Южно-Уральский государственный университет | Национальный исследовательский университет

Модель ARIMA

Авторегрессия AR

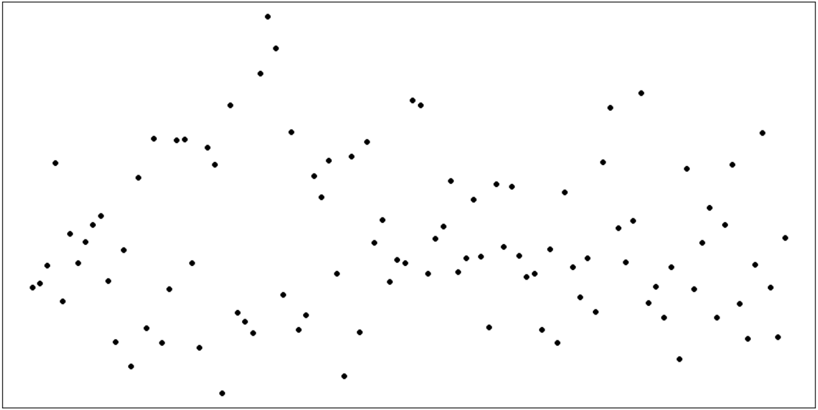
Модель авторегрессии порядка p:

$$AR(p): y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t,$$

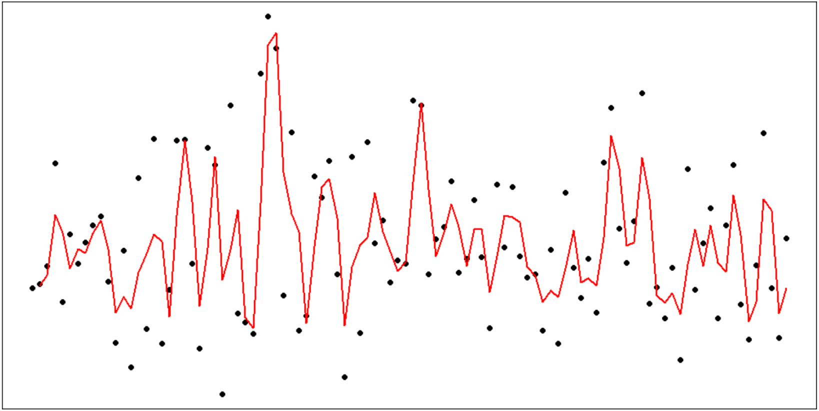
y_t — линейная комбинация p предыдущих значений ряда и шумовой компоненты $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$.

Скользящее среднее

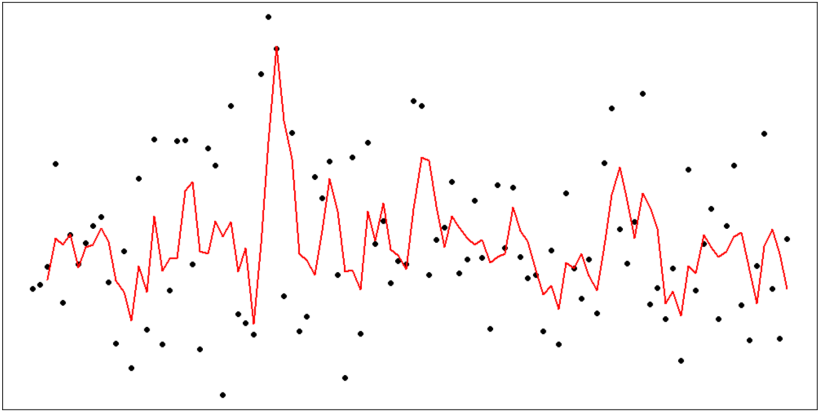
Пусть у нас есть независимый одинаково распределённый во времени шум ε_t :



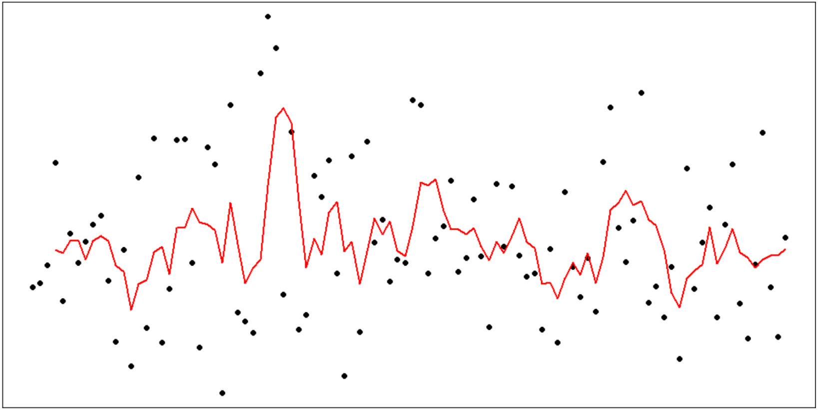
Скользящее среднее по двум соседним точкам



Скользящее среднее по трем соседним точкам



Скользящее среднее по четырем соседним точкам



Скользящее среднее MA

Модель скользящего среднего порядка q:

$$MA(q) : y_t = \alpha + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q},$$

y_t - линейная комбинация q последних значений шумовой компоненты.

ARMA (Autoregressive moving average)

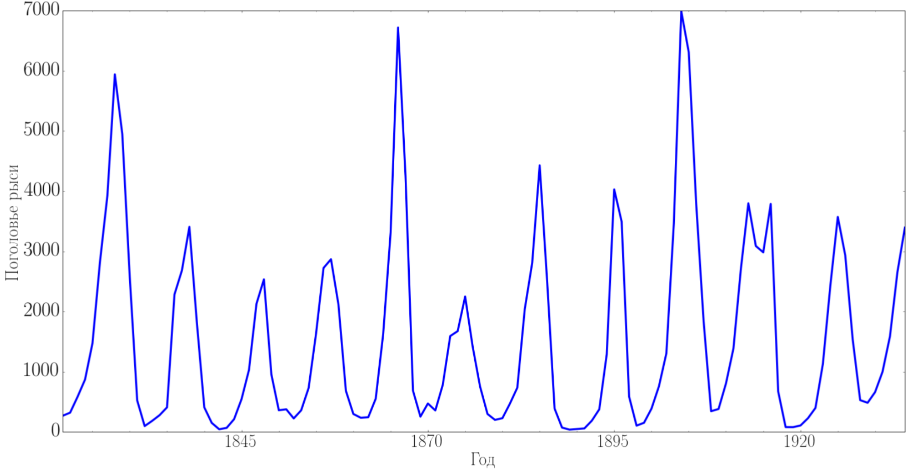
Модель $ARMA(p, q)$:

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

Теорема Вольда:

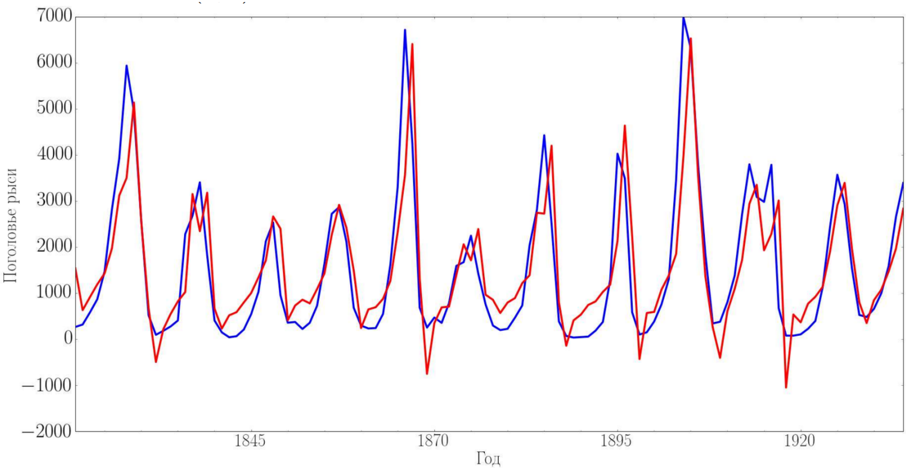
любой стационарный ряд может быть описан моделью $ARMA(p, q)$ с любой точностью.

Пример ARMA



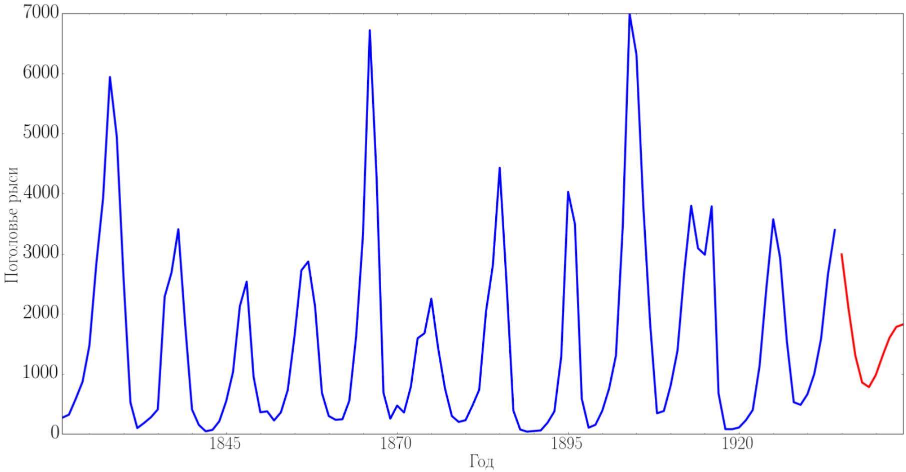
7000
6000
5000
4000
3000
2000
1000
0
Поголовье рыси
1845
1870
1895
1920
Год

ARMA(2,2)



7000
6000
5000
4000
3000
2000
1000
0
−1000
−2000
Поголовье рыси
1845
1870
1895
1920
Год

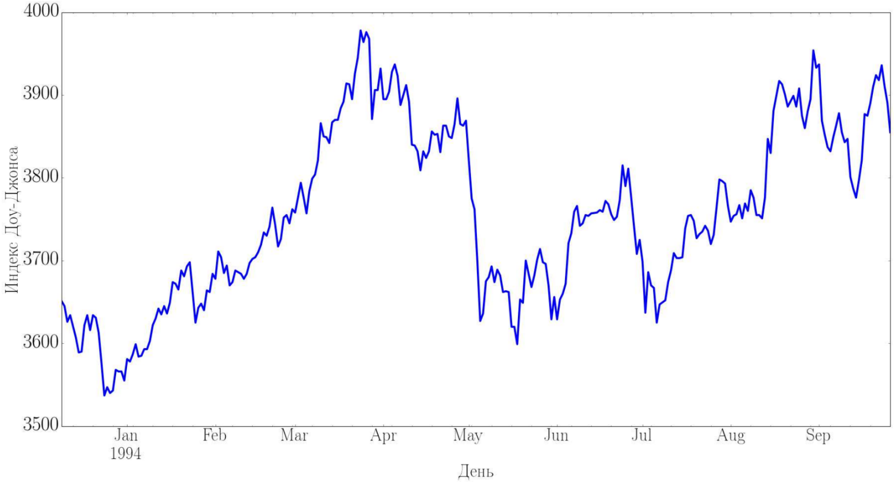
ARMA для прогнозирования



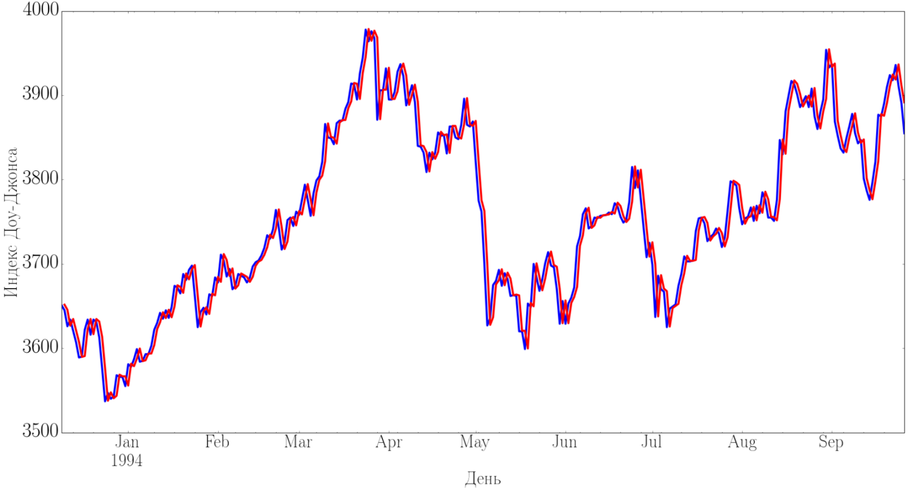
ARIMA (Autoregressive integrated moving average)

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

Пример ARIMA



ARIMA(0,1,0)



4000
3900
3800
3700
3600
3500
Индекс Доу-Джонса
Jan
1994
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
День

SARMA

Пусть ряд имеет сезонный период длины S.
Возьмём модель $ARMA(p, q)$:

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

и добавим P авторегрессионных компонент:

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

и Q компонент скользящего среднего:

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

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

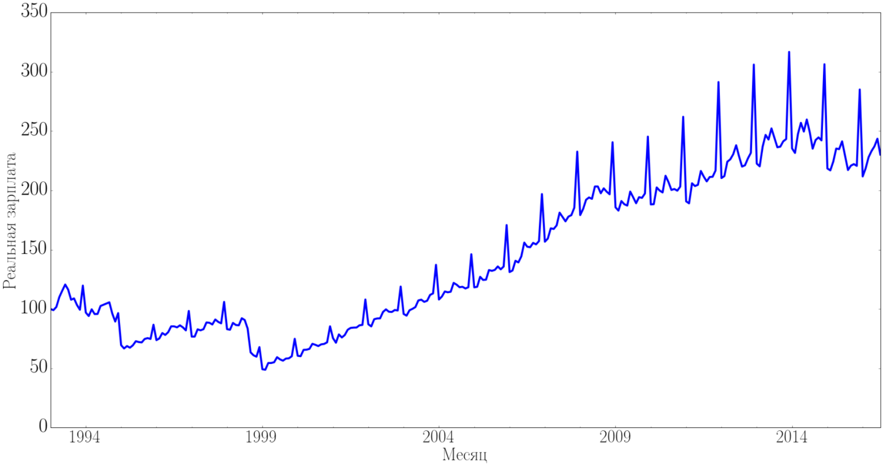
SARIMA

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.

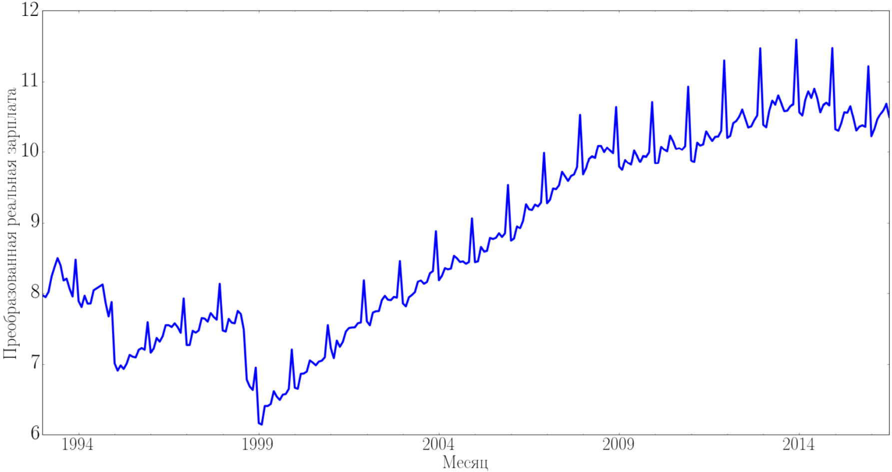
Пример



Критерий Дики-Фуллера: $p = 0.2265$.

Пример

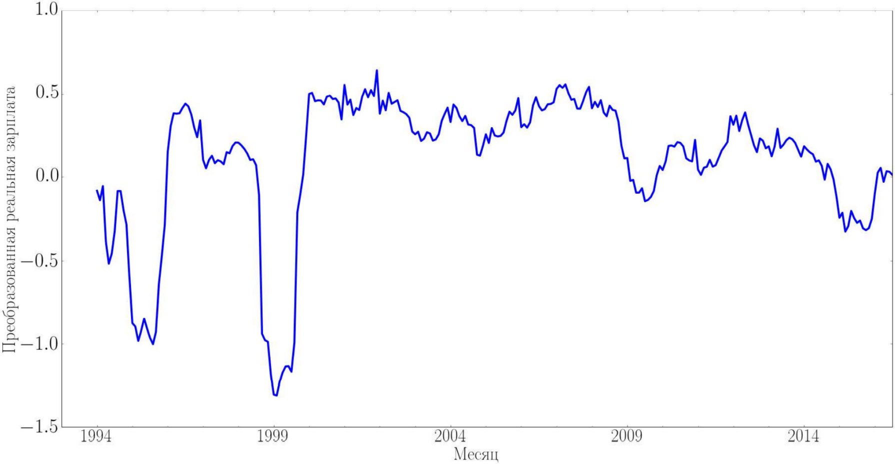
После преобразования Бокса-Кокса с $\lambda = 0.22$:



Критерий Дики-Фуллера: $p = 0.1661$.

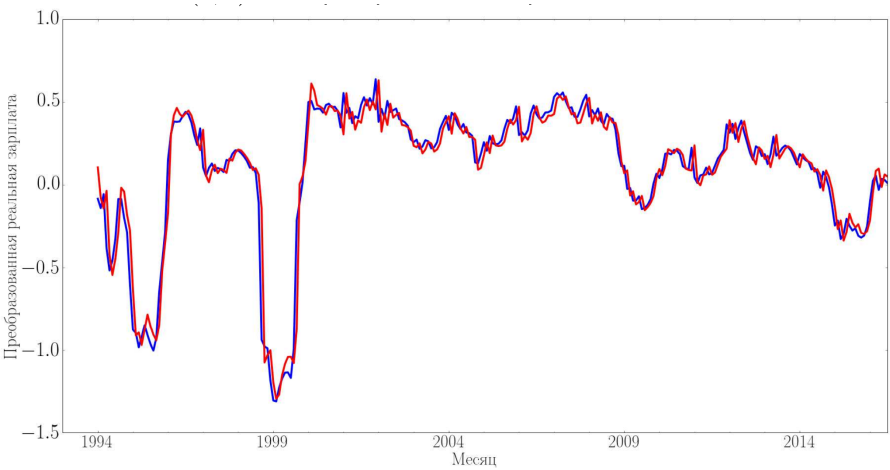
Пример

После сезонного дифференцирования:



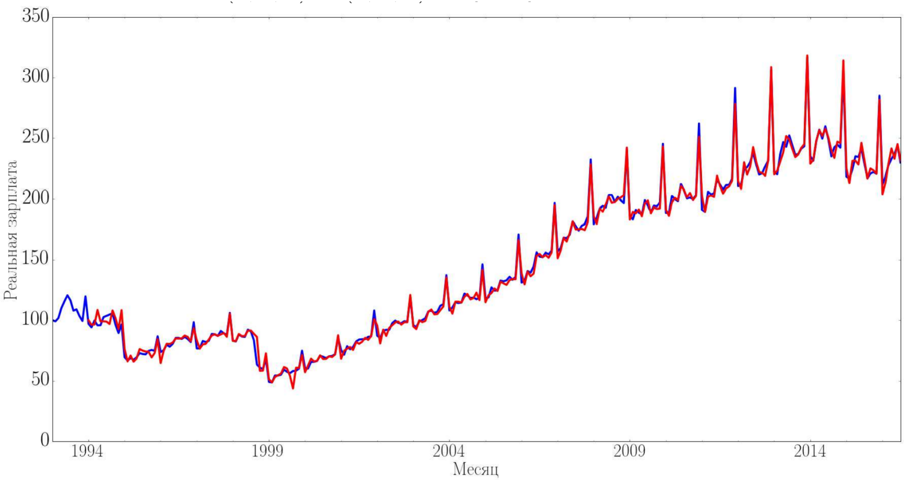
Критерий Дики-Фуллера: $p = 0.01$.

Модель $ARMA(2,2)$



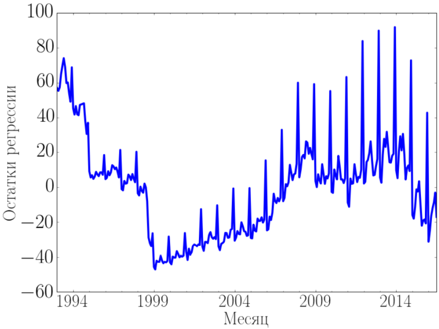
Пример

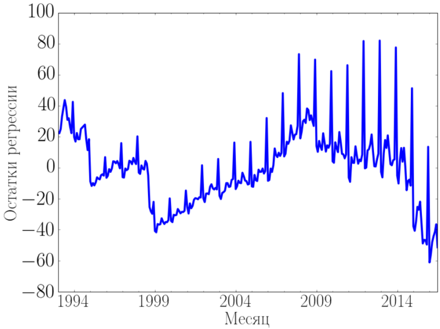
Модель $SARIMA(2,0,2) \times (0,1,0)$ с преобразованием Бокса-Кокса:



Пример

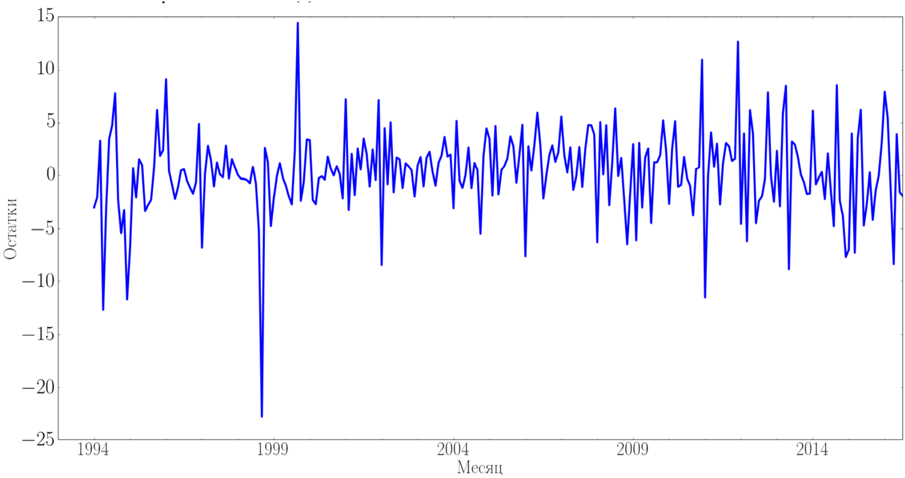
Остатки регрессий на время:





Пример

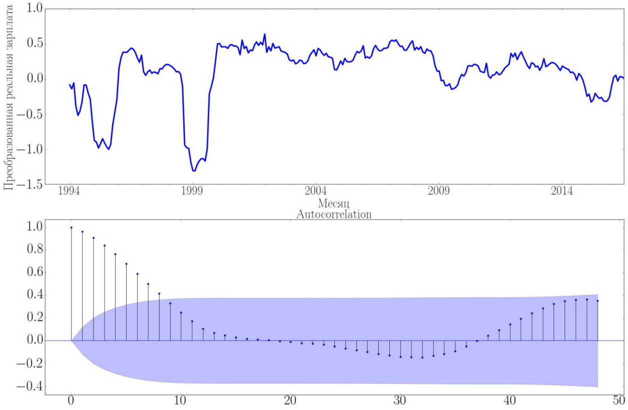
Остатки модели *SARIMA*:



Подбор параметров модели d, D

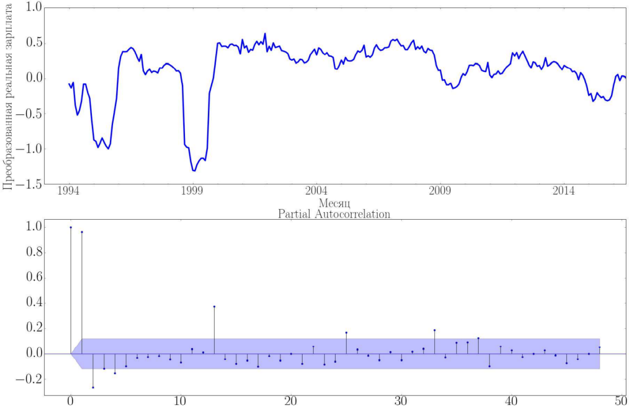
- Порядки дифференцирования d, D подбираются так, чтобы ряд стал стационарным.
- Если ряд сезонный, рекомендуется начинать с сезонного дифференцирования.
- Чем меньше раз мы продифференцируем, тем меньше будет дисперсия итогового прогноза.

Подбор параметров модели q, Q



$Q * S$ — номер последнего сезонного лага, при котором автокорреляция значима (здесь 0).
q — номер последнего несезонного лага, меньшего величин периода, при котором автокорреляция значима (здесь 8).

Подбор параметров модели p, P

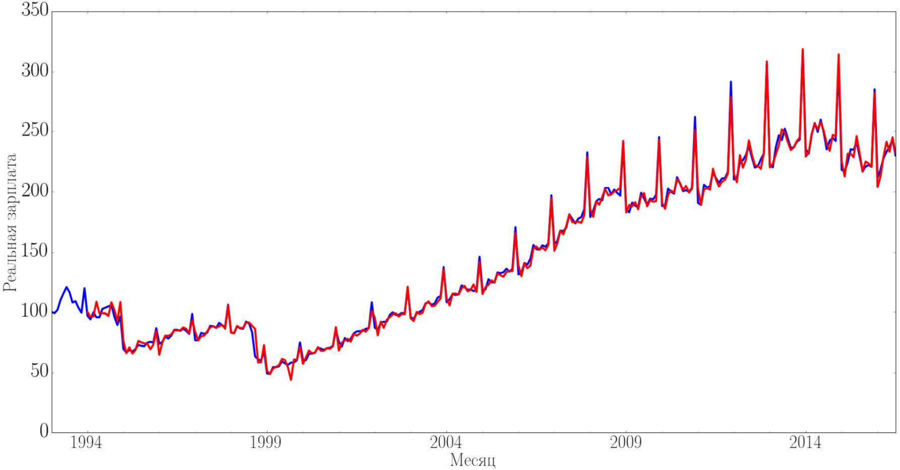


$P * S$ — номер последнего сезонного лага, при котором частичная автокорреляция значима (здесь 0).

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

Пример

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

1. Смотрим на ряд. При необходимости стабилизируем дисперсию.
2. Если ряд нестационарен, подбираем порядок дифференцирования.
3. Анализируем ACF/PACF, определяем примерные p, q, P, Q.
4. Обучаем модели-кандидаты, сравниваем их по AIC (информационному критерия Акаике), выбираем победителя.
5. Смотрим на остатки $y_t - \hat{y}_t$ полученной модели, если они плохие, пробуем что-то поменять.

- несмещенные
- стационарные