Временные ряды

Модели и прогнозирование

План

- Понятие временного ряда
- Модели временных рядов
- Примеры задач прогнозирования
- Свойства временных рядов
- Модель ARIMA
- Метрики точности прогноза
- Методы прогнозирования

Исходные статистические данные

Если рассматривать значения одного признака у одного объекта в равноотстоящие моменты времени, то последовательность $\boldsymbol{x(t_1), x(t_2), \dots, x(t_N)}$ называют *одномерным временным рядом.*

Если регистрировать значения *p* признаков у одного объекта, то говорят о статистическом анализе *многомерного временного ряда*

$\boldsymbol{X(t) = (x^1(t_k), x^2(t_k), \dots, x^p(t_k))}$, *k=1,2, …, N*

Исходные статистические данные

Говоря о проблеме прогнозирования на основе одномерных временных рядов, обычно имеется ввиду ***кратко-*** и ***среднесрочный*** прогноз, поскольку построение долгосрочного прогноза подразумевает обязательное использование методов организации и статистического анализа *специальных экспертных оценок.*

Использование доступных к моменту $t=N$ наблюдений временного ряда $x(t)$ для прогнозирования может явиться основой для:

- планирования в экономике, производстве, торговле
- управления и оптимизации социально-экономических процессов
- принятия оптимальных решений в бизнесе
- частичного управления параметрами демографических процессов

Модели временных рядов

При построении эконометрических регрессионных моделей для временных рядов возникает ряд особенностей, которые необходимо учесть:

- ✓ упорядоченность во времени (хронологический порядок);
- ✓ зависимость от прошлого («память», серийная или автокорреляция);
- ✓ различаются краткосрочные и долгосрочные зависимости и модели;
- ✓ часто встречается феномен «ложной регрессии»;
- ✓ бывает небольшое число наблюдений (как правило при работе с макроданными), которое невозможно увеличить (т.к. изменяется вид или структура зависимости)

Модели временных рядов

Наиболее распространённые модели временных рядов (одномерные и многомерные):

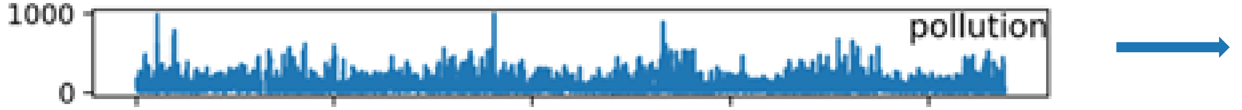
1. стационарные ряды;
2. стационарные относительно тренда или TS-ряды;
3. ряды с единичным корнем или DS-ряды;
4. ряды с переменной волатильностью или с условной гетероскедастичностью.

Для каждой модели существуют свои подходы к оцениванию и по строению регрессий

Одномерное и многомерное прогнозирование

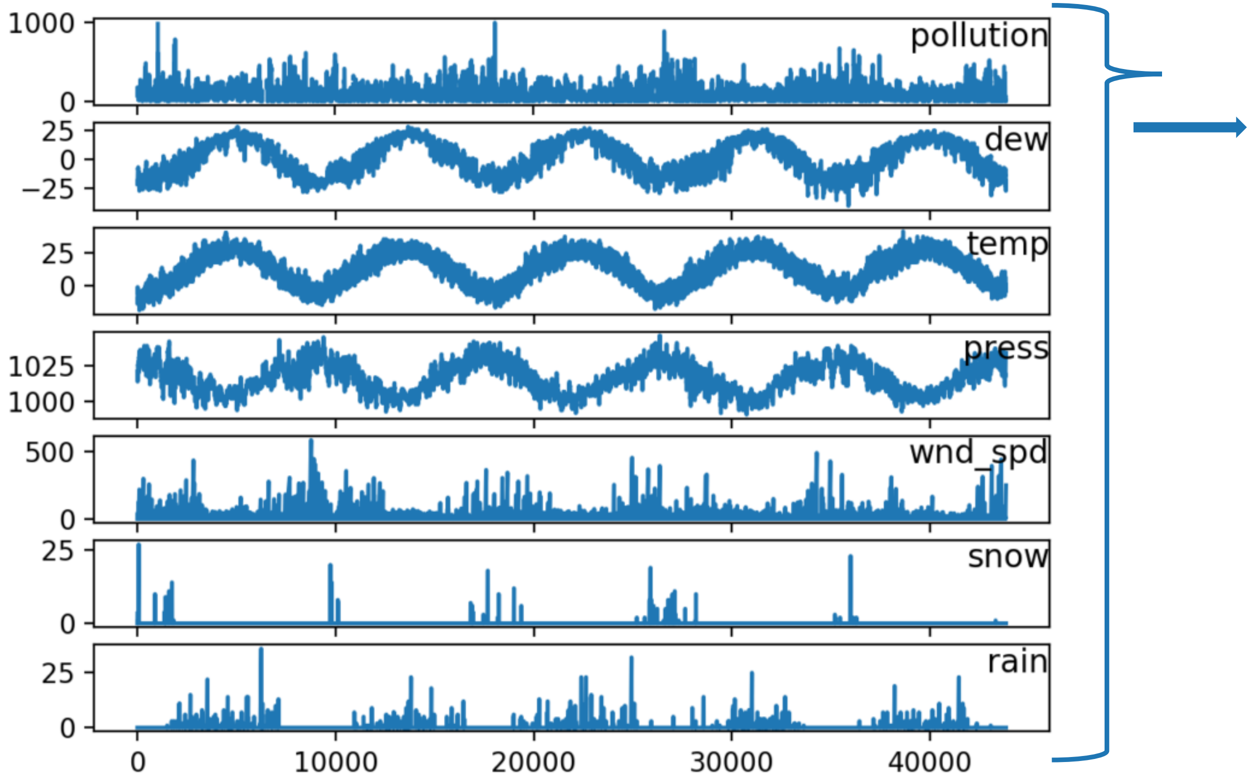
Одномерный (Univariate):

- Один целевой временной ряд
- Прогнозирование только на его основе



Многомерное (Multivariate):

- Один целевой временной ряд
- Несколько характеристик за один и тот же период времени, которые могут повлиять на результат (курс валюты, температура, уровень безработицы и др.)
- Прогноз на основе полных данных



Общие модели временных рядов

Аддитивная форма

$$\boldsymbol{x(t) = \lambda_1 f(t) + \lambda_2 \varphi(t) + \lambda_3 \psi(t) + \varepsilon(t)}, \quad \textbf{(1)}$$

$$\lambda_i = \begin{cases} 1, & \text{фактор участвует в формировании уровней ряда } x(t) \\ 0, & \text{в противном случае} \end{cases}$$

Мультипликативная форма

$$\boldsymbol{x(t) = f(t)^{\lambda_1} * \varphi(t)^{\lambda_2} * \psi(t)^{\lambda_3} * \varepsilon(t)}$$

$$\boldsymbol{ln\, x(t) = \lambda_1\, ln\, f(t) + \lambda_2\, ln\, \varphi(t) + \lambda_3\, ln\, \psi(t) + ln\, \varepsilon(t)}$$

Задачи анализа временных рядов

По имеющейся траектории анализируемого временного ряда $x(t)$ требуется:

- определить какие из неслучайных составляющих $f(t)$, $\varphi(t)$ и $\psi(t)$ присутствуют в разложении (1)
- Построить «хорошие» оценки для тех неслучайных функций, которые присутствуют в разложении (1)
- Подобрать модель, адекватно описывающую поведение «случайной» составляющей $S(t)$, и статистически оценить параметры этой модели

Основные факторы временных рядов

1. **Долговременные**, формирующие общую тенденцию в изменении анализируемого признака $x(t)$. Обычно описывается при помощи монотонной функции $f(t)$, называемой *трендом*.
2. **Сезонные**, формирующие периодически повторяющиеся в определенное время года колебания анализируемого признака. Описывается периодической функцией $\varphi(t)$ с периодом, кратным сезонам.
3. **Циклические**, формирующие изменения анализируемого признака, обусловленные действием долговременных циклов экономической, демографической или астрономической природы. Описывается функцией $\psi(t)$.
4. **Случайные**, не поддающиеся учету и регистрации. Их воздействие обуславливает *стохастическую природу* анализируемого признака. Обозначается $S(t)$.

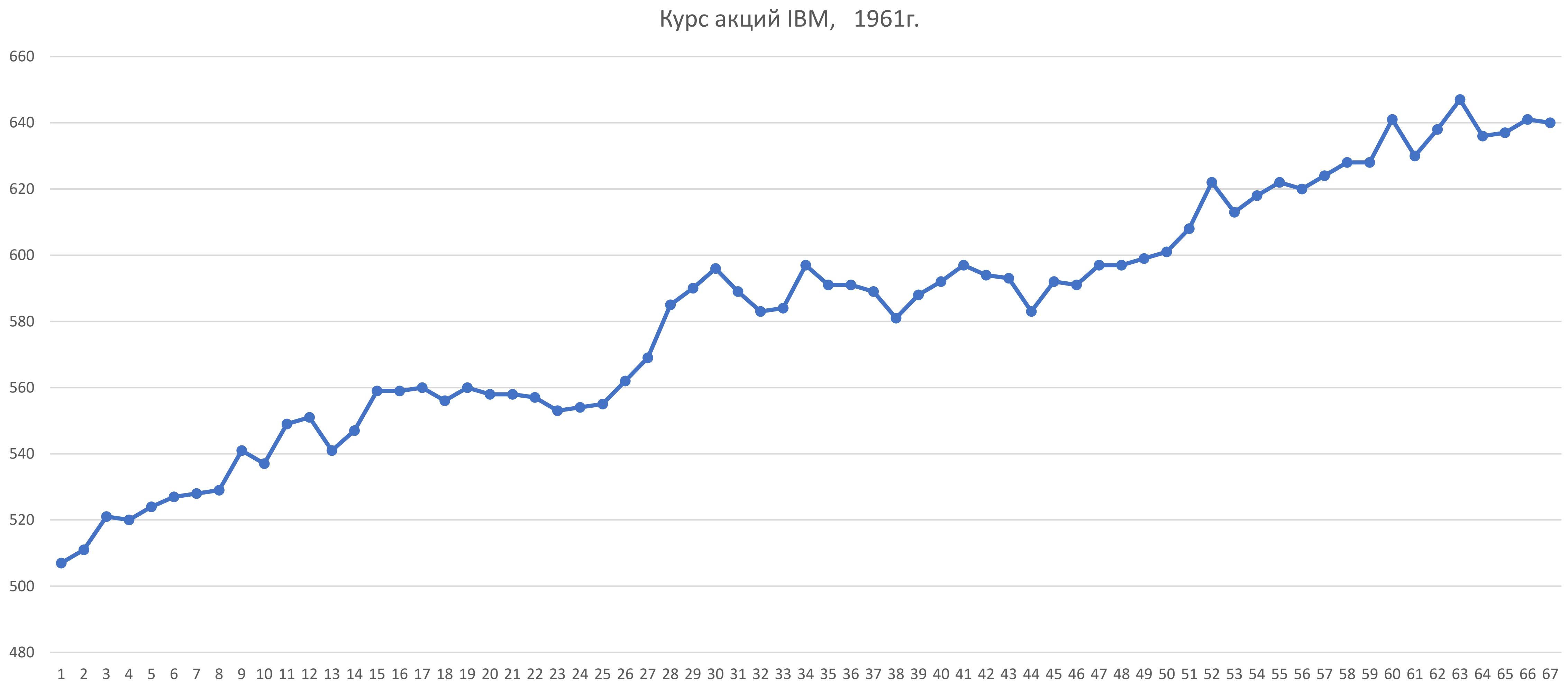
Примеры задач прогнозирования

В классической задаче анализа данных мы предполагаем независимость всех наблюдений. В случае анализа временных рядов мы исходим из гипотезы о том, что предсказываемое значение зависит от предыдущих.

Примеры задач прогнозирования:

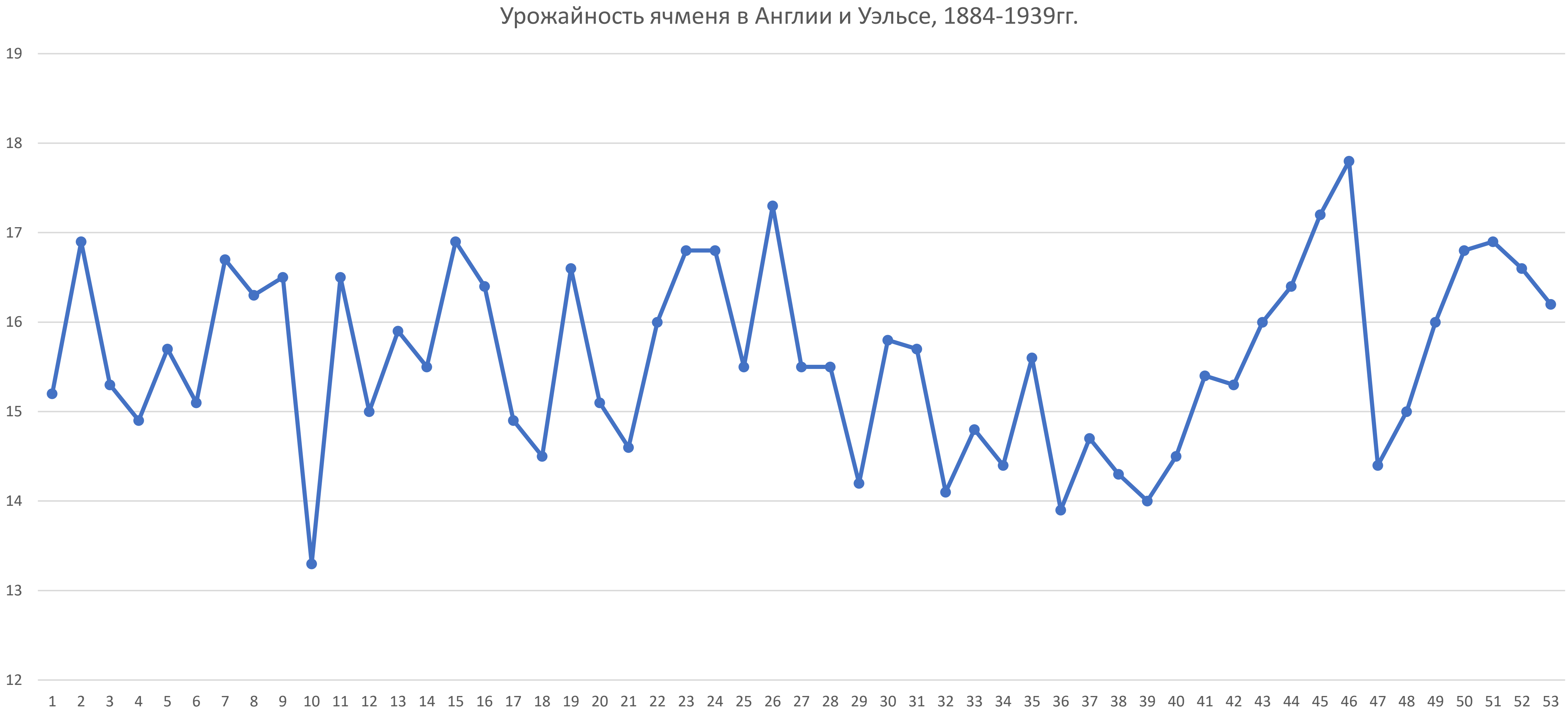
- курс валюты;
- стоимость акций компании "Яндекс";
- спрос на определённый продукт;
- количество студентов без долгов в определенный момент времени;
- процент посещаемости лекций по мат. анализу;
- уровень безработицы;
- ...

Примеры временных рядов

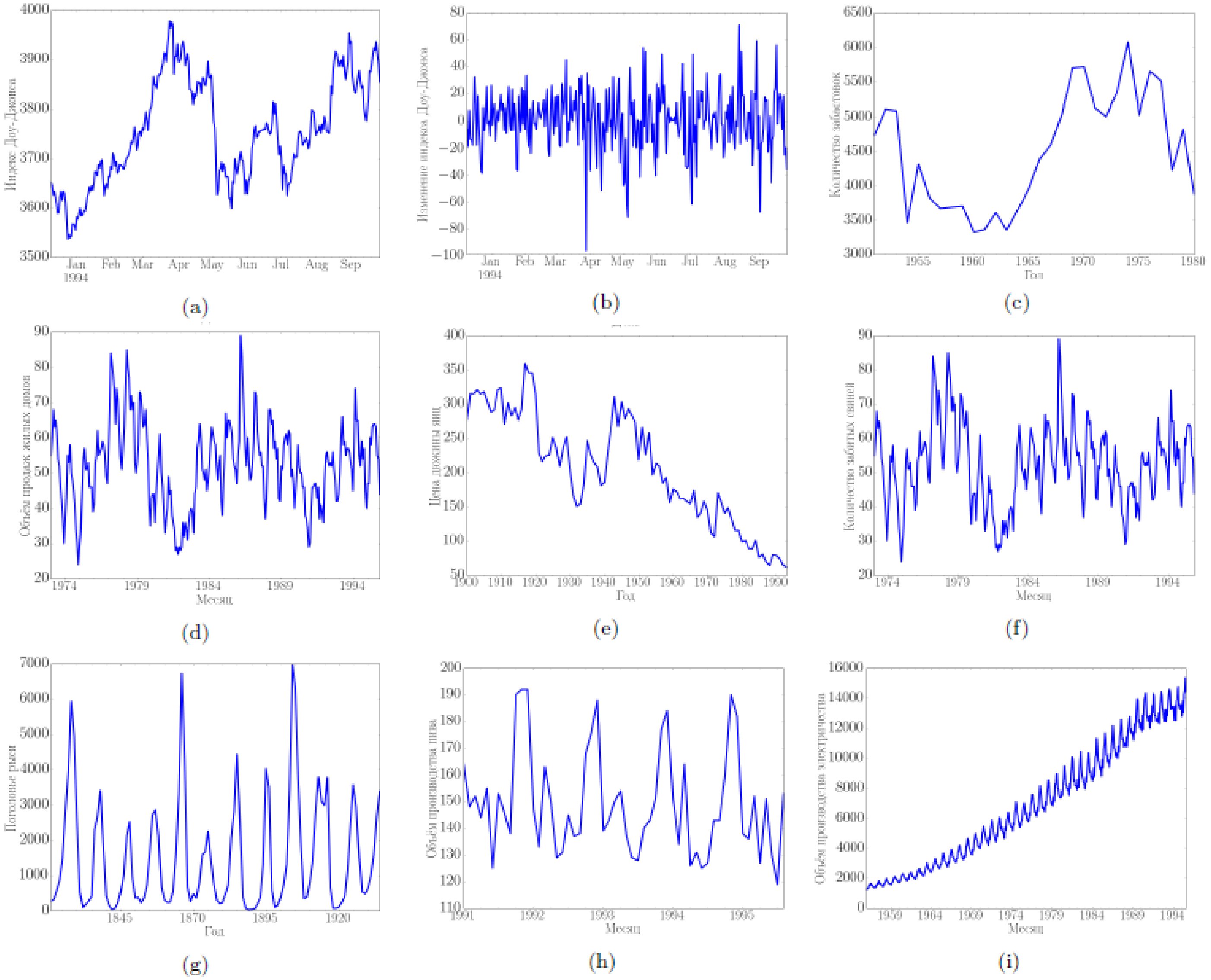


Примеры временных рядов

Примеры временных рядов



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



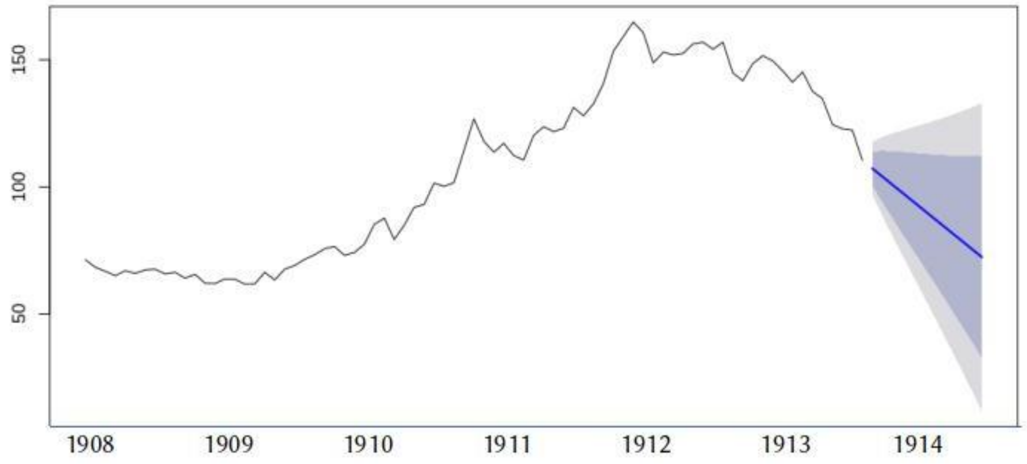
Анализ временных рядов

Задачи:

- Поиск аномалий
- Поиск локальных трендов
- (локальные) максимумы и минимумы
- Корреляция с внешними характеристиками (новости, внешние переменные, стоимость валюты и т. д.)
- ПРОГНОЗИРОВАНИЕ

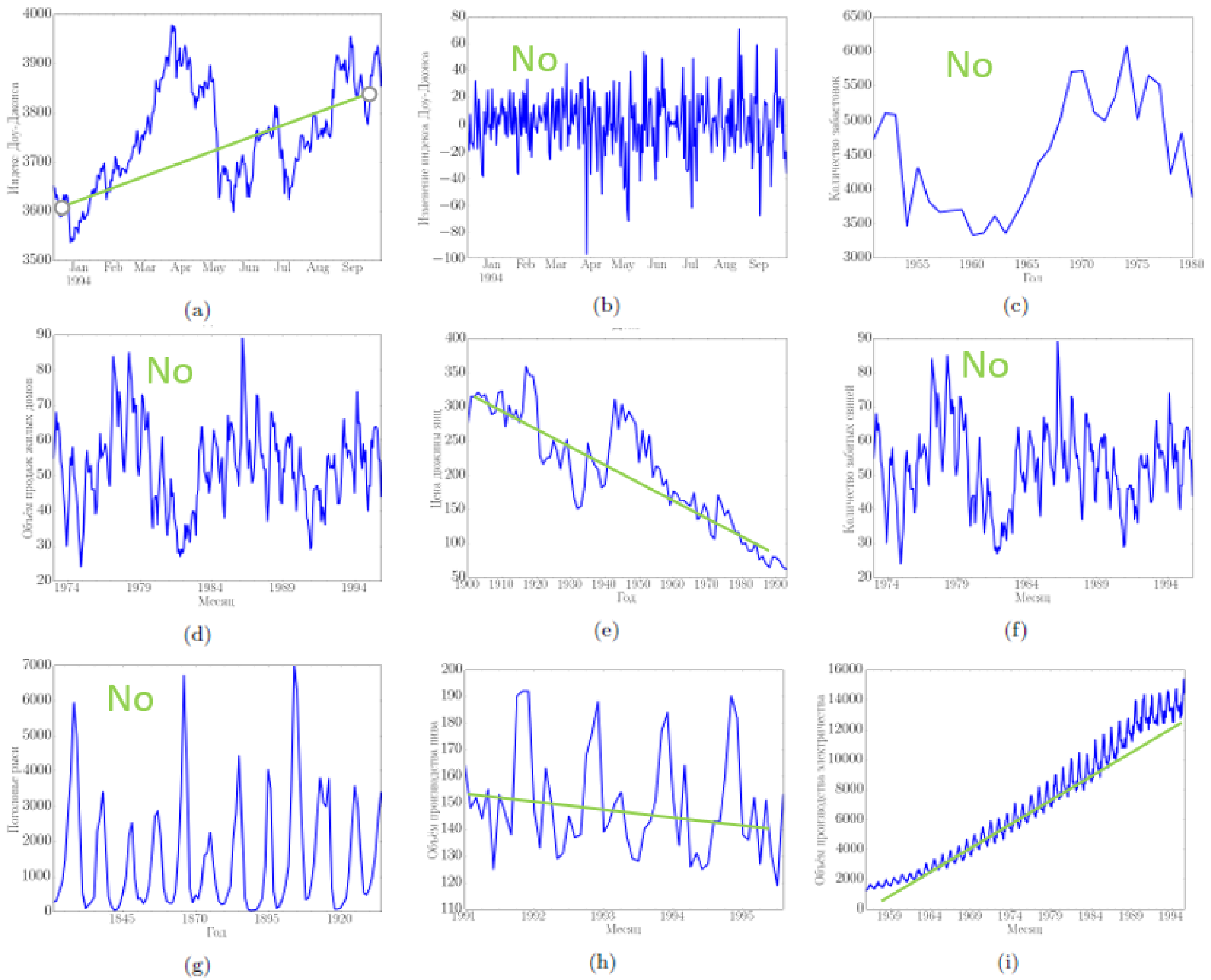
Свойства временных рядов:

- Тренд
- Сезонность
- Цикл(ы)
- Ошибки (шум)
- Стационарность



Свойства временных рядов: тренд

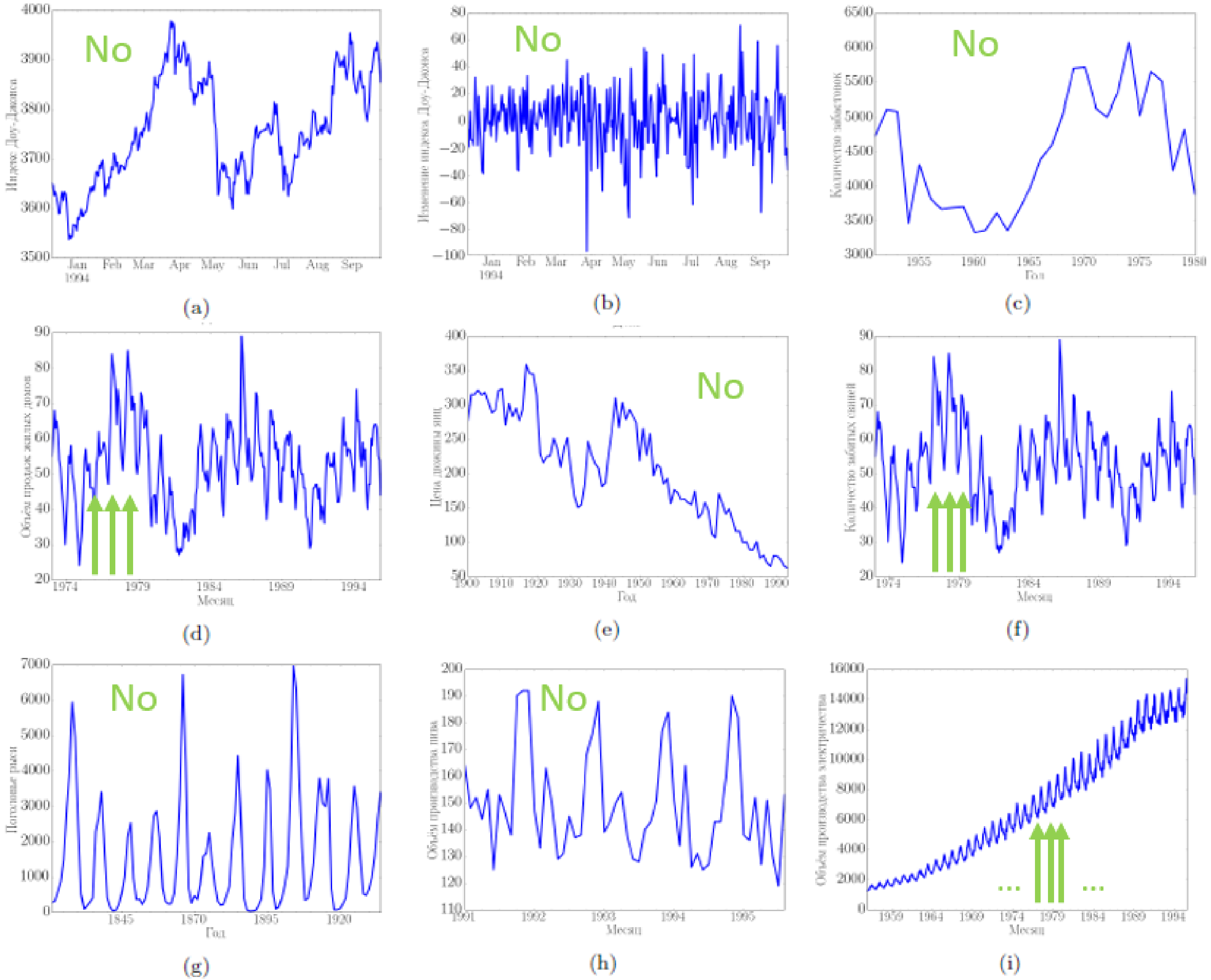
Тренд - изменение значений ряда в долгосрочной перспективе



Свойства временных рядов: сезонность и цикл

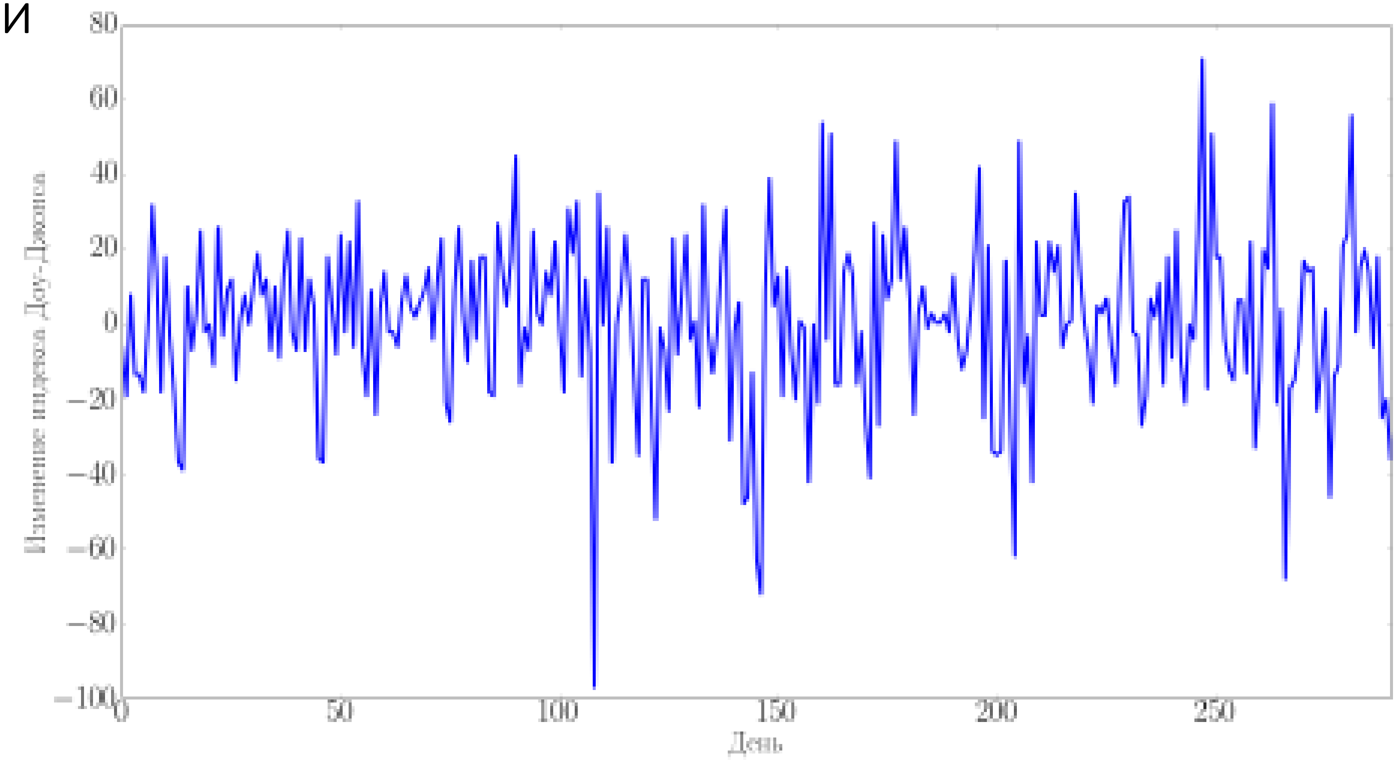
Сезонность – это циклические изменения уровня ряда с постоянным периодом.

Цикл – это изменение уровня ряда с переменным периодом.



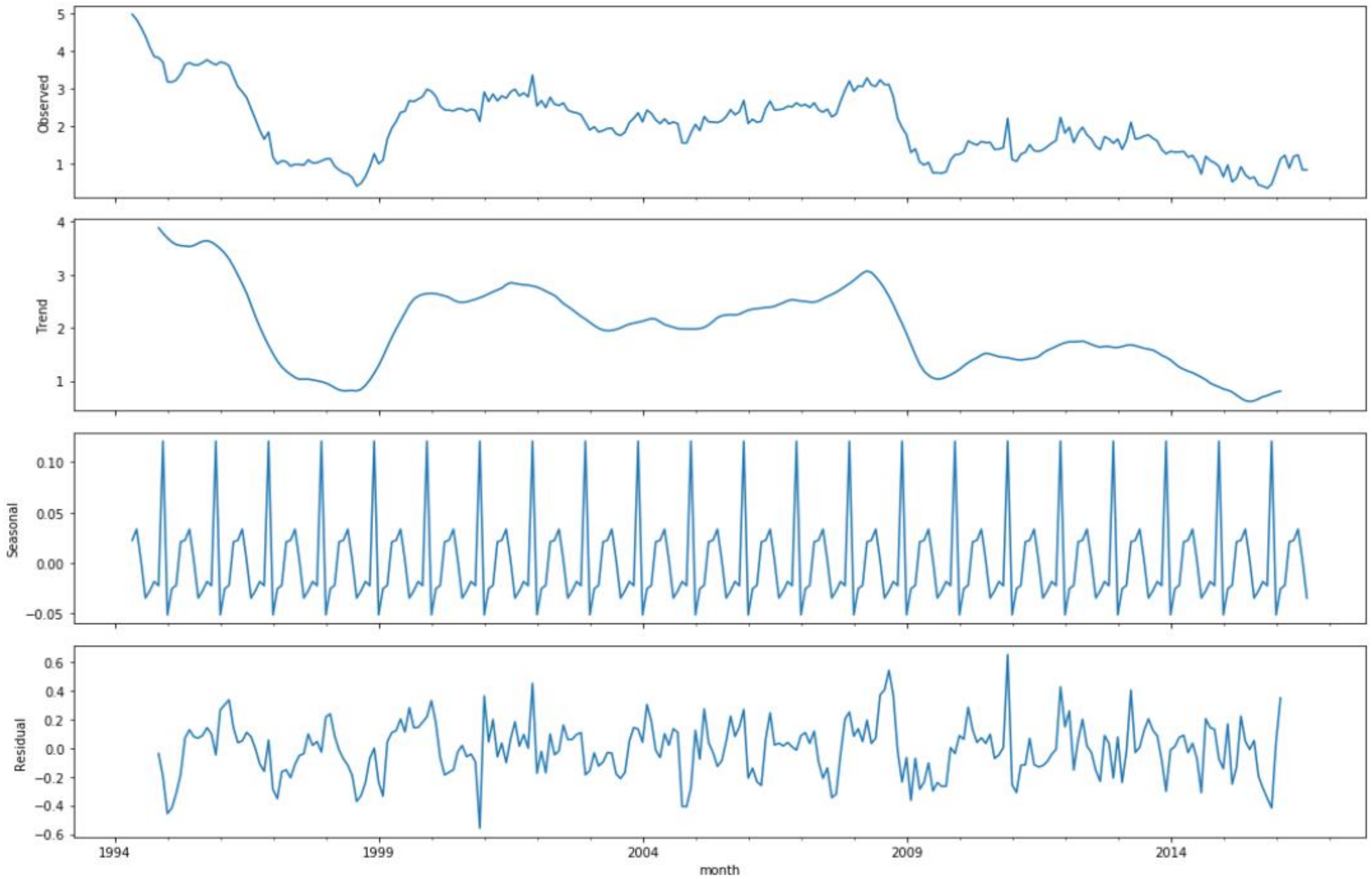
Свойства временных рядов: шум

Шум – это непредсказуемый случайный компонент временных рядов.



- несистематическое поведение: нет тренда, нет сезонности, нет циклов…
- случайная составляющая;
- ~ небольшие отклонения;

Компоненты временных рядов



Автокорреляция (I)

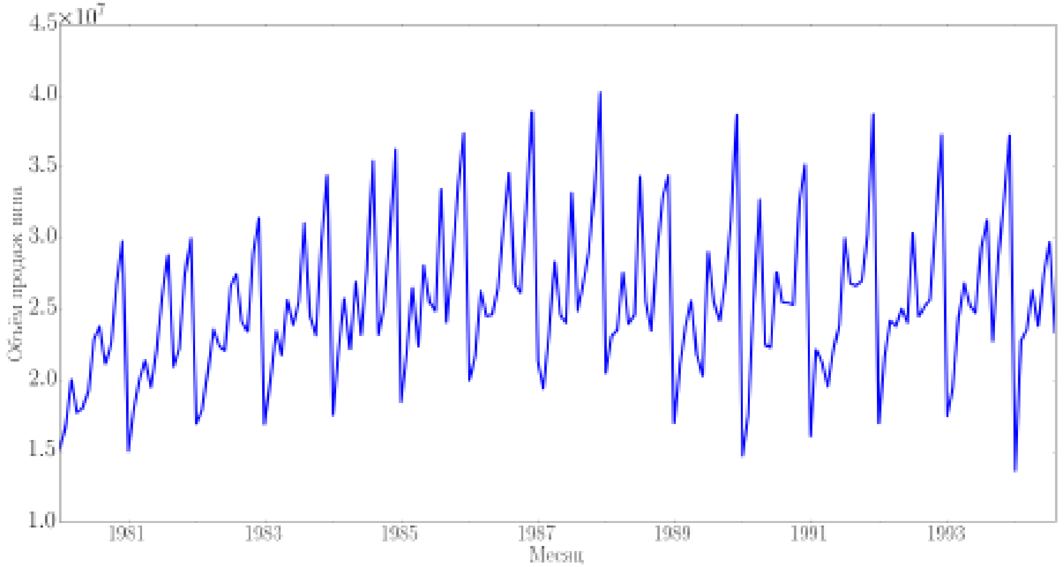
Автокорреляция – это статистическая взаимосвязь между последовательностями величин одного ряда, взятыми со сдвигом.

Автокорреляционная функция для лага τ:

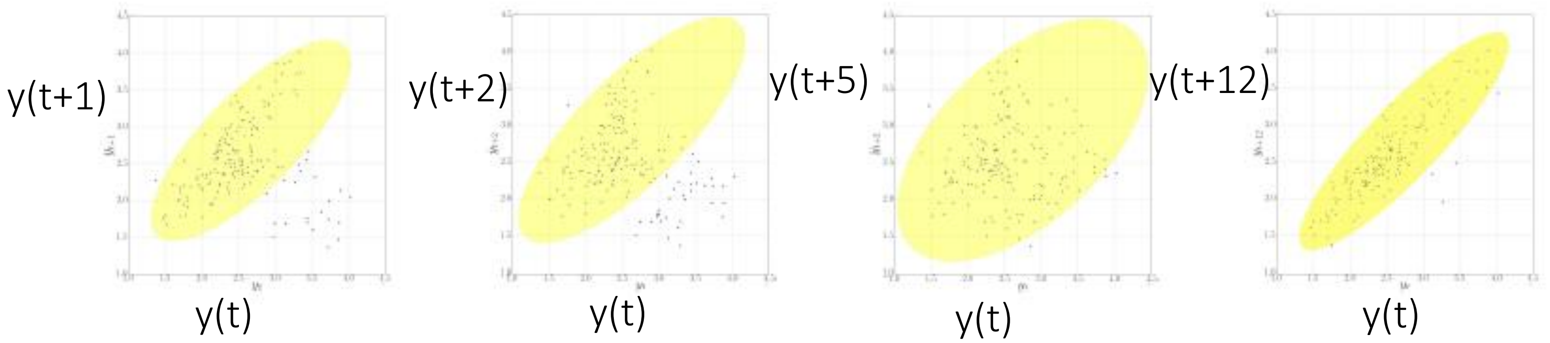
$$r_\tau = \frac{\sum_{t=1}^{T-\tau}(y_t - \bar{y})(y_{t+\tau} - \mathbb{E}y)}{\sum_{t=1}^{T-\tau}((y_t - \bar{y}))^2}.$$

Корреляционная функция Пирсона между значением временного ряда в момент времени (t) и $(t + \tau)$.

Ежемесячный объем продаж вина в Австралии (# бутылки)

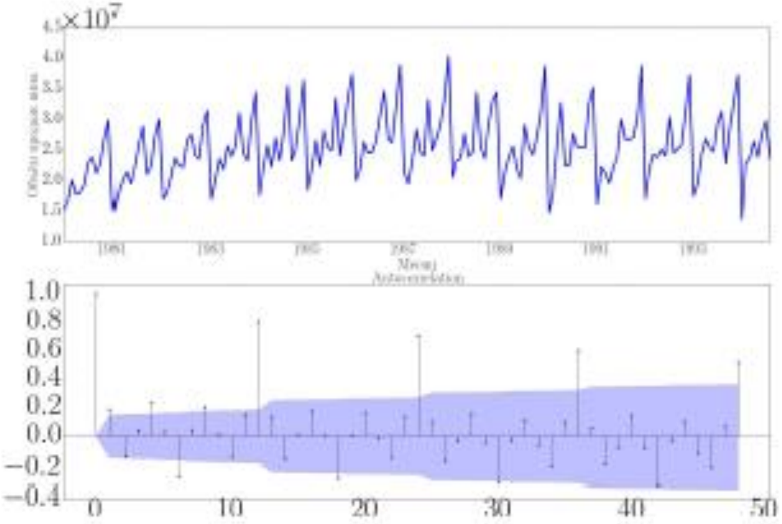


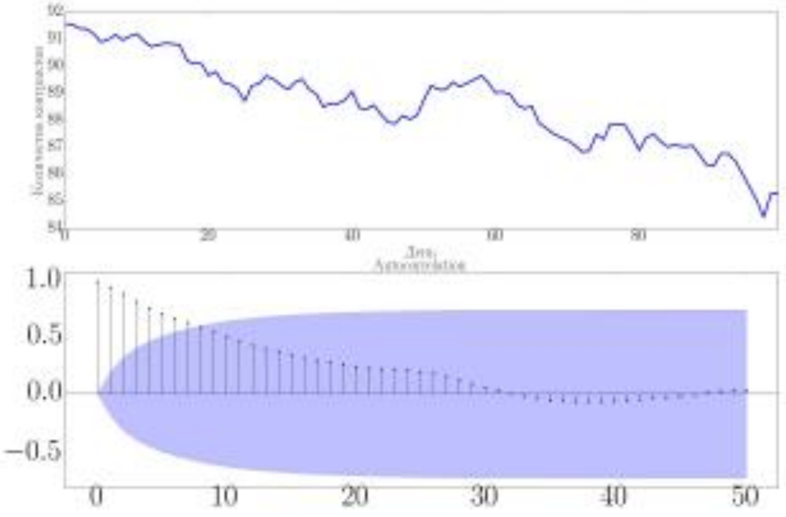
Зависимость значений от предыдущих шагов

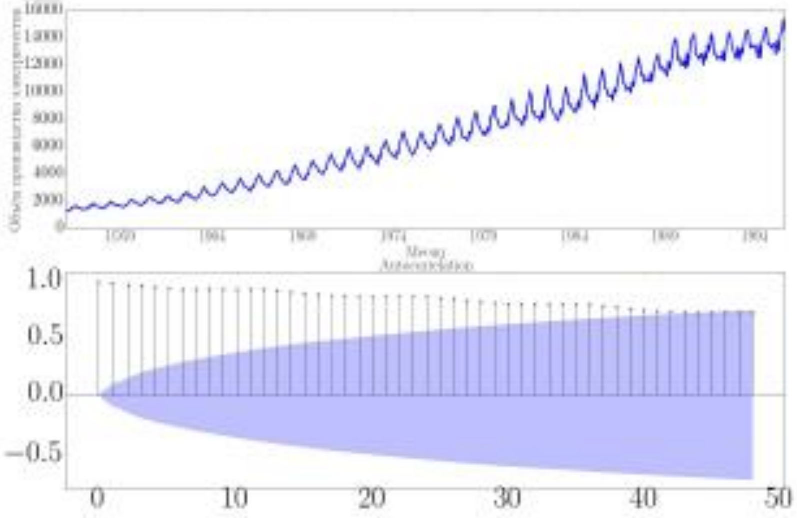


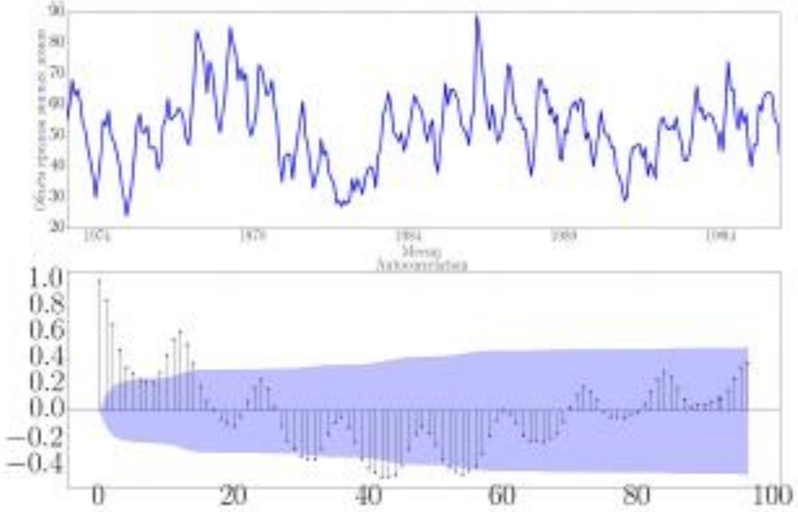
Автокорреляция (II)

Примеры:





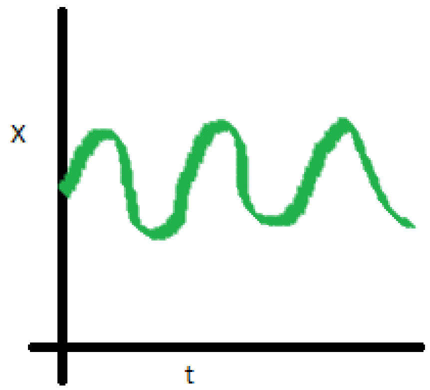




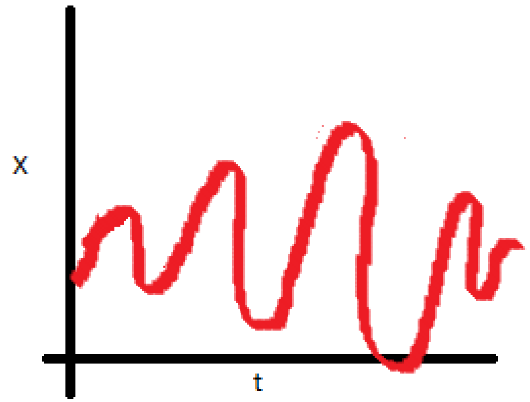
Свойства временных рядов: стационарность

Стационарность – это свойство процесса не менять своих статистических характеристик с течением времени, а именно постоянство матожидания, постоянство дисперсии (гомоскедастичность) и независимость ковариационной функции от времени (должна зависеть только от расстояния между наблюдениями).

Изменение дисперсии

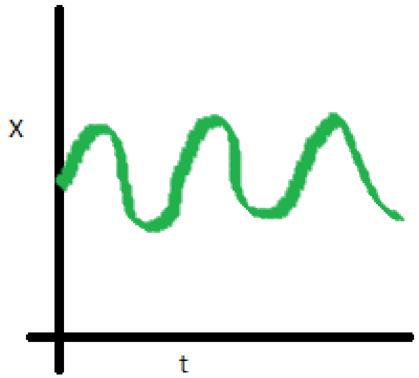


Stationary series

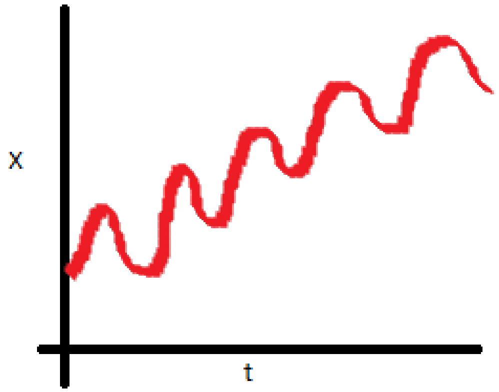


Non-Stationary series

Изменение матожидания

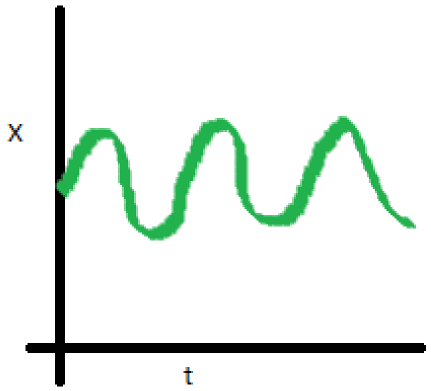


Stationary series

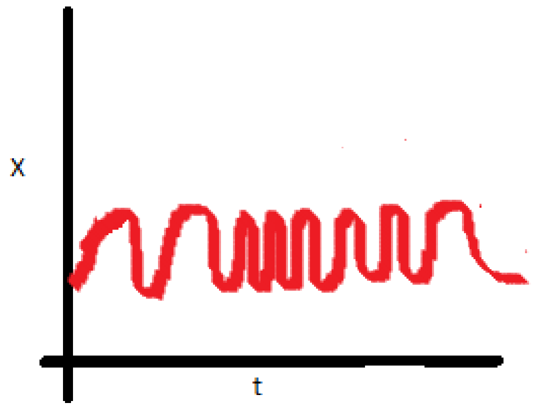


Non-Stationary series

Непостоянство ковариаций



Stationary series



Non-Stationary series

Операции с временными рядами

- Дифференцирование (derivative):

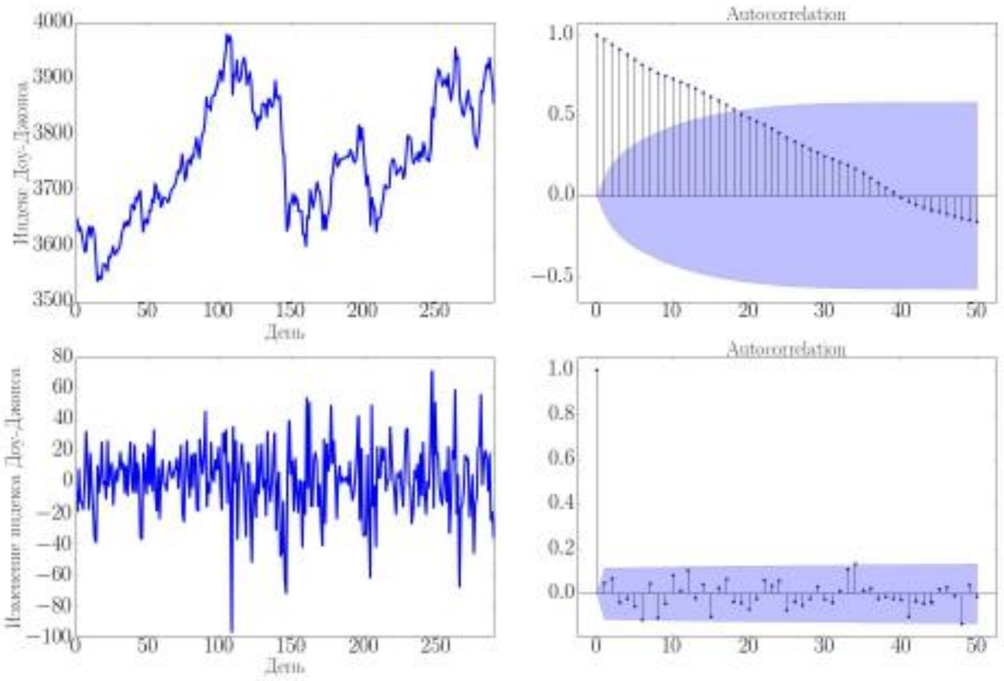
$$y' = y_t - y_{t-1}.$$

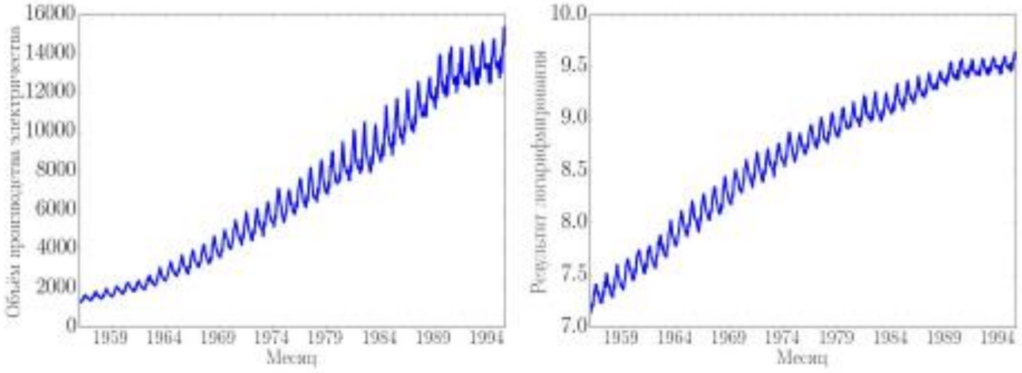
- Сезонное дифференцирование
Seasonal derivative: $y'_t = y_t - y_{t-s}.$

- Нормализация дисперсии
(преобразование Бокса-Кокса):

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

- Тест на стационарность
(Критерий Дики-Фуллера):
  H_0 – non-stationarity
  H_1 – stationarity





Модель ARIMA (I)

autoregressive integrated moving average

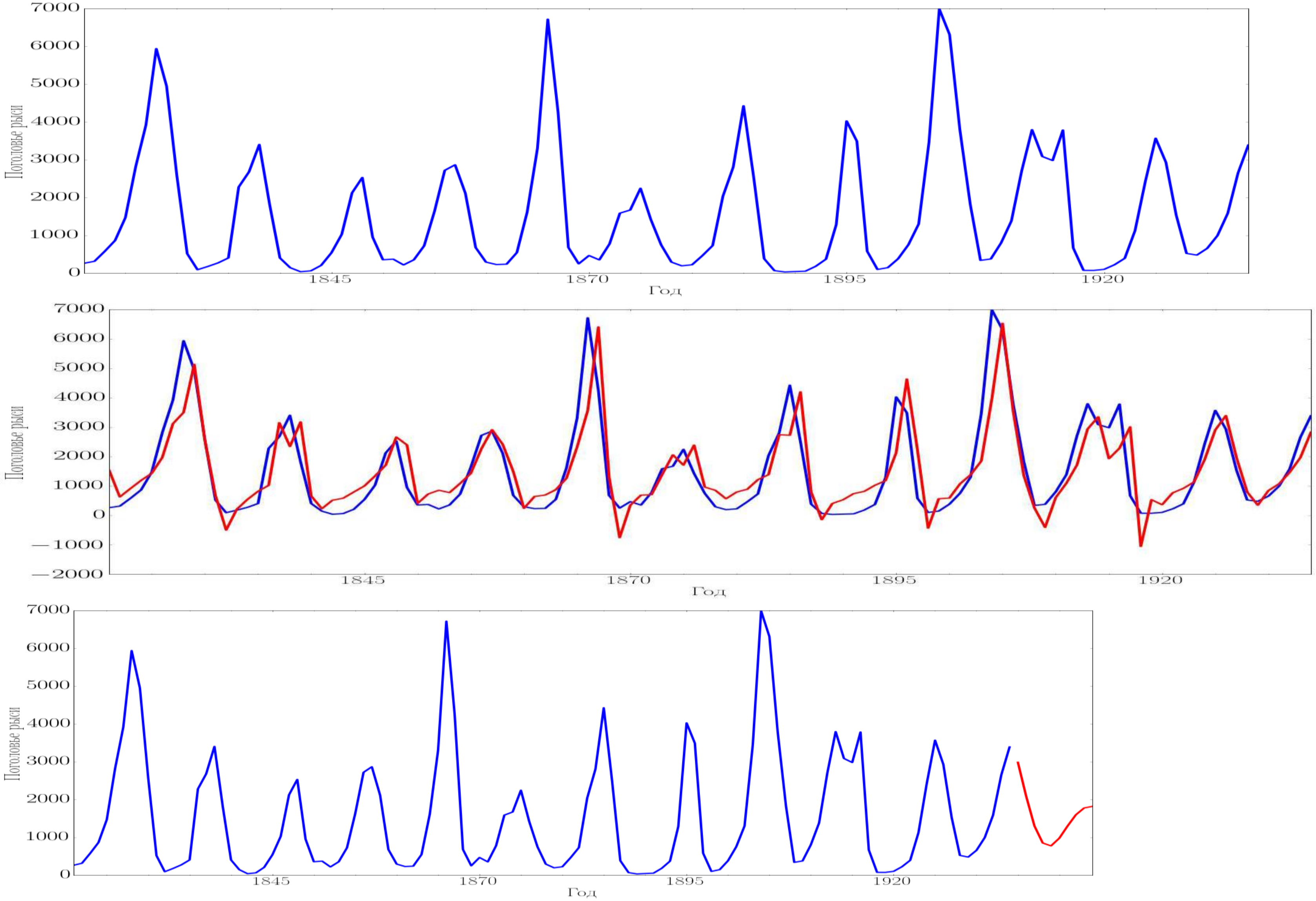
- Показывает хорошие результаты в прогнозировании авторегрессионных временных рядов с сильной сезонностью;
- Необходима индивидуальная тонкая настройка для каждого нового примера.

Компоненты:

- AR(p), авторегрессионная компонента: $y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t.$
- MA(q), компонента скользящего среднего: $y_t = \alpha + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q},$
- ARMA(p,q): $y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}.$

Модель ARIMA (II)

ARMA(2,2)



Модель ARIMA (III)

Wold's theorem:

Каждый стационарный временной ряд может быть аппроксимирован моделью ARMA (p, q) с заданной точностью.

Временной ряд должен быть **стационарен:**

- Преобразование Бокса-Кокса (log)
- Дифференцирование (одношаговое или сезонное)

$\Rightarrow$ ARIMA(p,d,q) – модель ARMA для временных рядов, где d-**порядок дифференцирования (взятия последовательной разности)**

Сезонность

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

+ P components with period S

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

+ Q components with period S

$\Downarrow$

SARMA(p,q)x(P,Q)

Модель ARIMA (IV)

Необходимо найти значения (P,Q,p,q).

Минимизация информационного критерия Акаике (Akaike info criterion): AIC = 2 lnL + 2k

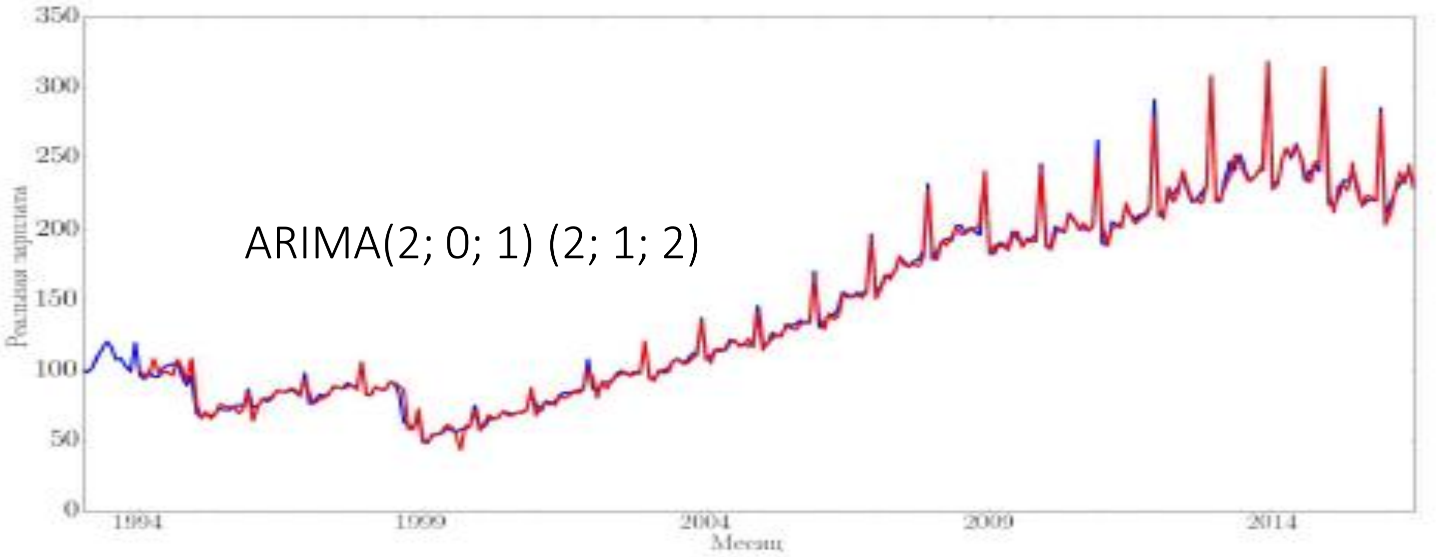
L - Функция правдоподобия

k = P + Q + p + q + 1 – число параметров модели

Лучшая модель - модель ARIMA(p,q)x(P,Q) с минимальным значением AIC.

SARMA(p,q)x(P,Q)
+ d – порядок дифференцирования
+ D – порядок сезонного дифференцирования

= модель **SARIMA(p,d,q)x(P,D,Q)**

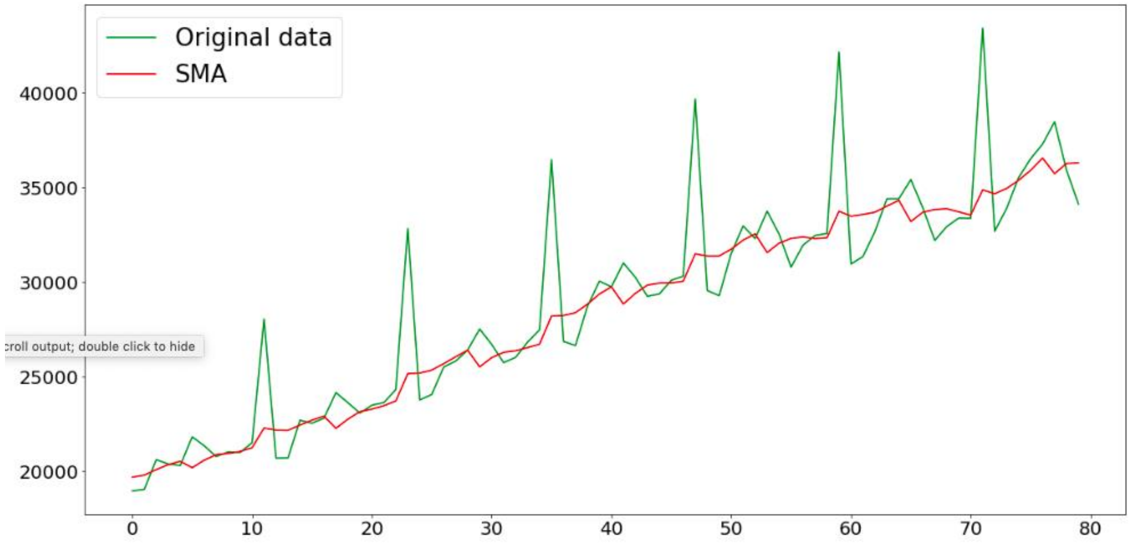


Метрики точности прогноза

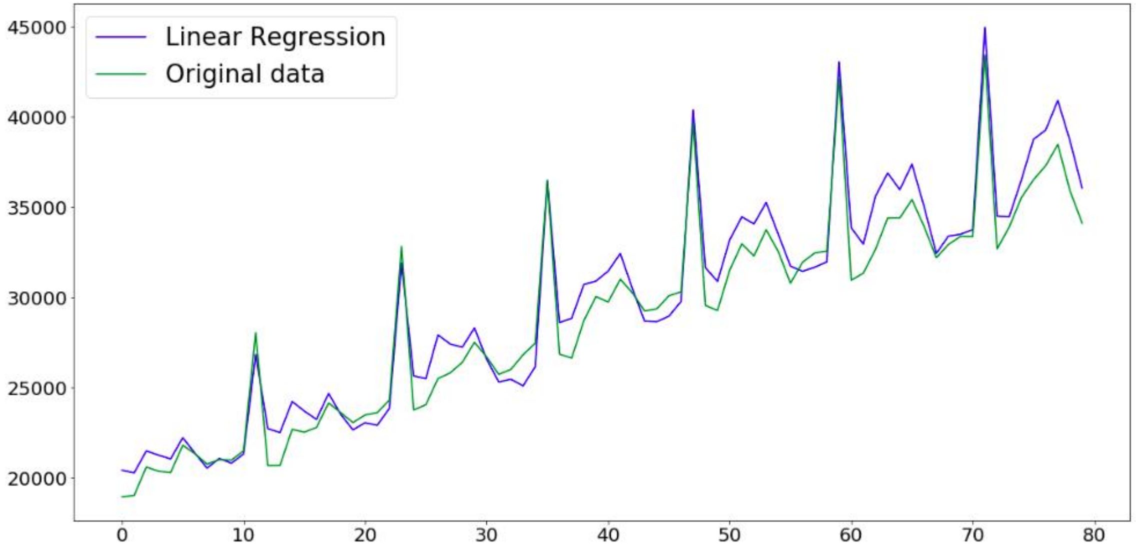
Пример. Сравним две модели:

- линейная регрессия
- скользящее среднее значение.

Скользящее среднее



Линейная регрессия



Метрики точности прогноза

Метрики оценки точности прогноза:

- R^2
- MSE (RMSE) – mean squared error – среднеквадратичная ошибка
- MAE – mean absolute error – средняя абсолютная ошибка
- MAPE – mean absolute percentage error – средняя абсолютная ошибка в %
- SMAPE – symmetric mean absolute percentage error – симметричная средняя абсолютная ошибка в %

Метрики точности прогноза: R^2

"R квадрат" или коэффициент детерминации – это доля дисперсии зависимой переменной, которую можно спрогнозировать на основе независимых переменных.

- Обычно используется для моделей линейной регрессии
- $0 \leq R^2 \leq 1$
- Чем выше значение, тем лучше.

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$RSS = \sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2$$

$$TSS = \sum_{t=1}^{n} (y_t - \bar{y})^2$$

$$\bar{y} = \frac{1}{n} \sum_{t=1}^{n} y_t$$

```python
from sklearn.metrics import r2_score

print("Linear Regression R^2:",  round(r2_score(y, y_pred_lr), 3))
print("SMA R^2:",  round(r2_score(y , y_sma), 3))
```

```
Linear Regression R^2: 0.942
SMA R^2: 0.822
```

Метрики точности прогноза: MSE

Среднеквадратичная ошибка (MSE) измеряет среднее значение квадратов ошибок, то есть среднеквадратичную разность между прогнозируемыми и фактическими значениями.

- Всегда неотрицательна.
- Значения ближе к нулю лучше.

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2$$

```python
from sklearn.metrics import mean_squared_error

print("Linear Regression MSE:", round(mean_squared_error(y, y_pred_lr), 3))
print("SMA MSE:",  round(mean_squared_error(y , y_sma), 3))
```

```
Linear Regression MSE: 1882343.713
SMA MSE: 5774211.042
```

Метрики точности прогноза: RMSE

Среднеквадратичная ошибка - это корень из среднего квадрата разности между прогнозируемыми и фактическими значениями.

- Всегда неотрицательна.
- Значения ближе к нулю лучше.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$$

```python
from sklearn.metrics import mean_squared_error

print("Linear Regression RMSE:", round(np.sqrt(mean_squared_error(y, y_pred_lr)), 3))
print("SMA RMSE:",  round(np.sqrt(mean_squared_error(y , y_sma)), 3))
```

```
Linear Regression RMSE: 1371.985
SMA RMSE: 2402.959
```

Метрики точности прогноза: MAE

Средняя абсолютная ошибка - это среднее расстояние по вертикали между каждой прогнозируемой точкой и фактической линией.

$$MAE = \frac{1}{n}\sum_{t=1}^{n} | y_t - \hat{y}_t |$$

```
from sklearn.metrics import mean_absolute_error

print("Linear Regression MAE:",  round(mean_absolute_error(y, y_pred_lr), 3))
print("SMA MAE:", round(mean_absolute_error(y , y_sma), 3))
```

```
Linear Regression MAE: 1148.816
SMA MAE: 1341.285
```

Метрики точности прогноза: MAPE

Средняя абсолютная процентная ошибка (MAPE) показывает среднюю долю ошибки относительно фактического значения. MAPE обычно выражает точность в процентах.

$$MAPE = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{y_t - \hat{y}_t}{y_t} \right|$$

Нельзя использовать, если есть нулевые значения, потому что будет деление на ноль. Для слишком низких прогнозов процентная ошибка не может превышать 100%, но для слишком высоких прогнозов нет верхнего предела процентной ошибки.

```
def mean_absolute_percentage_error(y_true, y_pred):
    return round(np.mean(np.abs((y_true - y_pred) / y_true)) * 100, 3)

print("Linear Regression MAPE:",  mean_absolute_percentage_error(y, y_pred_lr))
print("SMA MAPE:", mean_absolute_percentage_error(y , y_sma))

```

```
Linear Regression MAPE: 4.0
SMA MAPE: 22.493
```

Метрики точности прогноза: SMAPE

Симметричная средняя абсолютная ошибка в процентах - это показатель точности, основанный на процентах.

$$SMAPE = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|\hat{y}_t - y_t|}{\frac{1}{2} * (|y_t| + |\hat{y}_t|)}$$

Абсолютная разница между фактическим значением и прогнозируемым значением делится на половину суммы абсолютных значений фактического значения и прогнозируемого значения. Значение этого вычисления суммируется для каждой подобранной точки t и снова делится на количество подобранных точек n.

```
def smape(y_true, y_pred):
    return round(np.mean(np.abs((y_pred - y_true))/((np.abs(y_true)) + np.abs((y_pred)) / 2)), 3 )

print("Linear Regression SMAPE:",  smape(y, y_pred_lr))
print("SMA SMAPE:", smape(y , y_sma))
```

```
Linear Regression SMAPE: 0.026
SMA SMAPE: 0.147
```

Прогнозирование временных рядов

Методы экстраполяции - это методы, которые используют значения предыдущих периодов для прогнозирования будущих значений ряда.

Методы машинного обучения - это методы прогнозирования, которые используют данные прошлых значений ряда в качестве входных данных для модели обучения.

Прогнозирование с помощью SARIMA - это метод прогнозирования, который использует модель SARIMA (Seasonal AutoRegressive Integrated Moving Average).

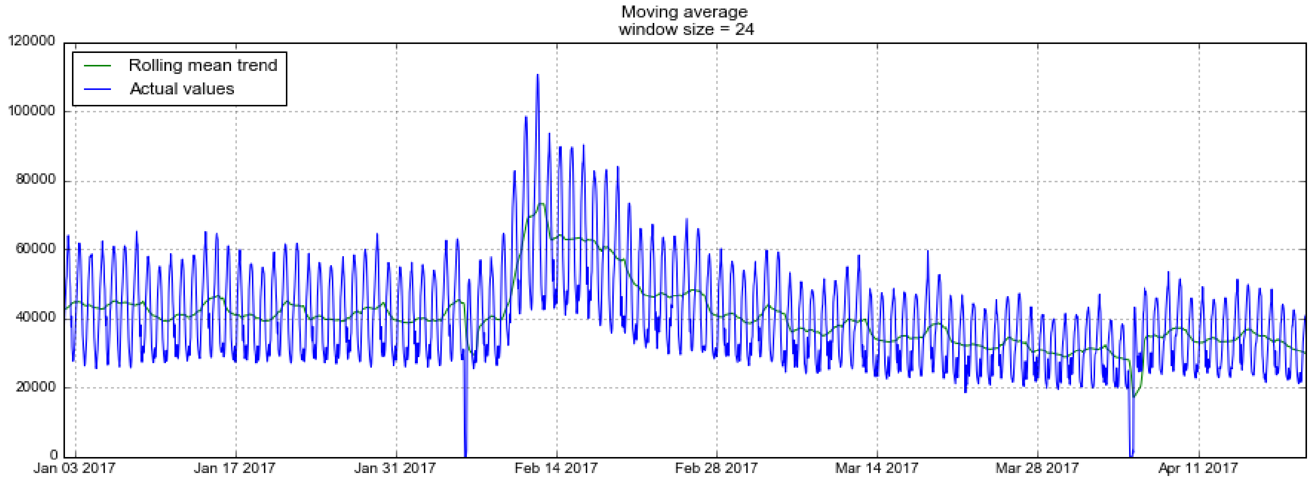
Метод скользящего среднего

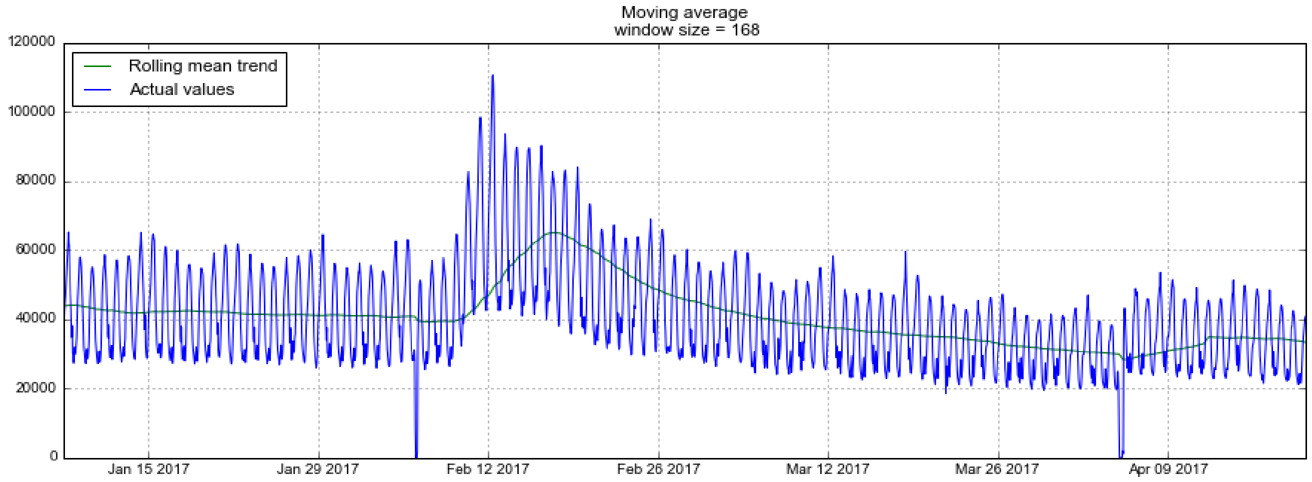
- Простая скользящая средняя

$$\hat{y}_t = \frac{1}{k} \sum_{n=0}^{k-1} y_{t-n}$$

- Взвешенная скользящая средняя

$$\hat{y}_t = \sum_{n=1}^{k} \omega_n y_{t+1-n}$$





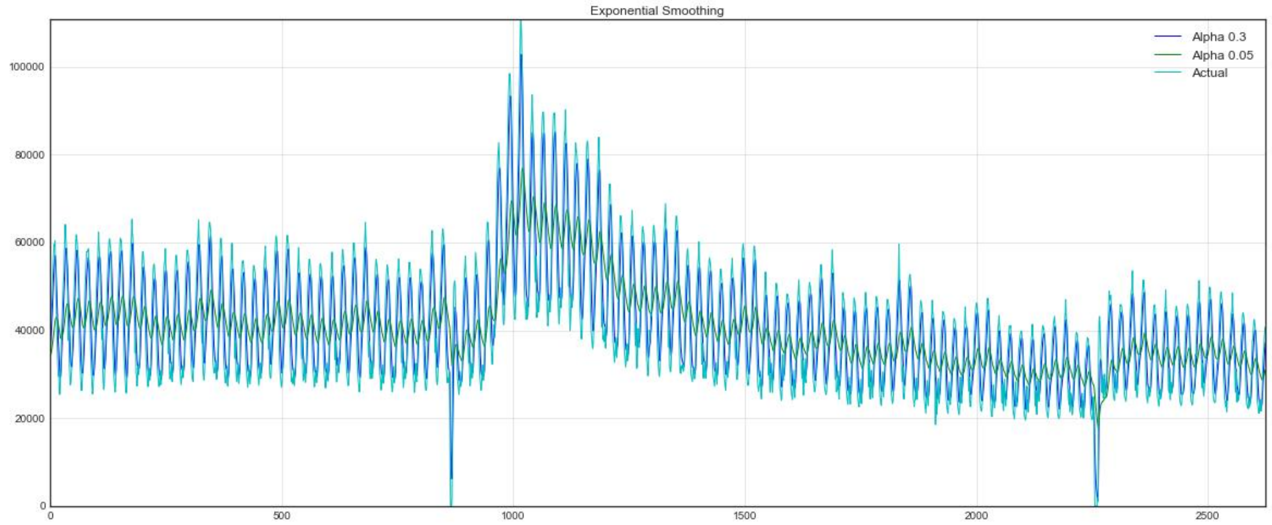
Экспоненциальное сглаживание, модель Хольта-Винтерса

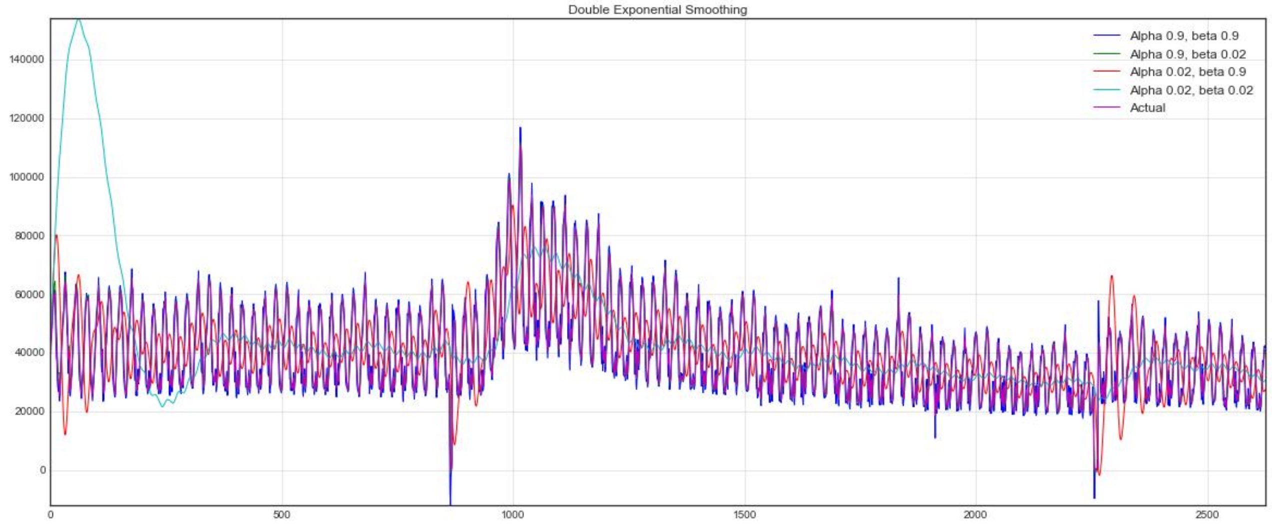
- Простое экспоненциальное сглаживание $\hat{y}_t = \alpha \cdot y_t + (1-\alpha) \cdot \hat{y}_{t-1}$
- Двойное экспоненциальное сглаживание

$$\begin{aligned}
\ell_x &= \alpha(y_x - s_{x-L}) + (1-\alpha)(\ell_{x-1} + b_{x-1}) \\
b_x &= \beta(\ell_x - \ell_{x-1}) + (1-\beta)b_{x-1} \\
s_x &= \gamma(y_x - \ell_x) + (1-\gamma)s_{x-L} \\
\hat{y}_{x+m} &= \ell_x + mb_x + s_{x-L+1+(m-1)modL}
\end{aligned}$$

- Тройное экспоненциальное сглаживание a.k.a. Holt-Winters

$$\begin{aligned}
\hat{y}_{max_x} &= \ell_{x-1} + b_{x-1} + s_{x-T} + m \cdot d_{t-T} \\
\hat{y}_{min_x} &= \ell_{x-1} + b_{x-1} + s_{x-T} - m \cdot d_{t-T} \\
d_t &= \gamma \mid y_t - \hat{y}_t \mid + (1-\gamma)d_{t-T},
\end{aligned}$$





Прогнозирование как задача машинного обучения

Прогнозирование на один шаг вперед. Задача обучения с учителем.

Необходимые данные:

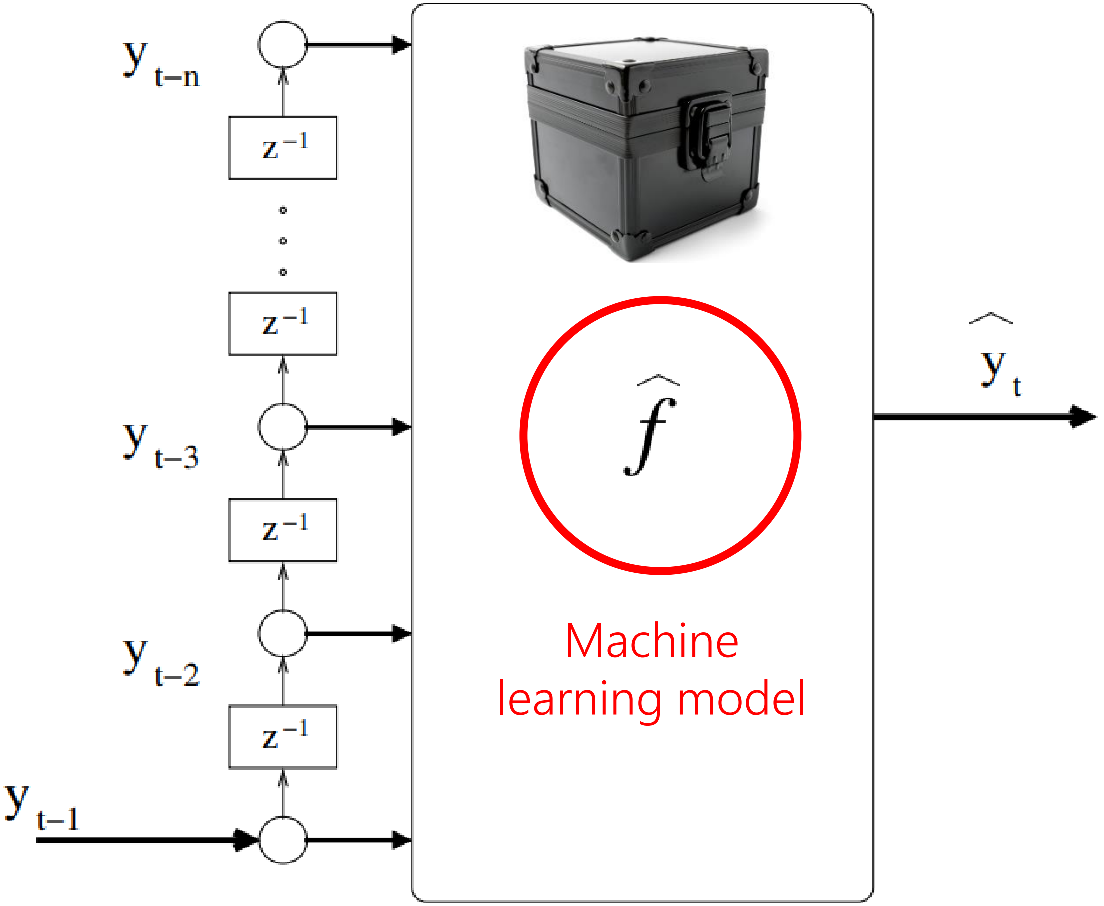
- обучающий набор (входы)
- метки (выходам)
- и тестовый набор

Временной ряд: **S: [$y_0, y_1, \ldots, y_{t-2}, y_{t-1}$]**

Предсказываем **$\langle y_t \rangle$**

$$X = \begin{bmatrix} y_{N-1} & y_{N-2} & \cdots & y_{N-n-1} \\ y_{N-2} & y_{N-3} & \cdots & y_{N-n-2} \\ \vdots & \vdots & \vdots & \vdots \\ y_n & y_{n-1} & \cdots & y_1 \end{bmatrix} \qquad Y = \begin{bmatrix} y_N \\ y_{N-1} \\ \vdots \\ y_{n+1} \end{bmatrix}$$

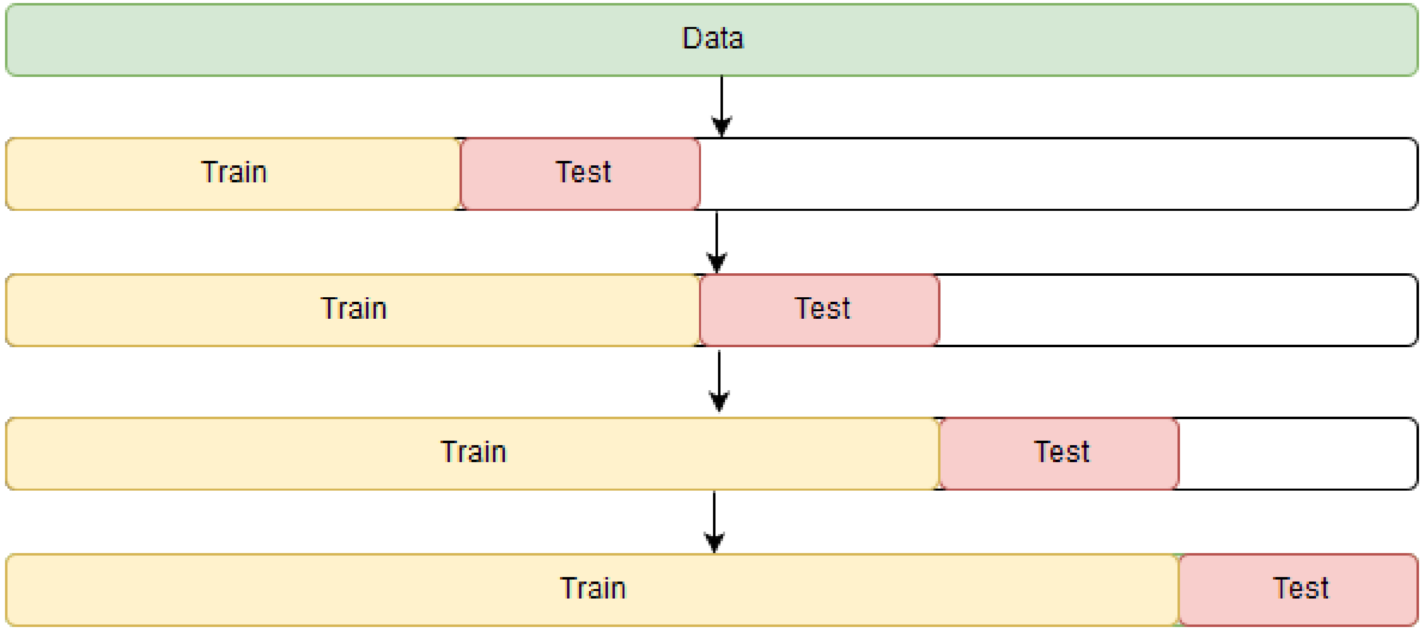
входы выходы



Кросс-валидация на временных рядах

Временной ряд имеет временную структуру, поэтому случайно перемешивать в фолдах значения всего ряда без сохранения этой структуры нельзя, так как в процессе потеряются все взаимосвязи наблюдений.

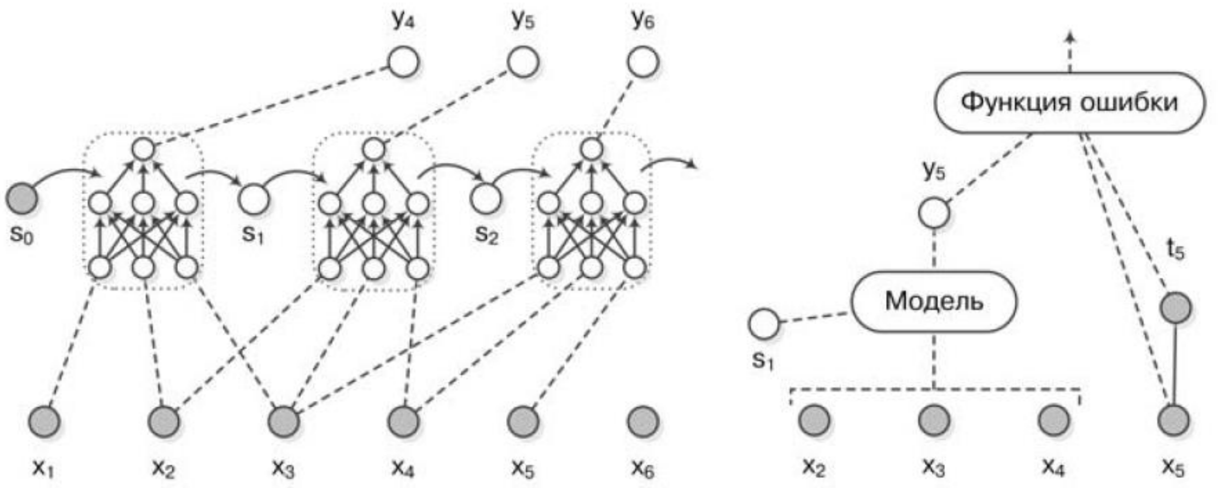
Для этого модель обучается и тестируется на последовательных интервалах данных



Нейронные сети в задачах предсказания временных рядов

- Модель должна «помнить» элементы последовательности с целью использовать их в дальнейшем;
- Необходимо фиксировать зависимости с большим временным окном.

Рекуррентные нейронные сети - это вид нейронных сетей, где связи между элементами образуют направленную последовательность. Благодаря этому появляется возможность обрабатывать серии событий во времени или последовательные пространственные цепочки.



Вопросы для самоконтроля

1) Для оценки точности прогноза с нулевыми значениями в фактических данных нельзя использовать:

    1. R^2
    2. MAPE
    3. MSE

2) Что не относится к операциям, которые используются для преобразования временного ряда к стационарному:

    1. Дифференцирование
    2. Масштабирование
    3. Преобразование Бокса-Кокса

Информационные источники

- Афанасьев В.Н., Юзбашев М.М. Анализ временных рядов и прогнозирование - М.: Финансы и статистика, 2001. — 228 с.:
- Портал https://machinelearningmastery.com/
- Статья: https://habr.com/ru/company/ods/blog/327242/
- Статья: https://habr.com/ru/companies/otus/articles/732080/
- Образовательный ресурс: https://blog.skillfactory.ru/analiz-vremennyh-ryadov-polnoe-rukovodstvo/
- Образовательный ресурс: https://ranalytics.github.io/tsa-with-r/ch-visualisation.html
- Образовательный ресурс: https://studfile.net/preview/2967731/page:5/
- Образовательный ресурс: https://studfile.net/preview/6322082/
- Образовательный ресурс: https://www.dmitrymakarov.ru/intro/time-series/
- Первичный анализ данных с Pandas
- Визуальный анализ данных с Python
- Классификация, деревья решений и метод ближайших соседей
- Линейные модели классификации и регрессии
- Композиции: бэггинг, случайный лес. Кривые валидации и обучения
- Построение и отбор признаков
- Обучение без учителя: PCA, кластеризация
- Обучаемся на гигабайтах с Vowpal Wabbit
- Анализ временных рядов с помощью Python
- Градиентный бустинг

Конец лекции