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

Перевірив

д.т.н., професор, с.н.с.,

Архипов Олександр Євгенович

Київ — 2016 року

ЗМІСТ

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 зглажування для побудови тренду часових рядів та методу авторегресії для оцінки невипадкових помилок.

1 ПРОГНОЗУВАННЯ ЧАСОВИХ РЯДІВ

На вході дано часовий ряд Y з шумом E

$$Y = Z + E,$$

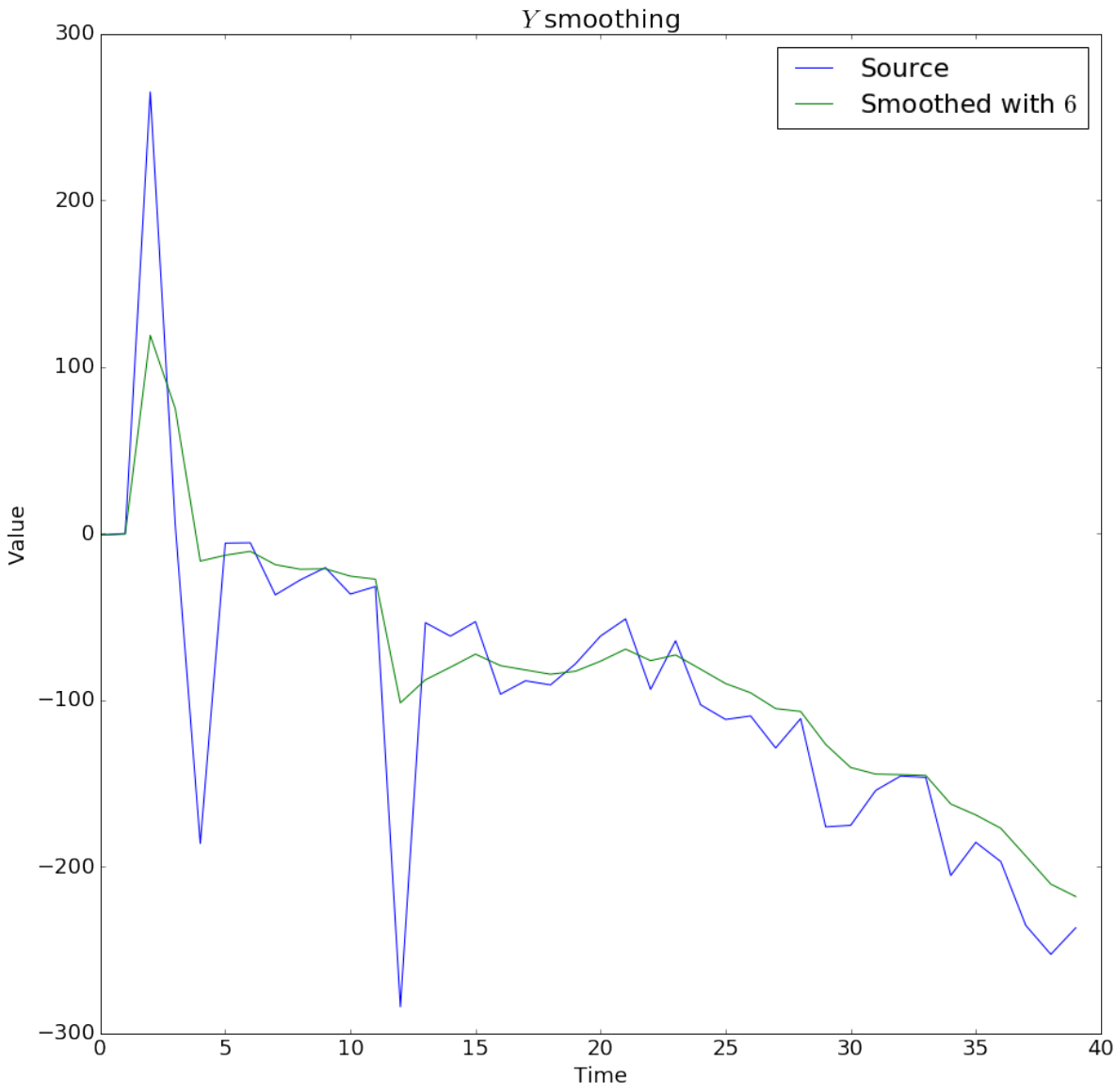
де

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

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

1.1 Видалення аномалій

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



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 зглажування для Y з якістю зглажування його виправленого аналогу. Логічно, що можна порівняти

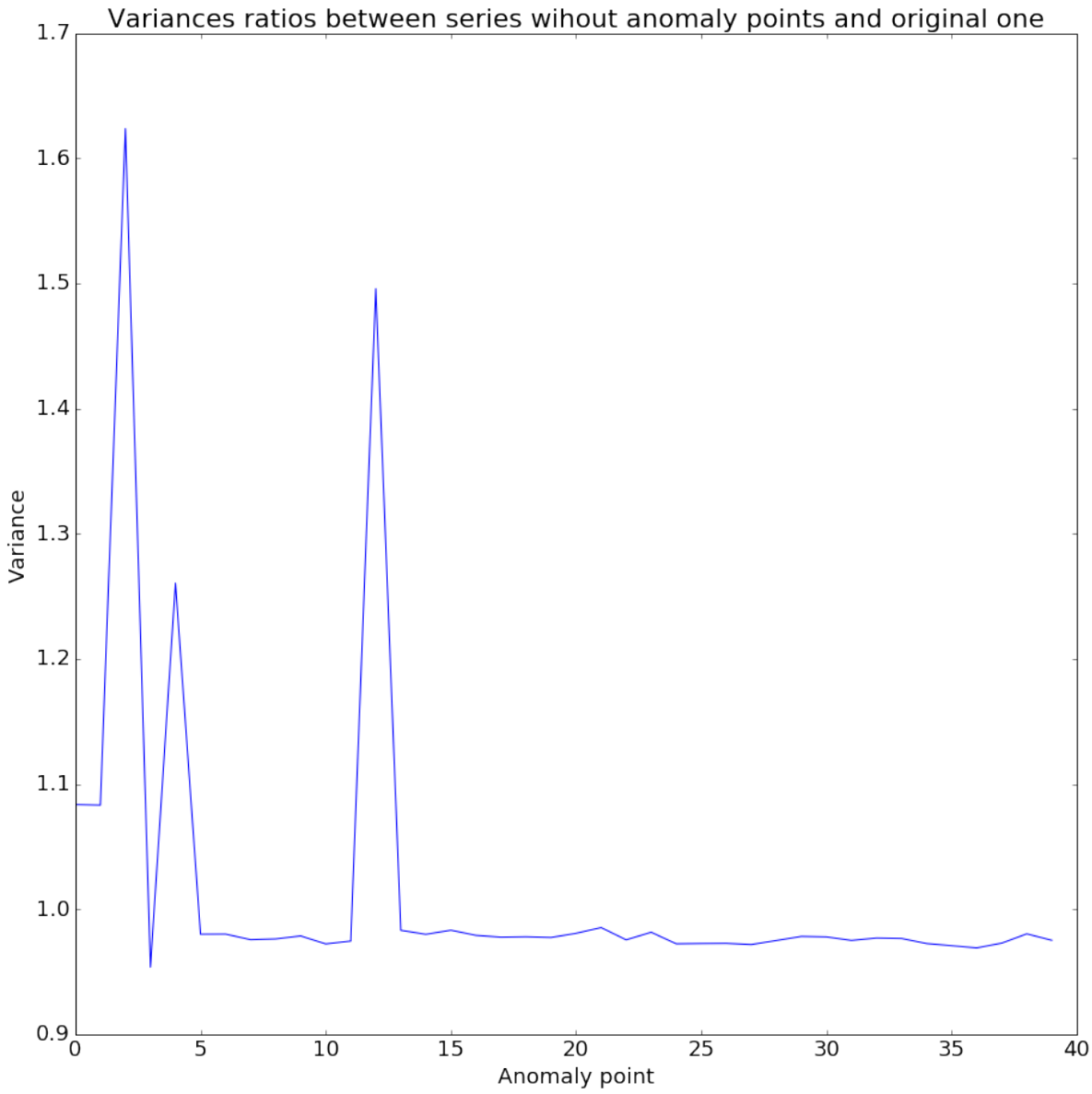
їх співвідношення з певним граничним значенням

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

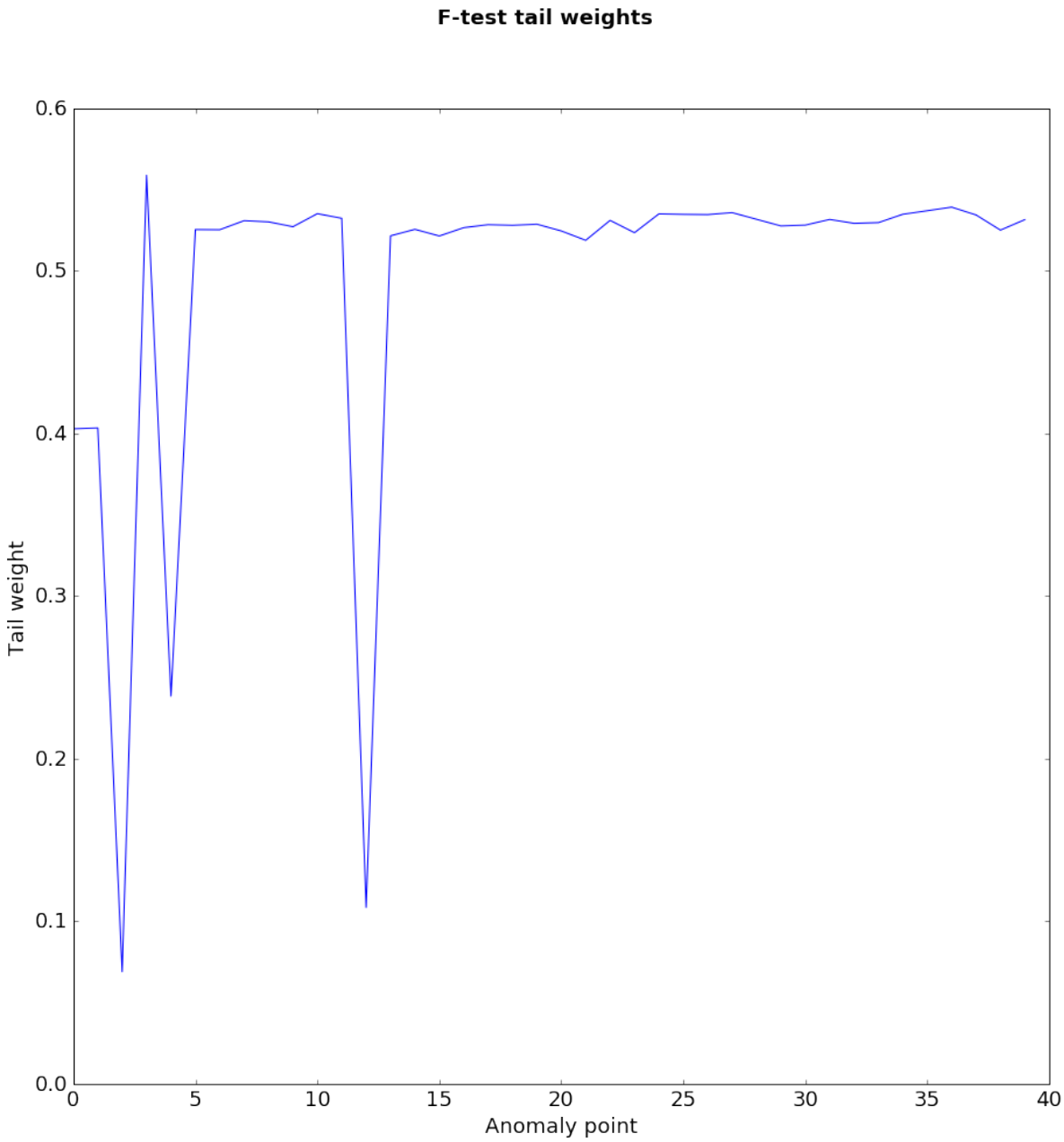
Отримали вираз, що відомий як F-тест [1]: маємо вибіркові дисперсії двох вибірок, що за умовою мають нормальний закон розподілення. Аномальні точки будемо знаходити одну за одною — знаходити найгіршу та виправляти її, якщо вона дійсно аномальна

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

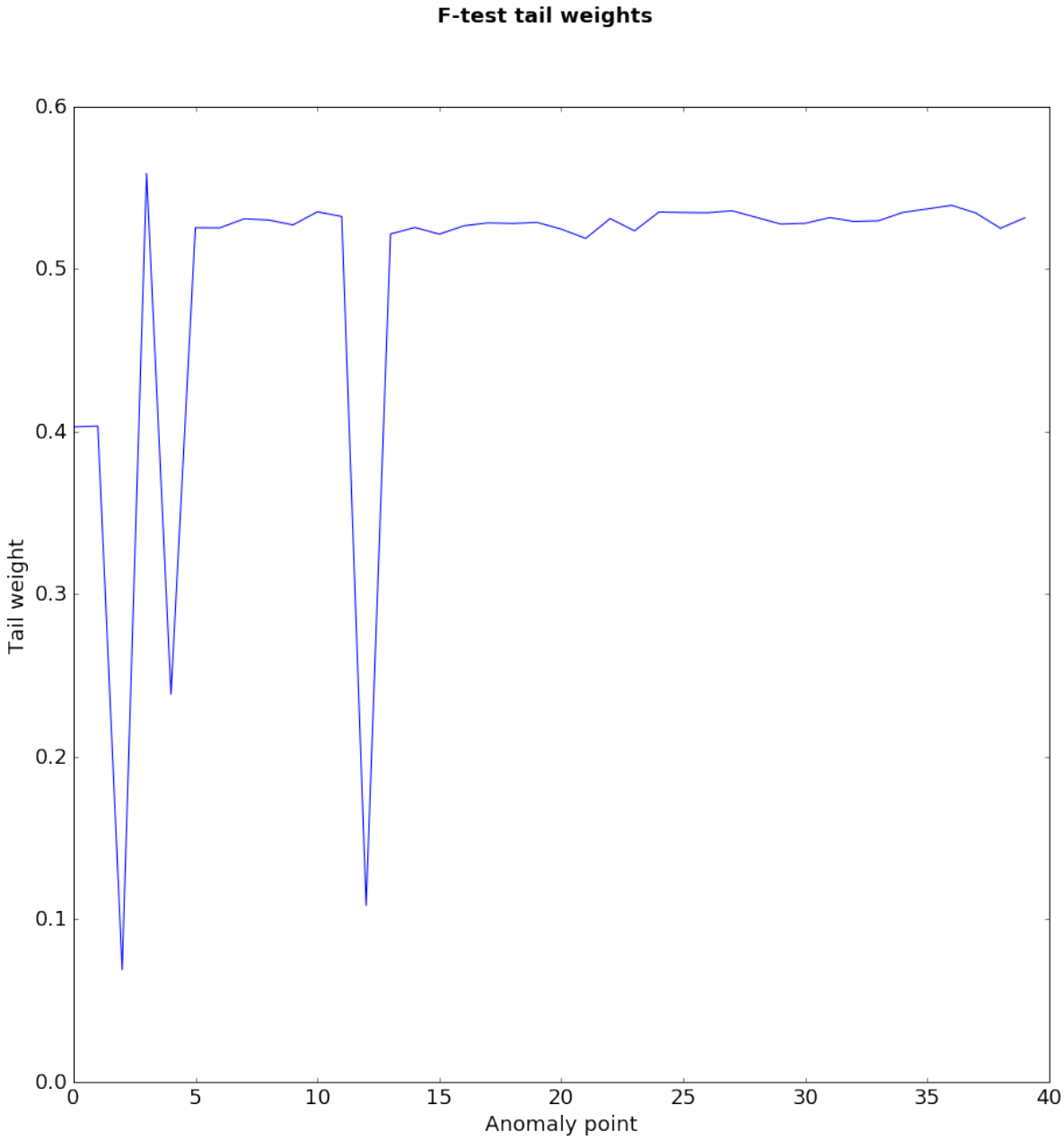
Поглянемо на те, як змінюється вибіркова дисперсія помилки при відкиданні кожної точки.

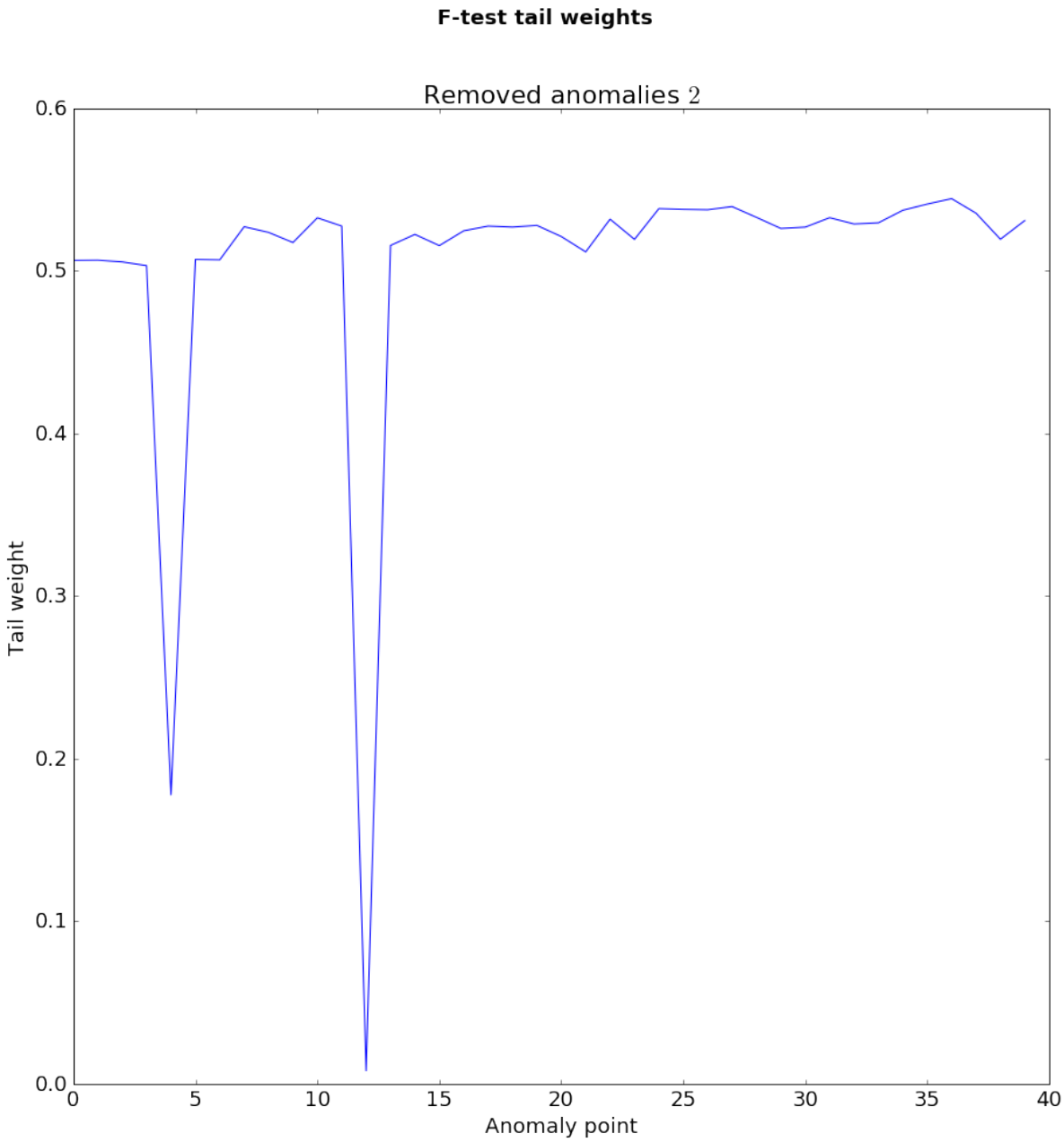


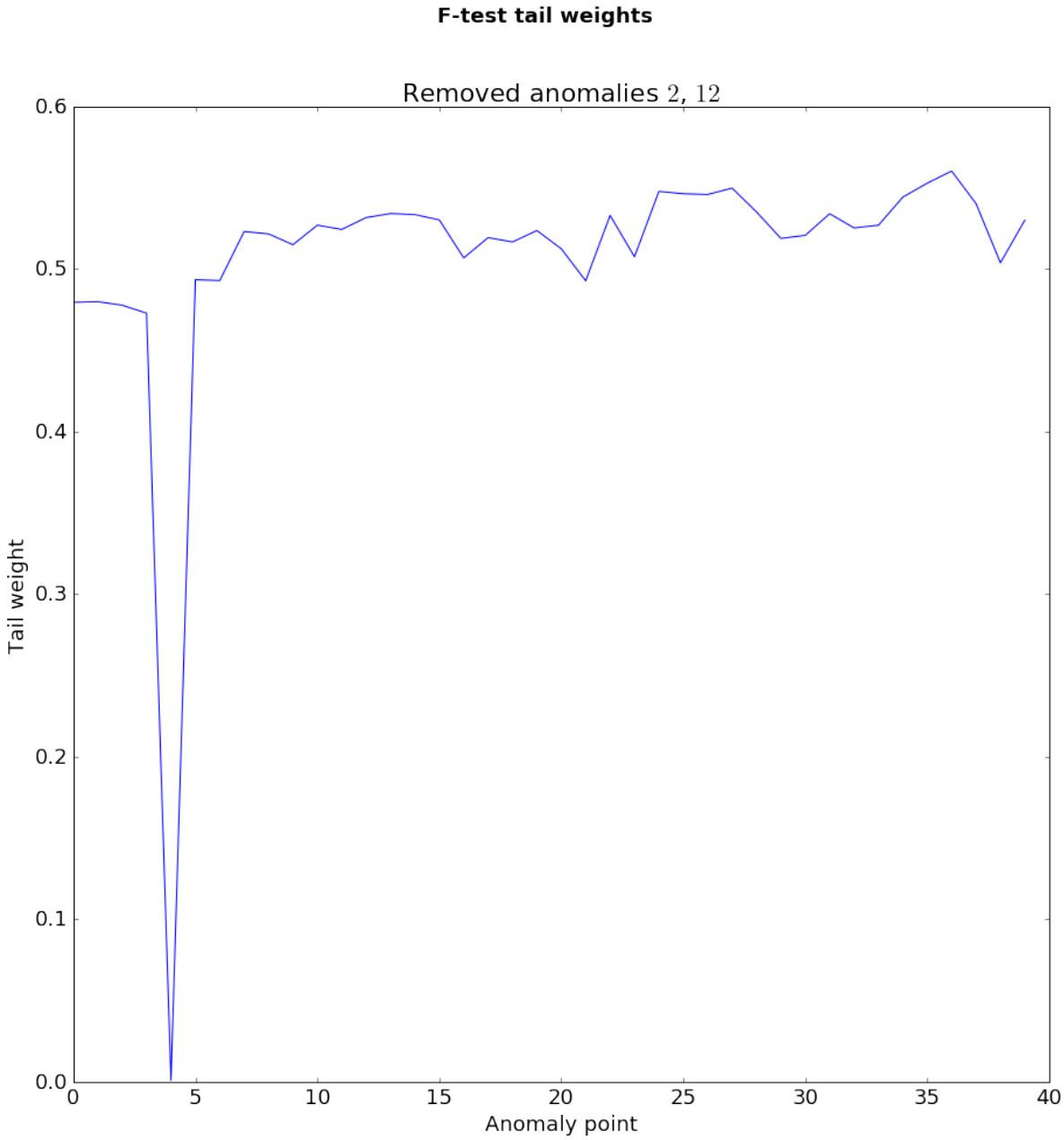
З рисунку видно піки саме в тих місцях, де передбачалась наявність аномальних точок. Проте зараз в нас є міра їх “аномальності” у вигляді ваги хвостів розподілу Фішера.

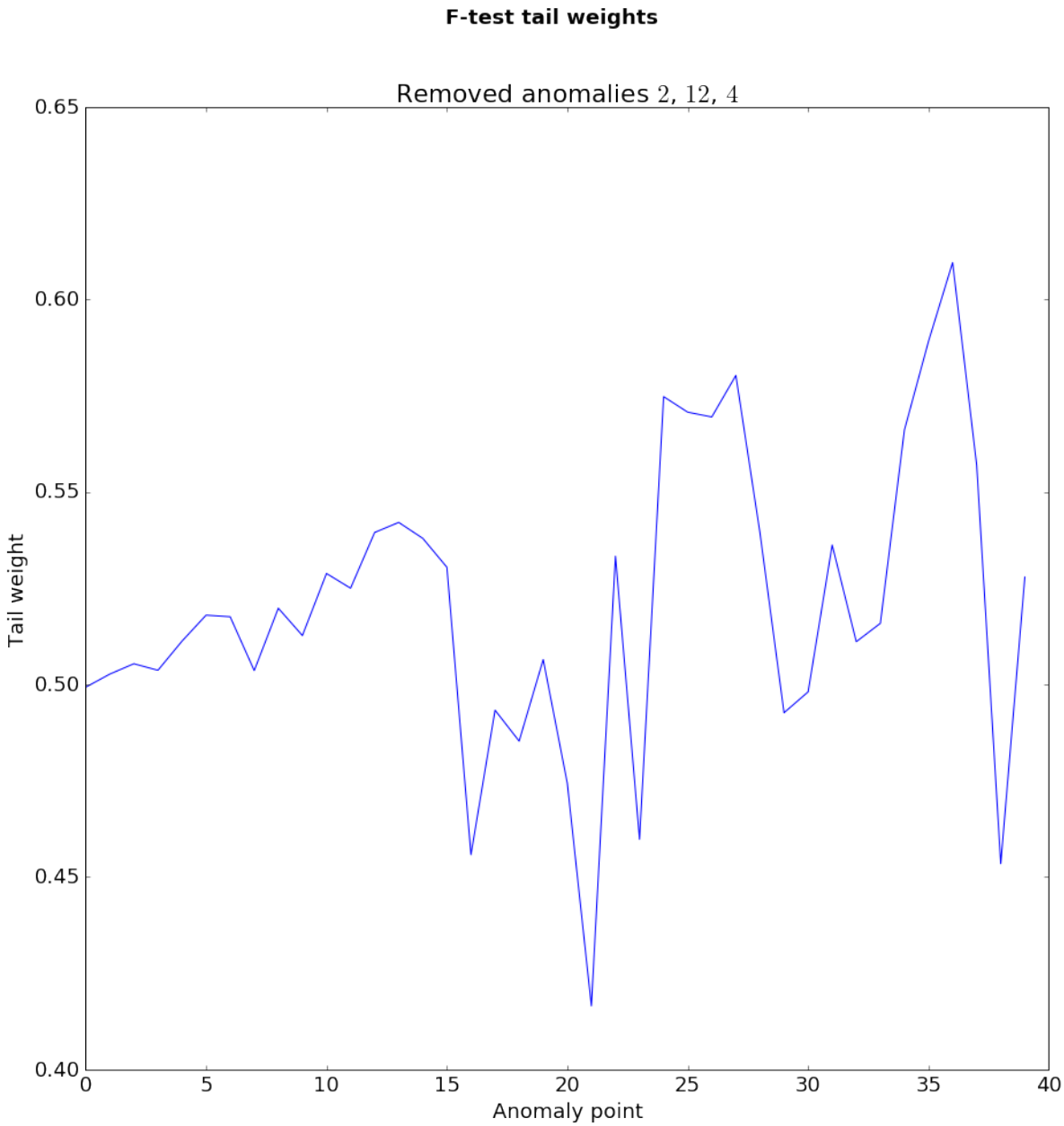


Відкидаємо одну за одною точки, при яких дисперсія значно змінюється, а саме такі, що з ймовірністю 0.9 це дисперсія іншої виборки.

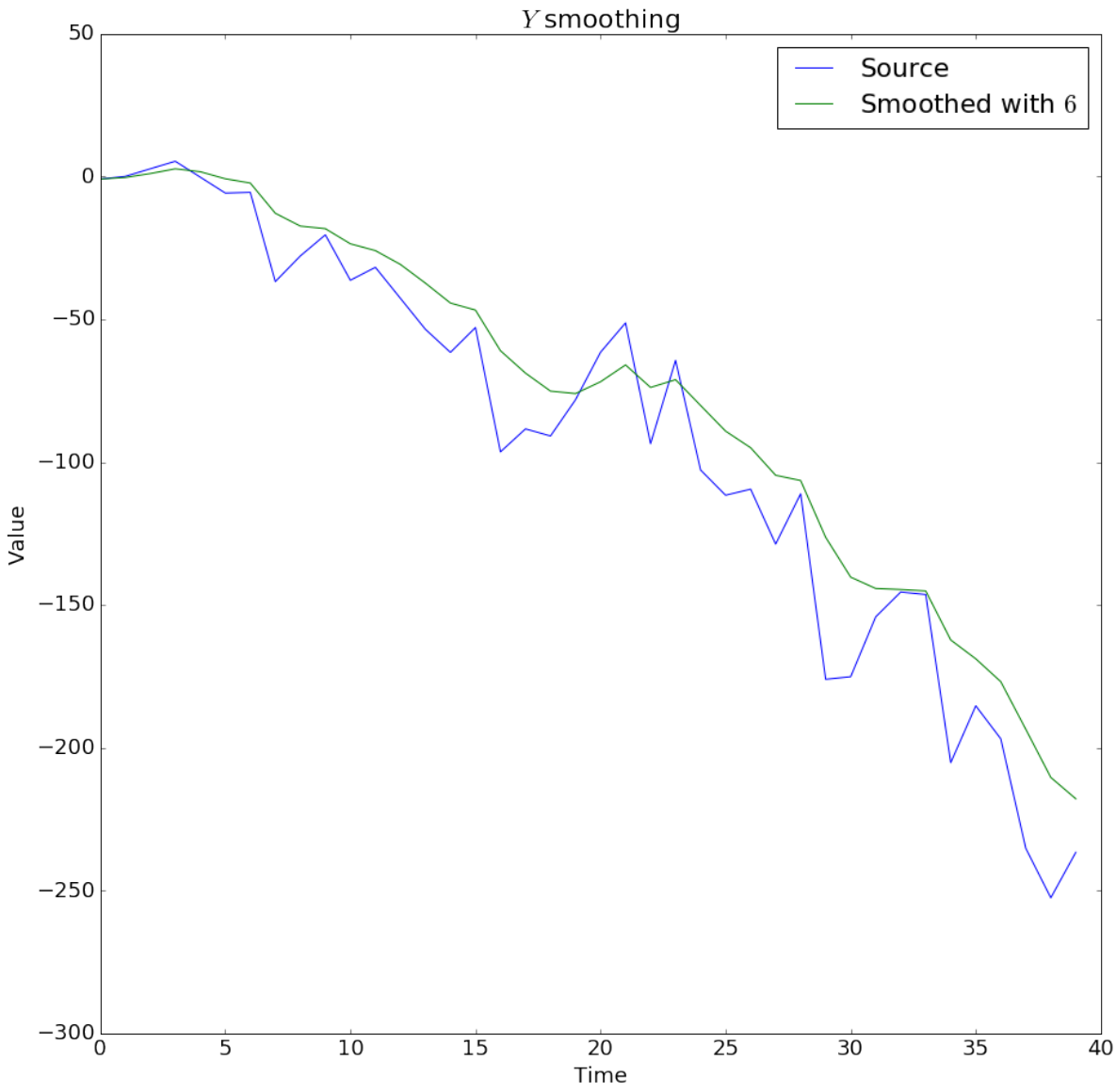

F-test tail weights
Tail weight
Anomaly point
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
5
10
15
20
25
30
35
40


F-test tail weights
Removed anomalies 2
Tail weight
Anomaly point
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
5
10
15
20
25
30
35
40


F-test tail weights
Removed anomalies 2, 12
Tail weight
Anomaly point
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
5
10
15
20
25
30
35
40



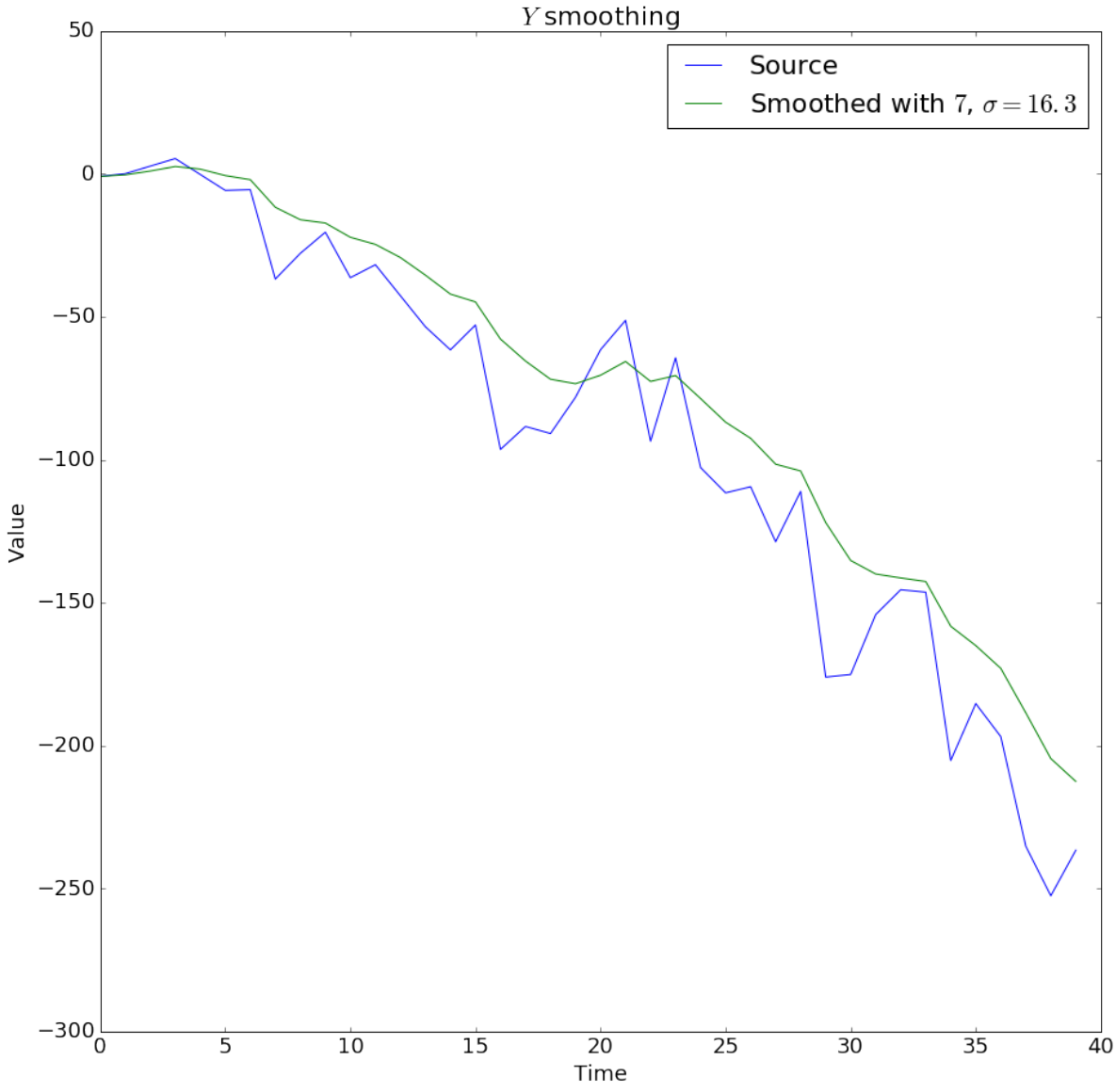
Аномальні дані знаходилися в точках 2, 12 і 4. Їх було виправлено за принципом найближчих сусідів — середнє арифметичне найближчих точок. Маємо графік з даними без аномалій та зглажуванням, яке тепер значно краще описує ряд



1.2 Вибір правильного зглажуючого віконця

Для подальших обчислень потрібно обрати коректні параметри зглажування, а саме параметр α для експоненційно зглаженого 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’Ауґустіно Ральфа [2]. Цей тест на виході дає відстань розподілу даної виборки до класу нормально розподілених виборок.

Ми хочемо обрати таке зглажування, щоб розподіл помилок був якомого більш гаусовим. Потрібно обрати таке α, для якого відстань між розподілом помилок та нормальним розподілом була наймешною. Ця умова винонується для віконця шириною 7, що відповідає $\alpha = \frac{1}{4}$. Маємо похибку $\sigma = 16.3$.



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 $\tilde{Y}$.

Тоді

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

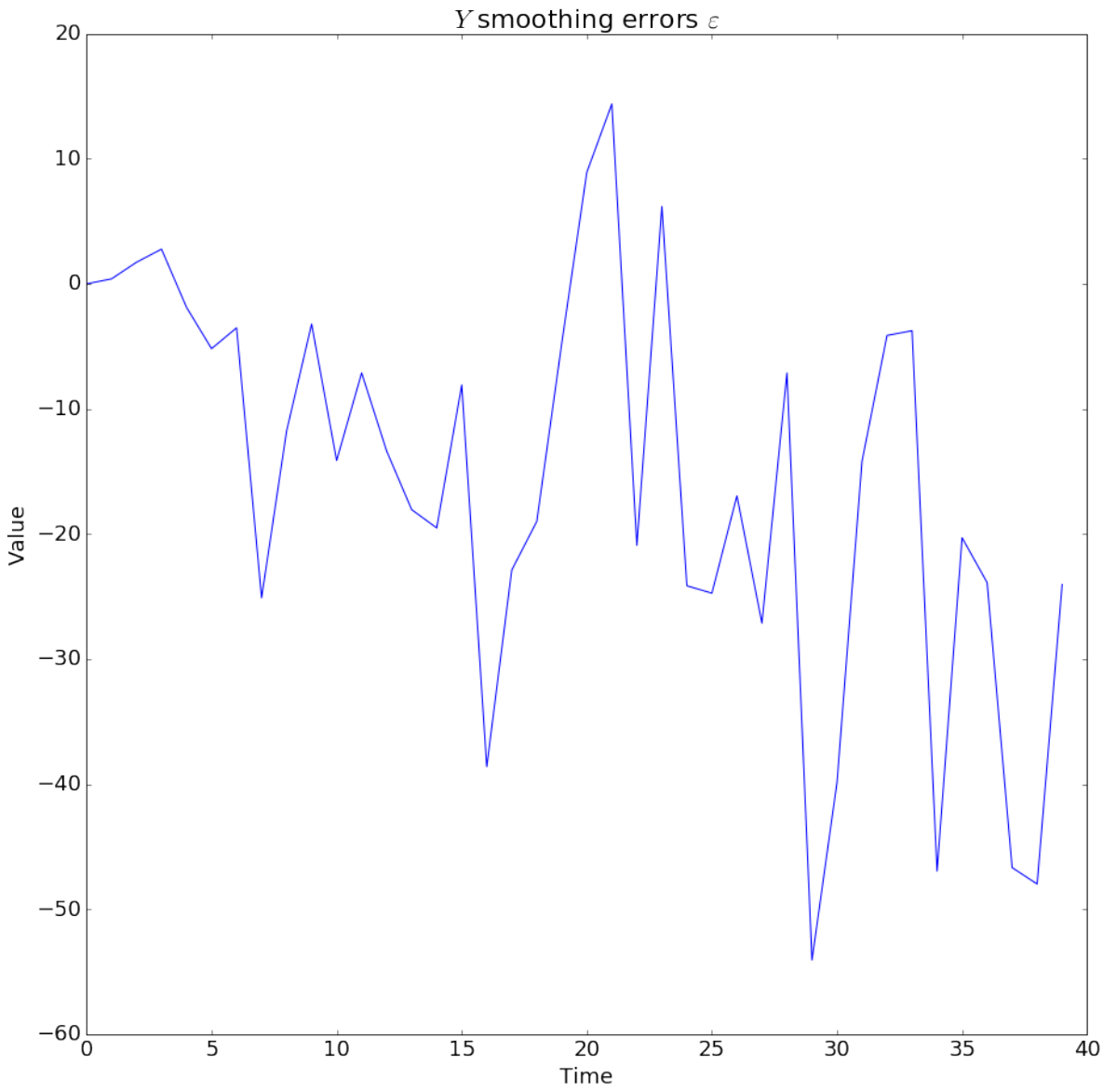
Це призводить до того, що деякі помилки невипадкові, тобто між ними є певна математична залежність. Проаналізуємо простий випадок, коли присутній деякий лаг τ, і помилки впливають не тільки на сусідні, але й через певний час τ. Тоді помилка має випадкову e та невипадкову частину δ

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

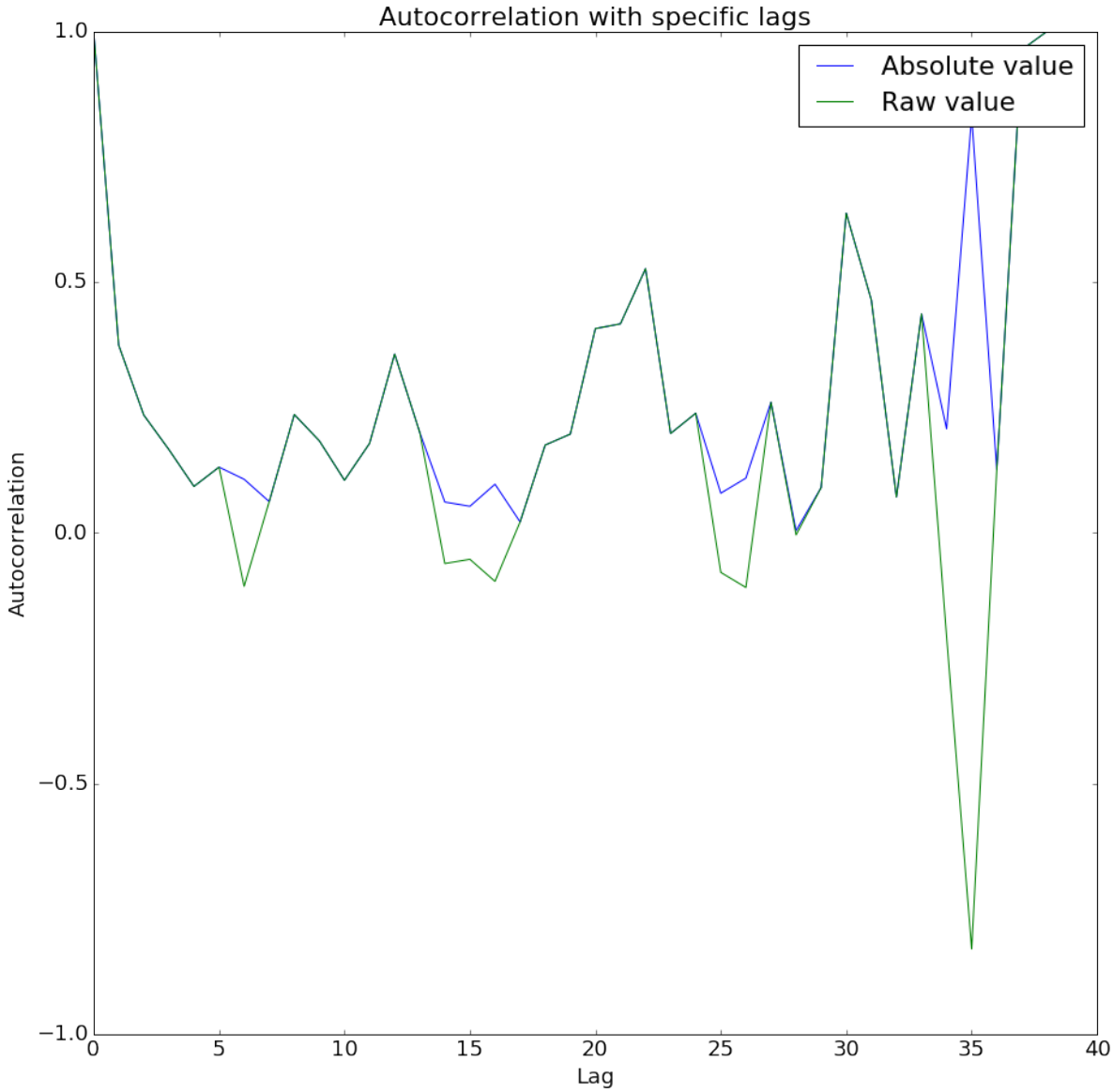
Аналітичний вираз для невипадкової частини — лінійна функція

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

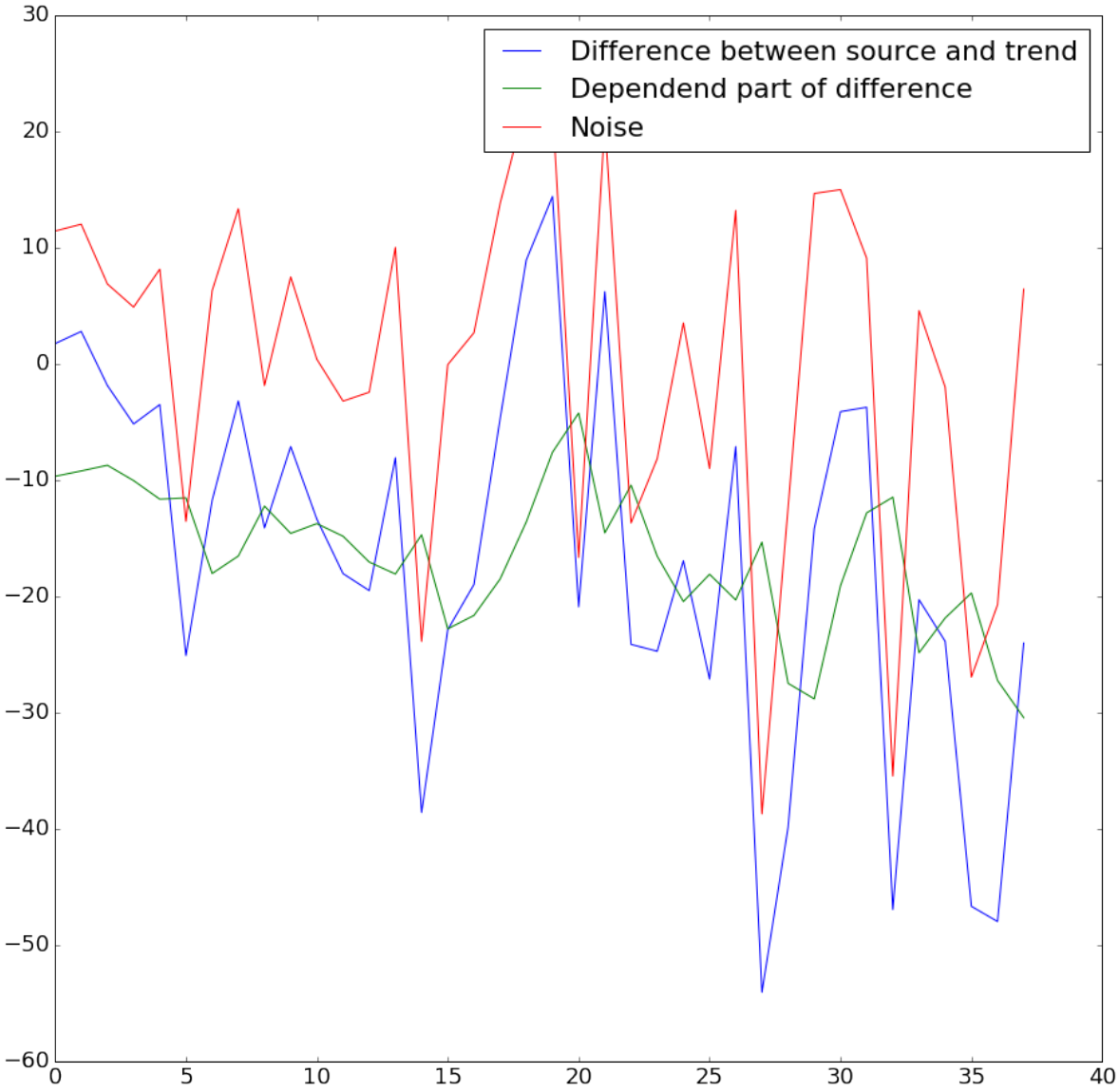
Дійсно видно, що середнє значення похибки ненульове, тобто її можно як мінімум зсунути по вертикальній осі.



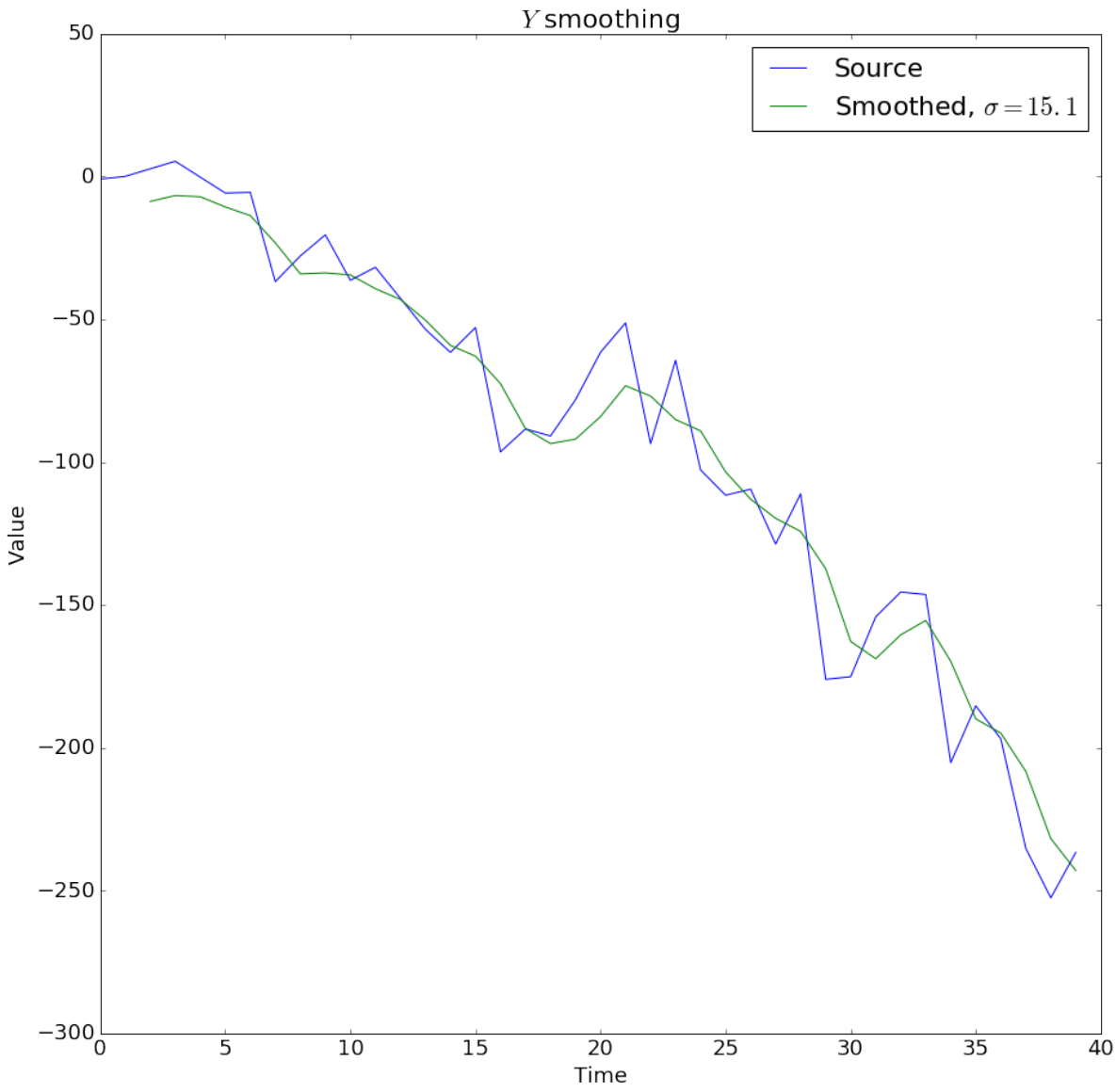
Щоб визначити величину лагу, потрібно порахувати авторегресії різних порядків. Далі методом найменших квадратів розраховуються коефіцієнти для рівняння (1.1).



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.



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



1.4 Прогнозування

Наявна лінія тренду може бути спрогнозована за формулою

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

де

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

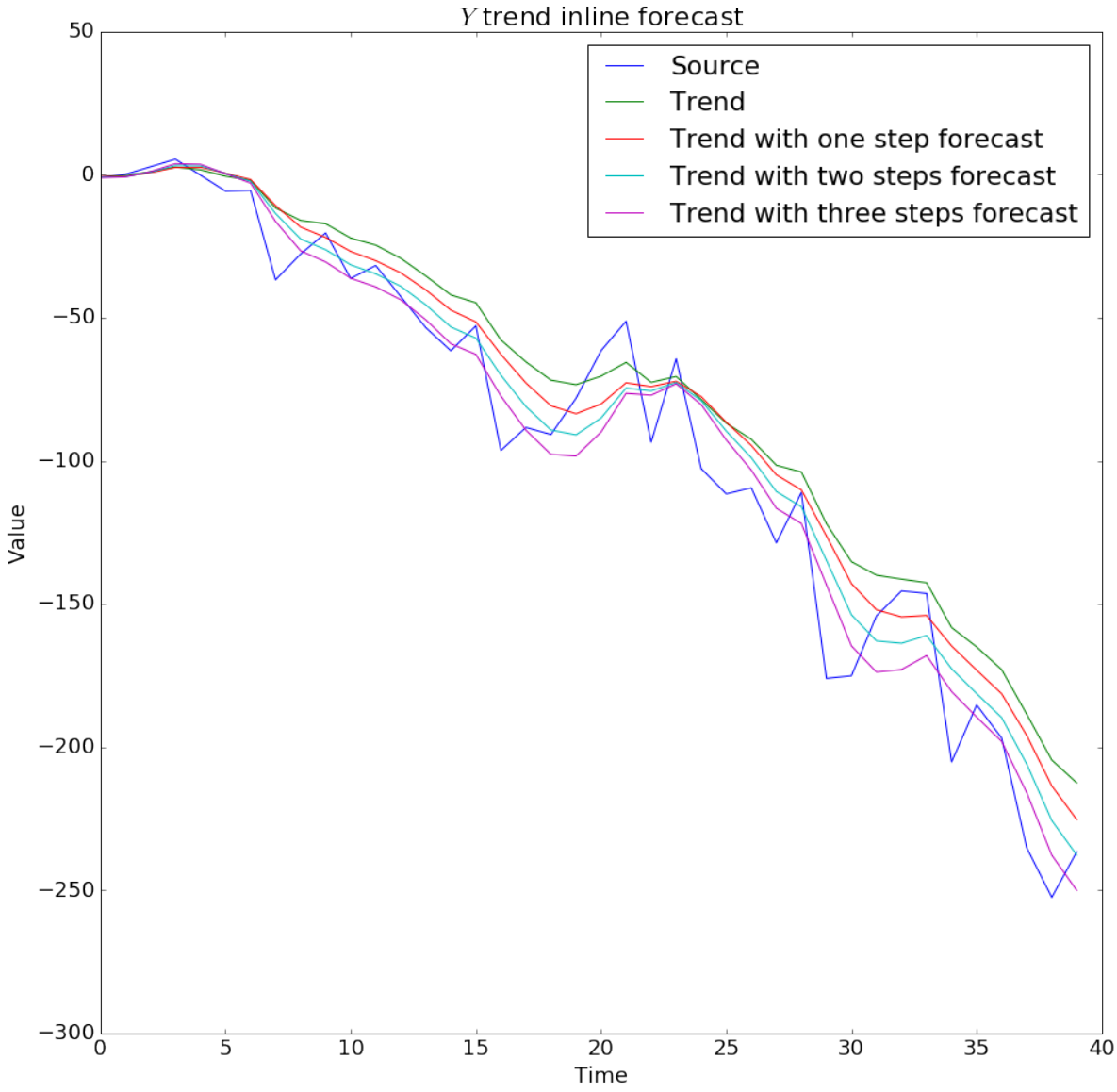
і

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

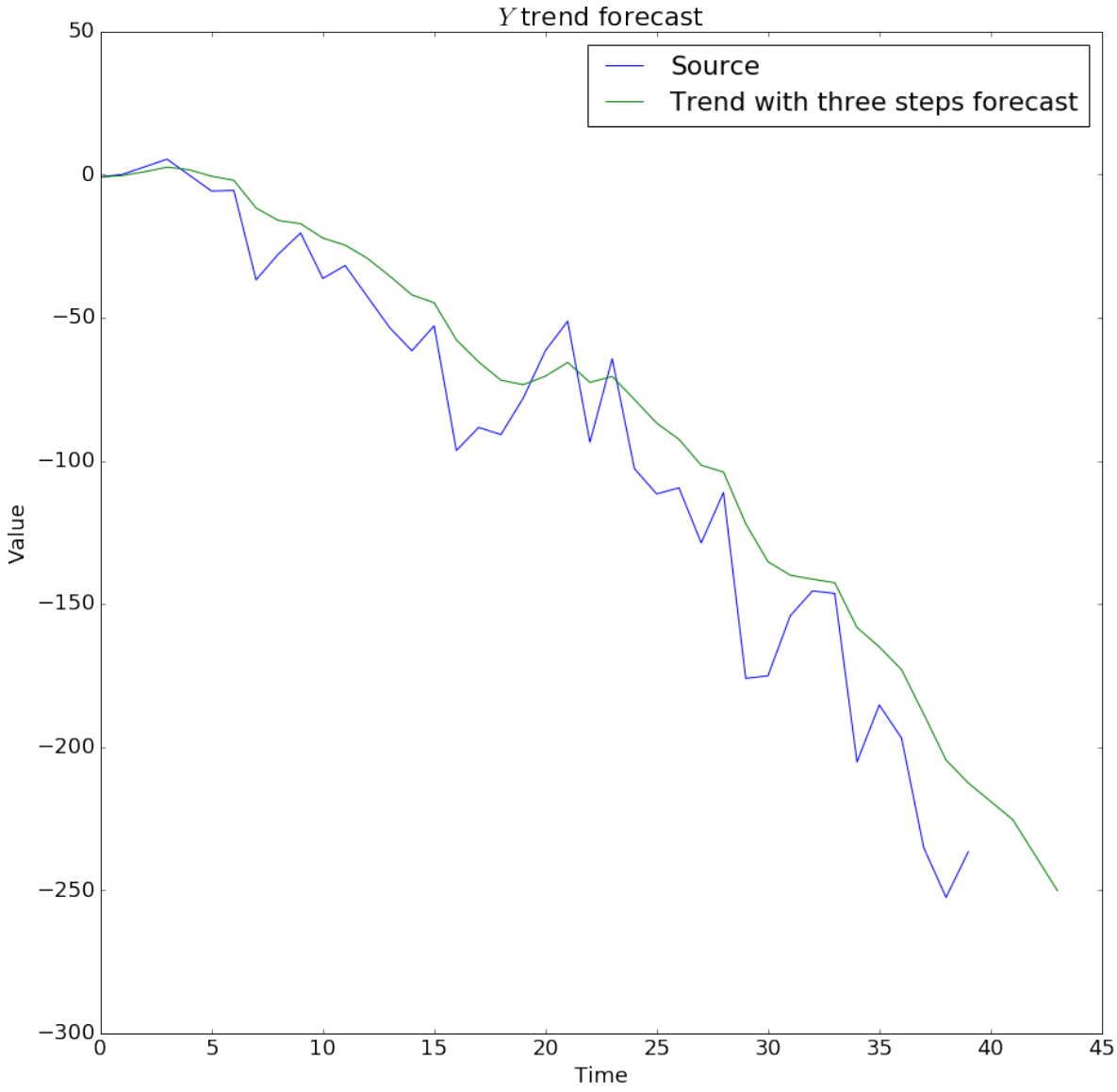
Проте нам вистачить лінійного прогнозу, тому член формули з квадратом відкинемо

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

Щоб приблизно зрозуміти якість прогнозу, можемо побудувати його для вже існуючих даних та порівняти



Прогноз на три кроки після останньої точки має вигляд.

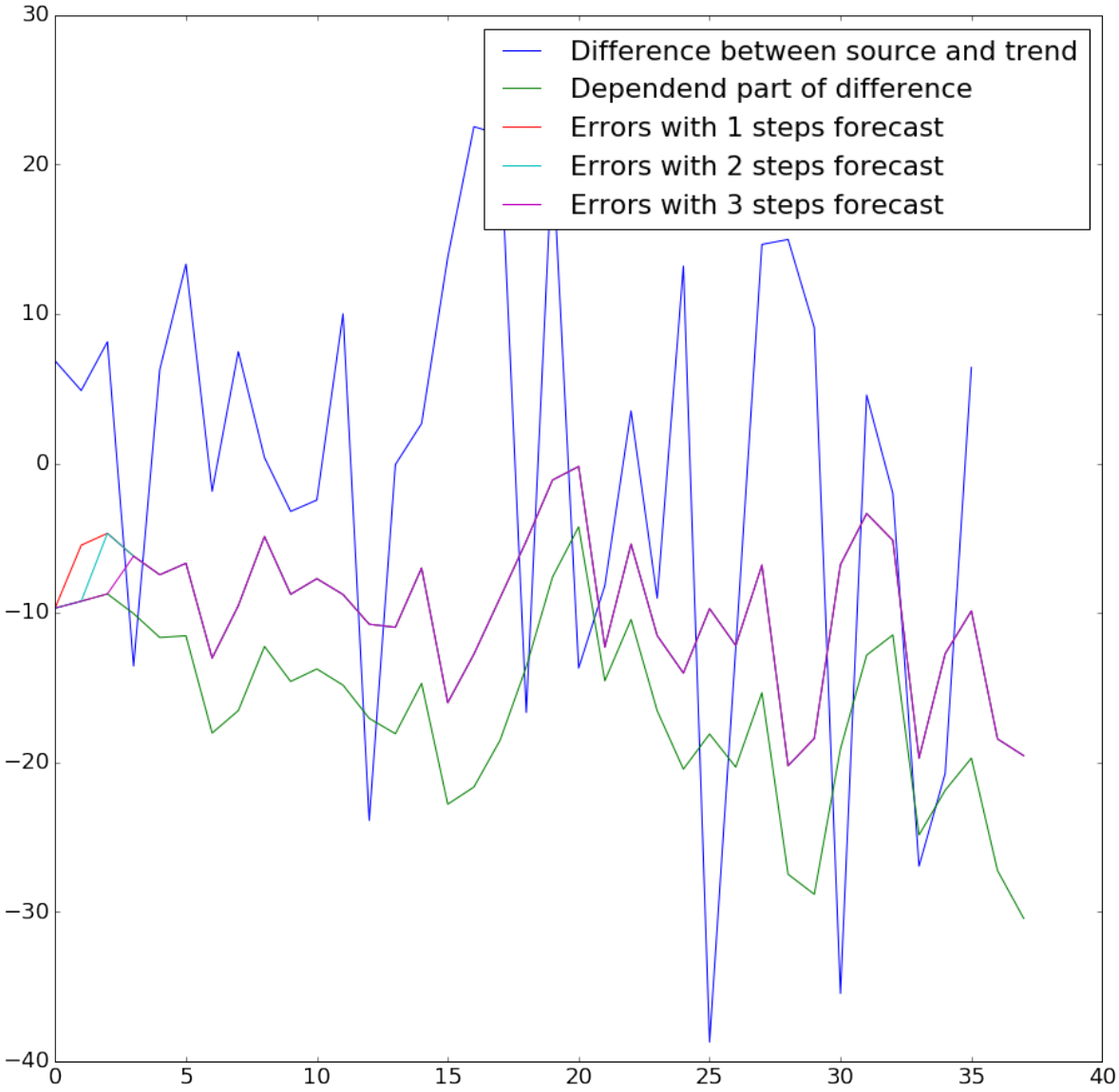


1.5 Прогнозування невипадкових помилок

Ми маємо лінійну формулу для розрахунку залежності між невипадковими помилками

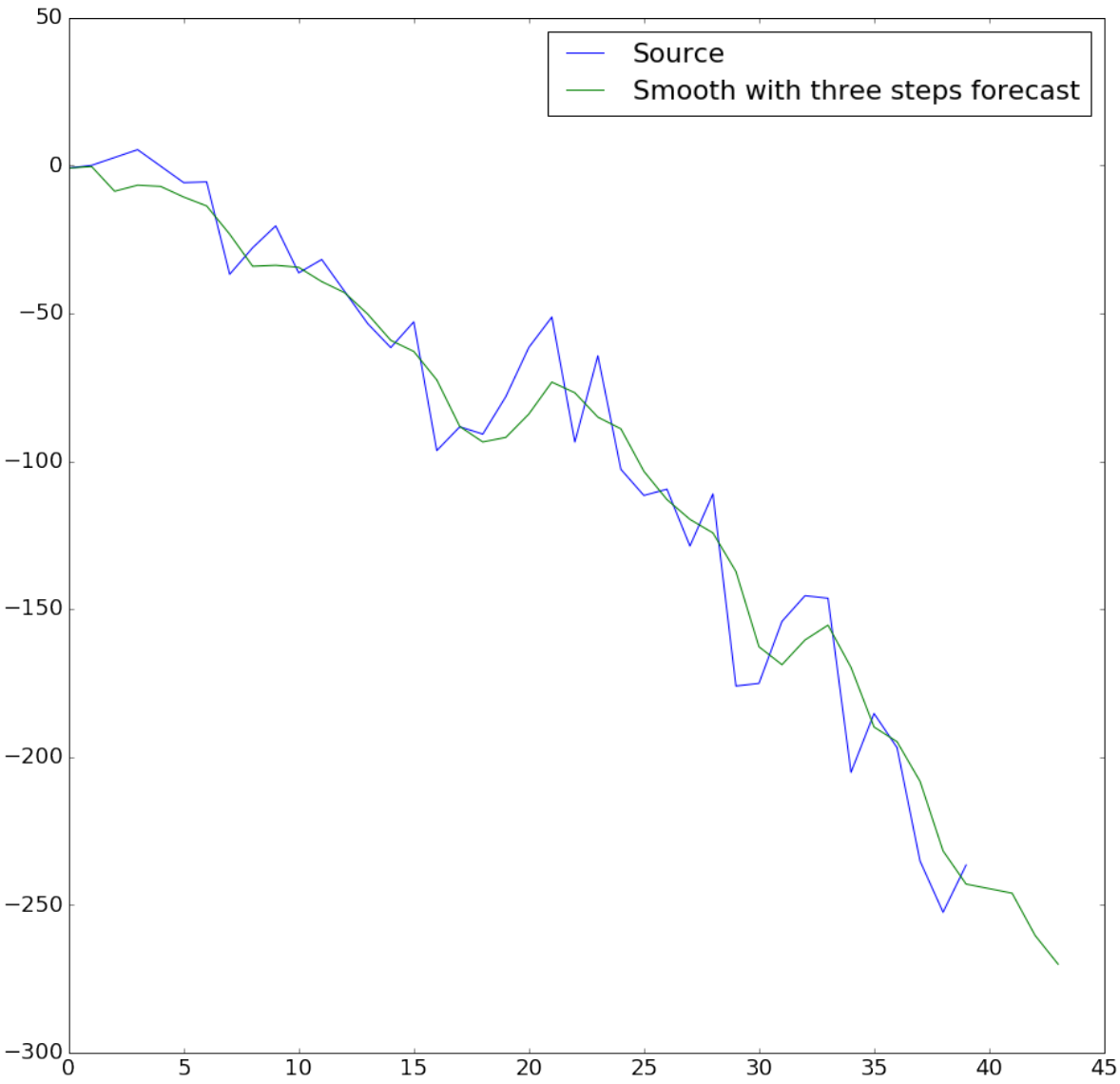
$$\delta_{t+2} = a_0 + \delta_t \cdot a_1 + \delta_{t+1} \cdot a_2.$$

Перевіримо якість прогнозу у той самий спосіб — порівняємо його з наявними даними.



1.6 Прогноз процесу

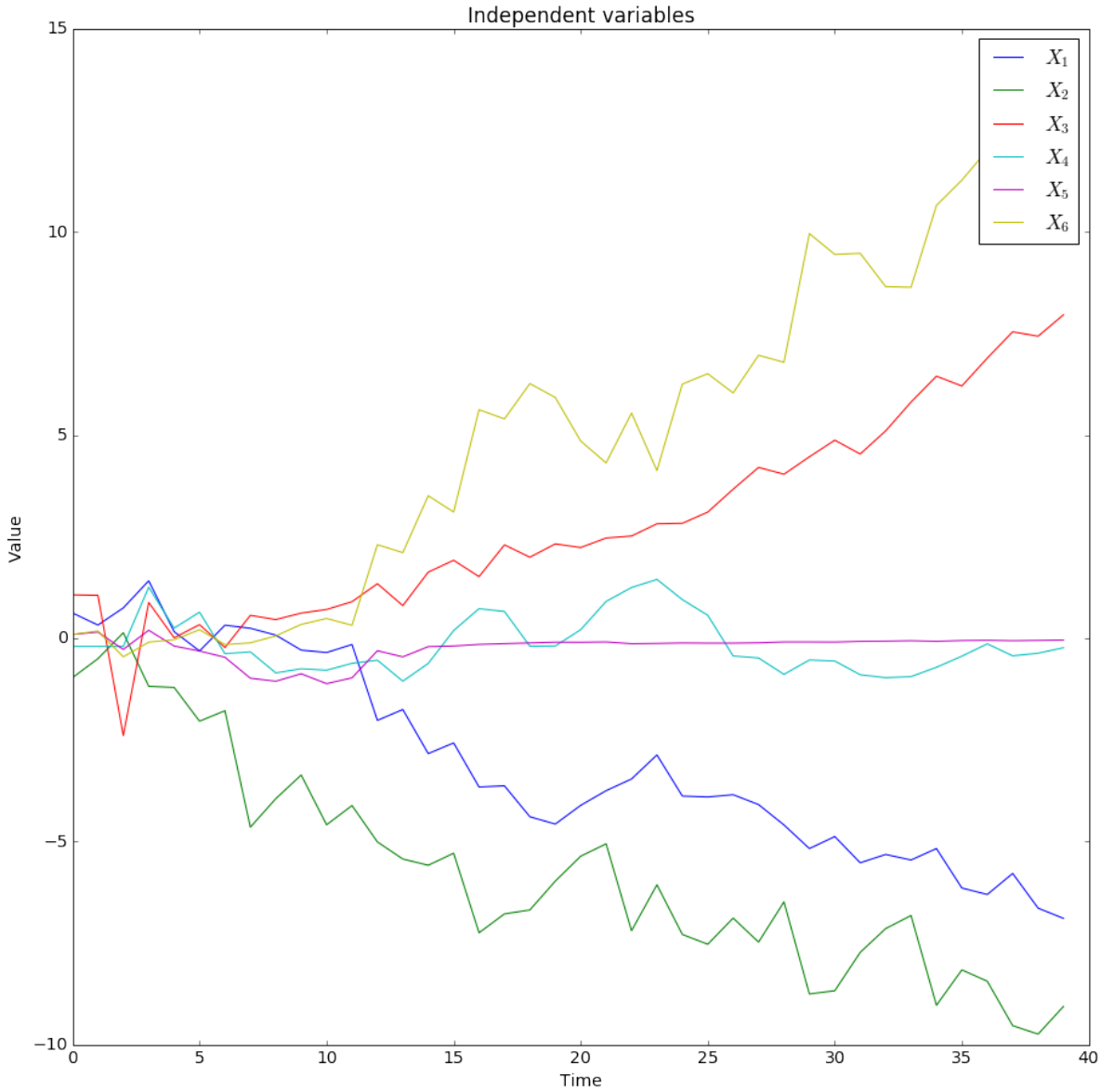
Кінцевий прогноз виглядає наступним чином.


Source
Smooth with three steps forecast
50
0
−50
−100
−150
−200
−250
−300
0
5
10
15
20
25
30
35
40
45

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.

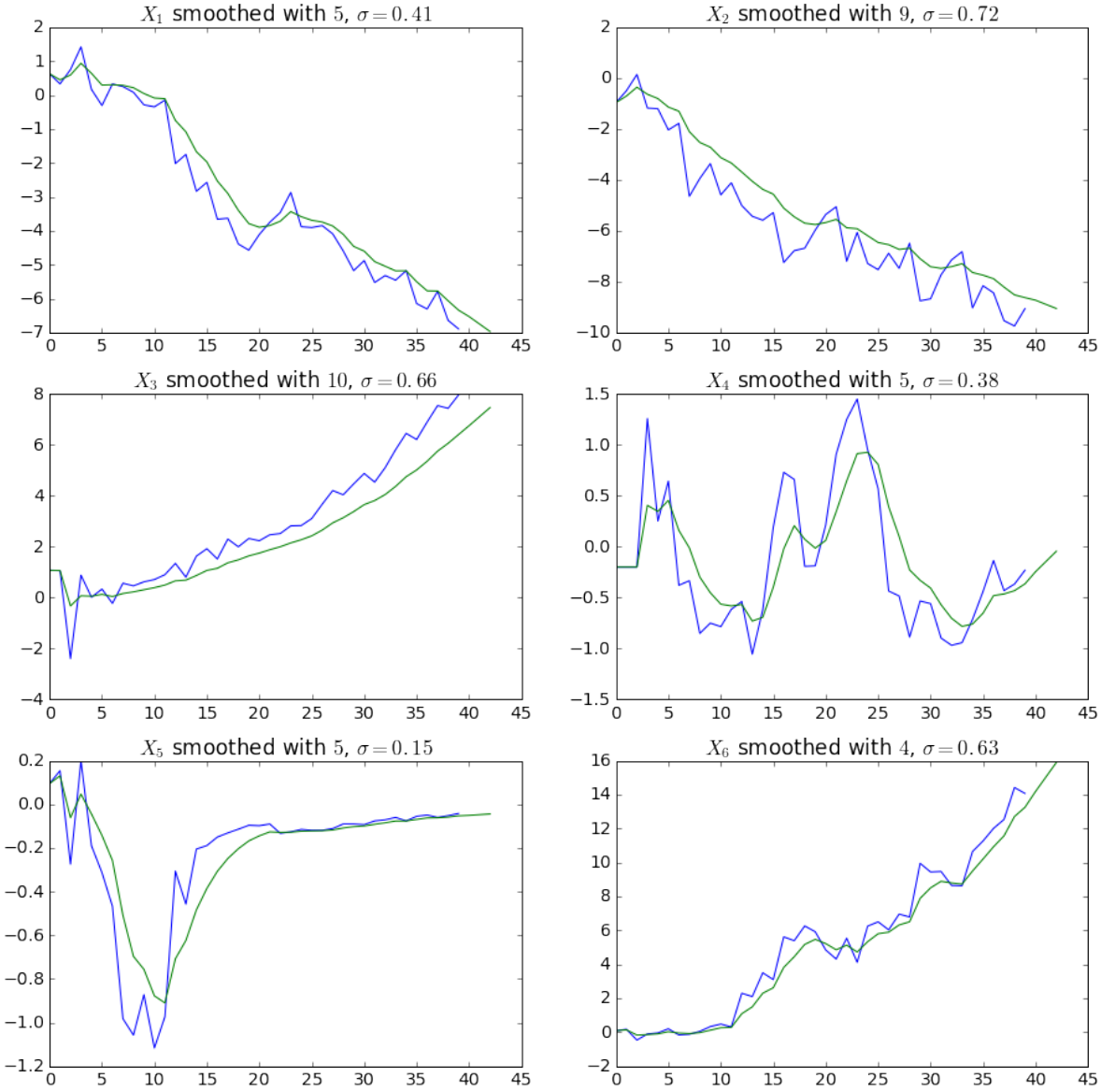
Зобразимо дані процеси графічно.



2.1 Зглажування

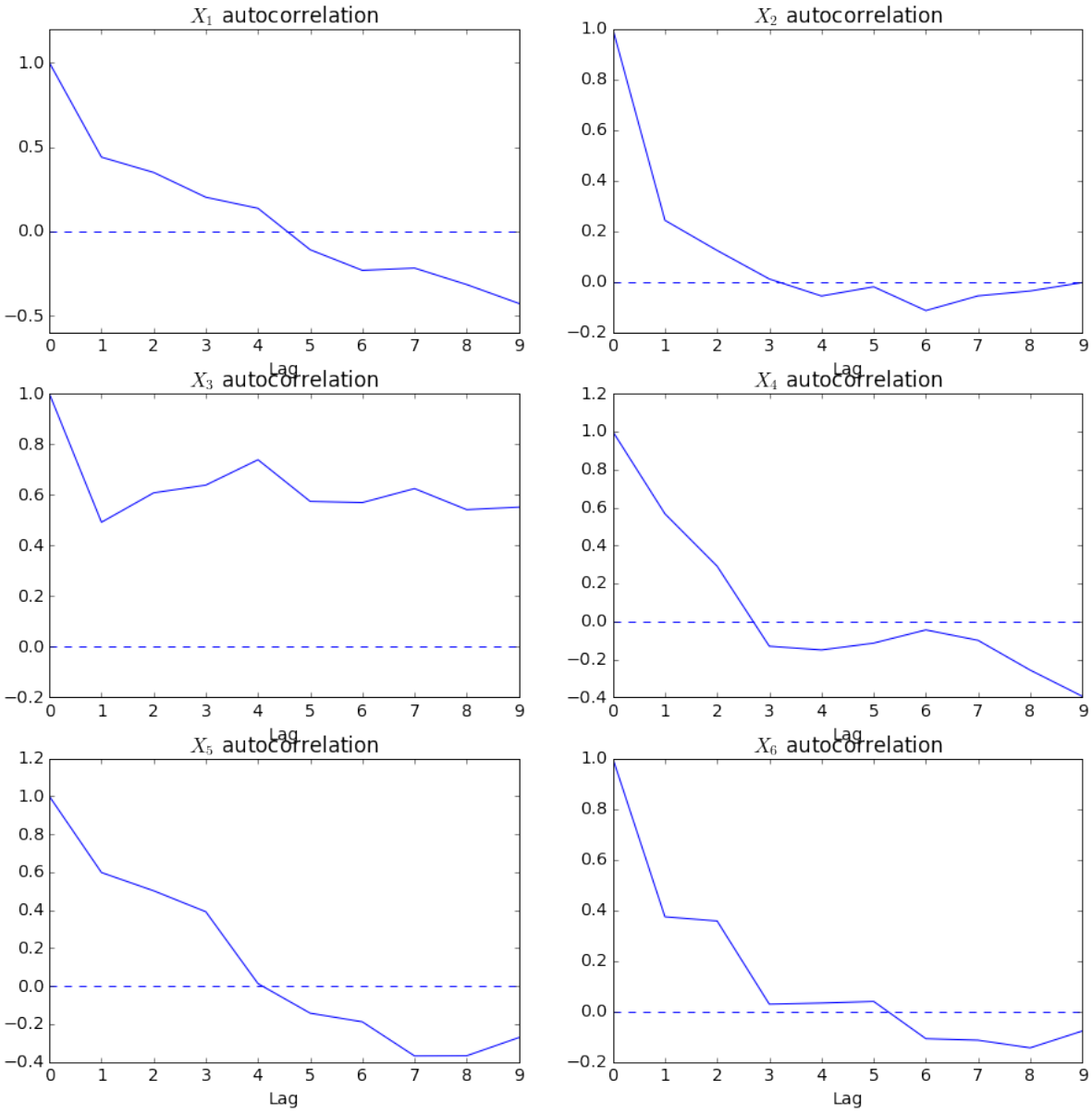
В даному розділі не будемо детально зупинятися на тому, что вже було розглянуто при аналізі Y.

Лінії тренду для всіх X знайдені за допомогою експоненційно зваженого ковзкого середнього.

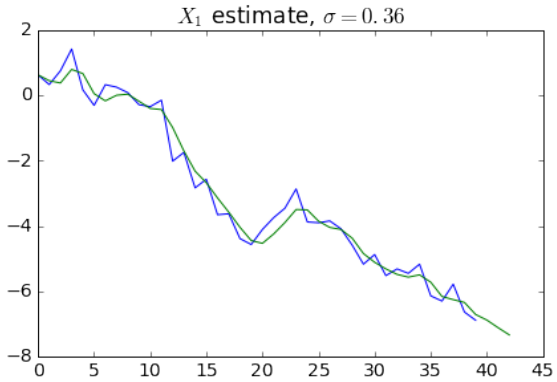


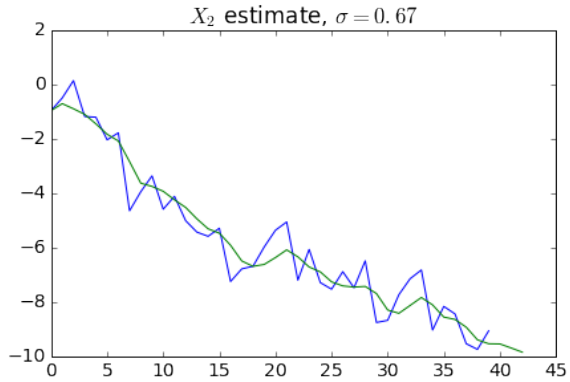
Графіки автокореляцій показали, що достатньо обрати два минулі члені послідовності, щоб зробити прогноз для наступного. Щоправда, X_3 має дуже багато залежностей, це може означати, що кожне наступне значення залежить

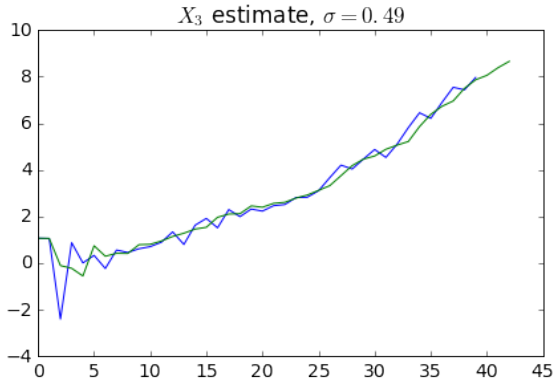
від попереднього і ми маємо Марковський ланцюг.

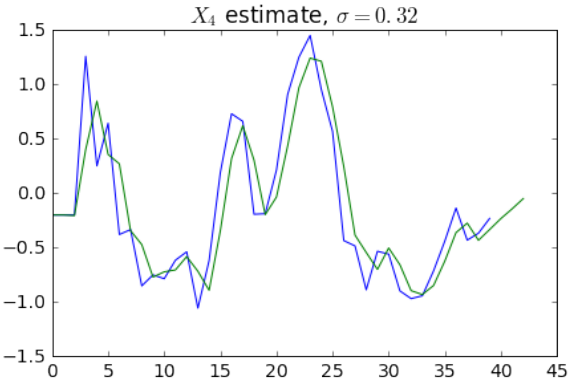


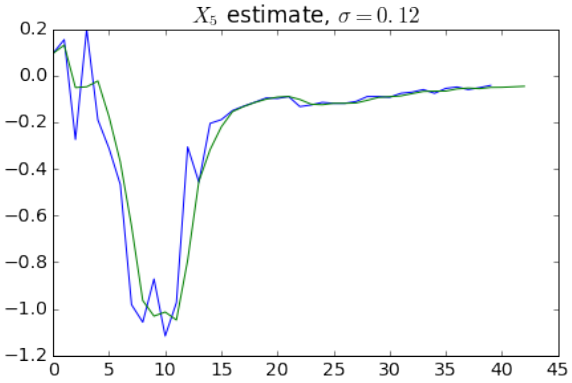
Фильтр, що враховує невипадкові помилки, дуже добре відповідає дійсним даним.

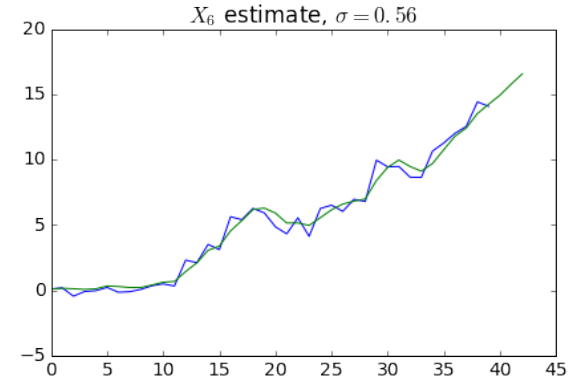

X_1 estimate, $\sigma = 0.36$


X_2 estimate, $\sigma = 0.67$

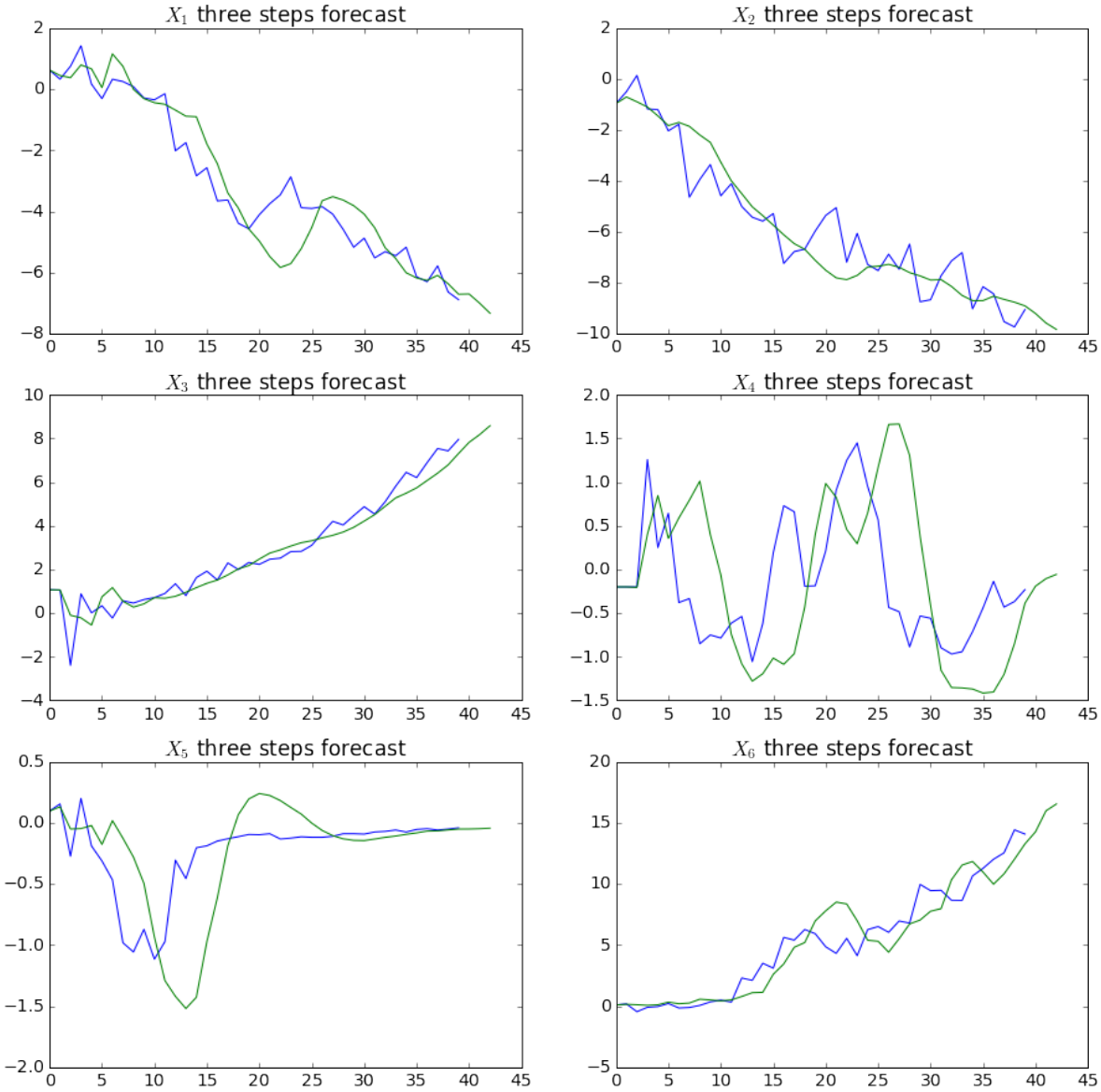

X_3 estimate, $\sigma = 0.49$


X_4 estimate, $\sigma = 0.32$


X_5 estimate, $\sigma = 0.12$


X_6 estimate, $\sigma = 0.56$

2.2 Прогноз часових рядів



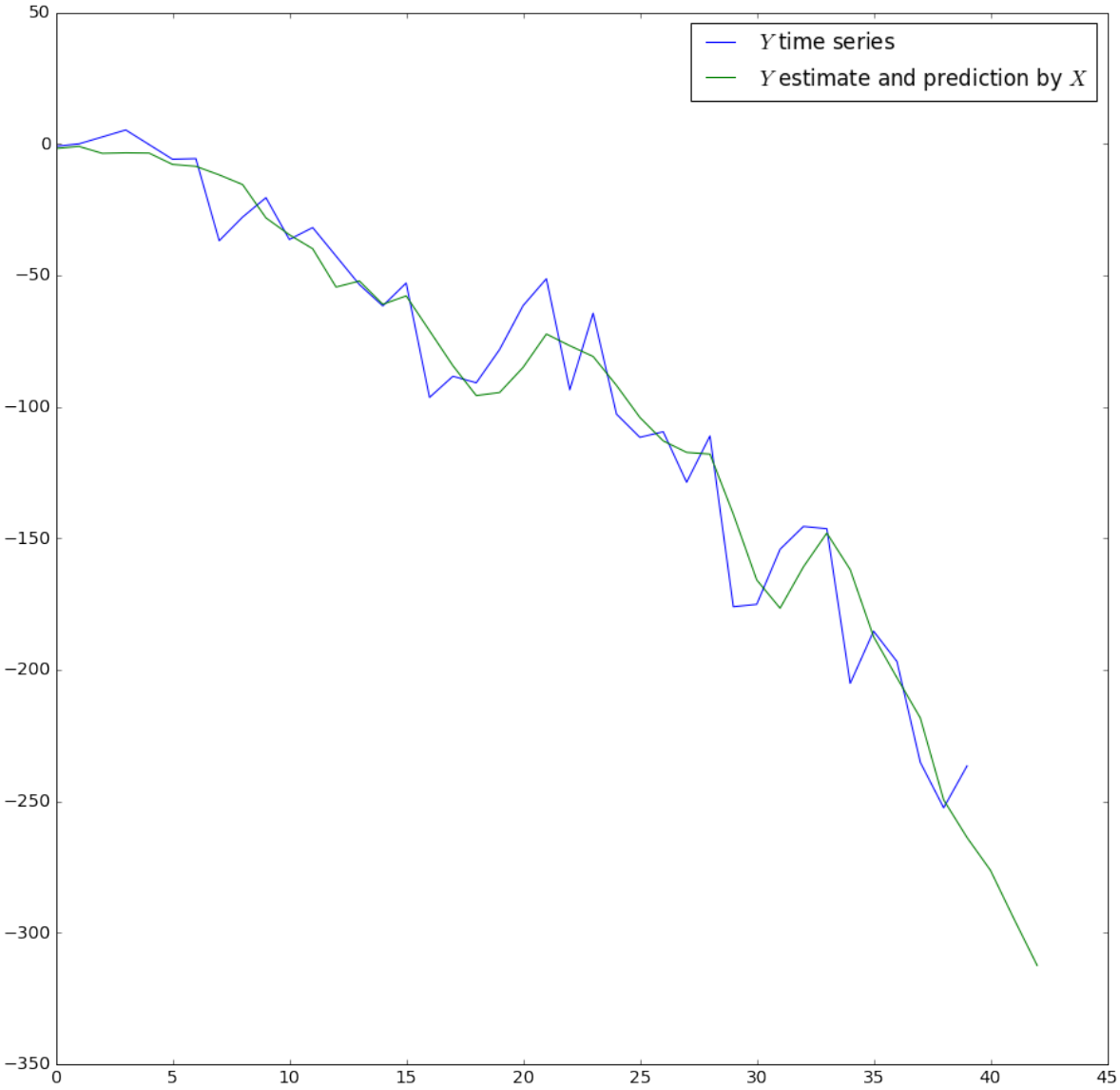
2.3 Побудова моделі

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 регресорами.

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.


Y time series
Y estimate and prediction by X
50
0
−50
−100
−150
−200
−250
−300
−350
0
5
10
15
20
25
30
35
40
45

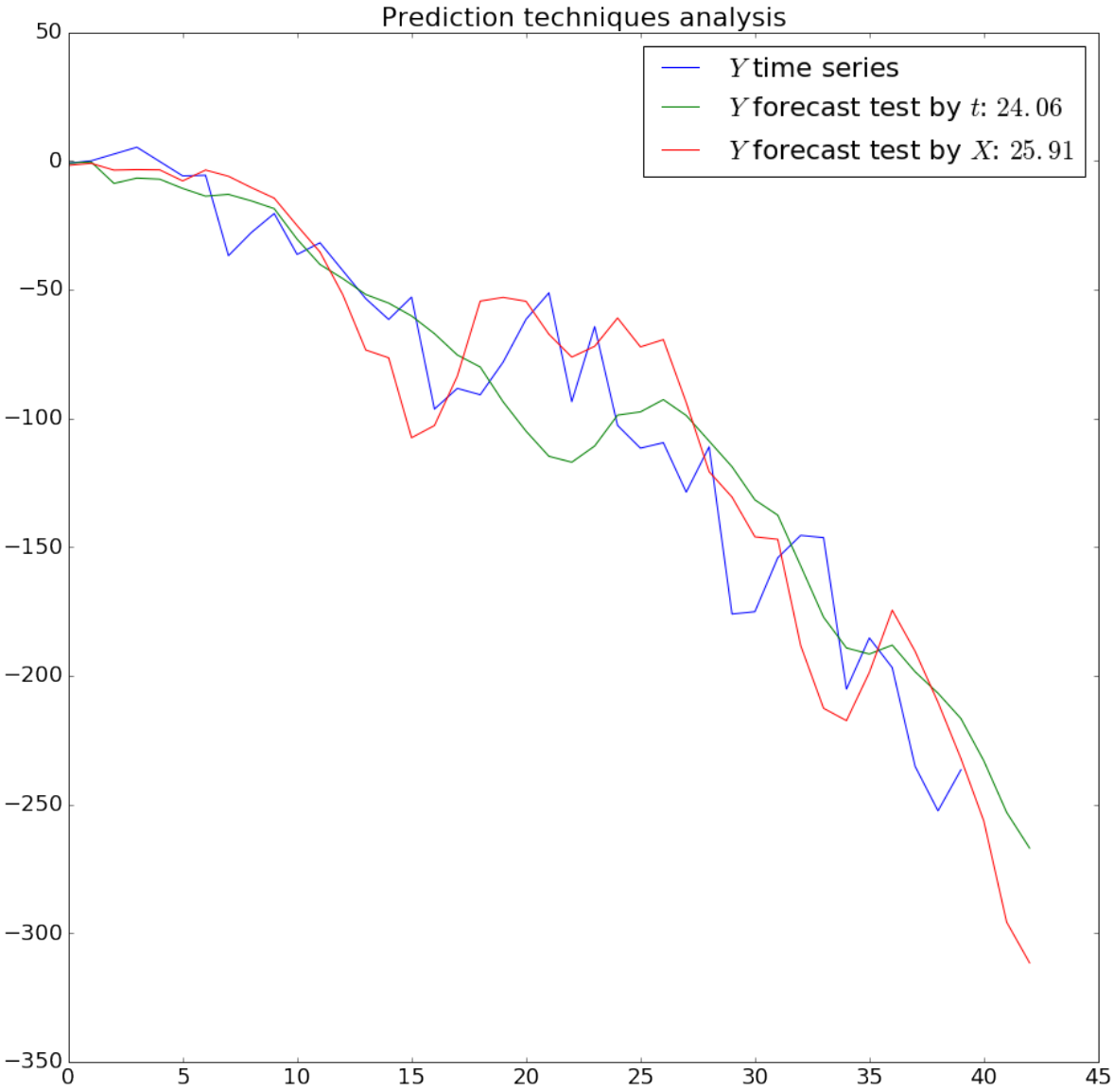
ВИСНОВКИ

В результаті виконання роботи вдалося

1) визначити оцінку (зглажування) 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 ретроданих 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 $\sigma = 25.91$. 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.


Prediction techniques analysis
Y time series
Y forecast test by t: 24.06
Y forecast test by X: 25.91
50
0
−50
−100
−150
−200
−250
−300
−350
0
5
10
15
20
25
30
35
40
45

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.