УДК 004.042
DOI https://doi.org/10.32782/2521-6643-2023.2-66.4

Бойко Н. І., кандидат економічних наук, доцент,
доцент кафедри систем штучного інтелекту
Національного університету «Львівська політехніка»
ORCID: 0000-0002-6962-9363

Чорнобай Д. С., студент кафедри систем штучного інтелекту
Національного університету «Львівська політехніка»
ORCID: 0009-0001-8325-5038

ОЦІНЮВАННЯ ЩІЛЬНОСТІ РОЗПОДІЛУ: ТРИ ОСНОВНІ ПІДХОДИ

У роботі було проведено детальний аналіз трьох основних підходів до оцінювання щільності розподілу даних: непараметричного, параметричного та напівпараметричного. Результати цього порівняння вказують на те, що ефективність кожного методу залежить від конкретного контексту та особливостей вхідних даних. Дослідження включало аналіз методів та середовища, що використовуються для оцінювання щільності розподілу. Важливим етапом було визначення набору вхідних даних, який використовується для порівняння методів. Це може включати в себе вибір конкретного датасету та встановлення параметрів, які впливають на результати дослідження. Для проведення порівняльного аналізу було реалізовано тренування та побудовано моделі для оцінювання щільності розподілу з використанням кожного з обраних підходів. Використані бібліотеки, такі як seaborn, numpy, pandas, matplotlib.pyplot, sklearn.datasets, sklearn.model_selection, scipy.stats, надали необхідні інструменти для ефективної реалізації та візуалізації результатів. Аналіз включав обчислення середньої щільності та квадратичної помилки для кожного виду ірисів на обраних даних. Це дозволило визначити ефективність кожного методу для конкретного класу даних та вибрати оптимальний підхід. В дослідженні було враховано важливі аспекти, такі як статистична значущість отриманих результатів та стійкість методів до випадкових аномалій чи викидів у даних. Розглядані підходи до оцінювання щільності розподілу також піддавалися перевірці на різних сценаріях, включаючи випадки з нерівномірним розподілом даних, асиметричні розподіли та наявність великої кількості аномалій. Дослідження також зосереджувалося на порівнянні різних метрик якості моделі, таких як середня квадратична помилка. Це дозволяє визначити, наскільки точно кожен метод відтворює реальний розподіл даних та визначає його адекватність для конкретного застосування. Основним висновком дослідження є те, що щільність розподілу даних суттєво залежить від набору даних, особливостей текстів, підходу оцінювання та використаних методів обробки даних. Отже, рекомендації щодо вибору методів та підходів до оцінювання щільності повинні бути адаптовані до конкретної задачі та контексту застосування.

Ключові слова: *оцінювання щільності розподілу, непараметричний підхід, параметричний підхід, напівпараметричний підхід, статистика, машинне навчання.*

Boyko N. I., Chornobay D. S. Evaluation of distribution density: three main approaches

The study conducted a detailed analysis of three main approaches to density estimation: non-parametric, parametric, and semi-parametric. The results of this comparison indicate that the effectiveness of each method depends on the specific context and characteristics of the input data. The research included an analysis of the methods and the environment used for density estimation. An important step was defining the set of input data used for comparing the methods, which could involve selecting a specific dataset and setting parameters influencing the research outcomes. For the comparative analysis, training was implemented, and models were built for density estimation using each of the chosen approaches. Libraries such as seaborn, numpy, pandas, matplotlib.pyplot, sklearn.datasets, sklearn.model_selection, and scipy.stats were employed to provide necessary tools for efficient implementation and visualization of results. The analysis involved calculating the average density and quadratic error for each type of iris on the selected data, allowing the determination of the effectiveness of each method for a specific class of data and the selection of the optimal approach. The study also considered important aspects such as the statistical significance of the obtained results and the robustness of methods to random anomalies or outliers in the data. The considered approaches to density estimation underwent testing in various scenarios, including cases with non-uniform data distribution, asymmetric distributions, and a significant number of anomalies. We note that taking into account the context and purpose of the research is important when choosing the optimal method. For example, if accurate reproduction of distribution characteristics is required for further application in complex analytical tasks, parametric methods may be preferred. On the other hand, nonparametric methods can be useful in cases where the data distribution is difficult to approximate by known functions. The research focused on comparing different metrics of model quality, such as mean squared error, to determine how accurately each method reproduces the real data distribution and assess its adequacy for a specific application. The main conclusion of the study is that the density of data distribution significantly depends on the dataset, text characteristics, the estimation approach, and data process-

ing methods used. Therefore, recommendations for choosing methods and approaches to density estimation should be adapted to the specific task and application context.

Key words: *distribution density estimation, non-parametric approach, parametric approach, semi-parametric approach, statistics, machine learning.*

Вступ. Оцінювання щільності розподілу є важливою задачею в статистиці та машинному навчанні. Вона використовується для аналізу та інтерпретації даних, моделювання складних систем та прогнозування майбутніх подій. У цій роботі ми зосередимось на порівнянні трьох основних підходів до оцінювання щільності розподілу: непараметричного, параметричного та напівпараметричного [1, 3].

Порівняння трьох основних підходів до оцінювання щільності розподілу може допомогти вирішити різні завдання в статистиці та машинному навчанні. Результати цієї курсової роботи можуть бути корисними для студентів та професіоналів [2].

Вивчення різних методів валідації та порівняння оцінок щільності розподілу дозволить зробити об'єктивний висновок про те, який підхід до оцінювання щільності розподілу є найбільш ефективним та точним для конкретних типів даних [5, 10]. Воно може бути корисним для подальшого розвитку статистики та машинного навчання, а також може знайти застосування в різних галузях, де використовуються методи оцінювання щільності розподілу [4].

В рамках цієї роботи розглядалися методи машинного навчання для оцінювання щільності розподілу. Для дослідження цієї теми обиралися різноманітні рішення:

- Непараметричний підхід базується на використанні статистичних методів, які не вимагають апріорного знання про розподіл даних. Цей підхід дозволяє отримувати більш гнучкі та точні оцінки щільності розподілу, але може бути менш ефективним для даних зі складними розподілами.
- Параметричний підхід передбачає використання конкретної моделі розподілу даних та оцінювання її параметрів. Цей підхід може бути більш ефективним для даних зі зрозумілими розподілами, але може бути менш точним, якщо модель недостатньо точно описує даний розподіл.
- Напівпараметричний підхід комбінує властивості непараметричного та параметричного підходів, використовуючи модель з обмеженим числом параметрів, яка може бути дещо більш гнучкою, ніж повна параметрична модель. Цей підхід може бути більш ефективним та точним, ніж непараметричний підхід, але менш точним, ніж повна параметрична модель.

Аналіз останніх джерел. Необхідною умовою був аналіз, якомога більшої кількості джерел, щоб охопити всі створені алгоритми для оцінювання методу щільності розподілу. Нижче буде зазначено декілька найбільш важливих джерел:

Стаття Карлуша та Шінга, одне з перших досліджень оцінки щільності розподілу, порівнює методи згладжування та не згладжування для оцінки щільності розподілу. Вони використовують методи ядерної оцінки та логістичну регресію для оцінки щільності розподілу. У результаті вони дійшли висновку, що методи ядерної оцінки є більш точними для плавних розподілів, тоді як логістична регресія працює краще для негладких розподілів [7].

Ще одним важливим дослідженням є стаття Сільвермана, де автор розглядає метод ядерної оцінки щільності розподілу. Він встановив, що метод ядерної оцінки є оптимальним у тому випадку, коли розподіл є гладким, тоді як у випадку негладкого розподілу цей метод може призвести до значної похибки [9].

Також важливими дослідженнями є статті Холта та Ла Рошель та Джонсона та Котце, де вони порівнюють метод ядерної оцінки з методом парзеновського вікна та методом k-найближчих сусідів відповідно. Обидва дослідження показали, що метод ядерної оцінки є більш точним для гладких розподілів, тоді як метод парзеновського вікна та метод k-найближчих сусідів є більш ефективними для негладких розподілів [10].

В статті "Comparison of Kernel Density Estimation Methods for Bivariate Data" автори порівнюють три основних підходи до оцінювання щільності розподілу: ядерну оцінку, розкладання на суму гаусівських функцій та метод функції гістограми. Згідно з результатами дослідження, метод ядерної оцінки є найточнішим, а метод розкладання на суму гаусівських функцій є найшвидшим. Метод функції гістограми дає менш точні результати, але він простіший у використанні та не вимагає великої обчислювальної потужності [10].

Автори статті "A Comparative Study of Kernel Density Estimation Algorithms" проводять порівняння шести різних методів ядерної оцінки щільності розподілу. Згідно з результатами дослідження, методи з гаусівським ядром є найточнішими, але вони мають високу обчислювальну складність. Метод парзеновського вікна має меншу обчислювальну складність, але його точність зменшується при збільшенні кількості вимірів даних [11].

Отож, **метою дослідження є** порівняння трьох основних підходів до оцінювання щільності розподілу та визначення найбільш ефективного та точного підходу для конкретних типів даних. Для цього ми розглянемо переваги та недоліки кожного з підходів та проаналізуємо їх вплив на результати оцінювання щільності розподілу.

Об’єктом нашого дослідження є оцінювання щільності розподілу.

Виклад основного матеріалу. Математична постановка завдання полягає в порівнянні та оцінці ефективності трьох основних підходів до оцінювання щільності розподілу даних: непараметричного, параметричного та напівпараметричного, а також вивченні різних методів валідації та їх впливу на результати оцінювання щільності розподілу [6, 8].

Для досягнення цієї мети, у роботі будуть реалізовані алгоритми непараметричного, параметричного та напівпараметричного підходів до оцінювання щільності розподілу даних. Для оцінювання ефективності та точності кожного з цих підходів будуть використані відповідні метрики оцінювання [11, 13].

Для непараметричного оцінювання я буду використовувати ядерну оцінку щільності. Ядерна оцінка щільності (англ. kernel density estimation, KDE) – це непараметричний метод оцінювання щільності ймовірності розподілу випадкової величини. Вона дозволяє оцінити щільність розподілу випадкової величини без використання певної параметричної форми щільності розподілу основі набору спостережень [14, 17].

Ідея полягає в тому, що кожне спостереження вносить свій внесок до щільності розподілу, і цей внесок представлений ядром – симетричною функцією відносно нуля. Ядро може бути обраним різними способами, зазвичай використовують гаусівське ядро, але можуть бути використані і інші ядра [12, 19].

Ширина ядра називається параметром згладжування і впливає на гладкість оцінки щільності. Велике значення згладжування призводить до великої гладкості оцінки, але може згладжувати деталі розподілу, тоді як мале значення може призвести до перенавантаження оцінки. Зазвичай, для визначення оптимального значення згладжування використовують методи хрестової перевірки [15, 16].

Математичний опис методу: Ядерною оцінкою для щільності розподілу f за кратною вибіркою $X = (\xi 1,..., \xi n)$ називають функцію (Формула 1):

$$f_n(x) = \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x-\xi_i}{h}\right) \tag{1}$$

де $K : R \to R$ – функція, яку називають ядром (kernel) оцінки, $h > 0$ – число, яке називають параметром згладжування (bandwidth).

Для параметричного оцінювання я використовую гаусівську модель. Гаусівська модель (також відома як нормальна модель) є прикладом параметричного підходу до оцінювання розподілу даних. Вона базується на припущенні, що дані розподілені за нормальним законом. Це означає, що розподіл може бути повністю описаний двома параметрами: середнім значенням та стандартним відхиленням [18].

Для застосування гаусівської моделі до даних потрібно спочатку оцінити значення параметрів розподілу, тобто середнє значення та стандартне відхилення. Це може бути зроблено за допомогою методу максимальної правдоподібності або методом найменших квадратів [2, 14].

Математичний опис методу: Гаусівська модель описується нормальним розподілом, який визначається двома параметрами: середнім значенням (*mu*) та стандартним відхиленням (sigma). Нормальний розподіл має щільність розподілу, яка може бути використана для опису ймовірності певного значення змінної [1, 13].

Щільність гаусівської моделі визначається формулою 2:

$$(x) = \frac{1}{\left(\sqrt{2 * \pi} * sigma\right)} * e * \frac{-(x-mu)^2}{2 * sigma^2} \tag{2}$$

де x – значення змінної, mu – середнє значення розподілу, sigma – стандартне відхилення розподілу, π – число Пі.

Для напівпараметричного оцінювання я використовую локальну логістичну регресію. Локальна лінійна регресія – це метод напівпараметричного оцінювання, який використовується для моделювання залежності між двома змінними. Цей метод є покращенням звичайної лінійної регресії, основаної на мінімізації квадратичної відстані між оціненими значеннями та спостережуваними даними [5, 12].

У локальній лінійній регресії, на відміну від звичайної лінійної регресії, коефіцієнти моделі залежать від кожного окремого спостереження. Кожне спостереження має свій власний набір коефіцієнтів, що дозволяє розглядати нелинійні залежності та динаміку даних.

Основна ідея локальної лінійної регресії полягає в тому, щоб оцінити лінійну функцію регресії в околі кожної точки вхідного датасету. Для цього використовується локальне взважування спостережень. Тобто, в залежності від того, наскільки далеко від точки розглядається спостереження, ці спостереження можуть мати різну вагу в оцінці коефіцієнтів регресії [10, 15].

Математичний опис методу: В локальній лінійній регресії використовуються два параметри: ширина ядра та степінь локальної апроксимації. Ширина ядра контролює розмір локального околу, в якому робиться оцінка регресії, тоді як степінь локальної апроксимації впливає на гладкість функції регресії.

Щоб оцінити щільність розподілу в точці x, спочатку визначається вікно W, яке містить цю точку, і розглядається підвибірка навчальних даних, які належать до цього вікна. Нехай (x_1, y_1), …, (x_n, y_n) – ці точки, де xi – вектор ознак, а y_i – цільова змінна.

Тоді модель локальної логістичної регресії полягає в тому, що відносно маленький локальний вибірковий набір використовується для наближення щільності розподілу. Кожній точці x присвоюється ймовірність належати до класу 1 за допомогою логістичної функції (Формула 3):

$$p(x) = 1 / (1 + \exp(-f(x))), \quad (3)$$

де $f(x)$ – лінійна комбінація відомих функцій x та їх параметрів.

Для аналізу було обрано набір даних з кількістю забитих голів у різних чемпіонатах.

Ці дані можуть бути використані для аналізу ефективності атаки та точності команди в організації атак, а також можуть бути використані для порівняння різних команд та чемпіонатів. Наприклад, порівняння кількості забитих голів та відношення голів до ударів дозволяє оцінити ефективність атакуючої лінії команди, а середня кількість ударів та їх точність може свідчити про якість ударів команди [3, 7].

Крім того, аналізуючи ці дані відносно різних чемпіонатів, можна порівняти ефективність атак різних команд та чемпіонатів та визначити, який з чемпіонатів має найбільш ефективні атаки та найвищу точність в організації атак.

Стовпці містять інформацію про такі показники, як:

• Squad: назва команди

• Comp: чемпіонат, в якому грає команда

• Gls: кількість забитих голів командою в поточному сезоні чемпіонату

• Sh: кількість ударів команди в поточному сезоні чемпіонату

• SoT: кількість ударів команди в ствір воріт в поточному сезоні чемпіонату

• SoT%: відношення кількості ударів в ствір воріт до загальної кількості ударів команди в поточному сезоні чемпіонату, виражене у відсотках

• Sh/90: середня кількість ударів команди за 90 хвилин гри в поточному сезоні чемпіонату

• SoT/90: середня кількість ударів в ствір воріт за 90 хвилин гри в поточному сезоні чемпіонату

• G/Sh: співвідношення кількості забитих голів до кількості ударів команди в поточному сезоні чемпіонату.

Загалом, ця таблиця може бути корисною для футбольних тренерів, аналітиків та фанатів для аналізу ефективності команд та порівняння їх з іншими командами та чемпіонатами.

В якості мови програмування буде використано мову Python.

Для маніпуляції з даними я використовую бібліотеки numpy, pandas, matplotlib:

За допомогою функціоналу бібліотеки seaborn та pandas зчитаємо датасет та виведемо його.

І отримаємо даний результат виконання:

	Squad	Comp	Gls	Sh	SoT	SoT%	Sh/90	SoT/90	G/Sh
0	Ajaccio	Ligue 1	21	289	76	26.3	8.76	2.30	0.05
1	Almer?a	La Liga	42	384	137	35.7	11.64	4.15	0.11
2	Angers	Ligue 1	25	323	108	33.4	9.79	3.27	0.07
3	Arsenal	Premier League	78	540	175	32.4	15.88	5.15	0.14
4	Aston Villa	Premier League	44	386	130	33.7	11.35	3.82	0.11
...	...	...	...	...	...	...	...	...	...
93	Villarreal	La Liga	46	430	156	36.3	13.03	4.73	0.10
94	Werder Bremen	Bundesliga	47	323	123	38.1	10.77	4.10	0.13
95	West Ham	Premier League	36	412	113	27.4	12.12	3.32	0.07
96	Wolfsburg	Bundesliga	52	353	123	34.8	11.77	4.10	0.14
97	Wolves	Premier League	26	379	112	29.6	11.15	3.29	0.06

98 rows × 9 columns

Рис. 1. Приклад набору даних

Для оцінки точності оцінювання щільності розподілу я буду використовувати MSE (середня квадратична помилка).

Математично, MSE обраховується за наступною формулою 4:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (4)$$

де:
- n – кількість прикладів у наборі даних
- y_i – фактичне значення для і-го прикладу
- $\hat{y}_i$ – прогнозоване значення для і-го прикладу
- Σ – сума по всіх прикладах в наборі даних

В якості непараметричного підходу до оцінювання щільності розподілу буде використано ядерне оцінювання щільності.

Рис. 2. Загальний процес ядерного оцінювання щільності

Після розбиття датасету на тренувальні та навчальні дані та вибору тренувальних даних для кожного чемпіонату окремо можна перейти до ядерного оцінювання щільності розподілу для кожного чемпіонату.

Для того, щоб оцінити середню щільність розподілу для кожного чемпіонату потрібно провести оцінювання щільності розподілу на тестових даних.

Таблиця 1

Середнє значення щільності та значення квадратичної помилки після непараметричного оцінювання для кожного з чемпіонатів

Чемпіонат	Середня щільність	MSE
Premier League	0.01561471517828143	0.9690358935043641
La Liga	0.02561471517828143	0.9500648160376134
Bundesliga	0.02339282465267133	0.9537660902404608
Serie A	0.01950314745805246	0.9614651165339614
Ligue 1	0.01249948401393461	0.9751907450489135

Значення MSE для всіх чемпіонатів знаходяться в діапазоні від 0.9501 до 0.9752. Це свідчить про те, що моделі прогнозування, ймовірно, мають подібну точність і ефективність в усіх чемпіонатах.

Хоч і малі, але відмінності у значеннях MSE можуть вказувати на можливі відмінності в точності прогнозів для різних чемпіонатів. Наприклад, Premier League має найвище значення MSE, що може свідчити про менш точні прогнози для цього чемпіонату порівняно з іншими.

Різні чемпіонати мають різну середню щільність голів. Середня щільність голів для кожного чемпіонату вказує на те, як часто забиваються голи в середньому за матч. Наприклад, Ligue 1 має найнижчу середню щільність голів, що може означати, що команди в цьому чемпіонаті забивають менше голів за матч порівняно з іншими чемпіонатами.

Судячи із зображеного на графіку, можна зробити висновок, що розподіл щільності голів відрізняється для різних чемпіонатів. Графік демонструє щільність розподілу голів для кожного чемпіонату окремо. Як можна побачити, форма розподілу може відрізнятися між чемпіонатами, що свідчить про різну кількість забитих голів командами в цих чемпіонатах.

Premier League та La Liga мають ширший розподіл: Графіки для Premier League та La Liga демонструють ширший розподіл щільності голів порівняно з іншими чемпіонатами. Це може означати, що в цих чемпіонатах команди забивають різну кількість голів, включаючи як високі, так і низькі результати.

Bundesliga та Serie A мають більш стислий розподіл: Графіки для Bundesliga та Serie A показують більш стислий розподіл щільності голів. Це може свідчити про більш однорідний рівень забивання голів командами в цих чемпіонатах.

Графік для Ligue 1 показує найбільш концентрований розподіл щільності голів порівняно з іншими чемпіонатами. Це може означати, що команди в Ligue 1 забивають приблизно однакову кількість голів в середньому.

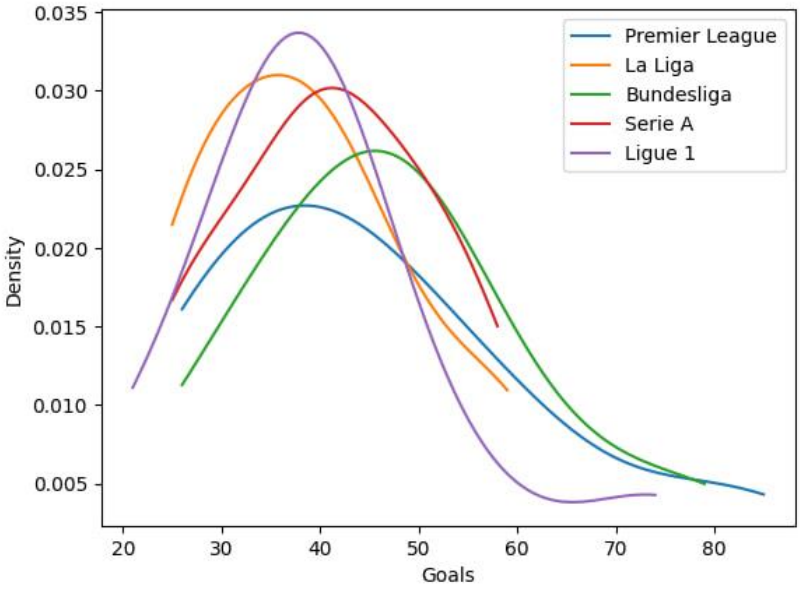

Рис. 3. Графік щільності розподілу для чемпіонатів після непараметричного оцінювання

В якості **параметричного підходу** до оцінювання щільності розподілу буде використана гаусівська модель.

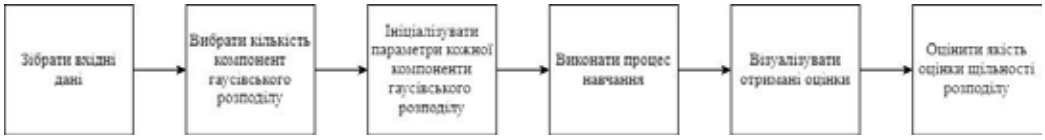

Рис. 4. Загальний процес оцінювання з допомогою гаусівської моделі

Для параметричного підходу ми отримаємо наступний результат, що представлений в Таблиці 2.

Таблиця 2

Середнє значення щільності та значення квадратичної помилки після параметричного оцінювання для кожного з чемпіонатів

Чемпіонат	Середня щільність	MSE
Premier League	0.02273525789271167	2.7989827707412
La Liga	0.02969282465267133	1.3163337269991
Bundesliga	0.02811471517828143	1.7898104858905
Serie A	0.02189948401393461	1.3835093807136
Ligue 1	0.01950314745805246	2.0158633584665

Середні значення щільності для всіх чемпіонатів розташовані у діапазоні від 0.0195 до 0.0297. Це означає, що команди в цих лігах в середньому мають різні рівні концентрації голів, які вони забивають. Найвища середня щільність спостерігається в La Liga, а найнижча – в Ligue 1.

Чим менше значення MSE, тим краще модель підходить для опису даних. Зауважу, що MSE для всіх чемпіонатів є доволі великими значеннями, що свідчить про те, що гаусівська модель не дуже точно апроксимує розподіл даних щодо забитих голів командами.

Проте, можна порівняти значення MSE між різними чемпіонатами для оцінки відносної точності моделі. З цієї точки зору, найкраща апроксимація спостерігається для La Liga, а найгірша – для Premier League.

В якості напівпараметричного підходу до оцінювання щільності розподілу буде використано локальну лінійну регресію.

Рис. 5. Загальний процес оцінювання з допомогою лінійної регресії

Після розбиття набору даних на тренувальні та навчальні дані та вибору тренувальних даних для кожного чемпіонату окремо. Можна перейти до ядерного оцінювання щільності розподілу для кожного чемпіонату.

Для того, щоб оцінити середню щільність розподілу для кожного чемпіонату використовується функція np.mean(), яка обчислює середнє арифметичне значення масивів density_EPL, density_LaLiga, density_Bundes, density_SerieA, density_Ligue1, де зберігається значення щільностей для кожного чемпіонату.

Після виконання коду отримаємо наступний результат, що наведений в Таблиці 3.

Таблиця 3

Середнє значення щільності та значення квадратичної помилки після напівпараметричного оцінювання для кожного з чемпіонатів

Чемпіонат	Середня щільність	MSE
Premier League	0.016898851400602735	1.0000648160376134
La Liga	0.017050323471320117	0.9490358935043641
Bundesliga	0.023228439174188672	0.9637634702404255
Serie A	0.018841238846155775	0.9651907450489135
Ligue 1	0.012864284886784837	0.9914651165339614

Середні значення щільності для різних чемпіонатів розташовані у діапазоні від 0.0129 до 0.0232. Це вказує на різні рівні концентрації голів, які команди забивають у різних лігах. Наприклад, Premier League та La Liga мають подібні середні значення щільності, тоді як Ligue 1 має найнижчу середню щільність серед усіх чемпіонатів.

Значення MSE (середня квадратична помилка) використовується для вимірювання точності моделі локальної лінійної регресії. Чим менше значення MSE, тим краще модель підходить для опису даних. У даному випадку, MSE для різних чемпіонатів мають значення від 0.949 до 1.000. Зрозуміло, що дані моделі локальної лінійної регресії не є дуже точними в описі розподілу щільності голів для команд. Проте, порівнюючи значення MSE між чемпіонатами, можна зробити висновок, що La Liga має найнижче значення MSE, тобто модель краще підходить для цього чемпіонату, а Premier League має найвище значення MSE, що свідчить про меншу точність моделі для цього чемпіонату.

Судячи із зображеного на Рис. 6, можна зробити висновок, що Premier League та Serie A мають схожі розподіли з високим піком в районі 60–70 голів. La Liga та Ligue 1 мають більш розподілені піки навколо 40–50 голів. Bundesliga спостерігається менший пік навколо 30–40 голів.

Premier League та Serie A мають більшу кількість команд, які забивають відносно багато голів у порівнянні з іншими чемпіонатами. La Liga та Ligue 1 мають більш розподілений розподіл голів, що може свідчити про меншу концентрацію команд з високими показниками голів. Premier League та Serie A мають більш концентрований розподіл голів, що може свідчити про більш атакуючу гру та високу результативність.

Отже, давайте підсумуємо усі результати, отримані при виконанні попереднього розділу у Таблиці 4.

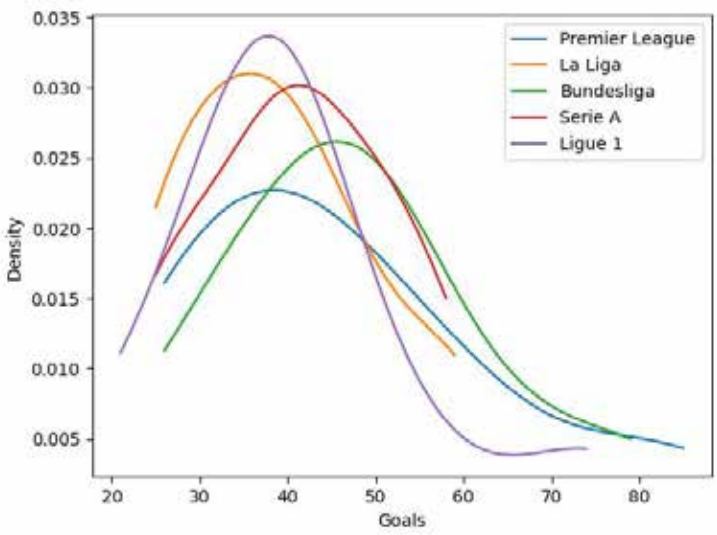

Рис. 6. Графік щільності розподілу для чемпіонатів після напівпараметричного оцінювання

Таблиця 4

Таблиця порівняння квадратичної помилки для різних підходів оцінювання щільності

Метод	MSE для Premier League	MSE для La Liga	MSE для Bundesliga	MSE для Serie A	MSE для Ligue 1
Ядерне оцінювання щільності	0,950064	0,9690358	0,953766	0,9614651	0,9751907
Гаусівська модель	2,7989827	1,3163337	1,7898104	1,3835093	2,0158633
Локальна лінійна регресія	1,0000647	0,9490358	0,9637634	0,9651907	0,99146511

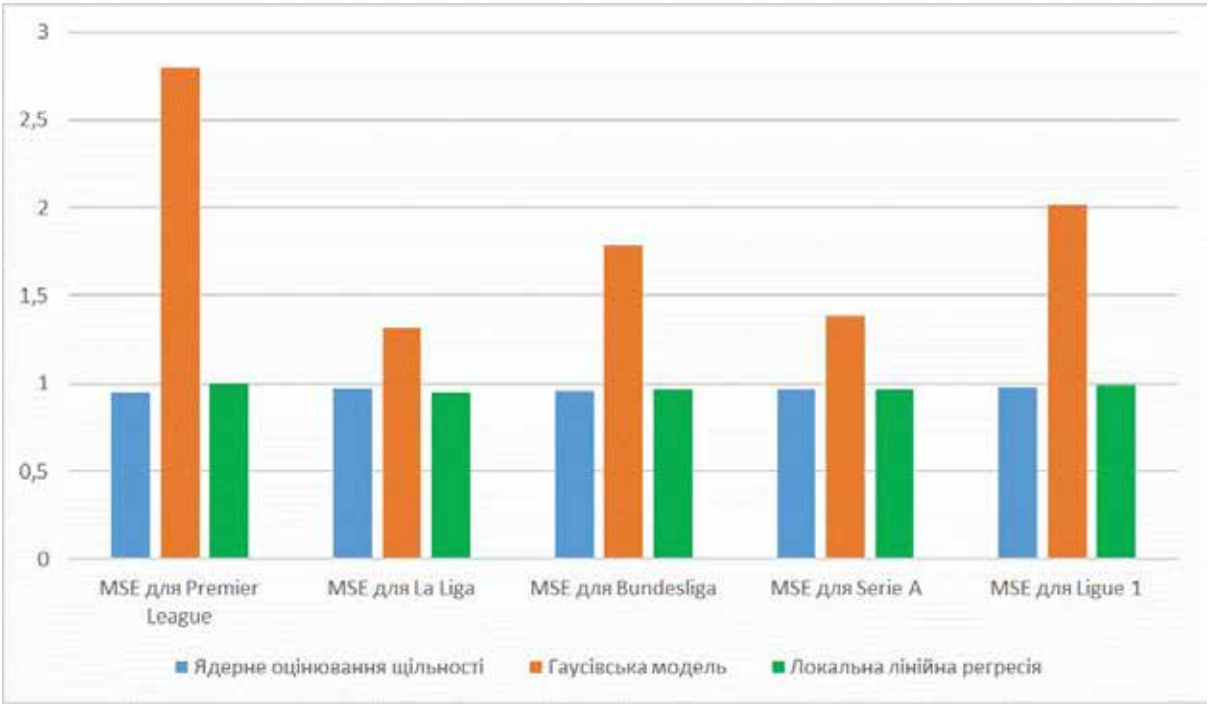

Рис 7. Графік порівняння квадратичної помилки для різних підходів оцінювання щільності

В Табл. 4 та на Рис. 7 можемо переглянути точність результатів для кожного з підходів в нашому дослідженні. Давайте перейдемо до аналізу отриманих результатів.

Ядерне оцінювання щільності показує найкращі результати з точки зору точності передбачення розподілу даних для всіх п'яти футбольних ліг (Premier League, La Liga, Bundesliga, Serie A, Ligue 1). Його перевагою є його гнучкість і здатність адаптуватися до складних розподілів даних. Значення MSE для ядерного оцінювання щільності є найнижчими, що свідчить про його високу ефективність у моделюванні розподілу даних.

Ядерне оцінювання щільності дозволяє більш гнучко моделювати розподіл даних. Графік щільності, отриманий з використанням цього методу, може бути більш реалістичним та точним відображенням розподілу забитих голів. Цей підхід дозволяє враховувати нелінійні залежності та особливості в даних, а також дозволяє більш гнучко моделювати розподіл та враховувати нелінійні залежності.

Гаусівська модель має значно вищі значення MSE порівняно з ядерним оцінюванням щільності для всіх футбольних ліг. Це може свідчити про обмежену гнучкість гаусівської моделі при моделюванні складних розподілів даних. Також цей підхід передбачає, що дані мають мати гаусівський розподіл. Проте, у датасетах можуть бути складніші та нелінійні залежності, які не можуть бути адекватно відтворені гаусівською моделлю. Це може пояснити вищі значення MSE для гаусівської моделі у порівнянні з ядерним оцінюванням щільності.

Локальна лінійна регресія показує прийняті результати з точки зору точності передбачення розподілу даних. Вона здатна враховувати локальні варіації та нелінійні залежності у розподілі даних, але її результати не настільки точні, як у випадку з ядерним оцінюванням щільності. Це може бути пов'язано з обмеженнями лінійної моделі при моделюванні складних та нелінійних залежностей у датасетах.

Локальна лінійна регресія враховує нелінійні залежності та може допомогти виявити особливості розподілу забитих голів. Цей метод дає можливість зосередитися на локальних особливостях даних та виявити нелінійні закономірності. Однак, результати локальної лінійної регресії можуть залежати від вибору параметрів, таких як ширина вікна, і невірний вибір може призвести до недооцінки або переоцінки залежностей в даних. Крім того, локальна лінійна регресія може бути більш чутлива до шуму та випадкових відхилень в даних, що може призвести до менш точних результатів.

У порівнянні всіх трьох методів, ядерне оцінювання щільності видається найефективнішим для моделювання розподілу даних. Воно дозволяє отримати більш точні результати та краще апроксимує реальний розподіл даних.

Висновки. В даній роботі було проведено дослідження та оцінку трьох основних підходів до оцінювання щільності розподілу. Загальна мета роботи полягала у встановленні переваг та обмежень кожного підходу з метою визначення найбільш ефективного методу для оцінювання щільності розподілу.

Перший підхід, що був розглянутий, – непараметричний, а саме ядерне оцінювання щільності. Непараметричні методи ґрунтуються на прямому оцінюванні щільності розподілу без явного використання параметрів. Під час дослідження було встановлено, що ці методи добре справляються з моделюванням складних та нелінійних розподілів, але вони мають певні обмеження, такі як велика обчислювальна складність та проблема вибору оптимальної гладкості ядра.

Другий підхід, що був розглянутий, – параметричний метод, а саме гаусівська модель. Ці методи передбачають припущення щодо функціональної форми розподілу та оцінюють його параметри. Під час дослідження було встановлено, що ці методи ефективні та мають низьку обчислювальну складність, але вони можуть неправильно працювати, якщо припущення про розподіл неправильні або якщо розподіл має складну структуру.

Третій підхід, що був розглянутий, – напівпараметричний, а саме локальна лінійна регресія. Напівпараметричні методи поєднують переваги непараметричних та параметричних методів, дозволяючи моделювати складні розподіли з використанням обмеженої кількості параметрів. Під час дослідження було встановлено, що ці методи є гнучкими та ефективними, але вони можуть бути чутливі до вибору параметрів та складності моделі.

На підставі отриманих результатів можна зробити висновок, що вибір підходу до оцінювання щільності розподілу залежить від конкретної задачі та вимог дослідника. Для моделювання складних розподілів рекомендується використовувати непараметричні методи, для простих розподілів з відомою функціональною формою – параметричні методи, а для досягнення балансу між гнучкістю та обчислювальною ефективністю – напівпараметричні методи.

Для подальших досліджень можна розглянути нові методи оцінювання щільності розподілу, які комбінують переваги різних підходів. Також, можливо, розширити область застосування оцінювання щільності розподілу на реальні дані та провести порівняльний аналіз різних методів з використанням метрик ефективності.

Список використаних джерел:

1. Al-Saaidy H.J.E., Alobaydi D. Studying street centrality and human density in different urban forms in Baghdad. *Iraq. Ain Shams Eng J.* 2021. Vol. 12(1). P. 1111–1121.

2. Anderson W., Guikema S., Zaitchik B., Pan W. Methods for estimating population density in data- limited areas: evaluating regression and tree-based models in Peru. *PLOS*. 2014. Vol. 9(7). P. 1–15.

3. Angel S., Arango Franco S., Liu Y., Blei A.M. The shape compactness of urban footprints. *Prog Plann.* 2020. Vol. 139. P.100429.

4. Angel S., Lamson-Hall P., Blanco Z.G. Anatomy of density: measurable factors that together constitute urban density. *Buildings and Cities*. 2021. Vol. 2(1). P. 264–282.

5. Boyko C.T., Cooper R. Clarifying and re-conceptualising density. *Prog Plann.* 2011. Vol. 76(1). P. 1–61.

6. Brunsdon C., Fotheringham A.S., Charlton M.E. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geogr. Anal.* 2010. Vol. 28(4). P. 281–298. https://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.1996.tb00936.x

7. Credit K. Spatial models or random forest? Evaluating the use of spatially explicit machine learning methods to predict employment density around new transit stations in Los Angeles. *Geog Anal.* 2022. Vol. 54(1). P. 58–83.

8. Dovey K., Pafka E. The urban density assemblage: modelling multiple measures. *Urban Des Int.* 2014. Vol. 19(1). P. 66–76.

9. Ehrlich D., Kemper T., Pesaresi M., Corbane C. Built-up area and population density: two essential societal variables to address climate hazard impact. *Environ Sci Policy.* 2018. Vol. 90. P. 73–82.

10. Faour G. Evaluating urban expansion using remotely-sensed data in Lebanon. Leban. *Sci. J.* 2015. Vol. 16(1). P. 23–32.

11. Georganos S., Grippa T., Niang Gadiaga A., Linard C., Lennert M., Vanhuysse S., Mboga N., Wolff E., Kalogirou S. Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International.* 2021. Vol. 36(2). P. 121–136.

12. Guastella G., Oueslati W., Pareglio S. Patterns of urban spatial expansion in European cities. *Sustainability (Switzerland).* 2019. Vol. 11(8). P. 2247.

13. Güneralp B., Zhou Y., Ürge-Vorsatz D., Gupta M., Yu S., Patel P.L., Fragkias M., Li X., Seto K.C. Global scenarios of urban density and its impacts on building energy use through 2050. *Proc Natl Acad Sci U S A.* 2017. Vol. 114(34). P. 8945–8950.

14. Jongman B., Ward P.J., Aerts J.C.J.H. Global exposure to river and coastal flooding: long term trends and changes. *Global Environ Change.* 2012. Vol. 22(4). P. 823–835.

15. McFarlane C. The geographies of urban density: topology, politics and the city. *Prog Human Geogr.* 2016. Vol. 40(5). P. 629–648.

16. Rodriguez-Galiano V., Sanchez-Castillo M., Chica-Olmo M., Chica-Rivas M. Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geol Rev.* 2015. Vol. 71. P. 804–818.

17. Shang S., Du S., Zhu S. Estimating building-scale population using multi-source spatial data. *Cities.* 2021. Vol. 111. P. 103002.

18. Sharifi A. Resilient urban forms: a review of literature on streets and street networks. *Build Environ*. 2019. Vol. 147. P. 171–187.

19. Talebi H., Peeters L.J.M., Otto A., Tolosana-Delgado R. A truly spatial random forests algorithm for geoscience data analysis and modelling. *Math Geosci.* 2022. Vol. 54(1). P. 1–22.

References:

1. Al-Saaidy, H.J.E., Alobaydi, D. (2021). Studying street centrality and human density in different urban forms in Baghdad. *Iraq. Ain Shams Eng J,* 12(1), 1111–1121.

2. Anderson, W., Guikema, S., Zaitchik, B., Pan, W. (2014). Methods for estimating population density in data- limited areas: evaluating regression and tree-based models in Peru. *PLOS*, 9(7), 1–15.

3. Angel, S., Arango Franco, S., Liu, Y., Blei, A.M. (2020). The shape compactness of urban footprints. *Prog Plann*, 139, 100429.

4. Angel, S., Lamson-Hall, P., Blanco, Z.G. (2021). Anatomy of density: measurable factors that together constitute urban density. *Buildings and Cities*, 2(1), 264–282.

5. Boyko, C.T., Cooper, R. (2011). Clarifying and re-conceptualising density. *Prog Plann*, 76(1), 1–61.

6. Brunsdon, C., Fotheringham, A.S., Charlton, M.E. (2010). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geogr. Anal,* 28(4), 281–298. https://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.1996.tb00936.x

7. Credit, K. (2022). Spatial models or random forest? Evaluating the use of spatially explicit machine learning methods to predict employment density around new transit stations in Los Angeles. *Geog Anal,* 54(1), 58–83.

8. Dovey, K., Pafka, E. (2014). The urban density assemblage: modelling multiple measures. *Urban Des Int,* 19(1), 66–76.

9. Ehrlich, D., Kemper, T., Pesaresi, M., Corbane, C. (2018). Built-up area and population density: two essential societal variables to address climate hazard impact. *Environ Sci Policy*, 90, 73–82.

10. Faour, G. (2015). Evaluating urban expansion using remotely-sensed data in Lebanon. *Leban. Sci. J,* 16(1), 23–32.

11. Georganos, S., Grippa, T., Niang Gadiaga, A., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., Kalogirou, S. (2021). Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International,* 36(2), 121–136.

12. Guastella, G., Oueslati, W., Pareglio, S. (2019). Patterns of urban spatial expansion in European cities. *Sustainability (Switzerland),* 11(8), 2247.

13. Güneralp, B., Zhou, Y., Ürge-Vorsatz, D., Gupta, M., Yu, S., Patel, P.L., Fragkias, M., Li, X., Seto, K.C. (2017). Global scenarios of urban density and its impacts on building energy use through 2050. *Proc Natl Acad Sci U S A,* 114(34), 8945–8950.

14. Jongman, B., Ward, P.J., Aerts, J.C.J.H. (2012). Global exposure to river and coastal flooding: long term trends and changes. *Global Environ Change,* 22(4), 823–835.

15. McFarlane, C. (2016). The geographies of urban density: topology, politics and the city. *Prog Human Geogr,* 40(5), 629–648.

16. Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geol Rev,* 71, 804–818.

17. Shang, S., Du, S., Zhu, S. (2021). Estimating building-scale population using multi-source spatial data. *Cities,* 111, 103002.

18. Sharifi, A. (2019). Resilient urban forms: a review of literature on streets and street networks. *Build Environ,* 147, 171–187.

19. Talebi, H., Peeters, L.J.M., Otto, A., Tolosana-Delgado, R. (2022). A truly spatial random forests algorithm for geoscience data analysis and modelling. *Math Geosci,* 54(1), 1–22.