УДК 519.652:519.254
DOI https://doi.org/10.32851/tnv-tech.2021.6.3

ДОСЛІДЖЕННЯ ІМІТАЦІЇ ОДНОВИМІРНИХ ВИБІРОК ІЗ ВИКОРИСТАННЯМ ПОЛІНОМІАЛЬНИХ СПЛАЙНІВ

Зівакін В.Д. *– молодший науковий співробітник,*
аспірант кафедри прикладної математики
Національного авіаційного університету
ORCID ID: 0000-0002-0420-0558

Використання моделювання для вирішення різноманітних завдань обумовлено низкою причин: економією часових і матеріальних ресурсів, імітацією «критичних» режимів, що в умовах реальної експлуатації може бути небезпечно для досліджуваного об'єкта, можливістю дистанційного тренінгу тощо. Зокрема, моделювання не роботи деякої системи, а саме послідовностей даних певного вигляду могло б вирішити проблему нестачі таких даних (наприклад, у машинному навчанні), що є актуальним в разі роботи у багатовимірних просторах.

За імітаційного моделювання вибірок перше, з чого необхідно починати, – це модель розподілу, яку необхідно отримати. Модель може бути визначена деяким аналітичним законом розподілу (нормальним, Вейбула, рівномірним тощо), і в цьому вона залежить від параметрів (параметрична модель). Зазвичай обирають такі моделі, щоб їхні параметри мали деяку змістовну інтерпретацію (a, b – початок та кінець інтервалу в рівномірному розподілі; λ – інтенсивність в експоненціальному тощо). Іншим класом моделей, що відтворюють функції розподілу, є непараметричні (ядерні методи, гістограмні оцінки емпіричної функції розподілу, сплайн-апроксимація). Основною проблемою методів, які ґрунтуються на параметрах, є обмеженість, що особливо простежується в двох випадках:

1. За моделювання багатовимірних даних – у цьому випадку робота завжди призводить до переходу до багатовимірного нормального розподілу.

2. За моделювання неоднорідних вибірок, які є сумішшю декількох розподілів (необов'язково з одного класу), усічених або тих, що містять пропуски спостережень.

У такому контексті використання параметричних моделей об'єктивно є неможливим у чистому вигляді. Отже, наявність інструменту, який добре апроксимує неоднорідні дані, є бажаною для вирішення завдання генерації неоднорідних багатовимірних сукупностей.

Ключові слова: *моделювання, імітація, дані, сплайн, щільність розподілу, гістограмна оцінка, апроксимація.*

Zivakin V.D. Research of simulation of one-dimensional samples using polynomial splines

The use of modeling to solve various problems is due to a set of reasons: saving time and material resources, simulation of "critical" modes, which in real operation can be dangerous for the object under study, the possibility of distance learning and others. In particular, modeling not the operation of a system, but data sequences of a certain type could solve the problem of lack of such data (for example, in machine learning), which is relevant especially in the case of working in multidimensional spaces.

When simulating samples, the first thing to start from is the distribution model to be obtained. The model can be determined by some analytical law of distribution (normal, Weibull, uniform, etc.), and in this it depends on the parameters (parametric model). Models are usually chosen so that their parameters carry some meaningful interpretation (a, b – the beginning and end of the interval in a uniform distribution, λ – intensity in exponential, etc.). Another class of models that reproduce distribution functions are nonparametric (nuclear methods, histogram estimates of the empirical distribution function, spline approximation). The main problem with parameter-based methods is limited, especially in two cases:

1. When modeling multidimensional data – in this case, the work always leads to a transition to multidimensional normal distribution.

2. When modeling inhomogeneous samples, which are a mixture of several distributions (not necessarily from one class), truncated or those that contain gaps in observations.

In this context, the use of parametric models is objectively impossible in its purest form. Therefore, the presence of a tool that well approximates inhomogeneous data is desirable to solve the problem of generating inhomogeneous multidimensional sets.

Key words: *modeling, simulation, data, spline, distribution density, histogram estimation, approximation.*

Окремим видом імітаційного моделювання послідовностей даних є машинне навчання, а саме нейронні мережі (наприклад, рекурентні) [1]. Але для реалізації всього процесу навчання окремої штучної нейронної мережі необхідні не тільки час та численні потужності, але й великі набори даних для навчання, водночас як одна з цілей цього дослідження – проведення імітації за нестачі таких. Більш класичні методи (як параметричні, так і непараметричні) поділяються на ті, що діють на основі оцінки функції розподілу, та на ті, що базуються на оцінці функції щільності. До першої групи належить метод зворотної функції [3; 4], який є ефективним, коли процес отримання зворотної функції не є складним. Окрім того, саме у багатовимірному просторі доречне оцінювання функції щільності, адже це є простішим завданням, аніж оцінка функції розподілу. Методів, які належать до останньої групи, значно менше. Це метод суперпозиції [2; 4], який досить складно реалізувати програмно, та метод винятків [2], який через простоту реалізації було обрано для подальшого дослідження.

Для згладжування гістограмних оцінок будуть використані сплайн-оцінки [3]. Тут виникає ціла низка питань, пов'язана із правильним розбиттям вибірки на інтервали (класи) групування для адекватної оцінки. В джерелі [3] обґрунтовано, що як у багатовимірному, так і в одновимірному випадку цей апарат не поступається іншим за оцінкою функції щільності і функції розподілу, але й має свої переваги завдяки тому, що локальна апроксимація є більш гнучкою з точки зору врахування локальних особливостей цих функцій (особливо за їхньої неоднорідності). Тому є певний інтерес перевірити якість застосування методів на основі цього апарата, починаючи з одновимірного випадку. Не зменшуючи загальності, буде використовуватися однорідна сукупність змодельованих даних (за розподілом Вейбула). Гнучкість цього розподілу дозволяє імітувати різну асиметричність даних, а різний обсяг фактично імітує ситуацію з пропуском спостережень.

Врахувавши все вищезазначене, поставлено таке завдання: за наявності одновимірної вибірки $\Omega_{1,N}$ з деякої генеральної сукупності Ω розробити алгоритм і програмний додаток для моделювання нових вибірок $\Omega^*_{1,N}$, які б належали тій самій генеральній сукупності, та перевірити якість його роботи.

Як описано в [2], нехай g – двовимірна область, яка, з одного боку, обмежена деяким інтервалом [a; b] осі абсцис, а з іншого – графіком функції $f_\eta(y)$, яка є щільністю розподілу випадкової величини η, що моделюється рис. 1.

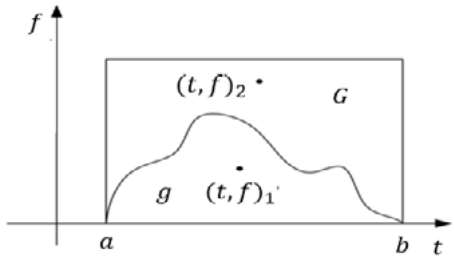

Рис. 1. Ілюстрація методу винятків

Помістимо область g всередину іншої області G (на рис. 1 G – прямокутна область, як це зазвичай буває на практиці; t – абсциси точок; f – ординати точок), що також обмежена на осі абсциси інтервалом [a; b], і нехай точка (t, f) – реалізація випадкового двовимірного вектора (η, ξ), рівномірно розподіленого

в області G. Тоді процедура полягає в тому, що за виконання нерівності $f_\eta(t) \geq f$ значення приймається як реалізація випадкової величини із заданим законом розподілу.

Обраний метод імітації нових вибірок (метод винятків) працює з функцією щільності, а на початку ми маємо лише деяку вибірку даних. Проте, використавши знання, що значення відносних частот класів гістограмної оцінки вибірки є усередненою оцінкою значень функції щільності, маємо змогу оцінити функцію щільності за допомогою згаданої вище сплайн-оцінки. Дотримуючись [3], одержуємо таке: нехай за рівномірним розбиттям

$$\Delta_h : t_i = ih \ \left(\tilde{\Delta}_h : t_i = (i+0,5)h\right),\ i \in Z,\ h > 0 \qquad (1)$$

осі реалізацій випадкової величини $\xi(\omega)$ на підставі вибірки проведено гістограмну оцінку, а отже, отримано

$$F_1,\ F_{1,N_i},\ t \in [t_i; t_{i+1}),\ i \in Z\ -$$

масиви оцінок усереднених значень функцій щільності та розподілу. Під час оцінювання функції щільності $f(t)$ для $\forall t \in [t_{\min}; t_{\max}]$ доцільним є застосування сплайнів на основі B-сплайнів другого та четвертого порядків, серед яких сплайн $S_{2,0}(f, t)$ має найпростішу обчислювальну схему:

$$S_{2,0}(f,t) = \frac{1}{8}\left((f_{i-1} - 2f_i + f_{i+1})x^2 + (-2f_{i-1} + 2f_{i+1})x + (f_{i-1} + 6f_i + f_{i+1})\right), \qquad (2)$$

де

$$x = \frac{2}{h}(t - t_i) - 1;\ i = \left[\frac{t - t_{\min}}{h}\right] + 1;\ [\square] - \text{ціла частина.}$$

Якщо за невизначені під час проведення гістограмної оцінки значення f_{-1}, f_m взяти величини $f_{-1} = 0, f_m = 0$.

У підсумку отримуємо такий алгоритм моделювання:

1. Провести розбиття вихідної вибірки на класи та визначити їх відносні частоти (проведення первинного гістограмного аналізу).

2. На основі отриманих частот відновити функцію щільності розподілу за допомогою поліноміального сплайна (2).

3. На основі відновленої функції щільності використати метод винятків для моделювання нової вибірки $\Omega^*_{1,100}$.

На основі цього алгоритму було проведено такий експеримент:

1. Моделюється вибірка розподілу Вейбула заданої кількості.

2. Для змодельованої вибірки проводиться розбиття на класи (кількість визначається автоматично) та вираховуються їхні відносні частоти (1).

3. На основі порахованих частот за допомогою поліноміального сплайна (2) відновлюються значення функції щільності.

4. На основі відновленої функції за допомогою методу винятків моделюється нова вибірка з 1000 елементів, і для неї проводиться первинний аналіз (кількість класів збігається з кількістю у першій вибірці).

5. Оцінюються параметри α та β [6] для розподілу Вейбула першої (змодельованої) і другої (імітованої) вибірки, аналізується відносна похибка таких параметрів між собою та визначеною програмою.

6. Пункти з 1 по 5 повторюються по 1000 разів для кожної кількості елементів у першій вибірці (15, 25, 40, 100, 400, 1000). У такий спосіб формуються вибірки з 1000 оцінених похибок для α та β.

7. Пункт 6 проводиться для чотирьох заданих пар параметрів розподілу Вейбула:

α	β
70	0.7
200	1.4
1000	2.4
8000	3.9

8. За отриманими вибірками відносних похибок оцінок параметрів від заданих пар параметрів оцінюється якість моделювання за описаним алгоритмом.

На основі вищенаведеної процедури було реалізоване програмне забезпечення мовою програмування C# у програмному середовищі «Visual Studio». У програмному забезпеченні реалізовані модулі моделювання одновимірних вибірок заданого закону розподілу та імітування нових вибірок на основі вже оброблених. Для виконання описаного вище експерименту була визначена окрема кнопка.

Для оцінки якості моделювання у програмному забезпеченні реалізовано обчислення відносних похибок двох видів:

1. Похибка оцінок перших змодельованих вибірок щодо заданих пар параметрів.
2. Похибка оцінок імітованих вибірок щодо заданих пар параметрів.

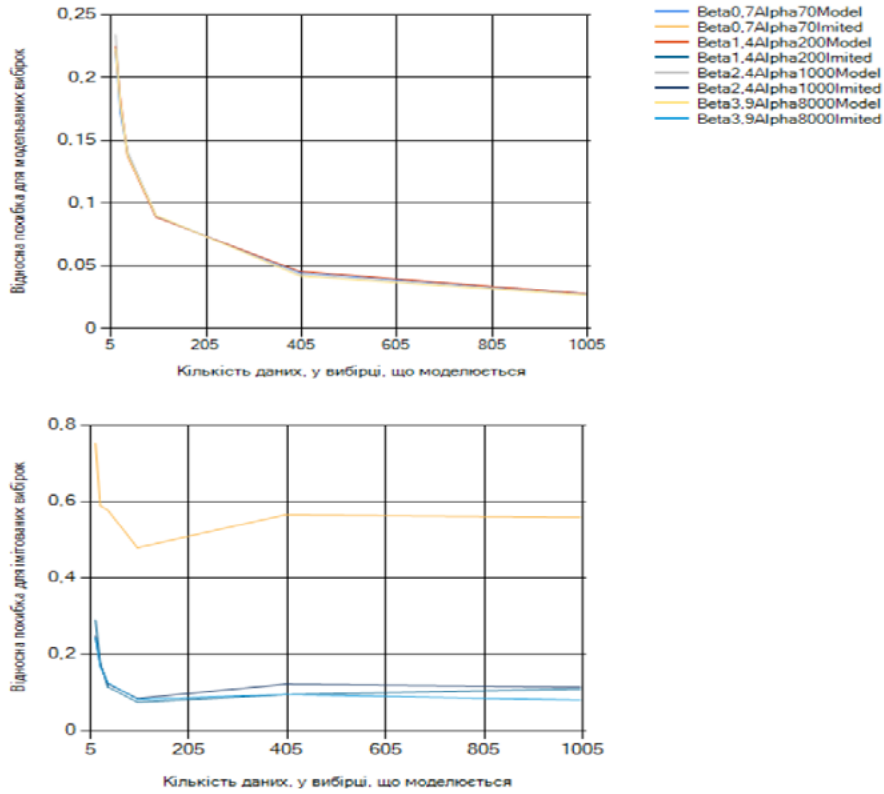

Рис. 2. Порівняльні графіки усереднених відносних похибок за розбиття першої вибірки на класи за формулою (3), залежно від кількості елементів у першій вибірці

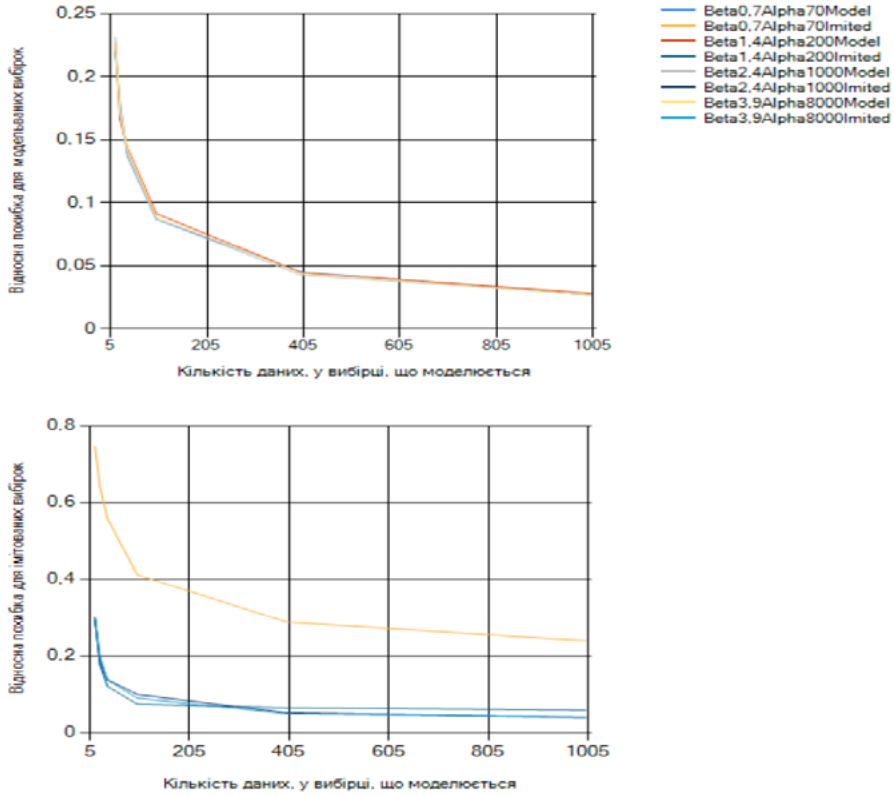

Рис. 3. Порівняльні графіки усереднених відносних похибок за розбиття першої вибірки на класи за формулою (4), залежно від кількості елементів у першій вибірці

Після описаної вище процедури для кожного випадку «пара параметрів – кількість у першочерговій вибірці» сформовано по дві вибірки відносних похибок. Параметр бета (тобто параметр форми) в цьому випадку є більш важливим, тож на рис. 2 та 3 відображено графіки зміни усередненої відносної похибки оціненого параметра бета у модельованій та імітованій вибірках, відповідно, зі зростанням кількості елементів у першій вибірці.

Для відновлення функції щільності за допомогою сплайнів важливою є початкова кількість точок, тобто кількість класів, на яку розбивається перша вибірка. Насамперед була використана стандартна формула розбиття, яка базується на кількості елементів у вибірці:

$$n = \begin{cases} [\sqrt{N}], & N \le 100 \\ [\sqrt[3]{N}], & N > 100 \end{cases}, \tag{3}$$

де n – кількість класів; N – кількість елементів у вибірці; [□] – ціла частина.

За такого розбиття першої вибірки та подальших дій відповідно до алгоритму були отримані вибірки, усереднені відносні похибки оцінок параметра бета яких представлено на рис. 2.

Як видно з графіків на рис. 2, отриманий результат був незадовільний. Тому розбиття на класи було проведено таким чином:

$$n = \left[\frac{\bar{E}+1.5}{6} * N^{0.4}\right], \text{ де } \bar{E} \text{ – коефіцієнт ексцесу.} \quad (4)$$

Для більш чіткого уявлення про результати цього варіанта сформовано порівняльні таблиці з відносними похибками оцінок обох параметрів.

Таблиця 1

Порівняльна таблиця усереднених відносних похибок за 15 елементів у першій вибірці

Пара параметрів		**Похибки α**		**Похибки β**	
α	β	**Модельовані**	**Імітовані**	**Модельовані**	**Імітовані**
70	0.7	0,3097	0,7431	0,2197	0,7462
200	1.4	0,1647	0,1626	0,2262	0,3
1000	2.4	0,1009	0,0971	0,231	0,2929
8000	3.9	0,0627	0,0602	0,2262	0,294

Таблиця 2

Порівняльна таблиця усереднених відносних похибок за 25 елементів у першій вибірці

Пара параметрів		**Похибки α**		**Похибки β**	
α	β	**Модельовані**	**Імітовані**	**Модельовані**	**Імітовані**
70	0.7	0,2479	0,6627	0,1775	0,642
200	1.4	0,1247	0,1302	0,1655	0,1768
1000	2.4	0,077	0,077	0,1705	0,1848
8000	3.9	0,0438	0,0457	0,1724	0,1985

Таблиця 3

Порівняльна таблиця усереднених відносних похибок за 40 елементів у першій вибірці

Пара параметрів		**Похибки α**		**Похибки β**	
α	β	**Модельовані**	**Імітовані**	**Модельовані**	**Імітовані**
70	0.7	0,1968	0,5152	0,1369	0,5584
200	1.4	0,1001	0,0999	0,1434	0,1199
1000	2.4	0,0584	0,0627	0,1382	0,137
8000	3.9	0,0357	0,0403	0,1419	0,136

Таблиця 4

Порівняльна таблиця усереднених відносних похибок за 100 елементів у першій вибірці

Пара параметрів		**Похибки α**		**Похибки β**	
α	β	**Модельовані**	**Імітовані**	**Модельовані**	**Імітовані**
70	0.7	0,1302	0,2942	0,0867	0,4112
200	1.4	0,0638	0,0703	0,0912	0,0739
1000	2.4	0,0375	0,0411	0,0866	0,0992
8000	3.9	0,0232	0,0257	0,0889	0,0897

Таблиця 5

Порівняльна таблиця усереднених відносних похибок за 400 елементів у першій вибірці

Пара параметрів		**Похибки α**		**Похибки β**	
α	β	**Модельовані**	**Імітовані**	**Модельовані**	**Імітовані**
70	0.7	0,0638	0,1366	0,0451	0,2879
200	1.4	0,0319	0,0414	0,0446	0,0644
1000	2.4	0,0176	0,0246	0,0426	0,051
8000	3.9	0,0113	0,0138	0,0441	0,0482

Таблиця 6

Порівняльна таблиця усереднених відносних похибок за 1000 елементів у першій вибірці

Пара параметрів		**Похибки α**		**Похибки β**	
α	β	**Модельовані**	**Імітовані**	**Модельовані**	**Імітовані**
70	0.7	0,0399	0,08	0,0268	0,2392
200	1.4	0,0196	0,0311	0,0281	0,0577
1000	2.4	0,0113	0,0175	0,027	0,0383
8000	3.9	0,007	0,0098	0,0269	0,0398

Як видно з графіка на рис. 3, зміна усередненої відносної похибки для імітованих даних, залежно від кількості елементів у першій вибірці, стала набувати спадаючого характеру. Тобто розбиття на класи першої вибірки за (4) є ефективнішим. Також із наведених таблиць видно, що в деяких випадках усереднена відносна похибка оцінок параметрів для імітованих вибірок є меншою за ті ж похибки у модельованих вибірках (див. табл. 3, 4), а в більшості випадків різниця між ними вимірюється в сотих. Також для обох випадків розбиття на класи видно, що пара параметрів (70, 0.7) піддається імітації гірше за інші (криві мають таку саму форму, але зміщені вгору по осі ординат). Це свідчить про те, що вибірки з сильною асиметрією треба піддавати додатковій обробці (виконувати перетворення даних) або враховувати асиметрію під час розбиття на класи.

Був розроблений алгоритм та програмний додаток для імітації вибірок на основі наявних із використанням поліноміальних сплайнів. З огляду на представлені результати можна стверджувати, що цей алгоритм є ефективним засобом імітації нових вибірок з тієї ж генеральної сукупності, що і вихідна.

Тематикою подальших досліджень можуть бути питання про інші види розбиття на класи для вихідної вибірки, необхідність її попередньої обробки, а також вибір іншого методу імітування на заміну використаного методу винятків.

Найголовнішим напрямом подальшої роботи є модифікація алгоритму для його використання у двовимірних та багатовимірних випадках.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ:

1. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение / пер. с англ. А. Слинкина. 2-е изд., испр. Москва : ДМК Пресс, 2018. 652 с.

2. Гельгор А.Л., Горлов А.И., Попов Е.А. Методы моделирования случайных величин и случайных процессов : учебное пособие. Санкт-Петербург : Издательство Политехнического университета, 2012. 217 с.

3. Приставка П.О. Поліноміальні сплайни під час обробки даних : монографія. Дніпро : Видавництво Дніпропетровського університету, 2004. С. 155–164.

4. Шмырин И.С. Математическое и программное обеспечение информационных, технических и экономических систем. *Труды Томского государственного университета. Серия физико-математическая* : материалы VII Международной молодежной научной конференции, г. Томск, 23–25 мая 2019 г. Томск : Издательский Дом Томского государственного университета, 2019. Т. 303. С. 73–84.

REFERENCES:

1. Goodfellow, Ya., Bengio, I., Courville, A. (2018) Glubokoe obuchenie [Deep Learning]. (2 ed.). Moscow : DMC press. (in Russian)

2. Gel'gor, A.L., Gorlov, A.I., Popov, E.A. (2012) Metody modelirovaniya sluchajnyh velichin i sluchajnyh processov [Methods for modeling random variables and random processes]. Tutorial. St. Petersburg Polytechnic University Publishing House. (in Russian)

3. Pristavka, P.O. (2004) Polinomial'ni splajni pri obrobci danih [Polynomial splines in data processing]. Monograph. Dnipro : Dnipropetrovsk University Publishing House, pp. 155–164. (in Ukrainian)

4. Shmyrin I.S. (2019) Matematicheskoe i programmnoe obespechenie informacionnyh, tekhnicheskih i ekonomicheskih system. Seriya fiziko-matematicheskaya : materialy VII Mezhdunarodnoj molodezhnoj nauchnoj konferencii [Physics and Mathematics Series: Mathematical and Software Support of Information, Technical and Economic Systems: Proceedings of the VII International Youth Scientific Conference]. Proceedings of Tomsk State University (Russia, Tomsk, May 23–25, 2019). Tomsk : Publishing House of Tomsk State University. T. 303. pp. 73–84. (in Russian)