

MML minor #13

Прогнозирование временных рядов

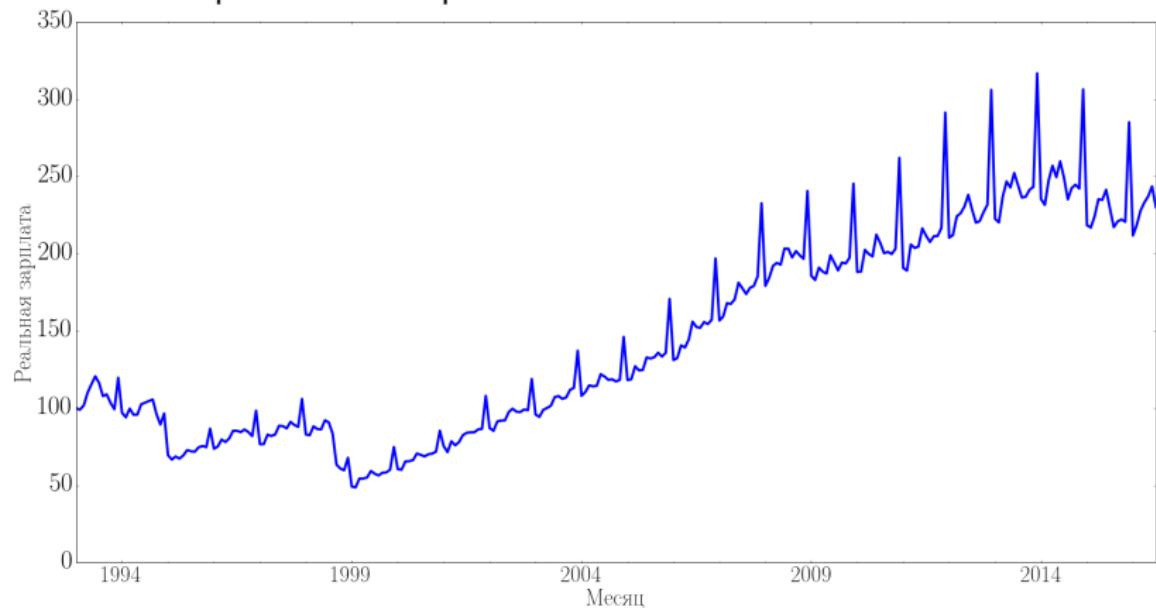
Thanks

Рябенко Евгений
riabenko.e@gmail.com

о дс

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

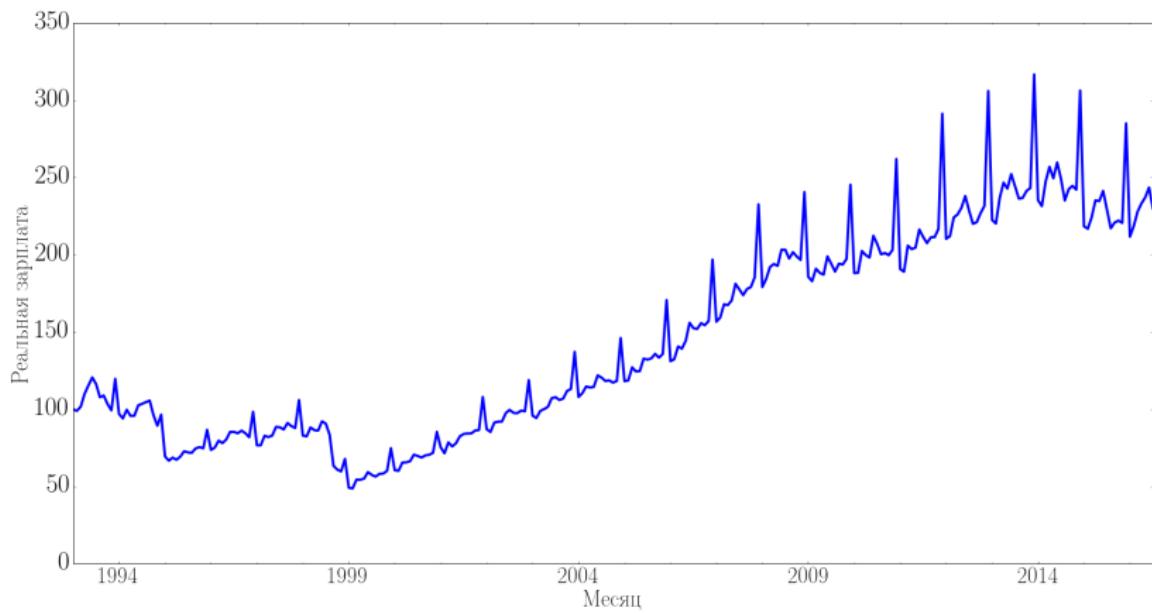
- В классических задачах анализа данных предполагается независимость наблюдений
 - При прогнозировании временных рядов, наоборот, мы надеемся, что значения ряда в прошлом содержат информацию о его поведении в будущем

Временные ряды

ARIMA

oooooooooooooooooooo

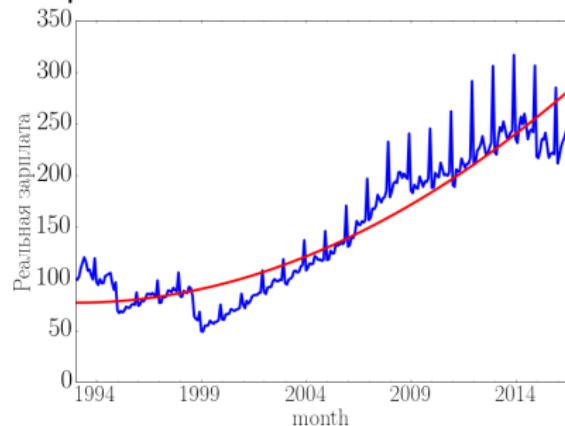
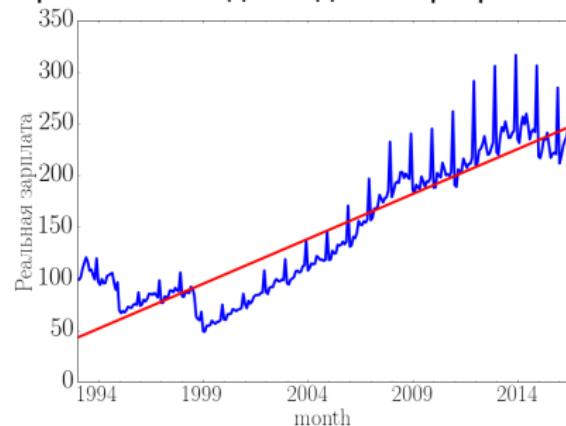
He i.i.d.



Это явно не случайная выборка!

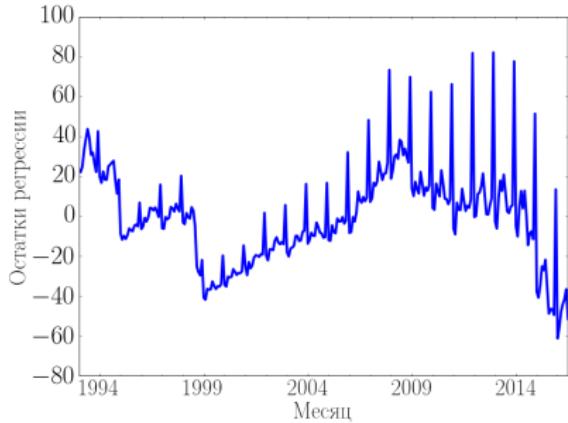
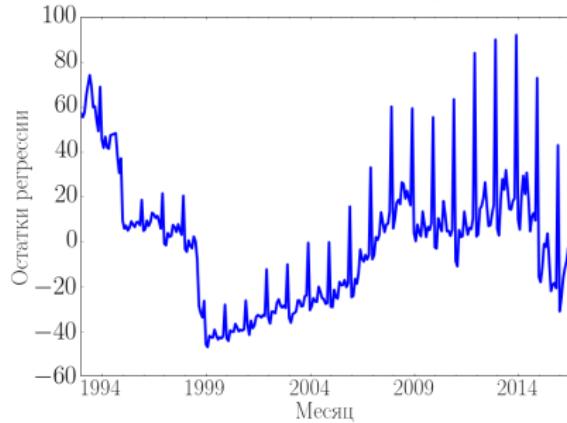
Регрессия?

Простейшая идея: сделать регрессию на время.



Регрессия?

Остатки не выглядят как шум:



Компоненты временных рядов

Тренд — плавное долгосрочное изменение уровня ряда.

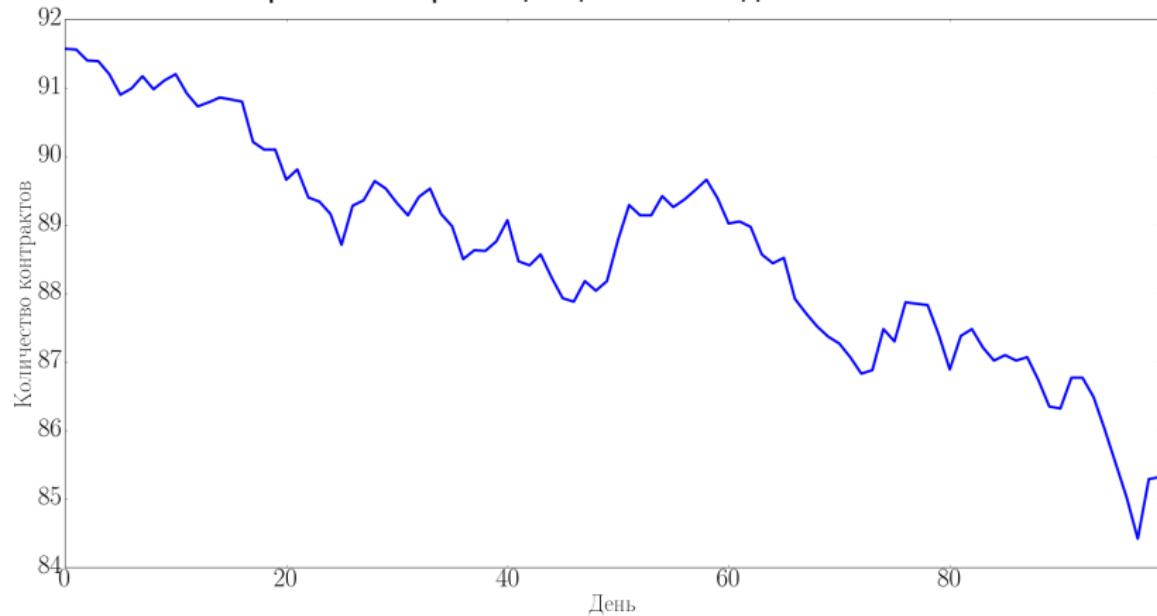
Сезонность — циклические изменения уровня ряда с постоянным периодом.

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

Ошибка — непрогнозируемая случайная компонента ряда.

Компоненты временных рядов

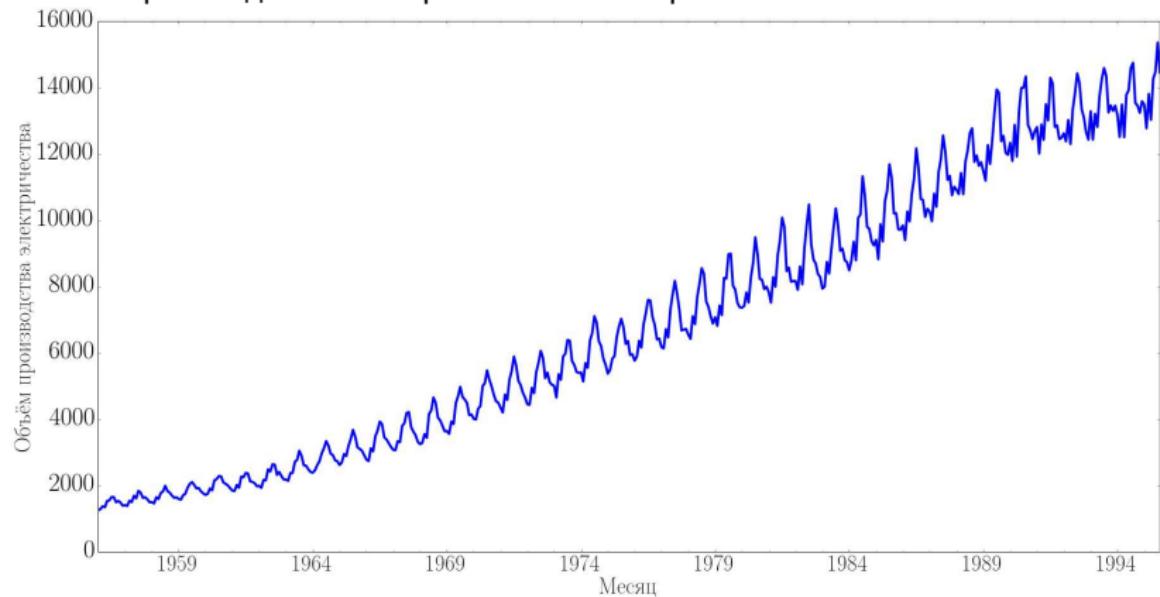
Количество контрактов сокровищницы США в день:



Тренд

Компоненты временных рядов

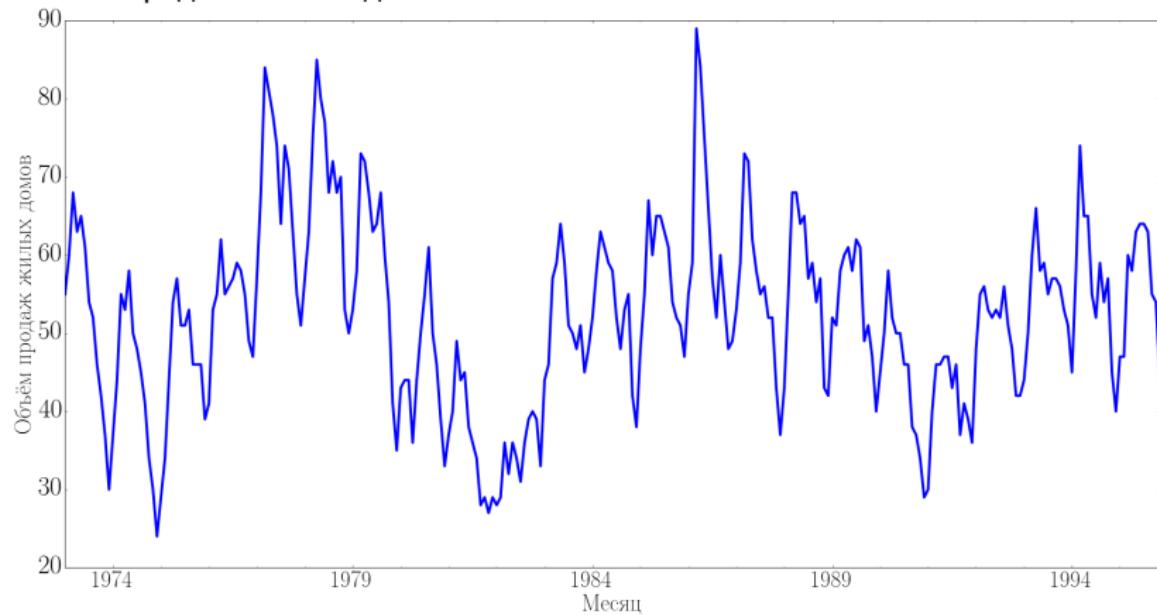
Объём производства электричества в Австралии:



Тренд, годовая сезонность

Компоненты временных рядов

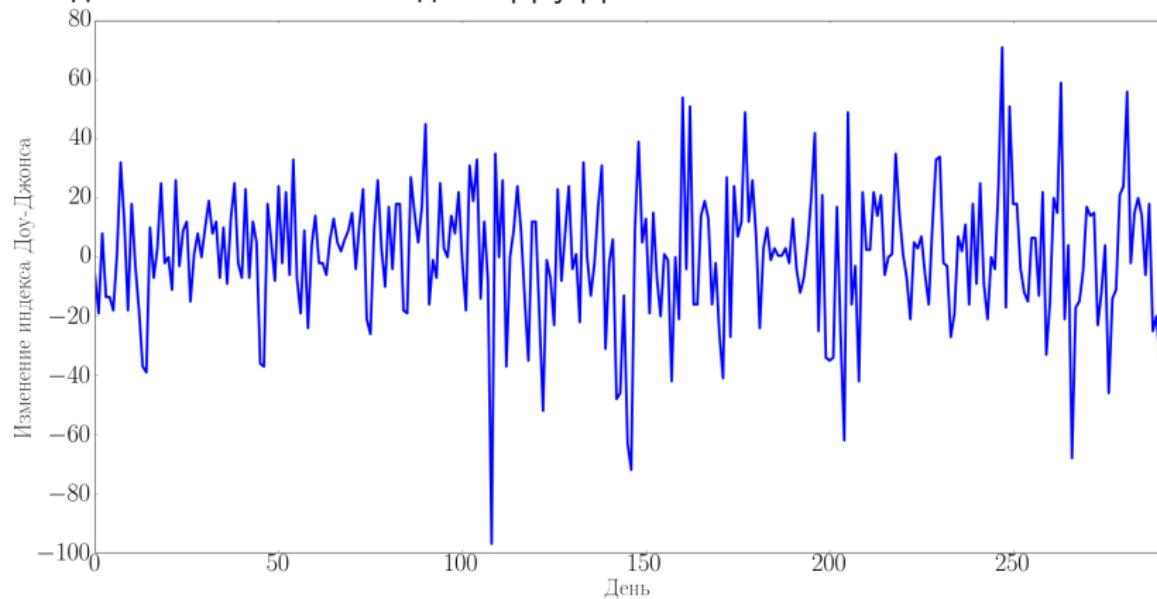
Объём продаж жилых домов:



Годовая сезонность, экономические циклы

Компоненты временных рядов

Ежедневные изменения индекса Доу-Джонса:



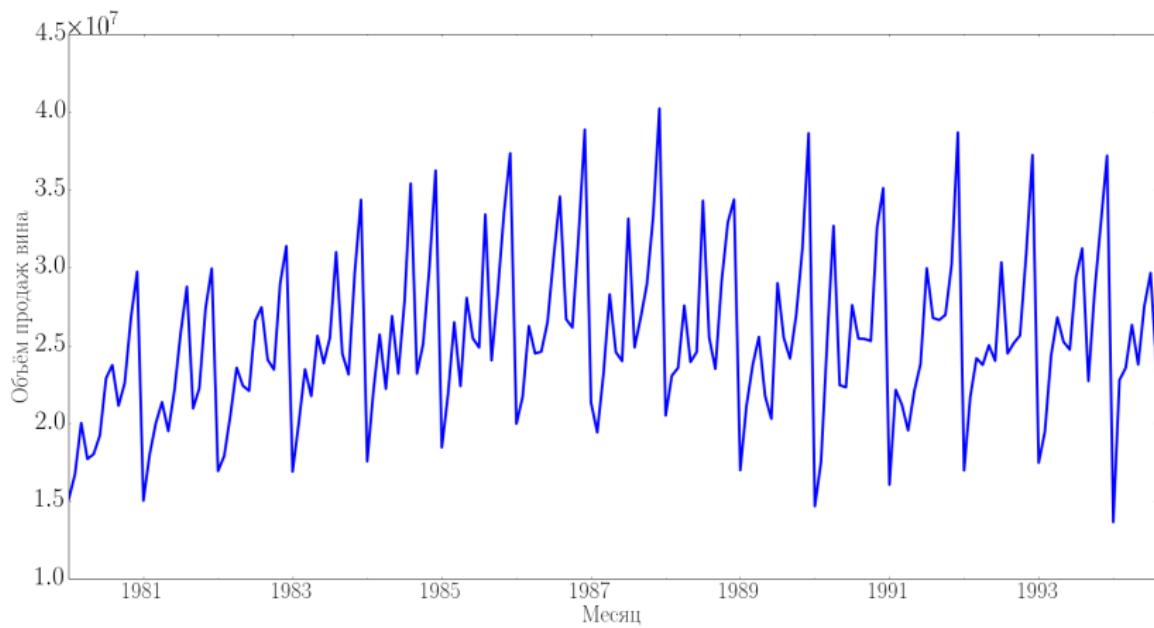
Ничего

Временные ряды

ARIMA

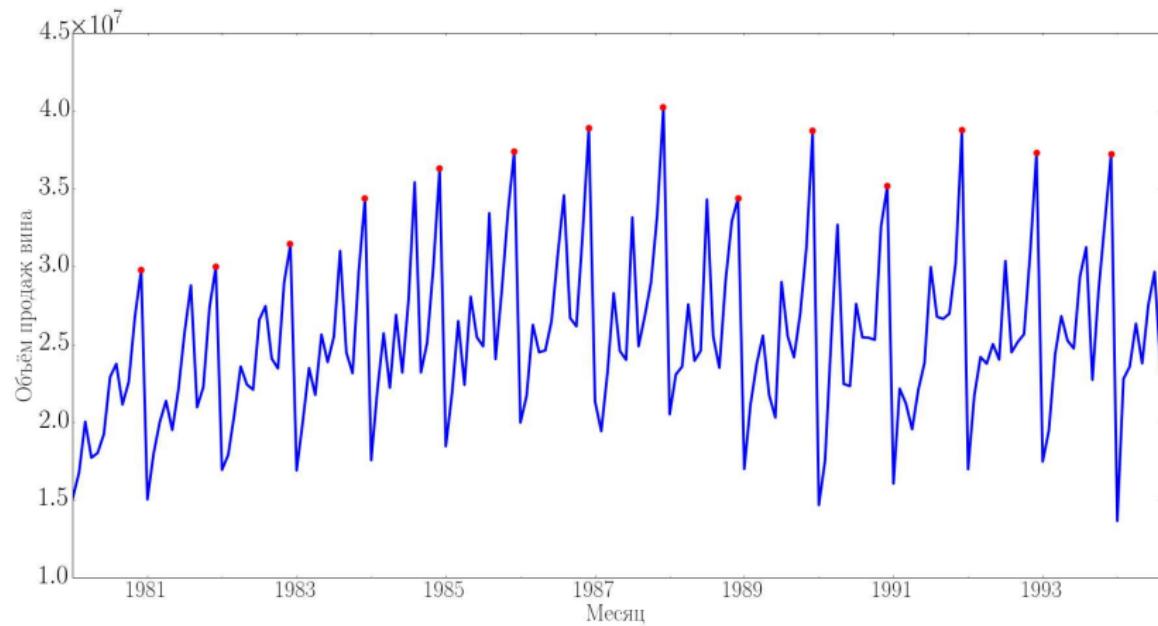
oooooooooooooooooooo

Продажи вина в Австралии



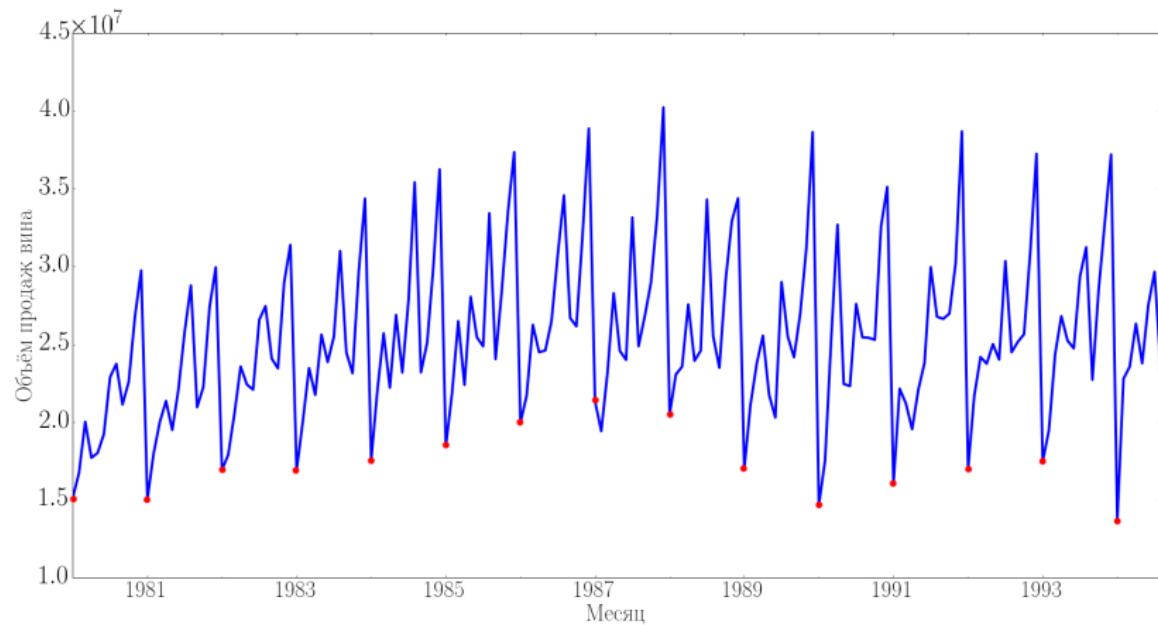
Продажи вина в Австралии

Каждый декабрь продажи большие:

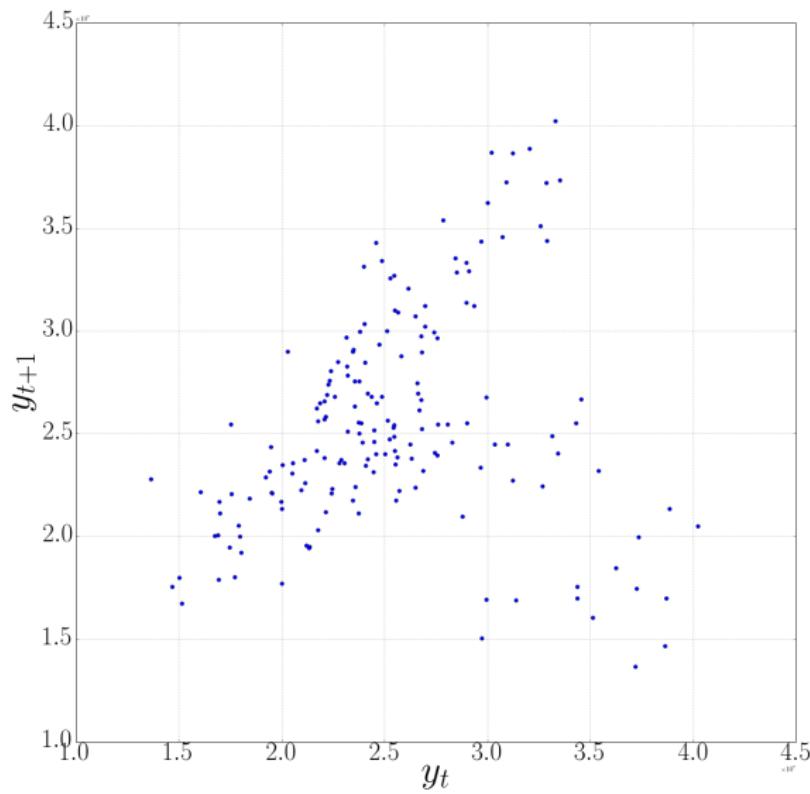


Продажи вина в Австралии

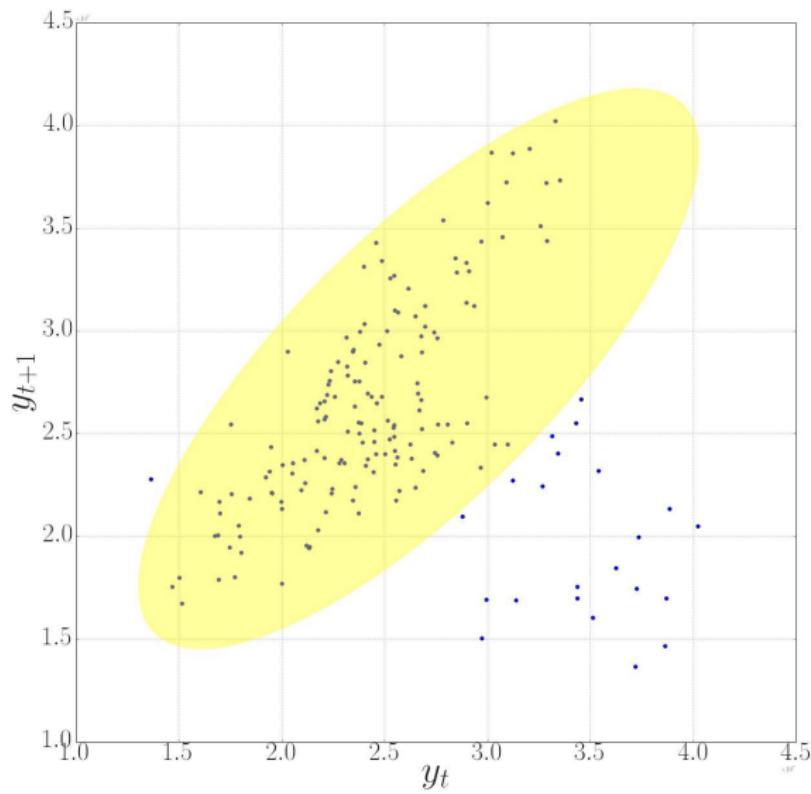
Каждый январь продажи падают:



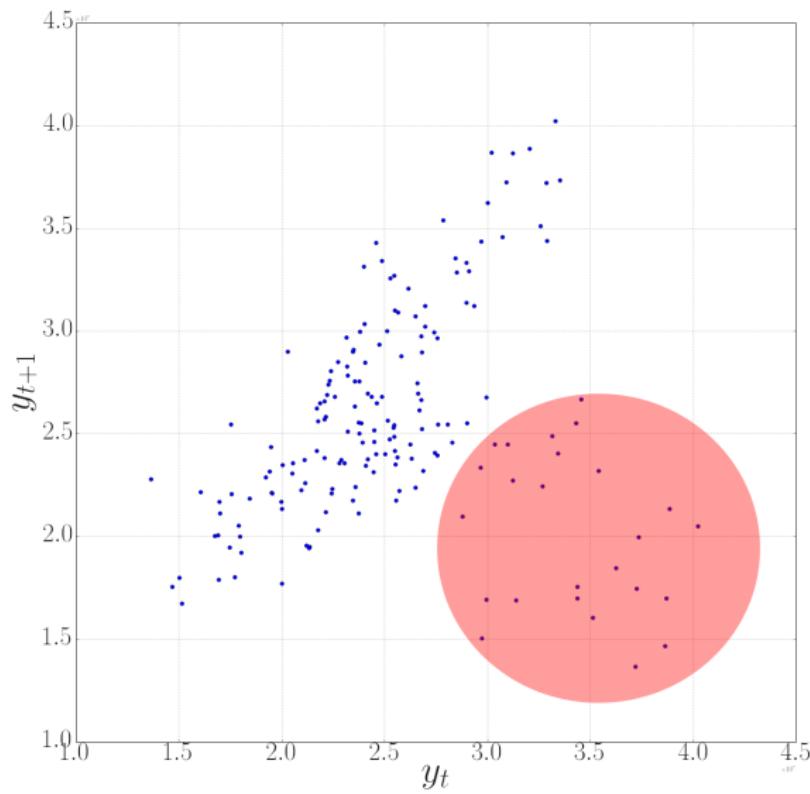
Продажи в соседние месяцы



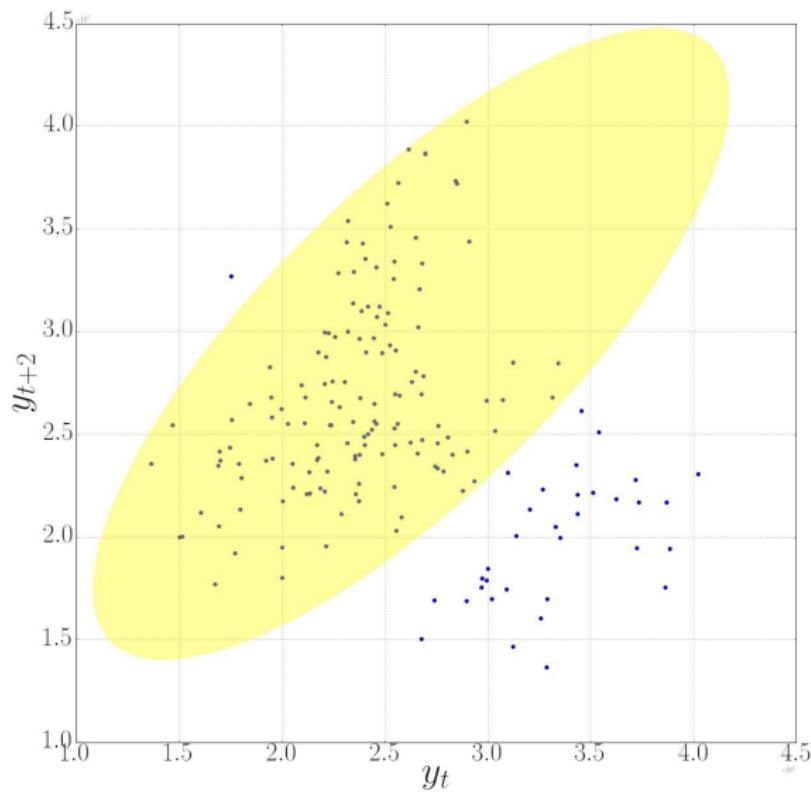
Продажи в соседние месяцы



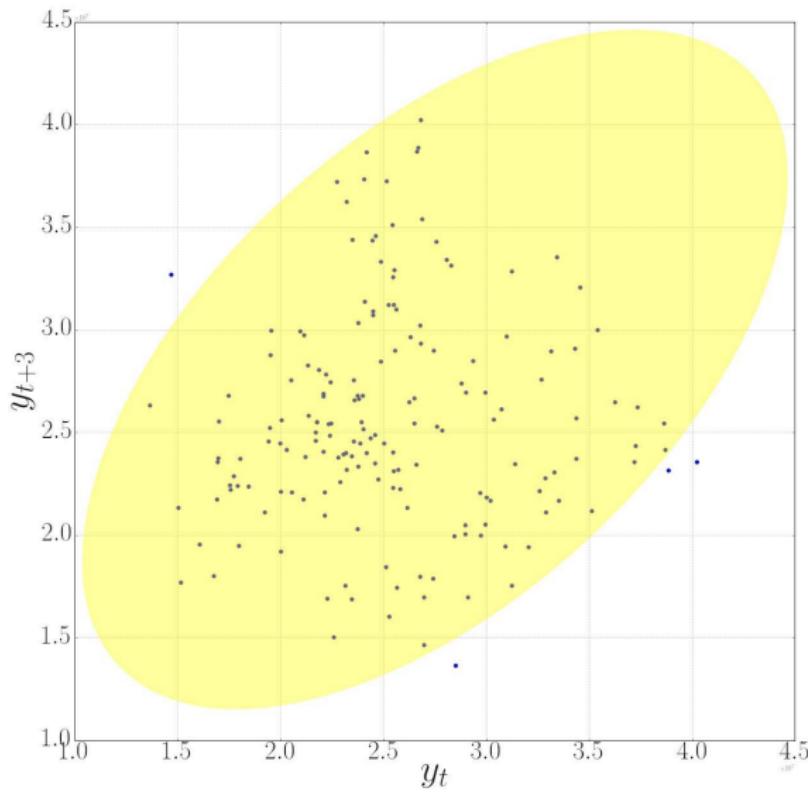
Продажи в соседние месяцы



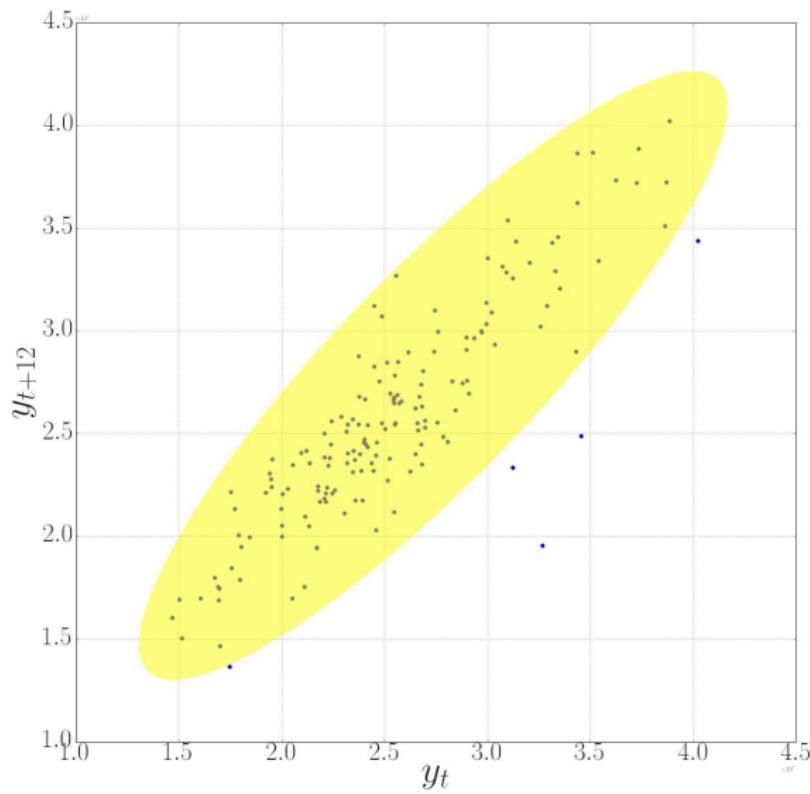
Продажи через 1 месяц



Продажи через 2 месяца



Продажи через год



Автокорреляция

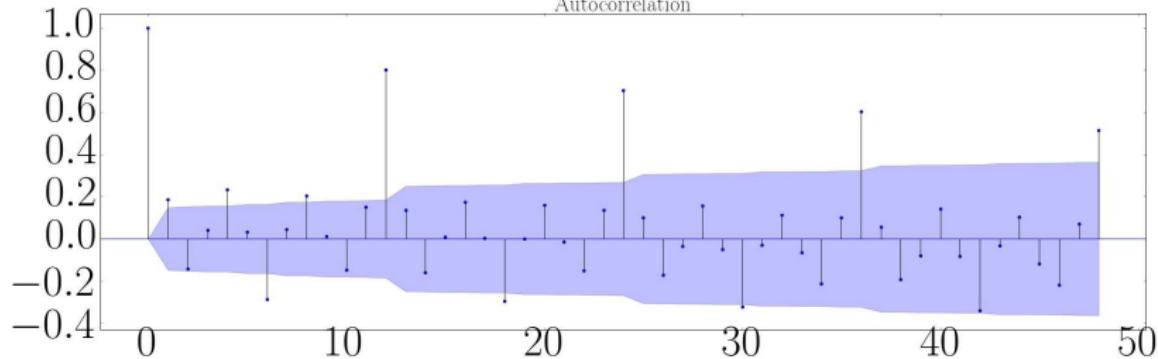
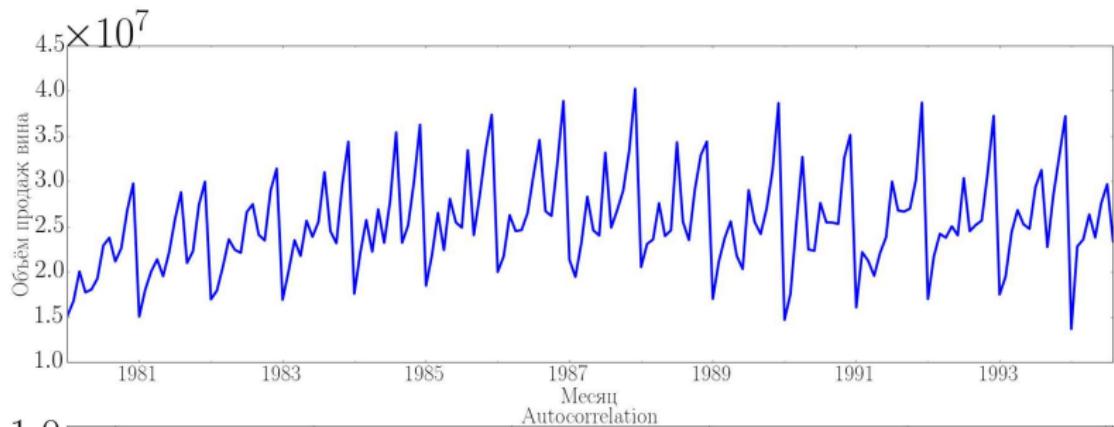
$$r_\tau = \frac{\mathbb{E} ((y_t - \mathbb{E} y) (y_{t+\tau} - \mathbb{E} y))}{\mathbb{D} y}.$$

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

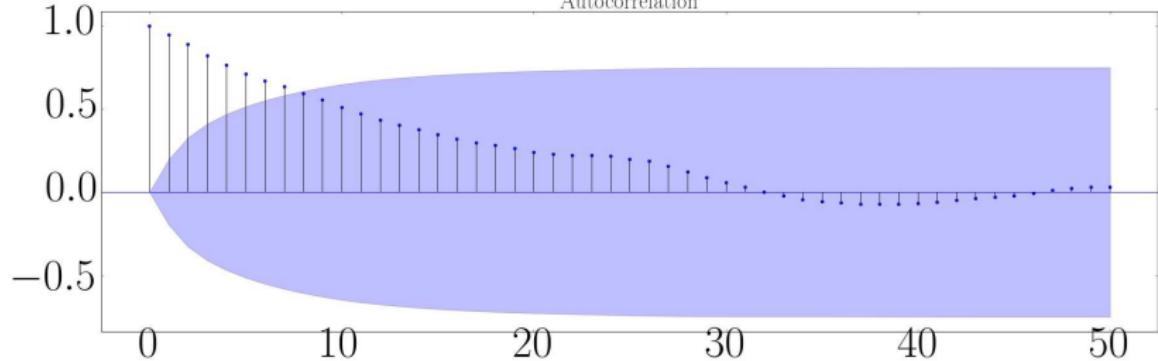
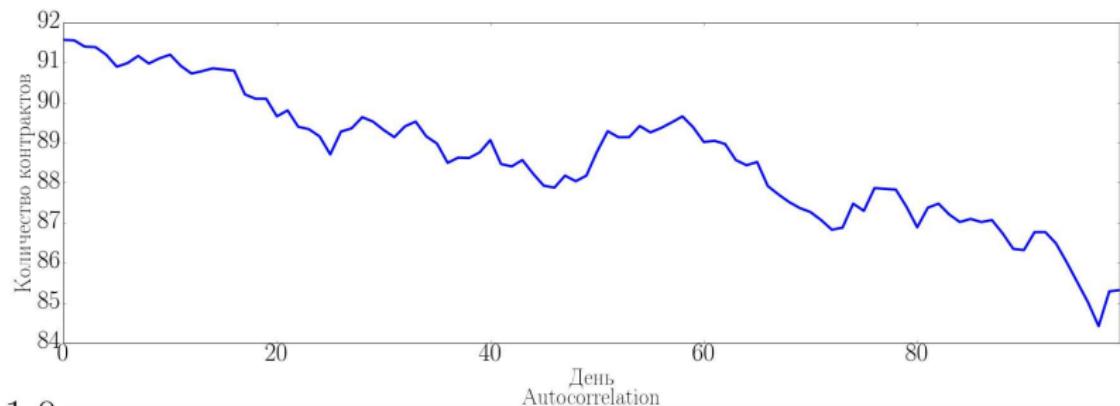
Выборочная автокорреляция:

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

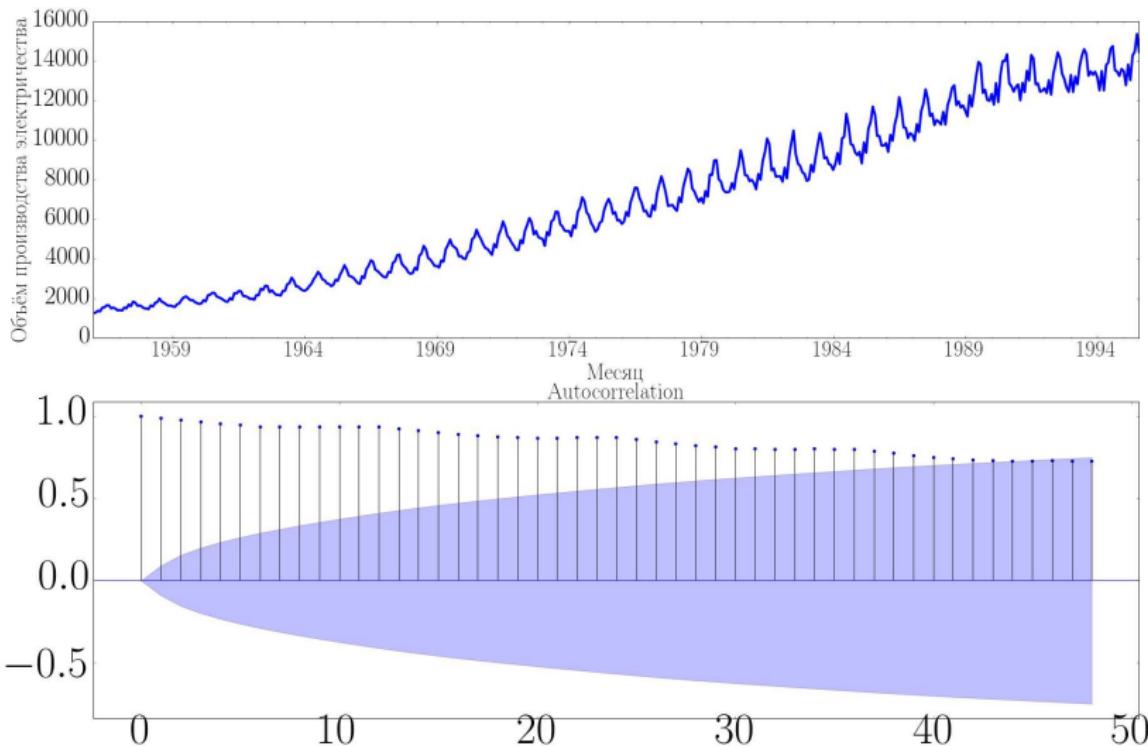
Коррелограммы



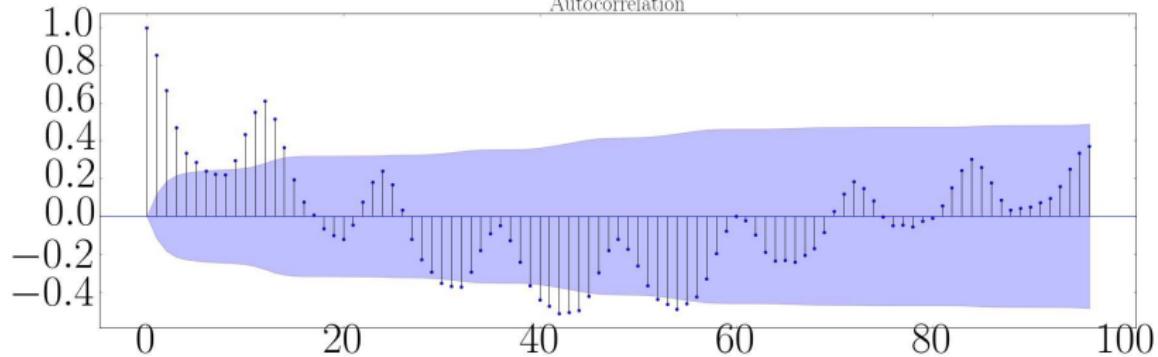
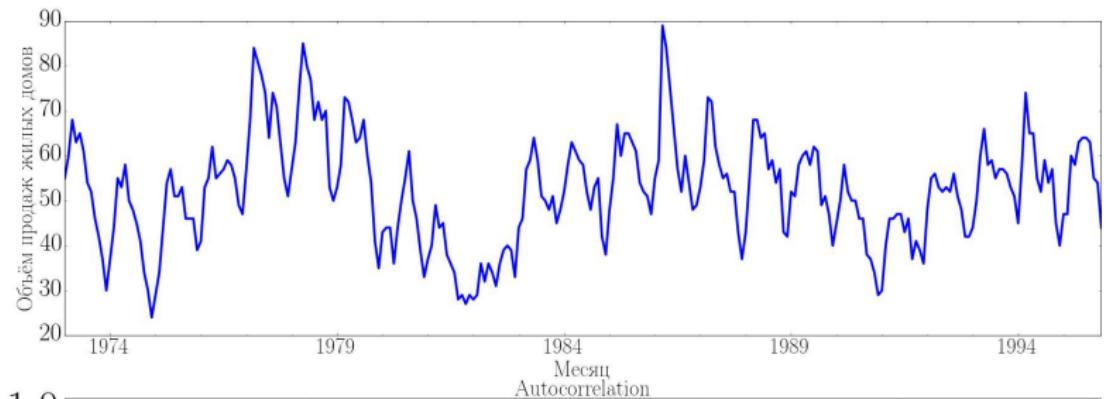
Коррелограммы



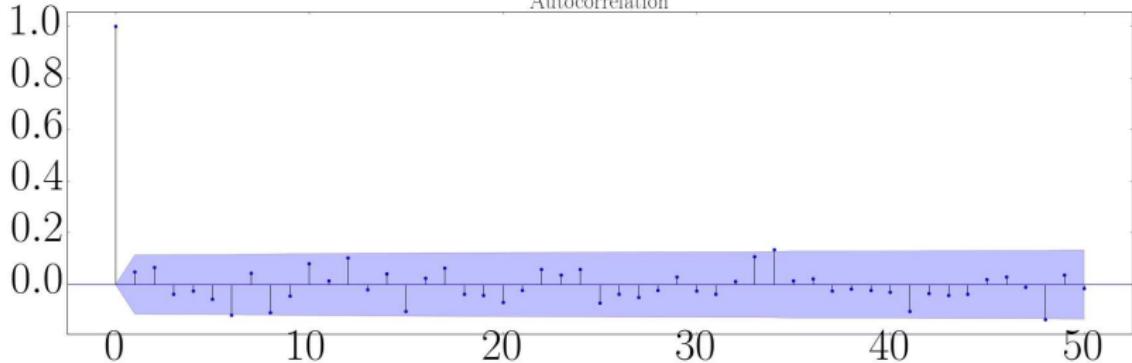
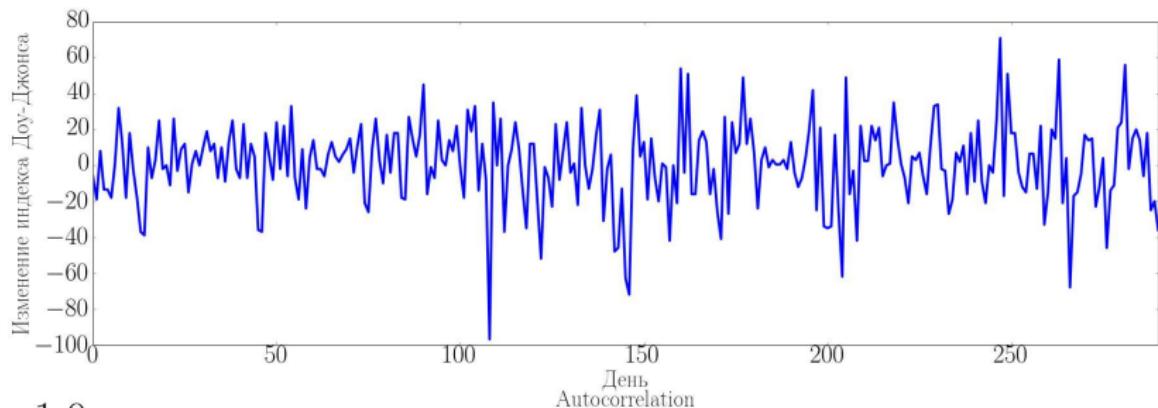
Коррелограммы



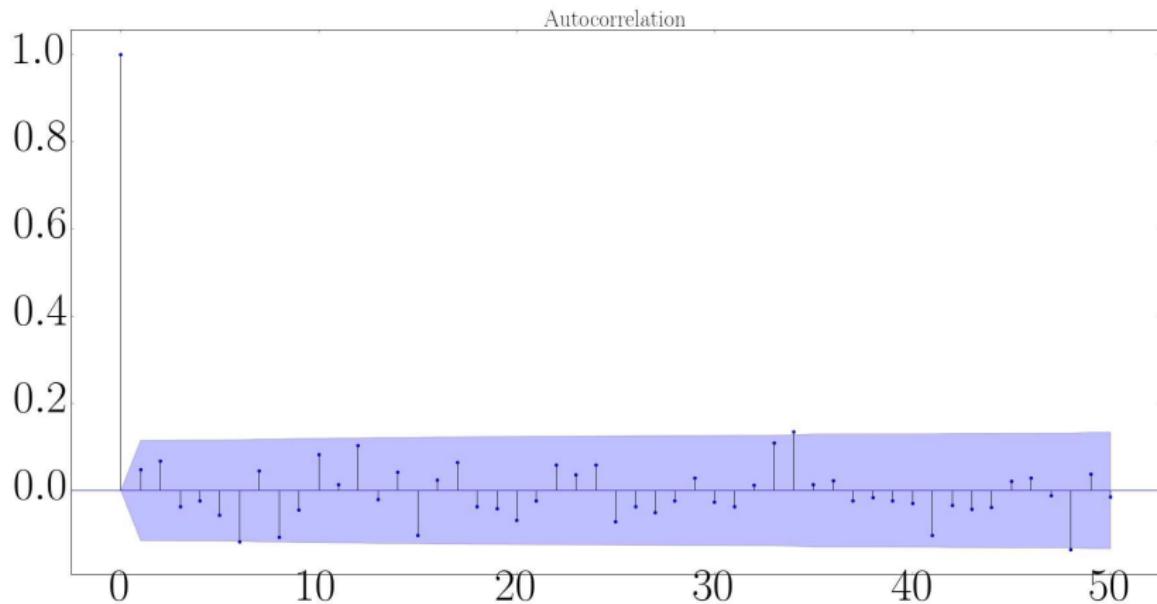
Коррелограммы



Коррелограммы



Значимость автокорреляции



Значимость автокорреляции

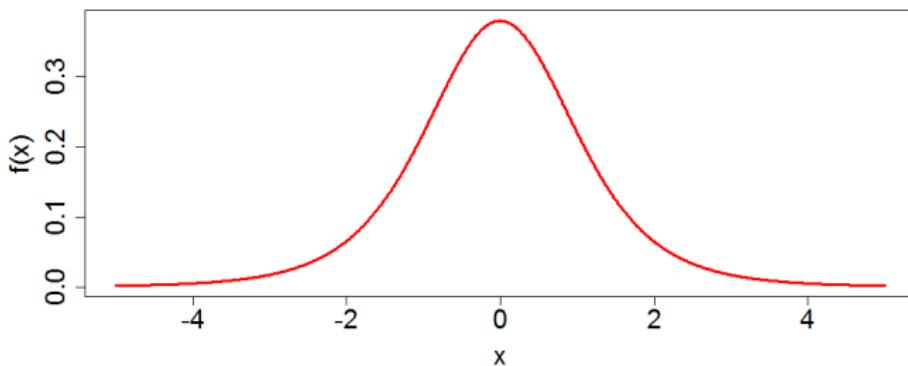
временной ряд: $y^T = y_1, \dots, y_T$;

нулевая гипотеза: $H_0: r_\tau = 0$;

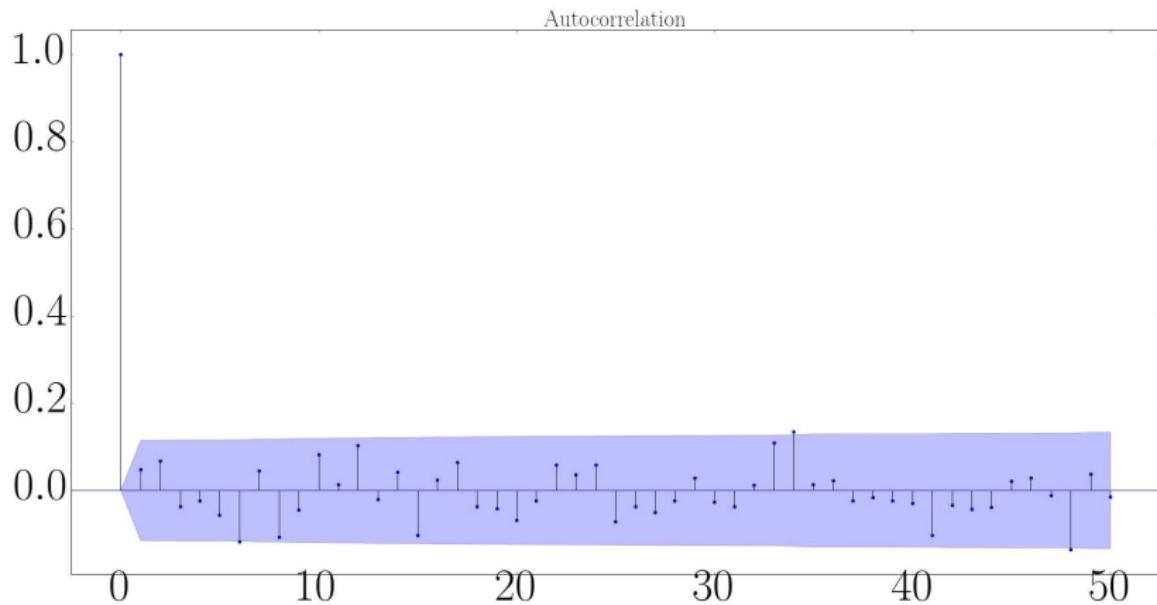
альтернатива: $H_1: r_\tau \neq 0$;

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 .



Значимость автокорреляции



Q-критерий Льюнга-Бокса

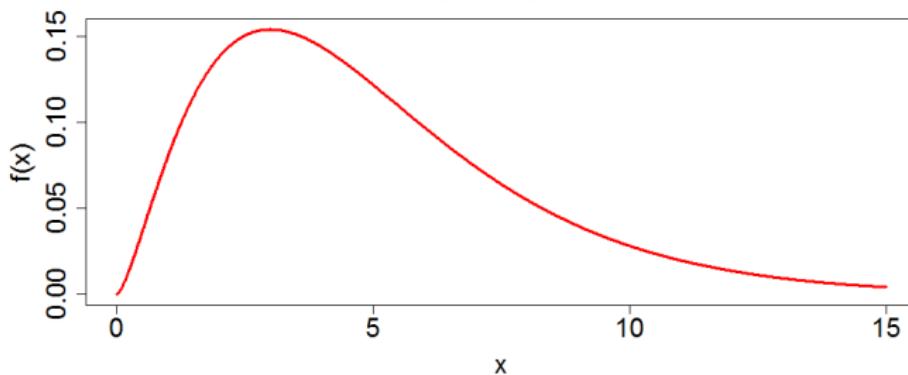
ряд ошибок прогноза: $\varepsilon^T = \varepsilon_1, \dots, \varepsilon_T$;

нулевая гипотеза: $H_0: r_1 = \dots = r_L = 0$;

альтернатива: $H_1: H_0$ неверна;

статистика: $Q(\varepsilon^T) = T(T+2) \sum_{\tau=1}^L \frac{r_\tau^2}{T-\tau}$;

нулевое распределение: χ^2_{L-K} , K — число настраиваемых параметров модели ряда.

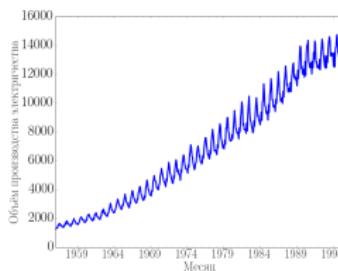
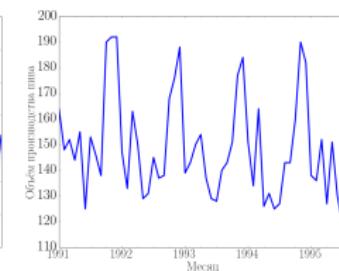
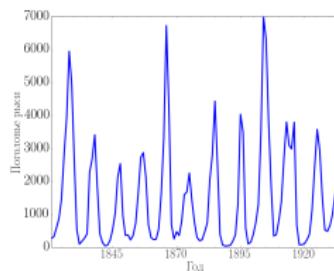
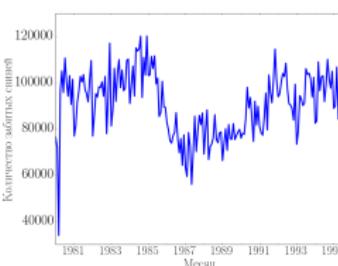
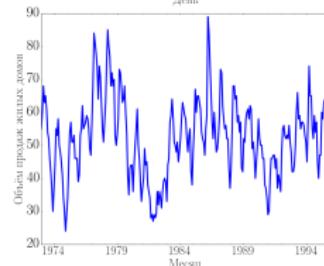
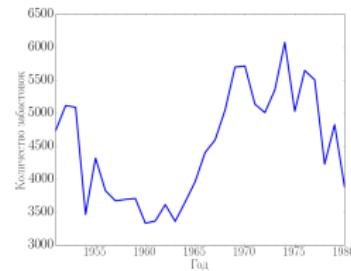
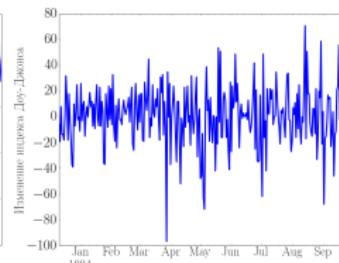
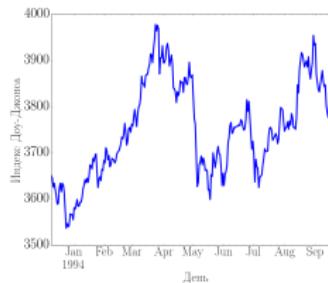


Стационарность

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.

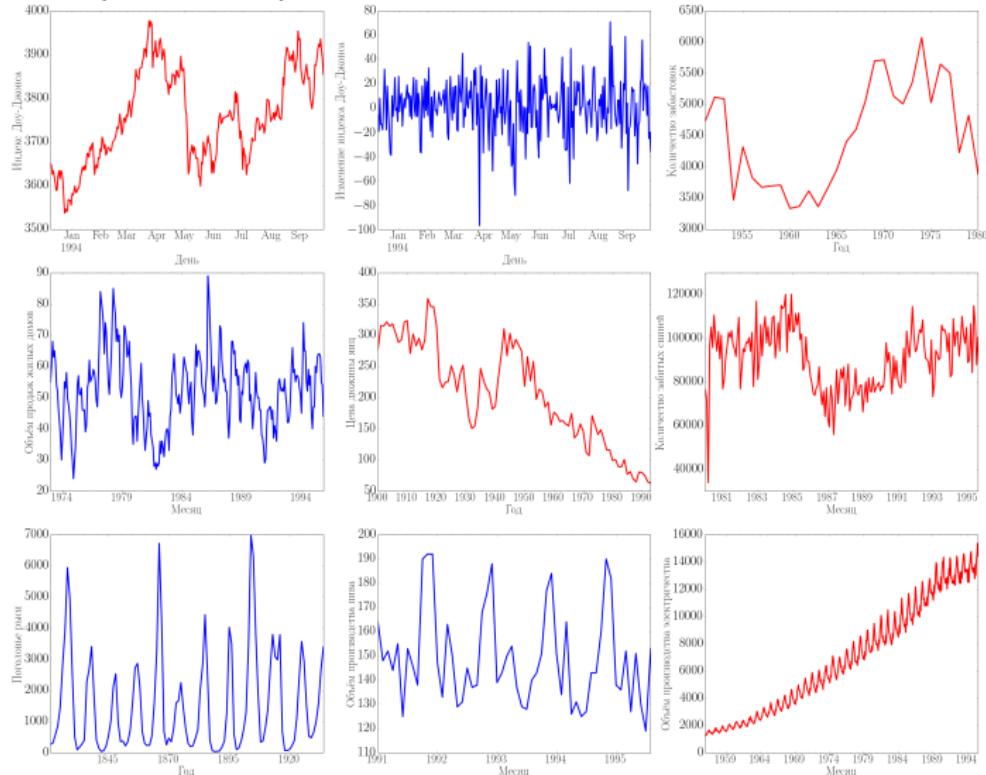
- тренд \Rightarrow нестационарность
- сезонность \Rightarrow нестационарность
- цикл \Rightarrow нестационарность (нельзя предсказать заранее, где будут находятся максимумы и минимумы)

Примеры



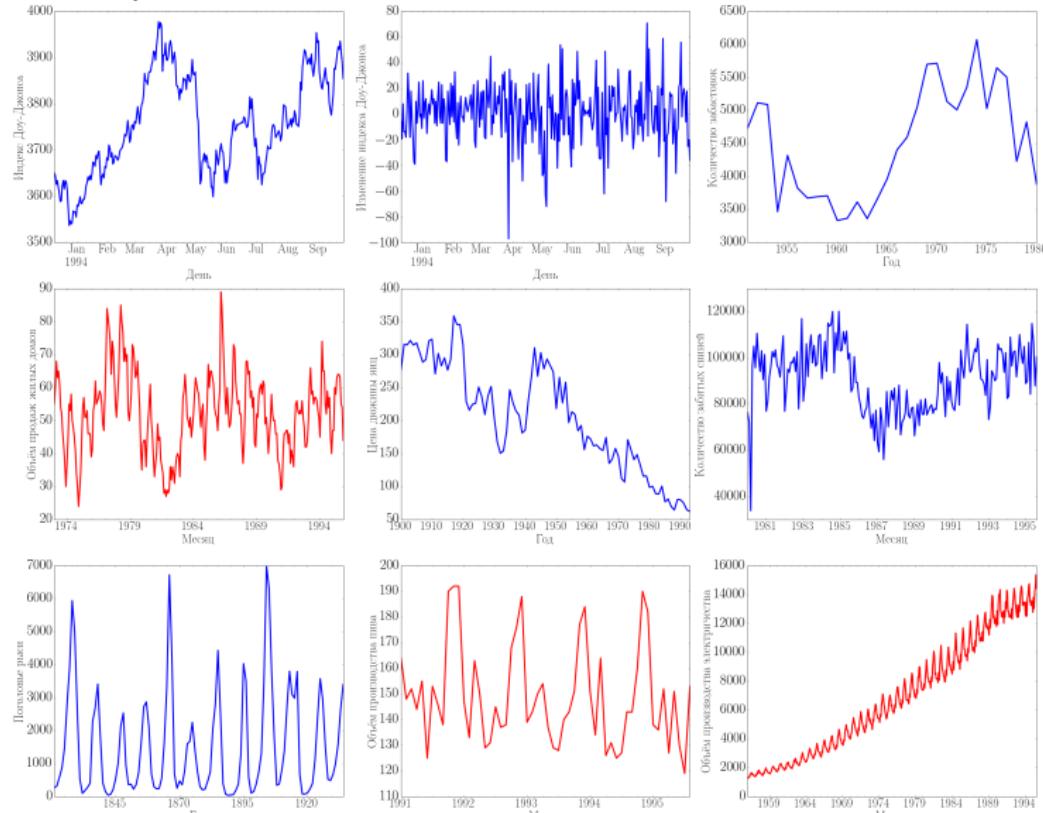
Примеры

Нестационарны из-за тренда:



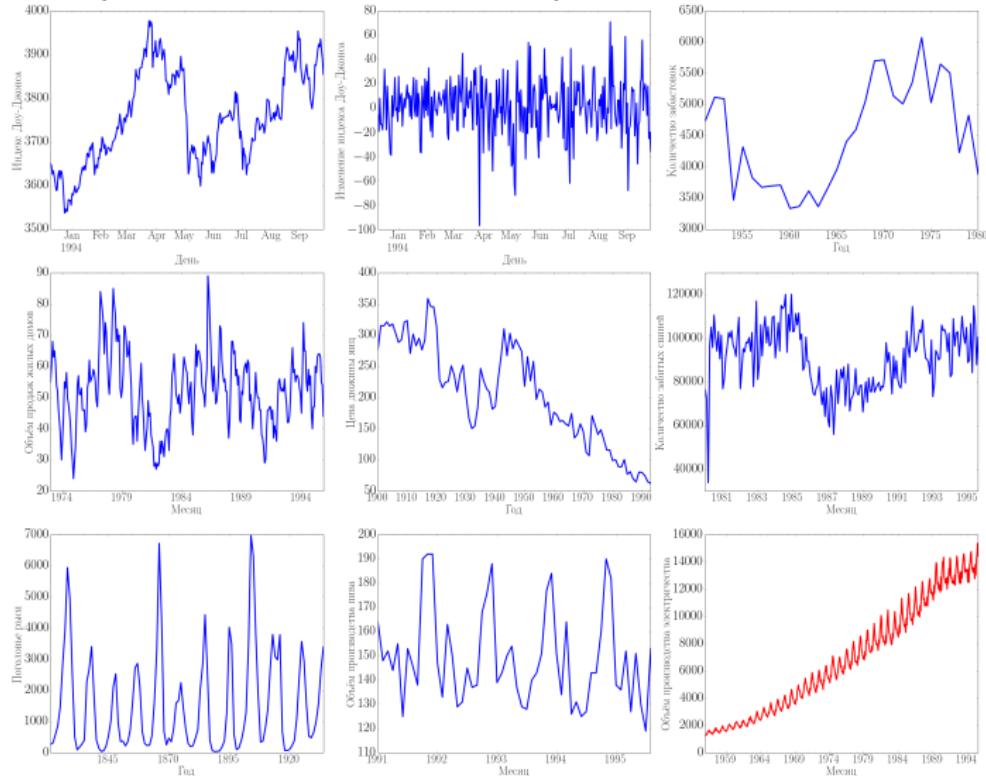
Примеры

Нестационарны из-за сезонности:



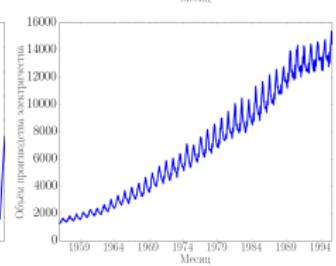
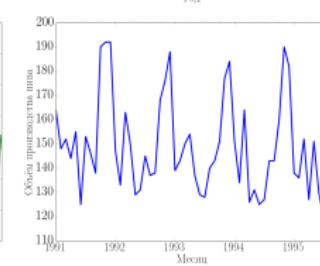
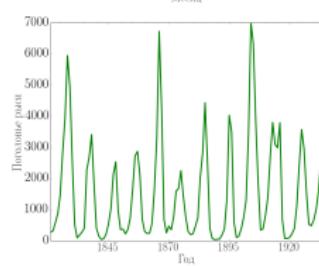
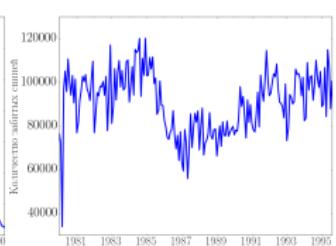
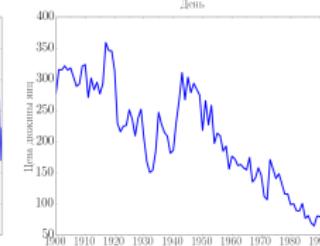
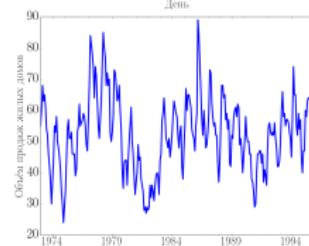
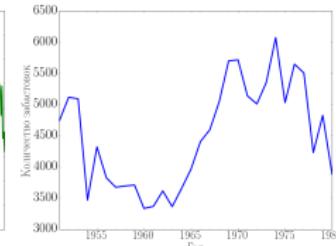
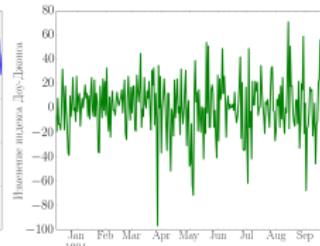
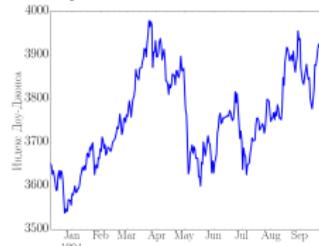
Примеры

Нестационарны из-за меняющейся дисперсии:



Примеры

Стационарны:



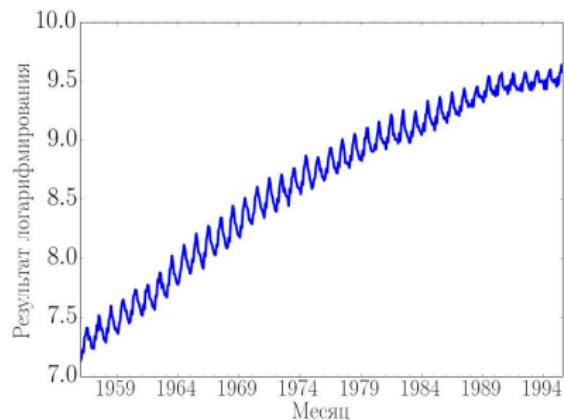
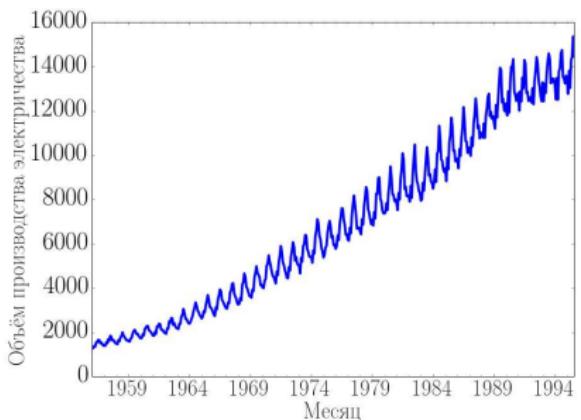
Критерий Дики-Фуллера

- 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.

Стабилизация дисперсии

Для рядов с монотонно меняющейся дисперсией можно использовать стабилизирующие преобразования.

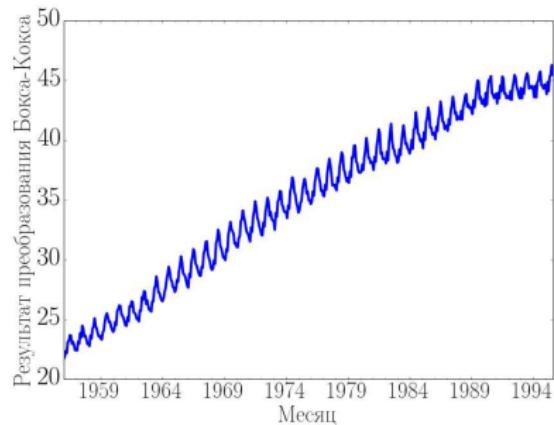
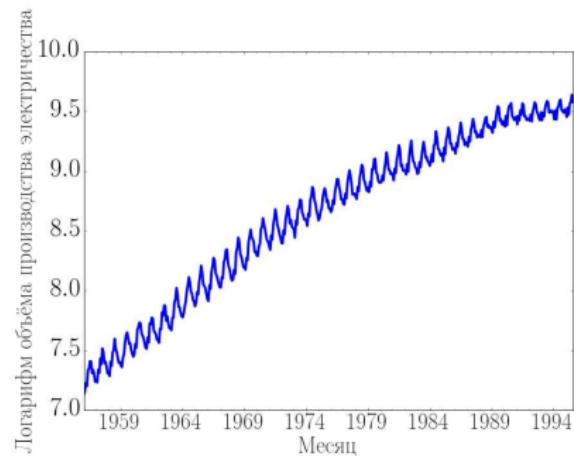
Часто используют логарифмирование:



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

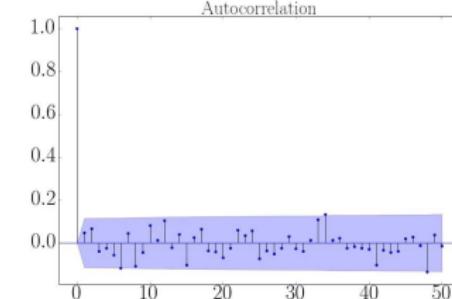
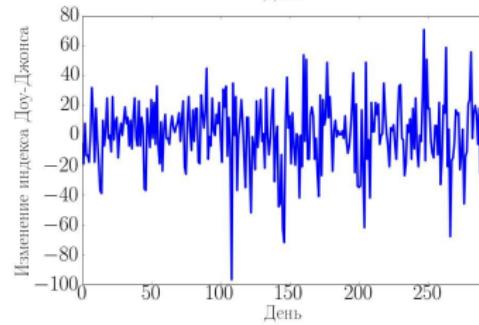
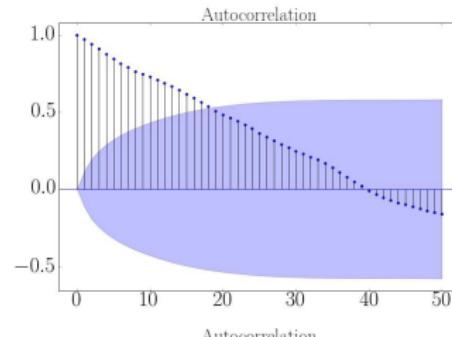
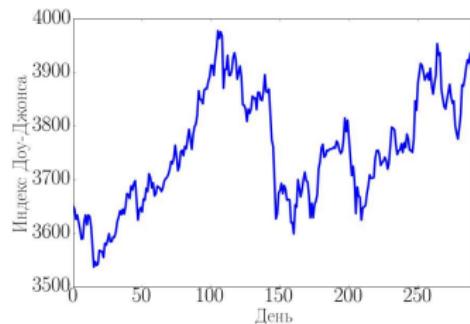
Дифференцирование

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?

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

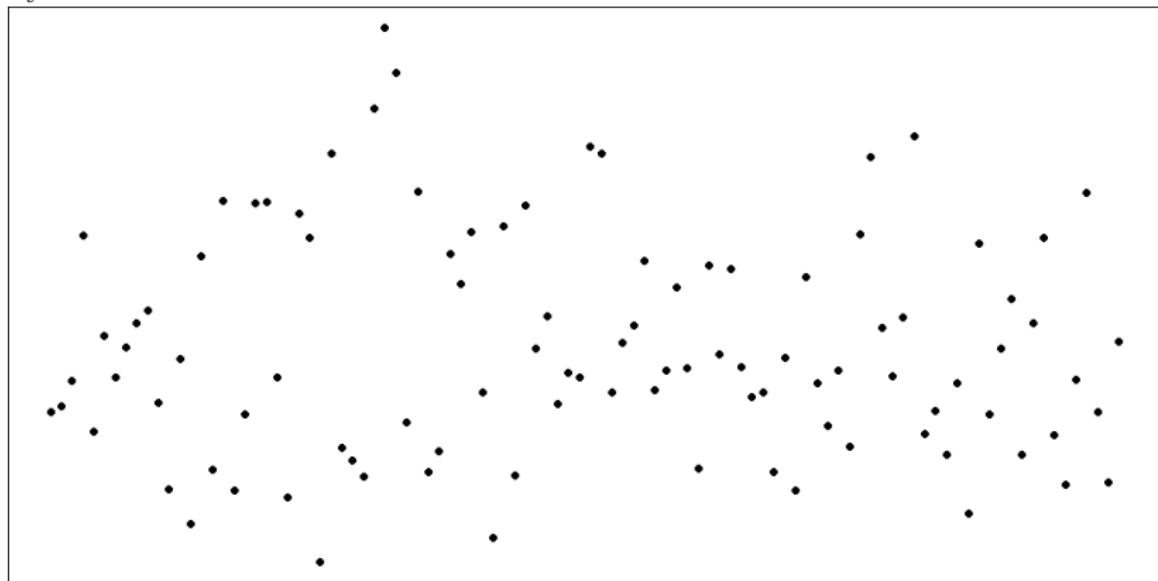
Модель авторегрессии порядка p ($AR(p)$):

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 :

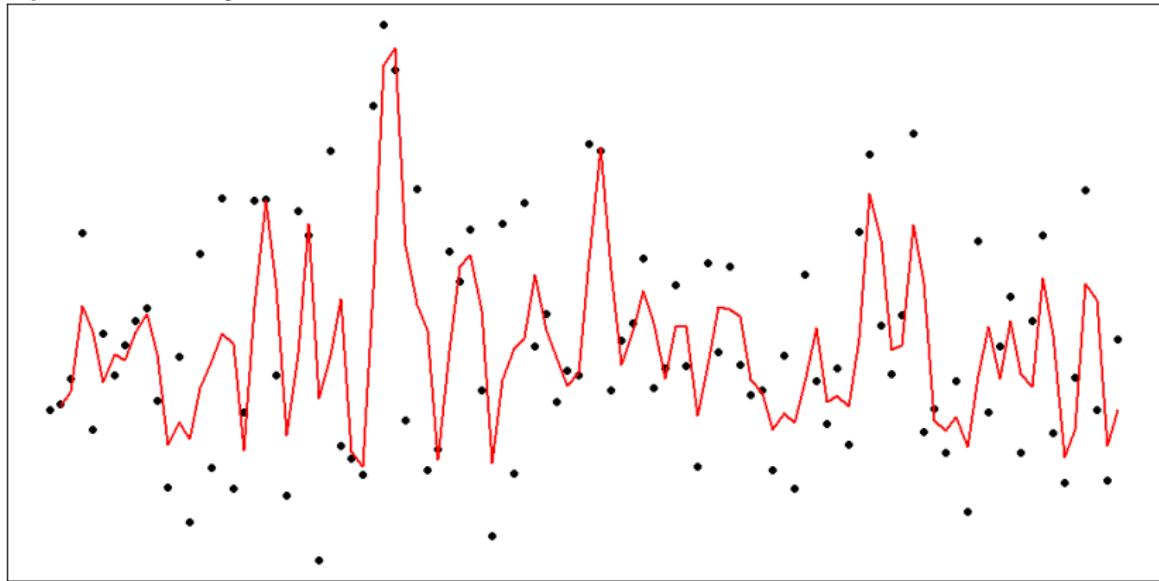


Временные ряды

ARIMA

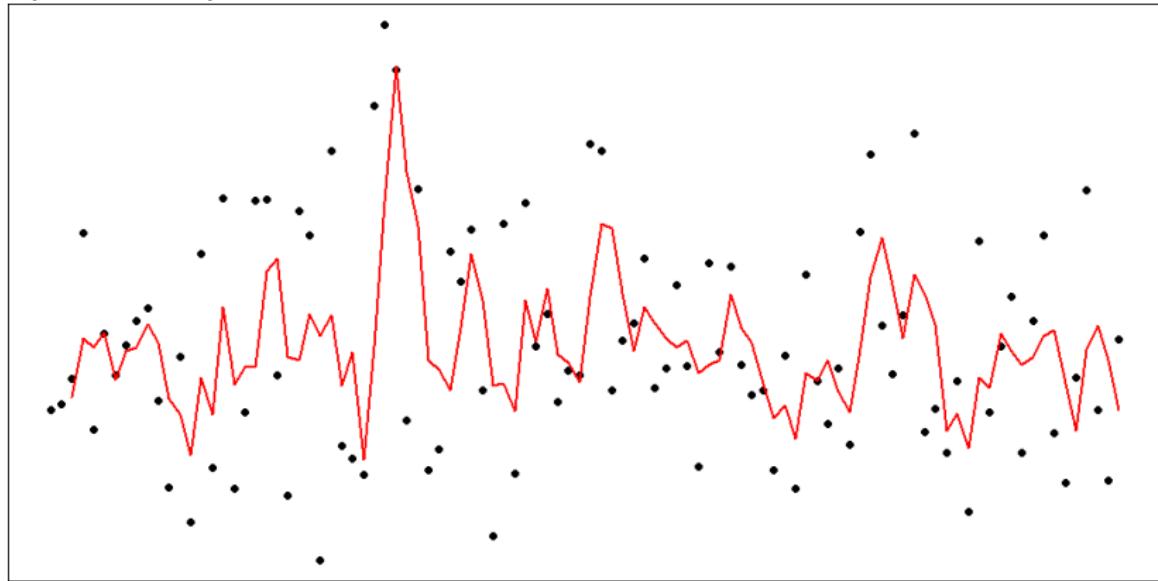
Скользящее среднее

Среднее по двум соседним точкам:



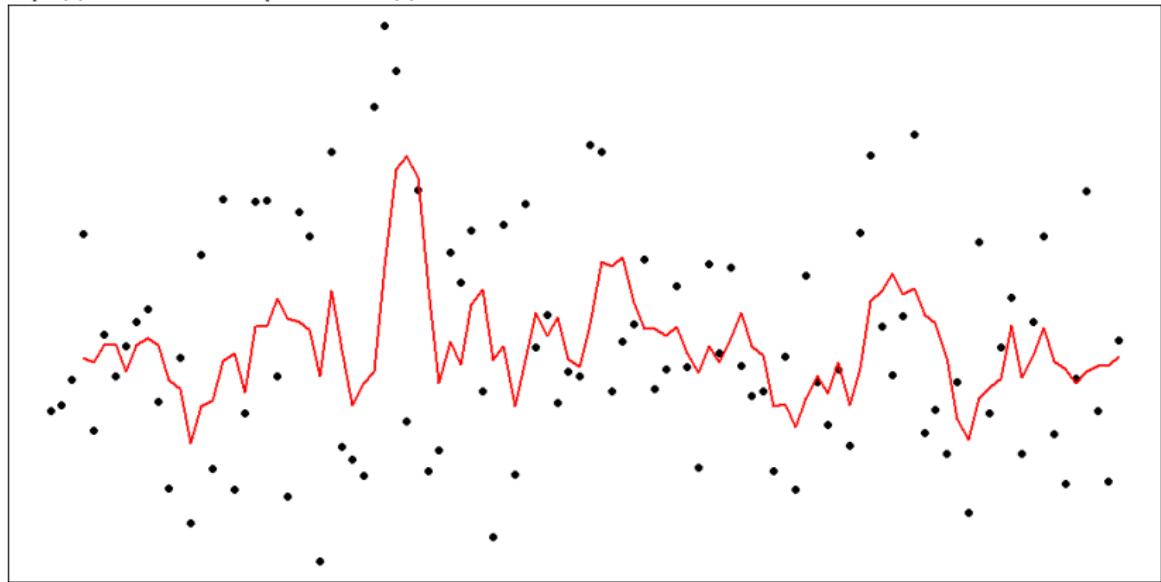
Скользящее среднее

Среднее по трём соседним точкам:



Скользящее среднее

Среднее по четырём соседним точкам:



Какая-то информация о Y в этом шуме есть

Скользящее среднее

Обобщим и добавим веса:

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

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

y_t — линейная комбинация q последних значений шумовой компоненты.

Модель $ARMA(p, q)$:

$$y_t = \alpha + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

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

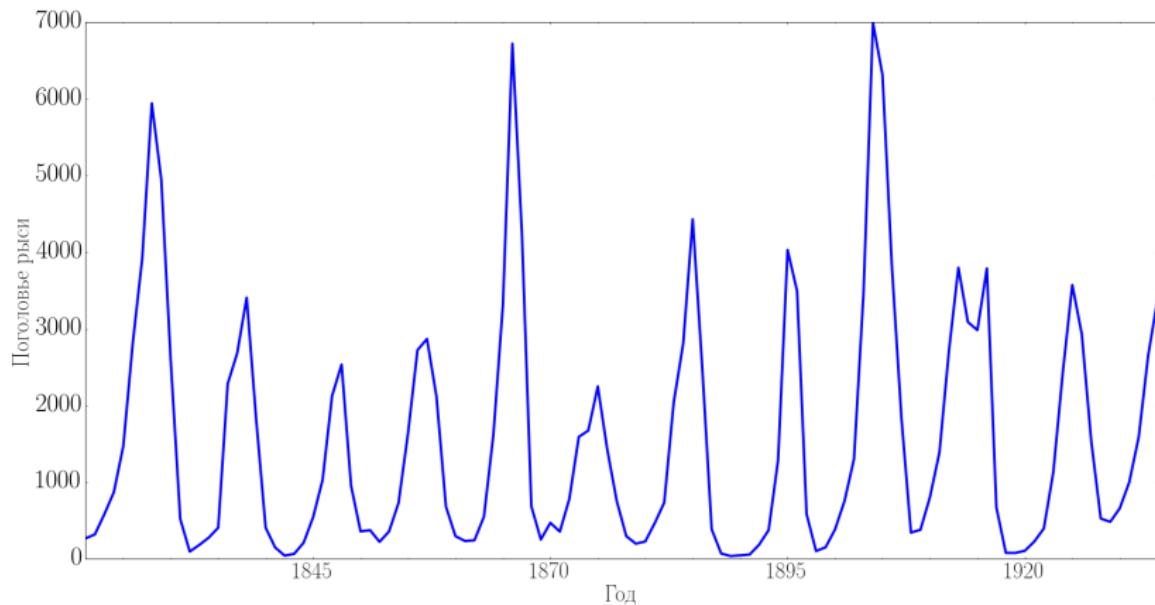
Временные ряды

oooooooooooooooooooo

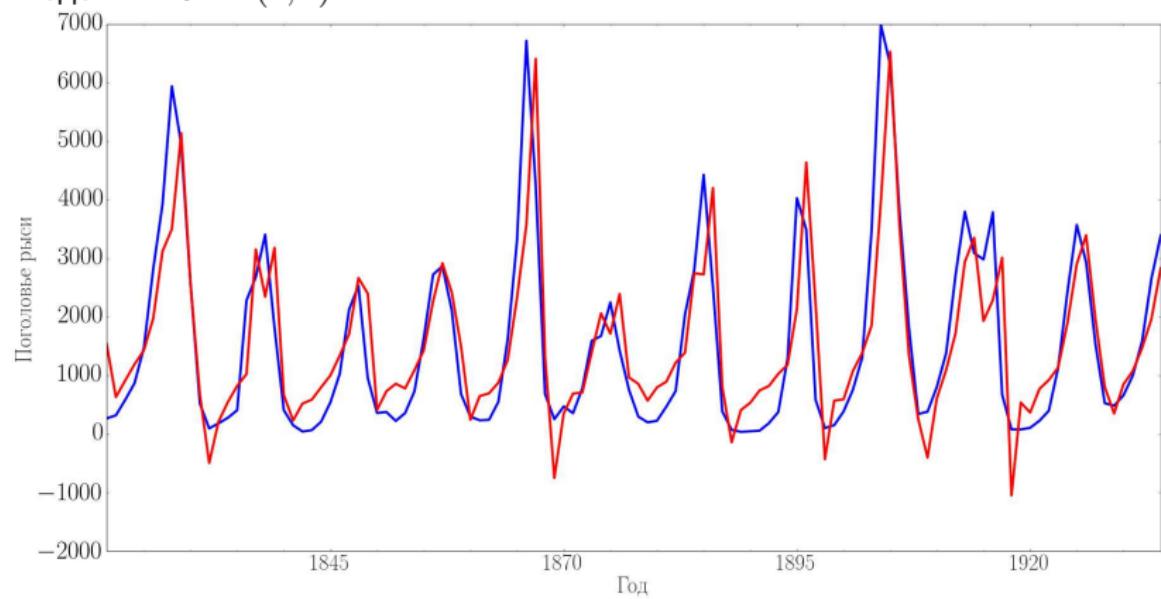
ARIMA

oooo●ooooooooooooo

Поголовье рыси



Поголовье рыси

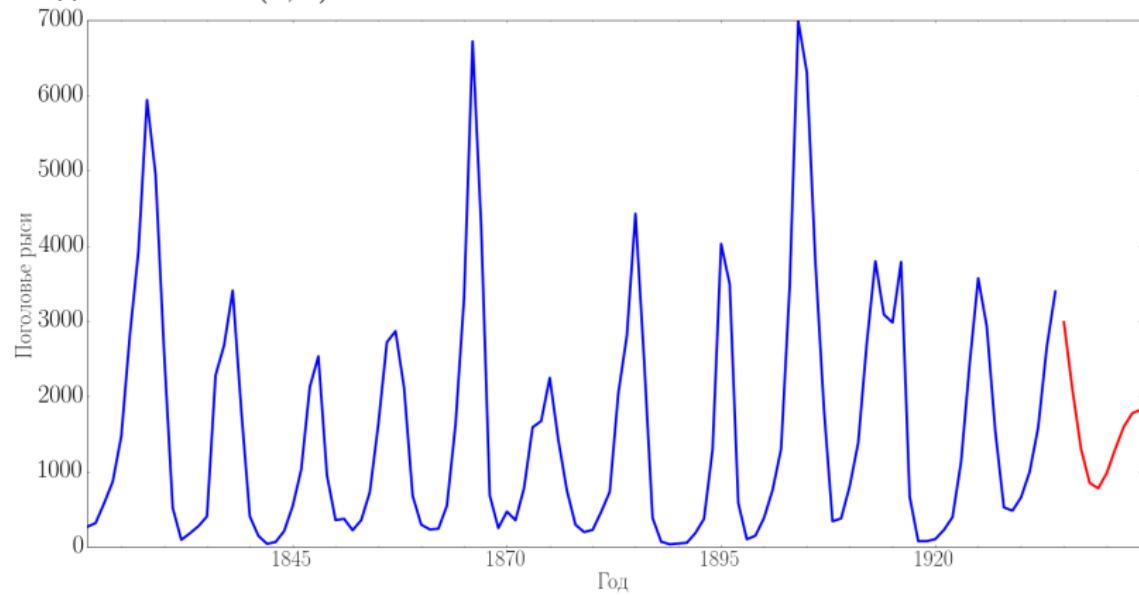
Модель $ARMA(2, 2)$:

Временные ряды
oooooooooooooooooooo

ARIMA
oooo●ooooooooooooooooo

Поголовье рыси

Модель $ARMA(2, 2)$:



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

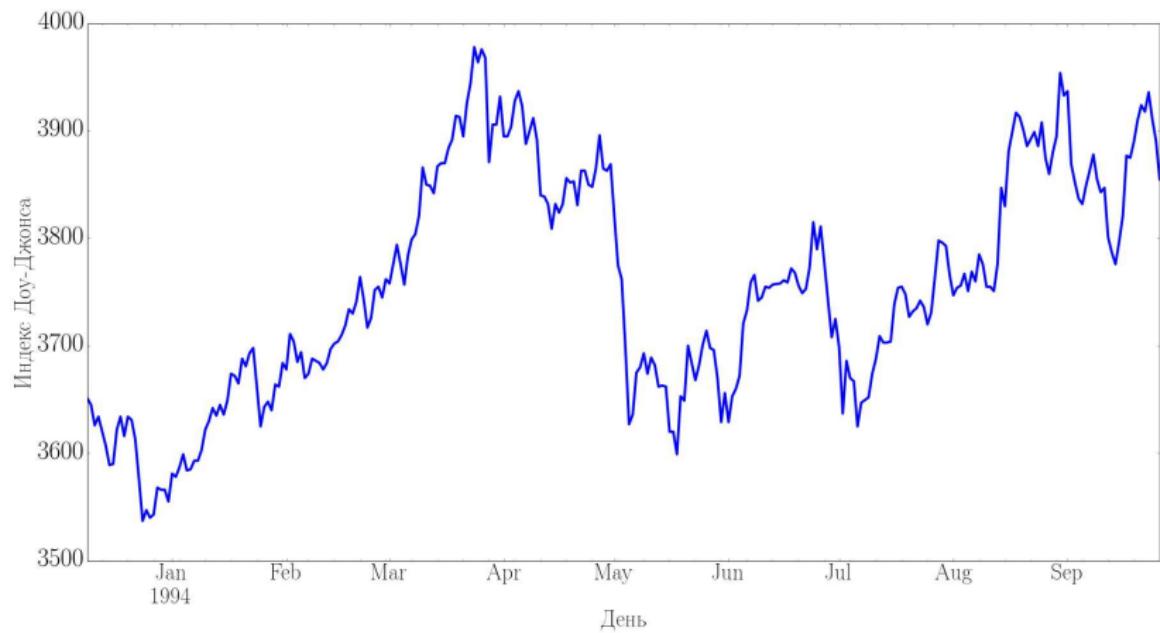
Временные ряды

oooooooooooooooooooo

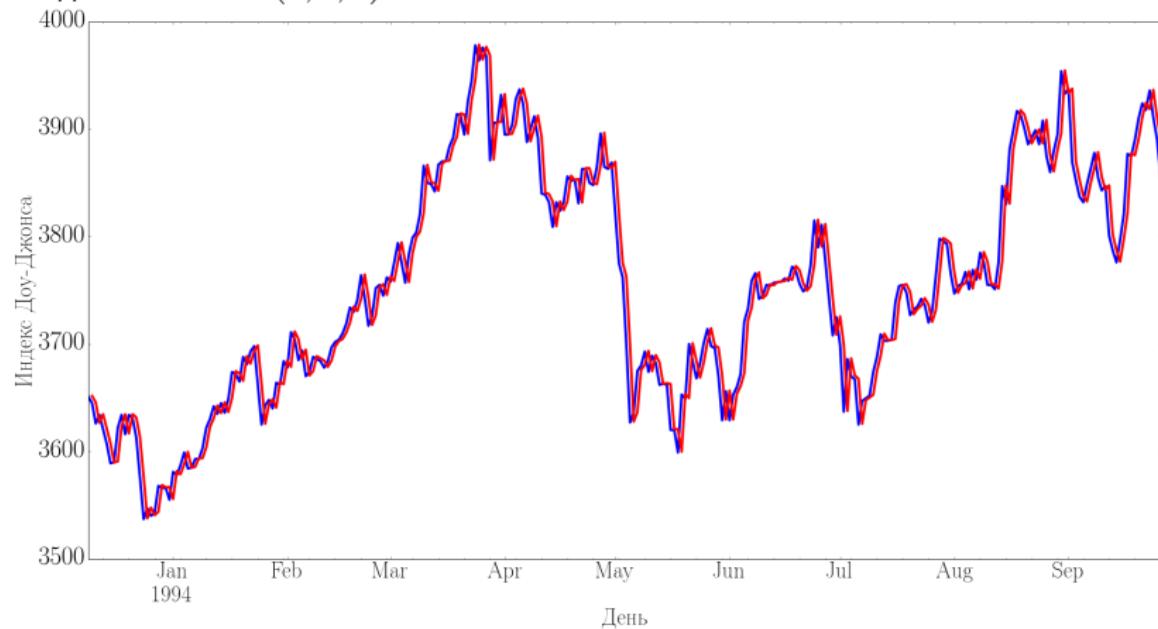
ARIMA

oooooooo●oooooooooooooooooooo

Индекс Доу-Джонса



Индекс Доу-Джонса

Модель $ARIMA(0, 1, 0)$:

Пусть ряд имеет сезонный период длины S .

Возьмём модель $ARMA(p, q)$:

$$y_t = \alpha + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

и добавим P авторегрессионных компонент:

$$+\phi_S y_{t-S} + \phi_{2S} y_{t-2S} + \cdots + \phi_{PS} y_{t-PS}$$

и Q компонент скользящего среднего:

$$+\theta_S \varepsilon_{t-S} + \theta_{2S} \varepsilon_{t-2S} + \cdots + \theta_{QS} \varepsilon_{t-QS}.$$

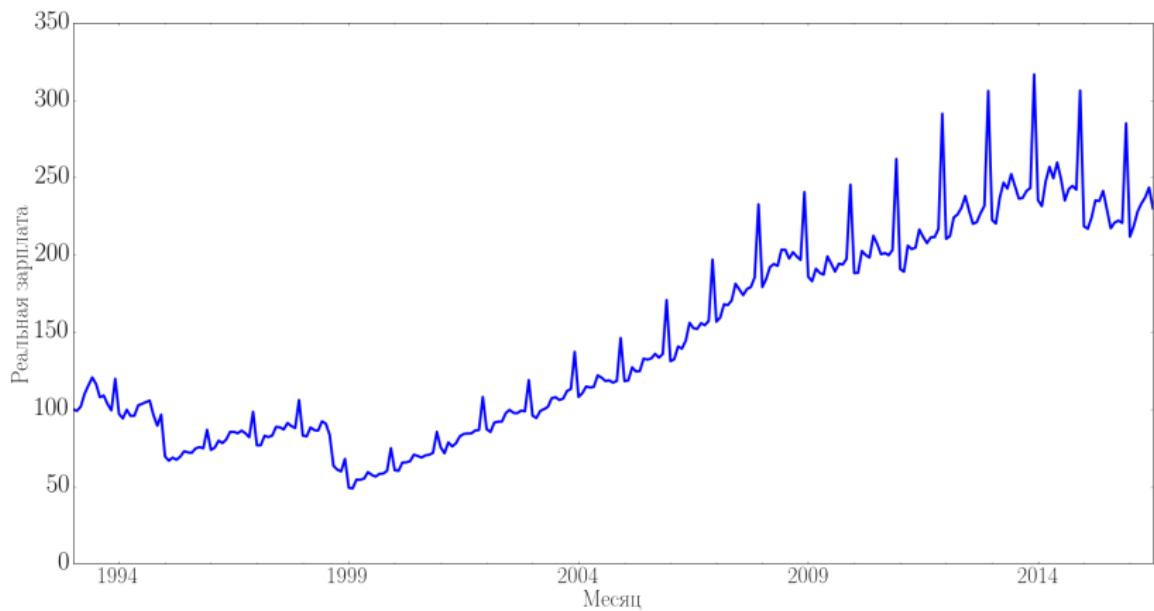
Это модель $SARMA(p, q) \times (P, Q)$

SARIMA

Модель $SARIMA(p, d, q) \times (P, D, Q)$ — модель $SARMA(p, q) \times (P, Q)$ для ряда, к которому d раз было применено обычное дифференцирование и D раз — сезонное.

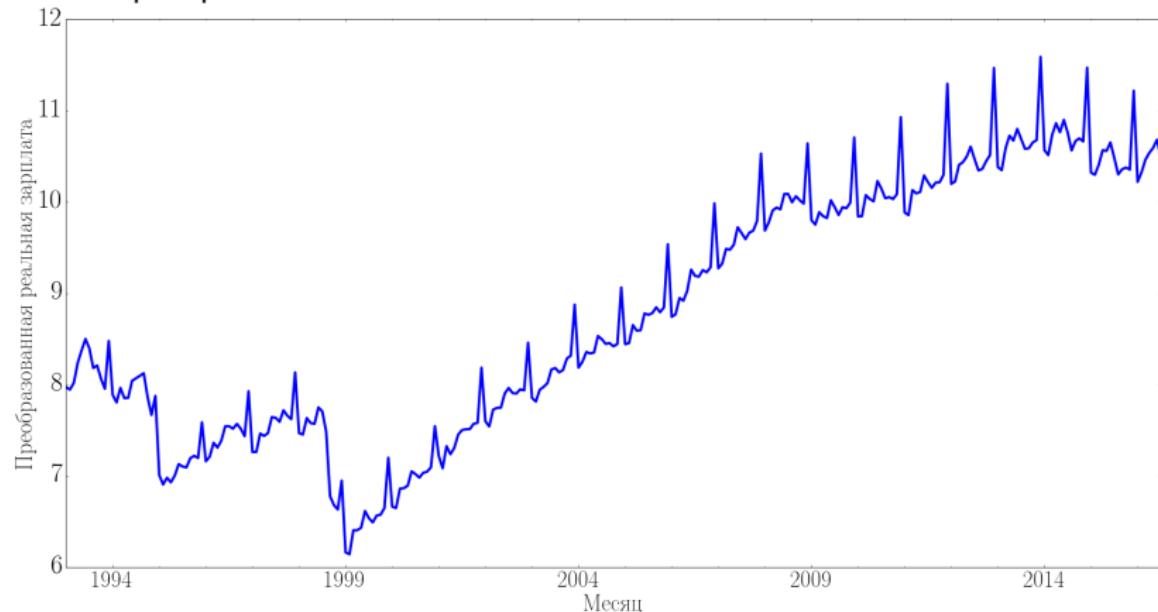
Часто называют просто ARIMA.

Реальная заработка плата



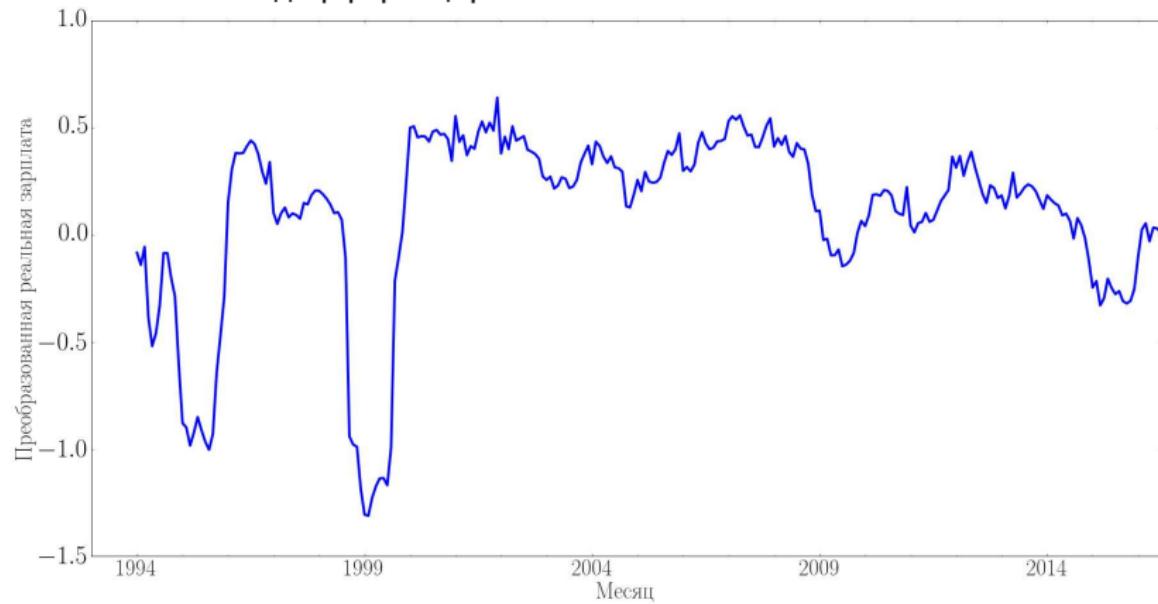
Критерий Дики-Фуллера: $p = 0.2265$.

Реальная заработка плата

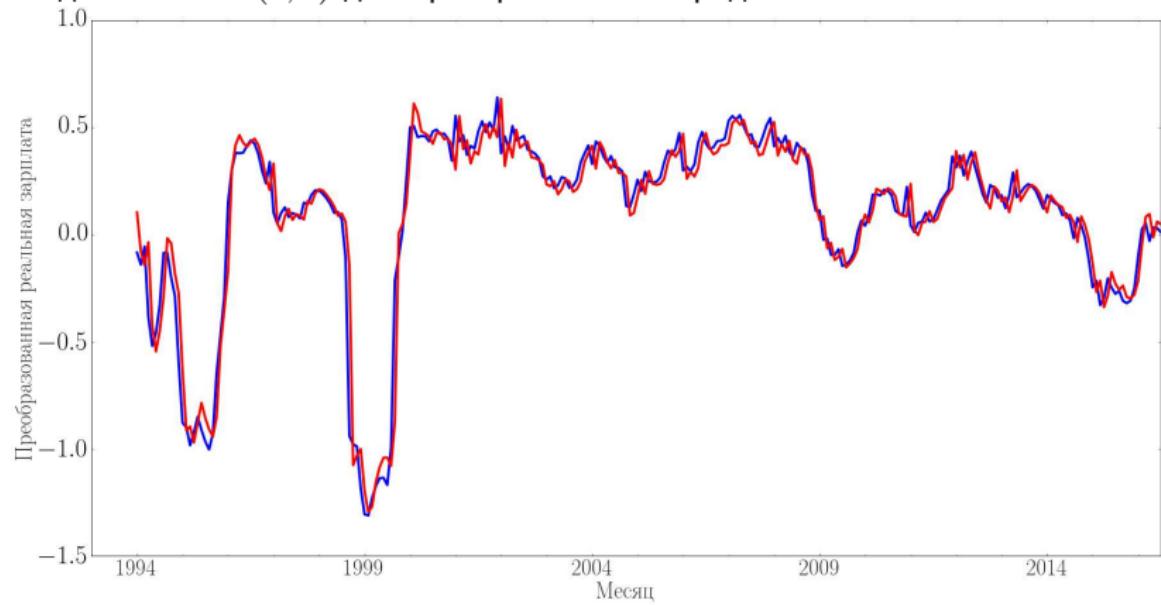
После преобразования Бокса-Кокса с $\lambda = 0.22$:Критерий Дики-Фуллера: $p = 0.1661$.

Реальная заработка платы

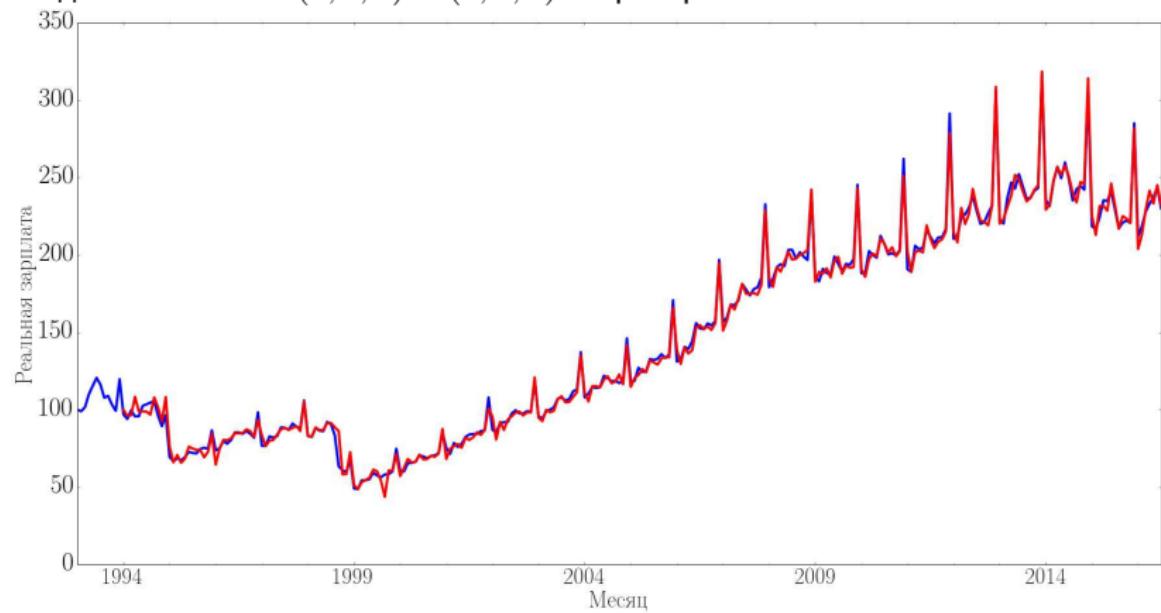
После сезонного дифференцирования:

Критерий Дики-Фуллера: $p = 0.01$.

Реальная заработка плата

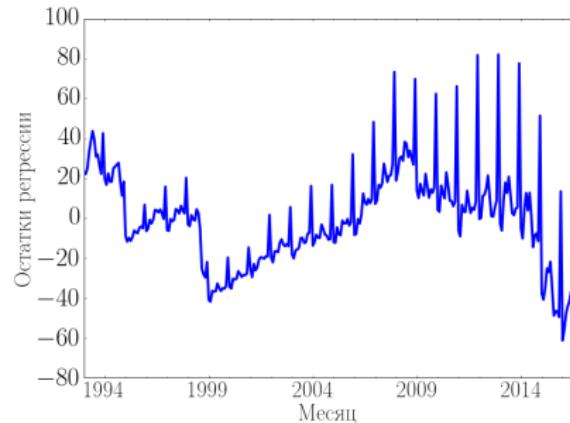
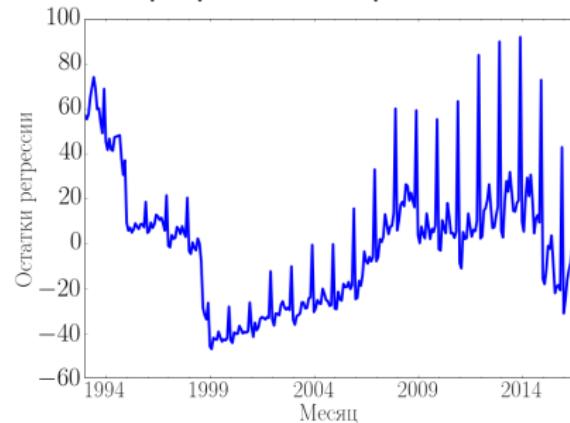
Модель $ARMA(2, 2)$ для преобразованного ряда:

Реальная заработка плата

Модель $SARIMA(2, 0, 2) \times (0, 1, 0)$ с преобразованием Бокса-Кокса:

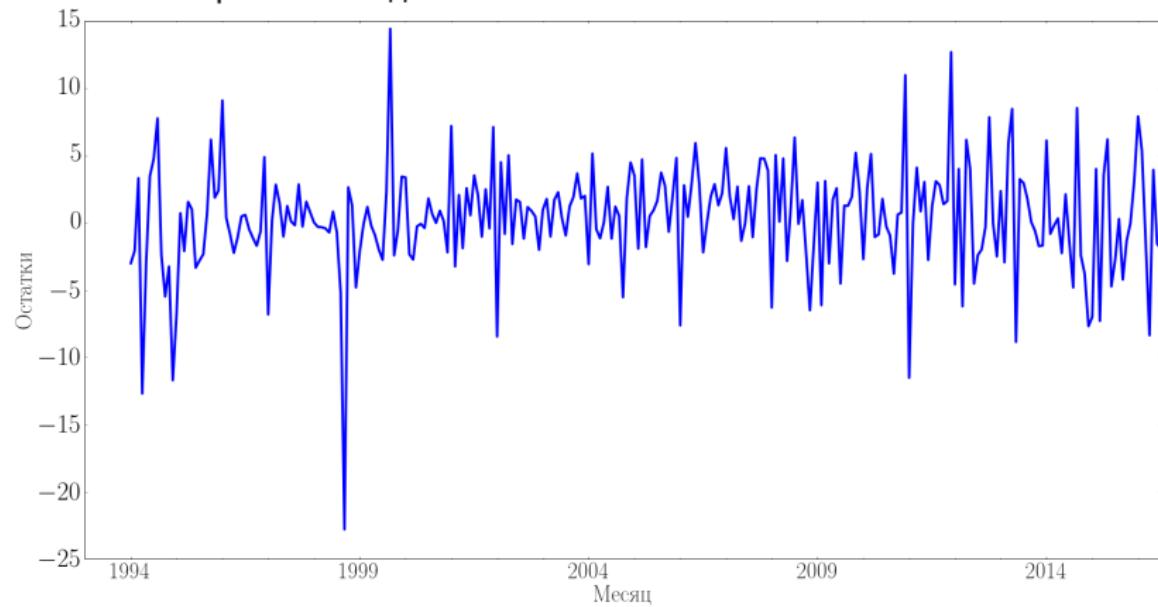
Реальная заработка плата

Остатки регрессий на время:



Реальная заработка плата

Остатки построенной модели:



Подбор параметров



Не рассматривали, пропускаем

- α, ϕ, θ
- d, D
- q, Q
- p, P

- 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 (независимый одинаково распределённый гауссовский), то МНК даёт оценки максимального правдоподобия.

- Порядки дифференцирования подбираются так, чтобы ряд стал стационарным.
- Ещё раз: если ряд сезонный, рекомендуется начинать с сезонного дифференцирования.
- Чем меньше раз мы продифференцируем, тем меньше будет дисперсия итогового прогноза.

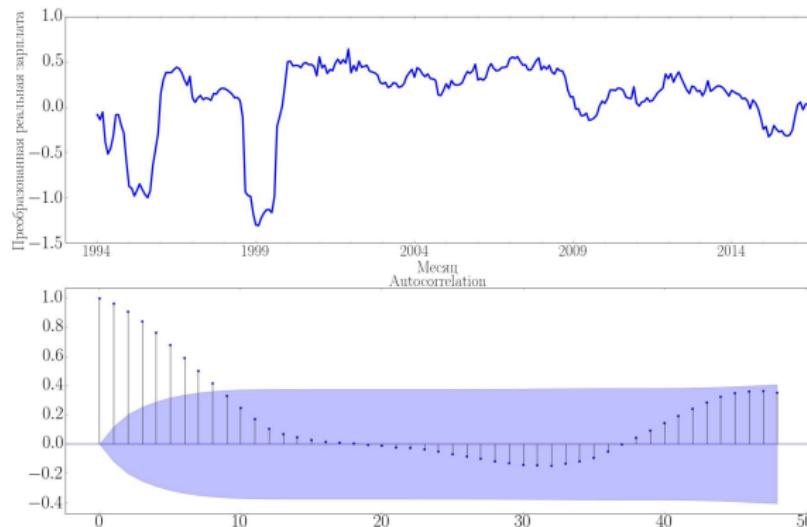
q, Q, p, P

- Гиперпараметры нельзя выбирать из принципа максимума правдоподобия: L всегда увеличивается с их ростом.
- Для сравнения моделей с разными q, Q, p, P можно использовать критерий Акаике:

$$AIC = -2 \log L + 2k,$$

$k = P + Q + p + q + 1$ — число параметров в модели.

- Начальные приближения можно выбрать с помощью автокорреляций.

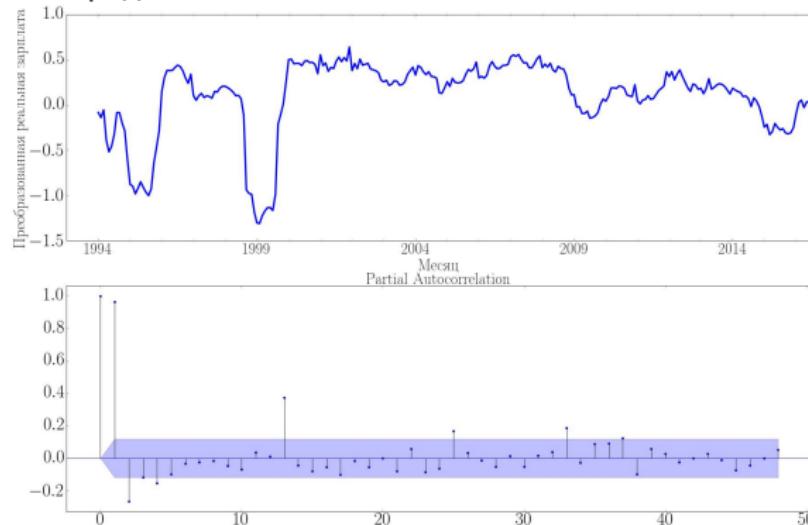
q, Q 

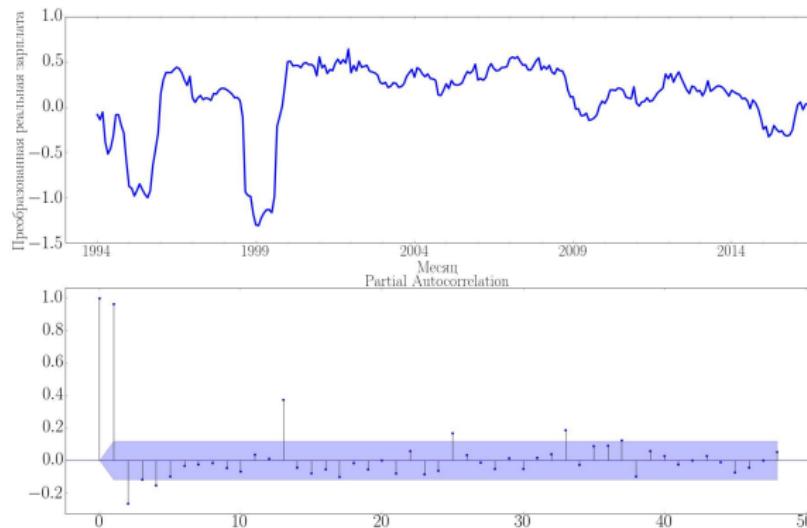
$Q * S$ — номер последнего сезонного лага, при котором автокорреляция значима (здесь 0).

q — номер последнего несезонного лага, при котором автокорреляция значима (здесь 8).

p, P

Частичная автокорреляция — автокорреляция после снятия авторегрессии предыдущего порядка.



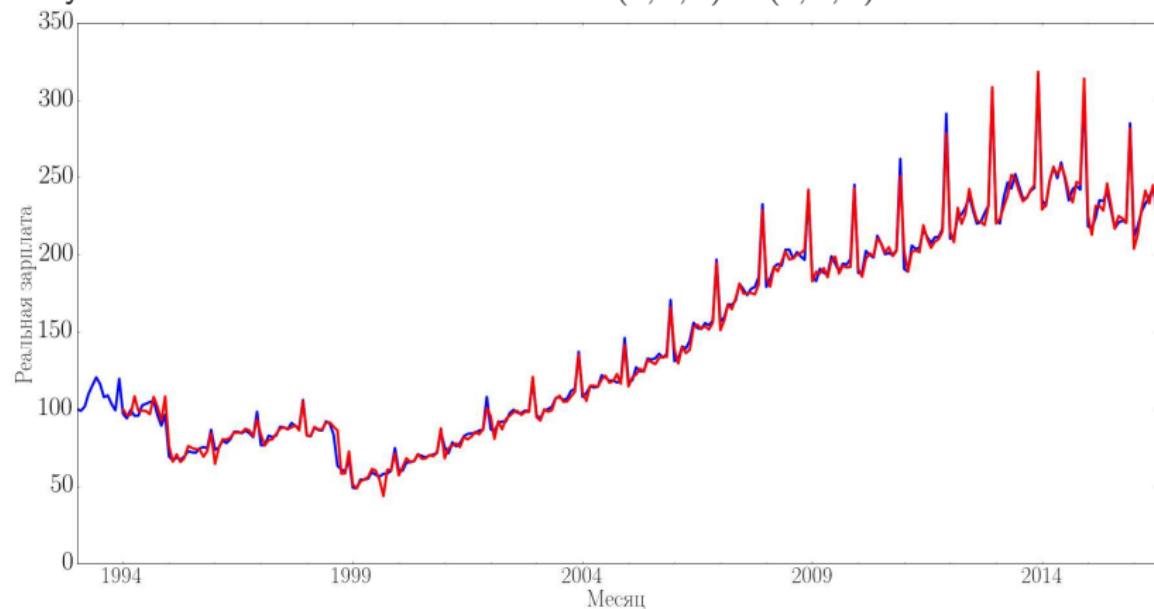
p, P 

$P * S$ — номер последнего сезонного лага, при котором частичная автокорреляция значима (здесь 2).

p — номер последнего несезонного лага, при котором частичная автокорреляция значима (здесь 2).

Реальная заработка плата

Перебирая модели с $D = 1$, $d = 0$ и преобразованием Бокса-Кокса, получаем наименьший AIC на $ARIMA(2, 0, 1) \times (2, 1, 2)$:



Подбор ARIMA

- ❶ Смотрим на ряд.
- ❷ При необходимости стабилизуем дисперсию.
- ❸ Если ряд нестационарен, подбираем порядок дифференцирования.
- ❹ Анализируем ACF/PACF, определяем примерные p, q, P, Q
- ❺ Обучаем модели-кандидаты, сравниваем их по AIC, выбираем победителя.
- ❻ Смотрим на остатки полученной модели, если они плохие, пробуем что-то поменять.

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:

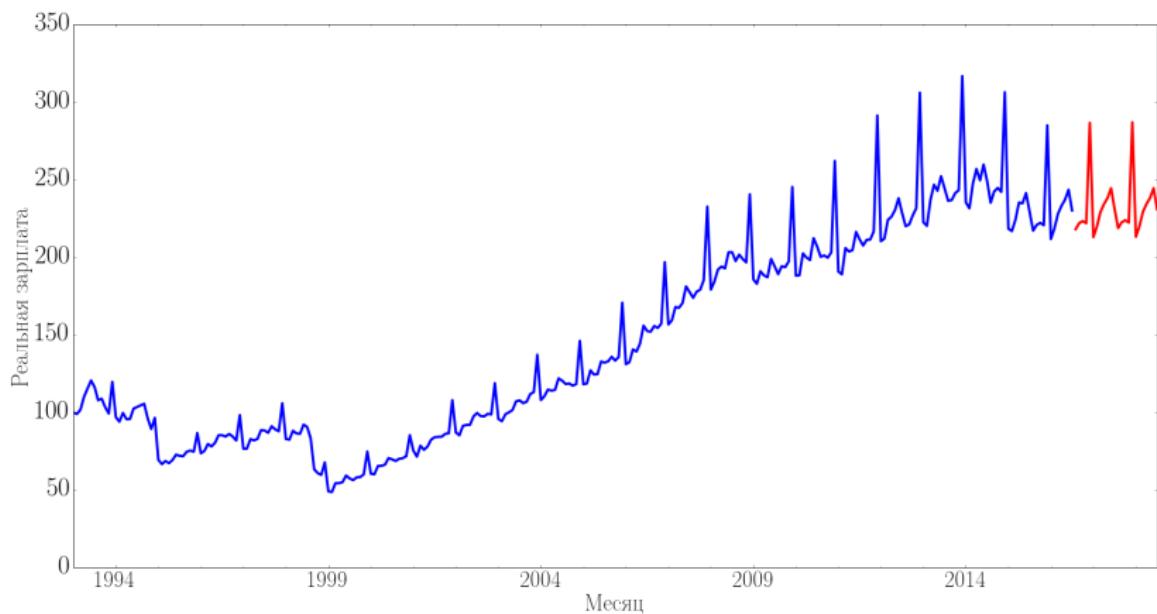
$$\hat{y}_{T+1|T} = \hat{\alpha} + \hat{\phi}_1 y_T + \cdots + \hat{\phi}_p y_{T+1-p} + \hat{\theta}_1 \hat{\varepsilon}_T + \cdots + \hat{\theta}_q \hat{\varepsilon}_{T+1-q}$$

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

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

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

Прогнозирование



Остатки

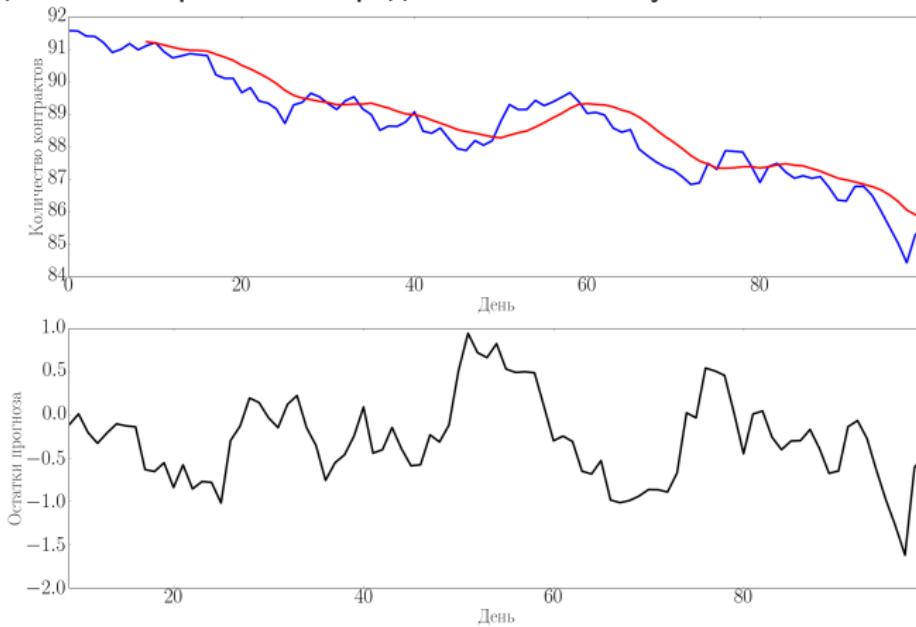
Остатки — разность между фактом и прогнозом:

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

Нужно проверять, обладают ли они некоторыми свойствами.

Несмешённость

Несмешённость — равенство среднего значения нулю:

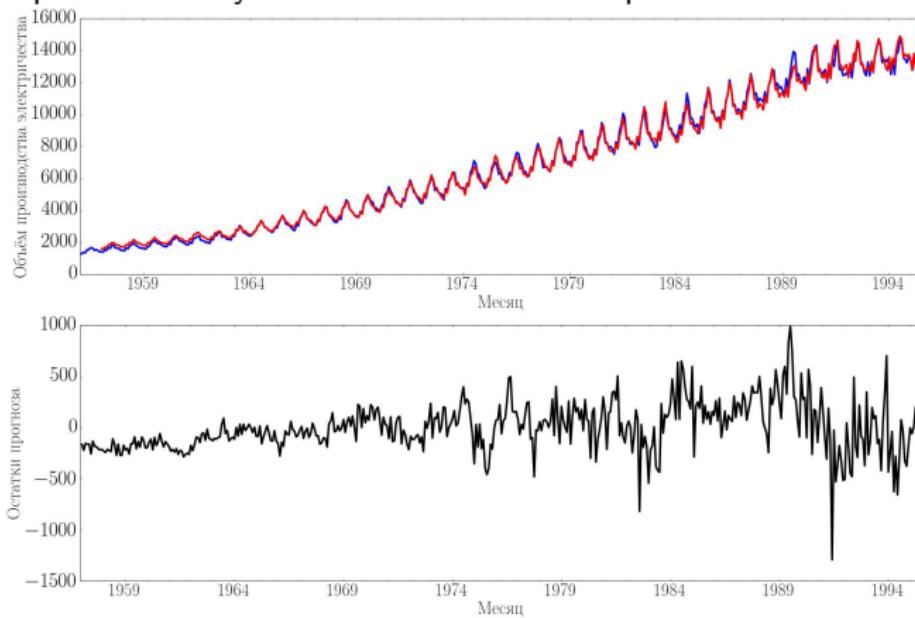


Несмешённость

- 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)

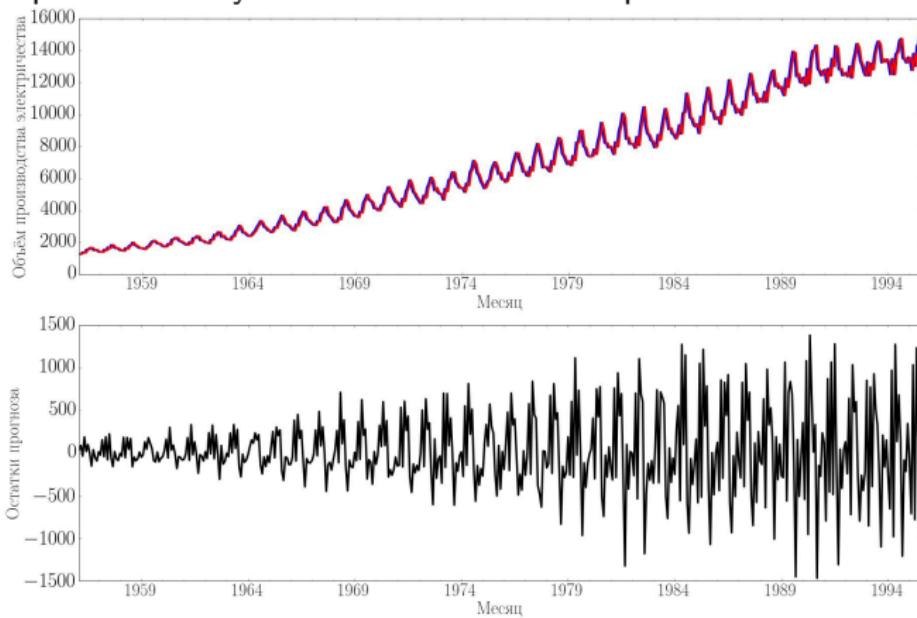
Стационарность

Стационарность — отсутствие зависимости от времени:



Стационарность

Стационарность — отсутствие зависимости от времени:

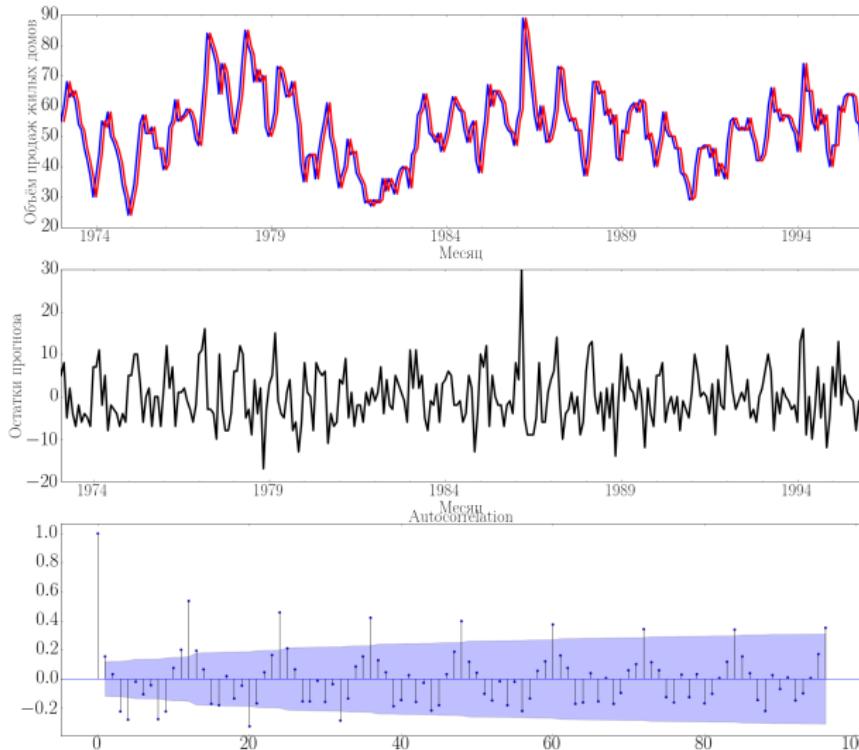


Стационарность

- 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:

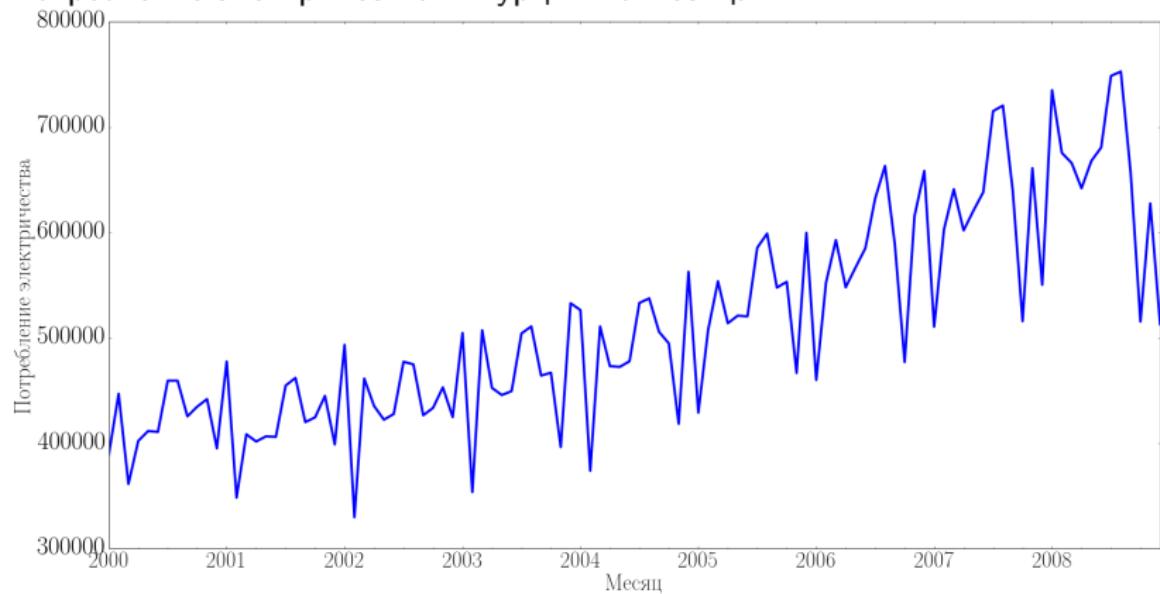


Неавтокоррелированность

- 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

Праздники

Потребление электричества в Турции по месяцам:



Падения соответствуют месяцам, на которые выпадают праздники по исламскому календарю (год примерно на 11 дней короче, чем в грегорианском)

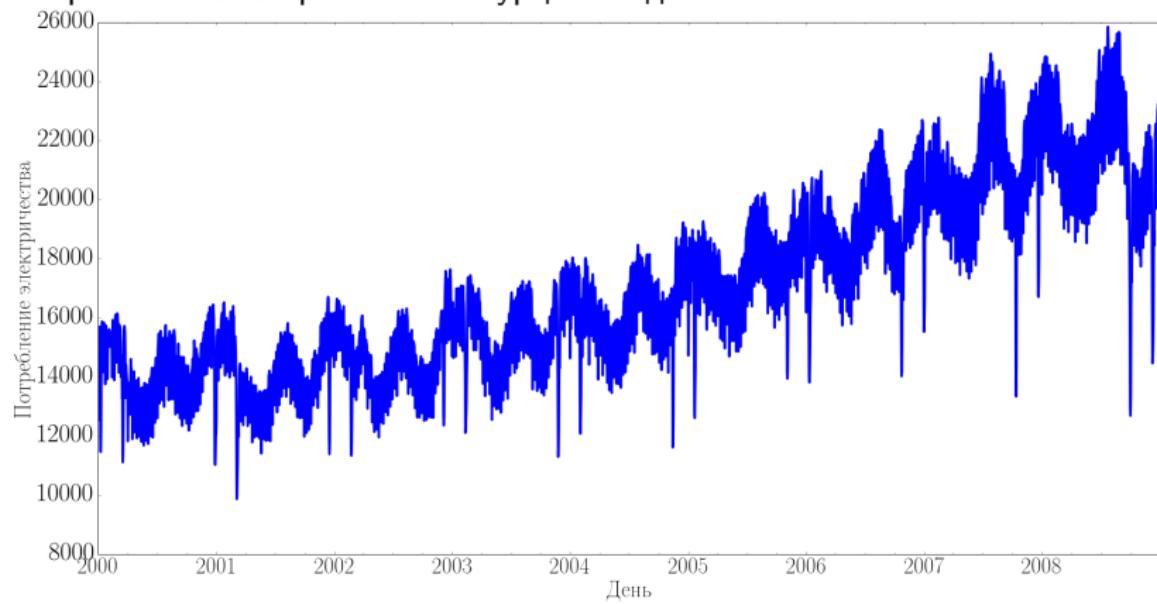
SARIMAX

$$y_t = \sum_{j=1}^k \beta_j x_{jt} + z_t,$$

$$\begin{aligned} z_t = & \alpha + \phi_1 z_{t-1} + \dots + \phi_p z_{t-p} + \\ & + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} + \\ & + \phi_S z_{t-S} + \dots + \phi_{PS} z_{t-PS} + \\ & + \theta_S \varepsilon_{t-S} + \dots + \theta_{PS} \varepsilon_{t-PS} + \varepsilon_t. \end{aligned}$$

Сложная сезонность

Потребление электричества в Турции по дням:



- недельная сезонность;
- годовая сезонность;
- праздники по исламскому календарю.

Сложная сезонность

Сложности:

- при длинных периодах сезонности в модели SARIMA получается слишком много параметров;
- поведение дневного ряда вряд ли определяется его значением ровно 365 дней назад;
- длина года — 365.25 дней и 52.18 недель.

Решение: брать в качестве S период самой короткой сезонности, а сезонность более высоких порядков учитывать регрессией на фурье-гармоники с периодами, например, 365.25, 365.25/2, 365.25/3 и т.д.

Регрессионные признаки

- гармоники по длинным периодам сезонности
- индикаторы номера периода в коротких сезонностях
- индикаторы праздников
- индикаторы пред- и постпраздничных дней
- тренды (линейный, квадратичный и т.д.)
- скользящие средние ряды за предыдущие периоды

При хорошем подборе признаков регрессии часто оказывается достаточно.

Массовое прогнозирование

Пример: дневные продажи товаров в магазинах.

Информация: продажи, остатки, цены, скидки, промо-акции, иерархия товаров, иерархия и расположение торговых точек.

Задача: построить прогнозы продаж всех товаров во всех магазинах.

Проблема: ручной подбор прогнозирующих моделей для каждой пары товар-магазин невозможен.

Решение: регрессионная модель с хорошо подобранными признаками.

Резюме

- в ARIMA можно учитывать внешние факторы
- при хорошем подборе внешних факторов специфические модели временных рядов часто оказываются не нужны.

Литература

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