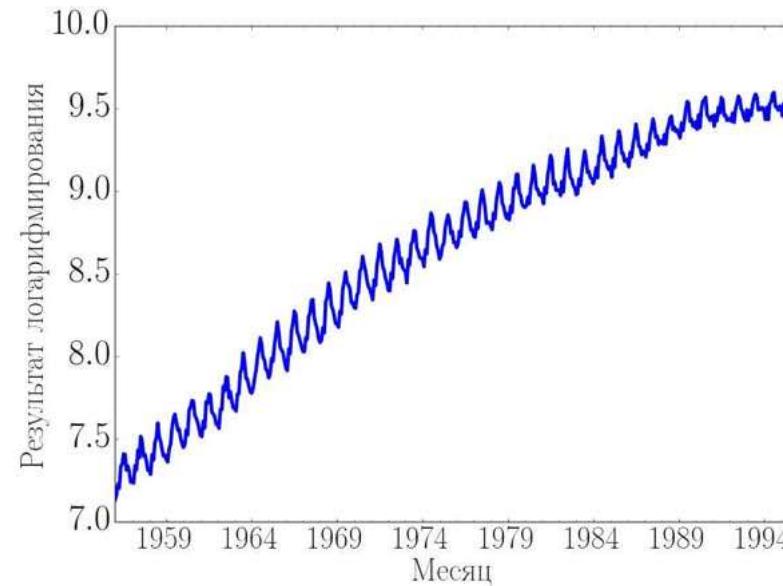
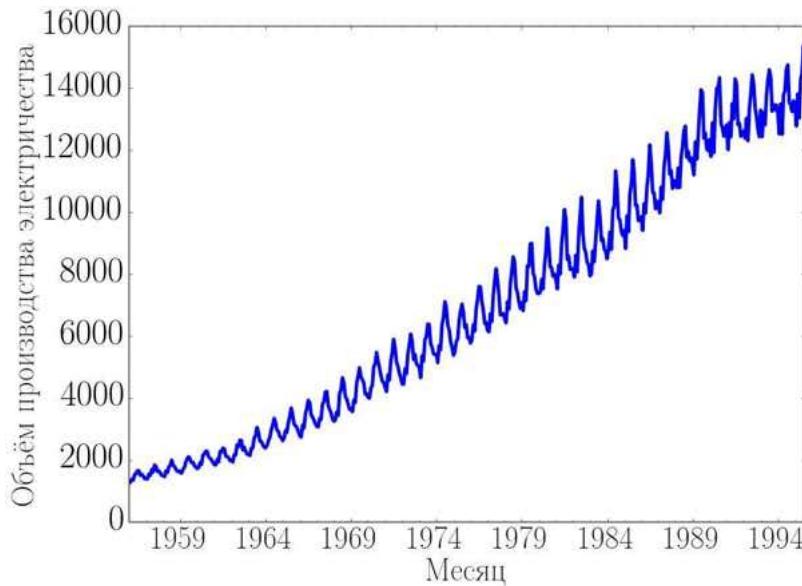
Зачем нам стационарность?

- Чтобы не зависеть от времени => применять регрессию по последним значениям

Стабилизация дисперсии

Для рядов с монотонно меняющейся дисперсией можно использовать стабилизирующие преобразования.

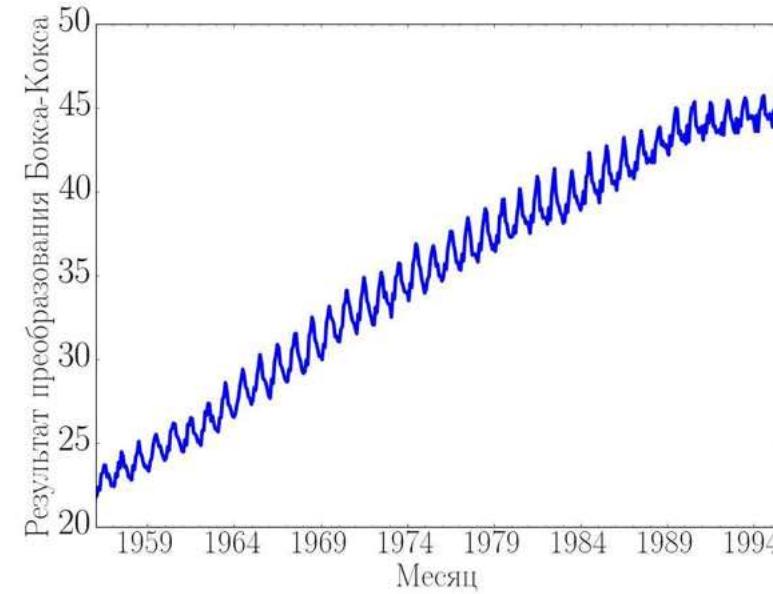
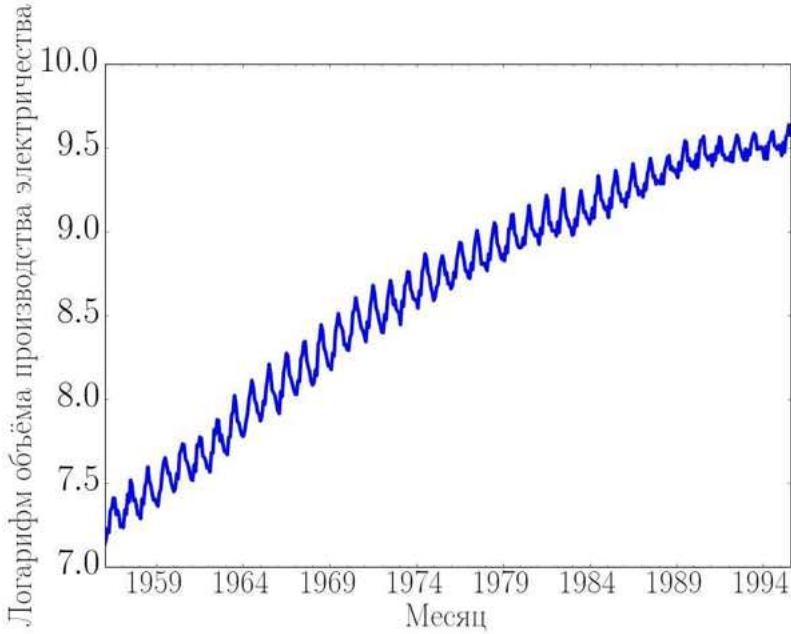
Часто используют логарифмирование:



Стабилизация дисперсии

Преобразования Бокса-Кокса:

$$y'_t = \begin{cases} \ln y_t, & \lambda = 0, \\ (y_t^\lambda - 1) / \lambda, & \lambda \neq 0. \end{cases}$$



Стабилизация дисперсии

После построения прогноза для трансформированного ряда его нужно преобразовать в прогноз исходного:

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

- Если некоторые $y_t \leq 0$, преобразования Бокса-Кокса невозможны (нужно прибавить к ряду константу).
- Можно округлять значение λ , чтобы упростить интерпретацию.

Дифференцирование

Дифференцирование ряда — переход к попарным разностям соседних значений:

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

- позволяет стабилизировать среднее значение ряда и избавиться от тренда
- может применяться неоднократно

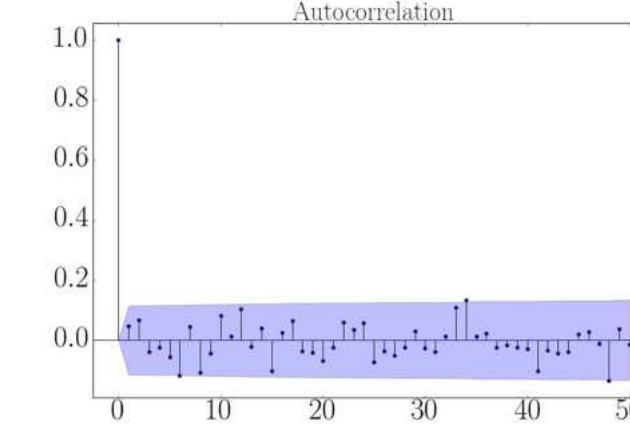
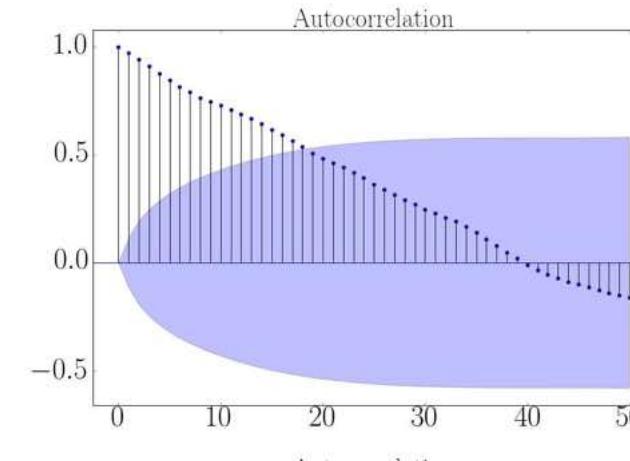
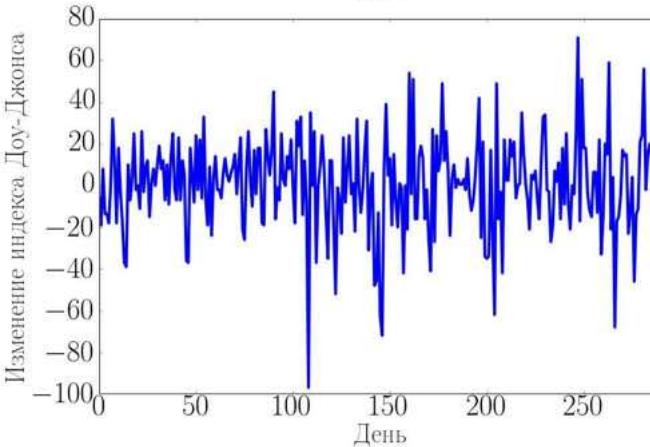
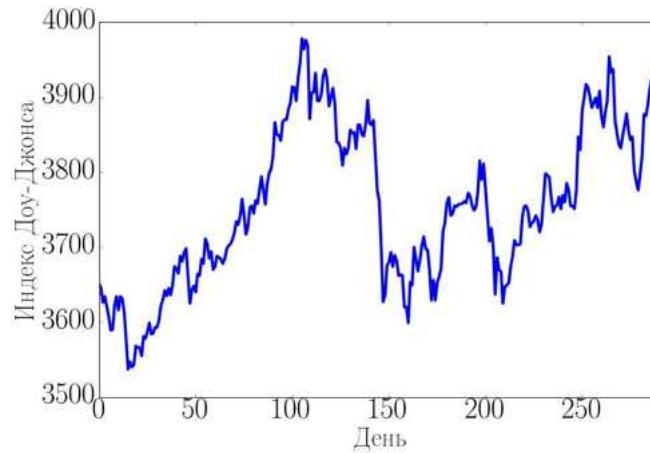
Дифференцирование

Сезонное дифференцирование ряда — переход к попарным разностям значений в соседних сезонах:

$$y_t' = y_t - y_{t-s}.$$

- убирает сезонность
- сезонное и обычное дифференцирование могут применяться к ряду в любом порядке
- если ряд имеет выраженный сезонный профиль, рекомендуется начинать с сезонного дифференцирования — после него ряд уже может оказаться стационарным

Дифференцирование



Критерий Дики-Фуллера: для исходного ряда $p = 0.3636$, для ряда первых разностей — $p = 5.2 \times 10^{-29}$.

Авторегрессия

Что если делать регрессию ряда на собственные значения в прошлом?

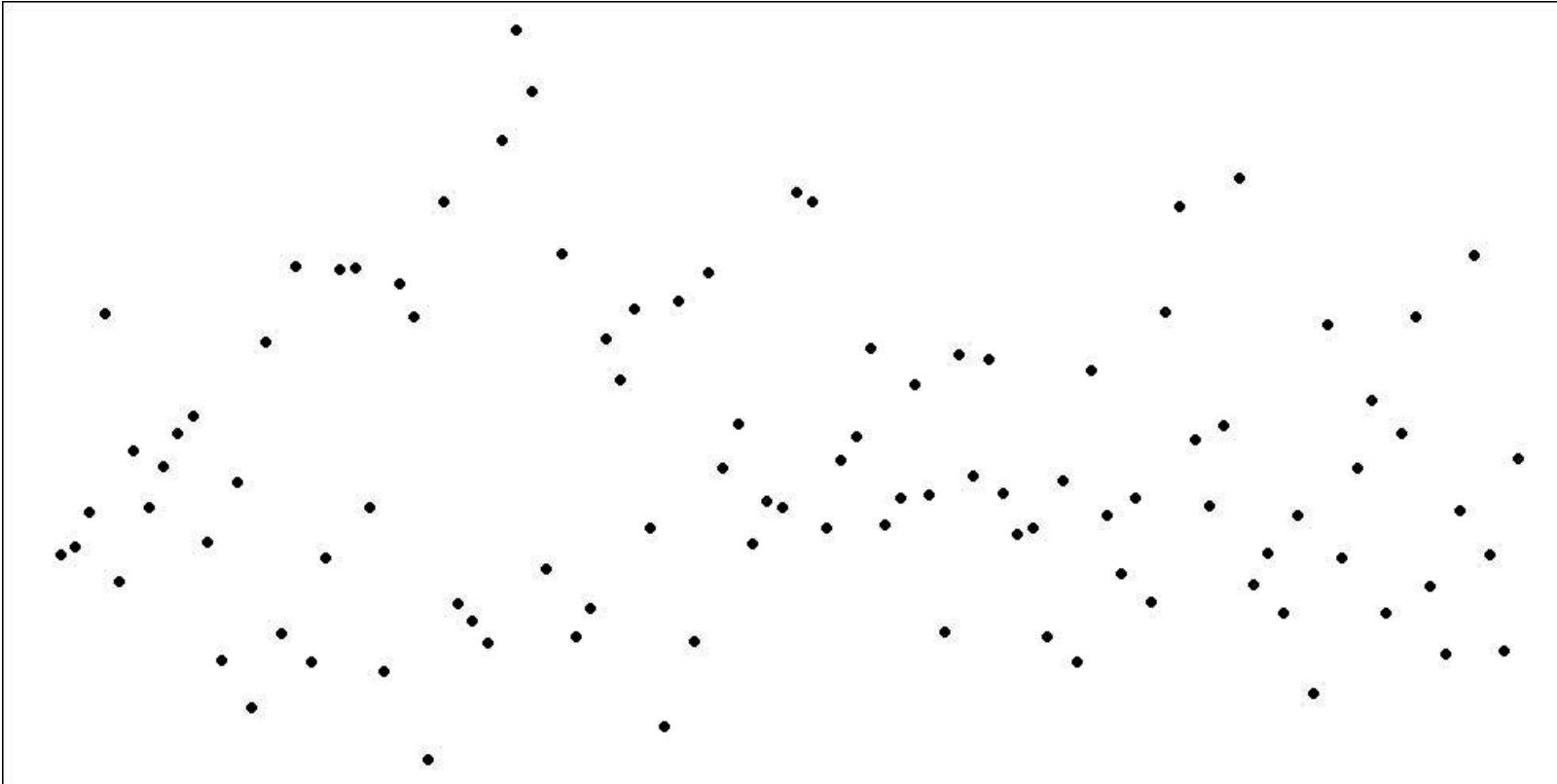
$$y_t = \alpha + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t$$

Модель авторегрессии порядка p ($AR(p)$):

y_t — линейная комбинация p предыдущих значений ряда и шумовой компоненты.

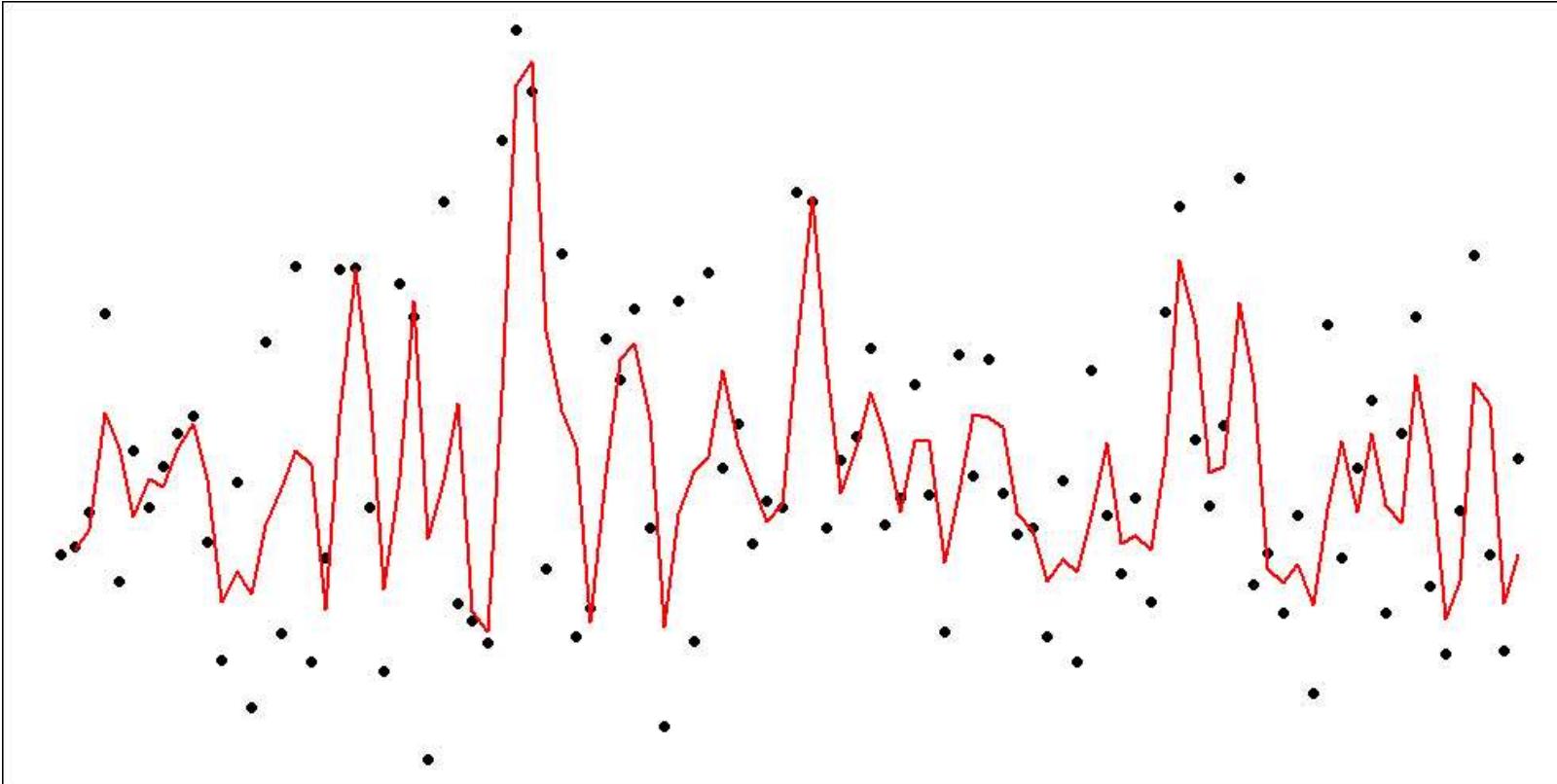
Скользящее среднее

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



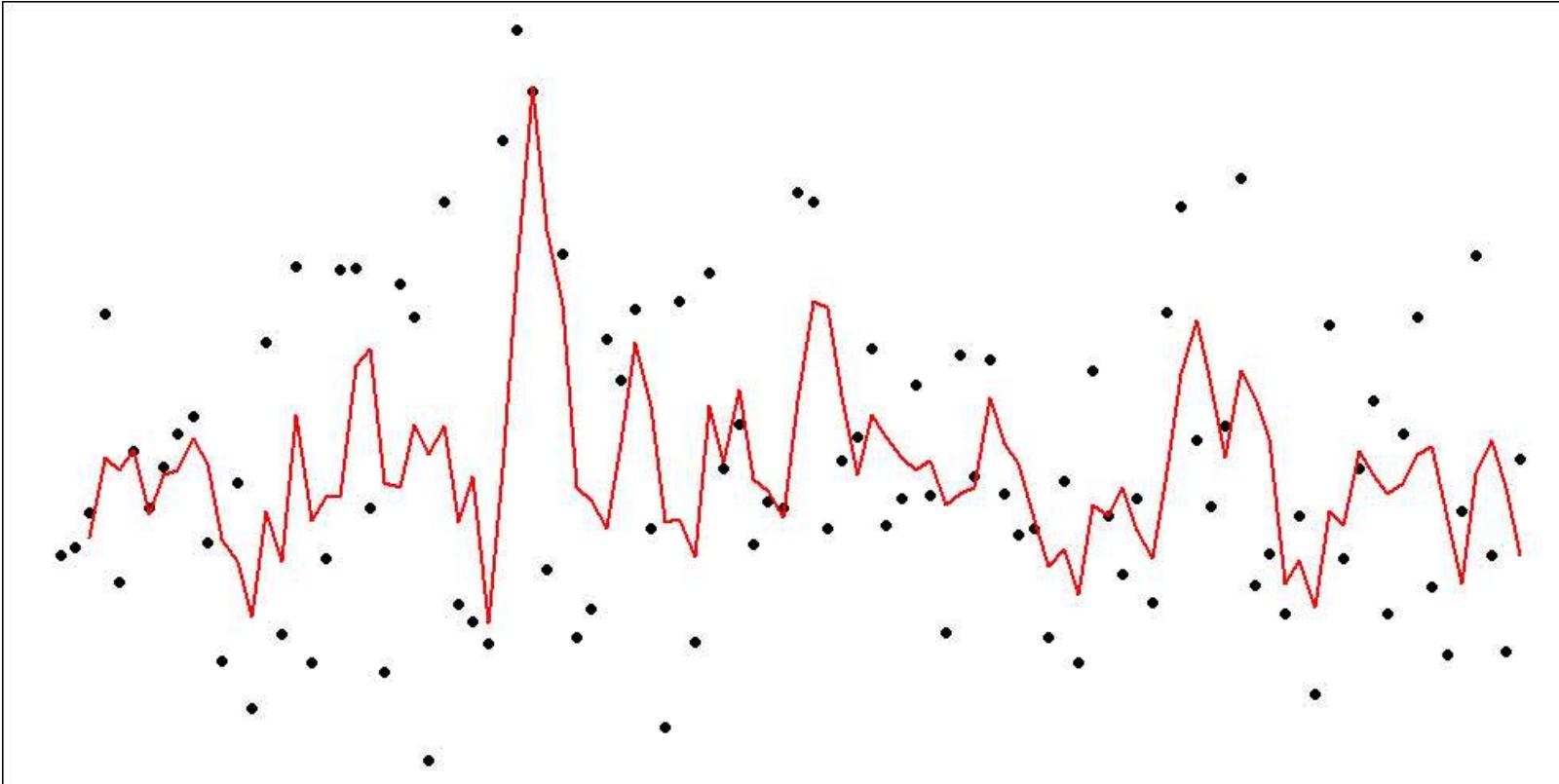
Скользящее среднее

Среднее по двум соседним точкам:



Скользящее среднее

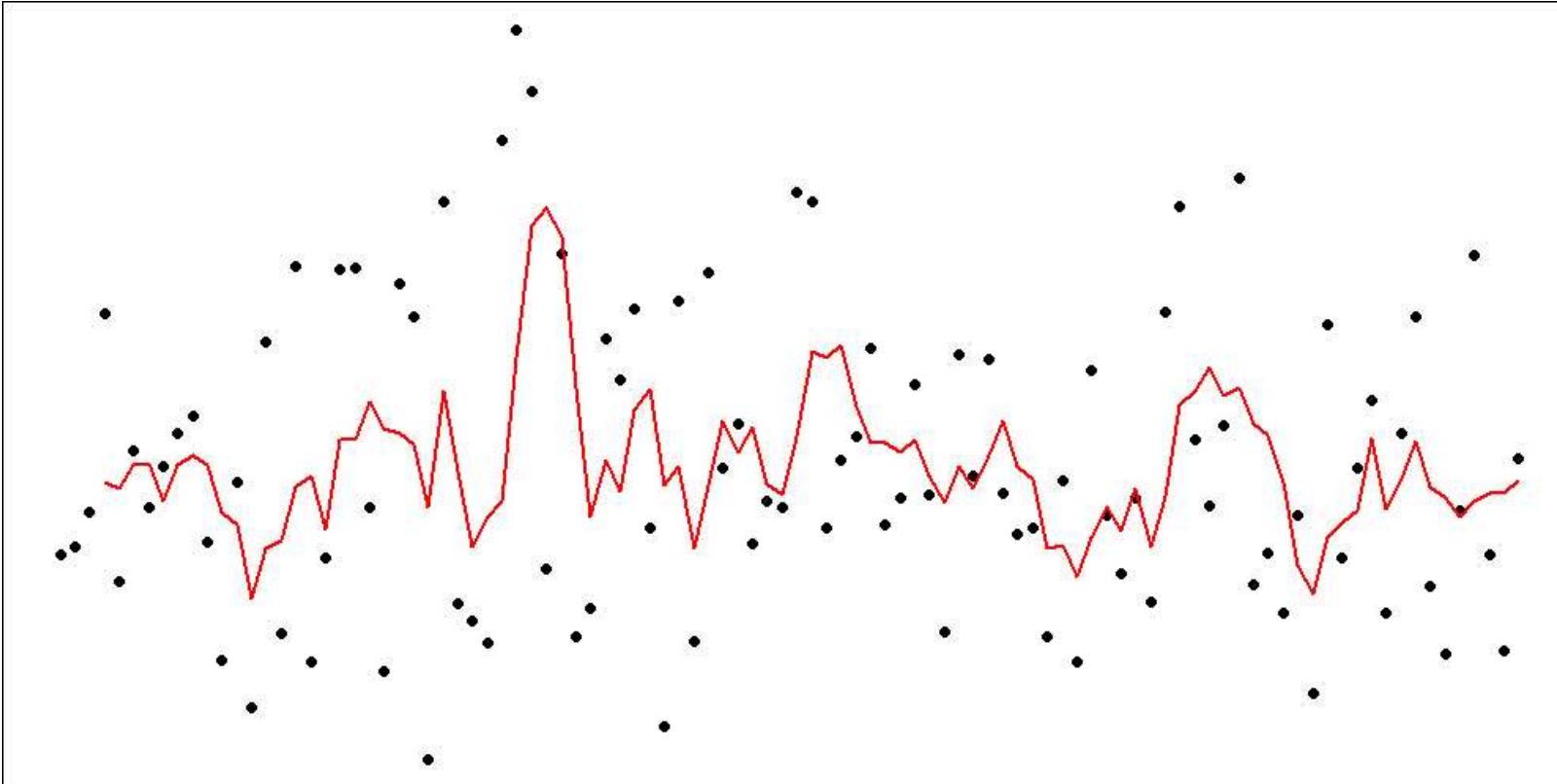
Среднее по трём соседним точкам:



Описывает какой-то ряд...

Скользящее среднее

Среднее по четырём соседним точкам:



Скользящее среднее

Обобщим и добавим веса:

$$y_t = \alpha + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q}$$

Модель скользящего среднего порядка q ($MA(q)$):

y_t — линейная комбинация q последних значений шумовой компоненты.

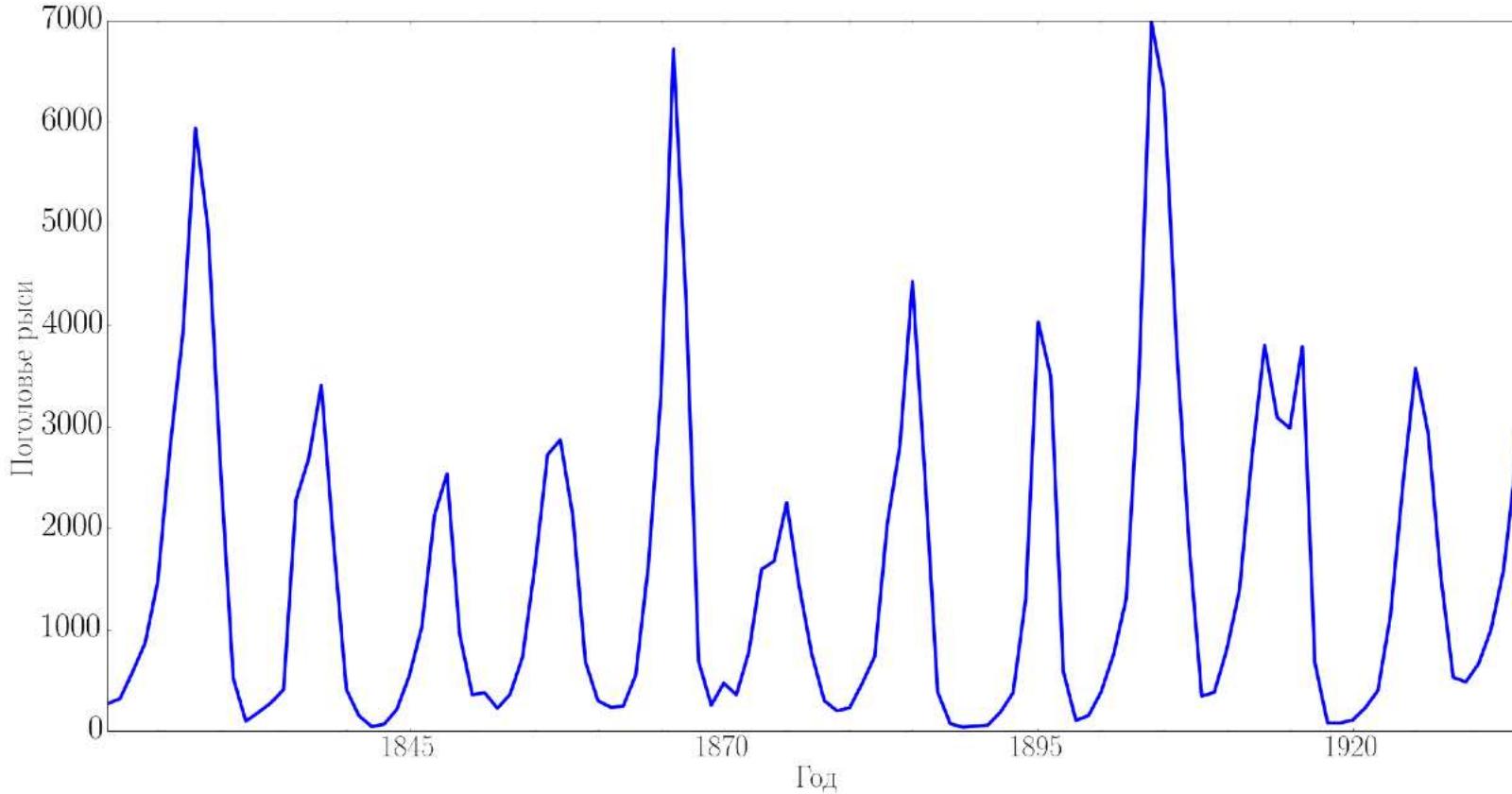
ARMA

Модель $ARMA(p, q)$:

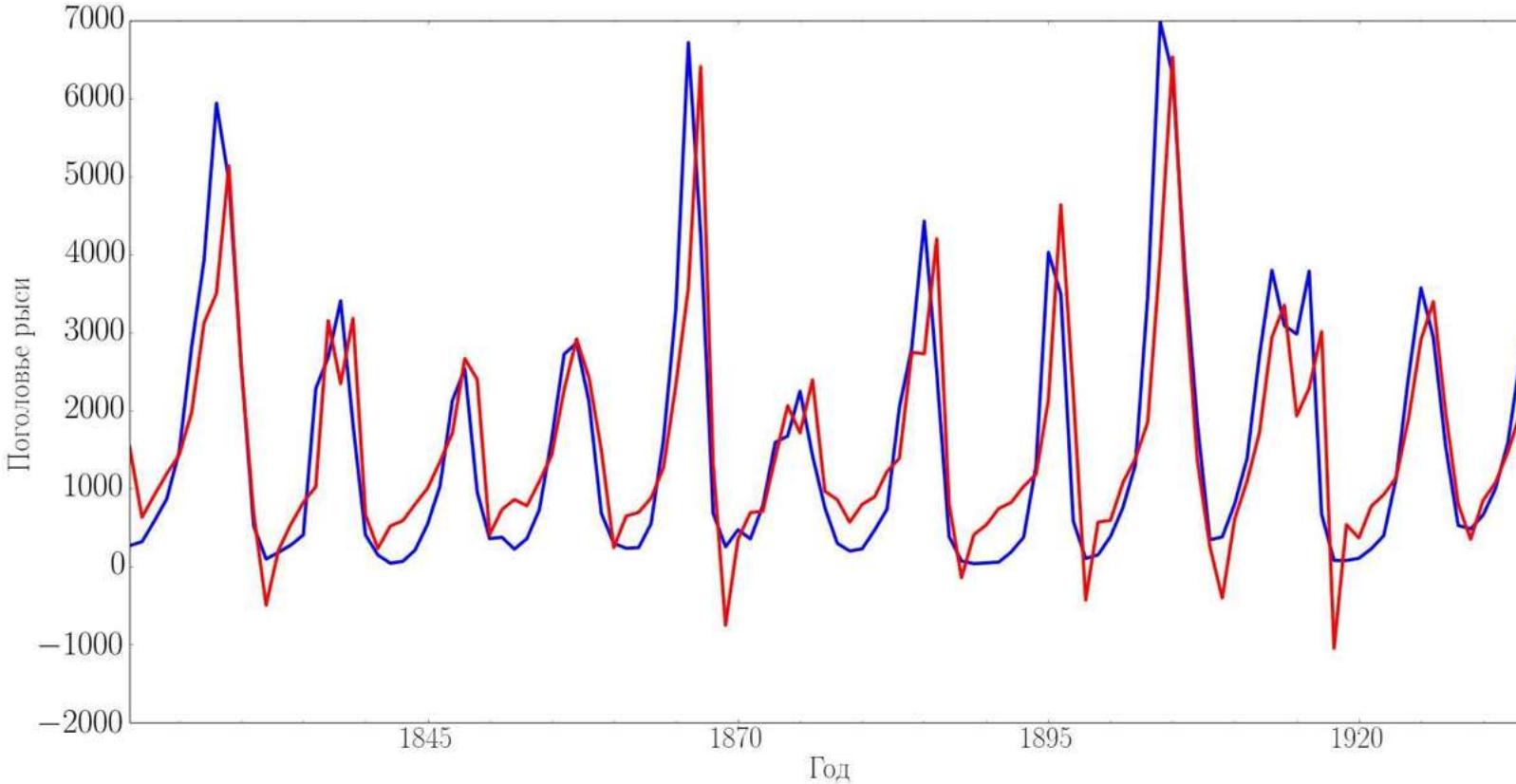
$$y_t = \alpha + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

Теорема Вольда: любой стационарный ряд может быть описан моделью $ARMA(p, q)$.

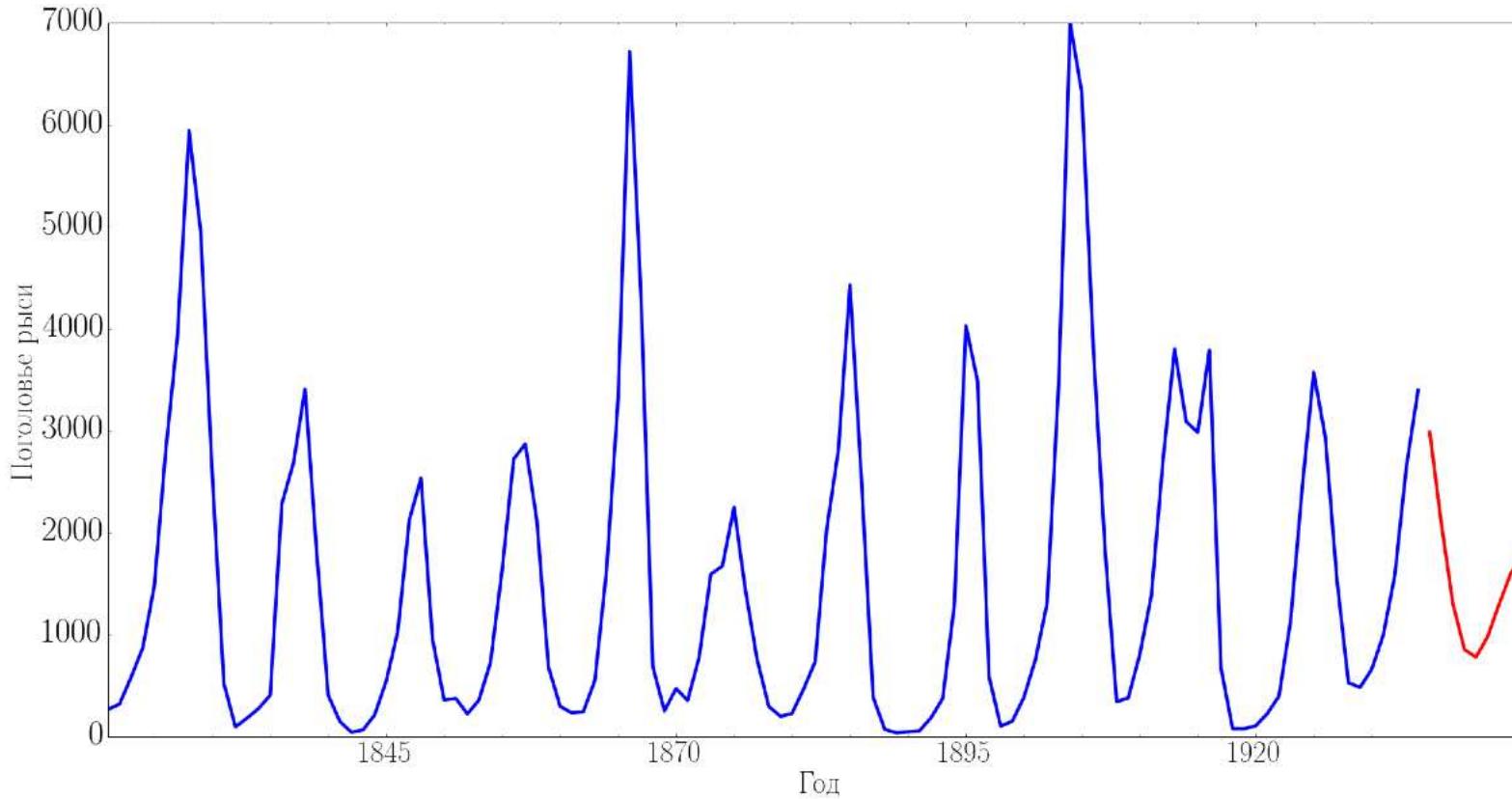
Поголовье рыси



Поголовье рыси

Модель $ARMA(2, 2)$:

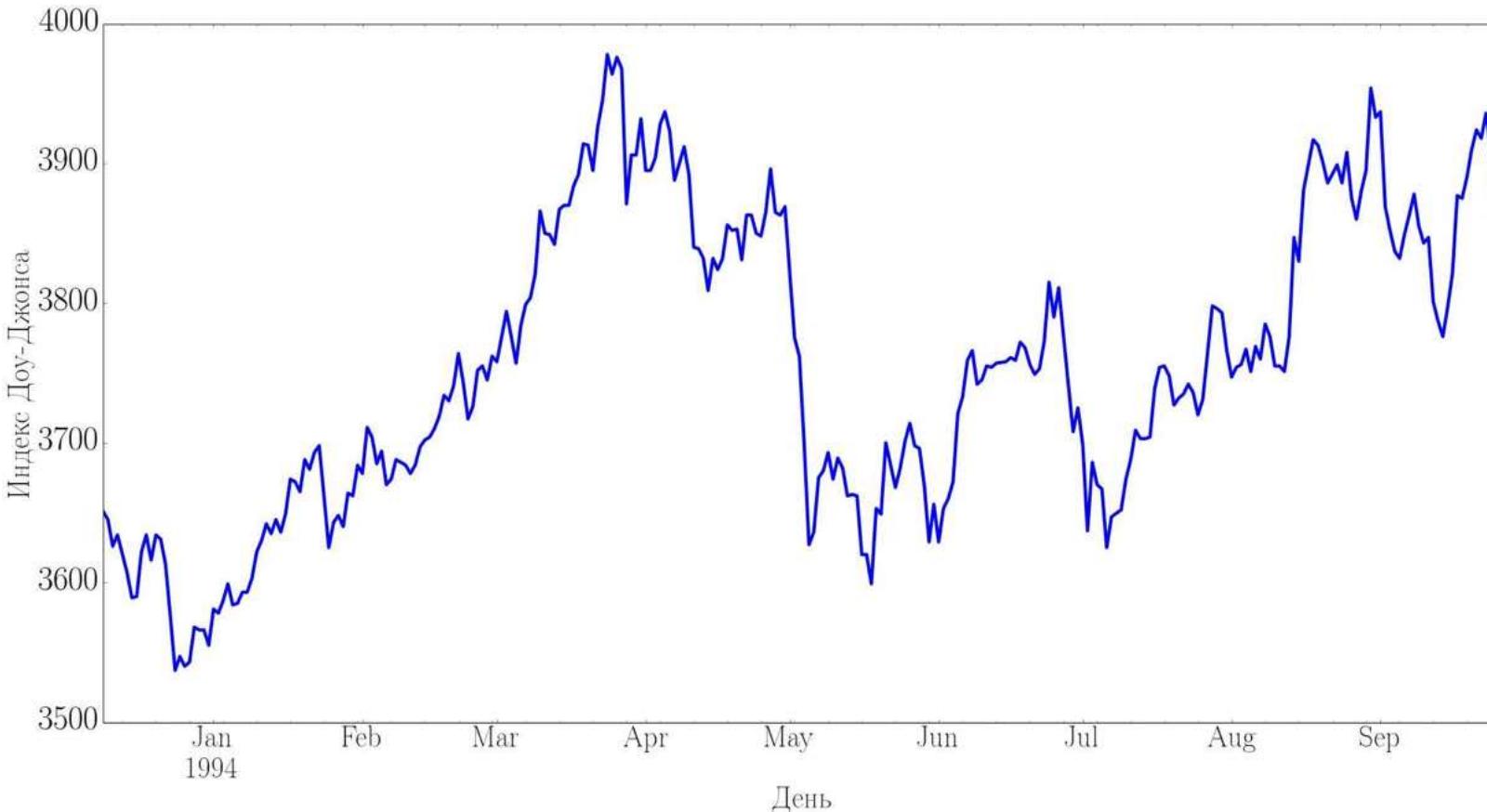
Поголовье рыси

Прогноз модели $ARMA(2, 2)$:

ARIMA

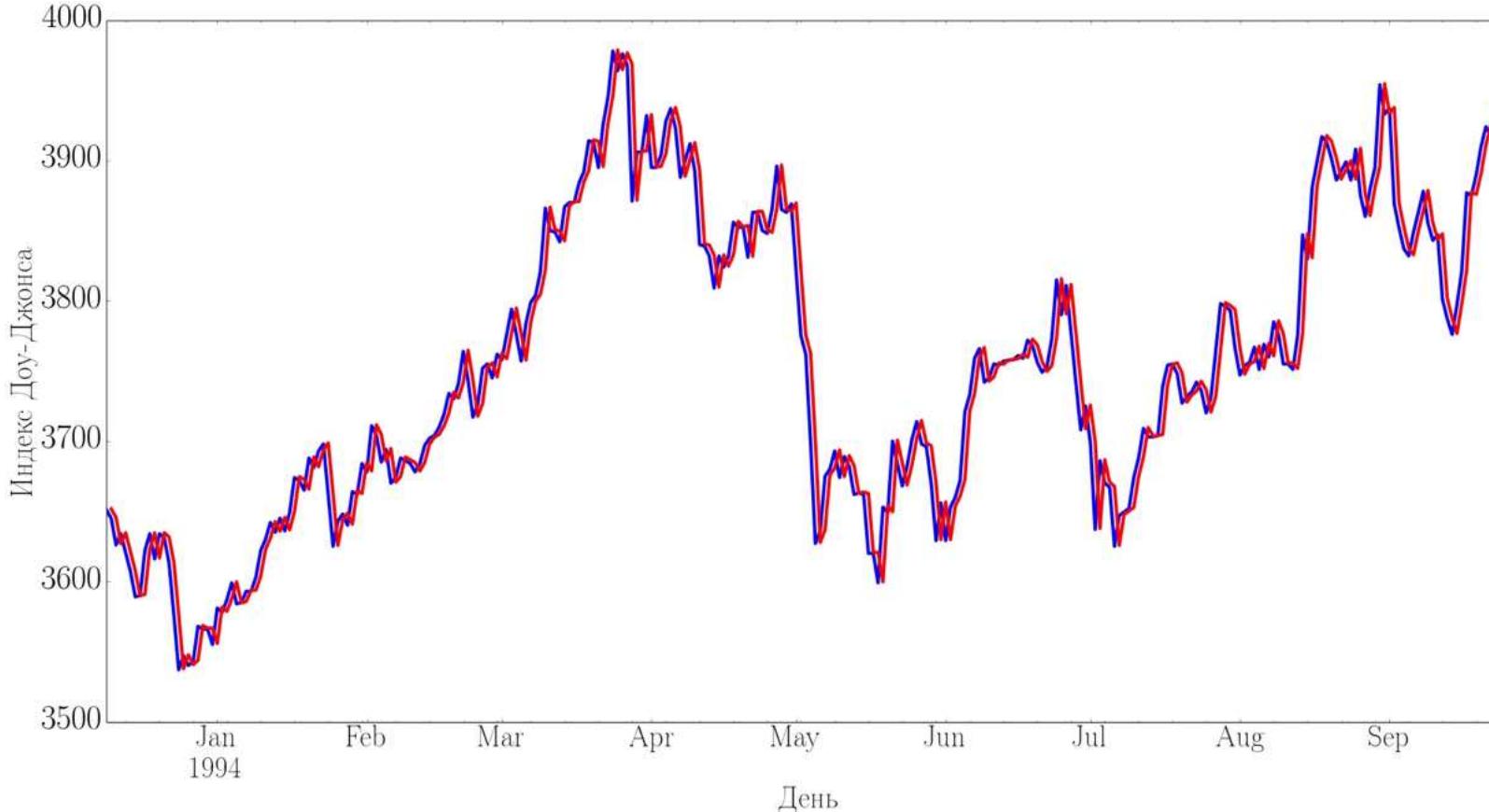
Модель $ARIMA(p, d, q)$ — модель $ARMA(p, q)$ для d раз продифференцированного ряда.

Индекс Доу-Джонса



Индекс Доу-Джонса

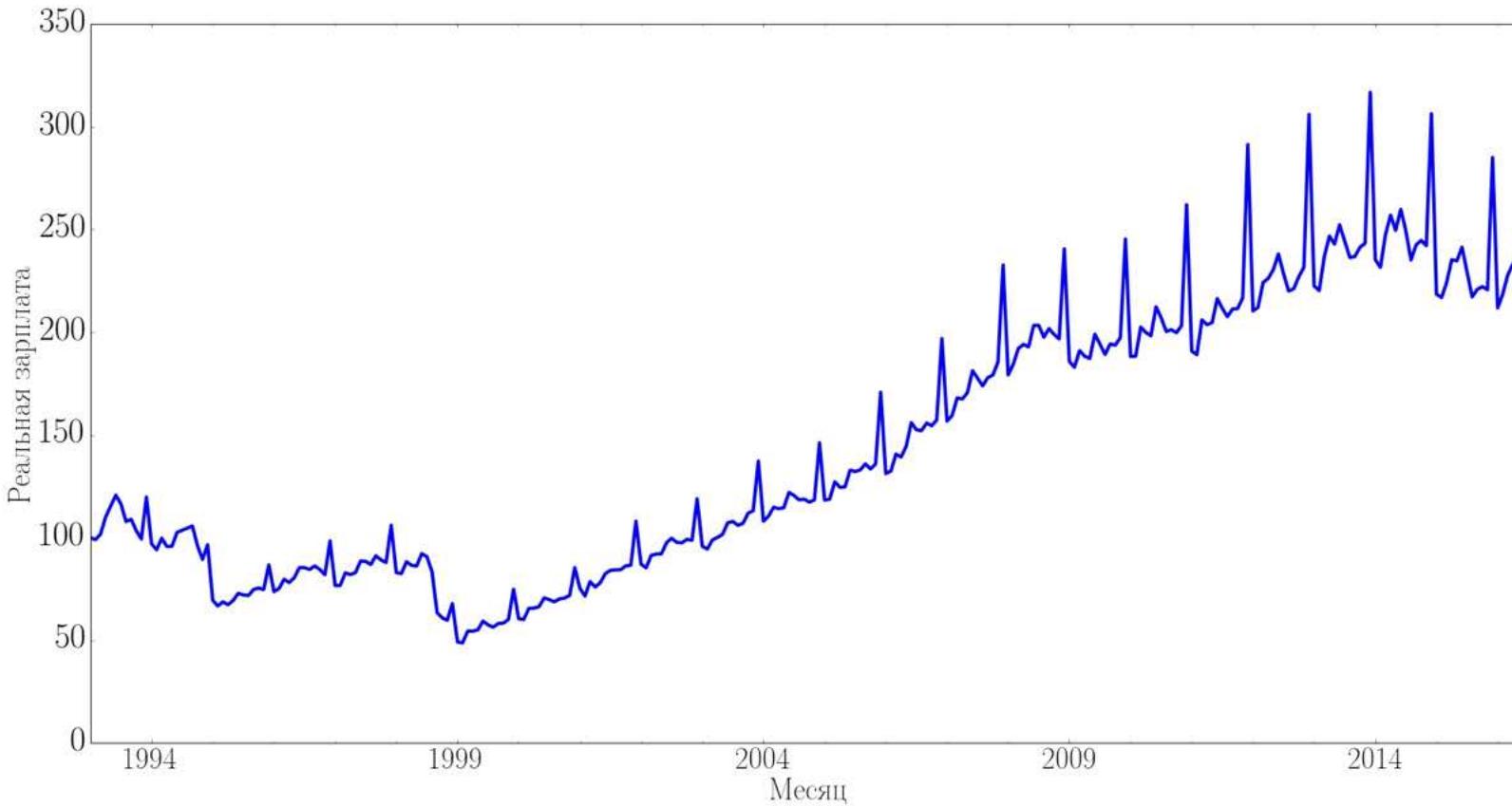
Модель $ARIMA(0, 1, 0)$:



$ARMA(0, 0) = \text{шум}$

$ARIMA(0, 1, 0) = \text{случайное блуждание (шаг с шумом)}$

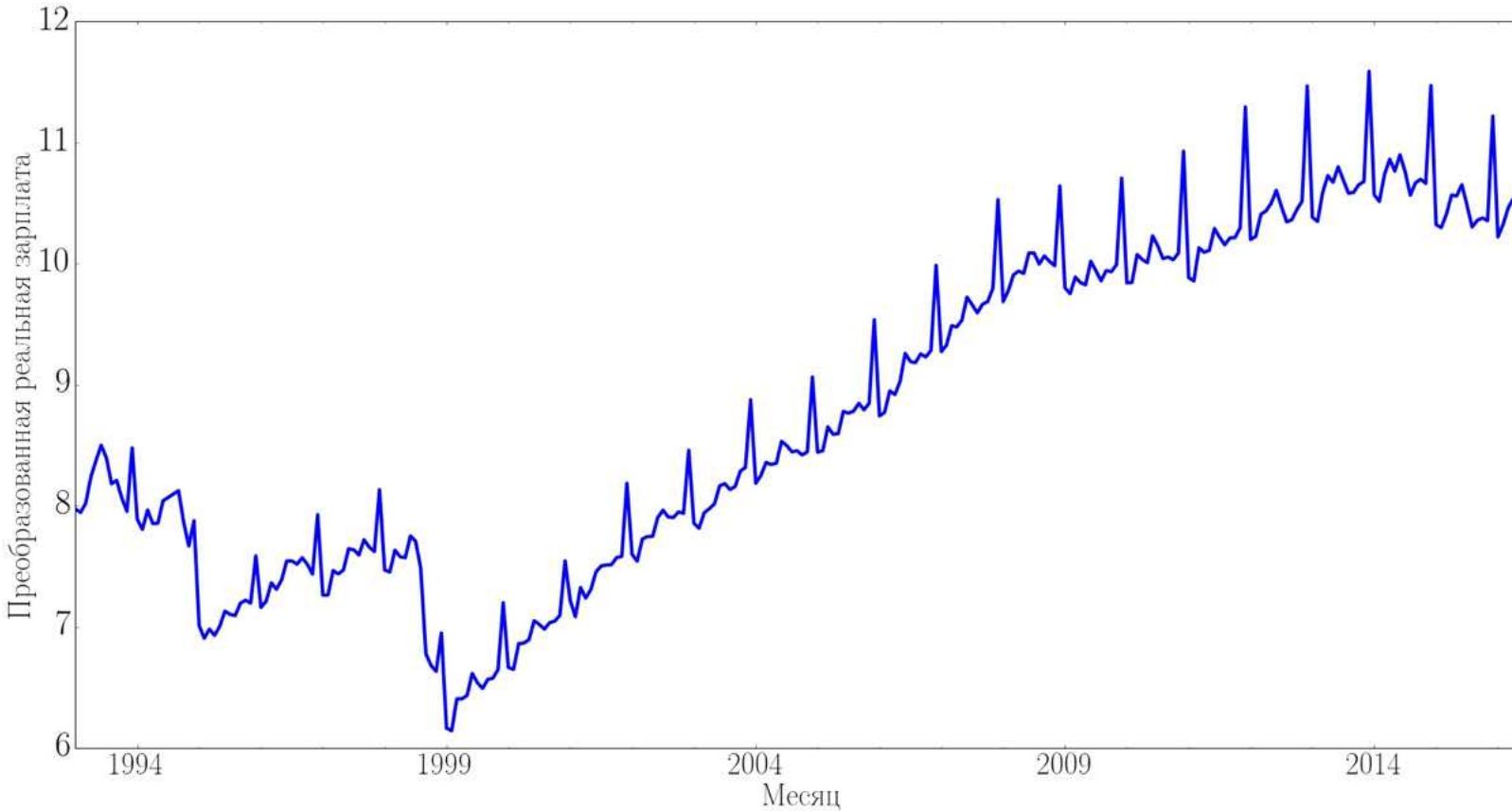
Реальная заработная плата



Критерий Дики-Фуллера: $p = 0.2265$.

Реальная заработная плата

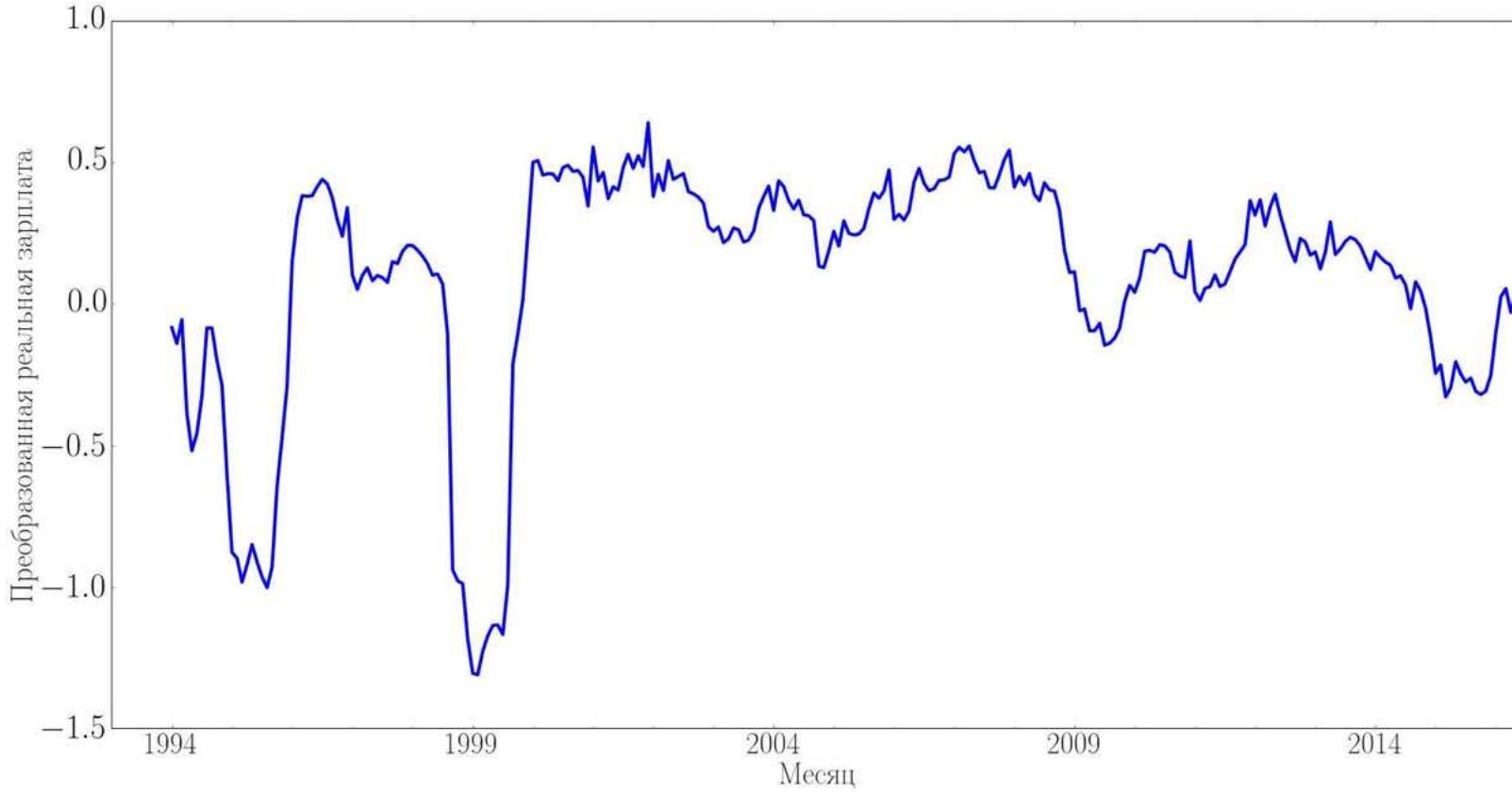
После преобразования Бокса-Кокса с $\lambda = 0.22$:



Критерий Дики-Фуллера: $p = 0.1661$.

Реальная заработная плата

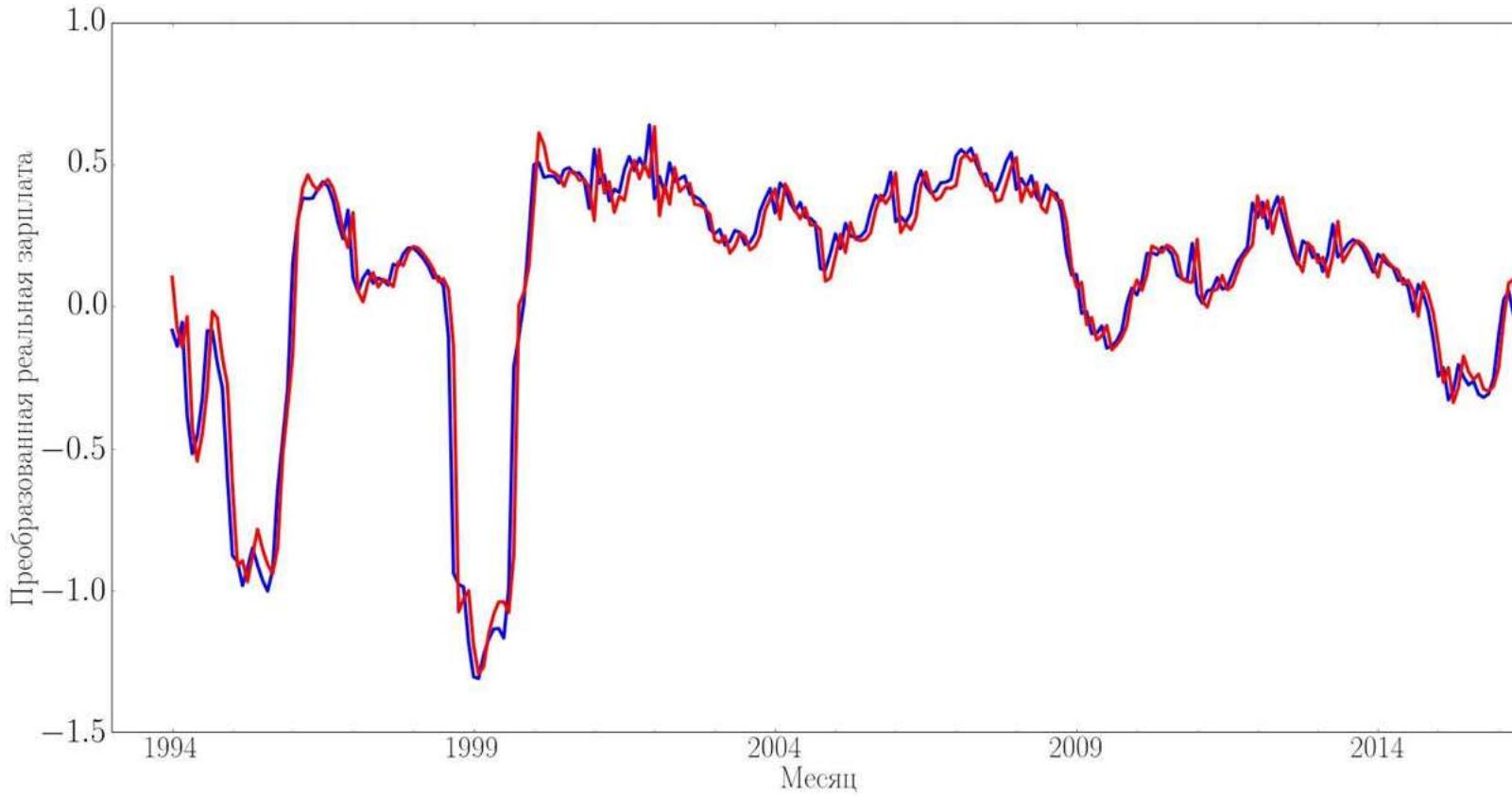
После сезонного дифференцирования:



Критерий Дики-Фуллера: $p = 0.01$.

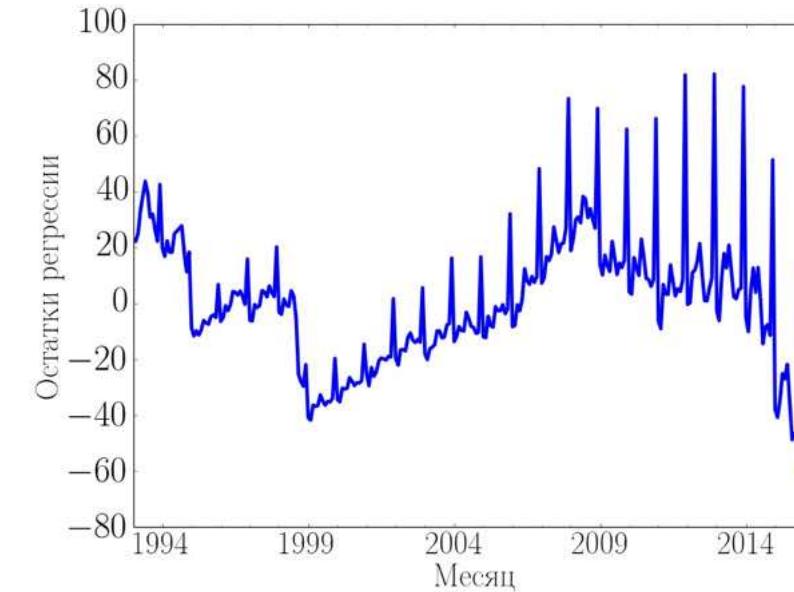
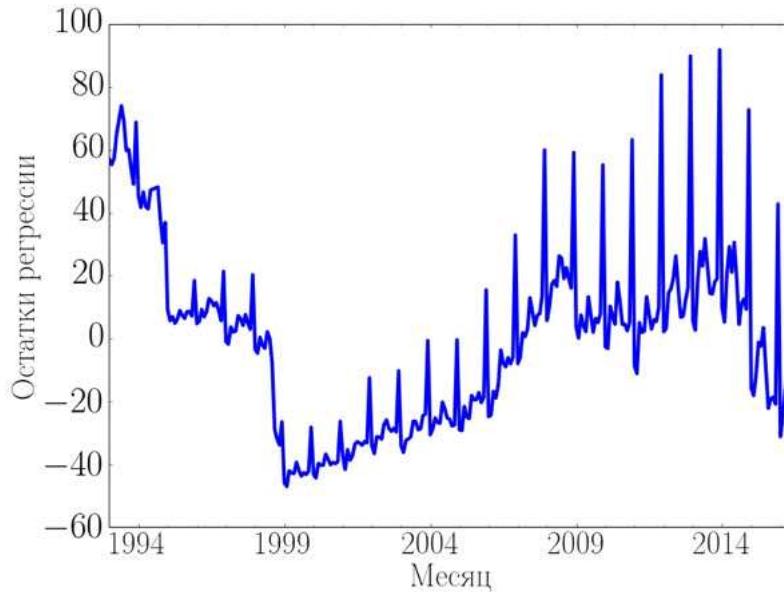
Реальная заработная плата

Модель $ARMA(2, 2)$ для преобразованного ряда:



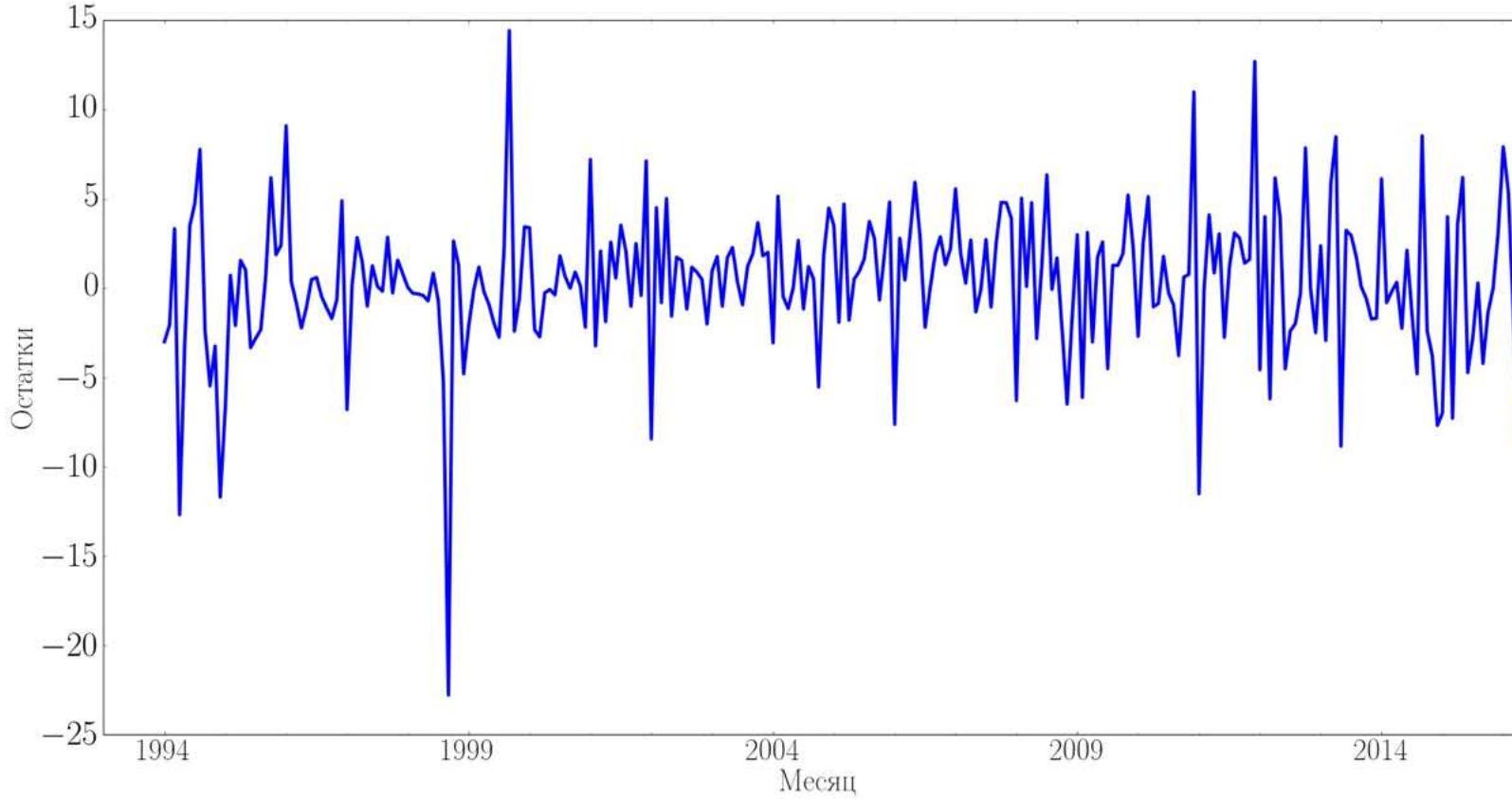
Реальная заработная плата

Остатки регрессий на время:



Реальная заработная плата

Остатки построенной модели:



Подбор параметров

- α, φ, θ
 - d, D
 - q, Q
 - p, P

Это долгая история...

Прогнозирование

$$y_t = \hat{\alpha} + \hat{\phi}_1 y_{t-1} + \cdots + \hat{\phi}_p y_{t-p} + \varepsilon_t + \hat{\theta}_1 \varepsilon_{t-1} + \cdots + \hat{\theta}_q \varepsilon_{t-q}$$

Заменяем t на $T+1$:

$$\hat{y}_{T+1|T} = \hat{\alpha} + \hat{\phi}_1 y_T + \cdots + \hat{\phi}_p y_{T+1-p} + \varepsilon_{T+1} + \hat{\theta}_1 \varepsilon_T + \cdots + \hat{\theta}_q \varepsilon_{T+1-q}$$

Заменяем будущие ошибки на нули:

$$\hat{y}_{T+1|T} = \hat{\alpha} + \hat{\phi}_1 y_T + \cdots + \hat{\phi}_p y_{T+1-p} + \hat{\theta}_1 \varepsilon_T + \cdots + \hat{\theta}_q \varepsilon_{T+1-q}$$

Заменяем прошлые ошибки на остатки:

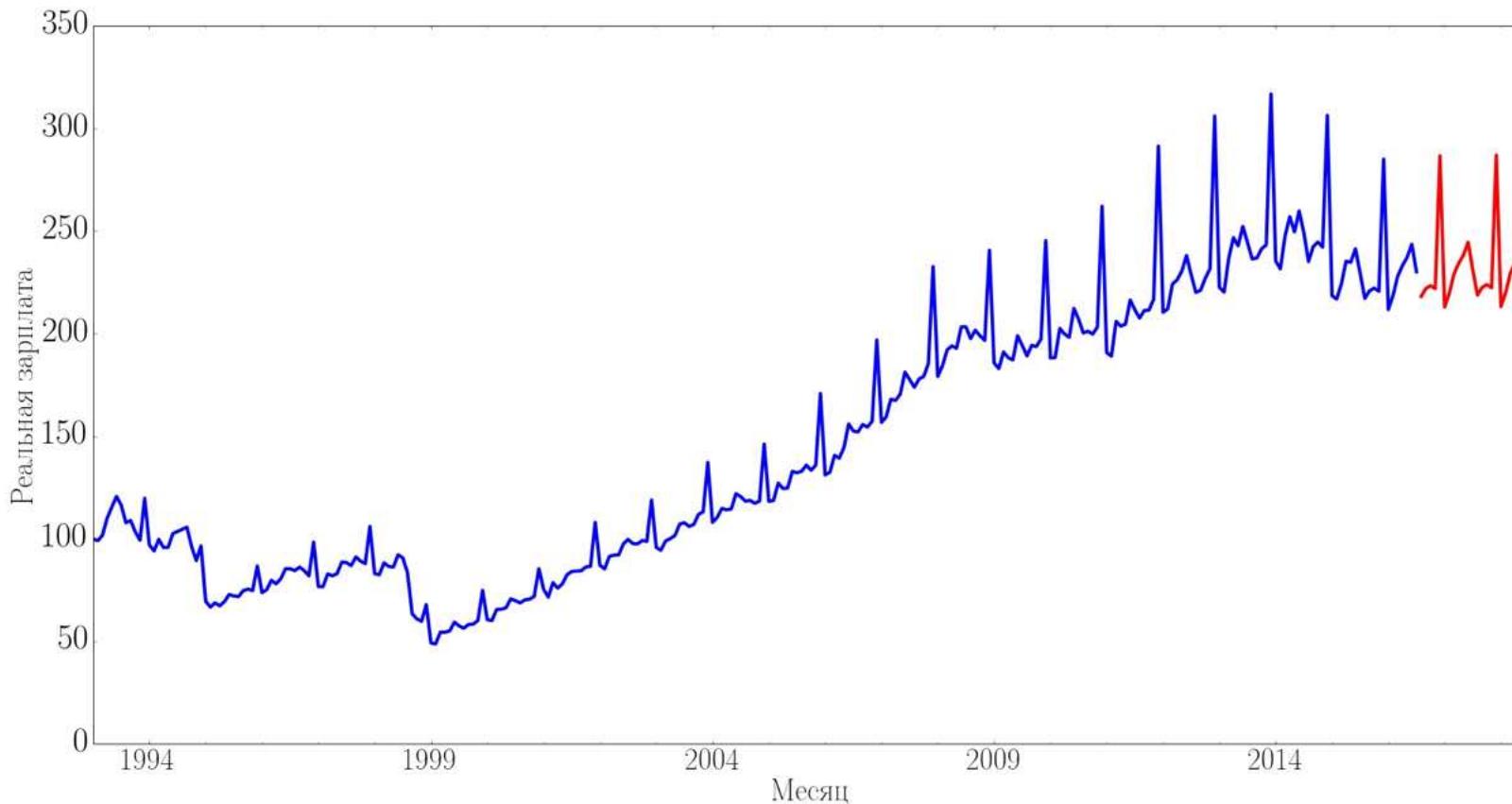
$$\hat{y}_{T+1|T} = \hat{\alpha} + \hat{\phi}_1 y_T + \cdots + \hat{\phi}_p y_{T+1-p} + \hat{\theta}_1 \hat{\varepsilon}_T + \cdots + \hat{\theta}_q \hat{\varepsilon}_{T+1-q}$$

Если мы прогнозируем на момент времени $T+2$, в формуле появляется значение ряда из будущего:

$$\hat{y}_{T+2|T} = \hat{\alpha} + \hat{\phi}_1 \color{red}{y_{T+1}} + \cdots + \hat{\phi}_p y_{T+2-p} + \hat{\theta}_1 \hat{\varepsilon}_{T+1} + \cdots + \hat{\theta}_q \hat{\varepsilon}_{T+2-q}$$

Заменяем его на прогноз $\hat{y}_{T+1|T}$.

Прогнозирование



Остатки

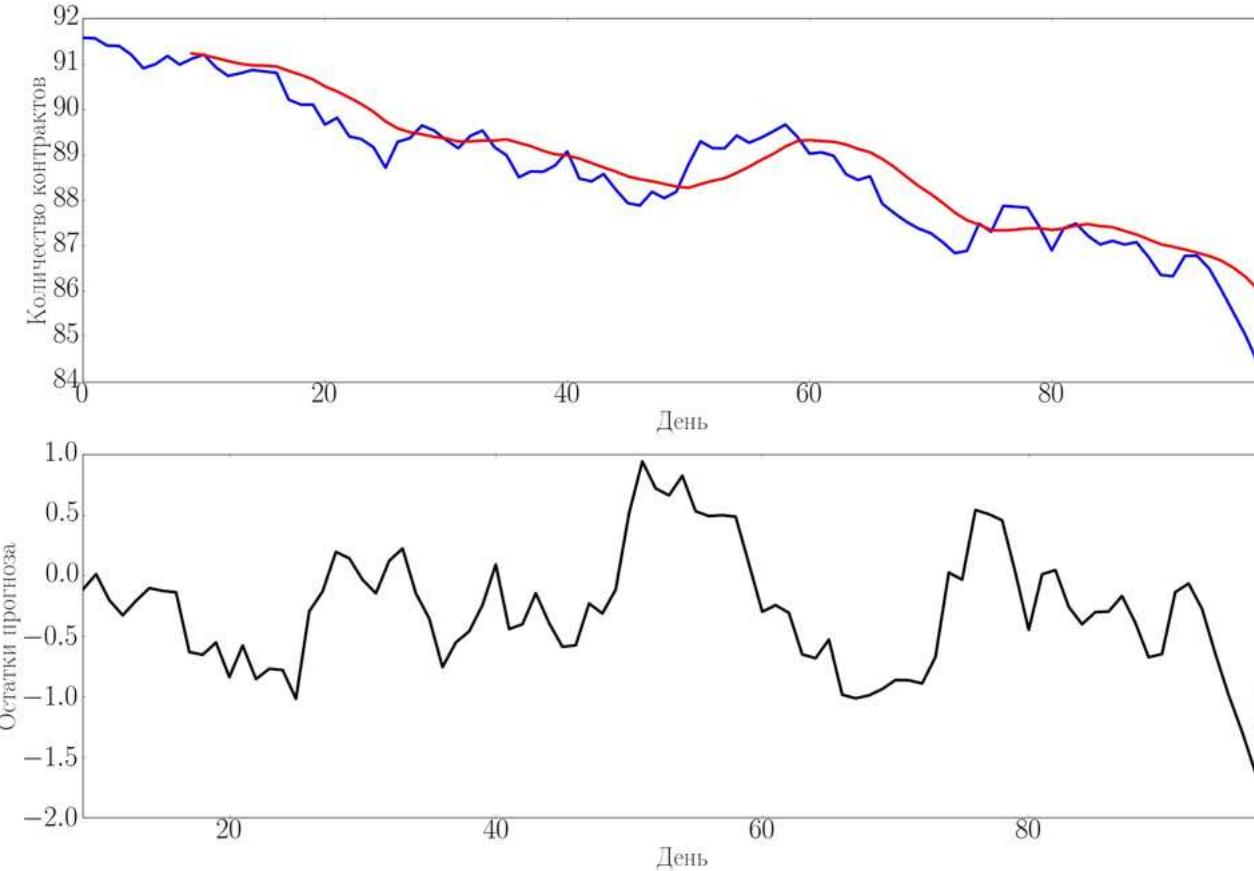
Остатки — разность между фактом и прогнозом:

$$\hat{\varepsilon}_t = y_t - \hat{y}_{t|t-1}.$$

Нужно проверять, обладают ли они некоторыми свойствами.

Несмешённость

Несмешённость — равенство среднего значения нулю:

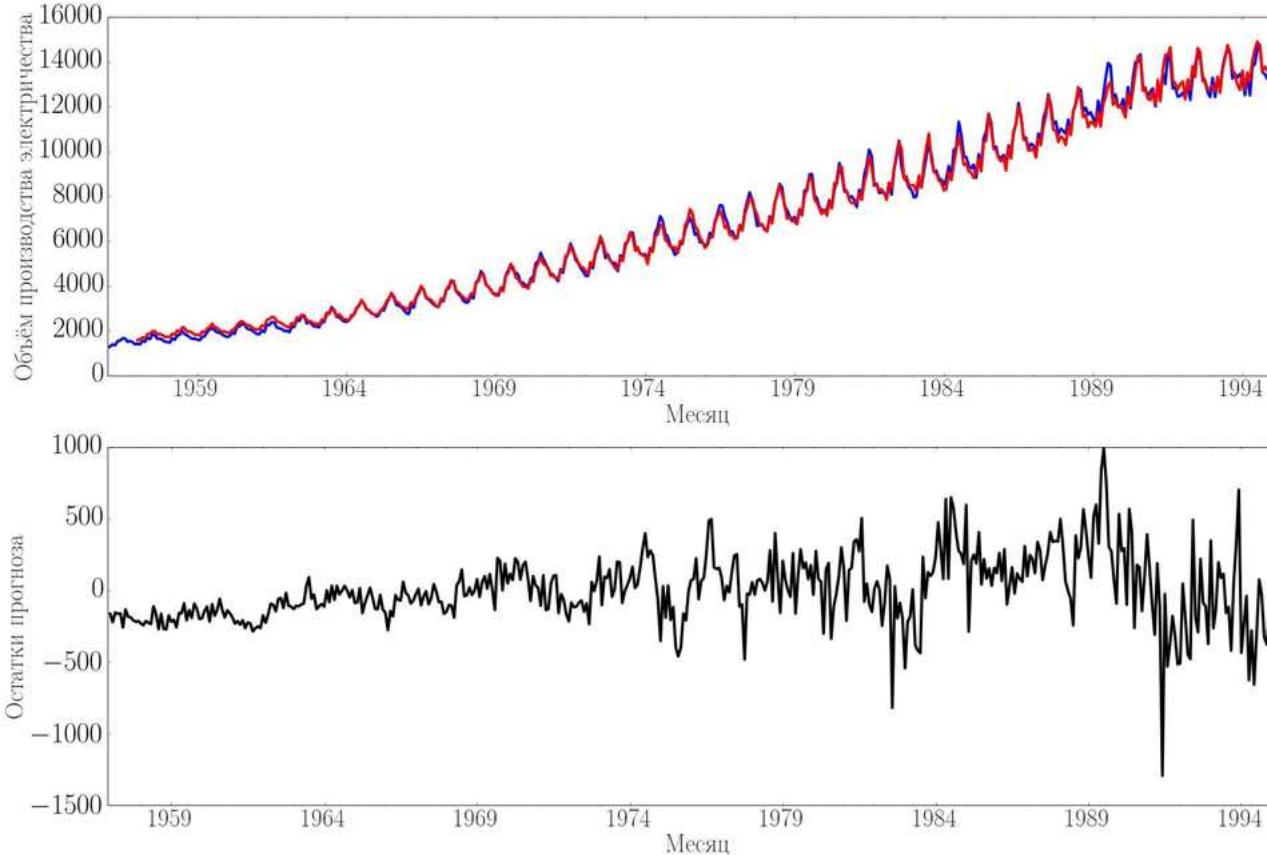


Несмешённость

- Можно проверить гипотезу $H_0 : \varepsilon = 0$ с помощью критерия Стьюдента или Уилкоксона
- Если не выполняется, с моделью что-то серьёзно не так (необходим визуальный анализ)

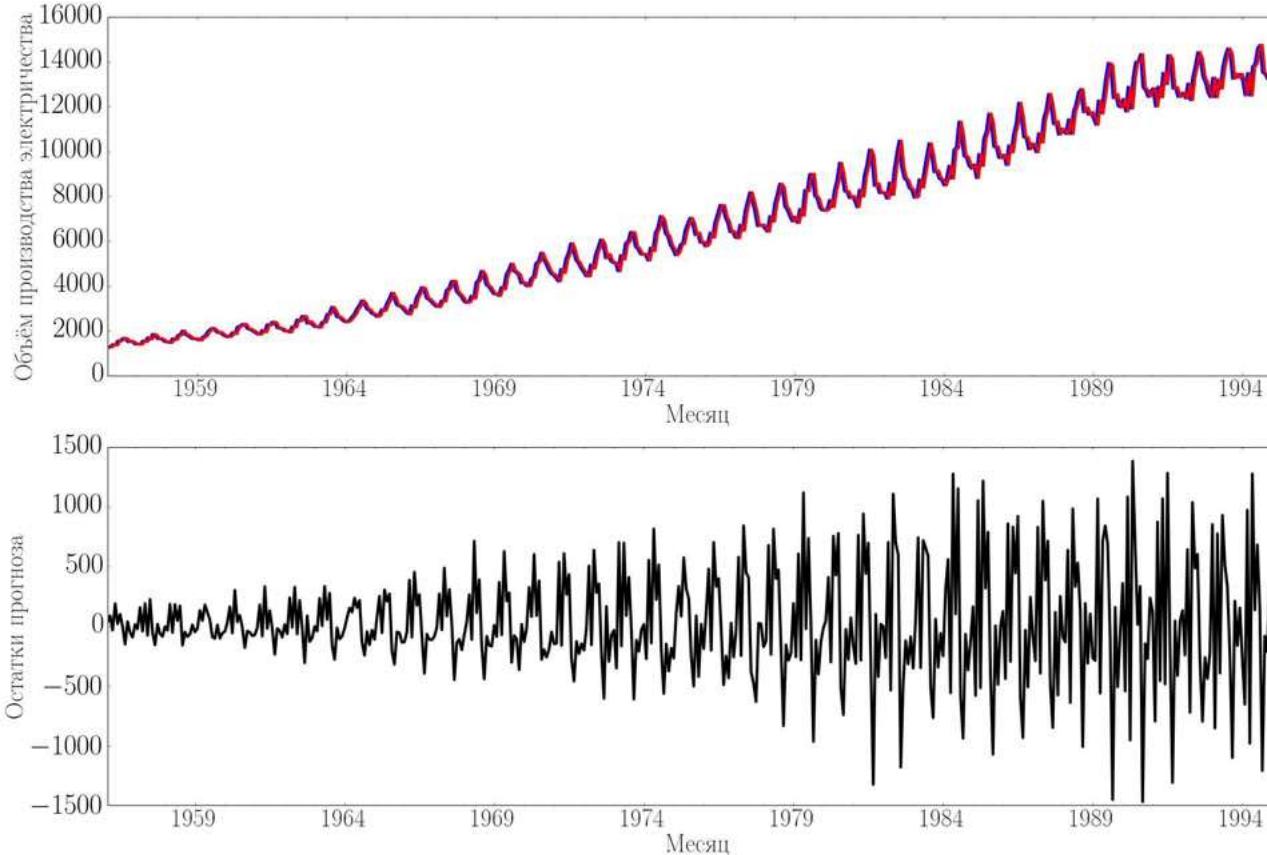
Стационарность

Стационарность — отсутствие зависимости от времени:



Стационарность

Стационарность — отсутствие зависимости от времени:

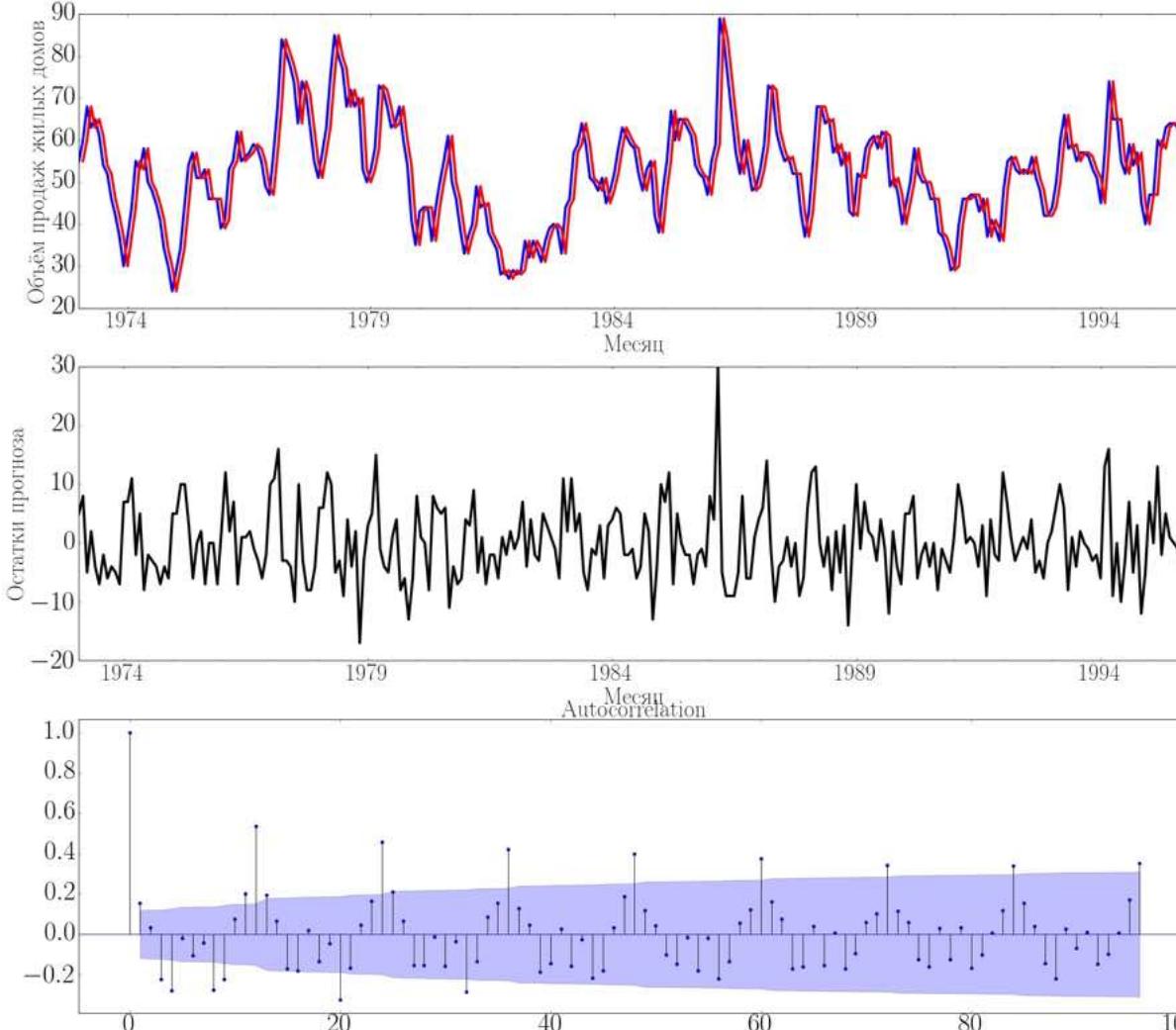


Стационарность

- Можно проверить с помощью критерия Дики-Фуллера
- Если не выполняется, значит, модель не одинаково точна в разные периоды (необходим визуальный анализ)

Неавтокоррелированность

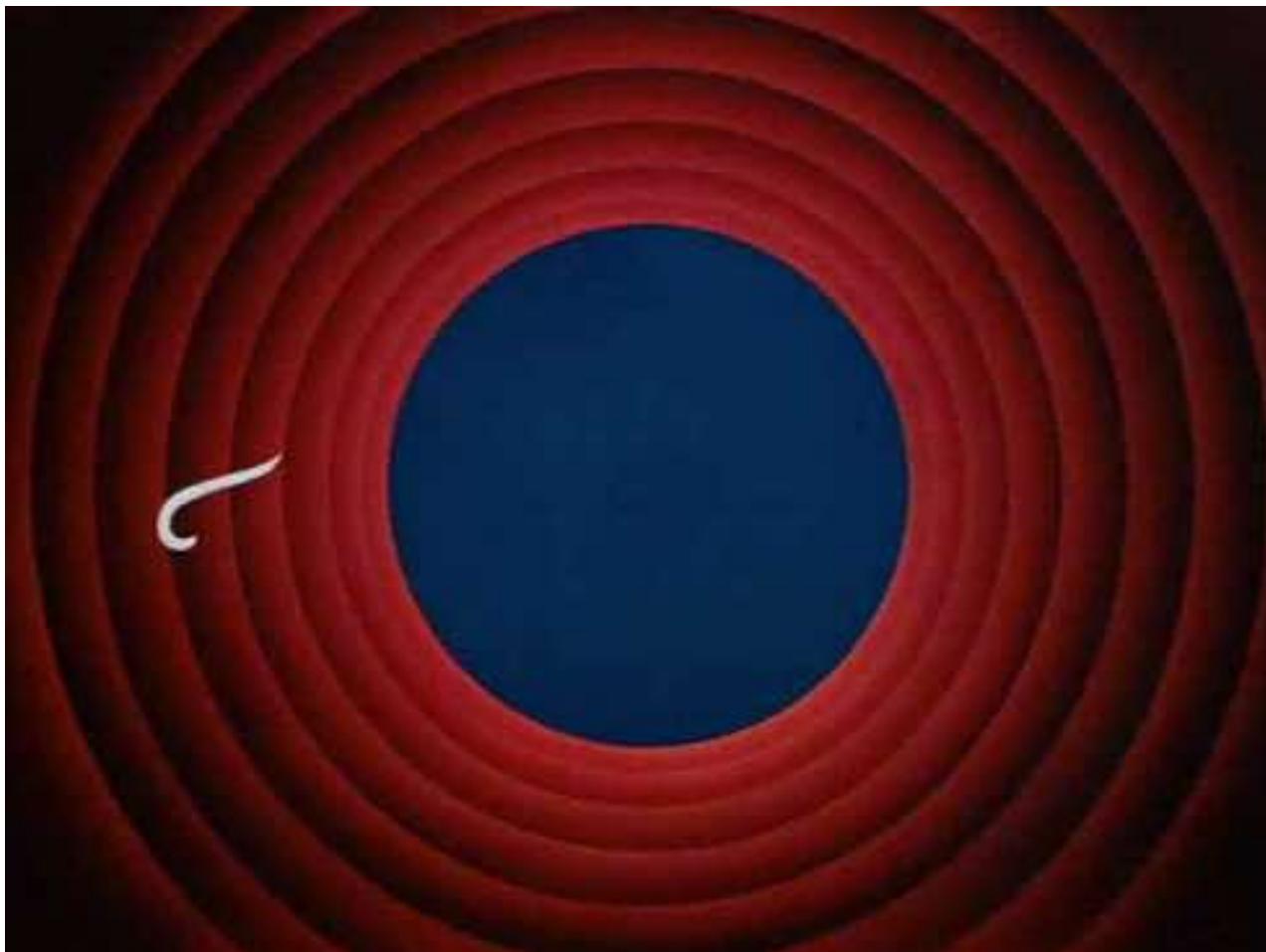
Неавтокоррелированность — отсутствие зависимости от предыдущих наблюдений:



Неавтокоррелированность

- Можно проверить на коррелограмме и с помощью Q-критерия Льюнга-Бокса
- Если не выполняется, значит, модель учитывает не все особенности данных — возможно, её можно улучшить

Литература



Hyndman R.J., Athanasopoulos G. *Forecasting: principles and practice*.—
OTexts, <https://www.otexts.org/book/fpp>