

*Баран Данило Романович, аспірант кафедри обчислювальної техніки,
факультету інформатики та обчислювальної
КПІ ім. Ігоря Сікорського, Україна*

*Туганських Олександр Антонович, здобувач вищої освіти,
КПІ ім. Ігоря Сікорського, Україна*

*Науковий керівник: Писарчук Олексій Олександрович, доктор технічних наук, професор,
професор кафедри обчислювальної техніки, факультету інформатики та обчислювальної
КПІ ім. Ігоря Сікорського, Україна*

МЕТОД АНАЛІТИЧНОГО ВИЗНАЧЕННЯ ПАРАМЕТРІВ НЕЛІНІЙНИХ МОДЕЛЕЙ ЗА СТАТИСТИЧНОЮ ВИБІРКОЮ ВИМІРІВ

Анотація. Запропоновано метод аналітичного визначення параметрів нелінійних моделей за статистичною вибіркою вимірів для задач Data Science. Метод базується на методології статистичного навчання в схемі диференціально-нетейлорівських перетворень. Наведено приклад застосування запропонованого методу з побудовою експоненційної моделі. Доведено ефективність запропонованого підходу, порівняно із класичними методами чисельних ітераційних розрахунків.

КЛЮЧОВІ СЛОВА: модель, статистичне навчання, Data Science, диференціальні перетворення.

Abstract. Analytical method of determining parameters for nonlinear models using statistical samples in Data Science is suggested here. It is based on statistical learning methodology in scheme of differential non-taylor transformations. As example of suggested method usage, the exponential model was built. The effectiveness of its application was confirmed using classical numerical iterative calculations as comparison.

KEY WORDS: model, statistical learning, Data Science, differential transformations.

Вступ. Сучасним прикладним задачам технологій Data Science притаманні нелінійність, значна динаміка зміни та надмірність статистичних вибірок досліджуваних процесів. При цьому висуваються жорсткі вимоги до оперативності, точності та достовірності результатів обробки, а також до інваріантності математичної моделі статистичного навчання до властивостей експериментальних даних. Це може бути досягнуто впровадженням аналітичних підходів до визначення параметрів нелінійних моделей. Традиційно, для вирішення цієї задачі використовуються чисельні методи апроксимації [1]. Проте вони мають значну обчислювальну складність та потребують апіорної інформації про межі зміни шуканих параметрів моделі. Це унеможливує їх використання для задач сучасних задач Data Science. Тому *актуальною* є задача удосконалення методів статистичного навчання у напрямку подолання недоліків традиційних підходів. *Метою* досліджень є розробка методу аналітичного визначення параметрів нелінійних моделей за статистичною вибіркою вимірів

Основна частина. Суть методу аналітичного визначення параметрів нелінійних моделей за статистичною вибіркою – імплементація теоретико-емпіричного підходу до аналізу нелінійних процесів. Він базується на перенесенні якостей спрощеної моделі досліджуваного процесу на нелінійну, яка є більш адекватною до теоретичного опису процесу. Для отримання параметрів спрощеної моделі можуть пропонуватися використати будь-який метод статистичного

навчання, наприклад метод найменших квадратів (МНК) Теоретичний опис математичної моделі може бути отриманий з аналітичних спостережень за природою вибірки даних. Значна, складність нелінійного теоретичного опису моделі унеможливує застосування класичних статистичних алгоритмів згладжування. Запропонований метод полягає в багатоетапному отримання параметрів нелінійної моделі з використанням

диференційних перетворень [3] для мінімізації розходжень поліноміальної та нелінійної моделей. Такий підхід дозволяє поєднати відносну простоту поліноміальної моделі з більш прогностично стійкою нелінійною моделлю.

Вихідними даними для аналізу нелінійного процесу є дискретна вибірка параметрів:

$$y = \{y_1, y_2, y_3, \dots, y_n\}, \quad (1)$$

Аналітично визначена нелінійна модель, позначено $f(t, c)$, та задана функціональною залежністю або диференціальним рівнянням. З теоретичної точки зору, це найбільш точно описує природу досліджуваного процесу. Необхідно побудувати нелінійну модель досліджуваного процесу $f(t, c)$ за вибіркою експериментальних даних (1).

Далі, використовуючи обраний статистичний алгоритм згладжування, наприклад МНК, формуються параметри експериментальної моделі (2).

$$z^{\wedge}(t) = \Phi(y), \quad (2)$$

де Φ позначає операції алгоритму визначення параметрів моделі $z^{\wedge}(t)$ за вибіркою експериментальних даних (1) відповідно до обраного методу статистичного аналізу. Форма моделі $z^{\wedge}(t)$ є, як правило, поліноміальною в обраному базисі, що має властивості і недоліки традиційних підходів статистичного навчання:

$$y(x) = \sum_{i=1}^m c_i \varphi_i(x_i), \quad (3)$$

де c_i – коефіцієнти полінома; $\varphi_i(x_i)$ – базисні функції полінома.

Таким чином отримано дві різні моделі. $f(t, c)$ є теоретичною формою, яка більш адекватно відображає властивості досліджуваного процесу, але для неї не відомі ключові параметри $c = \{c_i\}, i = 1 \dots m$. В той час $z^{\wedge}(t)$ є апроксимуючою експериментальною моделлю, форма і параметри якої відомі з використання класичних статистичних методів, але вона менш адекватна процесу та містить більший вплив випадкових помилок. Операція перенесення властивостей з однієї моделі на іншу є наближенням, реалізувати який можна методом балансу диференціальних спектрів (БДС) [2].

Сутність наближення експериментальної та теоретичної моделей полягає у визначенні параметрів більш складної нелінійної моделі за спрощеною поліноміальною. Для цього можливо використовувати диференційні перетворення з різними критеріями наближення (в альтернативу БДС) [3].

$$\delta(c) = D \left(\left| P \left[\hat{z}(t) \right]_i \Rightarrow \hat{z}(k) \right| - \left| P \left[f(t, c) \right]_i \Rightarrow F(k, c) \right| \right) = D \left(\left| P \left[\varepsilon(t, c) \right]_i \Rightarrow E(k, c) \right| \right) \rightarrow \min, \quad (4)$$

де D позначає операції обраного критерію наближення експериментальної моделі до теоретичної. Визначення параметрів $c = \{c_i\}$ нелінійної моделі $f(t, c)$ здійснюватиметься шляхом формування та розв'язку системи рівнянь:

$$\frac{d\delta(c)}{dc} = 0, \quad (5)$$

Покладемо, що процес, який має експоненційну природу заданий у формі набору вимірів (1) має загальну аналітичну форму:

$$f(t, a) = a_0 + a_1 t + a_2 e^{a_3 t}, \quad (6)$$

де t – це залежний параметр моделі; a – це вектор незалежних параметрів моделі.

Надалі проводиться згладжування вибірки (1) статистичним методом - МНК.

Визначення невідомих параметрів апроксимуючої функції за МНК реалізується за матричним виразом [2]:

$$C = (A^T A)^{-1} A^T Y, \quad (7)$$

де Y – вектор вибірки вимірюваних даних; A – матриця базисних функцій; C – вектор шуканих параметрів моделі.

Результатом застосування (7) є набір незалежних параметрів експериментальної моделі, статистично «навчені» за вибіркою вимірів. Надалі визначаються параметри нелінійної моделі за параметрами поліноміальної за схемою (5):

$$V(t, c) = c_0 + c_1 t + c_2 t^2 + c_3 t^3, \quad (8)$$

$$F(t, a) = (a_0 + a_2) + (a_1 + a_2 + a_3)t + \frac{a_2 a_3}{2} t^2 + \frac{a_2 a_3}{6} t^3, \quad (9)$$

де V, F поліноміальні моделі, утворені з диференціальних Р-спектрів початкових моделей.

Наближення двох моделей реалізується за методом БДС, що дає можливість утворити систему рівнянь:

$$\begin{cases} c_0 - (a_0 + a_2) = 0, \\ c_0 - (a_1 + a_2 + a_3) = 0, \\ c_0 - \frac{a_2 a_3^2}{2} = 0, \\ c_0 - \frac{a_2 a_3^3}{6} = 0. \end{cases} \quad (10)$$

Розв'язком системи рівнянь (10) є параметри нелінійної моделі (6), пов'язані із параметрами поліноміальної моделі (8)

$$\begin{cases} a_0 = c_0 - a_2, \\ a_1 = c_1 - (a_2 * a_3), \\ a_2 = \frac{2c_2}{a_3^2}, \\ a_3 = \frac{3c_3}{c_2}. \end{cases} \quad (11)$$

Для перевірки та порівняння якості отриманих результатів з іншими методами, було виконано моделювання з наступними параметрами:

Ідеальний тренд:

$$f(x) = 0.0000005x^2, x \in 0, 1 \dots 10000;$$

Стохастичні відхилення: нормальний закон, $\mu = 0, \sigma = 5$;

Аномальні відхилення: нормальний закон, $\mu = 0, \sigma = 15$;

На рис. 1 візуалізовано ідеальний тренд $f(x)$ з різними відхиленнями позначеними різними кольорами.

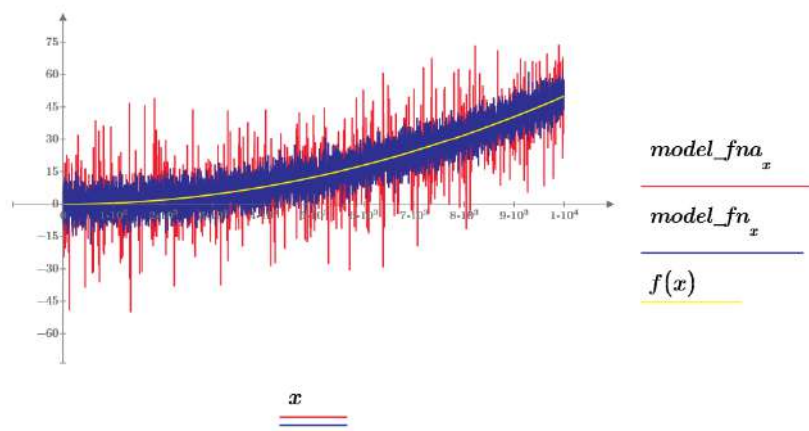


Рис. 1

До обробки моделі вхідних даних – визначення лінії тренду було застосовано три підходи:

- МНК - 1;
- запропонований метод - 2;
- чисельний метод з використанням алгоритму Левенберга-Маркардта -3;

Отримані результати кожного з методів були оцінені за показниками: середньоквадратичне відхилення (СКВ),

лінійного відхилення (ЛВ) та часу обрахунку T . Значення показники занесено в табл. 1.

Таблиця 1

Методи	СКВ	ЛВ	T (мкс)
1	37.842	0.102	436
2	44.489	0.098	568
3	37.808	0.102	4647

Висновки. Отримані дані розрахунків та моделювання показують, що застосування розробленого методу для визначення параметрів нелінійної моделі є ефективним за показниками часу та адекватності моделі, порівняно із чисельними підходами. Запропонований підхід базується на використанні методу диференціальних перетворень та

може бути масштабований для інших типів нелінійних моделей. Цей перехід дозволяє покращити також прогностичні характеристики моделі.

Список інформаційних джерел

1. Chapra, Steven C., and Raymond P. Canale. Numerical methods for engineers. Vol. 1221. New York: McGraw-hill, 2011. – 987p.
2. Пухов Г.Е. Дифференциальные преобразования и математическое моделирование физических процессов. – Киев: Наук. Думка, 1986. – 159 с.
3. Hastie, Trevor, et al. The elements of statistical learning: data mining, inference, and prediction. Vol. 2. New York: springer, 2009. – 764