

**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ
СІКОРСЬКОГО”**

**ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ
КАФЕДРА ІНФОРМАЦІЙНОЇ БЕЗПЕКИ**

КУРСОВА РОБОТА

Дисципліна: «Аналіз даних та статистична обробка сигналів»

Тема: «Застосування математичних моделей для прогнозування часових рядів»

Виконав

студент 6 курсу групи ФІ-51м

Кригін Валерій Михайлович

Перевірів

д.т.н., професор, с.н.с.,

Архипов Олександр Євгенович

ЗМІСТ

Вступ	3
1 Прогнозування часових рядів	4
1.1 Видалення аномалій	4
1.2 Вибір правильного зглажуючого віконця	6
1.3 Невипадкові похибки	8
1.4 Прогнозування	15
1.5 Прогнозування невинуватених помилок	19
1.6 Прогноз процесу	21
2 Прогнозування за моделлю залежності	22
2.1 Зглажування	23
2.2 Прогноз часових рядів	26
2.3 Побудова моделі	26
Висновки	29
Перелік посилань	31

ВСТУП

Об'єкт дослідження — часові ряди та методи їх прогнозування.

Предмет дослідження — якість прогнозування часових рядів різними методами.

Завдання:

- 1) зробити прогноз часового ряду на основі ретроданих;
- 2) зробити прогноз часового ряду на основі інших рядів, від яких він залежить;
- 3) порівняти ці два підходи.

Обов'язкове використання методу експоненційного згладжування для побудови тренду часових рядів та методу авторегресії для оцінки невинпадкових помилок.

1 ПРОГНОЗУВАННЯ ЧАСОВИХ РЯДІВ

На вході дано часовий ряд Y з шумом E

$$Z = Y + E,$$

де

$$\text{cov}(E_i, E_j) = \begin{cases} \sigma_e^2, & i = j, \\ 0, & i \neq j. \end{cases}$$

Також наявні аномальні точки, які заважають процесу згладжування та прогнозування.

1.1 Видалення аномалій

Для початку побудуємо згладжування на основі експоненційно зваженого ковзкого середнього з $\alpha = \frac{1}{3}$, при якому довжина віконця дорівнює 6 (рис. 1.1).

Одразу видно, де знаходяться аномальні точки, проте треба застосувати метод, який ґрунтується на здоровому глузді та математичній статистиці. Оскільки аномальні точки заважають будувати лінію тренду, потрібно порівняти якість фільтрації при видаленні різних точок даного ряду. Введемо $Y_{fixed}(t)$ як ряд, в якому “виправлено” точку t . Це означає, що дані в цій точці мають значення, яке не є аномальним. Під якістю фільтрації розуміємо середньоквадратичне відхилення між лінією тренду та самим рядом. Порівняємо якість згладжування для Y з якістю згладжування його виправленого аналогу. Логічно, що можна порівняти

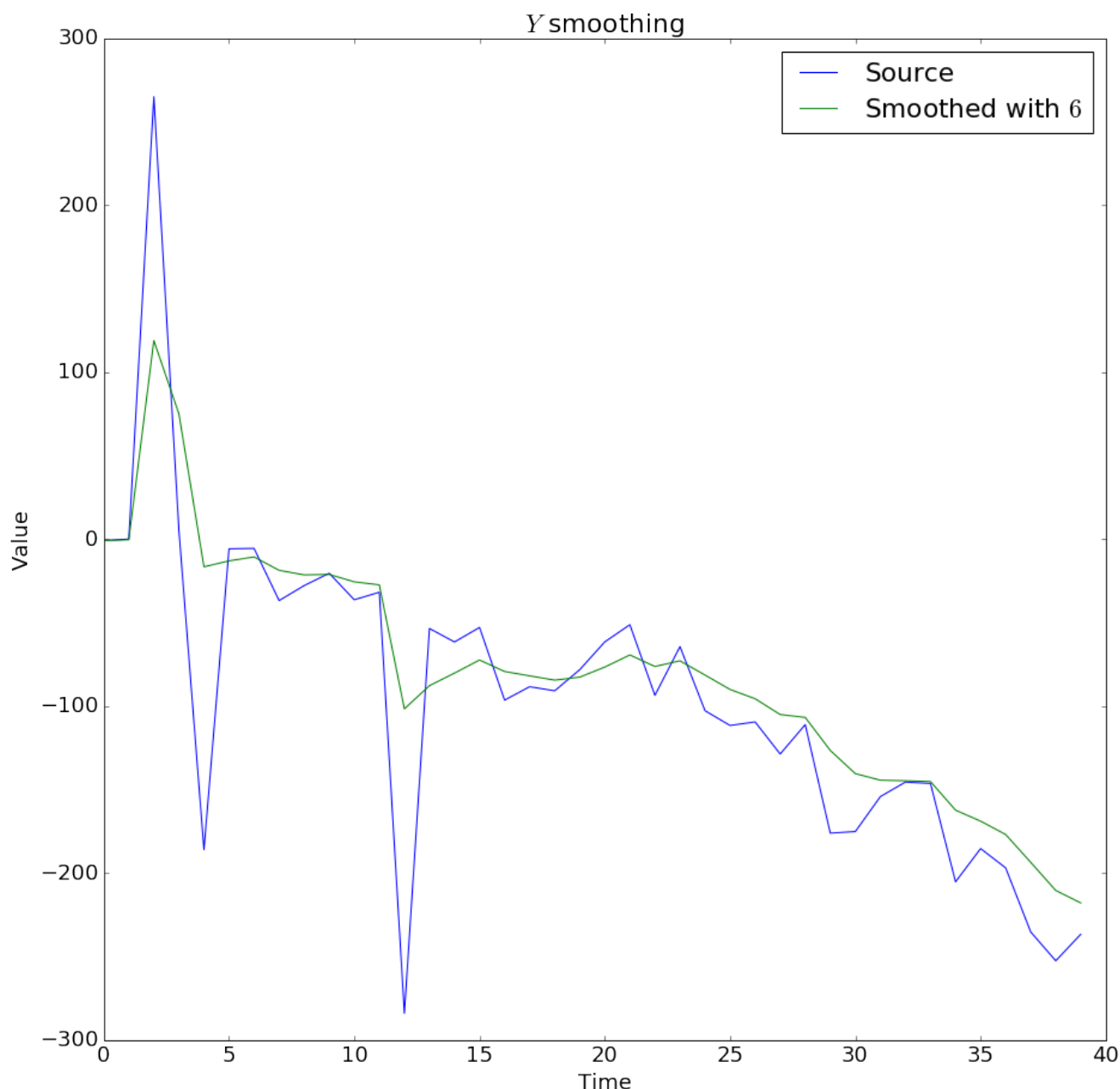


Рисунок 1.1 — Згладжування ряду з аномаліями

їх співвідношення з певним граничним значенням

$$V(t) = \frac{\text{Var}(Y - Y^{\text{smoothed}})}{\text{Var}(Y_{\text{fixed}}(t) - Y_{\text{fixed}}^{\text{smoothed}}(t))} > V_{\text{critical}}$$

Отримали вираз, що відомий як F -тест [1]: маємо вибіркові дисперсії двох вибірок, що за умовою мають нормальний закон розподілення. Аномальні точки

будемо знаходити одну за одною — знаходити найгіршу та виправляти її, якщо вона дійсно аномальна

$$\max_t F_{F,T,T-1}(V(t)) > F_{F,T,T-1}^{critical} \implies t_{anomaly} = \max_t F_{F,T,T-1}(V(t)).$$

Поглянемо на те, як змінюється вибіркова дисперсія помилки при відкиданні кожної точки на рис. 1.2.

З рисунку видно піки саме в тих місцях, де передбачалась наявність аномальних точок. Проте зараз в нас є міра їх “аномальності” у вигляді ваги хвостів розподілу Фішера (рис. 1.3).

Відкидаємо одну за одною точки, при яких дисперсія значно змінюється, а саме такі, що з ймовірністю 0.9 це дисперсія іншої виборки (рис. 1.4, 1.5, 1.6).

Аномальні дані знаходилися в точках 2, 12 і 4. Їх було виправлено за принципом найближчих сусідів — середнє арифметичне найближчих точок. Маємо графік з даними без аномалій та згладжуванням, яке тепер значно краще описує ряд

1.2 Вибір правильного згладжуючого віконця

Для подальших обчислень потрібно обрати коректні параметри згладжування, а саме параметр α для експоненційно згладженого ковзкого середнього. Зазвичай це робиться в залежності від природи даних — їх циклічності або характерного часу зміни. Цієї інформації немає, проте ми знаємо, що помилки мають гаусовий розподіл. Скористуємося відомим методом для визначення “нормальності” розподілу виборки — методом Д’Аугустіно Ральфа [2]. Цей тест на виході дає відстань розподілу даної виборки до класу нормально розподілених виборок. Ми хочемо обрати таке згладжування, щоб розподіл помилок був якомога більш

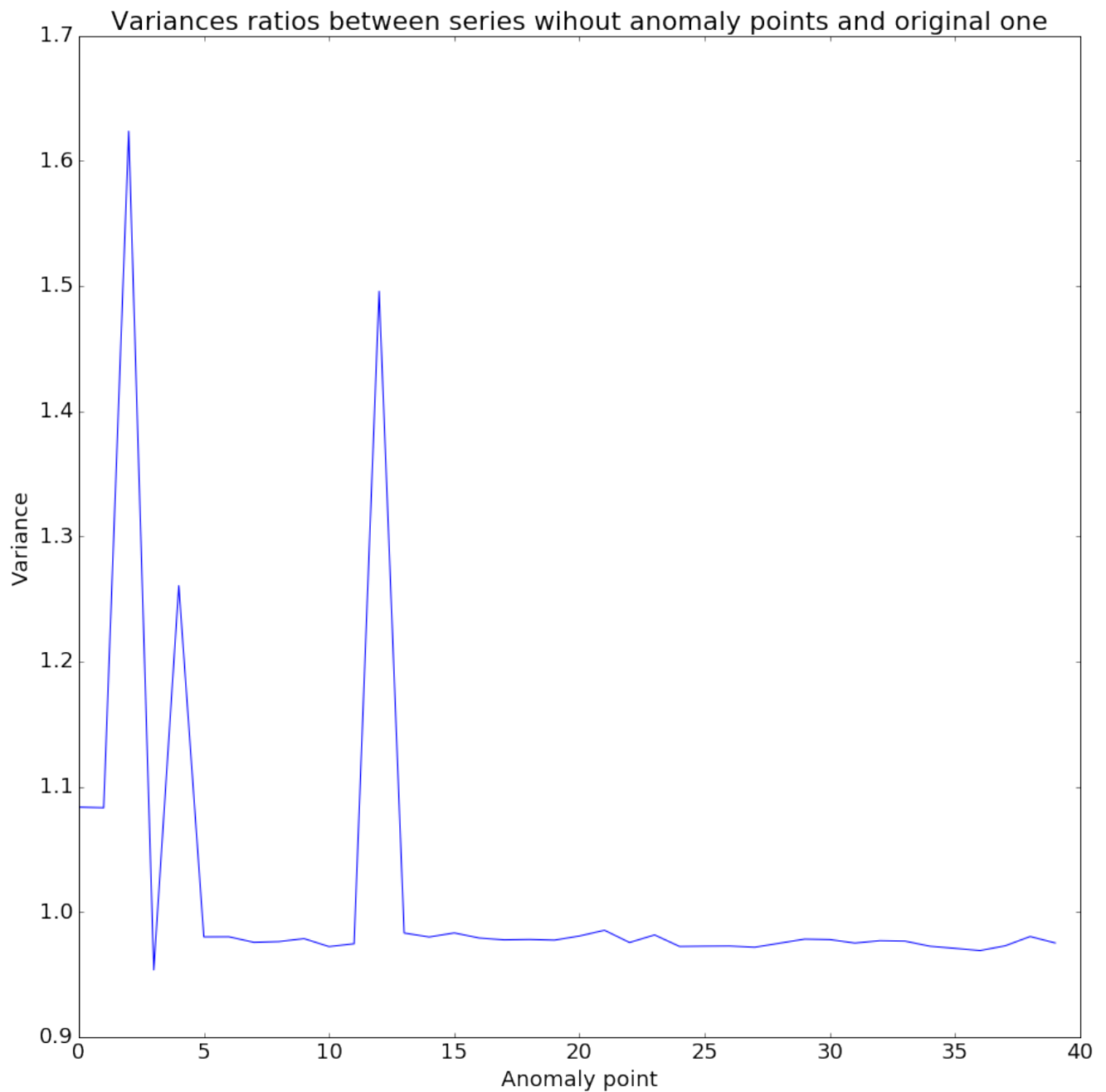


Рисунок 1.2 — Відношення дисперсій помилок за виправлення певних точок

гаусовим. Потрібно обрати таке α , для якого відстань між розподілом помилок та нормальним розподілом була найменшою. Ця умова виконується для віконця шириною 7, що відповідає $\alpha = \frac{1}{4}$. Маємо похибку $\sigma = 16.3$ (рис. 1.8).

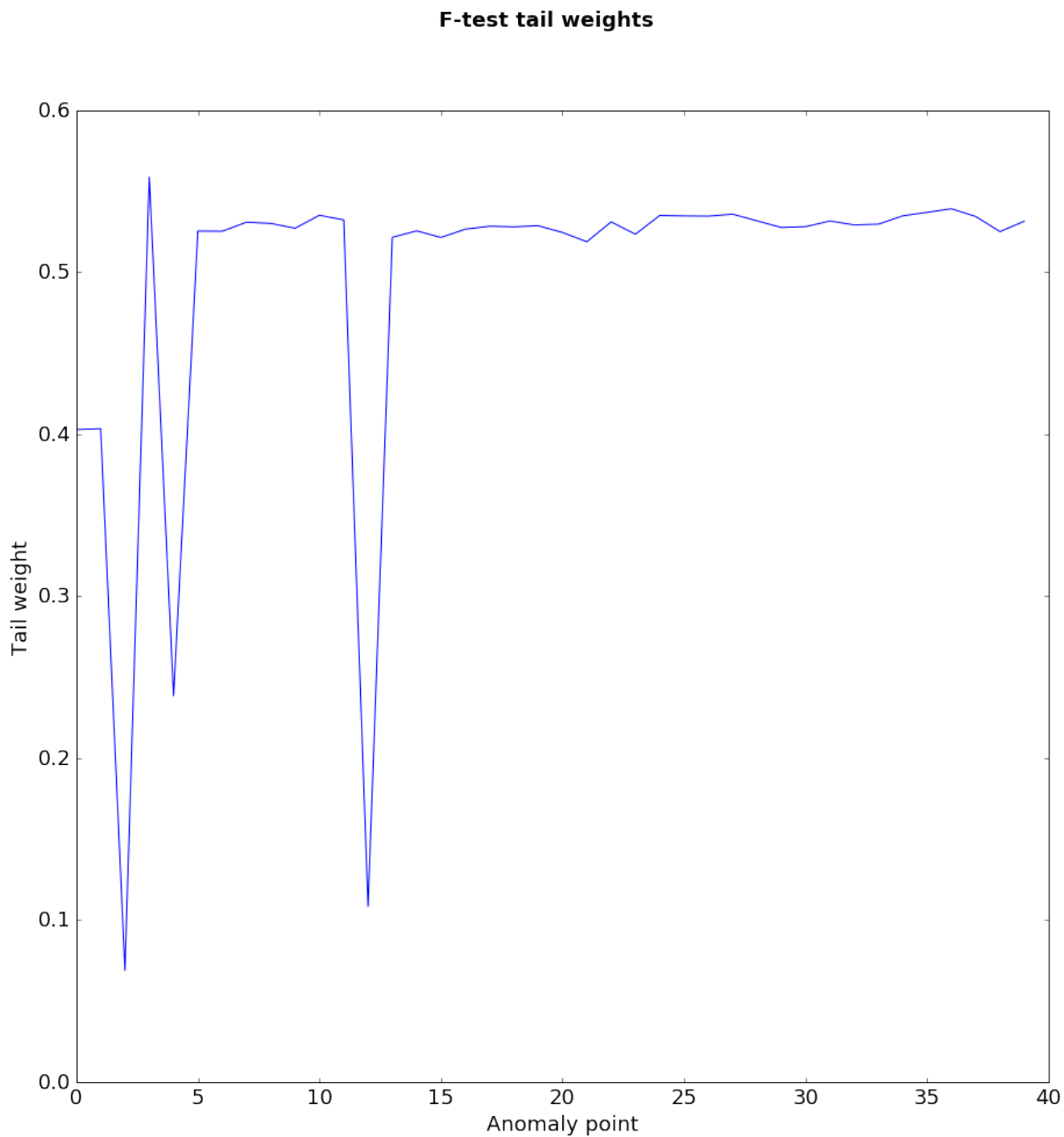


Рисунок 1.3 — Хвости розподілу Фішера для відношень дисперсій виправлених рядів

1.3 Невипадкові похибки

Лінія тренду може врахувати не всі закономірності ряду, особливо якщо невідома його природа. Наприклад, ми маємо ряд Y . Позначимо його оцінку \tilde{Y} .

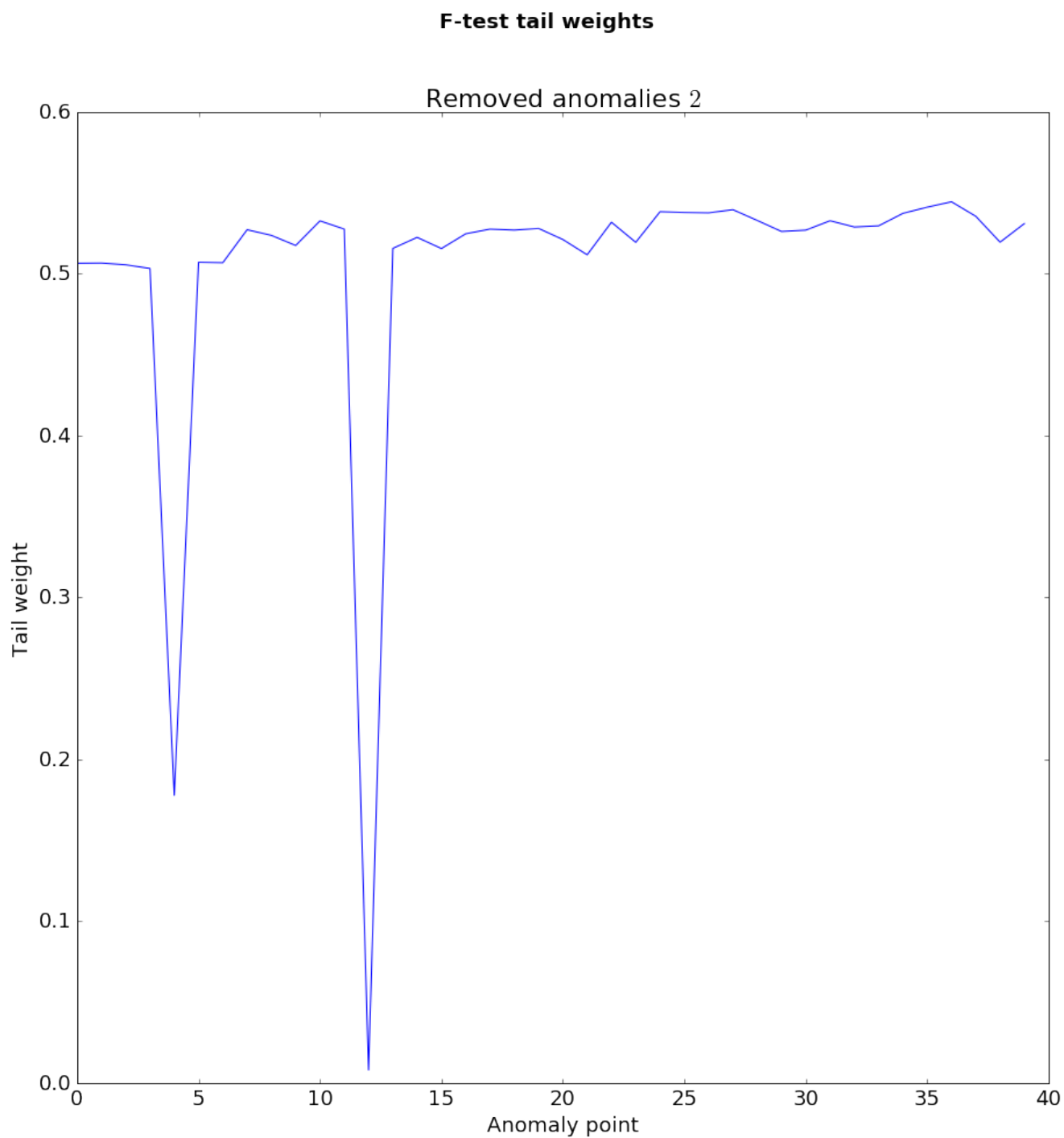


Рисунок 1.4 — Хвости розподілу Фішера для відношень дисперсій виправлених рядів

Тоді

$$\varepsilon = Y - \tilde{Y}.$$

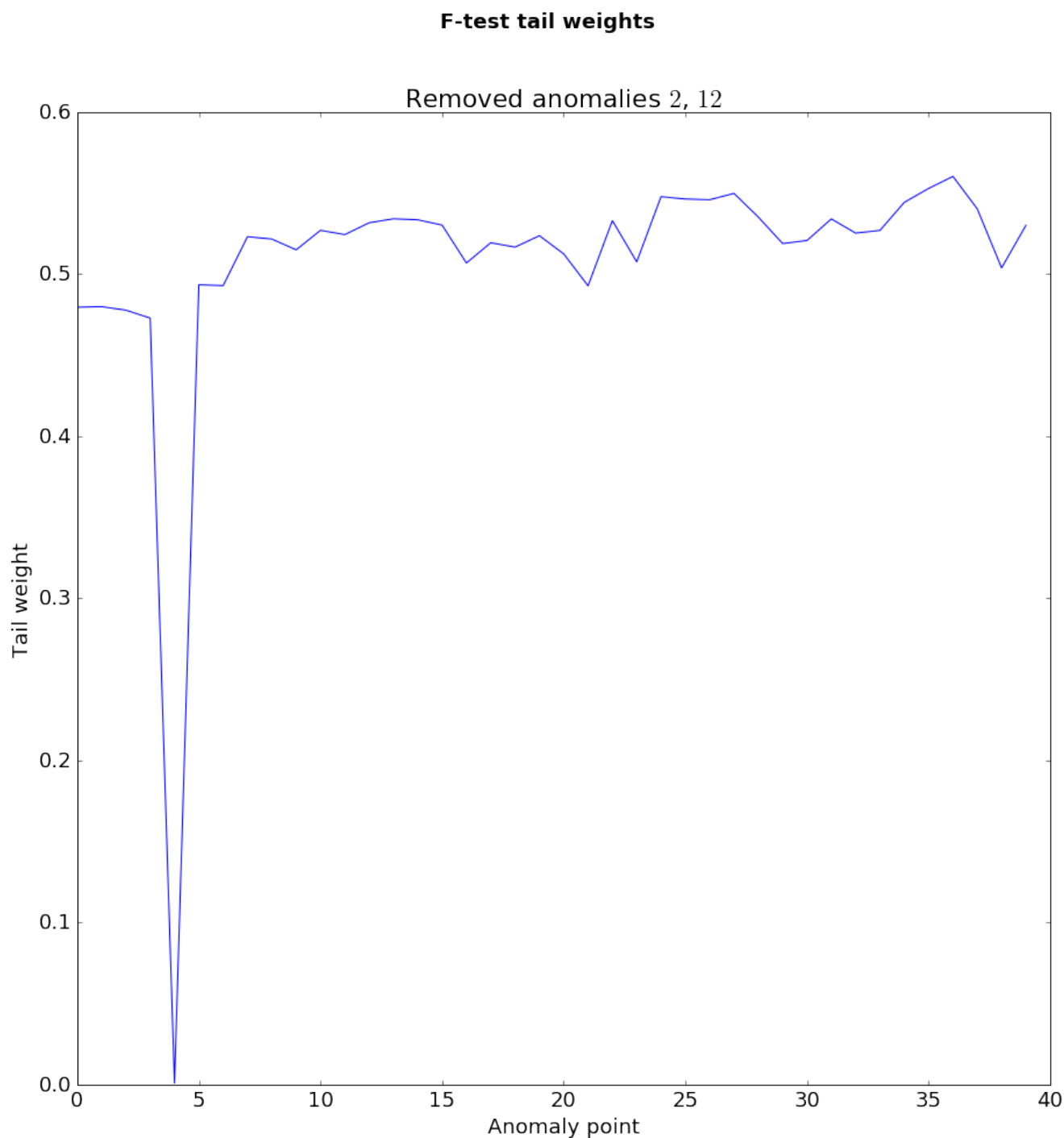


Рисунок 1.5 — Хвости розподілу Фішера для відношень дисперсій виправлених рядів

Це призводить до того, що деякі помилки не випадкові, тобто між ними є певна математична залежність. Проаналізуємо простий випадок, коли присутній деякий лаг τ , і помилки впливають не тільки на сусідні, але й через певний час τ .

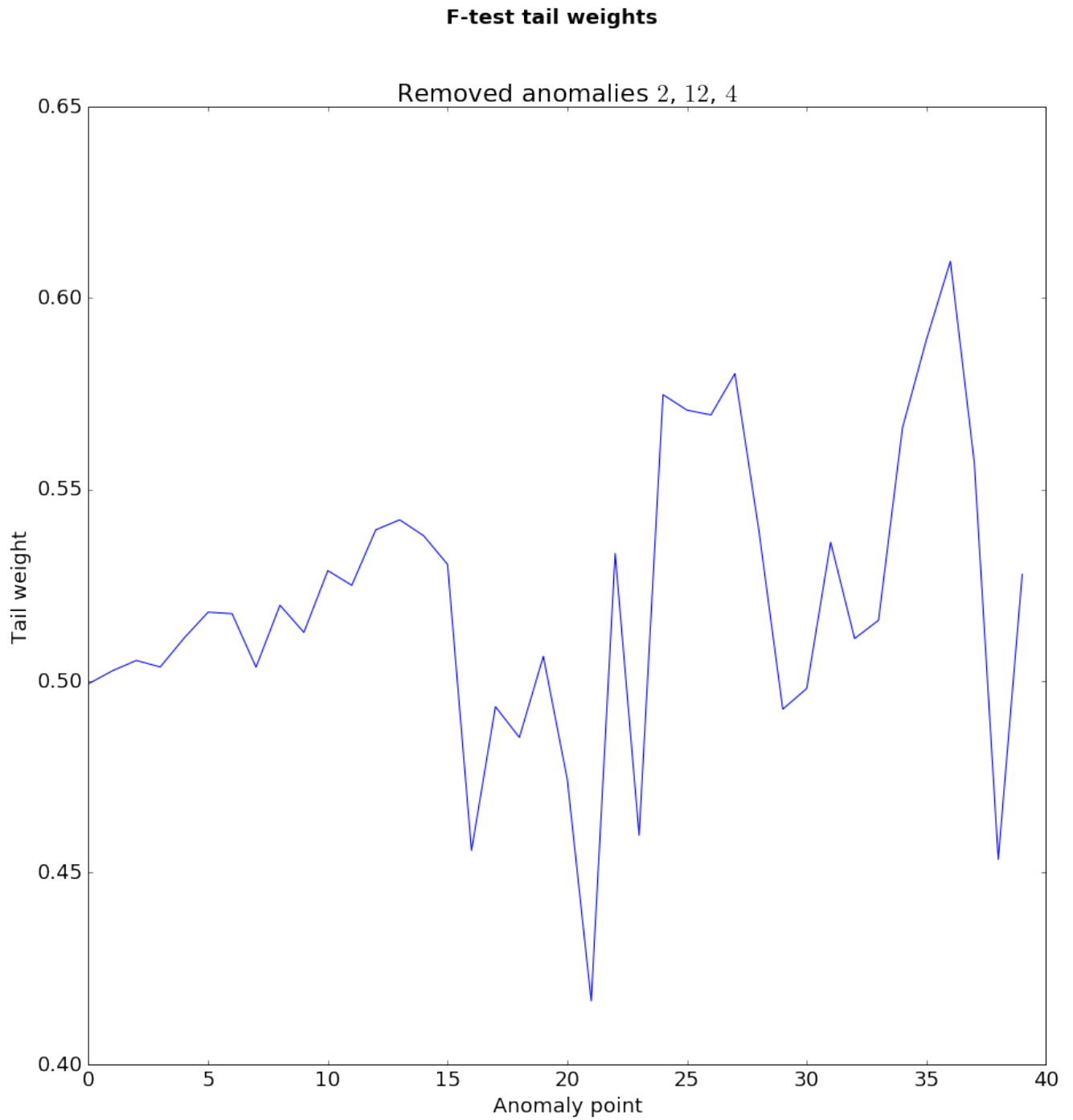


Рисунок 1.6 — Хвости розподілу Фішера для відношень дисперсій виправлених рядів

Тоді помилка має випадкову e та не випадкову частину δ

$$\varepsilon_{t+\tau} = \delta_{t+\tau} + e_{t+\tau}.$$

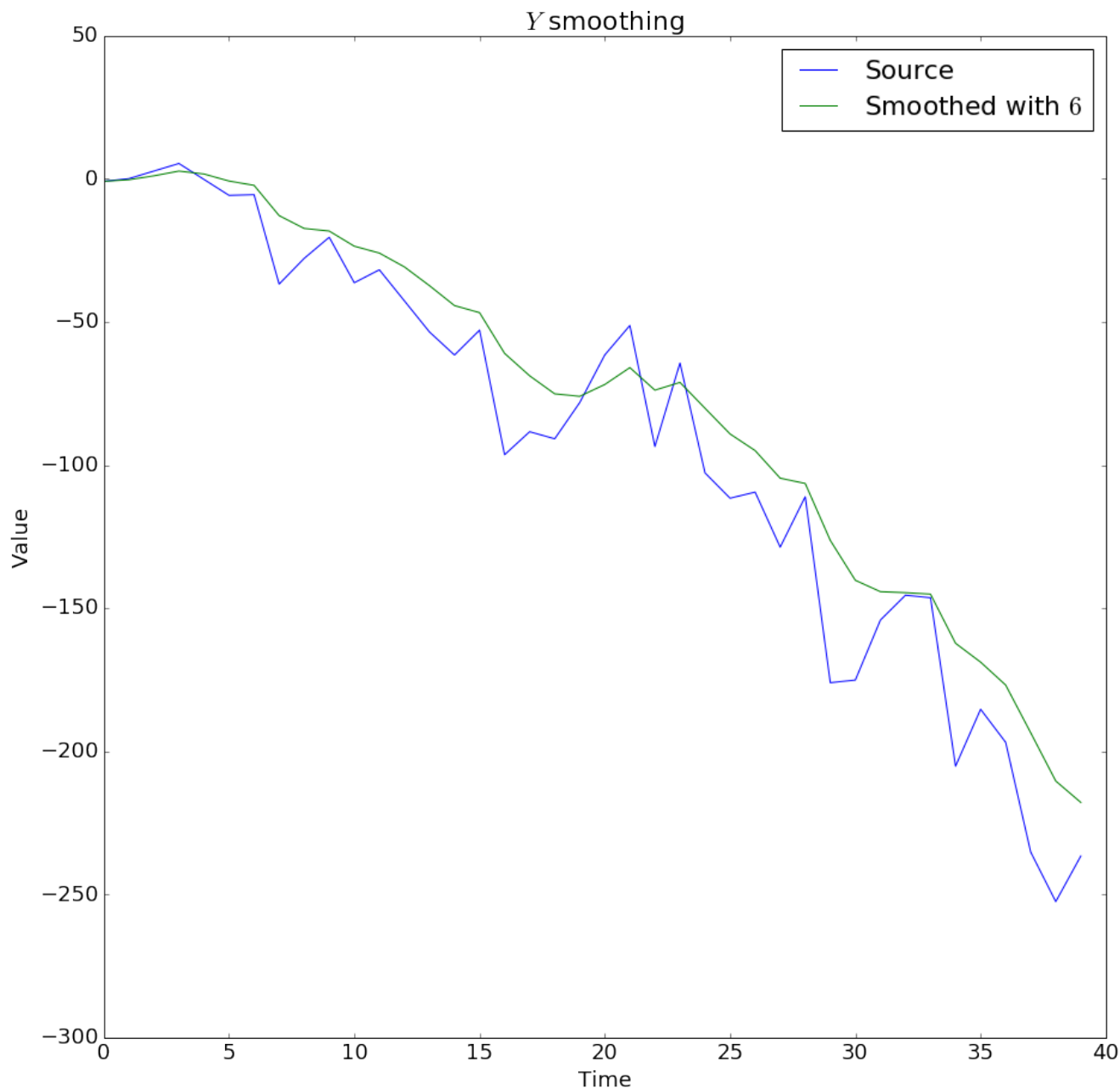


Рисунок 1.7 — Оцінка виправленого ряду

Аналітичний вираз для не випадкової частини — лінійна функція

$$\delta_{t+\tau} = a_0 + \sum_{i=0}^{\tau-1} \delta_{t+i} \cdot a_{i+1}. \quad (1.1)$$

Дійсно видно, що середнє значення похибки ненульове, тобто її можна як мінімум зсунути по вертикальній осі (рис. 1.9).

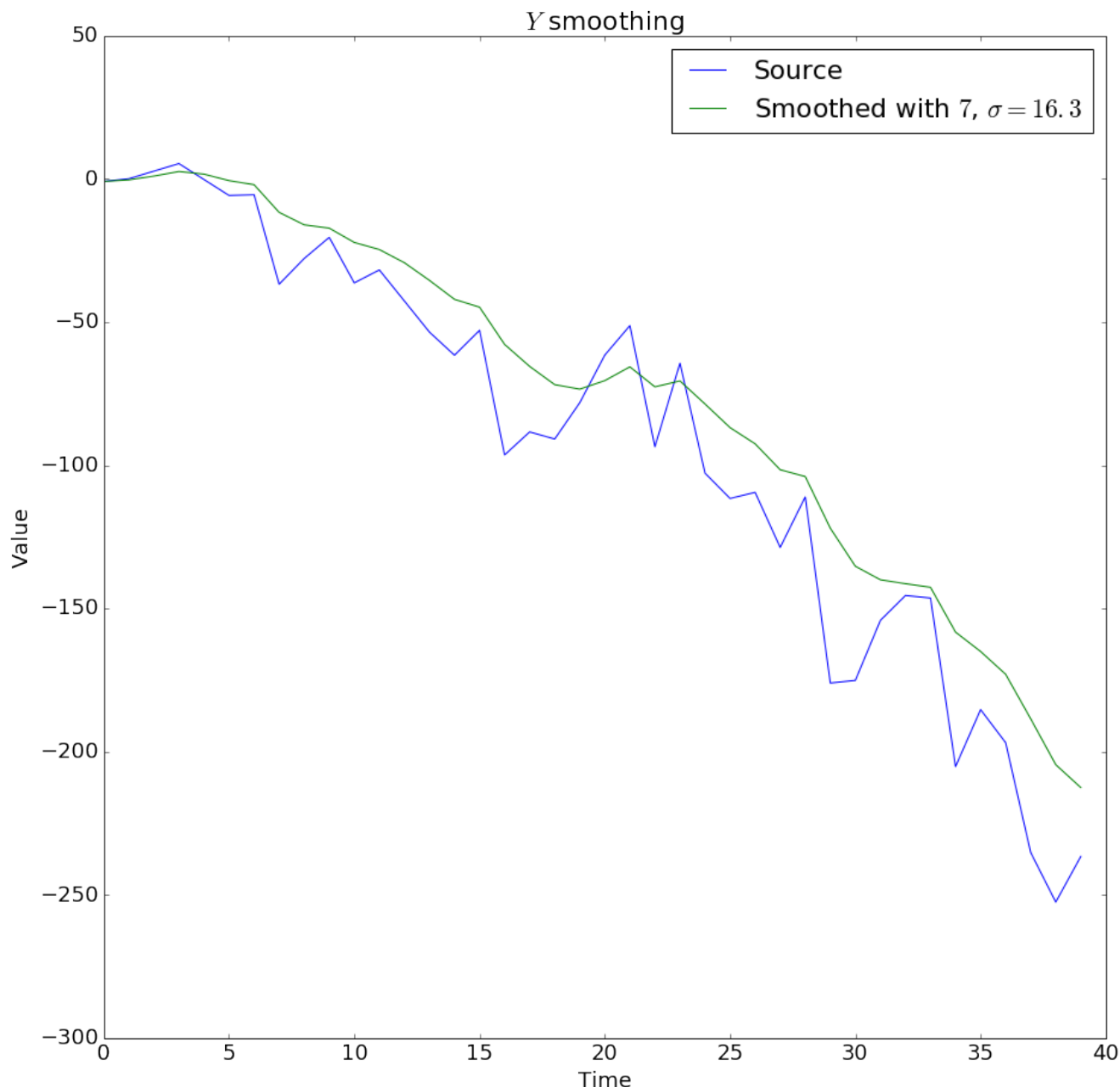


Рисунок 1.8 — Помилки тренду

Щоб визначити величину лагу, потрібно порахувати авторегресії різних порядків. Далі методом найменших квадратів розраховуються коефіцієнти для рівняння (1.1) (рис. 1.10).

Обравши крок 2, отримали хоч і невелику залежну частину, проте помилка піднялась і тепер її середнє нульове. Отримані коефіцієнти a_i : 0.12, 0.31, -9.81 (рис. 1.11).

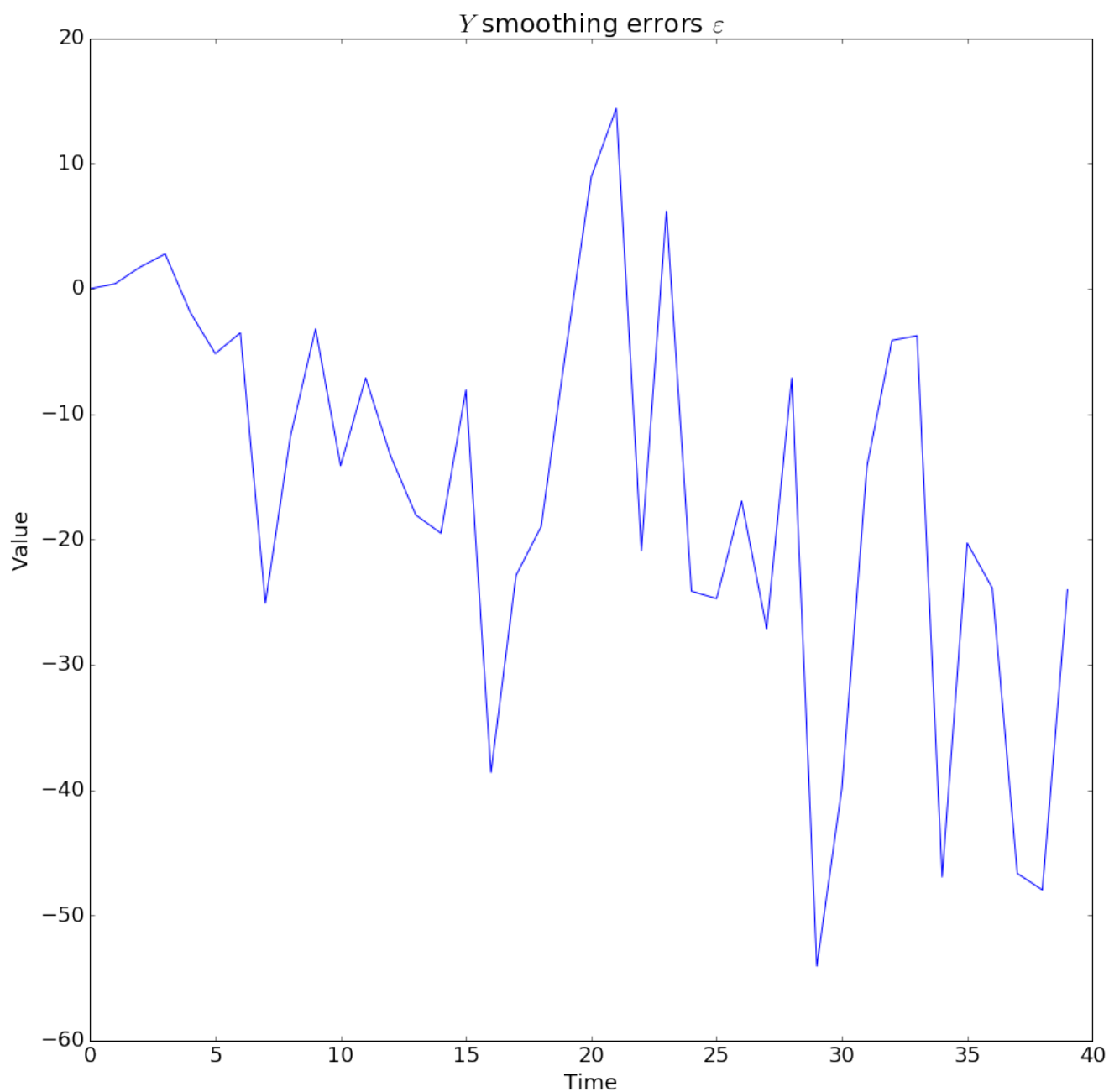


Рисунок 1.9 — Фільтр з кращими параметрами

Нова апроксимація ряду виглядає наступним чином та має меншу похибку $\sigma = 15.1$ (рис. 1.12).

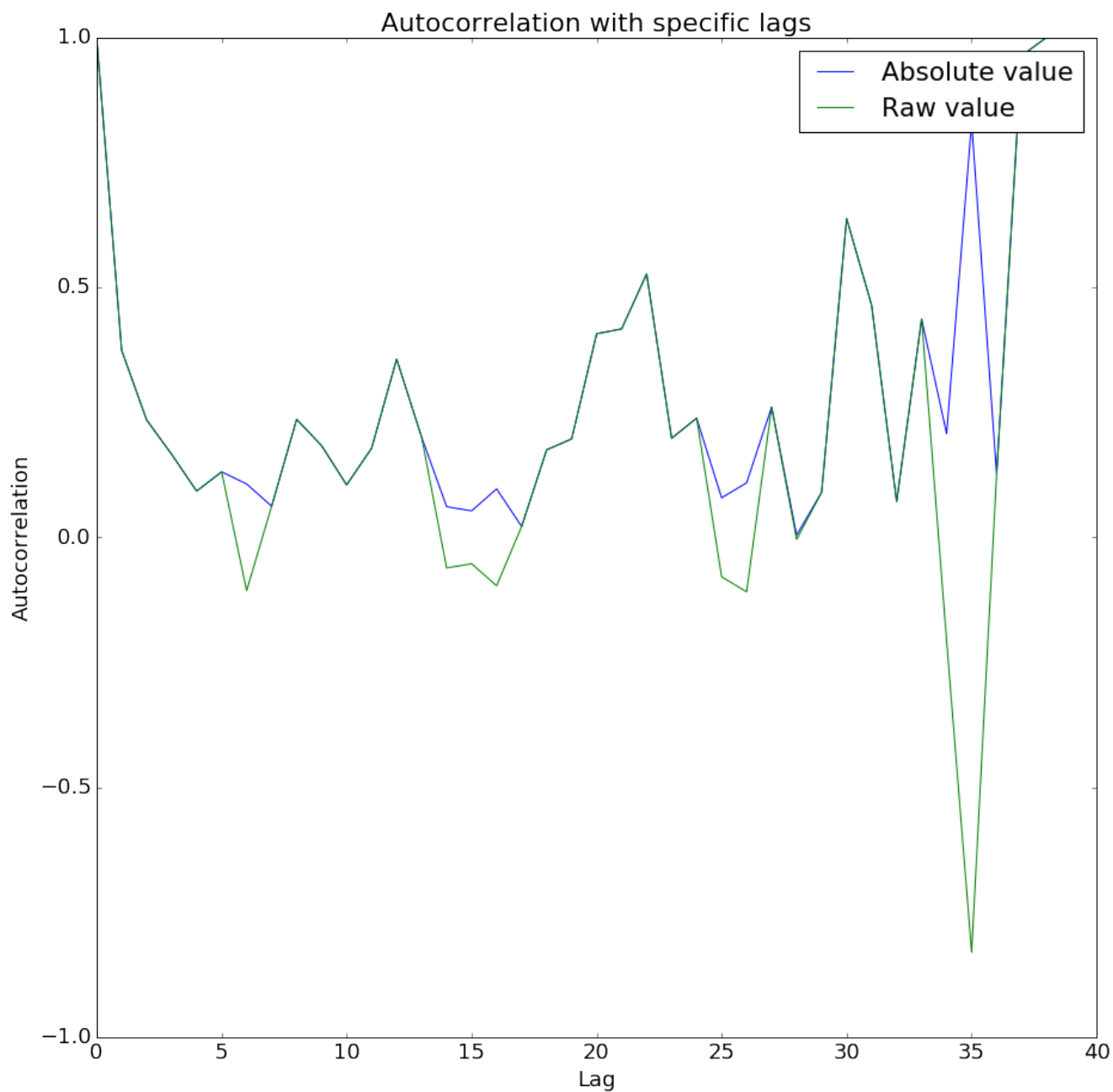


Рисунок 1.10 — Автокореляція

1.4 Прогнозування

Наявна лінія тренду може бути спрогнозована за формулою

$$Y_{t+L} = \alpha_{0t} + \alpha_{1t} \cdot L + \frac{1}{2} \cdot \alpha_{2t} \cdot L^2,$$

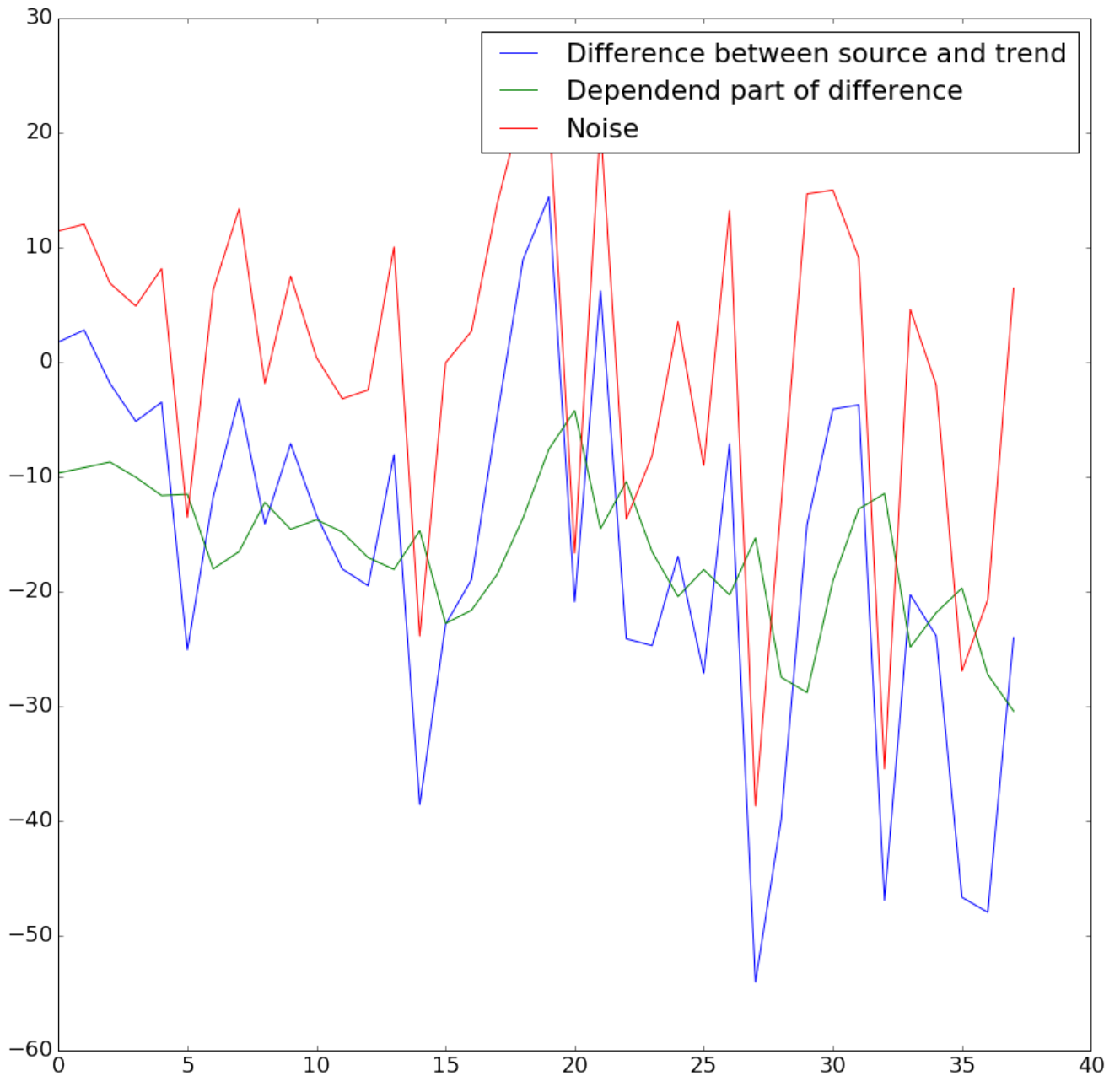


Рисунок 1.11 — Порівняння помилки з оцінкою її не випадкової частини

де

$$\begin{cases} \alpha_{0t} = 3 \cdot (S_t^1 - S_t^2) + S_t^3, \\ \alpha_{1t} = \frac{\alpha}{2 \cdot \beta^2} \cdot [(6 - 5 \cdot \alpha) \cdot S_t^1 - 2 \cdot (5 - 4 \cdot \alpha) \cdot S_t^2 + (4 - 3 \cdot \alpha) \cdot S_t^3], \\ \alpha_{2t} = \frac{\alpha^2}{\beta^2} \cdot (S_t^1 - 2 \cdot S_t^2) + S_t^3 \end{cases}$$

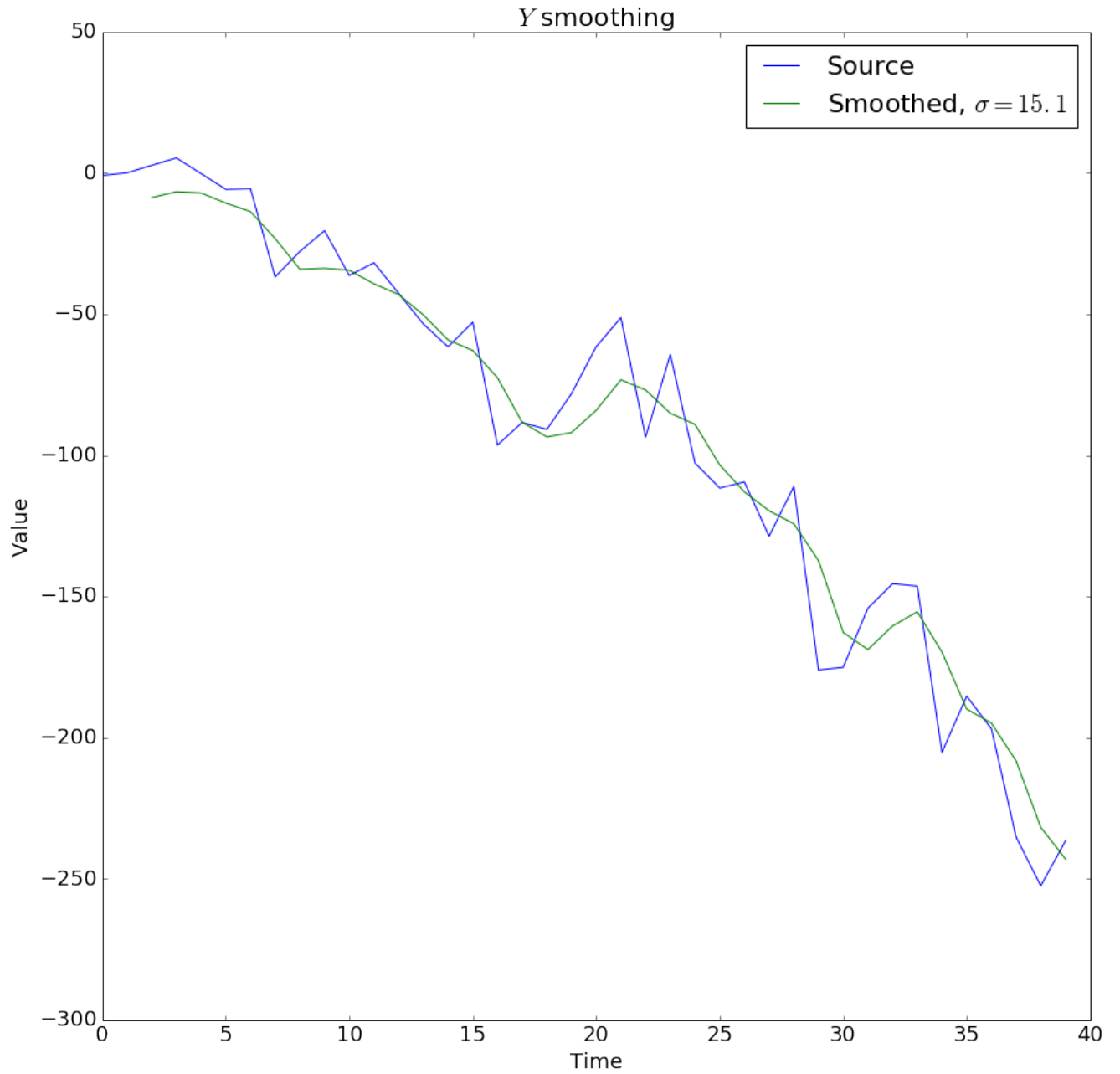


Рисунок 1.12 — Апроксимація з урахуванням не випадкових помилок

i

$$\begin{cases} S_t^1 = \alpha \cdot Y_t + \beta \cdot S_{t-1}^1 = \alpha \cdot \sum_{i=0}^{t-1} \beta^i \cdot Y_{t-i} + \beta^t \cdot Y_0, \\ S_t^2 = \alpha \cdot S_t^1 + \beta \cdot S_{t-1}^2, \\ S_t^3 = \alpha \cdot S_t^2 + \beta \cdot S_{t-1}^3. \end{cases}$$

Проте нам вистачить лінійного прогнозу, тому член формули з квадратом відкинемо

$$Y_{t+L} = \alpha_{0t} + \alpha_{1t} \cdot L.$$

Щоб приблизно визначити якість прогнозу, можемо побудувати його для вже існуючих даних та порівняти (рис. 1.13). Прогноз на три кроки після остан-

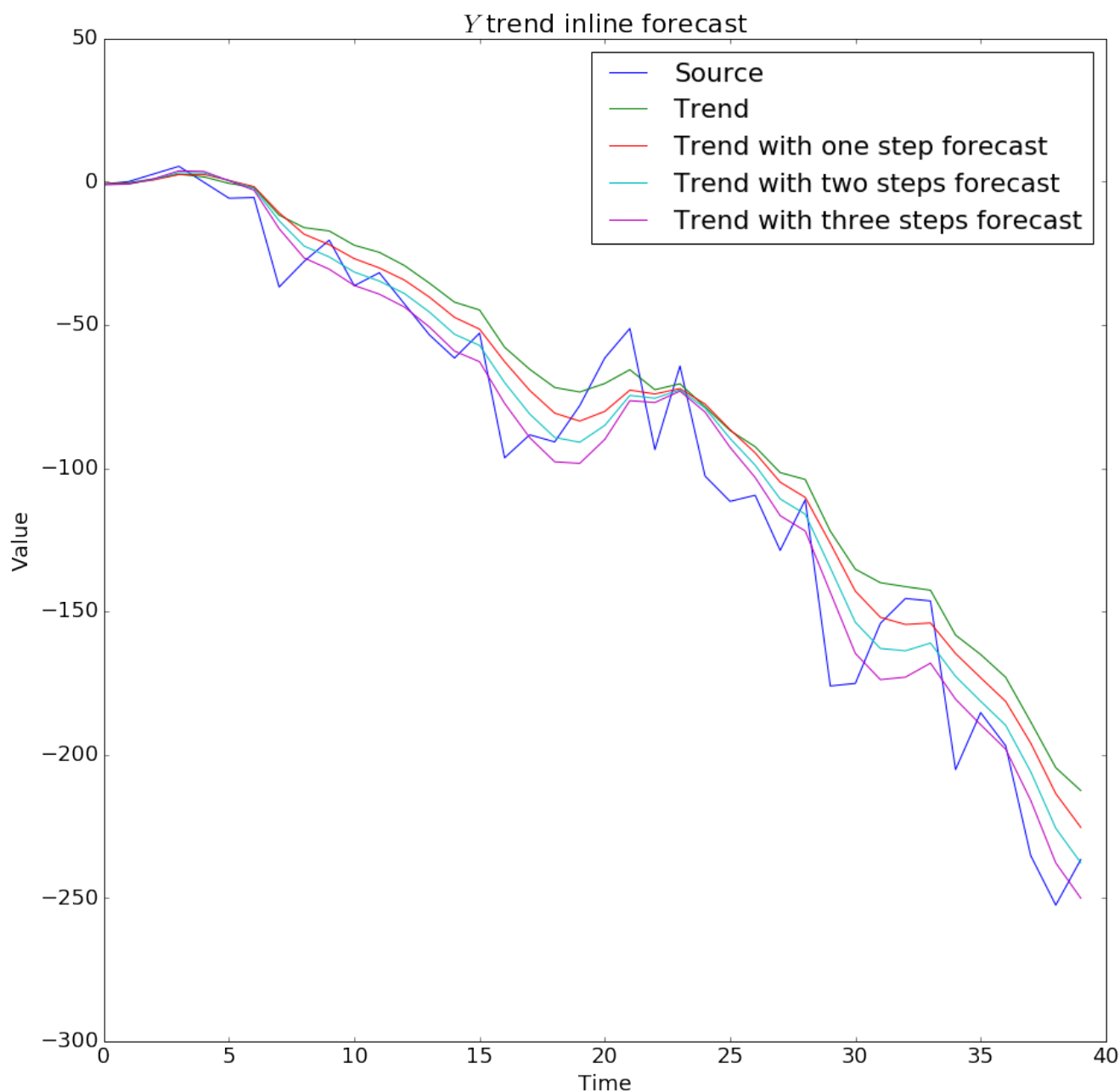


Рисунок 1.13 — Прогнозування ретроданих

ньої точки має вигляд (рис. 1.14).

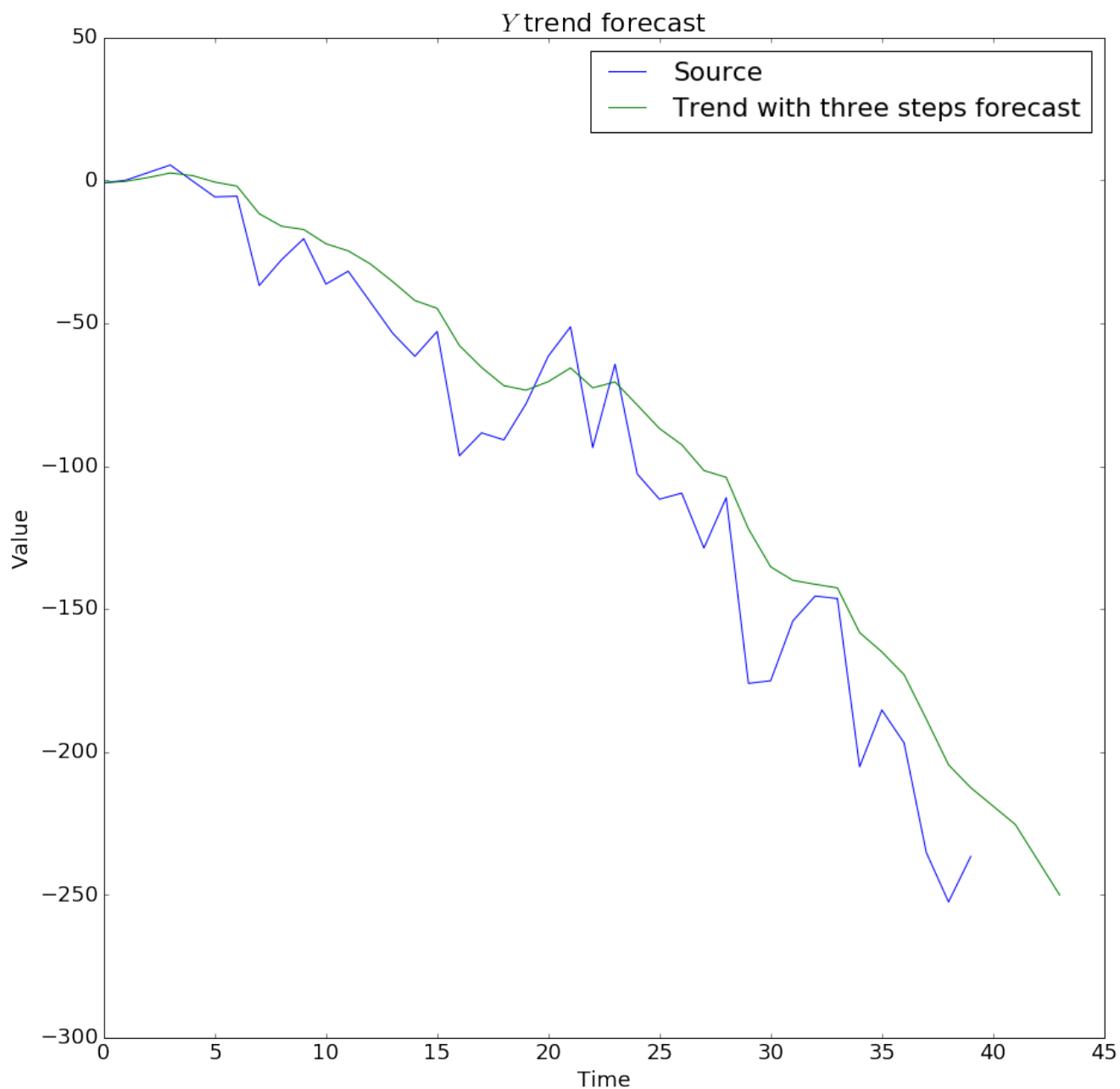


Рисунок 1.14 — Прогноз

1.5 Прогнозування не випадкових помилок

Ми маємо лінійну формулу для розрахунку залежності між не випадковими помилками

$$\delta_{t+2} = a_0 + \delta_t \cdot a_1 + \delta_{t+1} \cdot a_2.$$

Перевіримо якість прогнозу у той самий спосіб — порівняємо його з наявними даними (рис. 1.15).

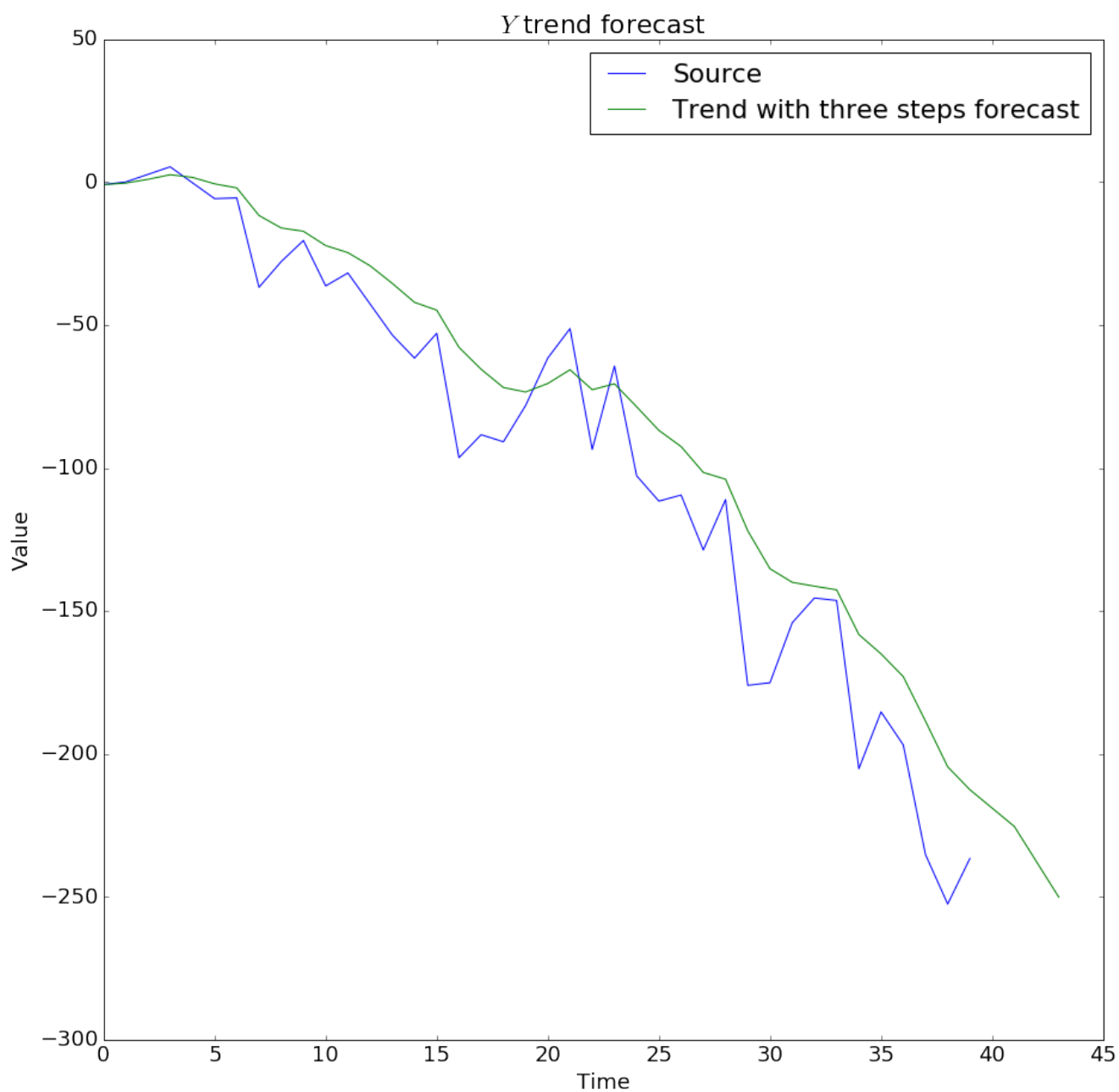


Рисунок 1.15 — Прогнозування ретроданих з урахуванням невинпадкових помилок

1.6 Прогноз процесу

Кінцевий прогноз виглядає наступним чином (рис. 1.16).

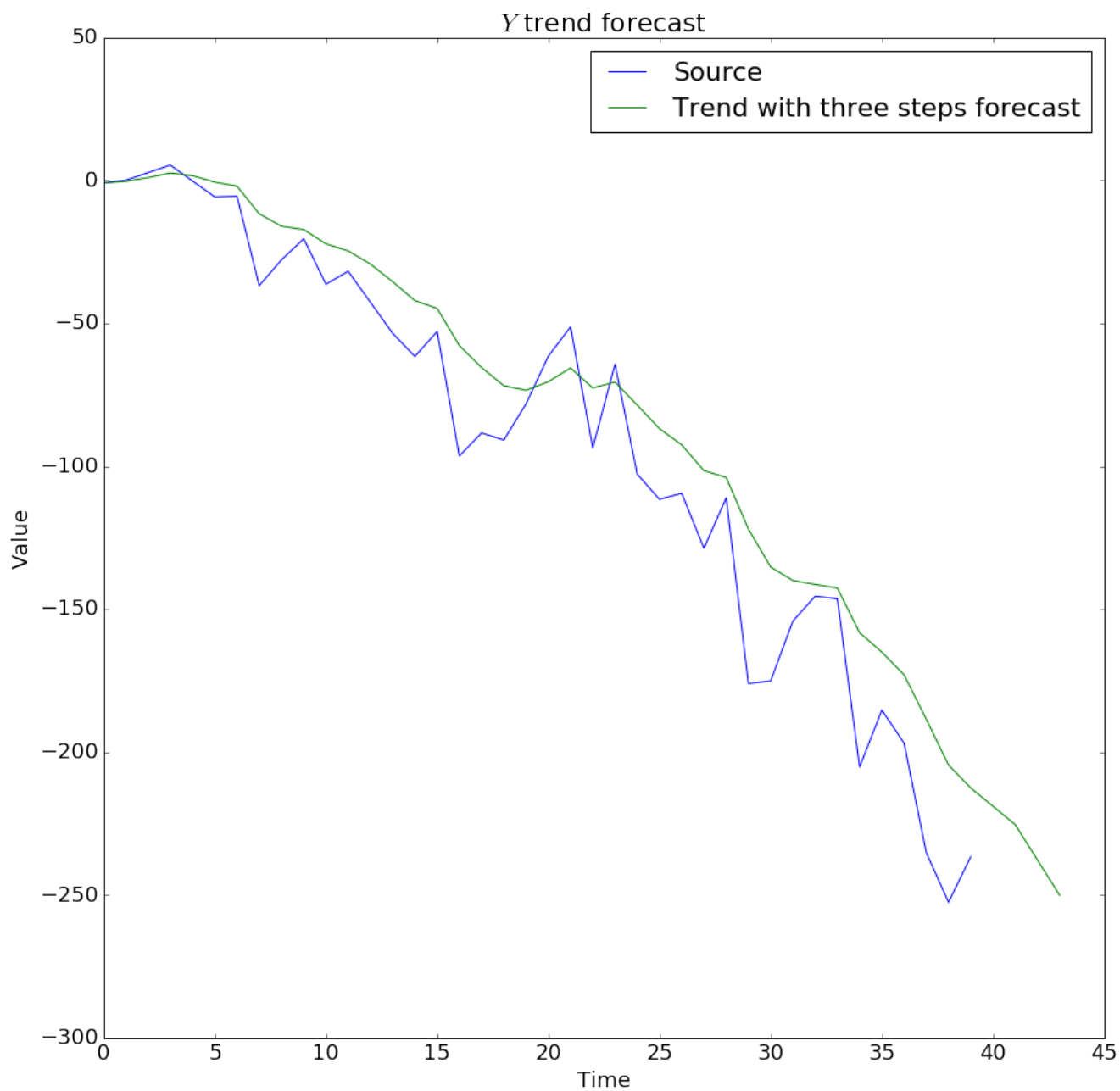


Рисунок 1.16 — Прогноз з урахуванням не випадкових помилок

2 ПРОГНОЗУВАННЯ ЗА МОДЕЛЛЮ ЗАЛЕЖНОСТІ

Маємо набір випадкових процесів $X_i, i = 1..6$, що не містять аномальних точок, проте містять помилки. Потрібно побудувати їх прогноз, модель залежності Y від цих процесів, та на основі цього створити прогноз для Y .

Зобразимо дані процеси графічно на рис. 2.1.

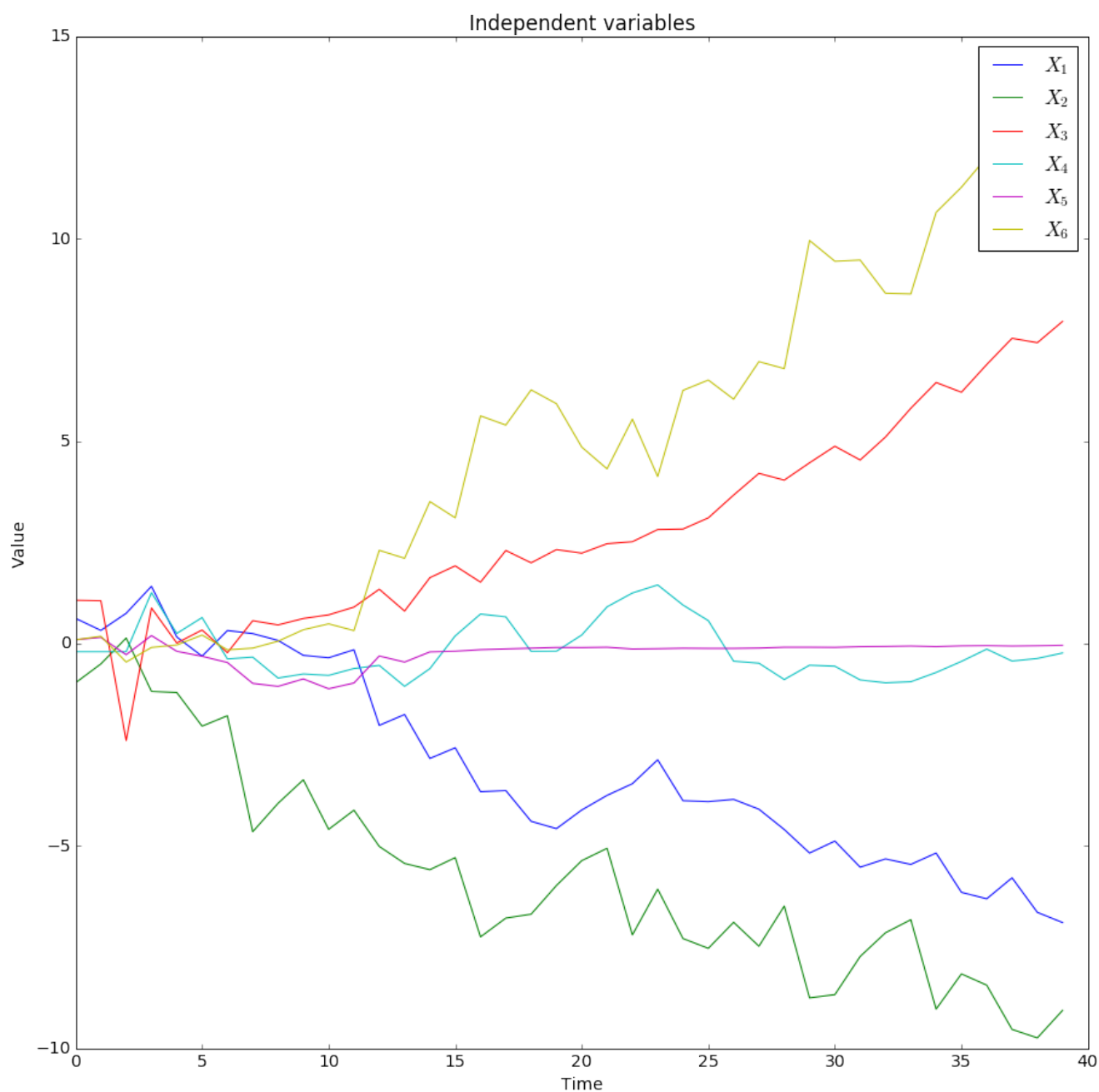


Рисунок 2.1 — Графіки процесів X

2.1 Згладжування

В даному розділі не будемо детально зупинятися на тому, що вже було розглянуто при аналізі Y .

Лінії тренду для всіх X знайдені за допомогою експоненційно зваженого ковзкого середнього (рис. 2.2).

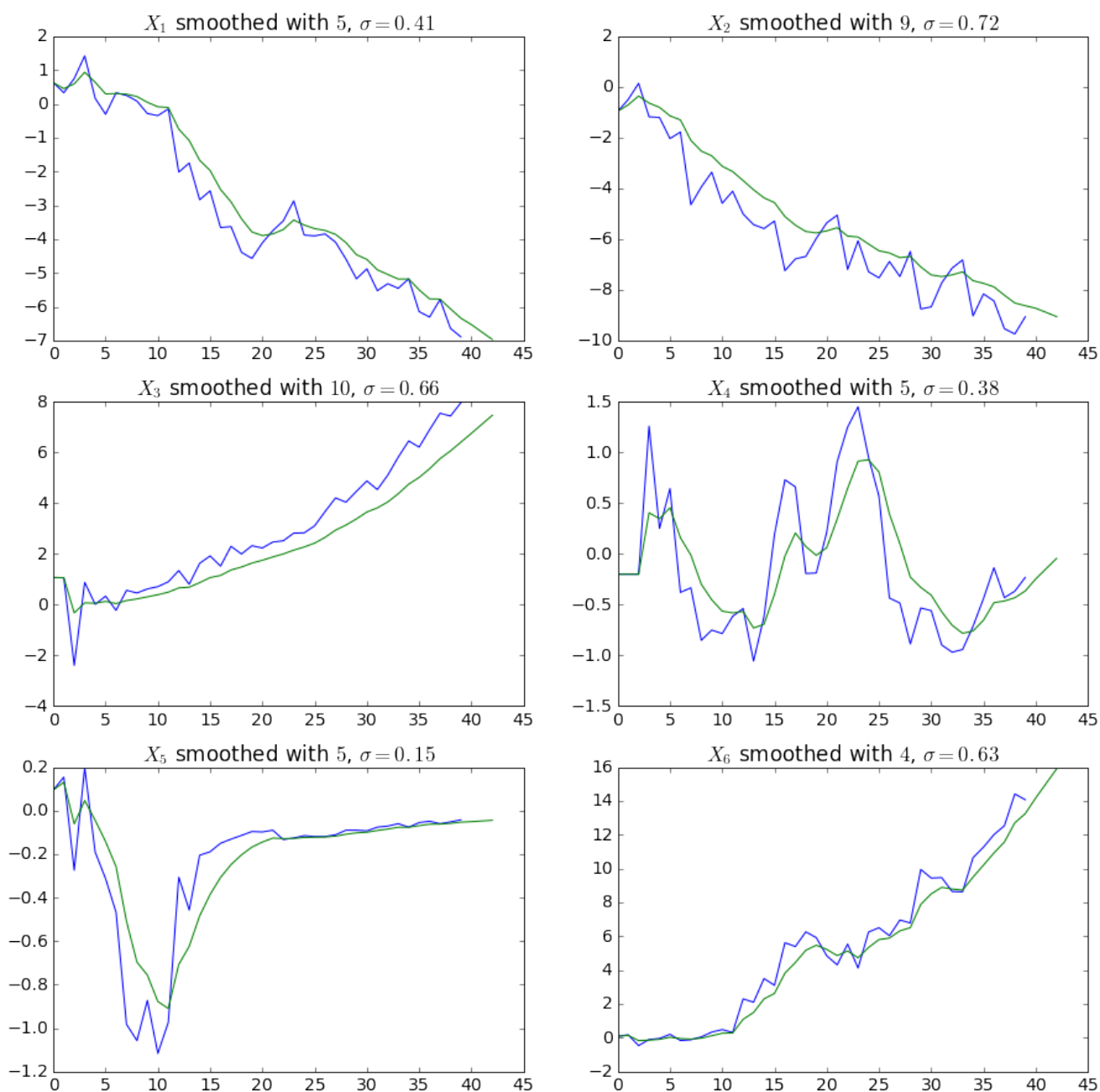


Рисунок 2.2 — Графіки процесів X та їх трендів

Графіки автокореляцій показали, що достатньо обрати два минулі члені послідовності, щоб зробити прогноз для наступного. Щоправда, X_3 має дуже багато залежностей, це може означати, що кожне наступне значення залежить від попереднього і ми маємо Марковський ланцюг (рис. 2.3).

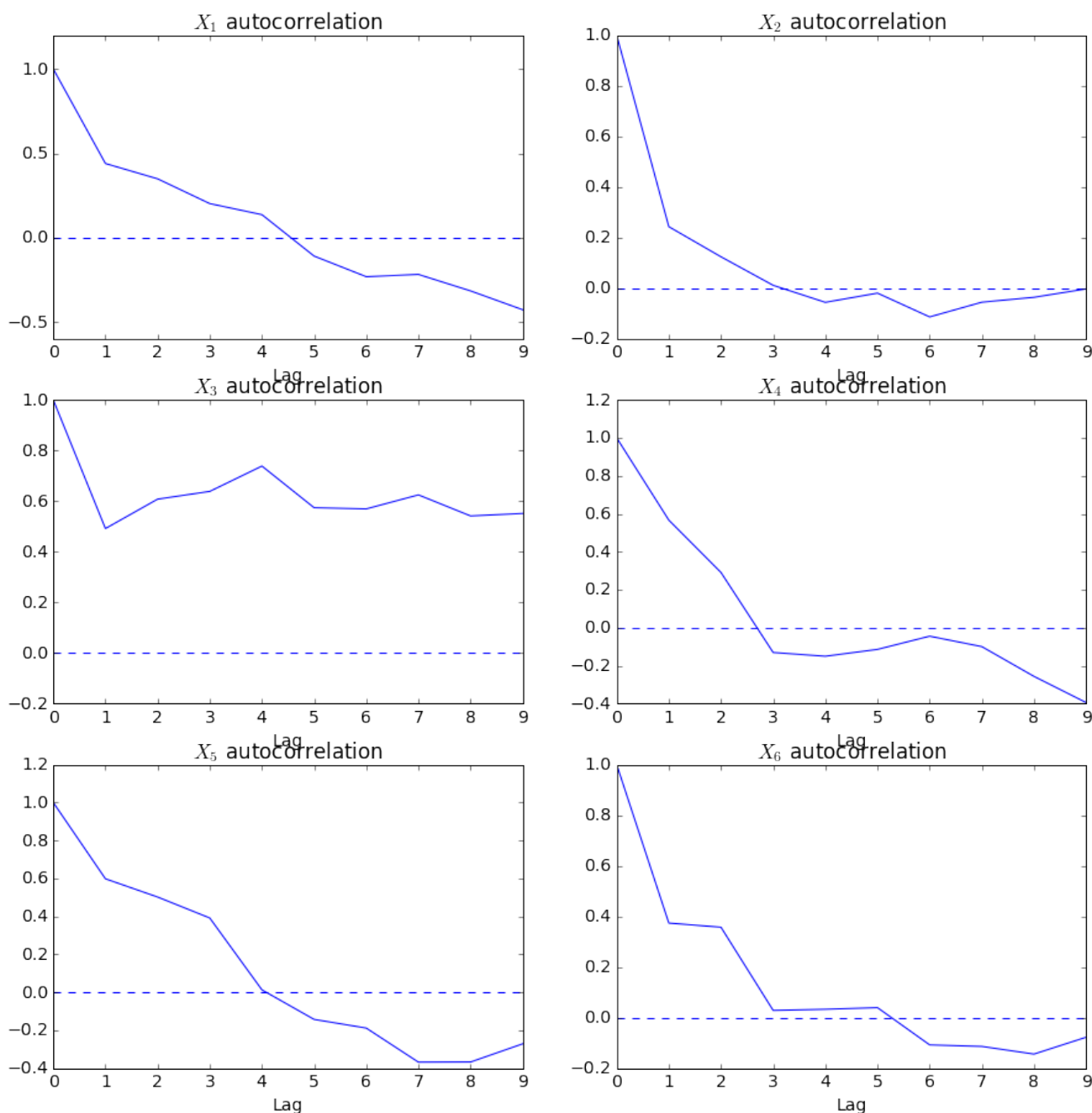


Рисунок 2.3 — Графіки автокореляцій процесів X за різних лагів τ

Фільтр, що враховує не випадкові помилки, дуже добре відповідає дійсним

даним (рис. 2.4.

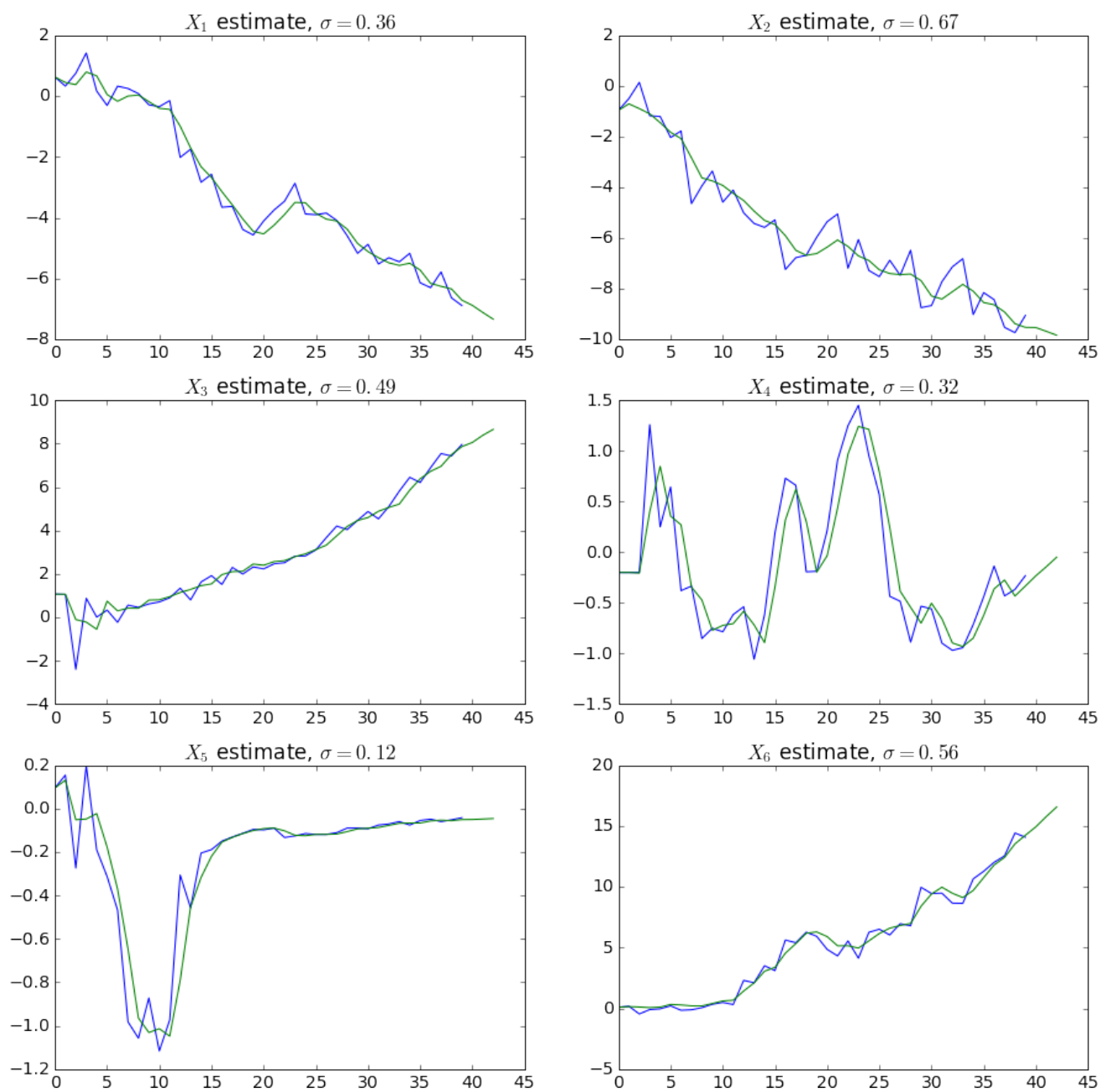
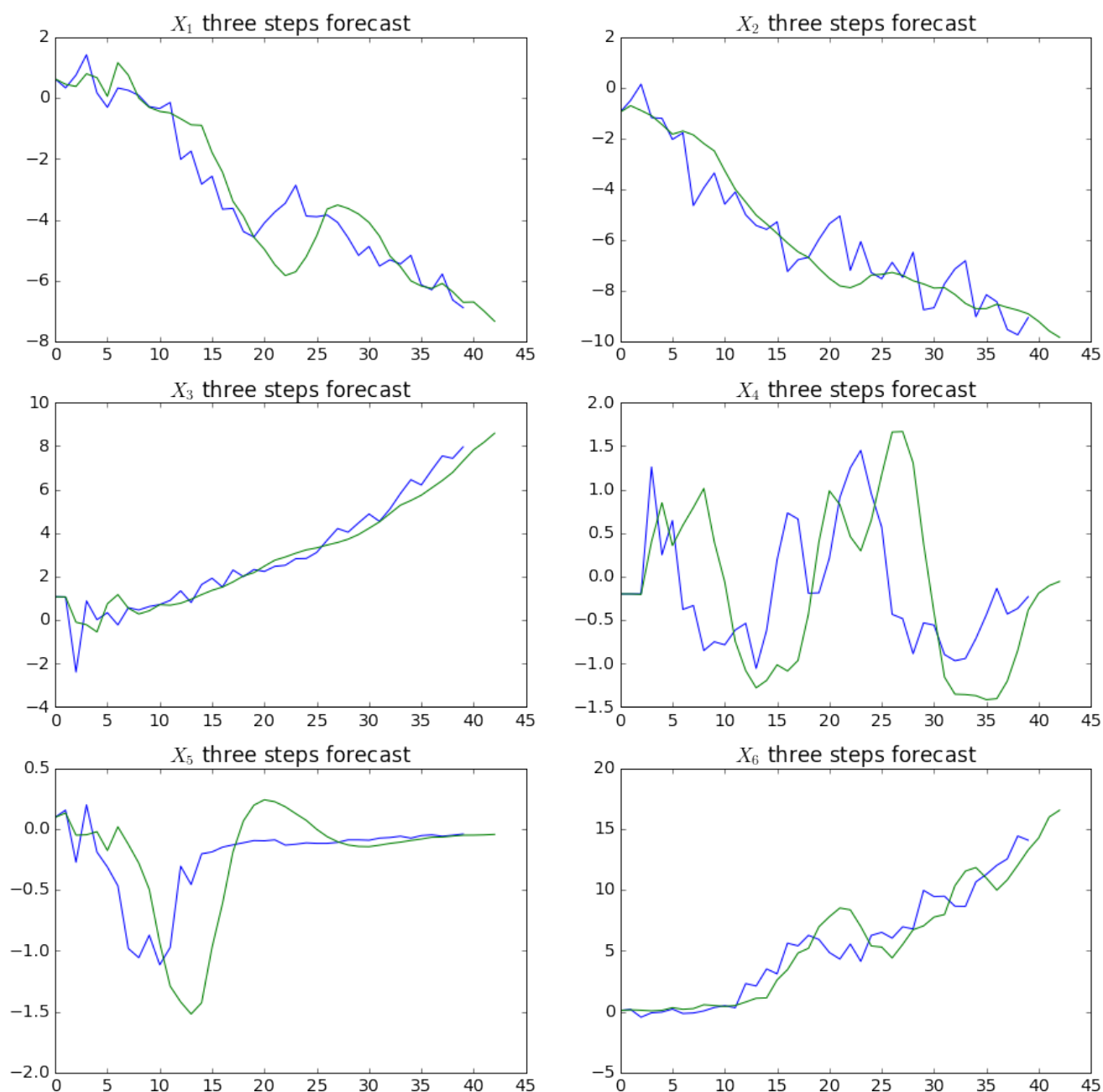


Рисунок 2.4 — Оцінки X з урахуванням не випадкових помилок

Рисунок 2.5 — Прогнози X

2.2 Прогноз часових рядів

2.3 Побудова моделі

Для того, щоб знайти залежність між Y та X , було використано не тільки X_i , але і їх попарні поелементні добутки $X_i \cdot X_j$. Ці значення називаємо регре-

сорами.

Ітеративний метод побудови моделі полягає в наступному:

- 1) першим регресором є ряд, що складається з одиниць (очевидно, що перший коефіцієнт — середнє значення Y);
- 2) з невикористаних регресорів обирається той, що найбільше корелює з залишками — різницею між вхідними даними Y та поточною моделлю \tilde{Y} ;
- 3) методом найменших квадратів перераховуються коефіцієнти, що множаться на обрані регресори;
- 4) за допомогою F -тесту перевіряється, чи дав цей новий регресор значний приріст якості;
- 5) якщо якість зросла і залишилися ще невикористані регресори, обираємо наступний регресор;
- 6) якщо ні, маємо модель.

Було отримано модель з регресорами ∞ , $X_2 \cdot X_6$, $X_3 \cdot X_5$, $X_4 \cdot X_5$ та коефіцієнтами -3.33 , 1.75 , 15.6 , 22.67 . Маємо графік з вхідним рядом Y та його прогнозом, що було побудовано на основі прогнозів X (рис. 2.6).

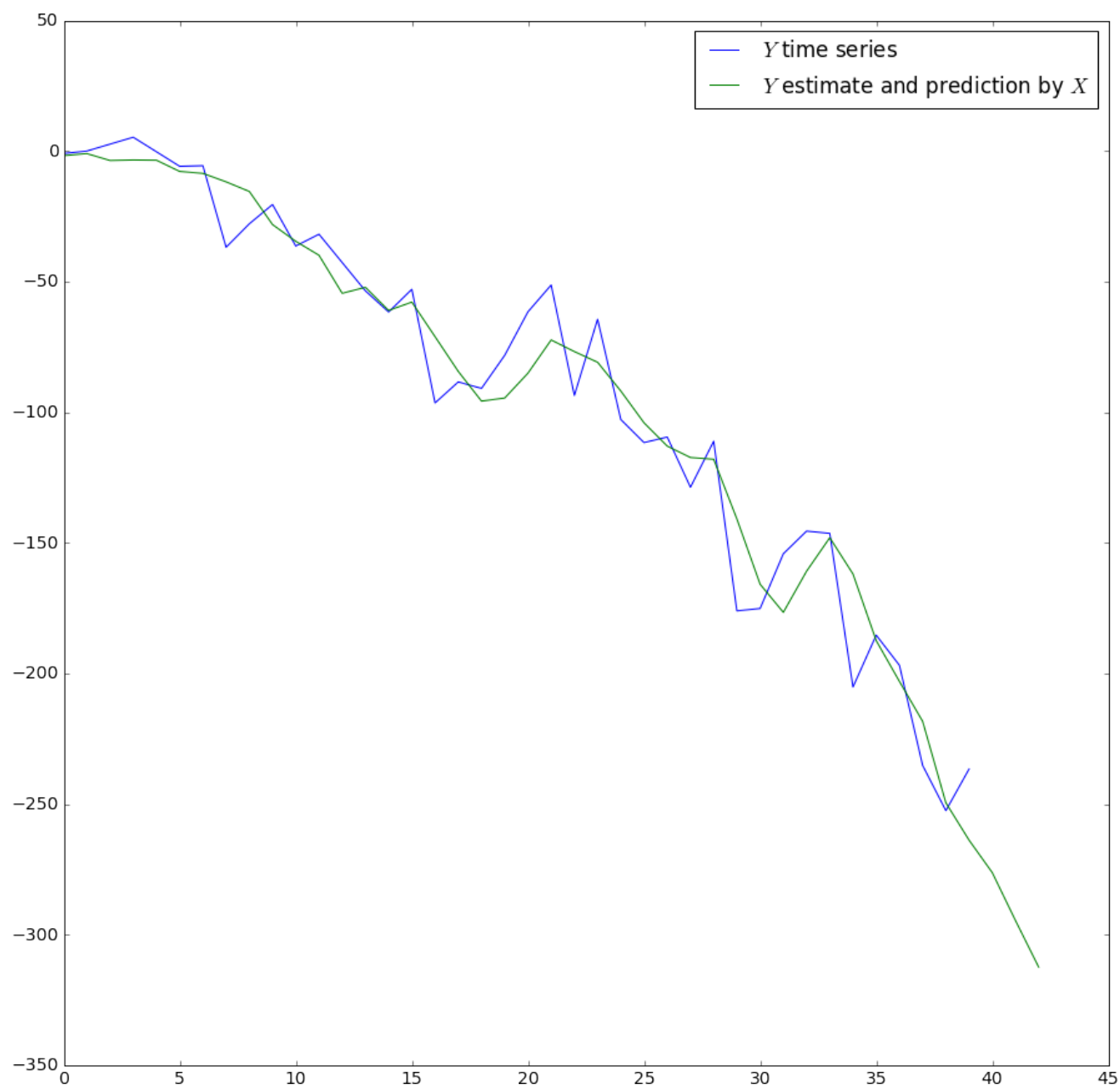


Рисунок 2.6 — Прогнози Y як моделі X

ВИСНОВКИ

В результаті виконання роботи вдалося

- 1) визначити оцінку (згладжування) часових рядів;
- 2) зробити прогноз часових рядів на основі ретроданих;
- 3) зробити прогноз часового ряду на основі моделі залежності від незалежних рядів.

Хоча обидва методи показали себе добре, проте прогноз на основі ретро-даних краще. Це було визначено за допомогою порівняння середньоквадратичного відхилення помилок прогнозу для існуючих даних. Прогноз за ретроданими дав помилку $\sigma = 24.06$, а прогноз за прогнозами моделі $\sigma = 25.91$ (рис. 2.7). Це могло статися через ряд причин:

- 1) кожен регресор мав свою похибку при прогнозі, потім ця похибка збільшилася при оцінці значення Y , коли оцінка Y за ретроданими має лише свою похибку прогнозу;
- 2) регресійна модель знайшла залежності, яких немає, але які допомагають покращити якість фільтрації, що також відомо як занадто точна регресія (overfitting); [3]
- 3) обрано некоректні регресори;
- 4) дані методи не підходять для цих даних.

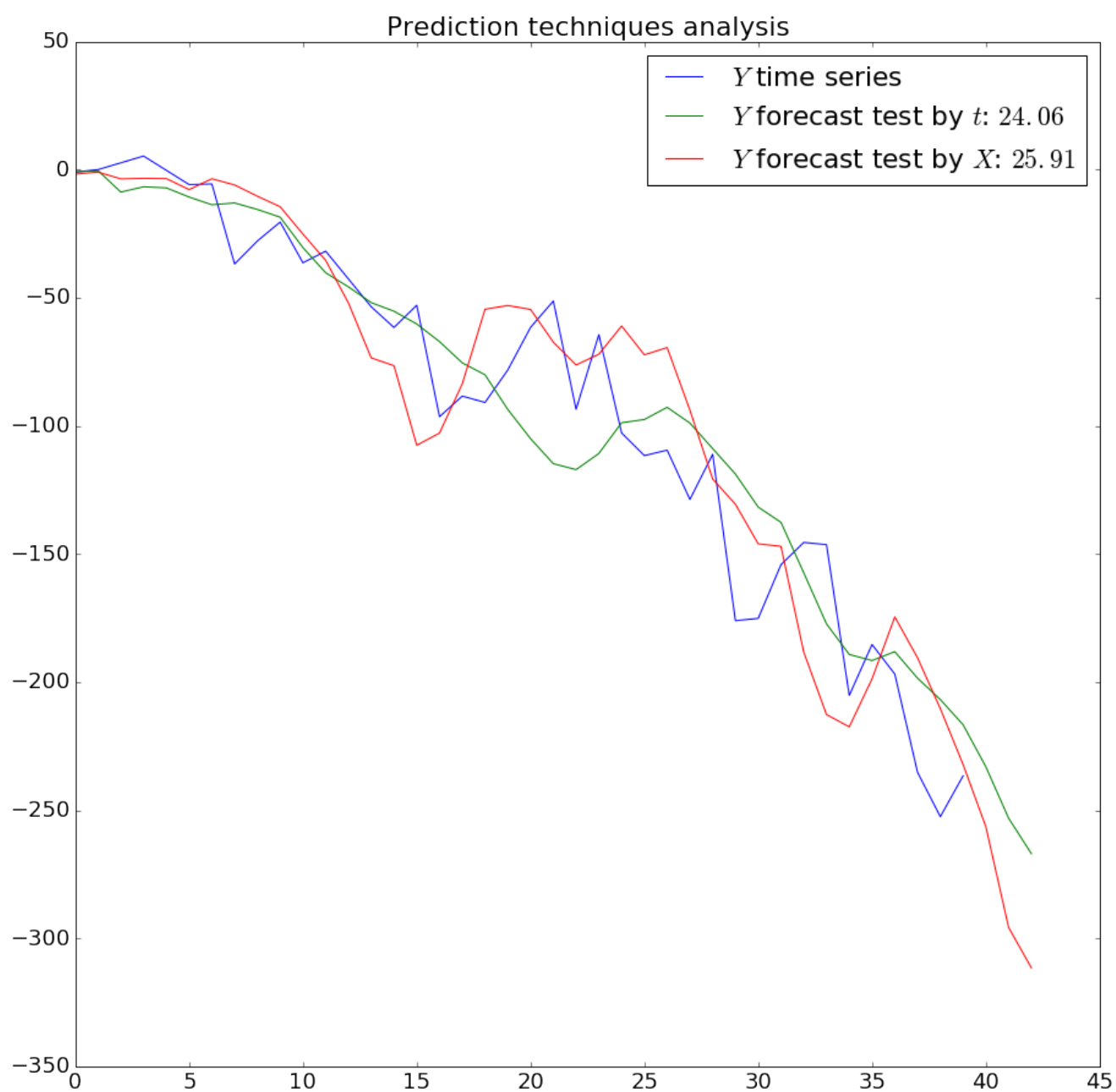


Рисунок 2.7 — Порівняння прогнозів Y за ретроданими та моделлю X

ПЕРЕЛІК ПОСИЛАНЬ

1. Lomax, R.G. Statistical Concepts: A Second Course / R.G. Lomax. — Lawrence Erlbaum Associates, 2007.
2. D'Agostino, Ralph B. A Suggestion for Using Powerful and Informative Tests of Normality / Ralph B. D'Agostino, Albert Belanger, Jr. D'Agostino // *The American Statistician*. — 1990. — Vol. 44, no. 4. — Pp. 316–321.
3. Hawkins, Douglas M. The Problem of Overfitting / Douglas M. Hawkins // *Journal of Chemical Information and Computer Sciences*. — 2004. — 1. — Vol. 44, no. 1. — Pp. 1–12.