УДК 519.233.2: 621.391.83 (045)
DOI: 10.18372/2310-5461.38.12834

В. М. Кузьмин, канд. техн. наук
Національний авіаційний університет
orcid.org/0000-0003-4461-9297
E-mail: kuzmin_vn@i.ua

М. Ю. Заліський, канд. техн. наук
Національний авіаційний університет
orcid.org/0000-0002-1535-4384
E-mail: maximus2812@ukr.net

СТАТИСТИЧНИЙ АНАЛІЗ ДАНИХ З ВИКОРИСТАННЯМ ДВОСЕГМЕНТНОЇ ПАРАБОЛІЧНОЇ РЕГРЕСІЇ

Вступ

Однією із задач побудови математичних моделей є обґрунтування найкращої моделі.

Як критерії, що дозволяють здійснити цей вибір, зазвичай використовують окремо або в комбінації такі критерії:

1. найменшу кількість коефіцієнтів, що сумісні з заданою похибкою;
2. найпростішу форму;
3. розумне фізичне обґрунтування (як наслідок виконання певних законів);
4. мінімальну суму квадратів відхилень між прогнозованими (апроксимованими) та емпіричними значеннями;
5. мінімальну дисперсію [1];
6. мінімізацію максимального відхилення.

Додатковим критерієм також можна вважати аналіз геометричної структури даних (існування точок перегину, дослідження екстремумів, дослідження на асимптотичність, наявність прямолінійних ділянок, існування точок перемикання тощо).

Математичним інструментом для побудови моделей може бути теорія апроксимації, яка переважно ґрунтується на використанні методу найменших квадратів.

Аналіз показує, що в теорії апроксимації необхідно забезпечити найбільш можливу гладкість використовуваних апроксимуючих функцій (тобто забезпечення неперервності та існування похідних вищих порядків на інтервалі апроксимації).

Проте на практиці можливі випадки, коли навіть перша похідна апроксимуючої функції має розриви, що призводить до незадовільних результатів під час побудови гладких математичних моделей для цих випадків.

Тому вибір апроксимуючих функцій є важливим етапом побудови математичних моделей.

Аналіз літератури та постановка проблеми

Аналіз літератури в галузі статистичної обробки емпіричних даних [1–6] показав, що для їх математичного опису зазвичай використовуються поліноми другого, третього та вищих порядків без точок перемикання навіть у випадках, коли сукупність даних різко змінює геометричну структуру (тобто має місце розрив похідної). Зміна структури передбачає наявність декількох сегментів зміни геометричної структури та точок їх перемикання з їх окремою апроксимацією. У таких випадках виникають задачі оптимізації абсцис точок з'єднання декількох сегментів.

У статті буде розглянуто приклад вирішення актуальної науково-технічної задачі обробки реальних статистичних даних щодо середньорічної температури повітря на відповідних градусах широти земної кулі з їх подальшою апроксимацією різноманітними конкуруючими функціями.

Математично задача дослідження може бути сформована так. Нехай для сукупності двовимірних статистичних даних $(x_i; y_i)$ існує певна множина апроксимуючих функцій $\hat{y}_i = f_n(x_i, \vec{a}_{m,n})$, що встановлює залежність між ними (де $\vec{a}_{m,n}$ — вектор m параметрів апроксимуючої функції; n — номер апроксимуючої функції). Для кожної апроксимуючої функції може бути розраховане стандартне відхилення σ між дійсними значеннями y_i та їх оцінками $\hat{y}_i$. Тоді вибір найкращої математичної моделі буде здійснюватися відповідно до наступного критерію

$$n = \inf\left(s \in N \,\forall j : \sigma(f_s(x_i, \vec{a}_{m,s})) \le \sigma(f_j(x_i, \vec{a}_{m,j}))\right).$$

Основна частина

Розглянемо реальний приклад конкретних статистичних даних. Вихідні дані щодо залежності середньорічної температури повітря *t* від широти земної кулі *n* наведені в табл. 1 [4].

Таблиця 1

Вихідні дані

n, °	–90	–80	–70	–60	–50	–40
t, °C	–32,6	–21,9	–12,2	–3,4	4,4	11,3
n, °	–30	–20	–10	10	20	30
t, °C	17,2	22,1	26,1	28,2	23,6	18,5
n, °	40	50	60	70	80	90
t, °C	12,7	6,4	–0,54	–8	–16,1	–24,8

Графічна інтерпретація вихідних даних показана на рис. 1.

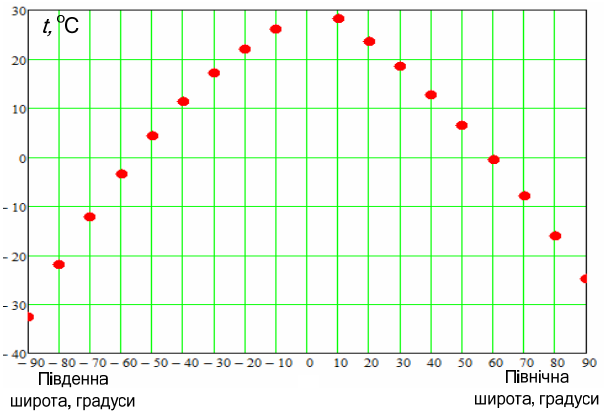


Рис. 1. Залежність середньорічної температури від широти земної кулі

Візуальний аналіз графіку вихідних даних (рис. 1) показує, що ці емпіричні дані можна умовно поділити на дві ділянки (сегменти). Така процедура була здійснена в роботі [4]. При цьому кожний сегмент (зліва та справа відносно нуля) був апроксимований окремими параболами другого порядку з використанням методу найменших квадратів. Далі була визначена точка перетину двох парабол, яка є піком тропічного екватору. Ця точка є дуже важливою під час вирішення задачах дослідження змін земного клімату. Знайдена величина положення тропічного екватора не є сталою величиною, вона змінюється в залежності від різних циклів змінювання середньої температури повітря на Землі.

У свою чергу систематичне щорічне визначення положення тропічного екватора може допомогти ідентифікувати зсув параметрів клімату Землі.

Необхідно також зазначити, що методика, запропонована Г. Н. Зайцевим [4], має певну невизначеність. Ця невизначеність полягає в тому, що середньорічна температура в точці географічного екватора не використовувалась під час проведення розрахунків, оскільки невідомо, у який сегмент її необхідно приєднати.

У статті буде розглянути метод, який передбачає більш загальний підхід, за якого розглядається вся сукупність даних одночасно, оскільки вони взаємопов'язані. При цьому використовується функція Хевісайда та методика знаходження оптимальної абсциси точки перемикання параболічних сегментів.

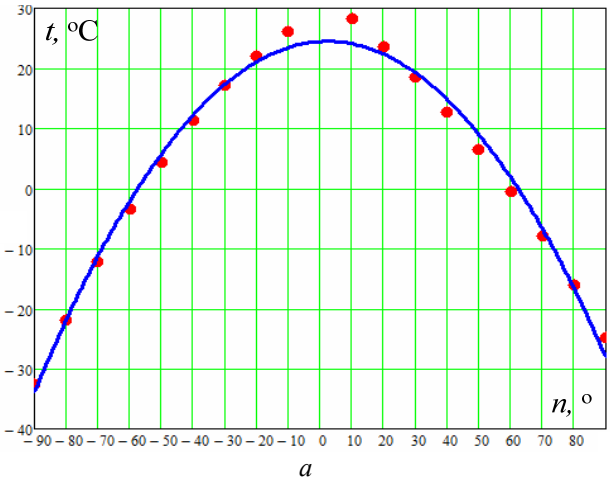
Розглянемо традиційні методи апроксимації з використанням поліномів другого та четвертого порядків.

Отримані рівняння апроксимуючих функцій мають вигляд

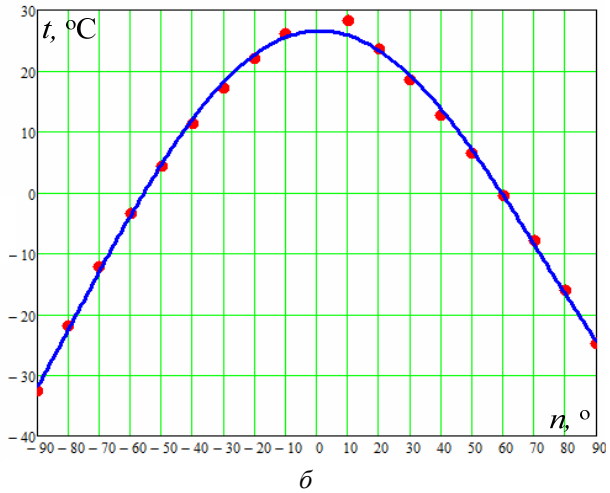
$$f_1(x) = 24,454 + 0,033x - 6,828 \cdot 10^{-3} x^2,$$

$$f_2(x) = 26,561 + 0,015x - 8,965 \cdot 10^{-3} x^2 + $$
$$+ 3,262 \cdot 10^{-6} x^3 + 2,735 \cdot 10^{-7} x^4.$$

Графічне представлення апроксимації вихідних даних поліномами другого та четвертого порядків наведено на рис. 2.



а



б

Рис. 2. Апроксимація вихідних даних поліномами другого (*а*) та четвертого (*б*) порядків

Стандартні відхилення для параболи складало 2,039, а для поліному четвертого порядку складало 0,990, тому апроксимація параболою призводить до незадовільних результатів. Навіть візуальний аналіз показує, що обидві апроксимуючі функції дають занижені результати середньорічної температури.

Числові значення максимальних середньорічних температур, розраховані за результатами цих апроксимацій: для параболи другого порядку – $t_{\max}$ = 24,493 якщо n = 2,412; для параболи четвертого порядку — $t_{\max}$ = 26,567 якщо n = 0.8587.

Як буде показано далі, ці значення суттєво відрізняються від отриманих відповідно до нової методики.

Розглянемо метод апроксимації з використанням двосегментної параболічної функції другого порядку

$$f_3(x) = a + bx + cx^2 + d(x - x_{sw})_+ + e(x - x_{sw})_+^2;$$

$$(x - x_{sw})_+ = (x - x_{sw})h(x - x_{sw}),$$

де $h(x - x_{sw})$ — функція Хевісайда; x_{sw} — абсциса точки перемикання (*switching*).

Під час апроксимації з використанням багатосегментних функцій завжди виникають задачі знаходження оптимальних абсцис точок перемикання сегментів.

Математично задача знаходження оптимальної абсциси точки перемикання $x_{\text{sw opt}}$ може бути сформульована відповідно до ткого критерію:

$$x_{\text{sw opt}} = \inf\left(s \in \mathrm{R} : \sigma(s) \le \sigma(x_{\text{sw}\,i})\right),$$

де $\sigma(x_{\text{sw}\,i})$ — стандартне відхилення статистичних даних від значень двосегментної параболічної функції для всіх можливих значень абсциси точки перемикання $x_{\text{sw}\,i}$.

Для здійснення оптимізації можна було б обмежитися п’ятьма варіантами апроксимації з різними значеннями абсцис точок перемикання, однак у розглянутому прикладі довелося використати сім варіантів (для підвищення точності знаходження оптимуму).

Отримані рівняння двосегментної параболічної функції мають такий вигляд:

$$f_3(x) = 31,399 + 0,339x - 4,104\cdot10^{-3}x^2 - 0,448(x+10)_+ - 9,935\cdot10^{-4}(x+10)_+^2;$$

$$f_3(x) = 31,445 + 0,347x - 4,004\cdot10^{-3}x^2 - 0,559(x+5)_+ - 3,008\cdot10^{-4}(x+5)_+^2;$$

$$f_3(x) = 30,650 + 0,318x - 4,233\cdot10^{-3}x^2 - 0,623(x+0)_+ + 7,395\cdot10^{-4}(x+0)_+^2;$$

$$f_3(x) = 29,113 + 0,254x - 4,798\cdot10^{-3}x^2 - 0,605(x-5)_+ + 1,868\cdot10^{-3}(x-5)_+^2;$$

$$f_3(x) = 27,428 + 0,179x - 5,479\cdot10^{-3}x^2 - 0,519(x-10)_+ + 2,764\cdot10^{-3}(x-10)_+^2;$$

$$f_3(x) = 26,741 + 0,144x - 5,806\cdot10^{-3}x^2 - 0,509(x-15)_+ + 3,740\cdot10^{-3}(x-15)_+^2;$$

$$f_3(x) = 25,906 + 0,097x - 6,279\cdot10^{-3}x^2 - 0,423(x-20)_+ + 4,352\cdot10^{-3}(x-20)_+^2.$$

У цих рівняннях значення абсциси точки перемикання послідовне приймає всі значення в діапазоні [–10; 20] з кроком 5 одиниць. Крім того, слід зазначити, що всі рівняння були отримані з використанням звичайного методу найменших квадратів.

Для отриманих рівнянь були розраховані стандартні відхилення σ, які наведені в табл. 2 з відповідними їм абсцисами точок перемикання x_{sw}.

Таблиця 2

Стандартні відхилення та відповідні їм абсциси точок перемикання

x_{sw}	–10	–5	0
σ	1,287	0,925	0,470

x_{sw}	5	10	15	20
σ	0,0221	0,408	0,695	1,085

Навіть візуальний аналіз даних із табл. 2 показує, що оптимальне значення абсциси точки перемикання повинно знаходитися поблизу значення $x_{sw} = 5$.

Для більш точного визначення оптимального значення абсциси точки перемикання виконаємо апроксимацію даних з використанням звичайної параболи другого порядку за методом найменших квадратів. Унаслідок цього було отримано рівняння:

$$\sigma(x_{sw}) = 0,413 - 0,052x_{sw} + 4,350\cdot10^{-3}{x_{sw}}^2.$$

Обчисливши першу похідну та прирівнявши її нулю, можна визначити оптимальне значення абсциси точки перемикання:

$$x_{sw\,\text{opt}} = 5,926.$$

Підставивши це оптимальне значення абсциси точки перемикання в апроксимуючу функцію для початкових даних із таблиці 1, отримаємо оптимальну двосегментну параболічну регресію наступного вигляду

$$f_{3\,\text{opt}}(x) = 28,793 + 0,240x - 4,924\cdot10^{-3}x^2 - 0,592(x-5,926)_+ + 2,058\cdot10^{-3}(x-5,926)_+^2.$$

Графічне подання апроксимації наведено на рис. 3. У результаті отримане стандартне відхилення склало 0,088. Це значення на порядок краще, ніж у випадку апроксимації поліномом четвертого порядку, який має таку саму кількість невідомих параметрів, як і прийнята в цій роботі двосегментна параболічна регресія.

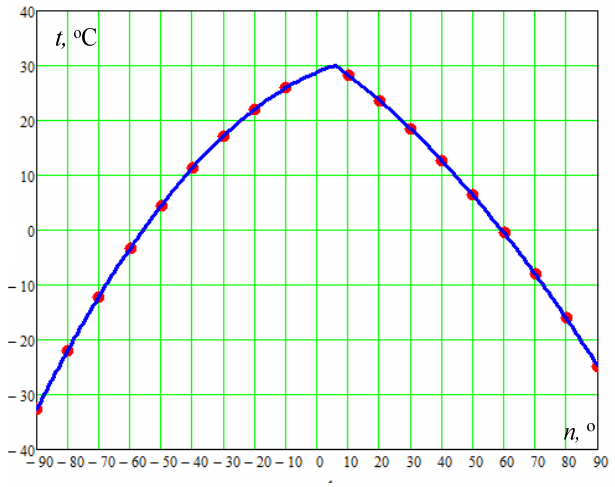


Рис. 3. Оптимальна двосегментна параболічна регресія

Максимальне значення температури, що відповідає точці перемикання становить $t_{max} = 30{,}042$.

Отже, запропонований варіант двосегментної параболічної регресії є найкращим серед розглянутих конкуруючих типів апроксимуючих функцій.

Висновок

Розглянута задача апроксимації емпіричних даних з використанням двосегментної параболічної регресії дозволила розробити більш коректну методику знаходження точки тропічного екватора, яка є дуже важливим параметром під час дослідження змін клімату земної кулі.

Запропонований метод апроксимації заснований на введенні двох додаткових аспектів:

1) використання функції Хевісайда для отримання загального математичного рівняння;

2) визначення оптимальної абсциси точки перемикання.

Такий підхід надав можливість отримання значно меншого значення стандартного відхилення в результаті апроксимації та відповідно більш коректне значення максимальної середньорічної температури, яка відповідає тропічному екватору. Таким чином, запропонована методика з використанням оптимальної двосегментної параболічної регресії може бути використана для щорічного уточнення положення тропічного екватору та відповідної йому середньорічної температури.

ЛІТЕРАТУРА

1. **Химмельблау Д**. Анализ процессов статистическими методами: пер. с англ. / Д. Химмельблау. — М. : Мир, 1973. — 960 с.

2. **Douglas C.** Montgomery, George C. Runger. Applied Statistics and Probability for Engineers, Fours Edition, NJ: John Wiley & Sons, 2007, 768 p.

3. **Mills C**. Frederick. Statistical Methods. New York: Pitman Publishing, 1965, 860 p.

4. **Зайцев Г. Н.** Математическая статистика в экспериментальной ботанике / Г. Н. Зайцев. — М. : Наука, 1984. — 424 с.

5. **Шараф М. А**. Хемометрика: пер. с англ. / М. А. Шараф, М. А. Иллмэн, Б. Р. Ковальски. — Л. : Химия, 1989. — 272 с.

6. *Reklaitis G. V*., Ravindran A., Ragsdell K .M. Engineering optimization. Methods and applications. New York: John Wiley and sons, 1983. — 688 p.

Кузьмин В. М., Заліський М. Ю.

СТАТИСТИЧНИЙ АНАЛІЗ ДАНИХ З ВИКОРИСТАННЯМ ДВОСЕГМЕНТНОЇ ПАРАБОЛІЧНОЇ РЕГРЕСІЇ

У статті розглянуто задачу апроксимації емпіричних даних з використанням двосегментної параболічної регресії. Проведено порівняльний аналіз цього типу апроксимації з іншими типами апроксимуючих функцій (односегментними поліномами другого та четвертого порядків), який дозволив обґрунтувати вибір найкращої математичної моделі. Запропонований метод апроксимації заснований на введенні двох додаткових аспектів: використання функції Хевісайда для отримання загального математичного рівняння та визначення оптимальної абсциси точки перемикання. Для знаходження оптимальної точки перемикання був використаний критерій мінімуму середньоквадратичного відхилення. Визначення невідомих коефіцієнтів апроксимуючих функцій здійснювався на основі використання звичайного методу найменших квадратів.

Ключові слова: апроксимація; двосегментна (параболічна) регресія; оптимізація абсцис точок перемикання; вибір найкращої моделі.

Kuzmin V. N., Zaliskyi M. Yu.

STATISTICAL DATA ANALYSIS WITH SEGMENTED PARABOLIC REGRESSION USAGE

The article deals with the problem of the approximation of empirical data using two-segmented parabolic regression. A comparative analysis of this type of approximation with other types of approximating functions (one-segmented polynomials of the second and fourth degrees) was carried out; this analysis allows substantiating the choice of the best mathematical model. The proposed approximation method is based on the introduction of two additional aspects: the use of the Heaviside function for obtaining a general mathematical equation and determining the optimal abscissa for switching point. In order to find the optimum switching point, the criterion for minimizing standard deviation was used.

The determination of unknown coefficients for approximating functions was carried out according to the ordinary least squares method.

Keywords: approximation; segmented (parabolic) regression; optimization of switching point's abscissa; the best mathematical model .

Кузьмин В. Н., Залисский М. Ю.

СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ДВУСЕГМЕНТНОЙ ПАРАБОЛИЧЕСКОЙ РЕГРЕССИИ

В статье рассматривается задача аппроксимации эмпирических данных с использованием двусегментной параболической регрессии. Проведен сравнительный анализ этого типа аппроксимации с другими типами аппроксимирующих функций (односегментной полиномами второго и четвертого порядков), который позволил обосновать выбор лучшей математической модели. Предложенный метод аппроксимации основан на введении двух дополнительных аспектов: использование функции Хевисайда для получения общего математического уравнения и определение оптимальной абсциссы точки переключения. Для нахождения оптимальной точки переключения был использован критерий минимума среднеквадратичного отклонения. Определение неизвестных коэффициентов аппроксимирующих функций осуществлялось на основе использования обычного метода наименьших квадратов.

Ключевые слова: аппроксимация; двусегментная (параболическая) регрессия; оптимизация абсцисс точек переключения; выбор наилучшей модели.

Стаття надійшла до редакції 18.05.2018 р.
Прийнято до друку 04.06.2018 р.
Рецензент — д-р техн. наук, проф. Конахович Г. Ф.