УДК 004.412:519.25 DOI https://doi.org/10.35546/kntu2078-4481.2025.4.2.18

О. С. ОРЄХОВ
аспірант кафедри програмного забезпечення автоматизованих систем
Національний університет кораблебудування імені адмірала Макарова
ORCID: 0000-0002-0001-0140

Т. А. ФАРІОНОВА
кандидат технічних наук,
доцент кафедри програмного забезпечення автоматизованих систем
Національний університет кораблебудування імені адмірала Макарова
ORCID: 0000-0003-3384-4712

П'ЯТИФАКТОРНА МАТЕМАТИЧНА МОДЕЛЬ ДЛЯ ОБРОБКИ ІНФОРМАЦІЇ З МЕТРИК КОДУ JAVA-ЗАСТОСУНКІВ ДЛЯ ОЦІНЮВАННЯ ЇХ РОЗМІРУ

Достовірна обробка інформації з метрик коду JAVA-застосунків на ранніх стадіях їх проєктування має велике значення, оскільки це безпосередньо впливає на прогнозування трудомісткості розробки. В роботі запропоновано математична модель, а саме п'ятифакторна нелінійна регресійна модель, для раннього оцінювання розміру JAVA-застосунків, а саме кількості рядків коду. Об'єктом дослідження є процес обробки інформації з метрик коду програмних JAVA-застосунків. Предметом дослідження є нелінійна регресійна модель для обробки інформації з метрик коду програмних JAVA-застосунків.

Метою роботи є підвищення достовірності обробки інформації з метрик коду, які доступні з UML-діаграми класів, для оцінювання параметру розміру JAVA застосунків на ранніх етапах проєктування програмного забезпечення шляхом побудови математичної моделі нелінійної регресії з п'ятьма факторами.

Для досягнення поставленої мети зібрано дані за метриками програмного коду 571 загальних JAVA-застосунків з відкритим кодом, які розташовані на платформі GitHub. Отриманий набір даних випадковим чином розділено на навчальну та тестові вибірки розміром 286 та 285 векторів характеристик. В процесі попередньої обробки даних, для побудови математичної моделі нелінійної регресії з п'ятьма факторами, вперше розділено параметр загальної кількості класів та інтерфейсів на окремі метрики та обрані середні значення видимих методів класів та інтерфейсів, полів класів та зв'язності між класами та інтерфейсами, що дозволило уникнути проблем з мультиколінеарністю для побудови регресійної моделі. Нормалізацію багатовимірних даних проведено за допомогою шестивимірного нормалізуючого перетворення Бокса-Кокса. Отримана математична модель має кращі показники якості, а саме R^2, MMRE та PRED(0,25), у порівнянні з існуючими трьохфакторною та чотирьохфакторними нелінійними регресійними моделями для обробки інформації з метрик коду JAVA-застосунків для оцінювання їх розміру.

Ключові слова: *математична модель, обробка інформації, метрика програмного коду, Java, нормалізуюче перетворення, нелінійна регресія, негаусівські дані.*

O. S. ORIEKHOV
Postgraduate Student at the Department of Automated Systems Software
Admiral Makarov National University of Shipbuilding
ORCID: 0000-0002-0001-0140

T. A. FARIONOVA
Candidate of Technical Sciences,
Associate Professor at the Automated Systems Software Department
Admiral Makarov National University of Shipbuilding
ORCID: 0000-0003-3384-4712

FIVE-FACTOR MATHEMATICAL MODEL FOR PROCESSING INFORMATION FROM CODE METRICS OF JAVA APPLICATIONS FOR ESTIMATING THEIR SIZE

Reliable information processing from JAVA application code metrics at the early software design stages is crucial, because it directly impacts software development effort estimation. This paper proposes a mathematical model, specifically a five-factor nonlinear regression model, for the early code lines size estimation of JAVA applications. The object of the

study is the process of information processing from code metrics of JAVA software applications. The subject of the study is a nonlinear regression model for information processing from code metrics of JAVA applications.

The aim of the study is to increase the reliability of information processing from code metrics that are available on the UML class diagram for estimating the size parameter of JAVA applications at the early stages of software design by building a five-factor nonlinear regression mathematical model.

To achieve this goal, data on software code metrics of 571 general open-source JAVA applications hosted on the GitHub platform were collected. The obtained dataset was randomly separated into training and test samples of 286 and 285 vectors, respectively. During data preprocessing for building the mathematical model, the parameter of the total number of classes and interfaces was separated into individual metrics for the first time. Additionally, average values of visible methods of classes and interfaces, class fields, and coupling between classes and interfaces were selected, which allowed avoiding multicollinearity during the regression model construction. Normalization of multidimensional data was performed using a six-dimensional normalizing Box-Cox transformation. The obtained mathematical model demonstrates better quality indicators, namely R^2, MMRE, and PRED(0.25), compared to existing three-factor and four-factor nonlinear regression models for information processing from JAVA application code metrics to estimate their size.

Key words: *mathematical model, information processing, software code metric, Java, normalizing transformation, nonlinear regression, non-Gaussian data.*

Постановка проблеми

Достовірна обробка інформації з метрик коду JAVA-застосунків на ранніх стадіях їх проєктування має велике значення, безпосередньо впливає на прогнозування трудомісткості розробки, та в цілому на терміни, успішність та вартість програмних проєктів [1-4].

Мова програмування JAVA використовується в багатьох сферах, таких як розробка веб-застосунків, наукові обчислення, розробка штучного інтелекту, розробка мобільних застосунків та ігор та є однією з найбільш поширених мов [5,6]. Отже, підвищення достовірності обробки інформації з метрик програмного коду для оцінювання параметру розміру JAVA-застосунків є актуальною науково-практичною задачею, яка потребує вирішення.

Аналіз останніх досліджень і публікацій

Для обробки інформації з метрик коду для оцінювання параметру розміру JAVA-застосунків побудовано як лінійні [7,8] так і нелінійні [9-14] регресійні рівняння та моделі в залежності від різних метрик коду та мов програмування.

В роботах [12-14] доведено, що регресійні рівняння та моделі [7-11] не забезпечують необхідних рівень достовірності оцінювання параметру розміру відповідно до оцінок критеріїв якості нелінійних регресійних моделей, таких як R^2, MMRE, and PRED(0.25) [15]. Трьохфакторна [13] та чотирьохфакторна [14] моделі удосконалюють існуючі трьохфакторні моделі [9], чотирьохфакторну нелінійну регресійну модель [11], та мають кращі показники якості математичної моделі. Але, не дивлячись на збільшення точності обробки інформації з метрик коду для оцінювання параметру розміру, моделі [9,11,13-14] побудовані із використанням метрики RFC, яка має певні складності визначення на ранніх стадіях проєктування ПЗ. Недоліком цієї метрики є те, що вона вимагає інформацію про унікальні виклики всіх залежних методів незалежно від модифікатору доступу до методу (public, protected, private), на виклик певного методу даного класу. Однак, UML діаграма класів не може надати достатньої деталізації для цієї метрики, оскільки в ній відсутня інформація про внутрішні виклики методів, послідовність викликів і алгоритми, які реалізовані в методах [16].

Формулювання мети дослідження

Метою роботи є підвищення достовірності обробки інформації з метрик коду, які доступні з UML-діаграми класів, для оцінювання параметру розміру JAVA застосунків на ранніх етапах проєктування програмного забезпечення шляхом побудови математичної моделі нелінійної регресії з п’ятьма факторами.

Викладення основного матеріалу дослідження

Для досягнення мети дослідження були зібрані дані за метриками програмного коду 571 загальних JAVA-застосунків з відкритим кодом, розташованих на платформі GitHub (https://github.com). За допомогою інструменту CK (https://github.com/mauricioaniche/ck) отримані наступні метрики кількість рядків коду (KLOC), загальна кількість класів та інтерфейсів (CLASS) та їх тип (TYPE), загальна кількість видимих методів (VMQ), загальна кількість атрибутів класів (TFQ), зв’язність між класами (CBO). Метрику CLASS, із застосування метрики типу класів TYPE розділено на дві окремі метрики які представляють фактичну кількість класів (CLS) та фактичну кількість інтерфейсів (INFC) окремо. Отриманий набір даних був розділений випадковим чином на навчальну і тестову вибірки з розмірами в 286 та 285 рядків даних відповідно. Дані навчальної вибірки використовуються для побудови математичної моделі, а тестовий набір даних для незалежного контролю достовірності обробки інформації з метрик коду для оцінювання параметру розміру.

Для побудови нелінійної регресійної моделі для обробки інформації з оцінювання розміру KLOC обрана комбінація метрик CLS, INFC, VMQ, TFQ та CBO.

Багатовимірні дані навчальної вибірки були перевірені на наявність мультиколінеарності за коефіцієнтами впливу дисперсії (VIFs) між незалежними факторами. Для багатовимірних даних з k незалежними факторами (X_i), $i = 1,2,...,k$ коефіцієнти VIFs представлені діагональними елементами оберненої коваріаційної $k \times k$ матриці.

Якщо значення коефіцієнту VIF перевищує 10, то це свідчить про наявність проблем із мультиколінеарністю [17]. Для факторі CLS, INFC, VMQ, TFQ та CBO значення коефіцієнтів VIFs дорівнюють 16,4, 2,3, 13,2, 10,8 та 20,9 відповідно, що свідчить про наявність значної мультиколінеарності між незалежними факторами. Для вирішення цієї проблеми, абсолютні значення метрик VMQ, TFQ та CBO замінені на їх середні значення відносно суми загальної кількості класів та інтерфейсів з програмного проєкту – aVMQ, aTFQ та aCBO відповідно. Для незалежних факторів CLS, INFC, aVMQ, aTFQ та aCBO розраховані значення оцінок коефіцієнтів VIFs дорівнюють 2,2, 2,3, 1,1, 1,2 та 1,1, що свідчить про відсутність мультиколінеарності між незалежними факторами. Характеристики розподілу навчальної та тестової вибірок наведені в таблиці 1 та таблиці 2.

Таблиця 1

Характеристики розподілу факторів навчальної вибірки

Параметри розподілу вибірки	KLOC	CLS	INFC	aVMQ	aTFQ	aCBO
Середнє	68,813	1158,699	118,210	5,194	2,276	5,937
Медіана	29,746	601,000	46,000	4,986	2,123	5,961
Середньоквадратичне відхилення	106,011	1626,298	219,993	2,097	0,947	1,628
Асиметрія	3,205	2,924	4,612	2,438	0,917	0,054
Ексцес	12,133	9,943	29,786	12,527	1,485	0,930
Мінімальне значення	1,586	33	0	1,705	0,133	0,118
Максимальне значення	699,674	10610	2128	21,055	6,337	12,590

Таблиця 2

Характеристики розподілу факторів навчальної вибірки

Параметри розподілу вибірки	KLOC	CLS	INFC	aVMQ	aTFQ	aCBO
Середнє	74,136	1136,039	103,263	5,061	2,387	6,017
Медіана	27,032	527,000	35,000	4,696	2,173	5,847
Середньоквадратичне відхилення	117,831	1610,205	173,496	2,046	1,134	1,487
Асиметрія	2,870	2,738	3,200	1,751	1,820	0,268
Ексцес	9,308	8,687	13,005	5,331	6,614	0,508
Мінімальне значення	1,201	29	0	1,589	0,017	2,444
Максимальне значення	777,036	9475	1245	16,995	9,160	11,768

Для оцінювання параметру кількості рядків коду існують різні підходи із застосуванням математичних моделей, у тому числі регресійних. Зазвичай дані за метриками програмного коду не розподілені за нормальним законом, що робить обмежену можливість використання лінійних регресійних моделей для оцінювання розміру рядків коду. Теоретичною умовою застосування лінійних регресійних моделей є нормальний розподіл залишків регресії ε або нормальний розподіл багатовимірних даних.

Дані навчальної вибірки перевірено на відповідність нормальному розподілу даних із застосуванням багатовимірного тесту Мардія [18]. Згідно з цим тестом виявлено, що розподіл шестивимірних даних CLS, INFC, aVMQ, aTFQ, aCBO і KLOC навчальної вибірки не є нормальним, оскільки оцінка багатовимірної асиметрії $N\beta_1 / 6 = 5134,45$ перевищує значення квантила $\chi^2 = 86,99$ для 56 ступенів свободи та рівня значущості $\alpha = 0,005$, а оцінка багатовимірного ексцесу $\beta_2 = 219,63$ перевищує значення квантилю розподілу Гауса, яке дорівнює 50,77 для математичного сподівання 48 і дисперсії 1,16. Виникає необхідність побудови нелінійної регресійної моделі для обробки інформації з оцінювання параметру розміру JAVA-застосунків.

Застосування нормалізуючих перетворень дозволяє перейти від нелінійної регресійної моделі до побудови лінійної регресійної моделі.

Введемо наступні позначення для факторів: X_1 - CLS, X_2 – INFC, X_3 – aVMQ, X_4 – aTFQ, X_5 – aCBO, Y – KLOC.

Згідно з запропонованим підходом [19,20], процес побудови математичної моделі є ітеративним та включає шість етапів.

На **першому етапі** застосовується нормалізуюче перетворення негаусівського випадкового вектору $P = \{Y, X_1, X_2, ..., X_5\}^T$ у гаусівський випадковий вектор $T = \{Z_Y, Z_1, Z_2, ..., Z_5\}^T$:

$$T = \psi(P). \tag{1}$$

Обернене перетворення до (1) має вигляд

$$P = \psi^{-1}(T), \tag{2}$$

де ψ – вектор взаємозворотніх функцій нормалізуючого перетворення, $\psi = \{\psi_Y, \psi_1, \psi_2, ..., \psi_5\}^T$.

На **другому етапі** виконується перевірка багатовимірних даних на відповідність нормальному розподілу за допомогою критерію Мардія [18], в залежності від параметрів багатовимірної асиметрії ($\beta_{1,5}$) та ексцесу ($\beta_{2,5}$). Визначення та вилучення викидів із багатовимірних нормалізованих даних здійснюється за квадратом відстані Махаланобіса. Якщо в багатовимірному негаусівському наборі даних є багатовимірний викид, то таку точку відкидають і відбувається перехід до першого етапу, інакше до наступного етапу.

На **третьому етапі**, для нормалізованих даних за перетворенням (1), виконується побудова лінійної регресійної моделі, яка має вигляд

$$\hat{Z}_y + \varepsilon = b_0 + b_1 Z_1 + b_2 Z_2 + ... + b_5 Z_5 + \varepsilon\,, \tag{3}$$

де ε – випадкова величина розподілена за нормальним законом, $\varepsilon \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$; $b_0, b_1, b_2, ..., b_5$ – параметри лінійного регресійного рівняння (3), які знаходяться за методом найменших квадратів.

На **четвертому етапі** залишки регресії ε побудованої лінійної регресійної моделі (3) перевіряються на відповідність нормальному закону розподілу. Якщо розподіл залишків ε лінійної регресійної моделі для нормалізованих даних не є нормальним, тоді необхідно відкинути точку даних, для якої модуль залишку є максимальним і повернутись до першого кроку. Перевірка нормальності розподілу залишків регресії здійснюється за критерієм Пірсона χ^2.

На **п'ятому етапі**, після побудови моделі лінійної регресії (3) для нормалізованих даних, застосувавши зворотне перетворення (2), нелінійна регресійна модель має наступний вигляд

$$Y = \psi_Y^{-1}\left(\hat{Z}_y + \varepsilon\right) = \psi_Y^{-1}\left(b_0 + b_1 Z_1 + b_2 Z_2 + ... + b_5 Z_5 + \varepsilon\right) \tag{4}$$

На **шостому етапі** побудови нелінійної регресійної моделі (4), розраховуються верхня та нижня границі інтервалів передбачення та проводиться пошук точок значень, які знаходяться поза їх межами. Якщо такі точки знайдені, то вони вважаються викидами і вилучаються із навчальної вибірки. Границі інтервалів передбачення нелінійної регресійної моделі визначаються за наступною формулою

$$\hat{Y}_{PI} = \psi_Y^{-1}\left(\hat{Z}_Y \pm t_{\alpha/2,\nu} S_{Z_Y} \left\{1 + \frac{1}{N} + \left(Z_X^+\right)^T S_Z^{-1}\left(Z_X^+\right)\right\}^{1/2}\right), \tag{5}$$

де $t_{a/2,\nu}$ – квантиль t-розподілу Стьюдента з $\nu = N - 5 - 1$ ступенями свободи та $\alpha/2$ – рівнем значущості; $S_{Z_Y}^2 = \frac{1}{\nu}\sum_{i=1}^{N}\left(Z_{Y_i} - \hat{Z}_{Y_i}\right)^2$; Z_X^+ – вектор центральних факторів(регресорів), який містить значення $\left\{Z_{1_i} - \bar{Z}_1, Z_{2_i} - \bar{Z}_2, ..., Z_{5_i} - \bar{Z}_5\right\}$; S_Z – 5×5 матриця

$$S_Z = \left[S_{Z_q} S_{Z_r}\right], \tag{6}$$

де $S_{X_q} S_{X_r} = \sum_{n-1}^{N}\left(Z_{q_i} - \bar{Z}_q\right)\left(Z_{r_i} - \bar{Z}_r\right)$, $q, r = 1, 2, \ldots, 5$.

Для нормалізації даних було обрано шестивимірне нормалізуюче перетворення Бокса-Кокса:

$$Z_j = \psi\left(X_j\right) = \begin{cases} \left(X_j^{\lambda_j} - 1\right)/\lambda_j, & якщо\, \lambda_j \neq 0 \\ ln\left(X_j\right), & якщо\, \lambda_j = 0 \end{cases}, \tag{7}$$

де, j змінюється від 1 до 6; X_j – випадкова негаусівська величина, яка нормалізується; Z_j – нормалізована гаусівська величина; λ_j – параметр перетворення Бокса-Кокса.

Для перетворення Бокса-Кокса, логарифмічна функція правдоподібності оцінювання параметрів перетворення [21] має вигляд

$$l\left(X, \theta\right) = \sum_{j=1}^{k}\left(\lambda_j - 1\right)\sum_{i=1}^{N} ln\left(X_{ji}\right) - \frac{N}{2} ln\left[det\left(S_N\right)\right], \tag{8}$$

де $l\left(P, \theta\right)$ – логарифмічна функція правдоподібності; θ – вектор параметрів перетворення, $\theta = \left\{\lambda_1, \lambda_2, ..., \lambda_6\right\}$; N – розмір вибірки.

S_N – вибіркова коваріаційна матриця для компонентів вектору Z:

$$S_N = \frac{1}{N}\sum_{i=1}^{N}\left(Z_i - \bar{Z}\right)\left(Z_i - \bar{Z}\right)^T, \tag{9}$$

де Z_i – гаусівський випадковий вектор, $Z_i = \{Z_{1i}, Z_{2i}, ..., Z_{6i}\}^T$; $\bar{Z}$ – вектор вибіркових середніх, $\bar{Z} = \{\bar{Z}_1, \bar{Z}_2, ..., \bar{Z}_6\}$.

Застосування нормалізуючого перетворення Бокса-Кокса вимагає збільшення метрики кількості інтерфейсів INFC на 1, оскільки це перетворення існує тільки для додатних чисел.

П'ятифакторну нелінійну регресійну модель побудовано за 26 ітерацій. При побудові моделі виключено 23 багатовимірні точки за квадратом відстані Махаланобіса та 2 багатовимірні точки були за межами інтервалу передбачення. Для останньої ітерації значення тестової статистики розподілу для багатовимірної асиметрії $B_1 N / 6$, яка дорівнює 81,82, що не перевищує значення квантилю розподілу χ^2 -квадрат, що становить 86,99 для 56 ступенів свободи та рівня значущості 0,005 та тестова статистика розподілу даної вибірки для багатовимірного ексцесу B_2, яка становить 47,60, не перевищує значення квантиля нормального розподілу, яке становить 50,84 для математичного сподівання 48 і дисперсії 1,21 та рівня значущості α 0,005, що свідчить про нормальну природу розподілу нормалізованих даних без викидів.

Оцінки параметрів багатовимірного нормалізуючого перетворення (7) отримані шляхом максимізації логарифмічної функції правдоподібності (8), та для залежного фактора Y і незалежних факторів X_1, X_2, X_3, X_4 та X_5 мають наступні значення $\hat{\lambda}_Y = 1{,}629679 \cdot 10^{-2}$, $\hat{\lambda}_{X_1} = 4{,}365094 \cdot 10^{-2}$, $\hat{\lambda}_{X_2} = 0{,}102789$, $\hat{\lambda}_{X_3} = -0{,}221238$, $\hat{\lambda}_{X_4} = 0{,}354485$ і $\hat{\lambda}_{X_5} = 0{,}894651$ відповідно для останньої ітерації. Оцінки параметрів п'ятифакторної лінійної регресійної моделі на основі нормалізованих даних, які отримані за методом найменших квадратів, становлять $\hat{b}_0 = -4{,}117734$, $\hat{b}_1 = 0{,}809663$, $\hat{b}_2 = -4{,}599656 \cdot 10^{-3}$, $\hat{b}_3 = 1{,}175418$, $\hat{b}_4 = 0{,}252310$ та $\hat{b}_5 = -1{,}608608 \cdot 10^{-2}$.

П'ятифакторна лінійна регресійна модель для нормалізованих даних має наступний вигляд:

$$\hat{Z}_Y + \varepsilon = \hat{b}_0 + \hat{b}_1 \psi_1(X_1) + \hat{b}_2 \psi_2(X_2 + 1) + \hat{b}_3 \psi_3(X_3) + \hat{b}_4 \psi_4(X_4) + \hat{b}_5 \psi_5(X_5) + \varepsilon \tag{10}$$

де $\psi_j(X_j) = \left(X_j^{\hat{\lambda}_{X_j}} - 1\right) / \hat{\lambda}_{X_j}$, – нормалізуюче перетворення Бокса-Кокса із значенням оцінок перетворення для незалежного фактору $X_j\,(j = 1, ..., 5)$. П'ятифакторна нелінійна регресійна модель $\hat{Y}$ має наступний вигляд

$$\hat{Y} = \left[\hat{\lambda}_Y \left(\hat{Z}_Y + \varepsilon\right) + 1\right]^{\frac{1}{\hat{\lambda}_Y}}, \tag{11}$$

де $\hat{Z}_Y$ – оцінка значення залежної змінної для нормалізованих даних за рівнянням (10), $\hat{Y}$ – оцінка значень залежної змінної для початкових даних.

Інтервал передбачення нелінійної регресійної моделі $\hat{Y}$ (11) заданий як

$$\hat{Y}_{PI} = \psi_Y^{-1}\left(\hat{Z}_Y \pm \hat{t}_{\alpha/2,\nu} \hat{S}_{Z_Y} \left\{1 + \frac{1}{N} + \left(Z_X^+\right)^T \hat{S}_Z^{-1} \left(Z_X^+\right)\right\}^{1/2}\right), \tag{12}$$

де $\hat{Z}_Y$ – оцінка параметру розміру для нормалізованих даних за (10).

Для останньої ітерації значення нормалізованих вибіркових середніх $\bar{Z}_1$, $\bar{Z}_2$, $\bar{Z}_3$, $\bar{Z}_4$ та $\bar{Z}_5$ становлять 7,411555, 4,778777, 1,312994, 0,882226, та 4,430020 відповідно. Квантиль t -розподілу Стьюдента $\hat{t}_{\alpha/2,\nu} = 2{,}5952$ для рівня значущості $\alpha = 0{,}01$ та 255 ступенів свободи; $\hat{S}_{Z_Y} = 0{,}171633$. Обернена коваріаційна матриця (6) має вигляд

$$\hat{S}_Z^{-1} = \begin{vmatrix} 8{,}0315 \cdot 10^{-3} & -5{,}2507 \cdot 10^{-3} & -2{,}5413 \cdot 10^{-3} & -1{,}0127 \cdot 10^{-3} & -5{,}4032 \cdot 10^{-4} \\ -5{,}2507 \cdot 10^{-3} & 4{,}2867 \cdot 10^{-3} & 3{,}4181 \cdot 10^{-4} & 1{,}2176 \cdot 10^{-3} & 3{,}2418 \cdot 10^{-4} \\ -2{,}5413 \cdot 10^{-3} & 3{,}4181 \cdot 10^{-4} & 1{,}0058 \cdot 10^{-1} & -1{,}2670 \cdot 10^{-2} & -5{,}2167 \cdot 10^{-3} \\ -1{,}0127 \cdot 10^{-3} & 1{,}2176 \cdot 10^{-3} & -1{,}2670 \cdot 10^{-2} & 1{,}9085 \cdot 10^{-2} & -2{,}0527 \cdot 10^{-3} \\ -5{,}4032 \cdot 10^{-4} & 3{,}2418 \cdot 10^{-4} & -5{,}2167 \cdot 10^{-3} & -2{,}0527 \cdot 10^{-3} & 3{,}3910 \cdot 10^{-3} \end{vmatrix}$$

Зменшення інтервалів передбачення можна досягти шляхом застосування квантилю t -розподілу Стьюдента $\hat{t}_{\alpha/2,\nu}$ з рівнем значущості $\alpha = 0{,}05$ та 255 ступенів свободи, який становить 1,9693.

Побудовану модель (11) перевірено за критеріями якості R^2, $MMRE$ та $PRED(0{,}25)$. Для отриманої п'ятифакторної нелінійної регресійної моделі $R^2 = 0{,}9647$, $MMRE = 0{,}1305$ та $PRED(0{,}25) = 0{,}8927$, що свідчить про високий рівень достовірності моделі для оцінювання параметру розміру JAVA-застосунків.

Проведемо порівняльний аналіз оцінок показників якості побудованої моделі з іншими моделями [13,14,22] на навчальній і тестовій вибірках JAVA-застосунків (див. табл 3, 4).

Таблиця 3

Порівняльний аналіз оцінок показників якості нелінійних регресійних моделей для навчальної вибірки

№	Нелінійна регресійна модель	Навчальна вибірка		
		R^2	MMRE	PRED
1	Трьохфакторна на базі перетворення Джонсона [13]	0,9179	0,1560	0,7867
2	Чотирьохфакторна на основі багатовимірного перетворення Бокса-Кокса [14]	0,9067	0,1526	0,8181
3	Чотирьохфакторна на основі багатовимірного перетворення Бокса-Кокса [22]	0,9150	0,1717	0,8146
4	Побудована п'ятифакторна нелінійна регресійна модель (11)	0,9294	0,1540	0,8462

Таблиця 4

Порівняльний аналіз оцінок показників якості нелінійних регресійних моделей для тестової вибірки

№	Нелінійна регресійна модель	Тестова вибірка		
		R^2	MMRE	PRED
1	Трьохфакторна на базі перетворення Джонсона [13]	0,9094	0,1583	0,8315
2	Чотирьохфакторна на основі багатовимірного перетворення Бокса-Кокса [14]	0,9052	0,1492	0,8421
3	Чотирьохфакторна на основі багатовимірного перетворення Бокса-Кокса [22]	0,9075	0,1854	0,7439
4	Побудована п'ятифакторна нелінійна регресійна модель (11)	0,9193	0,1742	0,7825

Отже, відповідно до отриманих оцінок параметрів якості нелінійних регресійних моделей для початкової навчальної та тестової вибірок JAVA-застосунків (табл. 3 та табл. 4), п'ятифакторна нелінійна регресійна модель (11) на базі шестивимірного перетворення Бокса-Кокса показала найкращу точність оцінювання серед запропонованих моделей, які дозволяють оцінити параметр розміру JAVA-застосунків.

Висновки

1. Вирішено задачу підвищення достовірності обробки інформації з метрик коду, які доступні на UML-діаграмі класів, для оцінювання параметру розміру JAVA застосунків на ранніх етапах проєктування програмного забезпечення шляхом побудови математичної моделі нелінійної регресії з п'ятьма факторами.

2. Наукова новизна отриманих результатів полягає в тому, що вперше побудована п'ятифакторна нелінійна регресійна модель на основі шестивимірного перетворення Бокса-Кокса із використанням метрик загальної кількості класів (CLS), загальної кількості інтерфейсів (INFC), середніх значення метрик видимих методів класів (aVMQ), полів класів (aTFQ) та значення зв'язності між класами (aCBO) відносно суми загальної кількості класів та інтерфейсів. Розділення метрики загальної кількості класів та інтерфейсів (CLASS) на 2 окремі метрики дозволило підвищення достовірності оцінювання параметру розміру із використанням п'ятифакторної моделі для загальних JAVA-застосунків.

3. Практичне значення отриманих результатів полягає у використанні побудованої п'ятифакторної нелінійної регресійної моделі для оцінювання параметру розміру із подальшим використанням у відповідних моделях оцінювання трудомісткості розробки програмного забезпечення.

Список використаної літератури

1. Boehm B., Abts C., Brown A.W., Chulani S., Clark B.K., Horowitz E., Madachy R., Reifer D., Steece B. Software cost estimation with COCOMO II. Prentice Hall. 2000.

2. Munialo S.W. A Review of Agile Software Effort Estimation Methods // *International Journal of Computer Applications Technology and Research. Association of Technology and Science*. 2016. Vol. 5. pp. 612–618. DOI:10.7753/IJCATR0509.1009.

3. Johnson J., Mulder H. Endless modernization: How infinite flow keeps software fresh. ResearchGate. 2021. URL: https://www.researchgate.net/publication/348849361_Endless_Modernization_How_Infinite_Flow_Keeps_Software_Fresh

4. The Standish Group. Chaos report 2015. 2015. URL: https://standishgroup.com/sample_research_files/CHAOSReport2015-Final.pdf

5. Oracle. Java. 2025. URL: https://www.oracle.com/java/

6. Cass S. Top programming languages 2024. IEEE Spectrum. 2024. URL: https://spectrum.ieee.org/top-programming-languages-2024

7. Tan H.B.K., Zhao Y., Zhang H. Estimating LOC for information systems from their conceptual data models // *Proceedings – International Conference on Software Engineering*. 2006. pp. 321-330. DOI:10.1145/1134285.1134331.

8. Tan H.B.K., Zhao Y., Zhang H. Conceptual Data Model-Based Software Size Estimation for Information Systems // *ACM Transactions of Software Engineering and Methodology*. 2009. Vol. 19. DOI:10.1145/1571629.1571630.

9. Приходько Н.В., Приходько С.Б. Нелінійна регресійна модель для оцінювання розміру програмного забезпечення промислових інформаційних систем на Java // *Моделювання та інформаційні технології.* 2018. Вип. 85. С. 81–88. URL: http://nbuv.gov.ua/UJRN/Mtit_2018_85_14

10. Макарова Л.М., Приходько Н.В., Кудін О.О. Побудова нелінійної регресійної моделі для оцінювання розміру веб-додатків, реалізованих мовою Java // *Вісник Херсонського національного технічного університету*. 2019. № 2 (69). С. 145–153. URL: http://eir.nuos.edu.ua/handle/123456789/4443

11. Приходько С.Б., Приходько Н.В., Смикодуб Т.Г. Чотирьохфакторна нелінійна регресійна модель для оцінювання розміру JAVA-застосунків з відкритим кодом // *Вчені записки ТНУ імені В.І. Вернадського. Серія: технічні науки* Том 31 (70) № 2 Частина 1. 2020. С. 157–162. DOI:10.32838/2663-5941/2020.2-1/25

12. Орєхов О.С., Фаріонова Т.А. Математичні моделі для оцінювання розміру JAVA-застосунків // *Вісник Херсонського національного технічного університету.* 2024. Т. 89, № 2. С. 196–203. DOI:10.35546/kntu2078-4481.2024.2.28.

13. Oriekhov O., Farionova T., Chernova L. Three-factor nonlinear regression model of estimating the size of JAVA-software // *Proceedings of the International Conference on Information Control Systems and Technologies (ICST-2024)*. 2024. pp. 506–518. URL: https://ceur-ws.org/Vol-3790/paper44.pdf

14. Орєхов О.С. Чотирьохфакторна нелінійна регресійна модель для оцінювання розміру JAVA-застосунків на ранніх стадіях розробки // *Розвиток інформаційно-керуючих систем та технологій.* 2024. С. 360–379. DOI:10.36059/978-966-397-422-4

15. Port D., Korte M. Comparative studies of the model evaluation criterions MMRE and PRED in software cost estimation research // *Proceedings of the 2nd ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, New York, 2008. pp. 51–60. DOI:10.1145/1414004.1414015

16. Subramanyam R., Krishnan M. Empirical analysis of CK metrics for object-oriented design complexity: Implications for software defects // *IEEE Transactions on Software Engineering*. 2003. Vol. 29, No. 4. pp. 297–310. DOI:10.1109/TSE.2003.1191795.

17. Olkin I., Sampson A.R. Multivariate Analysis: Overview // *International encyclopedia of social & behavioral sciences (eds.) 1st edn.*, Elsevier, Pergamon, 2001. pp. 10240–10247

18. Mardia K. V., Measures of multivariate skewness and kurtosis with applications // *Biometrika*. 1970. Vol. 57. pp. 519–530. DOI:10.1093/biomet/57.3.519.

19. Prykhodko S., Prykhodko N., Mathematical Modeling of Non-Gaussian Dependent Random Variables by Nonlinear Regression Models Based on the Multivariate Normalizing Transformations // *Mathematical Modeling and Simulation of Systems (MODS'2020). Advances in Intelligent Systems and Computing*. 2021. Vol. 1265. PP. 166-174. DOI:10.1007/978-3-030-58124-4_16

20. Prykhodko S., Prykhodko N., Makarova L., Pukhalevych A. Outlier Detection in Non-Linear Regression Analysis Based on the Normalizing Transformations // *2020 IEEE 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET)*. Lviv-Slavske, Ukraine, 2020. pp. 407–410, DOI:10.1109/TCSET49122.2020.235464.

21. Lindsey C., Sheather S. Power transformation via multivariate Box–Cox // *The Stata Journal*. 2010. Vol. 10, No. 1. pp. 69–81. DOI:10.1177/1536867X1001000108.

22. Oriekhov O., Farionova T., Chernova L., Vorona M. A Five-Factor Nonlinear Regression Model for JAVA Applications Size Estimation // *Proceedings of VI International Workshop on IT Project Management*. 2025. Vol. 4023. pp. 1–10. URL: https://ceur-ws.org/Vol-4023/paper1.pdf

References

1. Boehm, B., Abts, C., Brown, A. W., Chulani, S., Clark, B. K., Horowitz, E., Madachy, R., Reifer, D., & Steece, B. (2000). *Software cost estimation with COCOMO II*. Prentice Hall.

2. Munialo, S.W., & Muketha, G.M. (2016). A review of Agile Software effort estimation methods. *International Journal of Computer Applications Technology and Research*, 5 (9), 612–618. https://doi.org/10.7753/ijcatr0509.1009

3. Johnson, J., & Mulder, H. (2021, Jan). *Endless modernization: How infinite flow keeps software fresh.* ResearchGate. https://www.researchgate.net/publication/348849361_Endless_Modernization_How_Infinite_Flow_Keeps_Software_Fresh

4. The Standish Group. (2015). *Chaos report 2015.* https://standishgroup.com/sample_research_files/CHAOSReport2015-Final.pdf

5. Oracle. (2025, November 10). Java. Retrieved from https://www.oracle.com/java/

6. Cass, S. (2024, August 22). Top programming languages 2024. IEEE Spectrum. https://spectrum.ieee.org/top-programming-languages-2024

7. Tan, H. B. K., Zhao, Y., & Zhang, H. (2006). Estimating LOC for information systems from their conceptual data models. In *Proceedings of the 28th International Conference on Software Engineering* (pp. 321–330). Association for Computing Machinery. https://doi.org/10.1145/1134285.1134331

8. Tan, H. B. K., Zhao, Y., & Zhang, H. (2009). Conceptual data model-based software size estimation for information systems. *ACM Transactions of Software Engineering and Methodology, 19*(2), 1-37. https://doi.org/10.1145/1571629.1571630

9. Prykhodko, N.V. & Prykhodko S.B. (2018). A nonlinear regression model for estimation of the size of Java enterprise information systems software. *Modeling and Information Technologies*, Vol. 85, 81–88. URL: http://nbuv.gov.ua/UJRN/Mtit_2018_85_14 [in Ukrainian]

10. Makarova L.M., Prykhodko N.V. & Kudin O.O. (2019). Constructing the non-linear regression model for size estimation of WEB-applications implemented in JAVA. *Bulletin of Kherson National Technical University*, Vol. 69, P. 145–153. URL: http://eir.nuos.edu.ua/handle/123456789/4443 [in Ukrainian]

11. Prykhodko, S.B., Prykhodko, N.V. & T. G. Smykodub. (2020). Four-factor non-linear regression model to estimate the size of open source Java-based applications, *Scientific Notes of Taurida National V.I. Vernadsky University. Series: Technical Sciences*, Vol. 70, 157-162. https://doi.org/10.32838/2663-5941/2020.2-1/25 [in Ukrainian]

12. Oriekhov, O., & Farionova, T. (2024). Mathematical models for the size estimating of JAVA applications. *Herald (Kherson National Technical University), 89*, 196–203. https://doi.org/10.35546/kntu2078-4481.2024.2.28

13. Oriekhov, O., Farionova, T., & Chernova, L. (2024), Three-factor nonlinear regression model of estimating the size of JAVA-software. In *Proceedings of the International Conference on Information Control Systems and Technologies (ICST-2024)*, September 23-25, 2024. CEUR Workshop Proceedings. pp. 506-518. https://ceur-ws.org/Vol-3790/paper44.pdf

14. Oriekhov, O. (2024). The four-factor nonlinear regression model for early JAVA-applications size estimation. In *ICST-2024: Advances in Information Control Systems and Technologies*, 360–379. https://doi.org/10.36059/978-966-397-422-4

15. Port, D., & Korte, M. (2008). Comparative studies of the model evaluation criterions MMRE and PRED in software cost estimation research. In *Proceedings of the 2nd ACM-IEEE International Symposium on Empirical Software Engineering and Measurement* (pp. 51–60). Association for Computing Machinery. https://doi.org/10.1145/1414004.1414015

16. Subramanyam, R., & Krishnan, M. (2003). Empirical analysis of CK metrics for object-oriented design complexity: Implications for software defects. *IEEE Transactions on Software Engineering, 29*(4), 297–310. https://doi.org/10.1109/TSE.2003.1191795

17. Olkin, I., & Sampson, A. R. (2001). Multivariate analysis: Overview. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social & behavioral sciences* (1st ed., pp. 10240–10247). Elsevier.

18. Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*(3), 519–530. https://doi.org/10.1093/biomet/57.3.519

19. Prykhodko, S., & Prykhodko, N. (2021). Mathematical modeling of non-Gaussian dependent random variables by nonlinear regression models based on the multivariate normalizing transformations. In S. Shkarlet, V. Morozov, & A. Palagin (Eds.), *Mathematical modeling and simulation of systems (MODS'2020)* (pp. 166–174). Springer, Cham. https://doi.org/10.1007/978-3-030-58124-4_16

20. Prykhodko, S., Prykhodko, N., Makarova, L., & Pukhalevych, A. (2020). Outlier detection in non-linear regression analysis based on the normalizing transformations. In *Proceedings of 2020 IEEE 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET)* (с. 407–410). IEEE. https://doi.org/10.1109/TCSET49122.2020.235464

21. Lindsey, C., & Sheather, S. (2010). Power transformation via multivariate Box–Cox. *The Stata Journal*, 10(1), 69–81. https://doi.org/10.1177/1536867X1001000108

22. Oriekhov, O., Farionova, T., Chernova, L., & Vorona, M. (2025). A Five-Factor Nonlinear Regression Model for JAVA Applications Size Estimation. In *Proceedings of VI International Workshop on IT Project Management*, Kyiv, Ukraine, May 22, 2025. CEUR Workshop Proceedings, Vol. 4023, pp. 1-10. https://ceur-ws.org/Vol-4023/paper1.pdf

Дата першого надходження рукопису до видання: 15.11.2025
Дата прийнятого до друку рукопису після рецензування: 12.12.2025
Дата публікації: 31.12.2025