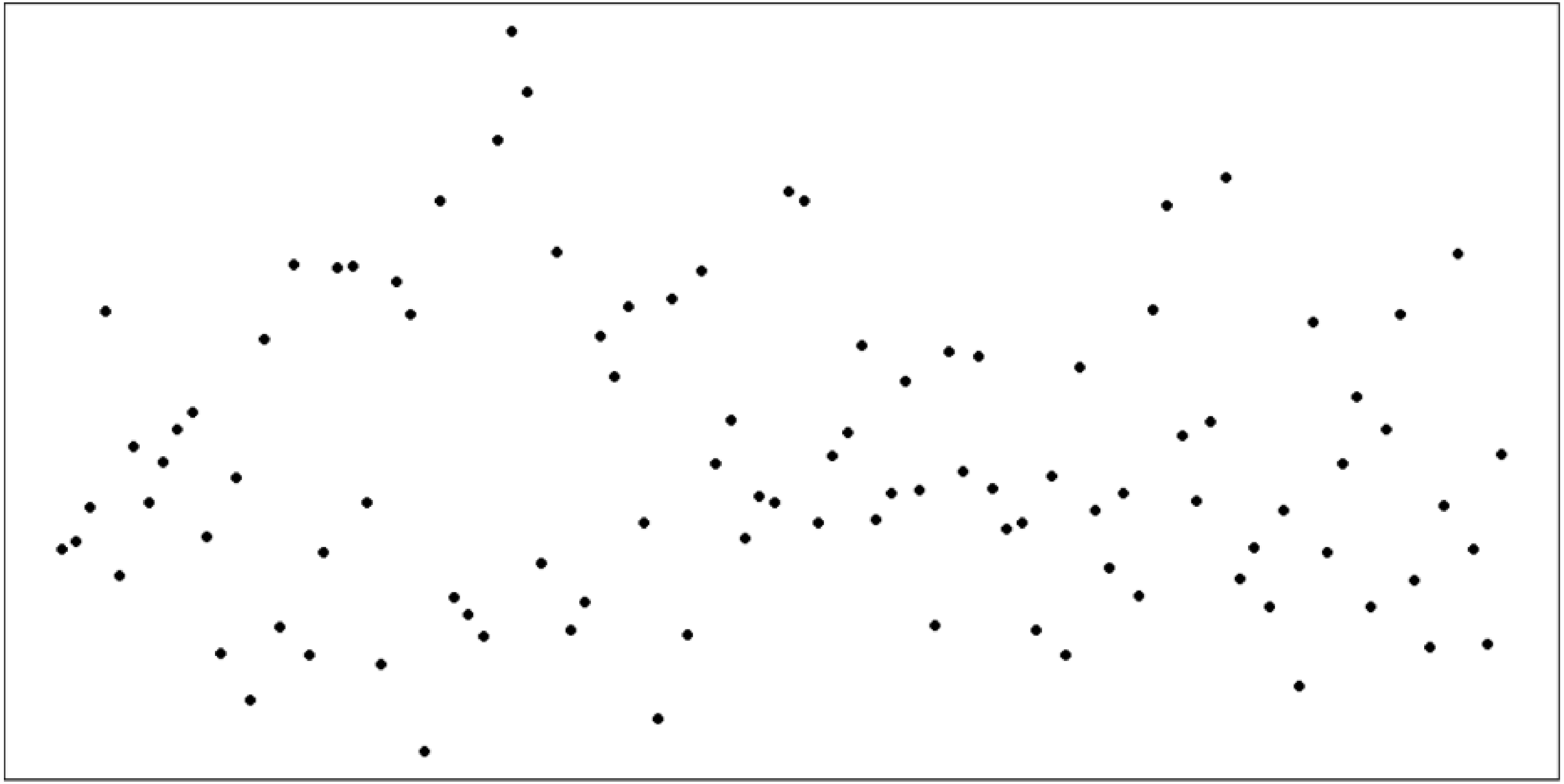
Что если делать регрессию ряда на собственные значения в прошлом?

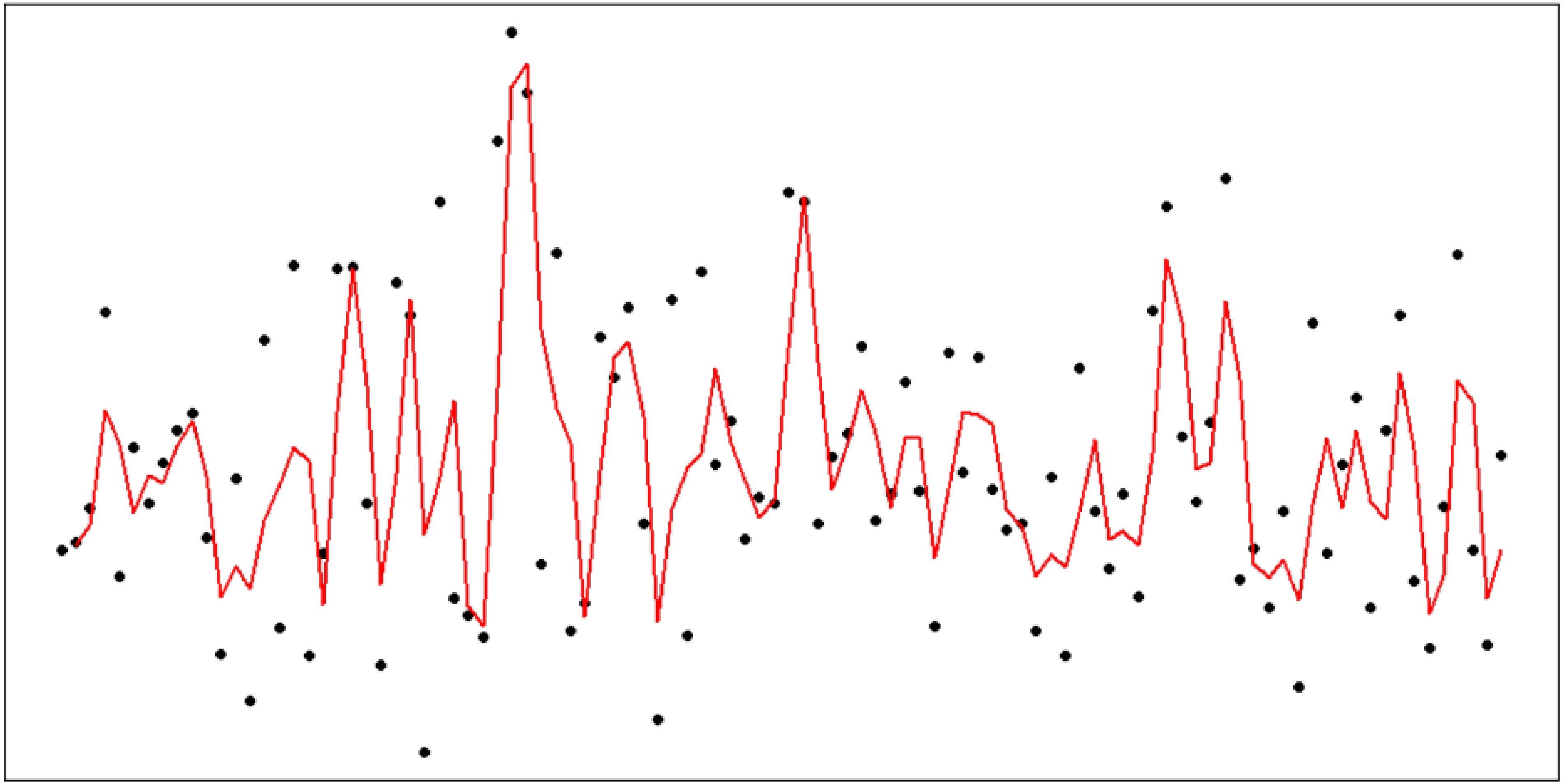
$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

Модель авторегрессии порядка p ($AR(p)$):
y_t — линейная комбинация p предыдущих значений ряда и шумовой компоненты.

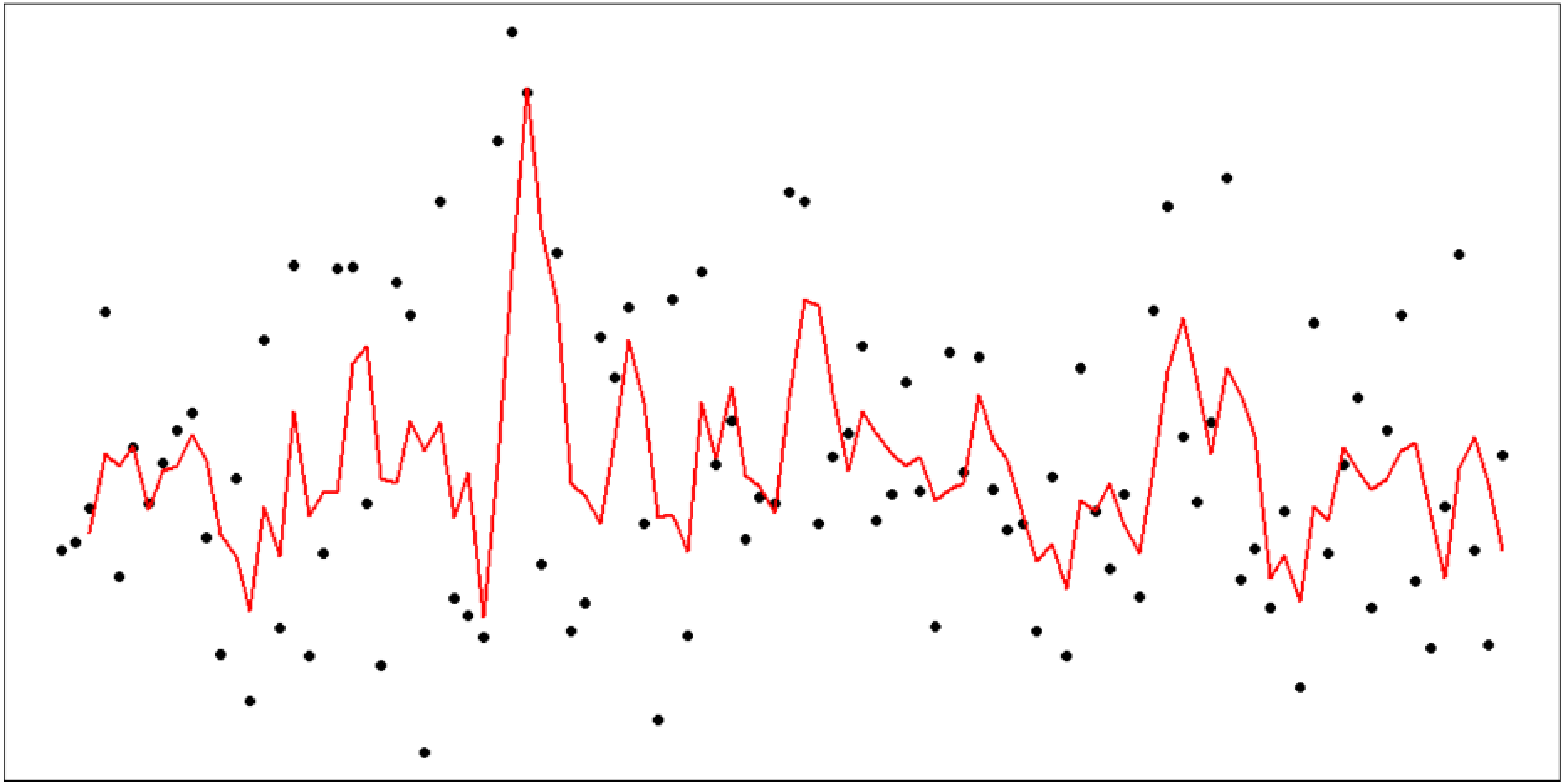
Пусть у нас есть независимый одинаково распределённый во времени шум ε_t:



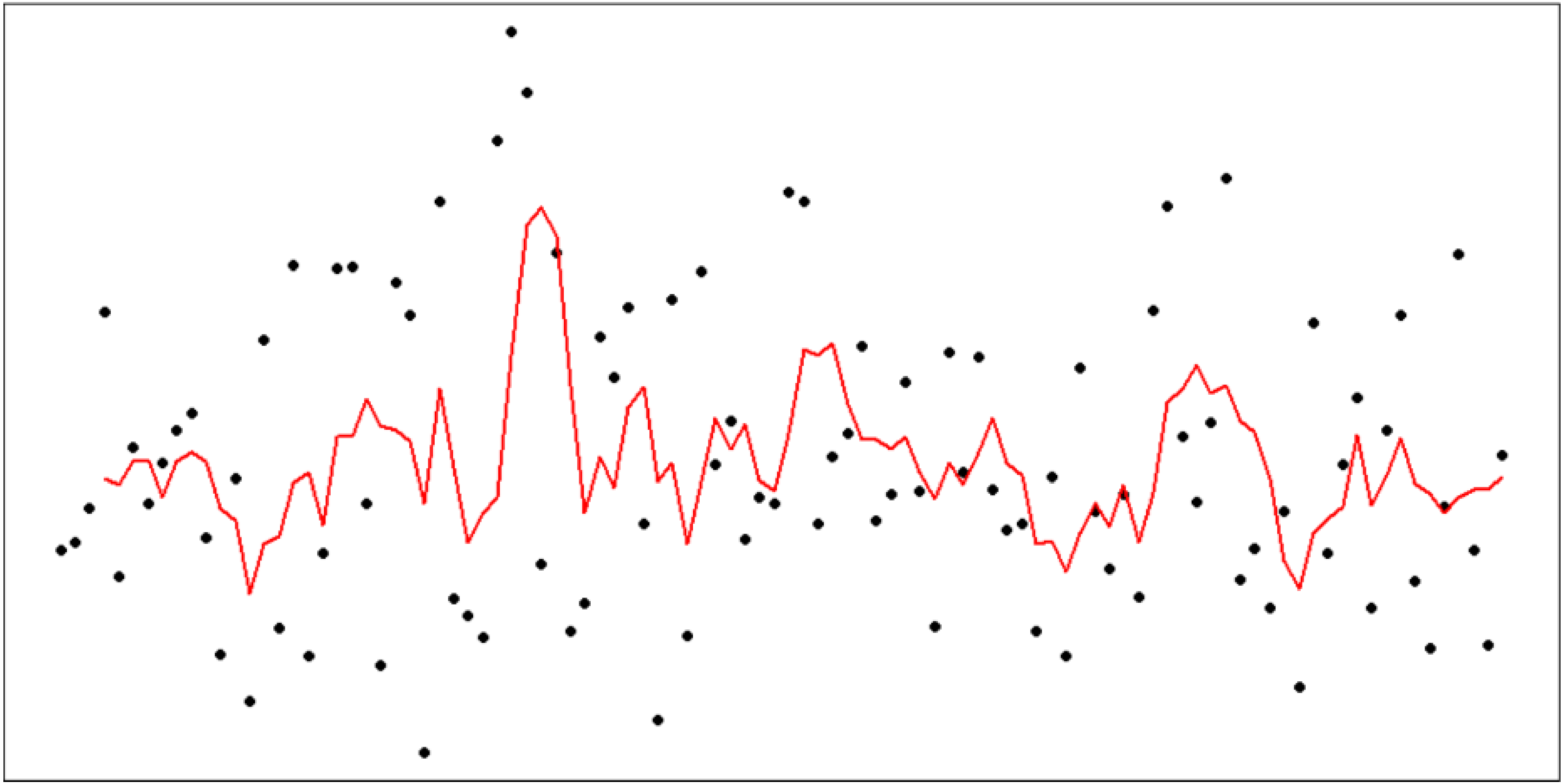
Среднее по двум соседним точкам:



Среднее по трём соседним точкам:



Среднее по четырём соседним точкам:



Обобщим и добавим веса:

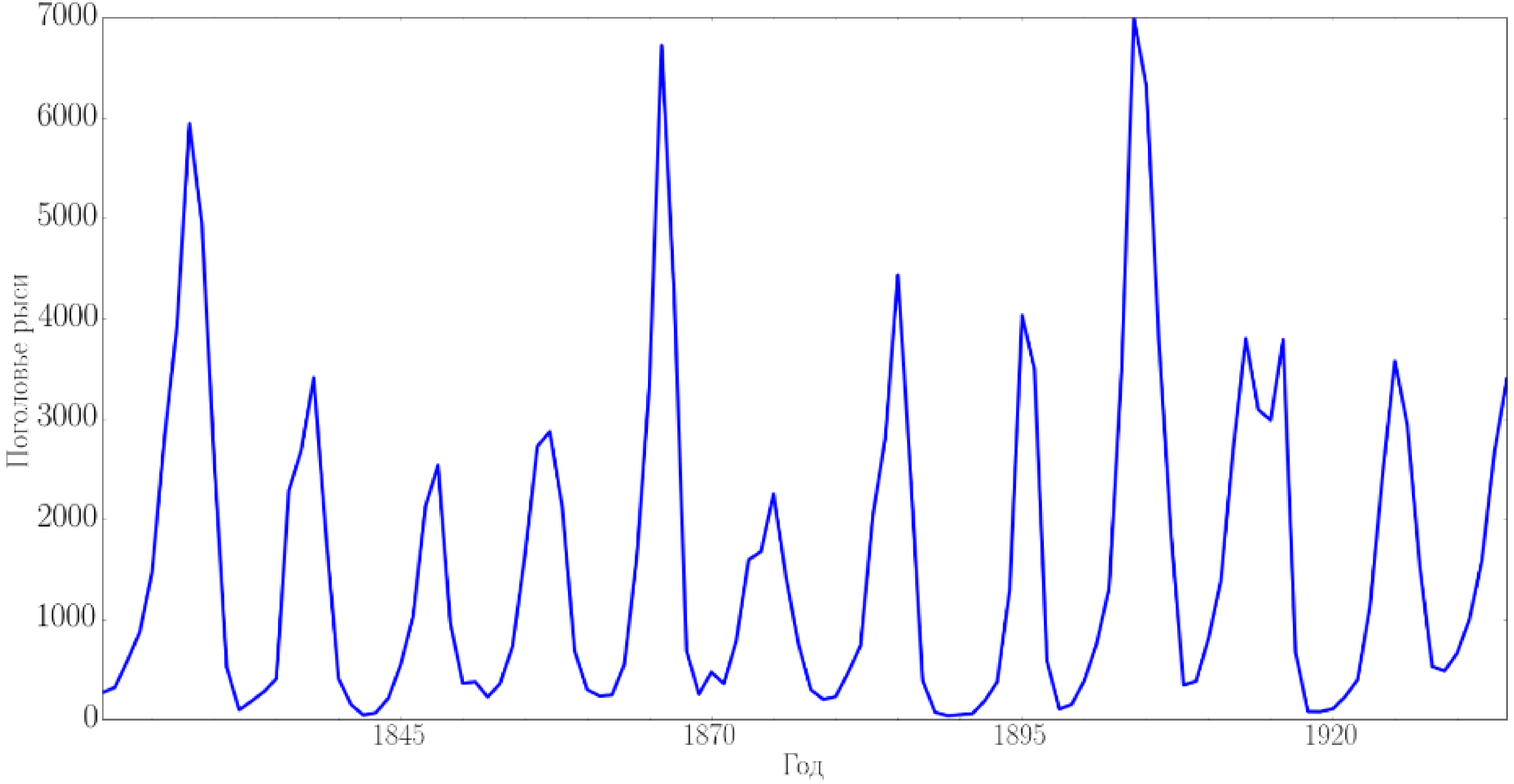
$$y_t = \alpha + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

Модель скользящего среднего порядка q ($MA(q)$):
y_t — линейная комбинация q последних значений шумовой компоненты.

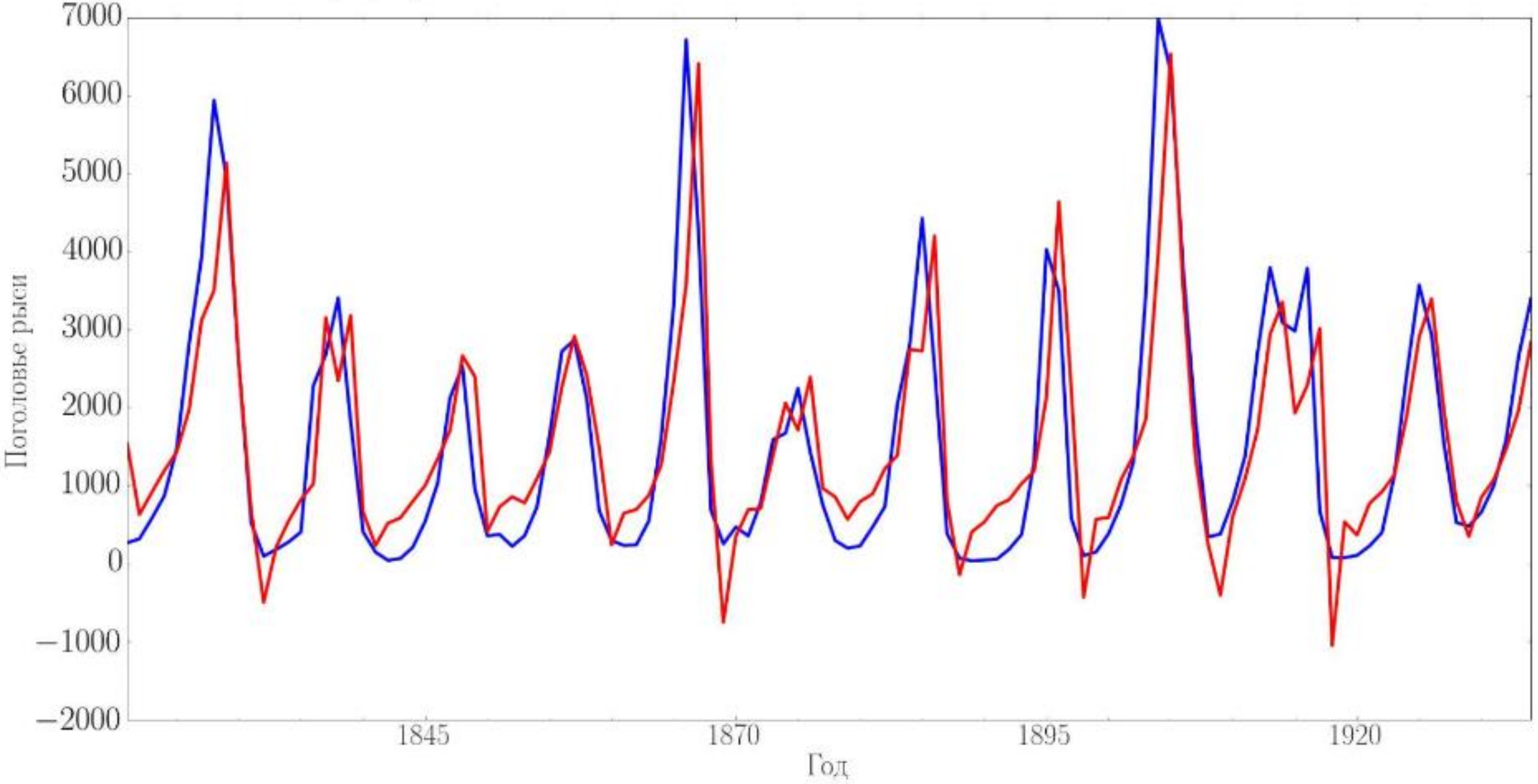
Модель $ARMA(p,q)$:

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

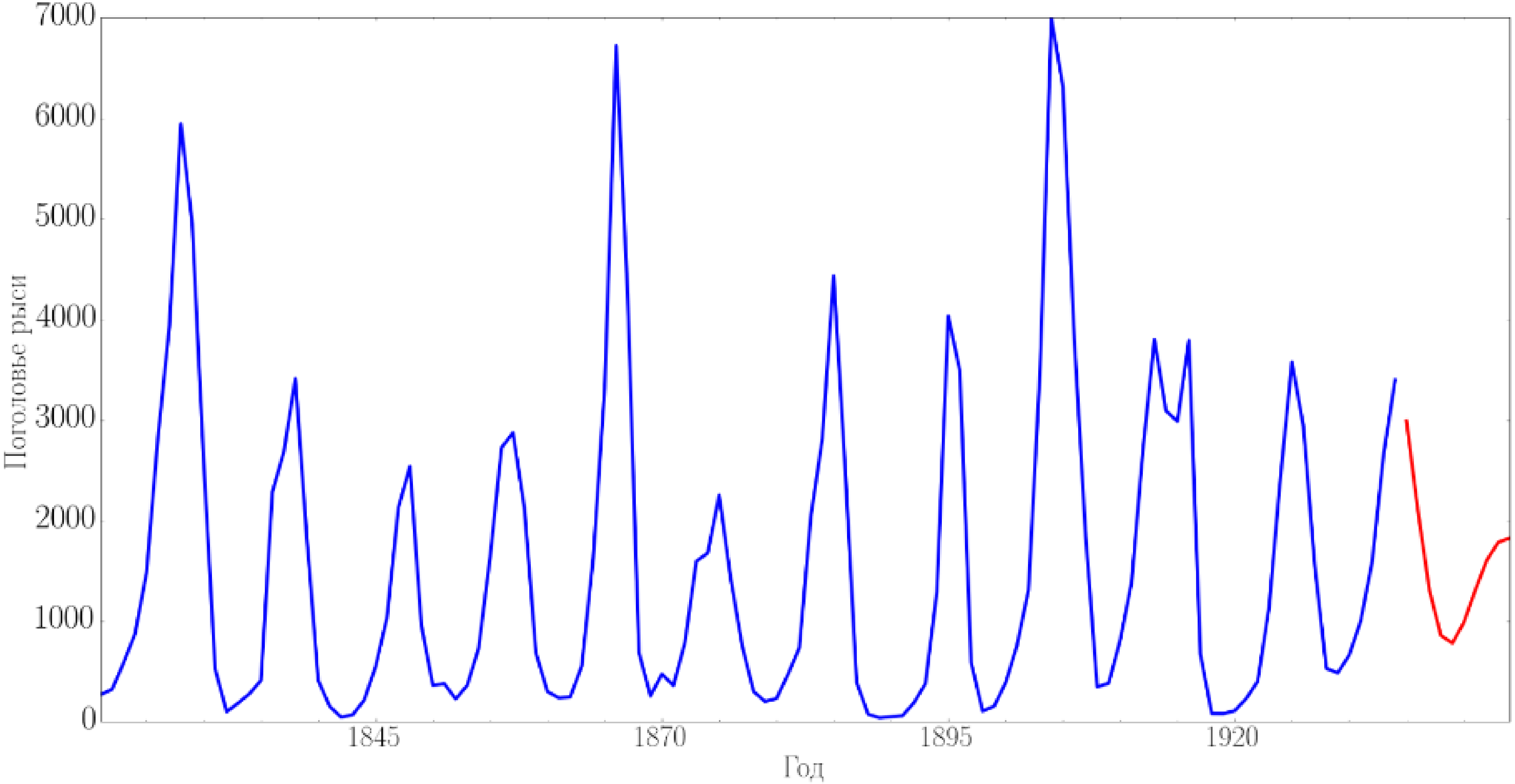
Теорема Вольда: любой стационарный ряд может быть описан моделью $ARMA(p,q)$.


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

Модель $ARMA(2, 2)$:

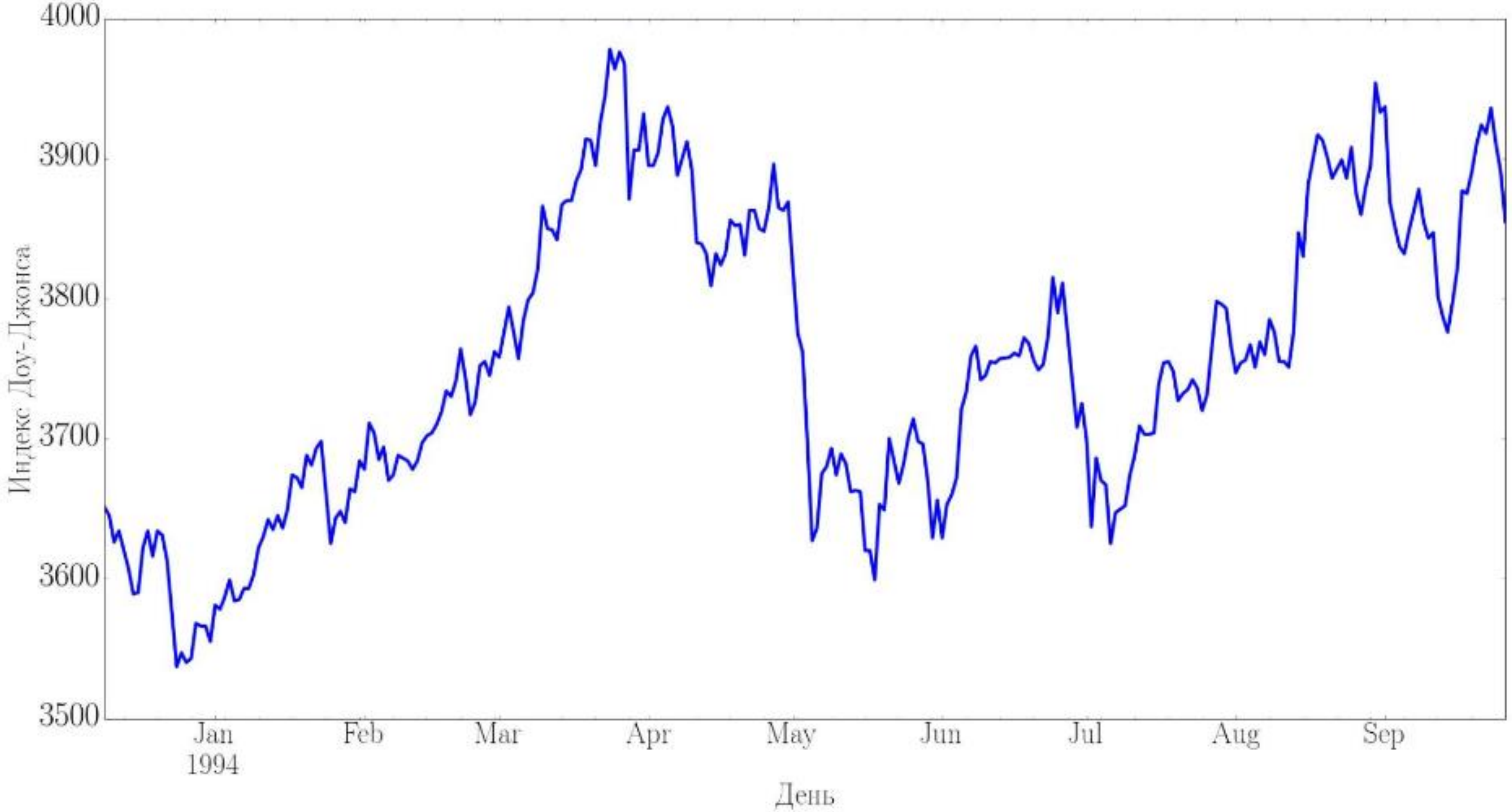


Модель $ARMA(2,2)$:

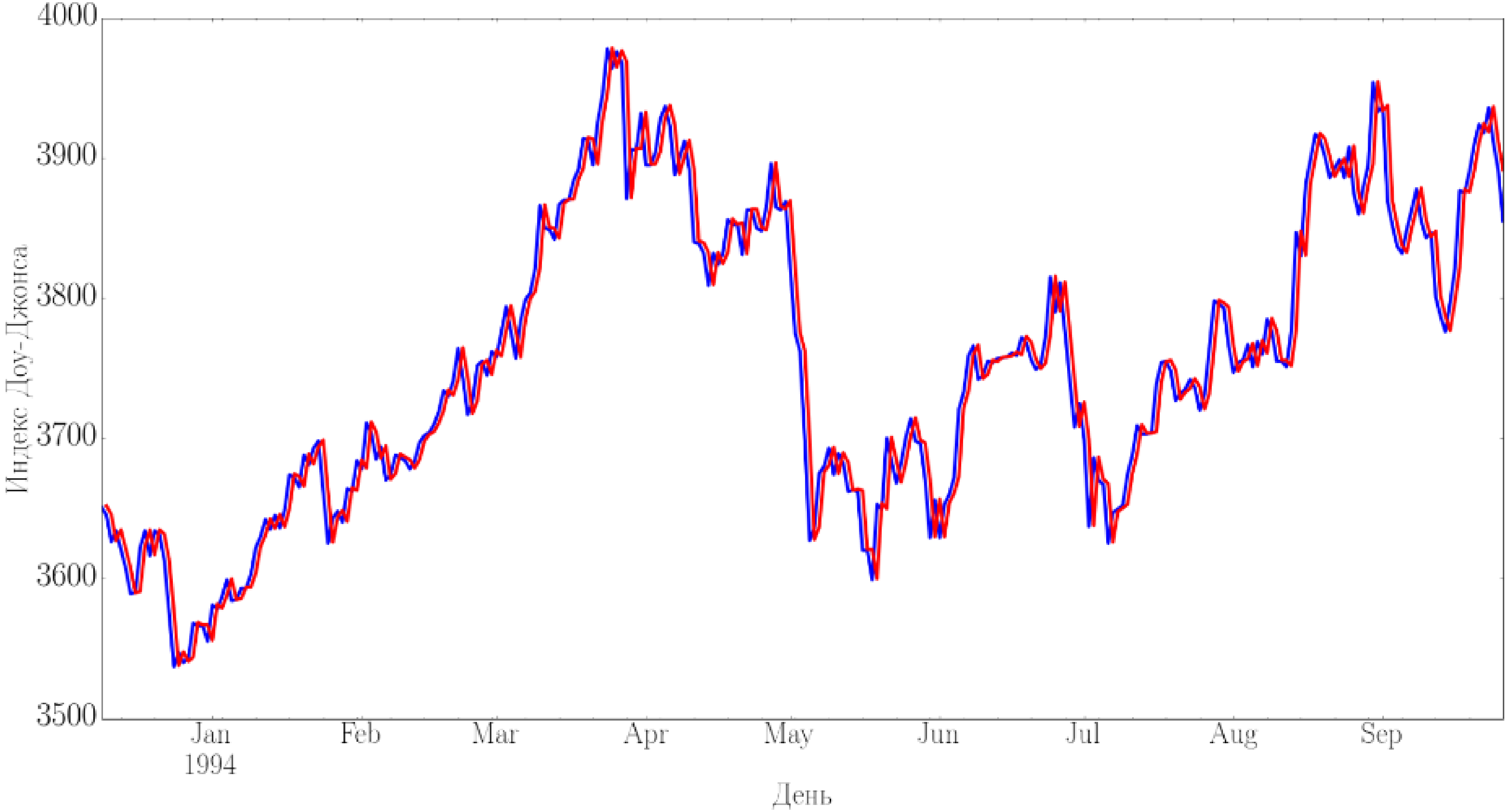


Модель $ARIMA(p, d, q)$ — модель $ARMA(p, q)$ для d раз продифференцированного ряда.

Индекс Доу-Джонса



Модель $ARIMA(0,1,0)$:



SARMA

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

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

и добавим P авторегрессионных компонент:

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

и Q компонент скользящего среднего:

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

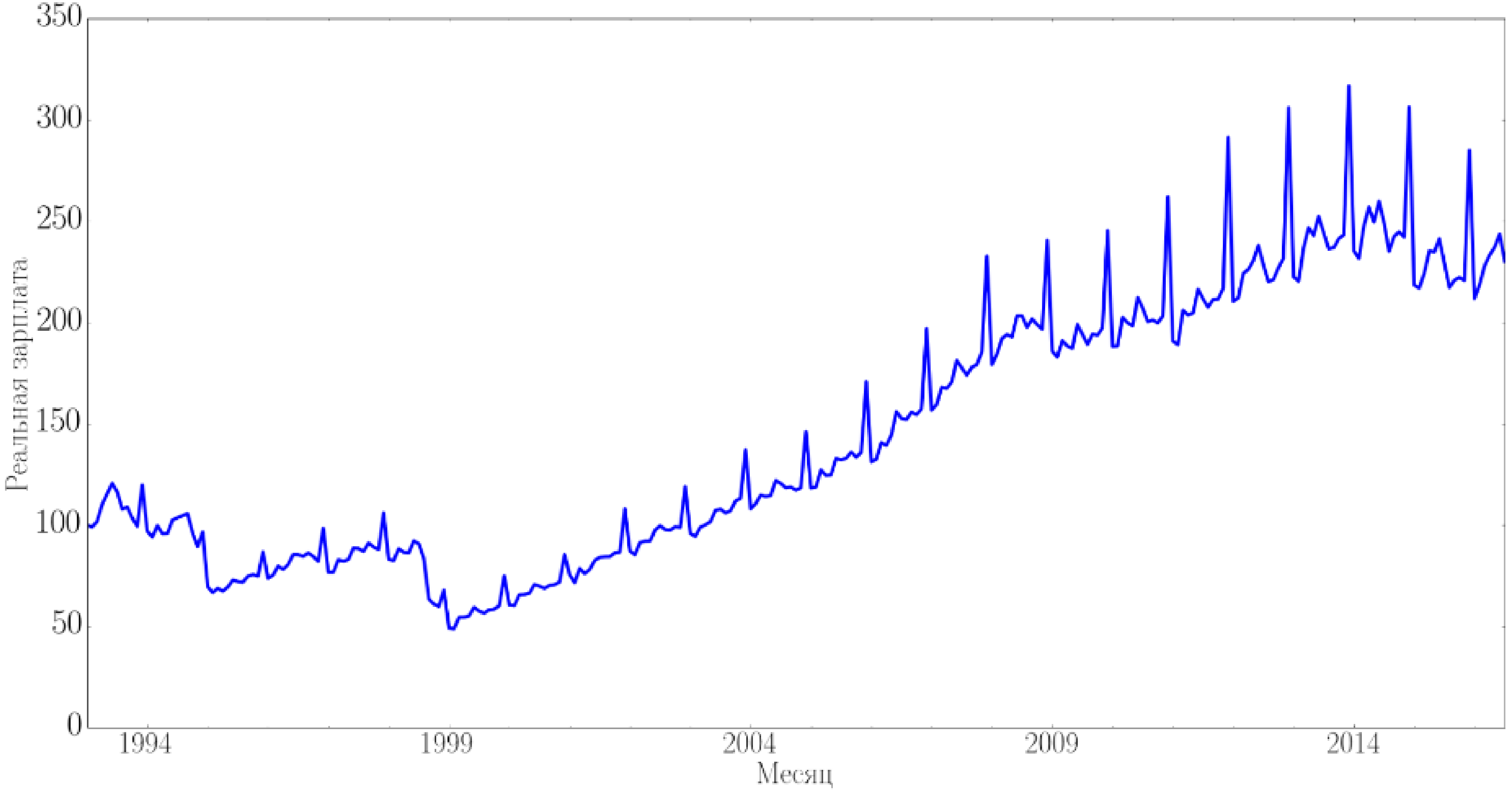
Это модель $SARMA(p,q) \times (P,Q)$

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.

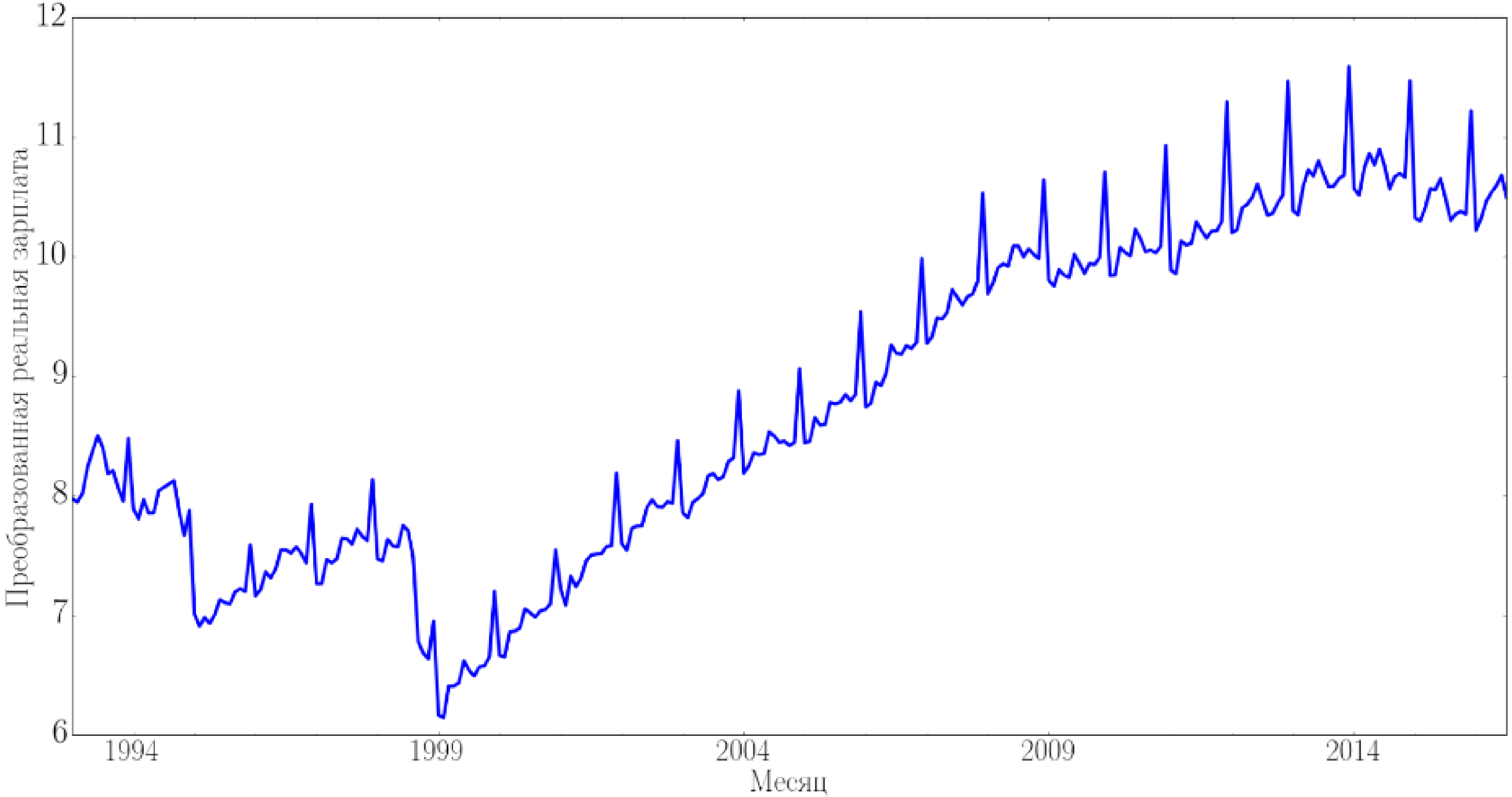
Реальная заработная плата



Критерий Дики-Фуллера: $p = 0.2265$.

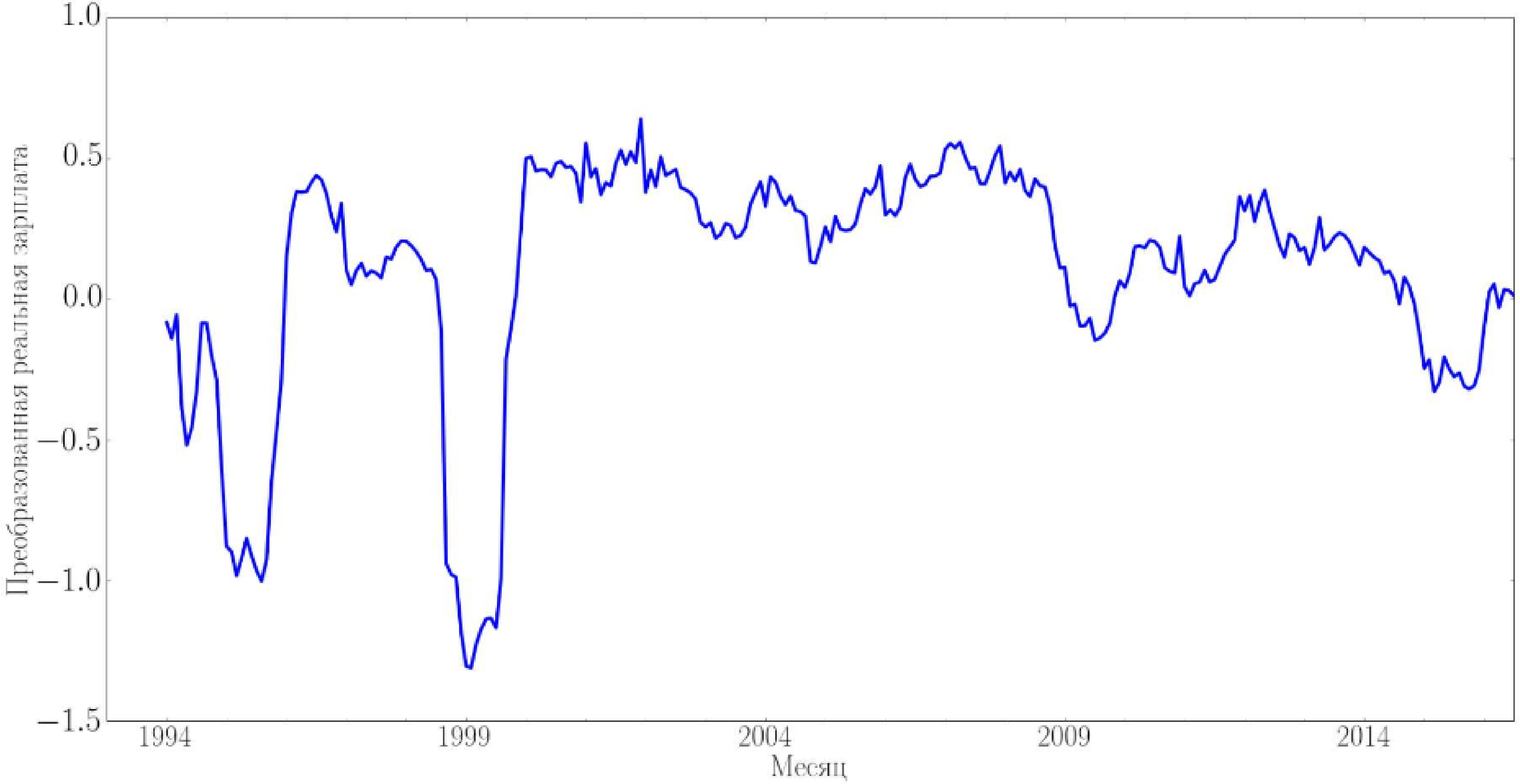
Реальная заработная плата

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



Критерий Дики-Фуллера: $p = 0.1661$.

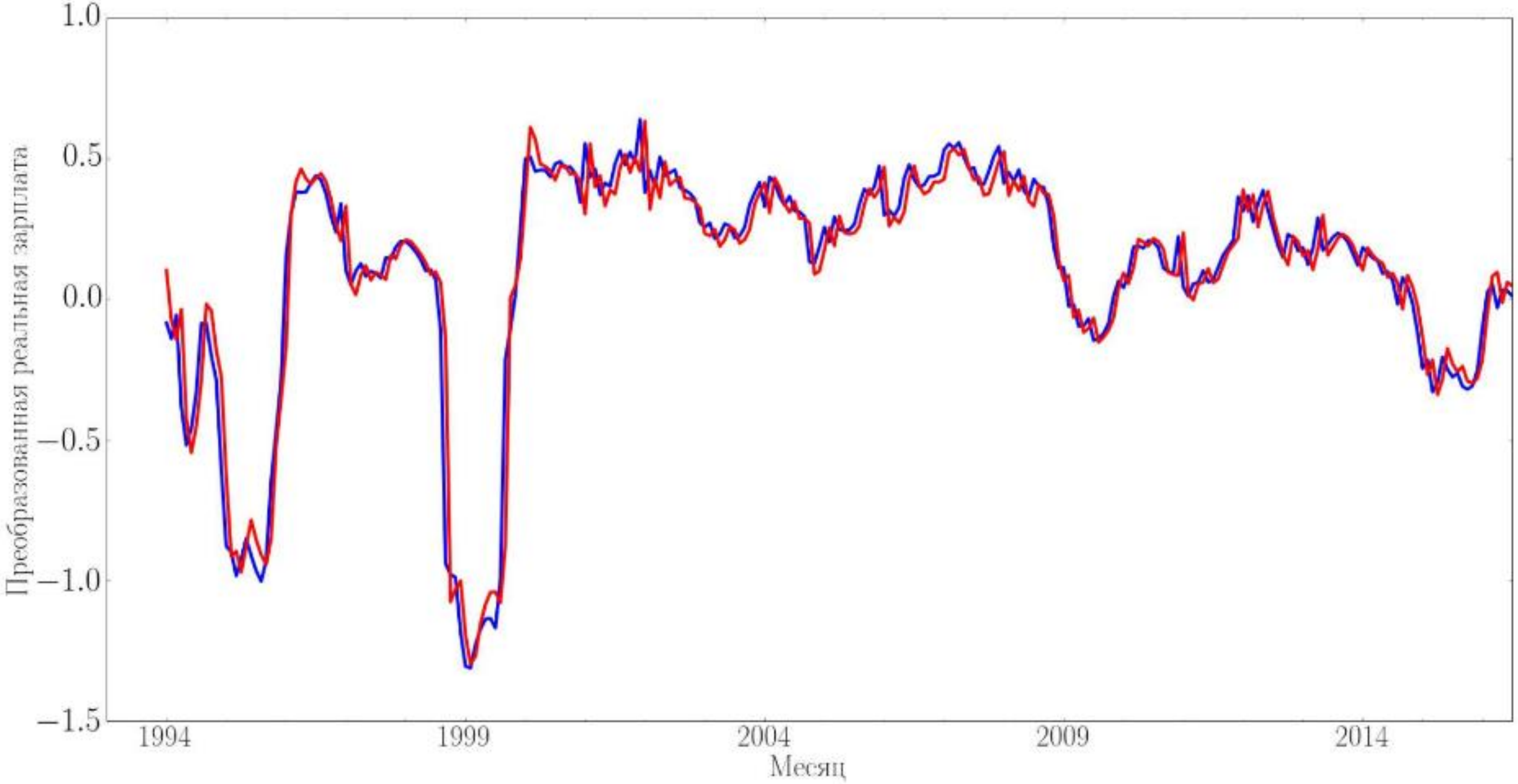
После сезонного дифференцирования:



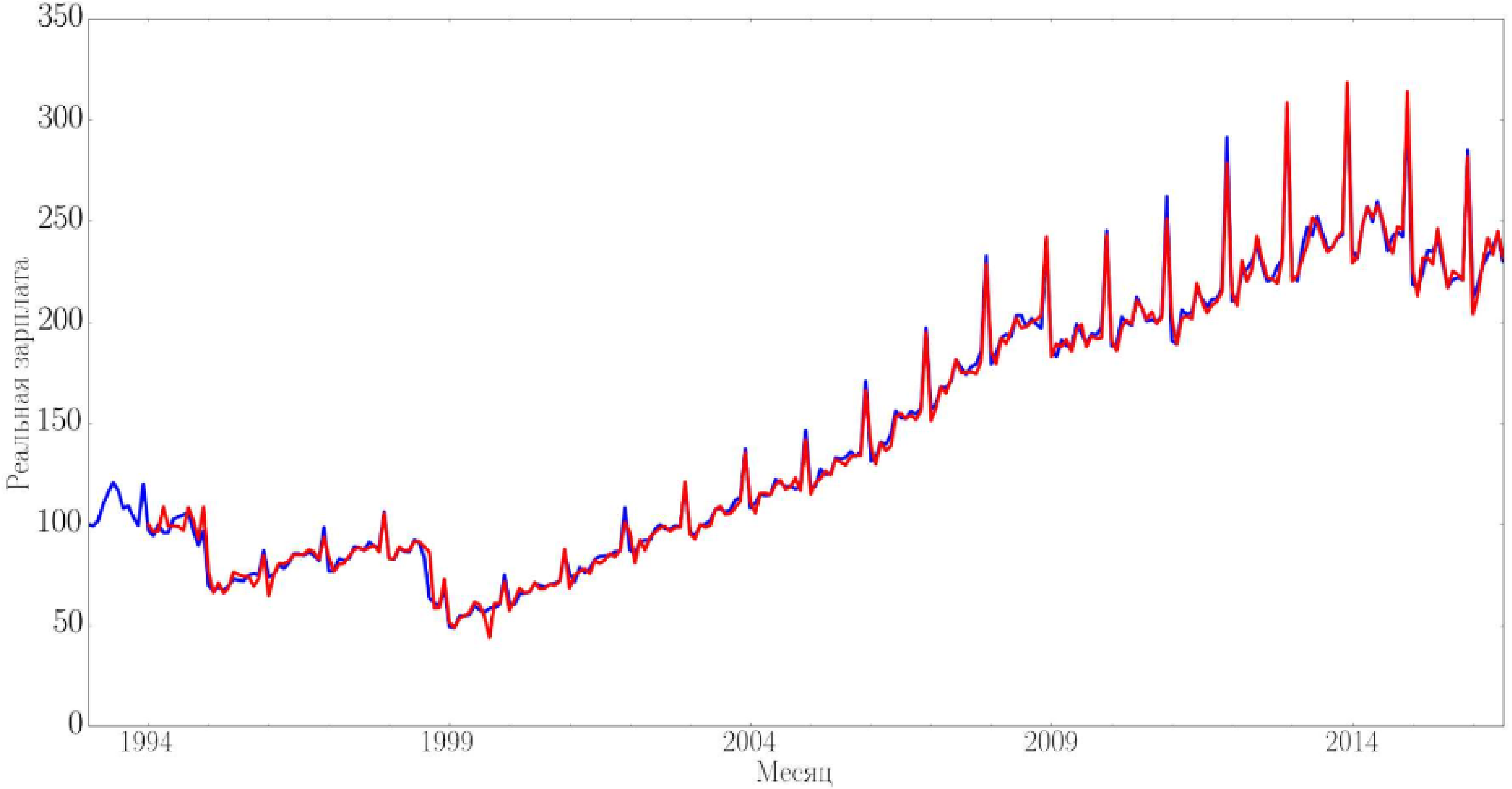
Критерий Дики-Фуллера: $p = 0.01$.

Реальная заработная плата

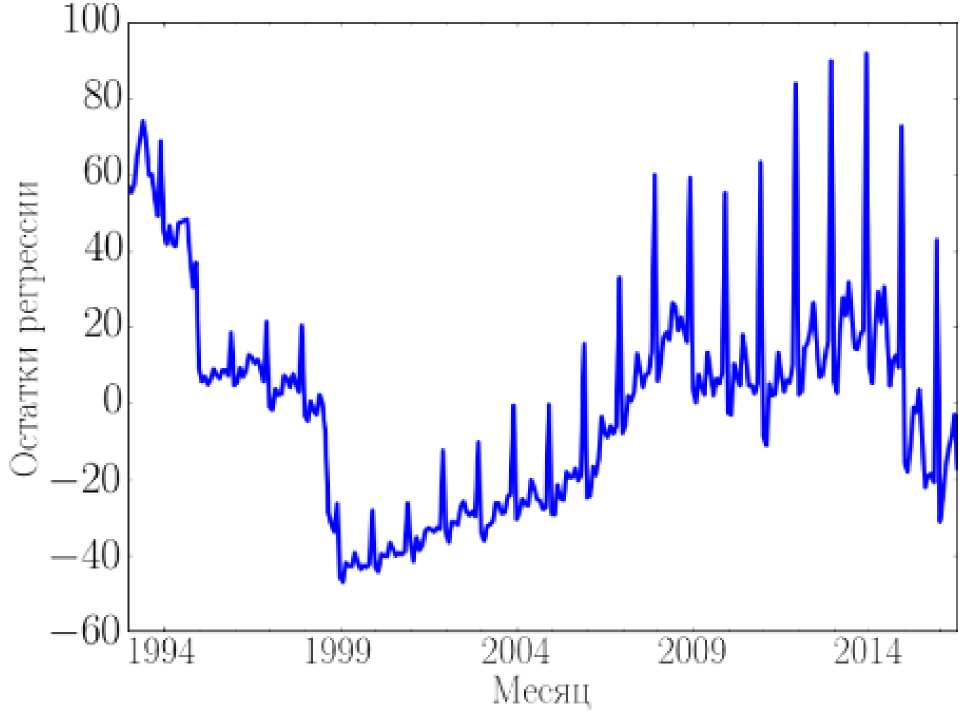
Модель $ARMA(2,2)$ для преобразованного ряда:

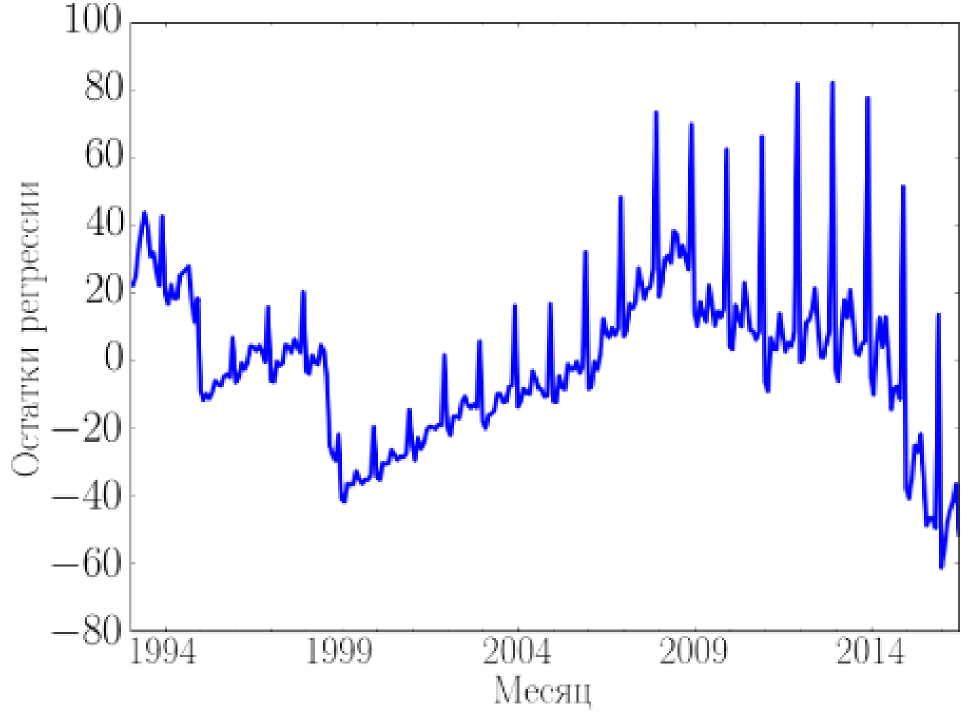


Модель $SARIMA(2,0,2)\times(0,1,0)$ с преобразованием Бокса-Кокса:

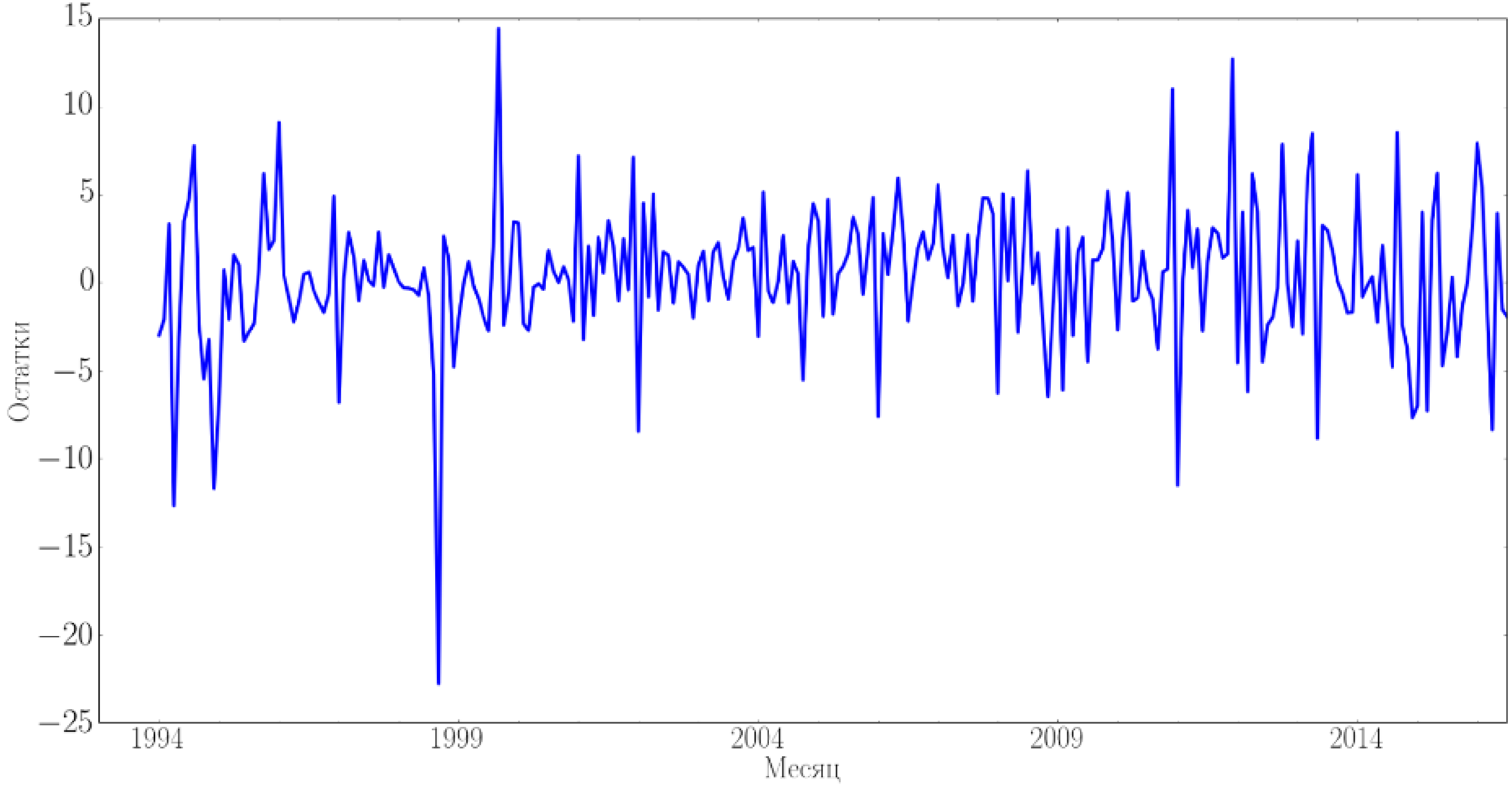


Остатки регрессий на время:





Остатки построенной модели:



Подбор параметров

- α, ϕ, θ
- d, D
- q, Q
- p, P

α, ϕ, θ

- 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.

d, D

- 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.

AIC — информационный критерий Акаике:

$$AIC = -2LL + 2k;$$

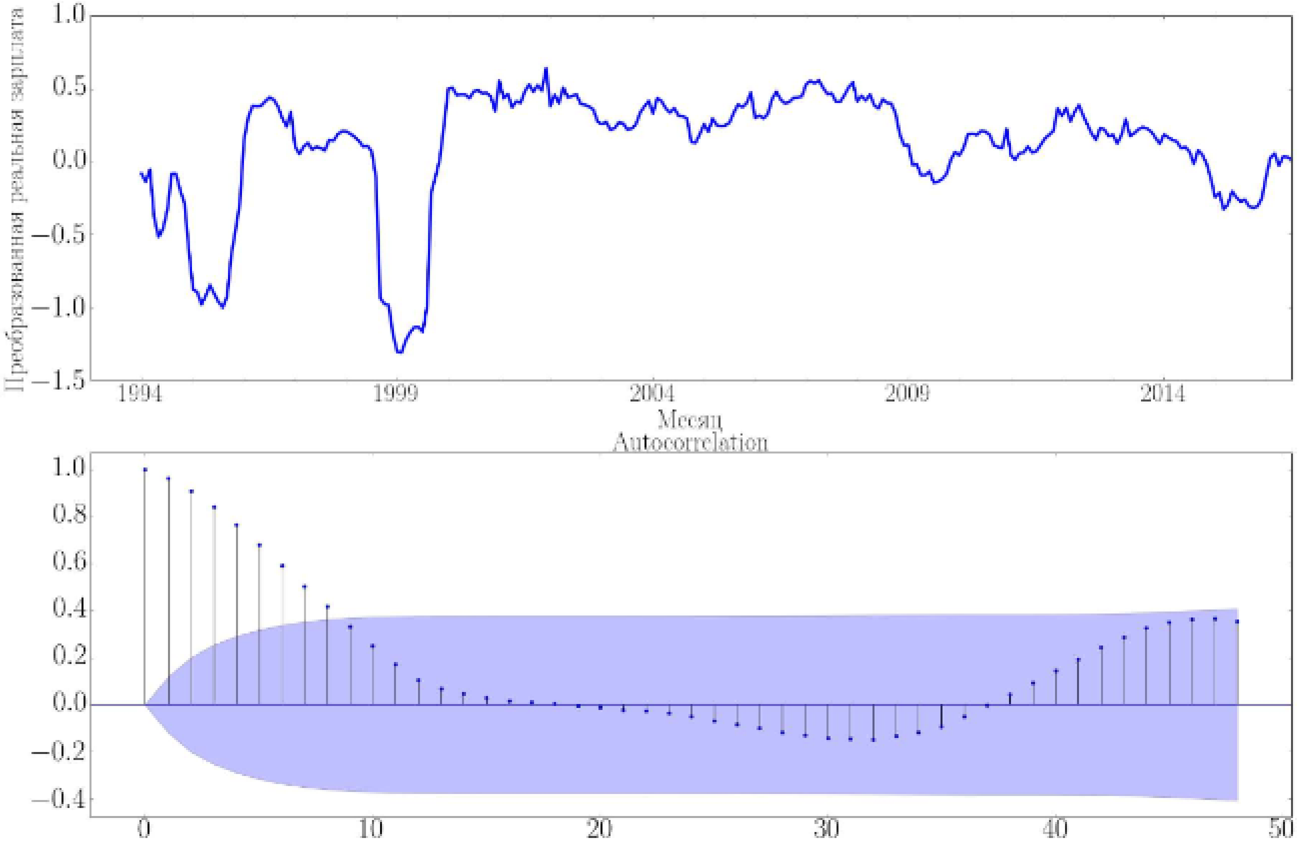
$AICc$ — он же с поправкой на случай небольшого размера выборки:

$$AICc = -2LL + \frac{2k^2)}{T - k - 1};$$

BIC (SIC) — байесовский (Шварца) информационный критерий:

$$BIC = -2LL + k\,(\log T - 2)\,.$$

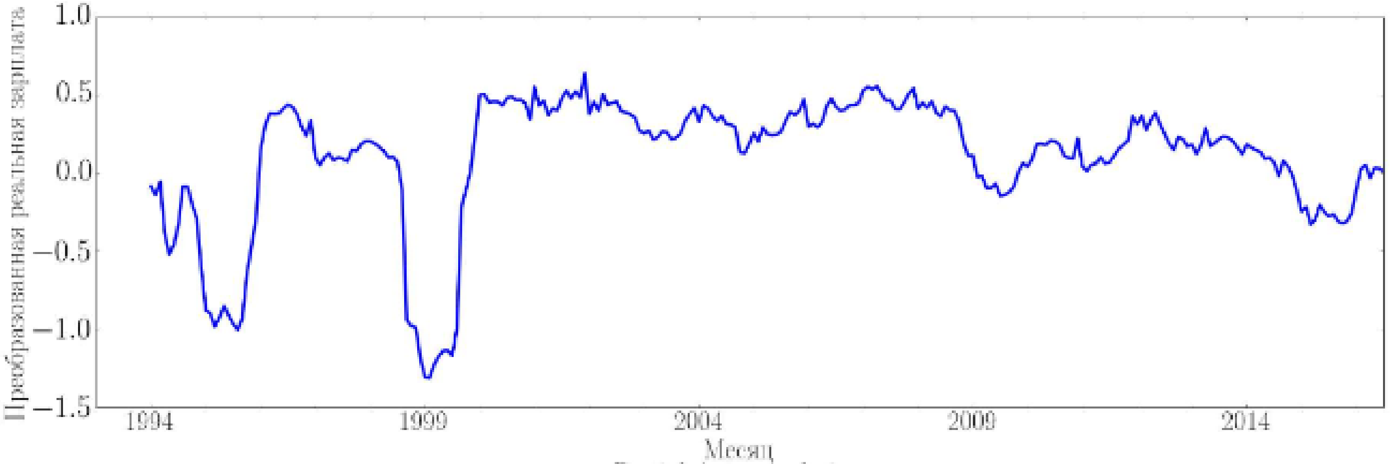
q, Q

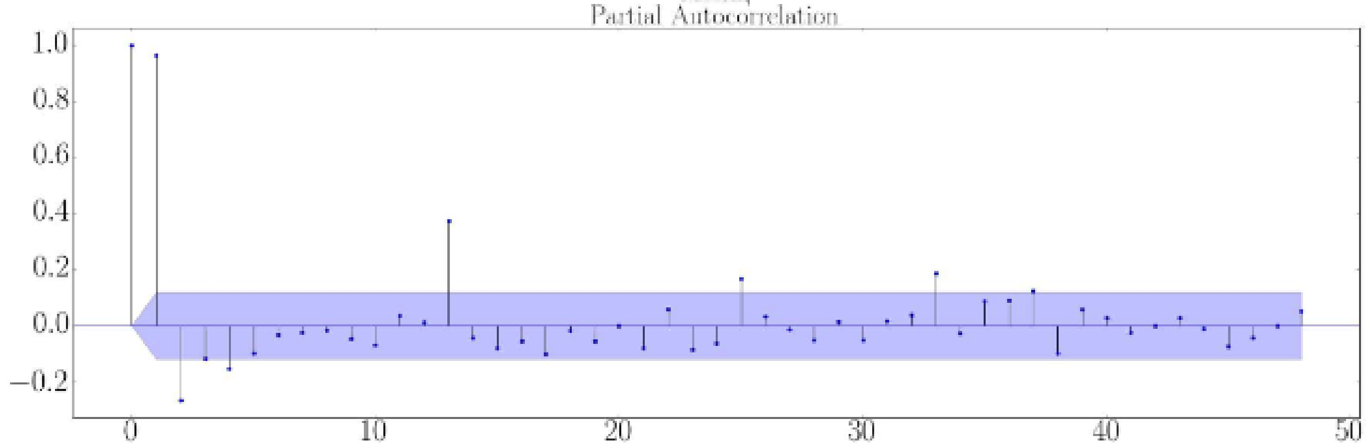


$Q * S$ — номер последнего сезонного лага, при котором автокорреляция значима (здесь 0).
q — номер последнего несезонного лага, при котором автокорреляция значима (здесь 8).

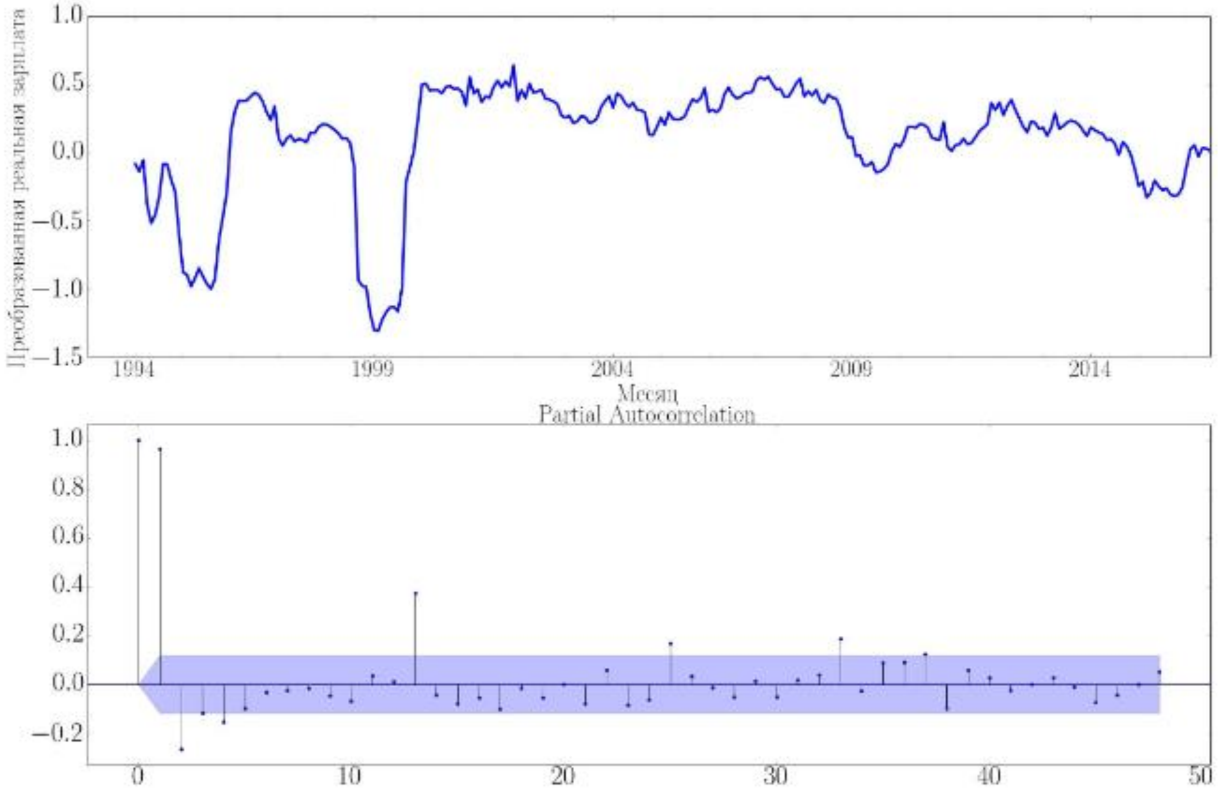
p, P

Частичная автокорреляция — автокорреляция после снятия авторегрессии предыдущего порядка.





p, P

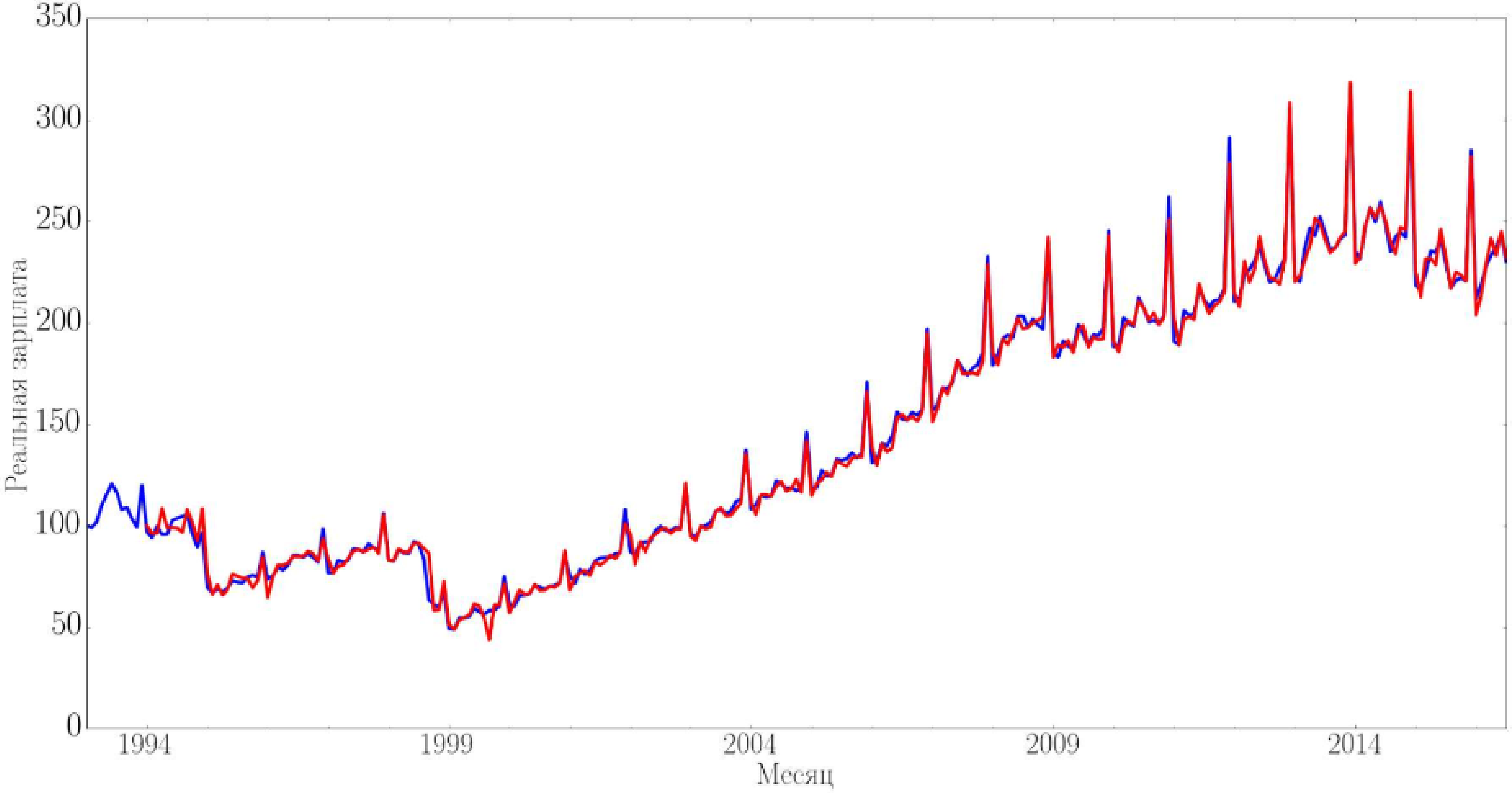


$P * S$ — номер последнего сезонного лага, при котором частичная автокорреляция значима (здесь 2).

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

Реальная заработная 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

1. Смотрим на ряд.
2. При необходимости стабилизируем дисперсию.
3. Если ряд нестационарен, подбираем порядок дифференцирования.
4. Анализируем ACF/PACF, определяем примерные p, q, P, Q
5. Обучаем модели-кандидаты, сравниваем их по AIC, выбираем победителя.
6. 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:

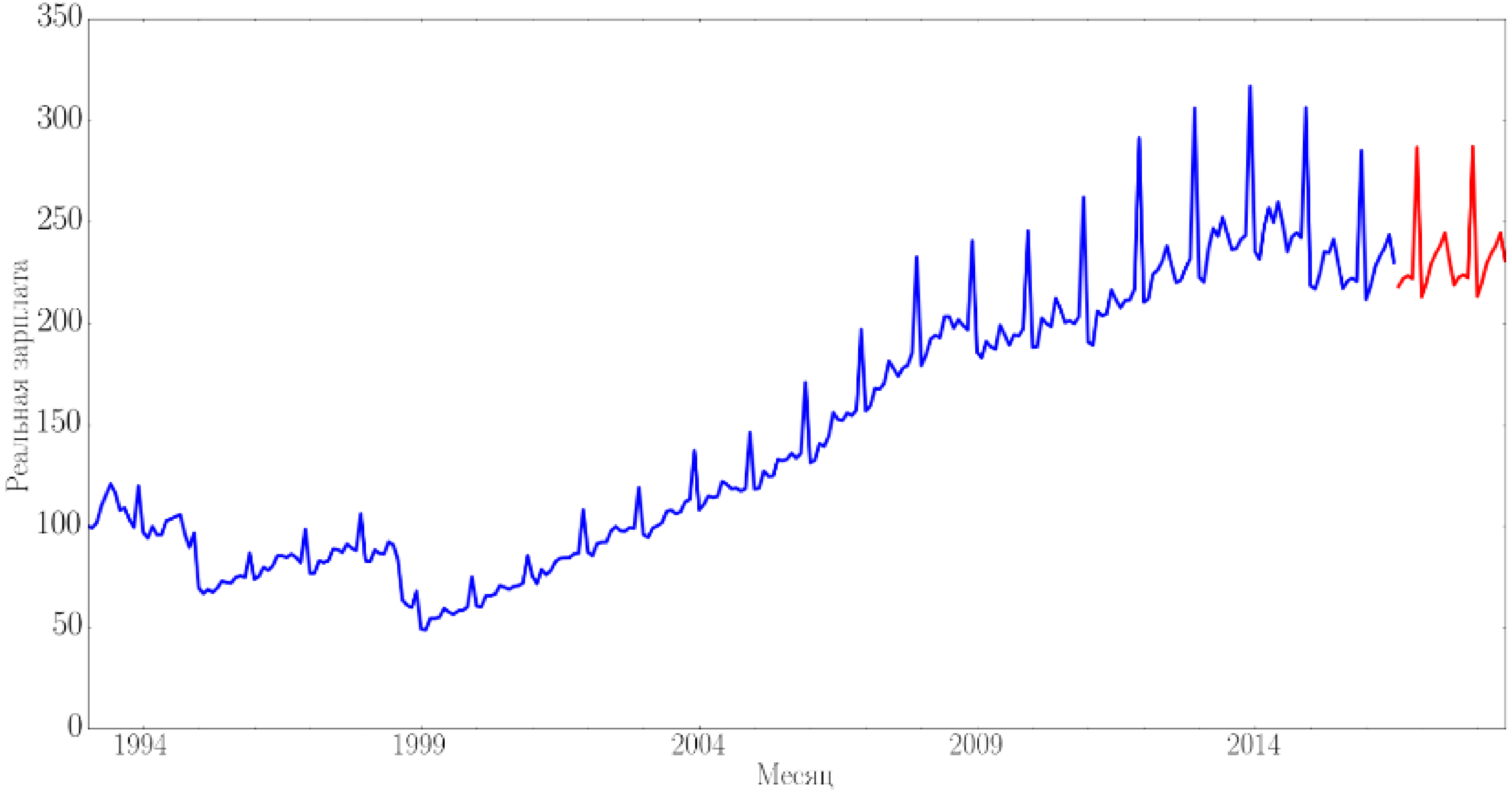
$$\hat{y}_{T+1|T} = \hat{\alpha} + \hat{\phi}_1 y_T + \cdots + \hat{\phi}_p y_{T+1-p} + \hat{\theta}_1 \hat{\varepsilon}_T + \cdots + \hat{\theta}_q \hat{\varepsilon}_{T+1-q}$$

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

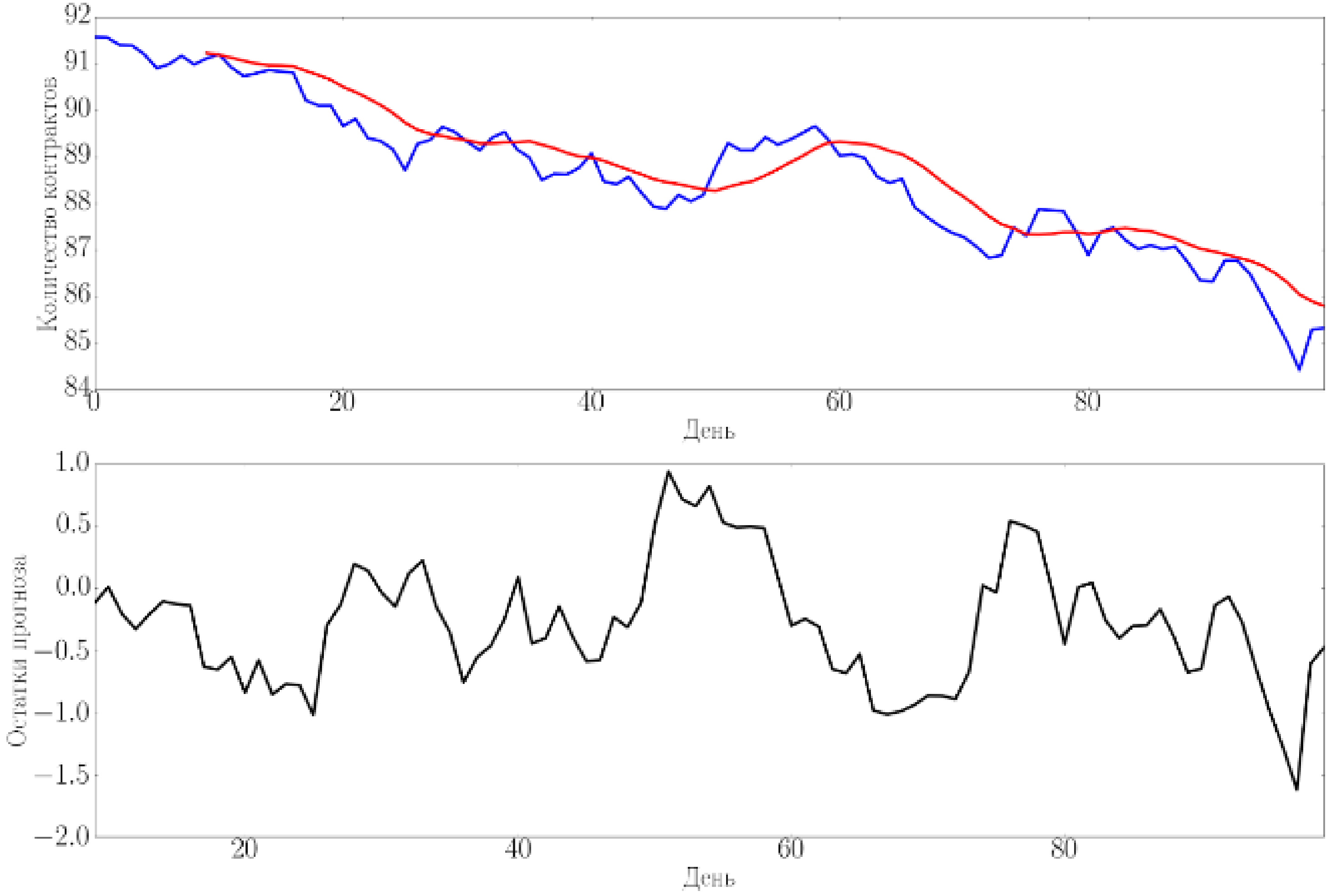


Остатки — разность между фактом и прогнозом:

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

Нужно проверять, обладают ли они некоторыми свойствами.

Несмещённость — равенство среднего значения нулю:

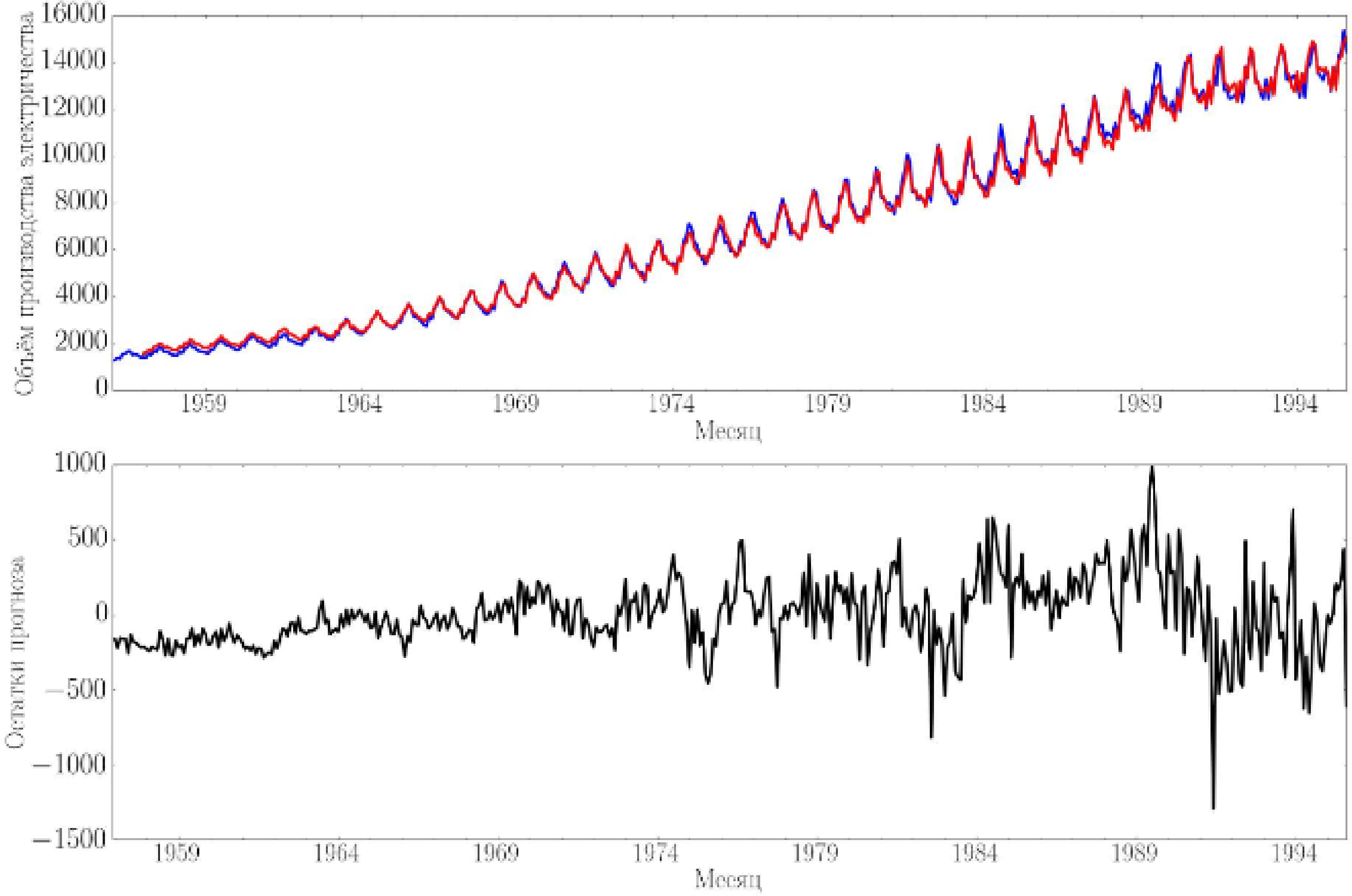


Несмещённость

- 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

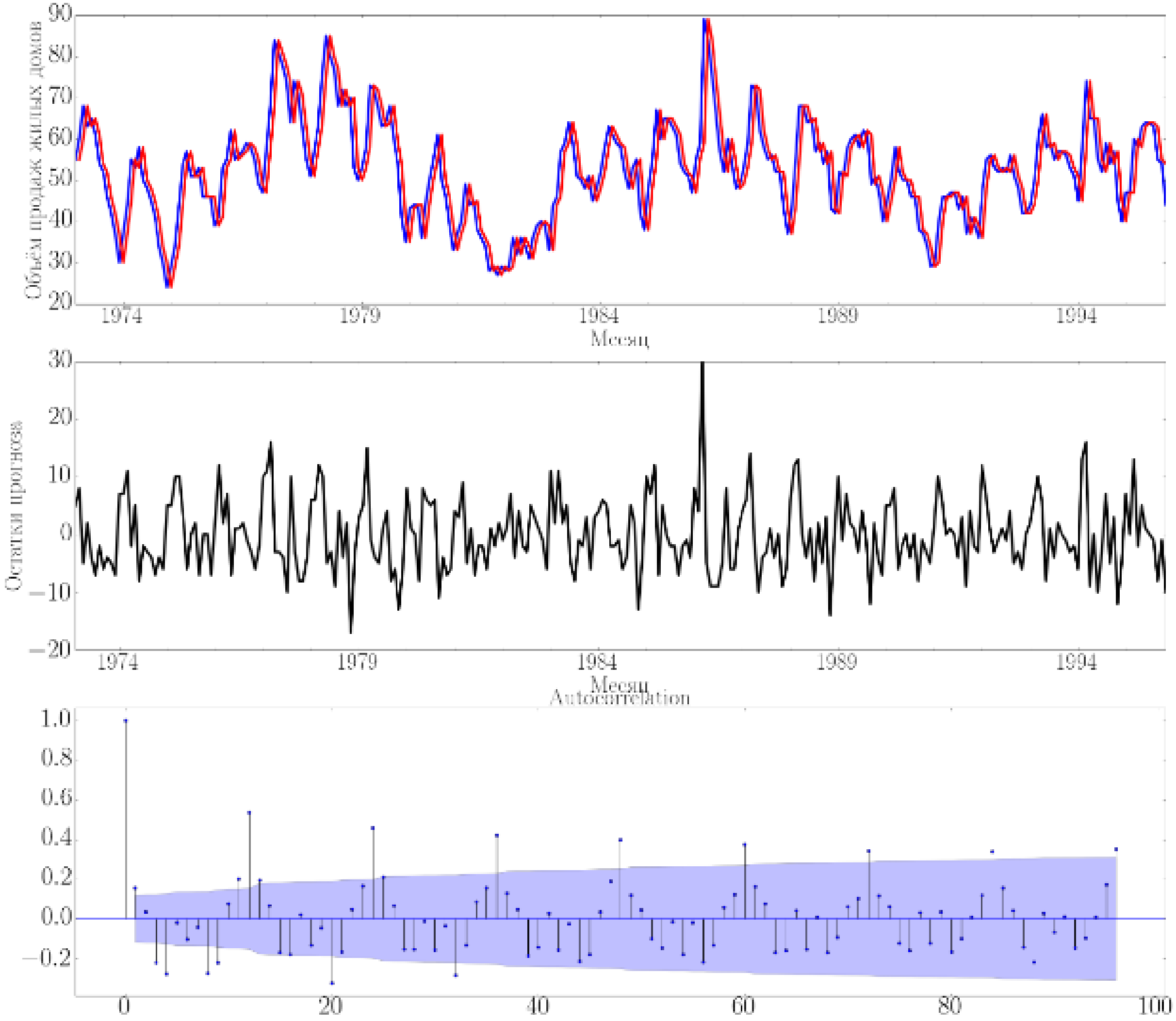
Стационарность — отсутствие зависимости от времени:



Стационарность

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

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