МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЧЕРНІВЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ЮРІЯ ФЕДЬКОВИЧА

Кваліфікаційна наукова праця
на правах рукопису

Кнігніцька Тетяна Василівна

УДК 519.246.8(043.5)

ДИСЕРТАЦІЯ

ОЦІНКИ ПАРАМЕТРІВ АВТОРЕГРЕСІЙНИХ МОДЕЛЕЙ

113 – Прикладна математика
11 – Математика та статистика

Подається на здобуття наукового ступеня доктора філософії.

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

_____ Т.В. Кнігніцька

Науковий керівник: **Малик Ігор Володимирович**, доктор фізико-математичних наук, професор

Чернівці – 2023

АНОТАЦІЯ

Кнігніцька Т.В. Оцінки параметрів авторегресійних моделей. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 113 – Прикладна математика. – Чернівецький національний університет імені Юрія Федьковича МОН України, Чернівці, 2023.

Дисертаційна робота присвячена знаходженню відстаней між вимірюваннями даних, які представлені часовими рядами, та визначенню оптимальної кількості кластерів на основі власних значень стохастичної матриці графа. Дисертація складається із вступу, трьох розділів, висновків та переліку використаних джерел.

У вступі обгрунтовано актуальність теми дослідження, сформульовано мету, завдання, предмет, об'єкт та методи дослідження, вказано наукову новизну, практичне значення отриманих результатів, зв'язок роботи з науковими дослідженнями та особистий внесок здобувача, а також наведено дані про те, де доповідались, обговорювались та були опубліковані основні результати дисертації.

У першому розділі здійснено огляд наукової літератури, присвяченої дослідженню часових рядів, зокрема, визначенню метрик подібності між часовими рядами та підходи до кластеризації даних, які представлені у вигляді неструктурованих типів даних. Детально проаналізовано хронологію розвитку наукових підходів до задач кластеризації, класифікації, зменшення розмірності часових рядів. Перший пункт розділу 1 відображає загальний огляд розвитку наукових досліджень при дослідженні часових рядів та існуючі метрики для встановлення подібності між часовими рядами. У другому пункті наведено методи дослідження структурних стрибків у часових рядах. У третьому пункті зроблено огляд наукових досліджень,

які стосуються неперервних часових рядів. Вибір оптимальної кількості кластерів при поділі даних на групи представлено у пункту четвертому.

У другому розділі запропоновано визначати подібність або відстань між часовими рядами за допомогою моделей часових рядів. Запропонований алгоритм для встановлення подібності двох наборів даних використовує параметри моделі, а не самі вимірювання. У якості моделей часових рядів розглянуто стаціонарні $ARMA$ моделі. Отриманий алгоритм порівнюється з уже існуючими метриками знаходження відстаней у випадку збільшення вимірювань часового ряду та у випадку зростання кількості викидів у вхідному часовому ряді. Отриманий алгоритм має меншу обчислювальну складність, ніж алгоритми Евкліда, DTW та ERP. Запропоновану відстань можна використовувати для кластеризації сильно зашумлених даних.

Наукову новизну висновків, зроблених на основі отриманих у другому розділі результатів, розкривають такі положення:

1. Описано алгоритм для знаходження відстані між часовими рядами на основі моделей часових рядів. Отримана відстань є більш стійкою до викидів у часових рядах. У випадку збільшення кількості викидів запропонований у дисертаційному дослідженні алгоритм дає кращі результати (відносна похибка зростає логарифмічно), ніж аналогічні алгоритми (Евклідова відстань, ERP, DTW) для знаходження відстані між часовими рядами (відносна похибка зростає лінійно).

2. Запропонований метод знаходження відстані між вимірюваннями часового ряду дає кращі результати для великих часових рядів, коли кількість вимірювань $T > 1000$. До того ж обчислювальна складність отриманого алгоритму є меншою за обчислювальну складність уже існуючих алгоритмів.

У третьому розділі розглянуто проблему кластеризації на графах на основі власних значень стохастичної матриці графа. Доведено, що власні значення стохастичної матриці для великих графів ($N > 100$) поділяються на три групи, одна із яких є визначальною для числа кластерів у графі. Використовуючи теорію випадкових матриць, вдалося показати, що асимптотичний розподіл підгрупи дійсних частин власних значень стохастичної матриці графу описується напівколовим розподілом Вігнера. Використання стохастичних матриць дало змогу точно локалізувати власні значення, що відповідають за кількість кластерів, чого не вдавалося зробити для матриць суміжності. Основні припущення моделі пов'язані з властивостями дискретних ланцюгів Маркова, що дозволяє розширити область застосування отриманих результатів на більш широкий клас об'єктів. Теоретичні результати перевірені на кластеризації часових рядів,що описують вартості $N = 470$ акцій $S\&P500$ в період з 2013 до 2018 року.

Наукову новизну висновків, зроблених на основі отриманих у третьому розділі результатів, розкривають такі положення:

1. У роботі запропоновано новий метод визначення оптимальної кількості кластерів k_{opt} при кластеризації об'єктів, що задаються неструктурованими даними (графами та часовими рядами) на основі спектрального аналізу стохастичної матриці даного графу.

2. Використовуючи метод Монте-Карло, вдалося показати, що запропонований метод дає кращі результати для визначення оптимальної кількості кластерів k_{opt} у порівнянні із деякими класичними методами.

3. Оскільки запропонований алгоритм є спектральним, то його складність збігається зі складністю знаходження власних значень для стохастичної матриці P.

4. Описаний алгоритм не є чутливим до кластерів різного розміру,тобто співвідношення між розмірами кластерів практично не впливають на точність алгоритму.

5. Теоретичні результати роботи перевірено на реальних даних ($N = 470$ акцій $S\&P500$, розглянутих в період з 2013 до 2018 року. Результати оцінки оптимального значення k_{opt} збіглися із відповідними оцінками для даних компаній в інший період часу.

Практичне значення отриманих результатів

Питання про визначення відстані між вимірюваннями часових рядів (даних) та знаходження оптимальної кількості кластерів залишається відкритим. Досі не існує універсального підходу для визначення метрики подібності між часовими рядами та встановлення оптимальної кількості кластерів для даних, який однаково добре працює для наборів даних з різних сфер життєдіяльності людини. У даному дисертаційному дослідженні описано нові ідеї та підходи до розв'язання вище згаданих проблем. Результати дисертації можуть бути використані для поділу на групи (кластеризації) даних, які представлені графами або часовими рядами. Кластеризація дозволяє групувати подібні дані в категорії або кластери, що спрощує їхнє вивчення і використання у майбутньому.

Результати, отримані у даному дисертаційному дослідженні, можуть бути використані при кластеризації медичних даних: за допомогою аналізу симптомів і медичних даних можна класифікувати пацієнтів за різними хворобами або ступенями важкості захворювань; підбирати індивідуальні підходи до лікування на основі схожості пацієнтів і їх реакції на терапію; розробляти програми попередження захворювань і проводити цільові медичні обстеження.

Використання запропонованих у дисертаційному дослідженні підходів до рекламної галузі: рекламодавці можуть створювати кластери споживачів на основі їхніх інтересів, демографічних характеристик і поведінки, щоб розробляти більш ефективні рекламні кампанії; кластеризація даних допомагає рекламодавцям створювати персоналізовану рекламу для кожного сегмента аудиторії; аналіз кластерів споживачів допомагає передбачати попит на продукти і послуги в майбутньому.

В економіці кластеризація даних корисна для: дослідження конкурентної ситуації та сегментації ринку, що дозволяють компаніям розробляти ефективні стратегії маркетингу та розвитку; для оцінки ризику та портфельного управління; прогнозування економічних трендів та розвитку стратегії під них.

У наш час науковці вивчають генетичні схожості і родові зв'язки саме за допомогою кластеризації. Кластеризація може допомогти у виділенні регіонів зі схожим кліматом для дослідження змін клімату. За допомогою кластеризації у соціологічних та психологічних дослідженнях виділяють групи осіб зі схожими характеристиками для причинно-наслідкового аналізу поведінки.

Усі ці приклади підкреслюють важливість кластеризації даних у великій кількості галузей життєдіяльності людини. Завдяки кластеризації даних можна приймати правильні рішення в бізнесі, підвищувати ефективність виробництва у промисловості, оптимізувати роботу розумних мереж тощо.

Ключові слова: марковське перемикання, параметри регресії, динаміка, модель, моделювання, часові ряди, машинне навчання, нейронні мережі, збурене випадкове блукання, стохастичні диференціальні рівняння, слабка збіжність, стійкість, стохастична модель оптимізації, ройовий алгоритм, подібність кластерів.

ABSTRACT

Knignitska T. V. Estimates of parameters of autoregressive models. – Qualifying scientific project on manuscript rights.

Thesis for obtaining the scientific degree of Doctor of Philosophy in specialty 113 – Applied mathematics. – Yury Fedkovich Chernivtsi National University named after, Ministry of Education and Science of Ukraine, Chernivtsi, 2023.

The dissertation paper is devoted to finding the distances between data measurements, which are represented by time series, and determining the optimal number of clusters based on the eigenvalues of the stochastic matrix of the graph. The dissertation consists of an introduction, three sections, conclusions, and a list of used sources.

The introduction substantiates the relevance of the research topic, formulates the goal, task, subject, object, and methods of the research, indicates the scientific novelty, the practical significance of the results obtained, the connection of the work with scientific research and the personal contribution of the recipient, and also provides data about where the main results of the dissertation were reported, discussed and published.

In the first chapter a review of the scientific literature devoted to the study of time series, in particular, the definition of similarity metrics between time series and approaches to clustering data, which are presented in the form of unstructured data types, is carried out. The chronology of the development of scientific approaches to the problems of clustering, classification, and dimensionality reduction of time series is analyzed in detail. The first paragraph of Section 1 reflects a general overview of the development of scientific research in the study of time series and existing metrics for establishing similarity between time series. The second point describes the methods of researching structural

jumps in time series. In the third point, an overview of scientific research related to continuous time series is made. The selection of the optimal number of clusters when dividing the data into groups is presented in the fourth point.

The second Chapter suggests determining the similarity or distance between time series using time series models. Stationary $ARMA$ models are considered, as time series models. The resulting algorithm is compared with already existing metrics for finding distances in the case of an increase in time series measurements and in the case of an increase in the number of outliers in the input time series. The resulting algorithm has lower computational complexity than the DTW and ERP algorithms. The proposed distance can be used for clustering highly noisy data.

The scientific novelty of the conclusions drawn on the basis of the results obtained in the second chapter is revealed by the following provisions:

1. An algorithm for finding the distance between time series based on time series models is described. The resulting distance is more robust to outliers in the time series. In the case of an increase in the number of emissions, the algorithm proposed in the dissertation research gives better results (the relative error increases logarithmically) than similar algorithms (Euclidean distance, ERP, DTW) for finding the distance between time series (the relative error increases linearly).

2. The proposed method of finding the distance between time series measurements gives better results for large time series when the number of measurements $T > 1000$. In addition, the computational complexity of the obtained algorithm is lower than the computational complexity of already existing algorithms.

The third Chapter deals with the problem of clustering on graphs based on the eigenvalues of the stochastic matrix of the graph. It is proved that the

eigenvalues of the stochastic matrix for large graphs ($N > 100$) are divided into three groups, one of which is decisive for the number of clusters in the graph. Using the theory of random matrices, it was possible to show that the asymptotic distribution of the subgroup of the real parts of the eigenvalues of the stochastic matrix of the graph is described by the semicircular Wigner distribution. The use of stochastic matrices made it possible to precisely localize the eigenvalues responsible for the number of clusters, which could not be done for adjacency matrices. The main assumptions of the model are related to the properties of discrete Markov chains, which makes it possible to expand the scope of the obtained results to a wider class of objects. The theoretical results were tested on the clustering of time series describing the values of $N = 470$ shares of $S\&P500$ in the period from 2013 to 2018.

The scientific novelty of the conclusions drawn on the basis of the results obtained in the third chapter is revealed by the following provisions:

1. The paper proposes a new method for determining the optimal number of clusters k_{opt} when clustering objects given by unstructured data (graphs and time series) based on the spectral analysis of the stochastic matrix of the given graph.

2. Using the Monte Carlo method, it was possible to show that the proposed method gives better results for determining the optimal number of clusters k_{opt} in comparison with some classical methods.

3. Since the proposed algorithm is spectral, its complexity coincides with the complexity of finding eigenvalues for the stochastic matrix P.

4. The described algorithm is not sensitive to clusters of different sizes, that is, the ratio between the sizes of clusters practically does not affect the accuracy of the algorithm.

5. The theoretical results of the work were verified on real data ($N = 470$ shares of $S\&P500$, considered in the period from 2013 to 2018. The results of estimating the optimal value of k_{opt} coincided with the corresponding estimates for these companies in another time period.

Practical significance of the obtained results

The question of determining the distance between measurements of time series (data) and finding the optimal number of clusters remains open. There is still no universal approach for determining the similarity metric between time series and establishing the optimal number of clusters for data, which works equally well for datasets from different areas of human activity. This dissertation study describes new ideas and approaches to solving the above-mentioned problems. The results of the thesis can be used to divide the data into groups (clustering), which are represented by graphs or time series. Clustering allows you to group similar data into categories or clusters, which simplifies their further study and use.

The results obtained in this dissertation research can be used in the clustering of medical data: by means of the analysis of symptoms and medical data, it is possible to classify patients according to different diseases or degrees of severity of diseases; select individual approaches to treatment based on the similarity of patients and their response to therapy; develop disease prevention programs and conduct targeted medical examinations.

Applying the approaches proposed in the dissertation research to the advertising industry: advertisers can create clusters of consumers based on their interests, demographics, and behaviors to design more effective advertising campaigns; data clustering helps advertisers create personalized ads for each audience segment; analysis of consumer clusters helps predict future demand for products and services.

In economics, data clustering is useful for: competitive situation research and market segmentation, allowing companies to develop effective marketing and development strategies; for risk assessment and portfolio management; forecasting economic trends and developing strategies for them.

Nowadays, scientists study genetic similarities and ancestral relationships precisely with the help of clustering. Clustering can help in identifying regions with similar climates for climate change research. With the help of clustering in sociological and psychological research, groups of individuals with similar characteristics are distinguished for further analysis of behavior.

All these examples highlight the importance of data clustering in a large number of industries, where it helps in understanding and using complex data sets to make decisions, improve efficiency and achieve greater understanding of key issues.

Key words: Markov switching, regression parameters, dynamics, model, simulation, time series, machine learning, neural networks, perturbed random walk, stochastic differential equations, weak convergence, robustness, stochastic optimization model, swarm algorithm, cluster similarity.

Список публікацій за темою дисертації

Наукові праці у періодичних наукових виданнях, проіндексованих у наукометричній базі даних Scopus:

1. Knignitskaya T. V. Estimate of time series similarity based on models. Journal of Automation and Information Sciences. 2019. Vol. 51 (№8).

2. Pavlyukovich N., Pavlyukovich O.V., Dubolazov O.V., Ushenko Yu.A., Tomka Yu. Ya., Zabolotna N.I., Soltys I.V., Drin Ya.M., Knignitska T.V., Talakh M.V., Dovgun A.Ya., Kotyra A., and Kozbakova A. Methods and means of "single-point"phasometry of microscopic images of optical-anisotropic biological objects. Proceedings of SPIE – The International Society for Optical Engineering. Vol. 11176, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physic Experiments 2019, 1117630 (6 November 2019).

Наукові праці у виданнях, внесених до переліку наукових фахових видань України:

3. Кнігніцька Т.В., Малик І.В., Горбатенко М.Ю. Кластеризація: марковський алгоритм Буковинський математичний журнал. 2020. 7(2). С. 59-75.

Наукові праці, які додатково відображають наукові результати дисертації:

4. Doroshenko I., Knihnitska T., Deretorska T. Comparison of machine learning algorithms for predicting mortality from COVID-19 virus. SWorld Journal. 2022. 2(11-02). С. 72–77.

5. Іванчук М.А., Малик І.В., Кнігніцька Т.В., Лукашів Т.О. Статистичний аналіз відносних величин у медицині // Клінічна та експериментальна патологія. – Чернівці, 2019. – Том 18, 4(70), 109 – 114.

6. Малик І.В., Кнігніцька Т.В. Методи машинного навчання для ста-

тистичної обробки медичних даних. Науковий вісник Чернівецького національного університету. Серія: Комп'ютерні системи та компоненти. 2017. Том 8, випуск 2. – С. 77 – 85.

7. Knignitska T. «From The Practice To Theory» Or How To Interest The Students By Mathematics. Physical and Mathematical Education ISSN 2413-1571 (print), ISSN 2413-158X (online): scientific journal. 2017. Issue 4(14). – P. 199-204.

Наукові праці, які засвідчують апробацію матеріалів дисертації:

1. Книгніцька Т.В. Підбір оптимальних параметрів для однієї задачі кластеризації. Міжвузівський науковий семінар "Прикладні задачі та ІТ-технології", присвячений 100-річчю з дня народження професора В.П. Рубаника (1917-1993) і 55-річчю кафедри прикладної математики та інформаційних технологій: матеріали семінару, 9 – 10 червня 2017 р. Чернівці: 2017. С. 64 – 65.

2. Knignitska T. Cluster analysis in data mining. Scientific Conference of Doctoral Students C̈ontemporary trends in the development of science: visions of young researchers̈: Materials of the Scientific Conference of Doctoral Students, 6th Edition, June 15, 2017. Volume 1. Universities of the Academy of Sciences of Moldova. P. 30 – 35.

3. Книгніцька Т.В., Малик І.В. Вибір оптимальної моделі для аналізу часових рядів. Праці VI-ї Міжнародної науково-практичної конференції «Проблеми інформатики та комп'ютерної техніки» (ПІКТ – 2017) (Чернівці, 5-8 жовтня 2017 року). Чернівці: Видавничий дім «Родовід», 2017. С. 32-33.

4. Knignitska T., Malyk I.V. Method for Evaluating Time Series Similarity. Сучасні проблеми математики та її застосування в природничих науках і інформаційних технологіях: матеріали міжнародної наукової конференції, присвяченої 50-річчю факультету математики та інформатики Чернівe-

цького національного університету імені Юрія Федьковича, 17 – 18 вересня 2018 р. Чернівці: 2018. С. 128.

5. Кнігніцька Т.В., Малик І.В., Лукашів Т.О. Алгоритми знаходження відстаней між часовими рядами. Праці VII-ї Міжнародної науково-практичної конференції «Проблеми інформатики та комп'ютерної техніки» (ПІКТ – 2018) (Чернівці, 11-14 жовтня 2018 року). Чернівці: Видавничий дім «Родовід», 2018. С. 27-29.

6. Кнігніцька Т.В., Малик І.В., Лукашів Т.О. Порівняння алгоритмів знаходження відстаней між часовими рядами. Праці VIII-ї Міжнародної науково-практичної конференції «Проблеми інформатики та комп'ютерної техніки» (ПІКТ – 2019) (Чернівці, 3-6 жовтня 2019 року). Чернівці: Видавничий дім «Родовід», 2019. С. 30-31.

7. Kyrychenko O.L., Knignitska T.V., Ostapov S.E. Stochastic models in artificial intelligence development. International conference "Modern stochastics: theory and applications V". Kyiv, June 1–4, 2021. P. 35.

8. Кнігніцька Т.В., Малик І.В. Оптимальна комбінація прогнозів для ієрархічних часових рядів. Міжнародна наукова конференція "Диференціально-функціональні рівняння та їх застосування" присвячена 80-річчю від дня народження професора В.І. Фодчука (1936–1992): матеріали конференції, 28 – 30 вересня, 2016. Чернівці: 2016. – С. 56.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

Абревіатура	**Трактування**
AIC	Інформаційний критерій Акаіке
ANOVA	Аналіз дисперсій
AR	Модель авторегресії
ARIMA	Інтегрована модель авторегресії ковзного середнього
ARMA	Модель авторегресії ковзного середнього
CARMA	Модель авторегресії ковзного середнього з неперервним часом
CTARMA	модель авторегресії ковзного середнього з неперервним часом та пороговим значенням
DBSCAN	Алгоритм просторової кластеризації, який ґрунтується на щільності даних
DNN	Глибокі нейронні мережі
DTW	Динамічна трансформація часової шкали
EDR	Відредагована відстань у реальній послідовності
ERP	Відредагована відстань з реальним штрафом
GARCH	Узагальнена авторегресійна модель умовної гетероскедастичності
MLE	Оцінка методу максимальної правдоподібності
LSE	оцінка методу найменших квадратів
MA	ковзне середнє
MAP	метод максимальної апостеріорної правдоподібності
MCL	Алгоритм кластеризації Маркова
MCMC	Методи Монте-Карло з Марковськими ланцюгами
RMT	Теорія випадкових матриць
VAR	Векторна авторегресія

ВСТУП

Актуальність теми дослідження. Моделі штучного інтелекту на даний час є найбільш поширеними в сфері прикладної математики, оскільки дозволяють підвищувати ефективність реальних процесів [1, 2]. Найбільш поширені задачі відносяться до задач класифікації, кластеризації та обробки великих даних, що зумовлює розв'язання допоміжних підзадач в даних областях штучного інтелекту [3]. При розгляді вищезгаданих задач основну роль відіграють поняття групи або кластеру, метрики в просторі об'єктів та оптимізаційних задач, що дозволяють розв'язувати задачі штучного інтелекту в цілому, так і задач машинного навчання та глибинного навчання зокрема.

Основні класичні методи штучного інтелекту стрімко розвиваються з 60-70 років минулого століття [1] внаслідок комп'ютеризації суспільства і виробництва. Першими задачами даного напрямку можна вважати задачі класифікації та кластеризації, які включають багато інших підзідач, пов'язаних з ними, для прийняття рішення групування про об'єктів. До таких підзадач відносяться питання підбору метрики в просторі об'єктів та вибір оптимальної кількості кластерів в задачах кластеризації [4, 5, 6] . Крім того, задачі пошуку метрики в просторі об'єктів ускладняються за умови класифікації та кластеризації неструктурованих даних, до яких можна віднести часові ряди, графи, тексти, рисунки тощо.

Задачами машинного навчання займалися та займаються багато видатних українських та зарубіжних науковців, серед яких слід відзначити роботи A. Ethem, A. Y. Ng., P. Hart, W.S. Sarle, T.M. Mitchell, C. Bishop, В. Вапніка, О.Червоненкіса, В. Марченка та Л. Пастура. Дані роботи присвячені як аналізу основник класичних моделей і створенню нових моделей, так і дослідженню суміжних теоретичних результатів, що дозволяють

більш точно описувати топологію реальних даних. Основна увага роботи присвячена задачам, суміжним із задачами класифікації та кластеризації, а саме задачам покращення метрики в просторі часових рядів та проблемі пошуку оптимальної кількості кластерів в задачах кластеризації.

Мета дослідження полягає у розв'язанні двох підзадач кластеризації даних – визначенні відстані між вимірюваннями даних та знаходженні оптимальної кількості кластерів даних при розбитті вхідного набору неструктурованих даних на підгрупи. Відносна похибка вимірювань для отриманої метрики повинна бути меншою у порівнянні з уже існуючими метриками при зростанні кількості викидів у неструктурованих даних та збільшенні кількості вимірювань часових рядів (для довгих часових рядів, кількість вимірювань є більшою за 1000). Порівняти запропоновані агоритми з уже існуючими на прикладі реальних даних та даних, симульованих за допомогою методу Монте-Карло.

Завдання, які виконувалися для досягнення сформульованої мети: 1. Розробити нову метрику в просторі стаціонарних часових рядів, яка має ряд переваг над класичними моделями метрик в просторі часових рядів та яка є більш стійкою до викидів і дає більш точні результати для часових рядів з великою кількістю вимірювань. 2. Побудувати новий метод визначення оптимальної кількості кластерів при розгляді задач кластеризації об'єктів, що задаються неструктурованими даними (графами, часовими рядами тощо) на основі спектрального аналізу стохастичної матриці графу. 3. Використати метод Монте-Карло для порівняння запропонованого методу підбору кількості кластерів у порівнянні з рядом класичних методів. 4. Визначити, чи розроблений алгоритм знаходження оптимальної кількості кластерів є менш чутливим до наявності кластерів різного розміру. 5. Зменшити обчислювальну складність алгоритму знаходження відстані між даними з використанням запропонованої метрики для N часових рядів.

Об'єктом дослідження є реальні дані, представлені часовими рядами. Будь-який тип інформації, представлений у вигляді впорядкованої послідовності, є часовим рядом. Його можна визначити як сукупність спостережень за одним суб'єктом, зібраних протягом різних, як правило, однакових проміжків часу. Іншими словами, об'єктом дослідження є дані, зібрані в різні моменти часу та організовані в хронологічному порядку. Часові ряди представляють інформацію про поширення захворювань у медицині, про прогнозування курсу валют у економіці, про групування користувачів при розсилці реклами у маркетингу тощо.

Предметом дослідження є метрики визначення відстаней між вимірюваннями неструктурованих даних, методи визначення оптимальної кількості кластерів та оцінка параметрів моделей часових рядів.

Методи дослідження. Пошук схожості, який включає визначення ступеня подібності між двома або більше наборами даних часових рядів, є фундаментальним завданням аналізу часових рядів. В аналізі часових рядів існують численні підходи для пошуку подібності, включаючи евклідову відстань, динамічну деформацію часу (DTW, Dynamic Time Warping) та відстань зі штрафом (EDR, Edit Distance with Real penalty). Обраний підхід визначається індивідуальною метою, обсягом і складністю збору даних, а також кількістю шуму та викидів у даних. Метод, описаний у дисертаційному дослідженні Книгніцької Т.В. передбачає використання ARMA(p,q) моделей часових рядів для визначення відстані між самими часовими рядами. У такому випадку відстань знаходиться не між вимірюваннями даних, як це відбувається у класичних вищезгаданих методах, а за допомогою параметрів процесів авторегресії та ковзного середнього часових рядів. Метод ліктя (Elbow Method) та метод силуету (Silhouette Method) є найбільш популярними методами для встановлення оптимальної кількості кластерів. Основна ідея методу ліктя полягає в тому, що зі

збільшенням кількості кластерів варіація всередині кожного кластера зменшується. Метод силуету передбачає, що кожен кластер представлений так званим силуетом, який заснований на порівнянні його щільності та розділеності. Тому, зазвичай, на практиці, ці методи показують різні результати. Для визначення оптимальної кількості кластерів у дисертаційному дослідженні Кнігніцької Т.В. використано стохастичні матриці графу та властивості дискретних ланцюгів Маркова. Використання стохастичних матриць графу дає змогу точно локалізувати власні значення, що відповідають за кількість кластерів. Отримані у дисертаційному дослідженні методи порівнюються з уже існуючими алгоритмами визначенні відстані та методами знаходження оптимальної кількості кластерів.

Зв'язок роботи з науковими програмами, планами, темами

Обраний напрям дослідження відповідає програмі наукової тематики кафедри прикладної математики та інформаційних технологій Чернівецького національного університету імені Юрія Федьковича "Математичне моделювання і числово-аналітичні методи дослідження динамічних та інформаційних процесів".

Наукова новизна одержаних у роботі результатів полягає в тому, що для розв'язання підзадач кластеризації уперше:

1. Запропоновано нову метрику для знаходження міри подібності в просторі стаціонарних часових рядів, представлених моделями $ARMA(p, q)$. Отримана метрика базується на параметрах моделі, а не на самих вимірюваннях часових рядів.

Для початку вхідні дані необхідно нормалізувати, підібрати порядок для параметрів процесу авторегресії p та процесу ковзного середнього q. Вибір вказаних параметрів здійснено на основі автокореляційної та частинної автокореляційної функції часового ряду.

2. Здійснено порівняння запропонованої метрики з класичними моде-

лями метрик в просторі часових рядів. 3. Показала, що отримана метрика є більш стійкою до викидів і дає більш точні результати для часових рядів з великою кількістю вимірювань.

4. Встановлено, що складність алгоритму обчислення з використанням запропонованої метрики для N часових рядів складає $O(T * N^2)$, в той же час аналогічна складність алгоритмів DTW, ERP становить $O(T^2N^2)$. За рахунок стійкості до викидів дана метрика дозволяє отримувати більш стійкі до шумів кластери.

5. Запропоновано новий метод визначення оптимальної кількості кластерів при розгляді задач кластеризації об'єктів, що задаються неструктурованими даними (графами, часовими рядами тощо) на основі спектрального аналізу стохастичної матриці даних.

6. Показано, що дійсні частини власних значень стохастичної матриці графу можна розділити на три групи. До першої групи відноситься 1, так як вона завжди присутня серед власних значень. До другої групи відносяться власні значення стохастичної матриці, які є близькими до нуля, але не є нулями. До третьої групи відносяться ті власні значення, які знаходяться між 0 та 1. Якраз кількість власних значень у третій групі і відповідає оптимальній кількості кластерів для вхідних даних.

7. За допомогою симуляції методом Монте-Карло, показано, що запропонований метод підбору кількості кластерів дає кращі результати для визначення оптимальної кількості кластерів у порівнянні з рядом класичних методів (Марковський алгоритм з двома типами параметрів та метод ліктя). Симуляція Монте-Карло використана для утворення багатовимірних даних — графу з фіксованою кількістю сукупностей (кластерів). Таким чином для порівняння запропонованого методу вибору оптимальної кількості кластерів з Марковським алгоритмом та методом ліктя оптимальна кількість кластерів була наперед заданою.

8. Встановлено, що розроблений алгоритм знаходження оптимальної кількості кластерів є менш чутливим до наявності кластерів різного розміру.

Практичне значення отриманих результатів

Практичне значення роботи важко переоцінити. Результати, отримані у дисертаційному дослідженні, можуть бути використані при кластеризації медичних даних: за допомогою аналізу симптомів і медичних даних можна класифікувати пацієнтів за різними хворобами або ступенями важкості захворювань; підбирати індивідуальні підходи до лікування на основі схожості пацієнтів і їх реакції на терапію; розробляти програми попередження захворювань і проводити цільові медичні обстеження. Важливо відзначити, що спеціаліст з аналізу даних не приймає рішення про методи лікування, а лише надає статистичні результати та висновки спеціалісту-медику.

Використання запропонованих у дисертаційному дослідженні підходів до рекламної галузі: рекламодавці можуть створювати кластери споживачів на основі їхніх інтересів, демографічних характеристик і поведінки, щоб розробляти більш ефективні рекламні кампанії; кластеризація даних допомагає рекламодавцям створювати персоналізовану рекламу для кожного сегмента аудиторії; аналіз кластерів споживачів допомагає передбачати попит на продукти і послуги в майбутньому.

У економіці кластеризація даних корисна для: дослідження конкурентної ситуації та сегментації ринку, що дозволяють компаніям розробляти ефективні стратегії маркетингу та розвитку; для оцінки ризику та портфельного управління; прогнозування економічних трендів та розвитку стратегії під них.

У наш час науковці вивчають генетичні схожості і родові зв'язки саме за допомогою кластеризації. Кластеризація може допомогти у виділенні регіонів з схожим кліматом для дослідження змін клімату. За допомогою

кластеризації у соціологічних та психологічних дослідженнях виділяють групи осіб зі схожими характеристиками для подальшого аналізу поведінки.

Усі ці приклади підкреслюють важливість кластеризації даних у великій кількості галузей, де вона допомагає у зрозумінні і використанні складних наборів даних для прийняття рішень, підвищення ефективності та досягнення більшого розуміння ключових питань.

Особистий внесок здобувача. У всіх працях дисертантка брала участь в обговоренні постановки задачі, визначенні мети роботи, виборі методів досліджень і підготовці матеріалів до публікації у наукових журналах та представленні на наукових конференціях. Основні результати та висновки обговорювалися з науковим керівником – доктором фіз.-мат. наук, професором Маликом І.В. (Чернівецький національний університет імені Юрія Федьковича).

Зокрема, у [7] дисертантка описала новий алгоритм знаходження відстані між часовими рядами, який базується на моделях часових рядів. У роботі [8] використано запропонований алгоритм [7] при аналізі фізичних даних. Алгоритм для знаходження оптимальної кількості кластерів описано у [9]. Дослідження [10] є прикладом застосування аналізу часових рядів до прогнозування смартності від вірусу COVID-19. Дисертантка описала основні статистичні підходи до аналізу медичних даних у роботі [11]. Порівняння методів машинного навчання описано у дослідженнях [12, 13].

Апробація матеріалів дисертації

1. Міжвузівський науковий семінар “Прикладні задачі та ІТ-технології”, присвячений 100-річчю з дня народження професора В.П. Рубаника (1917-1993) і 55-річчю кафедри прикладної математики та інформаційних технологій, 9 – 10 червня 2017, Чернівці, Україна.

2. Scientific conference of doctoral students "Contemporary trends in the

development of science: visions of young researchers June 15, 2017, Moldova.

3. VI Міжнародна науково-практична конференція «Проблеми інформатики та комп'ютерної техніки» (ПІКТ – 2017), 5-8 жовтня 2017, Чернівці, Україна.

4. Міжнародна наукова конференція «Сучасні проблеми математики та її застосування в природничих науках і інформаційних технологіях», присвяченої 50-річчю факультету математики та інформатики Чернівецького національного університету імені Юрія Федьковича, 17 – 18 вересня, 2018, Чернівці, Україна.

5. VII Міжнародна науково-практична конференція «Проблеми інформатики та комп'ютерної техніки» (ПІКТ – 2018), 11-14 жовтня, 2018, Чернівці, Україна.

6. VIII Міжнародна науково-практична конференція «Проблеми інформатики та комп'ютерної техніки» (ПІКТ – 2019), 3-6 жовтня, 2019, Чернівці, Україна.

7. International conference "Modern stochastics: theory and applications V", June 1–4, 2021, Kyiv, Ukraine.

8. Міжнародна наукова конференція "Диференціально-функціональні рівняння та їх застосування присвячена 80-річчю від дня народження професора В.І. Фодчука (1936–1992), 28 – 30 вересня, 2016, Чернівці, Україна.

Структура й обсяг дисертації Дисертаційне дослідження складається з анотацій двома мовами, списку опублікованих праць автора, переліку умовних позначень, вступу, трьох розділів, висновків, списку використаних джерел (193 позицій) і Додатку (список публікацій здобувачки за темою дисертації).

Загальний обсяг роботи – 152 сторінки, робота містить 2 таблиці та 10 рисунків.

РОЗДІЛ І. ОГЛЯД ІСНУЮЧИХ ДОСЛІДЖЕНЬ

Розвиток підходів до аналізу часових рядів та прогнозування є поетапним та пов'язаним з різними досягненнями в науці та технологіях. Робота в галузі статистики почалася у XIX столітті з розвитком методів для оцінки середнього руху та використання лінійної регресії для моделювання трендів у часових рядах. У XX столітті були розроблені моделі авторегресії та ковзного середнього, які дозволили більш точно моделювати та прогнозувати часові ряди. Застосування методів аналізу автокореляції та часткової автокореляції стало важливим кроком для виявлення залежностей між значеннями в часовому ряді.

XXI століття розпочалося з розвитку більш складних моделей, таких як векторна авторегресійна модель [14] (Vector Autoregression, VAR), GARCH [15] (узагальнена авторегресійна модель гетероскедастичності) та інші. Застосування такого роду моделей стало можливим завдяки розвитку комп'ютерів та програмного забезпечення для аналізу даних. На піку популярності у даний час знаходяться методи машинного навчання, такі як рекурентні нейронні мережі (Recurrent neural network, RNN) [16], глибокі нейронні мережі (Deep Neural Networks, DNN) [17] та їх використання для аналізу та прогнозування часових рядів. Природно, що всі вище згадані підходи є тісно пов'язаними між собою – розвиток певного напрямку досліджень спонукає нові відкриття у суміжних напрямках. Розвиток цієї галузі продовжується. Науковці та аналітики поєднують та порівнюють традиційні методи з новими технологіями для більш точного та ефектив-

ного аналізу часових рядів.

У даному розділі розглянуто розвиток наукових підходів до аналізу часових рядів, огляд існуючих метрик подібності між вимірюваннями часових рядів, визначення оптимальної кількості кластерів та напрямки застосування вказаних теоретичних напрацювань у різних сферах життєдіяльності людини.

1.1. Міри подібності між часовими рядами

Теорія багатьох методів машинного навчання фундаментально ґрунтується на концепції вимірювання відстаней. Одним із таких важливих показників відстані є відстань Махаланобіса [18], яка, якщо досліджувати її в контексті вихідного та головного компонентів, дає безцінне розуміння. Maesschalck та інші розглядали відстань Махаланобіса та її зв'язок із більш загальновідомою Евклідовою відстанню (метрикою). Крім того, автори описали різні методи, які використовують відстань Махаланобіса при застосовувані в різних сферах, включаючи багатовимірне калібрування, розпізнавання образів і керування процесом. Взагалі, відстань Махаланобіса -– це показник, який використовується для кількісного визначення відмінності між двома точками даних у багатовимірному наборі даних, враховуючи кореляції між змінними. На відміну від Евклідової відстані, яка розглядає змінні як некорельовані та однаково зважені, відстань Махаланобіса враховує коваріаційну структуру даних. Як наслідок, це забезпечує точнішу міру несхожості в багатовимірному просторі.

Відстань Махаланобіса фактично масштабує Евклідову відстань за допомогою коваріаційної матриці, дозволяючи їй враховувати змінні кореляції. Це коригування особливо цінне, коли розглядаються набори даних, де змінні не є незалежними. Ключова відмінність між Евклідовою відстанню та відстанню Махаланобіса полягає в тому, як вони обробляють змінні

та їхні кореляції. Евклідова відстань розглядає змінні як некорельовані та надає однакову важливість кожній змінній. І, навпаки, відстань Махаланобіса коригує змінні кореляції, враховуючи їх під час вимірювання відстані між точками даних. Це робить відстань Махаланобіса особливо корисною під час роботи з багатовимірними наборами даних, де змінні можуть бути взаємопов'язані.

Розпізнавання образів передбачає класифікацію об'єктів або зразків у заздалегідь визначені категорії на основі їхніх виміряних характеристик. Відстань Махаланобіса є цінним інструментом для завдань розпізнавання образів, оскільки вона враховує змінні кореляції. Це допомагає ефективно розрізняти різні класи або групи, враховуючи структуру коваріації даних. У промислових процесах моніторинг і контроль є важливими для забезпечення якості продукції та ефективності процесу. Відстань Махаланобіса використовується для виявлення аномалій і відхилень від нормальних умов експлуатації. Аналізуючи відстань Махаланобіса між новими вимірюваннями та історичними даними, відхилення процесу можна виявити на ранній стадії, дозволяючи вчасно вжити коригувальні дії. Розуміння та використання відстані Махаланобіса дає змогу дослідникам і практикам отримувати значущі ідеї з багатовимірних даних, що в кінцевому підсумку сприяє розвитку наукових і промислових підходів.

Міри подібності є фундаментальними при здійсненні кластеризації, оскільки вони визначають поняття близькості між точками даних. Різні показники подібності враховують різні аспекти зв'язків даних, і їх вибір залежить від характеристик набору даних і цілей аналізу. Для даних часових рядів, де часова кореляція часто є значущим фактором, вибір відповідної метрики подібності особливо важливий. Традиційна Евклідова відстань може не враховувати належним чином часові залежності та кореляції в даних, як уже було зазначено вище. Таким чином, альтернативні показники

подібності, такі як динамічна трансформація часової шкали (Dynamic Time Warping, DTW) або коефіцієнт кореляції Пірсона, стають важливими інструментами для аналізу кластеризації. У контексті побудови моделей енергоспоживання надзвичайно важливим є розуміння часової кореляції між точками даних споживання енергії. Для енергоефективності будівель ключовими є такі фактори, як щоденний графік використання приміщень, погодні умови та кількість людей, які перебувають у будівлі. Визначення цих моделей і групування подібних точок даних може призвести до більш точних моделей прогнозування і більш ефективних стратегій управління енергією.

На додаток до вивчення впливу різних показників подібності, стаття [19] представляє нову техніку перевірки під назвою "баланс кластерного вектора". Цей метод спрямований на оцінку та порівняння продуктивності алгоритмів кластеризації. Кластерний векторний баланс оцінює розподіл точок даних у кластерах і визначає, чи вони рівномірно збалансовані, чи перекошені. Рівномірний розподіл точок даних між кластерами вказує на надійне та ефективне рішення кластеризації, тоді як незбалансований розподіл може означати неоптимальний результат.

Вимірювання та аналіз деформації часової шкали для визначення подібності між часовими рядами представлені також у роботі [20]. У праці автора досліджується використання різних методів, включаючи мінімальну норму (L_1), найменші квадрати (L_2), фільтрацію Калмана та аналіз часових рядів (зокрема, моделі Бокса-Дженкінса) у контексті визначення деформації. Автор вирішує проблеми, пов'язані з отриманням точних сигналів із даних деформації, особливо при роботі з викидами та екологічними порушеннями, такими як вітер, дощ і коливання температури. На попередніх етапах аналізу деформації використовуються такі традиційні методи, як мінімальна норма (L_1) і методи найменших квадратів (L_2), які

часто доповнюються підходами надійної оцінки. Однак ці традиційні методи мають властиві обмеження при застосуванні до неперервних вимірювань деформації, зібраних за допомогою моторизованих тахеометрів, також відомих як роботи-геодезисти. Ці обмеження випливають із неврахування кореляції даних, яка є фундаментальною характеристикою даних часових рядів.

У зв'язку зі збільшенням можливостей накопичувальних пристроїв, зазвичай доводиться мати справу з високочастотними даними. Проблеми та підходи, пов'язані з виявленням знань у базах даних, зокрема в контексті розробки комп'ютерних засобів для дослідження великих сховищ даних, обговорено у роботі [21]. Електронні архіви даних швидко розширюються та охоплюють різноманітні типи даних з різних областей, включаючи комерційний і науковий сектори. Значна частина таких даних за своєю суттю є часовими рядами, як-от курс акцій S&P500 або дані телеметрії NASA. Основна увага у статті [21] зосереджена на завданні виявлення патернів у таких часових потоках даних, що є ключовим аспектом одержання знань.

Автори описали попередні експерименти, які включають підхід на основі динамічного програмування для виявлення шаблонів у часових даних. Основний алгоритм, який використовується для цієї мети, базується на техніці динамічного викривлення часу DTW (трансформація часової шкали), методі, який зазвичай використовується в галузі розпізнавання мови. У дослідженні вказано, що цей підхід розглядається як потенційне рішення для виявлення закономірностей у часових потоках даних, що дає змогу зазирнути в поточні дослідження чи експерименти в цій галузі.

Дослідження [15] присвячене представленню нової функції для визначення подібності часових рядів під назвою "редагування відстані із реальним штрафом" (Edit Distance with Real Penalty, ERP). ERP є значним доповненням до аналізу часових рядів, оскільки вона ефективно долає розрив

між L_1-нормою та дистанцією коригування. Вона демонструє характеристики як норми L_1, так і відстані коригування, що робить його цінною метричною функцією відстані для обробки даних часових рядів. ERP пропонує переваги метричної функції відстані, яка узгоджується з існуючими функціями відстані, такими як Dynamic Time Warping (DTW) і Longest Common Subsequence (LCSS), які зазвичай використовуються для аналізу часових рядів. Визначальною особливістю ERP є її здатність керувати локальним зсувом часу, що є ключовим аспектом аналізу даних часових рядів.

Крім того, у роботі [15] представлено нову нижню межу для функції відстані ERP. На відміну від існуючих нижніх меж, які вимагають багатовимірного індексування, ця запропонована нижня межа є одновимірною та може бути легко реалізована в структурі $B+$ дерева. Ця реалізація не тільки зменшує вимоги до пам'яті, але й мінімізує потенційні витрати на введення-виведення, сприяючи ефективності пошуку подібності часових рядів. Комбінація методів скорочення, розглянута в даній статті, включаючи використання нерівності трикутника та нижніх меж, представляє інноваційний підхід до індексування даних часових рядів. Це об'єднання можна розглядати як розширення структури GEMINI, що пропонує покращений метод для ефективного управління та запиту даних часових рядів у метричному просторі.

Таким чином, автори розробили надійну функцію метричної відстані для пошуку подібності часових рядів. Крім того, запропонована функція представляє нову одновимірну нижню межу, яку можна легко інтегрувати в структури індексування, такі як $B+$ дерева, підвищуючи ефективність операцій зберігання та запитів. Підхід, запропонований у статті [15], у наш час є значним прогресом у галузі індексування та запиту даних часових рядів за допомогою метричних функцій відстані.

Пізніше науковець Lei Chen та ін. зосередилися на вирішальному аспекті пошуку подібності траєкторій рухомих об'єктів – визначенні стійкої функції відстані [22]. Ключова проблема в цьому контексті полягає в тому, що існуючі функції відстані, як правило, чутливі до різних недосконалостей у даних траєкторії, таких як шум, зсув у часі та масштабування. На практиці такі недоліки часто виникають через збої датчиків, помилки в техніці виявлення, зовнішні сигнали перешкод або коливання частоти дискретизації. Не завжди можливо правильно очистити дані, щоб усунути лише вказані недоліки. Щоб пом'якшити ці проблеми, у статті [22] представлено нову функцію відстані – відредаговану відстань у реальній послідовності (Edit Distance on Real sequence, EDR). EDR спеціально розроблено для захисту від недосконалості даних, що робить його добре придатним для аналізу та порівняння траєкторій рухомих об'єктів, на які впливають шум, зсув по часовій шкалі та масштабування.

У статті [22] проведено поглиблений аналіз та порівняння EDR з кількома популярними функціями відстані, які зазвичай використовуються в цій області. До них належать Евклідова відстань, динамічне трансформація часової шкали (DTW), редагування відстані з реальним штрафом (ERP) і найдовші загальні підпослідовності (LCSS). Результати цього порівняльного аналізу показують, що EDR перевершує Евклідову відстань, DTW і ERP з точки зору стійкості. Крім того, показано, що EDR в середньому на 50% точніший, ніж LCSS. На додаток до введення функції відстані EDR, у статті запропоновано три методи скорочення, спрямовані на підвищення ефективності пошуку EDR. Ці методи призначені для оптимізації процесу пошуку шляхом ефективного звуження потенційних збігів. Експериментальні результати підтверджують ефективність цих комбінованих методів обрізання, підкреслюючи їх високу ефективність у відновленні траєкторій рухомих об'єктів на основі подібності.

Коротко кажучи, дослідження Lei Chen та ін. стосується критичної проблеми визначення стійкої функції відстані для пошуку траєкторій рухомих об'єктів, на які впливають різні недосконалості даних. Воно представляє функцію відстані EDR як рішення, яке чудово справляється з перешкодами та недосконалими даними траєкторії. Порівняльний аналіз і експериментальні результати, наведені в статті, обґрунтовують ефективність EDR порівняно з іншими популярними функціями відстані та підкреслюють підвищення ефективності, досягнуте за допомогою запропонованих методів.

Ще однією мірою подібності між двома наборами або множинами даних є коефіцієнт Танімото (Tanimoto coefficient), також відомий як Jaccard coefficient або Jaccard similarity. Коефіцієнт Танімото визначає, наскільки схожі дві множини, порівнюючи кількість спільних елементів з загальною кількістю елементів у обох множинах. Дана міра широко застосовується до хімічних структур. У дослідженні [23] наведено новий простий доказ того, що відстань Танімото задовольняє нерівність трикутника. Нерівність трикутника відображає геометричний принцип, згідно з яким найкоротший шлях між двома точками – це пряма лінія. Якщо цей принцип не виконується, то виникають геометрично неможливі ситуації. Нерівність трикутника допомагає гарантувати, що вимірювання відстаней мають сенс і коректно використовуються в аналізі даних.

Розроблену теорію щодо метрик визначення відстаней між наборами даних часових рядів вдало поєднали Mori та ін. у роботі [24], представивши пакет «TSdist», який служить комплексним інструментом для обчислення широкого діапазону показників подібності для даних часових рядів у середовищі програмування R. Автори обговорили важливість визначення міри відстані між даними часових рядів, що відіграє вирішальну роль у різних завданнях аналізу даних, таких як кластеризація та класифікація. У дослідженні представлено різноманітність способів вимірювань відстані

в часових рядах, які були розроблені протягом багатьох років для задоволення різних аналітичних потреб. Запропонований пакет «TSdist» має на меті забезпечити уніфіковану структуру для обчислення широкого спектру вимірювань подібності часових рядів, доступних у R, об'єднуючи ці показники в єдиний інструмент. Пакет містить деякі популярні заходи вимірювання відстані, які раніше були недоступні в R, розширюючи діапазон опцій, доступних для аналітиків.

Окрім впровадження нових опцій, пакет TSdist також пропонує оболонки для функцій, які вже включені в інші пакети R, спрощуючи процес роботи з мірами подібності. TSdist виходить за рамки обчислення відстані, підтримуючи застосування методів знаходження відстані в кластеризації та класифікації. Пакет TSdist не лише розширює можливості R, вводячи нові показники, а також надає уніфіковану платформу для роботи з існуючими показниками, що робить його цінним ресурсом для аналітиків і дослідників даних, які працюють з даними часових рядів.

1.2. Структурні стрибки у часових рядах

Багато реальних процесів, які представлені часовими рядами або графами, характеризуються різкою зміною у поведінці тренду. Такі зміни можуть бути зумовлені економічною кризою у світі, локдауном через поширення певного вірусу, політичними рішеннями, змінами у законодавстві, технологічними інноваціями, демографічними трансформаціями тощо . Внаслідок цього виникають так звані структурні розриви або стрибки у часових рядах. У такому випадку послідовність даних можна представляти у вигляді послідовності функцій або моделей часового ряду [25]. У роботі [25] описано деякі підходи для визначення структурних розривів у моделях часових рядів. Зокрема, автори показали, як процедури, засновані на куму-

лятивній сумі, можуть бути модифіковані для роботи з даними, що мають послідовну залежність. Охоплено як структурні розриви в безумовному та умовному середньому, так і в структурі дисперсії та коваріації/кореляції. Процедури кумулятивної суми є непараметричними. Якщо дані дозволяють параметричне моделювання, автори стверджують, що підходи ймовірності можуть бути використані для відновлення структурних розривів. Крім того, у роботі [25] описано оцінку множинних структурних розривів та спосіб роз'єднання структурних розривів.

Методологічні міркування в контексті оцінки, тестування та обчислення для моделей, які включають структурні зміни або стрибки у часовому ряді, представлено у статті [26]. Основна мета дослідження полягає в тому, щоб вивчити розробки та їх актуальність в економетричних додатках, зокрема в рамках лінійних моделей. У праці досліджується значний прогрес, досягнутий у розширенні сфери застосування моделей для різноманітних практичних напрямків. Ця інклюзивність поширюється на моделі, що охоплюють загальні стаціонарні регресори, помилки, що виявляють часову залежність і гетероскедастичність, змінні трендів, можливі одиничні корені, коінтегровані моделі тощо. У статті також розглядаються удосконалення обчислень, що стосуються побудови оцінок, пов'язаних із ними граничних розподілів, процедур тестування для виявлення структурних змін і методологій для визначення кількості цих змін у наборі даних.

Автори розглянули широкий спектр тем, включаючи останні розробки, такі як тестування на загальні розриви, моделі з ендогенними регресорами (з наголосом на перевагу методу найменших квадратів над методами інструментальних змінних), квантильні регресії, підходи на основі Ласо, панельні моделі даних, тестування змін у точності прогнозу, факторні моделі та методи висновку, засновані на асимптотичній системі неперервних записів. Основна увага статті зосереджена на «офлайнових методах», які

стосуються ретроспективного тестування структурних розривів у заданому наборі даних. Ці методи дозволяють будувати довірчі інтервали навколо передбачуваних дат структурного стрибка.

Дослідження [27] заклало основу для застосування статистичних моделей для розуміння та прогнозування даних часових рядів, які являють собою послідовності спостережень, зібраних протягом певного часу. У 1978 році Akaike описав використання параметричних моделей для аналізу та контролю даних часових рядів. Уже майже 50 років ця робота є значним внеском у галузі статистики та аналізу часових рядів. Одним із помітних внесків Akaike є розробка інформаційного критерію Akaike (Akaike information criterion, AIC), який є інструментом для вибору та порівняння моделей. AIC забезпечує кількісний спосіб оцінки відповідності різних параметричних моделей заданому набору даних часових рядів. AIC врівноважує компроміс між складністю моделі та придатністю, дозволяючи дослідникам вибрати найбільш прийнятну модель для своїх даних.

Підхід Akaike до аналізу часових рядів із використанням параметричних моделей працює наступним чином:

- На першому етапі відбувається збір даних часового ряду, який зазвичай складається зі спостережень, записаних через регулярні проміжки часу. Це можуть бути дані про ціни акцій, погодні показники, економічні показники або будь-які інші дані, які змінюються з часом.

- На другому етапі відбувається підбір моделі: підхід Akaike передбачає вибір параметричної моделі, яка могла б описати базову структуру даних часових рядів. Ці моделі можуть включати авторегресію (Autoregression, AR), ковзне середнє (Mooving Average, MA), авторегресійне інтегроване ковзне середню (Autoregression Integrated Moving Average, ARIMA) або більш складні моделі, такі як моделі просто-

ру станів або моделі GARCH (Generalized AutoRegressive Conditional Heteroskedasticity).

- На третьому етапі має місце оцінка параметрів моделі за допомогою доступних даних. Це включає такі методи, як оцінка максимальної правдоподібності (Maximum likelihood estimation, MLE) або оцінка найменших квадратів (Least Squares Estimation, LSE).

- На четвертому етапі здійснюється оцінка відповідності моделі даним. Ось тут і вступає в гру AIC. AIC враховує ймовірність даних, заданих моделлю, і штрафує за складність моделі (кількість параметрів). Нижчі значення AIC вказують на те, що моделі краще підходять.

Останні два описані кроки повторюють для різних моделей-кандидатів і порівнюють їхні значення AIC. Модель з найнижчим AIC часто вважається найкращою моделлю, оскільки вона забезпечує баланс між хорошим поясненням даних і не є надто складною. Після вибору найкращої моделі стає можливим використання її для прогнозування майбутніх значень часового ряду та для контролю чи прийняття рішень на основі прогнозованих значень.

У наш час підхід Akaike до аналізу часових рядів із використанням параметричних моделей широко застосовується в різних галузях, зокрема в економіці, фінансах, інженерії та науці про навколишнє середовище, для розуміння та прогнозування даних, що змінюються в часі. Він забезпечує систематичний і керований даними спосіб вибору відповідних моделей і прийняття обґрунтованих рішень на основі даних часових рядів [28, 29].

Багато статистичних методів, таких як t-тест, аналіз дисперсії (ANalysis Of VAriance, ANOVA), критерій χ^2, регресійний аналіз та інші, базуються на припущенні про нормальний розподіл даних. У 1972 Shapiro та Francia представлено модифіковану версію статистики Шапіро-Вілка [30], розро-

блену для оцінки нормальності, особливо придатну для великих розмірів вибірки [31]. Оригінальний тест Шапіро та Уїлка визначав коефіцієнти та критичні значення для розмірів вибірки до 50 вимірювань, але ці коефіцієнти вимагали наближення коваріаційної матриці статистики нормального порядку. У роботі [31], навпаки, запропонований тест спирається виключно на значенні звичайної статистики нормального розподілу, яка є відомою. На додаток, у статті коротко викладаються результати емпіричного вибіркового дослідження, у якому порівнюється чутливість запропонованої тестової статистики до традиційної W-тестової статистики.

У 1885 році Сер Френсіс Гальтон вперше визначив термін "регресія" та розробив теорію біваріативної кореляції, що стала важливим кроком у розвитку статистичних методів [32]. Десять років потому Карл Пірсон розробив індекс, який все ще активно використовується для вимірювання кореляції, відомий як коефіцієнт кореляції Пірсона. Стаття Rodgers та Nicewander [32], написана з нагоди відзначення 100-річчя першої важливої дискусії Гальтона про регресію та кореляцію, відзначила початок цього напрямку досліджень. Автори розпочинають з короткої історії, яка допомагає зрозуміти еволюцію статистичних понять кореляції. Потім презентують 13 різних формул, кожна з яких представляє різні обчислювальні та концептуальні визначення коефіцієнта кореляції r. Кожна формула вказує на різний кут сприйняття показника кореляції у сенсі алгебраїчних, геометричних і тригонометричних підходів. Автори показують, що коефіцієнт кореляції Пірсона (або прості функції r) можна різним чином розглядати як особливий тип середнього (математичного сподівання), особливий тип дисперсії, відношення двох середніх, відношення двох дисперсій, нахил лінії, косинус кута і тангенс еліпса, і може бути досліджений з кількох інших цікавих точок зору, розширюючи наше розуміння цього показника та його застосувань.

Вимірювання неперервного моніторингу за своєю суттю мають атрибути авторегресії та автокореляції, які можна ефективно дослідити за допомогою моделей часових рядів Бокса-Дженкінса. Тим не менш, використання складних методів, таких як авторегресійне інтегроване ковзне середнє (ARIMA) або його підмножини для фільтрації та прогнозування, може бути не ідеальним для автоматизованого аналізу деформації в реальному часі. У дослідженні Yam Khoon особлива увага приділяється застосуванню фільтра Калмана для вирішення проблем, пов'язаних із викидами та зашумленими даними. Систематичні ефекти, пов'язані з обертанням стовпа, змінами масштабу, варіаціями заломлення, а також змінами горизонтальних напрямків, нахильних відстаней і зенітних кутів, моделюються як компоненти вектора стану в рамках прямої фільтрації Калмана та зворотного згладжування. Дослідження успішно демонструє доцільність реалізації автоматизованого фільтра Калмана в режимі реального часу для аналізу деформації.

Щоб підтвердити ефективність фільтра Калмана, у роботі [20] були проведені випробування на, здавалося б, "стійкій" будівлі, розташованій у Наньянському технологічному університеті, Сінгапур. Дані, зібрані під час цих тестів, були оброблені за допомогою фільтра Калмана. За словами автора, будівля демонструвала рухи менше 1 мм, що підкреслює потенціал фільтра Калмана як цінного інструменту для аналізу деформації в режимі реального часу в сценаріях, де порушення навколишнього середовища та викиди створюють проблеми для точності вимірювань.

Аналіз і пошук траєкторій об'єктів у дво- або тривимірних просторових середовищах здійснено також у роботі [33]. Автори вважають, що траєкторії (шаблони або тренди) зазвичай супроводжуються значною кількістю шуму, що робить традиційні показники непридатними для точного аналізу. Щоб вирішити цю проблему, автори представили неметричні функції

подібності, які грунтуються на методі найдовшої спільної підпослідовності (Longest Common Subsequence, LCSS). Ці нові функції виявляють виняткову стійкість до шуму, пропонуючи інтуїтивно зрозумілу міру подібності між траєкторіями, приділяючи більше уваги відповідним сегментам у послідовності.

Запропонований у роботі підхід дозволяє розтягнути послідовності в часі та отримати глобальну трансляцію послідовностей у просторі, забезпечуючи гнучкість у обробці різноманітних даних траєкторії. Крім того, автори розробили ефективні наближені алгоритми, здатні швидко й ефективно обчислювати ці показники подібності. Щоб продемонструвати ефективність запропонованих методів, дослідники провели всебічні порівняння з добре відомими функціями відстані Евкліда та викривлення часу, використовуючи як реальні, так і синтетичні дані траєкторії. Висновки показують перевагу запропонованого підходу, особливо в сценаріях із високим рівнем шуму. Науковці також створили слабшу версію нерівності трикутника та використали її для побудови структури індексування для ефективної адресації запитів найближчих сусідів, підвищуючи практичну корисність отриманого підходу в задачах пошуку. Крім того, у роботі [33] представлено результати експериментальних оцінок, які служать для підтвердження як точності, так і ефективності описаного підходу. Ці експерименти обґрунтовують практичну застосовність і стійкість неметричних функцій подібності на основі LCSS, пропонуючи багатообіцяючі рішення для аналізу та пошуку траєкторій об'єктів у зашумлених реальних даних.

1.3. Неперервні часові ряди

Протягом кількох останніх десятиліть фізики та інженери цікавляться процесами авторегресії неперервними в часі (CAR). Перші наукові статті,

що досліджували властивості та здійснювали статистичний аналіз таких процесів, а також більш загальних процесів неперервної авторегресії ковзного середнього (Continuous-Time Autoregressive Moving Average CARMA), були написані Дубом (1944), Бартлеттом (1946), Філліпсом (1959) і Дурбіном (1961) [34, 35]. Зараз спостерігається відновлення інтересу до процесів, які є неперервними в часі, в основному через успішне використання моделей стохастичних диференціальних рівнянь у фінансах. Один з яскравих прикладів – це виведення формули ціни опціонів Блека-Шоулза та її узагальнення (Халл і Уайт, 1987). Багато прикладів економетричних застосувань моделей з неперервним часом можна знайти в книзі Бергстрома (1990) [36, 37]. Моделі з неперервним часом також успішно використовувалися для моделювання нерегулярно розподілених даних, як показано в роботах Jones (1981, 1985) та Jones і Ackerson (1990) [38, 39]. Водночас стає очевидним, що нелінійні моделі часових рядів краще відображають велику кількість емпірично спостережуваних часових рядів, ніж лінійні моделі. Особливий успіх мали порогові моделі ARMA (Autoregressive Moving Average) Тонга (1983, 1990) [40, 41, 42], які вдало представляли різноманітні набори даних. Моделі ARCH і GARCH, розроблені Енглом (1982) і Боллерслевом (1986) відповідно [43, 44, 45, 46], також були дуже ефективними у моделюванні фінансових даних. Нельсон (1990) презентував версії моделей ARCH і GARCH для неперервного часу [47, 48].

Використання порогових авторегресійних процесів з неперервним часом (CTAR) у моделюванні та прогнозуванні часових рядів показано також у роботі [49]. Подібно до лінійного випадку, модель неперервного часу виявляється корисною при аналізі нерівномірно розподілених даних. У роботі [50] розглянуто аналогічний пороговий процес $ARMA(p, q)$ з неперервним часом, де $0 \leq q \leq p$, що виражається в термінах базового p-вимірного процесу дифузії. Описані рекурсивні формули для ймовірності спостере-

жень $\{y(t_1), \ldots, y(t_n)\}$ у контексті ймовірностей переходу процесу дифузії. У випадку, коли білий шум базового процесу і коефіцієнти ковзного середнього є сталими, характеристична функція розподілу ймовірності переходу основного процесу дифузії може бути виражена за допомогою формули Кемерона-Мартіна-Гірсанова, яка представляє собою явний функціонал стандартного броунівського руху [51, 52, 53, 54], також у статті обговорено апроксимаційні чисельні методи для обчислення ймовірності Гауса [55], які застосовуються до моделювання реальних даних.

Моделі ARMA з неперервним часом, основні властивості, зв'язок моделей ARMA з дискретним та неперервним часом, а також висновки, зроблені на основі спостережень, здійснених у дискретному часі, та нелінійні процеси, що включають аналоги порогових моделей ARMA Тонга у неперервному часі також детально розглянуто у роботі [56] .

Дедалі частіше у науковій літературі зустрічаються процеси, які є неперервними у часі. Моделі таких неперервних часових рядів повинні враховувати зміни в реальному часі та контролювати динаміку. Такі моделі називаються CTARMA (Continuous-time Threshold ARMA). "Continuous-time" означає, що часовий ряд аналізується на неперервному часовому інтервалі, а "threshold" вказує на використання порогових значень для контролю динаміки моделі. Модель CTARMA поєднує в собі ідеї ARMA та порогового аналізу. Порогові значення регулюють, коли та які ARMA-компоненти активуються. Ці порогові значення можуть встановлюватися на підставі певних умов або динамічно мінятися з часом, в залежності від характеристик даних.

Визначення та дослідження конкретного класу процесів, відомих як неперервні порогові процеси ARMA (CTARMA) описано у роботі [57]. Автори запропонували унікальне визначення цих процесів, описуючи їх у термінах слабкого розв'язку певного стохастичного диференціального рівня-

ння. Автори статті основну увагу приділяють дослідженню властивостей стійкості процесів CTARMA. Наведено умови, які визначають різні аспекти стійкості, включаючи швидкоплинність, повторення Харріса та геометричну ергодичність для процесів CTARMA. Автори також підкреслюють важливість певних умов неперервності, яким задовольняють процеси CTARMA, що дозволяє аналізувати їх як незвідні -процеси. Дослідження сприяє розумінню нелінійних моделей у контексті аналізу часових рядів.

Динамічне байєсівське моделювання в контексті аналізу часових рядів та прогнозування добре описано у книзі [58], яка зосереджена насамперед на практичному застосуванні теоретичних результатів. Вона забезпечує комплексну структуру для розуміння та впровадження динамічних лінійних моделей (Dynamic linear model, DLM) і досліджує різні аспекти байєсівського моделювання для аналізу даних часових рядів. Автори заглиблюються в специфіку динамічних лінійних моделей, пояснюють їх структуру, описують їх використання для аналізу часових рядів і прогнозування. Дослідники порівнюють та протиставляють динамічні байєсівські моделі з класичними моделями часових рядів, надаючи розуміння переваг байєсівських підходів. Також представлено різні реальні застосування динамічного байєсівського моделювання. Загалом, автори пропонують всебічне дослідження динамічного байєсівського моделювання для аналізу часових рядів, від теорії до практичного застосування. Книга має на меті надати читачам знання та інструменти, необхідні для моделювання та прогнозування даних часових рядів за допомогою байєсівських методів, з акцентом на інтерактивному аналізі та практичній реалізації за допомогою програмного забезпечення BATS.

Вибірка Гіббса (Gibbs sampling) є одним із методів Марковського ланцюга Монте-Карло (Markov Chain Monte-Carlo, MCMC), який використовується для чисельної апроксимації розподілів, що складно обчислюються

аналітично. Даний стохастичний алгоритм використовується для отримання вибірки зі складного розподілу, якщо відомо умовні розподіли величин, з якими він пов'язаний. У контексті генеративної моделі для набору випадкових змінних вибірка Гіббса може бути ефективно узагальнена в два окремих етапи [59]. Перший етап полягає в одержанні розподілів ймовірностей та отриманні повної щільності для всіх випадкових змінних у моделі. Крім того, автор обчислив апостеріорні умовні розподіли ймовірностей для кожної окремої випадкової змінної в моделі. Другий етап містить генерацію вибірок з апостеріорного сукупного розподілу, грунтуючись на апостеріорних умовних розподілах. Автори використовують такі апостеріорні умовні розподіли ймовірностей для моделювання вибірок із апостеріорного розподілу. Дослідники продемонстрували практичне застосування вказаних етапів у контексті проблеми виявлення точки зміни. Запропонований підхід охоплює фундаментальну суть вибірки Гіббса, дозволяючи авторам ітеративно оновлювати та генерувати вибірку з апостеріорних розподілів кожної змінної, що зрештою призводить до всебічного розуміння сукупного розподілу та полегшує виявлення точки зміни.

Brockwell та Davis є відомими науковцями у галузі дослідження часових рядів. Їхня книга [60] зосереджена на забезпеченні всебічного вступу до аналізу часових рядів та прогнозування. Автори представили фундаментальні концепції даних часових рядів, пояснили їх відмінності від інших типів даних та описали важливості аналізу послідовних спостережень. У книзі розглядаються різні моделі часових рядів, включаючи як статистичні, так і математичні моделі. Ці моделі є фундаментальними для розуміння закономірностей і поведінки в даних часових рядів. Brockwell та Davis охоплюють широкий спектр методів прогнозування, які є важливими для створення прогнозів на основі історичних даних часових рядів. Видання насичене прикладами та ілюстраціями, які демонструють, як аналіз часо-

вих рядів і прогнозування застосовуються в реальному житті. У вигляді додатків в кінці книги наведено основні математичні знання, необхідні для розуміння аналізу часових рядів та прогнозування. Простіше кажучи, ця книга є відправною точкою на шляху до розуміння аналізу часових рядів і прогнозування, охоплюючи як теоретичні концепції, так і практичні застосування. Вона призначена для читачів, які хочуть отримати чітке розуміння методів аналізу часових рядів. Дана книга є цінним ресурсом для студентів і професіоналів у різних галузях.

Для апроксимації складних статистичних даних та розподілів, які обчислити аналітично важко або неможливо, використовують метод Монте-Карло, який використовує випадкові вибірки. Метод Монте-Карло дозволяє апроксимувати розподіли даних, включаючи ймовірнісні розподіли, з якими пов'язані статистичні задачі. Це дає змогу оцінювати різні параметри розподілів, будувати інтервали надійності та робити статистичні прогнози. У деяких статистичних задачах, особливо в складних моделях, немає аналітичних рішень. Метод Монте-Карло допомагає обчислювати числові оцінки та вибірки, навіть якщо немає аналітичних виразів, та моделювати різні статистичні сценарії та ризики. Це корисно для оцінки ймовірностей подій, які важко передбачити або виміряти в реальному житті. Крім того, моделювання випадкових подій та аналіз їх впливу на статистичні результати також здійснюється за допомогою цього методу.

Десята міжнародна конференція з методів Монте-Карло та квазі-Монте-Карло в наукових обчисленнях відбулася в Університеті Нового Південного Уельсу (Австралія) у лютому 2012 року з публікацією рецензованих матеріалів конференції [61]. Такі конференції відбуваються кожні два роки та мають велике значення в галузі досліджень методів Монте-Карло та квазі-Монте-Карло. Слід зауважити, що методи Монте-Карло використані також і в даному дисертаційному дослідженні для генерації вимірювань часових

рядів та значень з рівномірного розподілу.

Матеріали охоплюють збірку статей, взятих із запрошених лекцій і ретельно відібраних доповідей, що охоплюють широкий спектр як теоретичних принципів, так і практичного застосування методів Монте-Карло та квазі-Монте-Карло. Ці статті містять цінну інформацію про найновіші досягнення в тих областях дослідження, що динамічно розвиваються. Книга [61] є цінним довідником як для теоретиків, так і для практиків, які займаються розв'язанням обчислювальних проблем великого розміру. Це особливо актуально для тих, хто займається такими галузями, як фінанси, статистика та комп'ютерна графіка, де методи Монте-Карло та квазі-Монте-Карло знаходять широке застосування для вирішення складних обчислювальних задач.

Нова байєсівська непараметрична структура для кластерного аналізу представлена у роботі [62]. Отримана модель поєднує в собі два важливі компоненти: моделі суміші відбору зразків видів, засновані на розподілах Гауса, і процедуру детермінованої кластеризації, відому як DBSCAN. Фундаментальна ідея, яка лежить в основі запропонованого підходу, полягає в тому, що два спостереження в моделі суміші відбору зразків видів призначаються одному кластеру, якщо функції щільності, пов'язані з їхніми прихованими параметрами, ближчі, ніж задана порогова відстань. Щоб формалізувати цю концепцію, автори встановили відношення еквівалентності між мітками даних, що призводить до нового випадкового розділення даних. Примітно, що цей розподіл є більш "грубим" порівняно з тим, який створено моделлю суміші відбору проб. Однак вибір порогового значення є вирішальним для цієї процедури, тому автори пропонують стратегію визначення відповідного порогового значення.

Крім того, дослідники обговорюють практичну реалізацію та застосування запропонованої моделі. Автори застосовують його до двох наборів

даних: змодельованого двовимірного набору даних, що містить суміш двох щільностей із вигнутим кластером, та профілів експресії генів, зібраних у різні моменти часу, широко відомих як дані клітинного циклу дріжджів. Дослідники здійснили порівняльний аналіз із більш звичайними алгоритмами кластеризації. В обох випадках кластерні оцінки описаної моделі демонструють вищу ефективність. Ключовою перевагою розглянутої моделі є її застосовність до наборів даних, що характеризуються кластерами з важкими хвостами або вигнутими кластерами, де традиційні методи кластеризації можуть бути неефективними. Одержаний підхід пропонує багатообіцяюче рішення для таких складних структур даних, відчиняючи двері для різних практичних застосувань у кластерному аналізі.

Оцінка максимальної правдоподібності для параметра ковзного середнього θ в моделі $MA(1)$, коли $\theta = 1$ або є дуже близько до цього 1, описана у дослідженні [63]. Автори вперше представили виведення граничного розподілу оцінки $\hat{\theta}$ LM, яка визначається як найбільший із локальних максимізаторів ймовірності. Вказана теорія об'єднує випадки, коли справжній параметр знаходиться в межах одиничного кола та випадки, коли він лежить на одиничному колі. Автори також описали асимптотичний розподіл оцінки максимальної правдоподібності $\hat{\theta}$ MLE та показали, що він відрізняється від розподілу $\hat{\theta}$ LM, хоча ця різниця є досить невеликою. Важливим аспектом є той факт, що асимптотичний розподіл для будь-якої оцінки є дуже точним навіть для малих розмірів вибірки та для значень параметра ковзного середнього, які знаходяться далеко від одиничного кола.

Багато складних динамічних явищ можна успішно моделювати за допомогою системи, яка перемикається між набором умовно лінійних динамічних режимів. У даному контексті Fox та ін. розглянули дві такі моделі: перемикальну лінійну динамічну систему (Switching Dynamic Linear Models, SLDS) і процес перемикання VAR [64]. У даному дослідженні байє-

сівський непараметричний підхід використовує ієрархічний процес Діріхле для вивчення невідомої кількості сталих, плавних динамічних режимів. Автори також використовували автоматичне визначення релевантності, щоб виявити розріджений набір динамічних залежностей, що дозволило досліджувати SLDS із змінним розміром стану або процеси VAR із змінним порядком авторегресії. Автори розробили алгоритм вибірки, який поєднує в собі спрощене наближення до процесу Діріхле з ефективною спільною вибіркою послідовностей режиму та стану. Корисність і гнучкість отриманої моделі продемонстровані на прикладах синтетичних даних, послідовностей рухів медоносних бджіл, фондового індексу IBOVESPA та програми відстеження маневруючих цілей.

Ghosh та інші запропонували модель простору станів [65], в якій функціонали рівнянь спостереження та еволюційних рівнянь вважаються невідомими та розглядаються як випадкові функції, що змінюються з часом. Цей підхід зробив запропоновану модель непараметричною та розширив можливості традиційних параметричних моделей простору станів. Важливо відзначити, що автори уникнули обмеження припущення, що функціональні форми залишаються незмінними з часом, що є загальним при параметричних моделях. Традиційний підхід, який полягає в припущенні відомих параметричних функціональних форм, перетворився у сумнівний підхід, особливо в моделях простору станів, де необхідні дані як про спостережувані часові ряди, так і про латентні стани. Автори визначили процеси Гаусса як пріоритети випадкових функцій та використали "підхід таблиці перегляду" для ефективної обробки динамічної структури моделі. Такий підхід дозволив розглянути як однофакторні, так і багатофакторні сценарії, використовуючи метод MCMC для аналізу апостеріорних розподілів.

У випадку складних багатоваріантних сценаріїв Ghosh та ін. показали, що недавно розроблений MCMC на основі трансформації (Transiti-

onal MCMC, TMCMC) надає цікаві та ефективні альтернативи звичайним розподілам пропозицій. Автори продемонстрували ефективність отриманих методів на складному багатовимірному симульованому наборі даних, в якому справжні рівняння спостереження та еволюційні рівняння є нелінійними та розглядаються як невідомі. Отримані результати є обнадійливими та демонструють переваги запропонованого підходу. За допомогою процесу Гаусса, науковці також провели аналіз реального набору даних, який був раніше досліджений іншими авторами з використанням припущення про лінійність. Описаний у роботі [65] аналіз показав, що це припущення про лінійність стає неактуальним з плином часу.

У роботі [66] представлено підхід до вибору відповідної моделі для складних часових рядів та на численних прикладах продемонстровано, як цей запропонований підхід значно розширює клас часових рядів, які можуть бути ефективно змодельовані за допомогою авторегресійних та ковзних середніх моделей.

Байєсівський непараметричний підхід також використано до визначення кількості компонентів у моделі суміші [67]. Автори розглянули ієрархічну модель з відповідним непараметричним пріоритетом для складної структури даних. У роботі [67] процес Діріхле замінено на більш загальний непараметричний пріоритет, який отримано з узагальненого Гамма-процесу. Головною особливістю розглянутої моделі є наявність структури для обробки прихованих змінних типу Гіббса, яка відповідає добре відомим альтернативним моделям розподілу продуктів. У порівнянні зі звичайною сумішшю моделей процесу Діріхле, перевага досліджуваного узагальнення полягає в наявності додаткового параметра σ, який приймає значення з інтервалу $(0, 1)$. Показано, що цей параметр має значний вплив на поведінку кластеризації моделі. Значення σ, близьке до 1, призводить до утворення великої кількості кластерів, більшість з яких мають невеликий розмір.

Механізм посилення, контрольований параметром σ, впливає на розподіл маси, штрафуючи кластери малого розміру та віддаючи перевагу тим кільком групам, що містять велику кількість елементів. Ці характеристики виявляються дуже корисними в контексті моделювання суміші. Оскільки важко визначити апріорну швидкість армування, автори вказали пріоритет для параметра σ залежним від даних.

Останнім часом спостерігається значний інтерес до аналізу даних щодо експресії генів, які були зібрані протягом тривалого періоду часу. У роботі [68] представлено байєсівську модель ієрархічної суміші для обробки таких даних. Замість стандартного підходу до кластеризації спостережень дослідники використовували непараметричну модель, яка ґрунтується на випадковому блуканні та розподілі параметрів, визначених цією моделлю. Отримана модель відзначається гнучкістю, її можна налаштувати для врахування специфічних контекстів, беручи до уваги порядок спостережень у кожній кривій, помилки вимірювань та дозволяючи введення попередніх знань щодо параметрів. Кількість розділів також може бути розглянута як невідома величина, яка може бути виведена із даних. У статті [68] обчислення проводилися за допомогою алгоритму Монте-Карло ланцюга Маркова. Автори дослідили поведінку моделі на синтетичних даних, порівнюючи її з традиційними підходами, а потім застосовали отриману модель для аналізу даних експресії генів, які були зібрані протягом тривалого часу в генах дріжджів поділу.

1.4. Кластеризація даних

Часові ряди та графи є математичними зображеннями реальних процесів у різних напрямках життєдіяльності людини. Саме тому вони активно досліджуються та описуються у науковій літературі. Спираючись на

дані, можна отримати знання, досвід для ефективної діяльності, керування, відповідей на питання, які цікавлять кожного з нас. Тобто той, хто вміє опрацьовувати дані, володіє інформацією. А хто володіє інформацією – керує світом. Кластеризація даних, як один із найпопулярніших напрямків у дослідженні часових рядів, активно застосовується у наш час. Однією із підзадач кластеризації є розгляд латентних класів або кластерів – це концепція, де вважається, що в даному наборі даних існують підгрупи (кластери), але самі кластери не мають визначеної приналежності кожного спостереження до конкретного кластера. Замість цього, припускається, що кожне спостереження може належати до одного з декількох латентних (прихованих) кластерів, і завданням алгоритму аналізу є виявлення цих кластерів та призначення спостережень за ними.

Основна ідея латентного класу включає в себе наступне. Припускається, що дані містять кластери, але вони не відомі перед початком аналізу. Іншими словами, існують приховані структури у даних, які ми намагаємося виявити. Для аналізу даних з латентними класами, дослідники повинні вибрати математичну модель, яка описує розподіл даних в кластерах та ймовірність приналежності кожного спостереження до кожного з латентних кластерів. Популярною моделлю для цього є модель латентного розподілу Діріхле (Latent Dirichlet Allocation, LDA). Кінцевою метою є оцінка параметрів моделі, які найкраще пояснюють розподіл даних та приналежність до кластерів. Це може бути зроблено за допомогою різних алгоритмів, таких як MLE або метод максимальної апостеріорної правдоподібності (Maximum A Posteriori, MAP). Після навчання моделі і оцінки параметрів кожне спостереження може бути призначено до одного з латентних кластерів на основі ймовірностей, розрахованих моделлю.

Альтернативний підхід до проведення експлораторного аналізу латентних класів, який використовує моделі латентних класів на основі факторів

і порівнює його із більш традиційним підходом, що базується на моделях латентних класів, запропоновано у [69]. Аналіз декількох наборів даних свідчить, що LC (Local Context) факторні моделі зазвичай краще відповідають даним і надають результати, які легше інтерпретувати, ніж відповідні LC моделі кластерів. У дослідженні [69] введено новий графічний дисплей для LC факторних моделей, який порівнюється з аналогічними графіками, що використовуються в аналізі відповідностей, а також з баріцентричним координатним дисплеєм для LC моделей кластерів. Автори презентували нові результати щодо ідентифікації LC моделей. У підсумку дослідження описано різноманітні розширення моделей і підхід для усунення граничних рішень в ідентифікованих і неідентифікованих LC моделях.

Ще однією краплею у морі застосування кластеризації даних є прогнозування та моделювання енергетичних профілів будівель, що є критично важливим завданням у пошуках енергоефективності та стійкості [19]. Щоб досягти цього, важливо використовувати інструменти, здатні розкривати значущі закономірності у величезній кількості зібраної інформації. Одним із основних методів, що використовуються в цьому процесі, є кластеризація. Однак кластеризація не є універсальним підходом – це пов'язано з різноманітними складнощами та невизначеністю, що вимагає прийняття кількох важливих рішень протягом усього процесу. Одним із початкових і ключових рішень є вибір метрики подібності, яка визначає, як вимірюється відстань між двома незалежними векторами або точками даних. Це рішення значно впливає на результат процесу кластеризації. Метою статті [19] є дослідження впливу різних методів подібності даних при застосуванні методів кластеризації у випадках, коли кореляція відіграє вирішальну роль. Це особливо актуально при роботі з даними часових рядів, наприклад, при побудові моделей споживання енергії. Досліджуючи вплив показників подібності, автори прагнули покращити проектування та розробку

оптимізованих моделей на основі кластеризації, предикторів і контролерів для залежних від часу процесів.

Впроваджуючи кластерно-векторний баланс як інструмент перевірки, дослідники та практики в галузі енергетичного аналізу будівель можуть отримати глибше розуміння ефективності своїх моделей на основі кластеризації та приймати обґрунтовані рішення щодо вибору та параметризації моделей. Таким чином, вибір показників подібності в аналізі кластеризації є критичним фактором, особливо при роботі з даними часових рядів, як-от створення моделей споживання енергії. Розуміння та використання часових кореляцій у таких даних може призвести до більш точних прогнозів і покращення стратегій управління енергією. Крім того, запропонована техніка перевірки балансу кластерного вектора є цінним інструментом для оцінки ефективності кластеризації та забезпечення надійності моделей на основі кластеризації в контексті залежних від часу процесів.

Кластеризація є потужним інструментом також для систематизації великої кількості невпорядкованих текстових документів у зручний і зрозумілий набір кластерів. Ця методика надає можливість інтуїтивної та інформативної навігації та перегляду текстів. Алгоритми роздільної кластеризації стали більш популярними для обробки об'ємних наборів даних, порівняно з ієрархічною кластеризацією.

У дослідженні [70] використано різноманітні функції відстані та міри подібності, такі як квадрат Евклідової відстані, косинусна подібність та відносна ентропія для кластеризації текстових документів. Автори провели порівняння та аналіз ефективності цих методів роздільної кластеризації на семи наборах даних текстових документів, використовуючи стандартний алгоритм K-середніх. Автори також представили результати, отримані для п'яти різних мір відстаней та подібності, які найчастіше використовуються в задачах текстової кластеризації.

Новий підхід до кластеризації даних часових рядів за допомогою методу, який ґрунтується на основі моделі часового ряду, запропоновано у роботі [71]. Описаний метод до кластеризації даних використовує байєсівський непараметричний підхід, який пропонує гнучкість у виборі функцій, здійснює апостеріорний висновок через MCMC та вибирає найкращий кластер на основі статистичних критеріїв. Метод проілюстровано на основі набору даних про ціни акцій на мексиканській фондовій біржі. Використання Байєсівського непараметричного підходу означає, що він може обробляти змінну кількість кластерів без попередньо визначених параметрів. Ця гнучкість особливо корисна в ситуаціях, коли кількість кластерів не визначена. У дослідженні [71] також представлено особливий тип моделі, що називається моделлю суміші процесу Пуассона-Діріхле, яка використовується для створення кластеризації даних часових рядів. Ця модель може фіксувати різні характеристики, які зазвичай зустрічаються в часових рядах, такі як тренд, сезонні та часові компоненти.

Отримана модель дозволяє користувачам вибирати, які функції часового ряду повинні бути розглянутими для кластеризації. Апостеріорний висновок для моделі отримано за допомогою схеми MCMC. MCMC —- це статистичний метод для оцінки складних імовірнісних моделей. Запропонований метод вибирає найкращий кластер на основі показника неоднорідності та критерію відбору моделі, який називається логарифмом псевдограничної правдоподібності (Logarithm of the Pseudo Marginal Likelihood, LPML). Такий підхід гарантує, що результати кластеризації є значущими та статистично надійними.

У випадку, коли дані задаються неструктурованими типами даних – часовими рядами чи графами, для здійснення кластеризації дослідники часто працюють з матрицями суміжності, випадковими матрицями та стохастичними випадковими матрицями. Досліджуючи власні значення цих ма-

триць, можна отримати багато корисної інформації для визначення оптимальної кількості кластерів. Книга [72] якраз містить спектральну теорію багатовимірних випадкових матриць та застосування вказаної теорії для бездротового зв'язку і моделювання фінансових даних. У першій частині книги автор представив деякі основні теореми спектрального аналізу багатовимірних випадкових матриць, які отримані в умовах кінцевого моменту, таких як граничні спектральні розподіли матриці Вігнера та коваріаційної матриці вибірки, межі екстремальних власних значень і центральні граничні теореми для лінійної спектральної статистики. Друга частина містить деякі основні приклади застосування теорії випадкових матриць до бездротового зв'язку. Третя частина відображає застосування розглянутої теорії до статистичних фінансів. Для ознайомлення з основними поняттями графів -– означеннями, теоремами та прикладами цінним є підручник [73], де для переходу з однієї вершини графу в іншу використано ймовірності переходу.

Ефективність кластеризації часових рядів у наданні корисної інформації в різних областях застосування розглянуто у статті [74]. Автори провели огляд та узагальнення попередніх робіт, що досліджували кластеризацію даних часових рядів у різних сферах застосування. У статті представлено основи кластеризації часових рядів, включаючи загальнопризначені алгоритми кластеризації, які зазвичай використовуються в дослідженнях кластеризації часових рядів, критерії оцінки результатів кластеризації та міри визначення схожості/відмінності між двома часовими рядами, які порівнюються, чи то у формі вхідних даних, вилучених ознак або деяких параметрів моделей. Попередні дослідження розбито на три групи в залежності від того, чи вони працюють безпосередньо з вимірюваннями даних у часовому або частотному домені, опосередковано з вилученими ознаками з вимірювань даних або опосередковано з моделями, побудованими на осно-

ві даних. Розглянуто унікальність та обмеження попередніх досліджень, визначено кілька можливих напрямків для подальших досліджень. Крім того, узагальнено області застосування кластеризації часових рядів, включаючи джерела використаних даних.

У роботі [75] виведено точну форму спектрів власних значень кореляційних матриць, отриманих із набору зміщених у часі кінцевих броунівських випадкових блукань (часових рядів). Ці матриці можна розглядати як реальні, асиметричні випадкові матриці, де зсув у часі накладає певну структуру. Автори продемонстрували, що для великих матриць асоційований спектр власних значень є кругово-симетричним у комплексній площині. Цей факт дозволив точно обчислити щільність власних значень за допомогою зворотного перетворення Абеля щільності симетризованої задачі. Автори також показали справедливість цього підходу чисельно. У роботі проведено порівняння теоретичних висновків з густиною власних значень, отриманих з високочастотних (5 хв) даних S& P 500. Автори ідентифікували різні нетривіальні, невипадкові закономірності та знайшли асиметричні залежності, пов'язані з власними значеннями, які сильно відхиляються від прогнозу Гауса в уявній частині. Для тих самих часових рядів, без внеску ринку, Biely та Thurner виявили кластеризацію акцій у сектори причинно-наслідкових зв'язків.

Ретельний розгляд систем (мереж) масового обслуговування, неперервних і дискретних ланцюгів Маркова та їх моделювання здійснено у книзі [76]. Текст книги пропонує читачам як теорію, так і практичні вказівки, необхідні для проведення оцінки продуктивності та надійності комп'ютерних, комунікаційних і виробничих систем. Починаючи з базової теорії ймовірностей, текст закладає основу для більш складних задач мереж масового обслуговування та ланцюгів Маркова, використовуючи застосування та приклади для ілюстрації ключових моментів. У тексті для формування

практичних навичок аналізу ефективності представлено безліч проблем, які відображають реальні виклики галузі. У зв'язку зі стрімким зростанням складності комп'ютерних і комунікаційних систем потреба в цьому тексті, який майстерно поєднує теорію та практику, є надзвичайно великою.

Проблема стохастичної фільтрації має справу з оцінкою поточного стану процесу сигналу, враховуючи інформацію, надану асоційованим процесом, який зазвичай називають процесом спостереження. У статті [77] описано частинний алгоритм, розроблений для вирішення чисельних задач дискретної фільтрації. Алгоритм передбачає використання системи з n частинок, які еволюціонують (мутують) у кореляції одна з одною (взаємодіють) відповідно до закону сигнального процесу та у фіксований час народжують певну кількість нащадків залежно від процесу спостереження. Автори представили декілька можливих механізмів розгалуження та довели конвергенцію систем частинок (якщо n прямує до 1) до умовного розподілу сигналу, заданого спостереженням. У роботі наведено застосування результату до дискретної фільтрації та наведено кілька прикладів, коли результати можна застосувати.

Дисертаційна робота [78] описує теорію, реалізацію та абстрактне тестування потужного нового кластерного алгоритму для графів, який названо кластерним алгоритмом Маркова або алгоритмом MCL (Markov Cluster Algorithm). Алгоритм використовує (і фактично є не чим іншим, як оболонкою) алгебраїчний процес, визначений для графів Маркова, тобто графів, для яких відповідна матриця є стохастичною. У цьому процесі початковий граф послідовно перетворюється шляхом чергування двох операторів розширення та інфляції. Операції розширення та інфляції матриці використовуються для зміни розмірності матриці шляхом додавання або видалення рядків і стовпців.

Розширення – це отримання потужності матриці відповідно до класичного матричного добутку. Стохастично кажучи, це означає обчислення ймовірностей переходу, пов'язаних із багатокроковим співвідношенням. Інфляція збігається з визначенням потужності матриці відповідно до поелементного добутку Адамара Шура з подальшим масштабуванням по стовпцях, щоб кінцевим результатом знову була стохастична матриця (стовпців). Це незвичайний оператор у світі стохастики, його впровадження повністю мотивоване передбачуваною операцією над графами, де присутня кластерна структура.

Автор стверджує, що багатокрокові зв'язки, які відповідають парам точок, що лежать у природному кластері, матимуть більшу ймовірність переходу, ніж пари точок, точки яких лежать у різних кластерах. Оператор інфляції віддає перевагу багатокроковим зв'язкам із великою асоційованою ймовірністю та віддає перевагу багатокроковим зв'язкам із малою асоційованою ймовірністю. Таким чином, показано, що процес MCL створює та зберігає багатоетапні зв'язки, пов'язані з взаємозалежностями в одному кластері, і що знищує усі багатоетапні зв'язки, пов'язані зі зв'язками між різними кластерами. Процес MCL зазвичай збігається до ідемпотентної матриці, яка є дуже розрідженою та складається з кількох компонентів. Компоненти інтерпретуються як кластеризація початкового графа.

Оскільки оператор інфляції параметризований, кластеризації можна виявити на різних рівнях деталізації. Алгоритм MCL спочатку складається з кроку перетворення від заданого графу до стохастичного початкового графу, використовуючи стандартну концепцію випадкового блукання по графу. По-друге, це вимагає специфікації двох рядків значень, які визначають послідовні параметри розширення та інфляції. Врешті решт, алгоритм обчислює відповідний процес та інтерпретує результуючу межу. Ідея використання випадкових блукань для виявлення кластерної структури не є

новою, але метод реалізації цієї ідеї запропоновано вперше [78]. Показано, що існує зв'язок між комбінаторним та імовірнісним методами кластеризації, і що важливою відмінністю є крок локалізації, який зазвичай вводять імовірнісні методи.

Ряд останніх досліджень [79] зосереджено на статистичних властивостях мережевих систем, таких як соціальні мережі та Всесвітня павутина. Дослідники зосередилися особливо на кількох властивостях, які, здається, є спільними для багатьох мереж: властивість малого світу, степеневий розподіл ступенів і транзитивність мережі. Girvan та Newman розглянули іншу властивість, яку можна знайти в багатьох мережах – властивість структури спільноти (сукупності), в якій мережеві вузли об'єднані разом у тісно пов'язані групи, між якими існують лише слабші зв'язки. Автори запропонували метод виявлення таких спільнот, заснований на ідеї використання індексів центральності для визначення меж спільнот. Запропонований метод перевірено на згенерованих та реальних графах, структура яких відома. У статті [79] показано, що описаний метод виявляє цю відому структуру з високою чутливістю та надійністю. Автори також застосували цей метод до двох мереж, структура спільноти яких недостатньо відома – мережі співпраці та харчової мережі. В обох випадках метод виявив значні та інформативні розбіжності спільноти.

1.4.1. Застосування випадкових матриць до задач кластеризації

Теорія випадкових матриць отримана на основі поєднання математичної фізики, а також теорії ймовірностей і передбачає вивчення властивостей ансамблів матриць, що містять випадково розподілені елементи. Як правило вказується закон розподілу елементів. У цьому випадку досліджу-

ється статистика власних значень і власних векторів випадкових матриць [72]. Далі реальні процеси можна описувати за допомогою графів різної структури. Таким чином, деякі властивості стохастичної матриці вважаються ключовими поняттями для кластеризації графів, оскільки кластеризація графів залежить від функцій розширення та роздування матриці. За допомогою цих двох операцій можна визначити кластерну структуру графа. Теорія випадкових матриць є основним математичним інструментом, який використовується для розгляду кластеризації графів. В даний час стохастичні матриці належним чином вивчені щодо їх властивостей і областей застосування.

Концепція розумних мереж, які є модернізованими електричними мережами, які включають передові технології та системи зв'язку для ефективного управління виробництвом, розподілом і споживанням електроенергії, розглянута у роботі [80]. Інтелектуальні мережі спрямовані на підвищення надійності, стійкості та загальної продуктивності. Розумні мережі зазвичай представляються за допомогою великих даних. Представлення великих даних у вигляді матриць та опрацювання вхідних даних за допомогою матричних властивостей наведено у книзі [80].

Унікальний аспект цієї книги полягає в застосуванні теорії випадкової матриці (Random Matrix Theory, RMT) для аналізу даних у контексті розумних мереж. RMT – це математична основа, яка використовується для вивчення статистичних властивостей великих матриць із випадковими елементами. У цьому контексті автори застосовують RMT до аналізу даних, створених інтелектуальними мережами, щоб виявити закономірності, кореляції та іншу відповідну інформацію. Книга також охоплює методи обробки та оптимізації даних у розумних мережах, що включає методи ефективної обробки великих наборів даних і прийняття керованих даними рішень для покращення продуктивності мережі. Автори докладно описали,

як аналітику великих даних, зосереджену на теорії випадкових матриць, можна використовувати для підвищення ефективності, надійності та сталості інтелектуальних мереж. Автори навели як теоретичні основи, так і практичне застосування цих концепцій у контексті сучасних електричних мереж.

Стаття українських математиків Марченка та Пастура [81] присвячена математичному вивченню випадкових матриць. Ці матриці часто виникають у різних галузях науки, включаючи фізику, статистику та техніку. Однією з центральних тем статті є дослідження розподілу ймовірностей власних значень випадкових матриць. Власні значення є важливими математичними характеристиками матриць, а їх статистична поведінка може бути досить складною, коли елементи матриці є випадковими. Поняття універсальності є ключовою темою статті. Це стосується явища, коли певні статистичні властивості власних значень випадкових матриць стають незалежними від конкретних деталей ансамблю матриці, коли розмір матриці стає великим. Універсальність є чудовою рисою теорії випадкових матриць і має зв'язки з різними галузями фізики та математики.

Марченко та Пастур досліджували розподіл власних значень для випадкових матриць з незалежними та однаково розподіленими (i.i.d.) елементами. Вони вивели відомий закон півкола Вігнера, який описує граничний розподіл власних значень для великих випадкових матриць. Цей закон має важливе значення в квантовій механіці та статистичній фізиці. Стаття зосереджена на теоретичних аспектах теорії випадкових матриць. У наш час результати, отримані у статті, мають широке застосування у фізиці, зокрема у вивченні складних систем, невпорядкованих систем і квантового хаосу. Стаття Марченка та Пастура є основоположною роботою в теорії випадкових матриць, яка досліджує статистичну поведінку власних значень для певних класів випадкових матриць. Вона мала значний вплив

на різні наукові галузі, включаючи фізику, і продовжує залишатися важливою довідкою у вивченні складних і невпорядкованих систем.

Проблема оновлення сингулярної декомпозиції (SVD), яка є популярною технікою декомпозиції матриці та яка може зменшити велику матрицю на різні компоненти для спрощення матричних обчислень розглянута у [82]. SVD описується наступним чином:

$$A = U \cdot \sum \cdot V^T,$$

де A позначає дійсну матрицю $m \times n$, яку слід розкласти, U — це матриця $m \times m$, $\sum$ -— це $m \times n$ діагональну матрицю, а V^T представляє матрицю транспонування розміром $n \times n$. Ліві сингулярні вектори A представлені за допомогою стовпців U-матриці, тоді як праві сингулярні вектори представлені за допомогою стовпців V-матриці. Значення матриці $\sum$, обчислені на основі її діагоналей, є сингулярними значеннями A. SVD можна спостерігати, коли доповнюється "висока тонка" матриця, тобто прямокутна матриця $A \in R^{m \times n}$. Крім того, ми отримуємо ефективну техніку для обчислення та компетентного отримання SVD розширеної матриці $[AB] \in R^{m \times (n+n')}$, припускаючи, що SVD A було ідентифіковано раніше, і що існує відома матриця $B \in R^{m \times n'}$. Це важливий інструмент для двох типів програм. На тлі дослідження первинних компонентів провідні відсутні видатні вектори, отримані шляхом розкладання, утворюють ортонормований базис для найтоншого лінійного підпростору певної розмірності, тоді як для правих сингулярних векторів можна виділити ортонормований базис ядра матриці. SVD може допомогти в кластеризації даних зі змінною структурою. За допомогою SVD можна визначити рухомі кластери.

Gudowska–Nowak та ін. [83] досліджували випадкові матриці розвитку з аналогами, запозиченими з броунівського блукання, використовуючи підхід Ейнштейна–Смолуховського для досягнення дифузії. Автори позна-

чили спосіб, у який дивовижно вільні випадкові змінні (FRV) дозволяють вирішити випадкову некомутуючу матричну згортку. Цей підхід безпосередньо керує матричним аналогом центральної граничної теореми. Послідовність відкритих приростів може бути отримана на основі формалізму, отриманого з незалежних (вільних) елементів матриці. Автори також представили результати побудови рядів для вільних матричних приростів на основі адитивного аналога броунівського блукання. Ця техніка призводить до нескінченного структурованого добутку матриць, які не є комутуючими, при розгляді великих матриць. Таким чином, звичайне спостереження, отримане на основі класичного ймовірнісного припущення про те, що алгоритм розглядає та зводить вихідну проблему до адитивного випадку (логарифмічний нормальний розподіл), не виконується, і стає необхідним отримати інші детальні структури. Використовуючи підхід FRV, автори отримали просту рекомендацію щодо методичної функції, яка може генерувати кожен "момент" із згаданого спектрального розподілу. Далі вони помітили, що отримана аналітична функція подібна до голоморфної функції. Автори також сформулювали гіпотезу подвійності на основі вищезгаданого спостереження.

Були розроблені ітераційні підходи для отримання розв'язків великої кількості рівнянь у системі лінійних рівнянь; ці підходи нещодавно використовувалися для отримання власних значень розріджених матриць. Ці ітераційні методи уникають поматричного множення за допомогою ітераційного векторного множення на матрицю, тобто кожен вектор множиться на матрицю поетапно. Спочатку ми отримуємо добуток вектора b і матриці A і знову множимо цей добуток на матрицю А. Таким чином, A^2b отримуємо як добуток, і цю процедуру повторюємо для отримання наступних продуктів. В даний час описані ітераційні алгоритми вважаються найкращими підходами до множення матриць числовою лінійною алгеброю. Усі

алгоритми, які працюють таким чином, називаються підпросторовими методами Крилова [84]. Крилов вперше описав метод знаходження власних значень λ_i, а також векторів великої матриці A.

Підпростір r-порядку Крилова, змодельований за допомогою A, матриці з розмірністю $n \times n$ і вектора b з розмірністю n, є лінійним підпростором, розтягнутим на b під першим r степені A (починаючи з $A^0 = I$); тобто,

$$K_r(A, b) = span\left\{b,\ Ab,\ A^2b,\ \ldots,\ A^{r-1}b\right\}.$$

У роботі [85] запропоновано та математично обґрунтовано деякі методи аналізу обмеження стійкості, а також поведінки експоненціально великих матричних норм. Запропонований метод використовує методи підпростору Крилова [84] для розбиття матриці на дві частини. Перша частина відповідає власним значенням у правій половині комплексної площини, які отримані за допомогою методики Крилова [84] і для яких обмеження стійкості та експоненціальні норми можуть бути отримані за допомогою типових підходів. Друга частина, що містить набір значень і спроектована методом Ерміта Ланцоша, повинна належати до лівої половини комплексної площини. Цей метод має працювати належним чином, якщо блок, що відповідає крайнім правим власним значенням, малий. Як приклад, цей підхід можна застосувати до еліптичних функцій і секторних функцій. Тут ми розглядаємо оператори з полем значення сектора. Слід підкреслити необхідну умову, тобто щоб деякий сектор на правій половині комплексної площини був малим.

Dax [86] зробила спробу вдосконалити метод Крилова [84] для отримання наближень низького рангу у випадку великих матриць. Запропонований підхід був названий новим типом перезапущеного методу Крилова. Основна відмінність від методики Крилова полягає в тому, що Дакс не використовував алгоритм Ланцоша [87] або бідіагоналізацію Ланцоша. На

думку автора, такий підхід спрощує фундаментальну ітерацію та дозволяє виявити деякі нововведення. Одним із покращень є зменшення обчислювальної складності. Удосконалено початковий вектор, який використовується під час кожної ітерації побудови матриці Крилова. Крім того, для отримання матриці Крилова використовується рекурентне співвідношення, що містить три доданки. Ці модифікації призводять до швидкої конвергенції.

Власні значення та власні вектори великих матриць є ключовими показниками для прийняття рішень на основі теорії графів. Графіки можна використовувати для опису багатьох подій і об'єктів повсякденного життя. Теорія графів знаходить застосування в транспорті та комп'ютерних мережах, будівельному проектуванні, молекулярному моделюванні та геоінформаційних системах. Теорія випадкових матриць є основним математичним і статистичним інструментом для точного прийняття рішень при розгляді графів з різною структурою. Наукові дослідження в цій галузі зросли зі збільшенням кількості інформації на запам'ятовуючих пристроях. Більшість раніше проведених досліджень присвячено спектральному аналізу графів, тобто розбиванню матриць інцидентів і суміжності (матриць, що представляють графи) на менші частини та їх аналізу. Це можна пояснити великою розмірністю даних, яка вимагає великої обчислювальної потужності. Обата [88] вивчав графи спектрального аналізу. Автор використав квантово-ймовірнісний підхід [89] і використав квантові компоненти матриці суміжності для отримання матричного спектрального розподілу. Обата здогадався, що певна структура графа може бути пов'язана з певним поняттям незалежності. На основі центральної граничної теореми можна отримати асимптотичний спектральний розподіл для збільшення сімейства графів.

У деяких реальних завданнях графи можуть постійно змінюватися або

зростати. У таких випадках розглядаються зростаючі графи. Ці зростаючі графи представляють збільшення кількості вершин і/або ребер. Квантова ймовірність [89] розглядається у раніше проведеному дослідженні [90] разом із спектральним аналізом неперервно зростаючих графів. Дана робота присвячена вивченню матриць суміжності, зокрема спектральних і граничних розподілів цих матриць. Матриці суміжності можна використовувати для опису постійно зростаючих графів. Значну увагу приділено розкриттю некомутативності матриць суміжності, тобто розширенню теорії ймовірностей. Цей підхід створює нову основу для вивчення матриць суміжності. Допоміжні інструменти, які можна використовувати для дослідження спектральних розподілів матриць суміжності, включають поняття незалежності щодо квантової ймовірності, а також центральну граничну теорему.

Мережевий простір на основі даних базової структури мережі досліджено у роботі [91]. У цьому дослідженні матриці суміжності графів не розглядалися. Теоретичну основу цього дослідження становить теорія збурень [92], яка використовується для отримання спектральних проекцій матриць.Автори в основному розглянули вплив зв'язків між спрямованими ребрами в конкретному кластері і кластерах на проекцію вузлів. Далі науковці розробили алгоритм поділу графів. Спектральну кластеризацію використовували для досягнення розділення графів. Автори порівняли розроблений підхід з іншими сучасними методами кластеризації з використанням реальних наборів даних, продемонструвавши ефективність описаного алгоритму.

Висновки до розділу I

Розділ I дисертаційної роботи надає огляд актуальних наукових досліджень у галузі аналізу часових рядів, визначення відстаней між вимірюва-

ннями часових рядів та кластеризації часових рядів. Ця тематика важлива в багатьох сферах, включаючи фінанси, медицину, транспорт, метеорологію та інші. Огляд літератури розкриває різноманітні підходи та методи, які використовуються для розв'язання проблем, пов'язаних з аналізом часових рядів.

У цьому розділі було визначено основні проблеми, що стосуються аналізу часових рядів. Один з головних викликів полягає в обробці великих обсягів даних і виявленні закономірностей у цих даних. Це важливо для прийняття бізнесових рішень та прогнозування майбутніх подій. Однак, розвиток обчислювальної техніки та методів машинного навчання привів до нових можливостей у аналізі часових рядів. Зокрема, глибинне навчання, яке включає в себе нейронні мережі, дозволяє автоматизувати процес аналізу та виявлення закономірностей у великих обсягах даних. Це допомагає покращити точність прогнозування та робити більш складні моделі аналізу.

Щодо визначення відстаней між вимірюваннями часових рядів, у літературі вказано на важливість цього аспекту для подальшого аналізу. Відомі різні метрики відстаней, такі як Евклідова відстань, відстань Махаланобіса, DTW, EDR, ERP, кореляція та інші. Вибір певної метрики залежить від конкретного завдання та характеристик даних. Важливо враховувати, що невірно вибрана метрика може призвести до неправильних висновків. Кластеризація часових рядів є ще однією важливою частиною аналізу часових даних. Цей процес допомагає групувати схожі часові ряди, що може бути корисним для подальшого аналізу та класифікації.

В цілому, огляд літератури демонструє, що аналіз часових рядів є активною галуззю досліджень, яка постійно розвивається. Різноманітні методи та підходи спрямовані на вирішення різних завдань в цій області. Розуміння цих методів та їхнє використання може допомогти в покращенні

аналізу часових рядів, що має велике значення для багатьох галузей науки та промисловості.

РОЗДІЛ II. ОСНОВНІ МОДЕЛІ ЧАСОВИХ РЯДІВ

Вибір відстані між вимірюваннями в часових рядах є важливим аспектом аналізу часових рядів. Відстань між рядами визначає, наскільки схожі або відмінні ці ряди. Однак визначення правильної метрики відстані може бути складним завданням через різні особливості часових рядів.

Однією з популярних метрик відстані для часових рядів є відстань Евкліда. Вона обчислює відстань між відповідними точками в двох часових рядах та підсумовує ці відстані. Однак цей підхід може бути чутливим до зсувів та масштабування в часових рядах і не завжди є найкращим вибором для нестаціонарних даних.

Кореляційна відстань – це інша популярна метрика для часових рядів. Вона вимірює схожість між рядами на основі їх кореляційної структури. Кореляційна відстань може бути корисною для виявлення подібних патернів у часових рядах, але вона також має обмеження, такі як чутливість до зсувів та змін масштабу.

Ще однією метрикою відстані для часових рядів є динамічний часовий ряд (Dynamic Time Warping – DTW). Він дозволяє розраховувати відстань між рядами, враховуючи зсуви та зміни масштабу. DTW може бути дуже корисним для аналізу часових рядів зі складними зміщеннями та змінами масштабу, але він також є обчислювально витратним методом.

Загалом, вибір оптимальної кількості кластерів та визначення відстані між вимірюваннями часових рядів залишаються відкритими проблемами у галузі аналізу даних та машинного навчання. Розробка нових методів та

підходів для вирішення цих завдань продовжується, і це завдання привертає увагу дослідників і практиків з усього світу. Для досягнення кращих результатів у визначенні оптимальної кількості кластерів та вибору відстані між вимірюваннями часових рядів необхідно подальше дослідження і співпраця спеціалістів з різних галузей науки і технологій.

2.1. Основні означення та властивості

Ключову роль в аналізі часових рядів відіграють процеси, властивості яких або деякі з них не залежать від часу. Для прогнозування значень часових рядів необхідним є припущення, що щось не змінюється з часом. При екстраполяції детермінованих функцій прийнято вважати, що або сама функція, або одна з її похідних є сталою. Припущення про сталу першу похідну призводить до лінійної екстраполяції як засобу прогнозування. При аналізі часових рядів основною метою є прогнозування ряду, який зазвичай не є детермінованим, але містить випадковий компонент.

Означення 1 *Нехай X_t – часовий ряд з $E(X_t^2) < \infty$. Математичним сподіванням часового ряду X_t є*

$$\mu_X(t) = E(X_t).$$

Означення 2 *Коваріаційною функцією часового ряду X_t називається*

$$\gamma_X(r, s) = Cov(X_r, X_s) = E[(X_r - \mu_X(r))(X_s - \mu_X(s))]$$

для всіх цілих значень r та s.

Означення 3 *Часовий ряд X_t є (слабко або нестрого) стаціонарним, якщо*

1. *Математичне сподівання $\mu_X(t) = E(X_t)$ не залежить від t.*

2. *Коваріаційна функція $\gamma_X(t+h, t)$ не залежить від t для кожного h.*

Зауваження 1 *Сильна або строга стаціонарність часового ряду* $\{X_t, t = 0, \pm 1, ...\}$ *визначається умовою, що* $(X_1, ..., X_n)$ *і* $(X_{1+h}, ..., X_{n+h})$ *мають однаковий сумісний розподіл для всіх цілих чисел* h *та* $n > 0$. *Легко перевірити, що якщо* X_t *є строго стаціонарним, а* $EX_t^2 < \infty$ *для всіх* t, *то* X_t *є також нестрого стаціонарний.*

Далі у роботі під використанням терміну "стаціонарний" мається на увазі нестрога стаціонарність часового ряду, якщо конкретно не вказано інше.

Зауваження 2 *Зважаючи на умову 2) означення 3, щоразу, коли використовується коваріаційна функція стаціонарного часового ряду* X_t, *мається на увазі функція однієї змінної* γ_X, *яка визначається наступним співвідношенням*

$$\gamma_X(h) := \gamma_X(h, 0) = \gamma_X(t + h, t).$$

Означення 4 *Нехай* X_t *– стаціонарний часовий ряд. Тоді автоковаріаційною функцією* X_t *з лагом (зсувом)* h *називається*

$$\gamma_X(h) = Cov(X_{t+h}, X_t).$$

Автокореляційною функцією X_t *з лагом* h *називається*

$$\rho_X(h) \equiv \frac{\gamma_X(h)}{\gamma_X(0)} = Cor(X_{t+h}, X_t).$$

Далі в роботі зсув по часовій шкалі на h одиниць називатимемо лагом, а оператор зсуву по часу – лаговим оператором.

2.2. Процеси $ARMA$

Стаціонарний $ARMA$ процес умовно складається з двох частин – авторегресії (з англ. "autoregression" (AR)) та ковзного середнього (з англ.

"mooving average" (MA)). Складова AR позначає кількість регресорів у часовому ряді. Тобто $AR(p)$ означає, що вимірювання часового ряду залежить від p попередніх вимірювань. Складова $MA(q)$ позначає суму q попередніх значень білого шуму ε_t з відповідними ваговими коефіцієнтами.

Означення 5 *$ARMA$ процесом називається процес, який має наступний вигляд:*

$$X_t = \alpha_1 X_{t-1} + ... + \alpha_p X_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + ... + \beta_q \varepsilon_{t-q}.$$

Із означення 5 зрозуміло, що значення часового ряду X_t у момент часу t залежить від суми процесів $AR(p)$ та $MA(q)$.

Розглянемо детальніше для початку процес $MA(q)$. Дле нетермінованого стаціонарного у широкому розумінні випадкового процесу справедливою є наступна теорема.

Теорема 1 *(Вольда). Недетермінований стаціонарний у широкому розумінні випадковий процес можна представити наступним чином:*

$$X_t - \mu = \sum_{\tau=0}^{\infty} \psi_\tau \cdot \varepsilon_{t-\tau}, \tag{1}$$

де μ_t – математичне сподівання цього процесу, а ε_j – білий шум зі скінченними математичним сподіванням та дисперсією. Тобто, будь-який слабко стаціонарний процес можна представити у вигляді лінійної комбінації білих шумів з різними ваговими коефіцієнтами.

Так як вираз (1) лінійний – його часто називають лінійним фільтром. Для того, щоб вираз (1) мав сенс, необхідно, щоб виконувалася умова збіжності за ймовірністю:

$$\sum_{i=0}^{\infty} |\psi_i| < \infty,$$

причому $\psi_0 = 1$. Зрозуміло, що нескінченна сума доданків породжує технічні проблеми. Виявляється, у багатьох випадках достатньо розглядати не загальне представлення Вольда (1), а його частинні випадки, коли число доданків – скінченне.

Означення 6 *Стохастичний процес називається процесом ковзного середнього (Moving Average, MA) порядку q, якщо у розкладі Вольда (1) присутні тільки q доданків. Тобто*

$$MA(q) : X_t = \sum_{\tau=0}^{q} \psi_t \varepsilon_{t-\tau},$$

(для спрощення запису використовується позначення $x_t = X_t - \mu_X$, тобто x_t – відхилення процесу X_t від його математичного сподівання μ_X).

Еквівалентно даний ряд можна представити у наступному вигляді:

$$x_t = \varepsilon_t + \psi_1 \cdot \varepsilon_{t-1} + ... + \psi_q \cdot \varepsilon_{t-q}.$$

Назва "ковзне середнє" пояснюється тим, що поточне значення випадкового процесу визначається зваженим середнім q попередніх значень білого шуму. Процедуру ковзного середнього часто використовують для того, щоб згладити дані, які сильно коливаються. Розглянемо властивості даного процесу.

Процес MA – стаціонарний, так як він є частинним випадком декомпозиції Вольда (1) з математичним сподіванням $E(x_t) = 0$ та дисперсією:

$$V(x_t) = \sigma^2 \sum_{i=1}^{q} \psi_i^2.$$

Отже, математичне сподівання та дисперсія не залежать від часу. Знайдемо коваріацію процесу MA:

$$Cov(x_t, x_{t+\tau}) = \sigma^2 \sum_{i=0}^{q-k} \psi_i \psi_{i+r}, \tau = 0, 1, 2, ..., q.$$

При $\tau > q$ $Cov(x_t, x_{t+\tau}) = 0$. Для позначення коефіцієнтів скінченного ряду $MA(q)$ використовуватимемо β:

$$X_t = \sum_{\tau=0}^{q} \beta_\tau \cdot \varepsilon_{t-\tau}.$$

Це означає, що ψ_i до $i = q$ включно дорівнюють β_i, решта ψ_i – рівні нулю. Тоді вираз для коваріаційної функції MA процесу має вигляд:

$$Cov(X_t, X_{t+\tau}) = \gamma(\tau) = \sigma_\varepsilon^2 \sum_{i=0}^{q-\tau} \beta_i \cdot \beta_{i+\tau},$$

де $\tau \leq q$, а $\sigma_\varepsilon^2 = Var(\varepsilon_t)$.

Іншими словами, між значеннями часового ряду, які знаходяться достатньо далеко один від одного, кореляційний зв'язок відсутній.

Для опису властивостей часових рядів зручно використовувати лаговий оператор L. Визначимо $LX_t = X_{t-1}$, тобто дія оператора зсуву на часовий ряд повертає значення часового ряду у попередній момент часу. Застосувавши лаговий оператор p разів, отримуємо:

$$L^p X_t = L(L(L...)X_t = X_{t-p}.$$

Інколи зручно використовувати нульовий степінь оператора зсуву:

$$L^0 X_t = X_t.$$

За допомогою оператора зсуву (лагового оператора) зручно записати процес ковзного середнього MA наступним чином:

$$X_t = (1L^0 + \beta_1 L + \beta_2 L^2 + ... + \beta_q L^q) =$$

$$= (1 + \beta_1 L + \beta_2 L^2 + ... + \beta_q L^q) =$$

$$= \beta_q(L)\varepsilon_t.$$

Із умови нормування випливає, що коефіцієнт $L^0\varepsilon_t$ або ε_t завжди рівний 1. Перейдемо до розгляду процесу авторегресії AR.

Означення 7 *Якщо значення випадкового процесу визначається лінійною комбінацією скінченного числа його попередніх значень з додаванням білого шуму, то такий процес називається процесом авторегресії (autoregression) порядку p, а його загальне рівняння має вигляд:*

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + ... + \alpha_p X_{t-p} + \varepsilon_t,$$

де ε_t – білий шум.

Згідно даного формулювання процес, обернений до $MA(\infty)$, може бути позначений як $AR(\infty)$. Тим не менше немає жодних гарантій, що при будь-яких коефіцієнтах $\alpha_1, \alpha_2, ..., \alpha_p$ даний процес буде стаціонарним. Для того, щоб $AR(\infty)$ був стаціонарним, необхідно, щоб він був представлений у вигляді розкладу Вольда, тобто, щоб його можна було перевести у $MA(q)$. Перепишемо рівняння процесу $AR(p)$ наступним чином:

$$X_t - \alpha_1 X_{t-1} - \alpha_2 X_{t-2} - ... - \alpha_p X_{t-p} = \varepsilon_t,$$

$$\alpha(L)X = X_t - \alpha_1 X_{t-1} - \alpha_2 X_{t-2} - ... - \alpha_p X_{t-p}.$$

У деякому сенсі отримано "дзеркальне відображення" процесу $MA(q)$. Тепер операторний поліном діє на X, а не на ε і результат дорівнює ε_t. Тобто зліва отримано різницеве рівняння відносно X рівне випадковій правій частині. Загальний розв'язок однорідного рівняння має наступний вигляд:

$$X_t = \sum C_i(t)(\pi_i)^t,$$

де π_i – різні по величині корені характеристичного рівняння (можуть бути також комплексні), яке має вигляд:

$$\lambda^p - \alpha_1 \lambda^{p-1} - ... - \alpha_p = 0,$$

а $C_i(t)$ – поліноми, степінь яких на одницю менші кратності відповідного кореня. Розв'язок характеристичного рівняння буде стійким при умові

$|\pi_i| < 1$. При виконанні даної умови існує оператор, обернений до оператора $\alpha(L)$, тобто справедливим є наступний вираз:

$$X_t = \frac{1}{\alpha(L)}\varepsilon_t.$$

Звідси випливає, що процес X_t приймає вигляд, який відповідає теоремі Вольда – $\sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}$, з чого випливає, що даний ряд є стаціонарним. При $|\pi_i| \geq 1$ ряд не є стаціонарним.

У результаті умова про те, що всі корені характеристичного рівняння для процесу $AR(p)$ лежать всередині одиничного кола, є необхідною і достатньою для того, щоб ряд був стаціонарним.

Тут спостерігається деяка асиметрія. Процеси $AR(p)$ і $MA(q)$ пов'язані. Процес $MA(q)$ завжди стаціонарний. Процес $AR(p)$ – або стаціонарний і зводиться до процесу $MA(\infty)$, або взагалі не стаціонарний. Розглянемо властивості процесу $AR(p)$. Якщо процес $AR(p)$ – стаціонарний, то $E(X_t) = 0$. Автоковаріаційну функцію даного процесу можна вивести, використовуючи представлення процесу у вигляді $MA(\infty)$ за допомогою операторних поліномів. Можна отримати потрібний результат і іншим способом. Помножимо обидві частини виразу

$$X_t = \alpha_1 X_{t-1} + ... + \alpha_p X_{t-p} + \varepsilon_t$$

на $X_{t-\tau}$ і візьмемо математичне сподівання. Оскільки математичне сподівання процесу рівне 0, отримаємо рівняння для значень автоковаріаційної функції. Якщо $\tau > p$, то отримуємо наступну рівність:

$$\gamma(\tau) = \alpha_1\gamma(\tau - 1) + \alpha_2\gamma(\tau - 2) + ... + \alpha_p\gamma(\tau - p) + E\{X_{t-\tau} \cdot \varepsilon_t\}.$$

Зліва у рівнянні використано парність автоковаріаційної функції: $\gamma(\tau) = \gamma(-\tau)$. Так як процес стаціонарний, то $X_{t-\tau}$ представимо у вигляді нескінченної суми $\varepsilon_{t-\tau}$ та попередніх значень ε. Всі ε, які входять до рівності

$X_{t-\tau}$, не співпадають по часу з ε_t. Тому $E\{X_{t-\tau} \cdot \varepsilon_t\} = 0$, бо значення білого шуму не корелюють між собою. Іншими словами, отримано наступне співвідношення:

$$\gamma(\tau) = \alpha_1\gamma(\tau - 1) + ... + \alpha_p\gamma(\tau - p).$$

Якщо розглядати його як різницеве рівняння, то воно співпадає з однорідним рівнянням для процесу X_t. Тобто значення автоковаріаційної функції для $\tau > p$ задовільняє те ж різницеве рівняння. Оскільки різницеве рівняння стійке, то значення автоковаріаційної функції спадають за модулем, по крайній мірі зі значення з індексом $(p + 1)$.

Для $\tau < p$ ці рівняння потрібно модифікувати. Замість рекурентного співвідношення для перших p значень автоковаріаційної функції отримуємо систему p лінійних рівнянь з p невідомими.

Отже, якщо процес AR стаціонарний, то він окрім скінченного представлення $AR(p)$ має нескінченне представлення $MA(\infty)$. Якщо умова оберненості виконується, то скінченний процес $MA(q)$ має нескінченне представлення $AR(\infty)$. Такий дуалізм представлень часового ряду можна розглядати як узагальнення перетворення Койка. Важлива відмінність процесу AR від процесу MA полягає в тому, що для процесу MA автокореляційна функція після $i > q$ затухає або обривається (стає рівною нулю).

Важливо відзначити відмінність між стаціонарним та нестаціонарним процесом.

Твердження 1 *Якщо процес стаціонарний (є він $AR(p)$ чи $MA(q)$ процесом), то, починаючи з деякого моменту, автокореляційна функція починає затухати або зовсім зникає. Якщо ж процес не стаціонарний, то це твердження не є вірним.*

У моделях MA q є точним параметром, після якого автокореляційна функція затухає. У випадку $AR(p)$ моделей такого не відбувається. Тому

важливо отримати відповідну характеристику, яка вказуватиме на точний порядок p у рівнянні авторегресії. Такою характеристикою є частинна автокореляційна функція (Partial Autocorrelation function, PACF). У визначення автокореляційної функції

$$\rho(\tau) = \frac{1}{Var(X_t)} E\{(X_t - \mu)(X_{t-\tau} - \mu)\}$$

входить коваріація між значеннями процесу, які відстають на τ кроків один від одного. Тим не менше на поведінку процесу $AR(p)$ статистично впливає не тільки його значення в момент на τ одиниць назад, а й усі проміжні значення процесу між моментами t і $t - \tau$. Встановимо далі лінійну статистичну залежність між значеннями процесу у ці моменти часу за умови виключення впливу усіх проміжних значень. Тобто яким є "чистий" взаємозв'язок між цими значеннями.

Висновок 1 *Частинна автокореляційна функція авторегресійного процесу $AR(p)$ рівна 0 для $k > p$ і, взагалі кажучи, не рівна 0 при $k \leq p$.*

Отже, частинна автокореляційна функція для процесу AR відіграє таку ж важливу роль, як автокореляційна функція для процесу MA – вона перетворюється в нуль як тільки $k > p$.

Перейдемо до комбінації двох процесів AR та MA – процес авторегресії та ковзного середнього ($ARMA$).

Опишемо умови стаціонарності процесу $ARMA$ із означення 5. Як було встановлено вище, якщо всі корені полінома $\alpha_p(L)$ по модулю менші одиниці, то існує обернений оператор, тобто:

$$X_t = [\alpha_p(L)]^{-1} \beta_q(L) \varepsilon_t.$$

Обернений оператор можна розкласти у суму елементарних дробів, кожну з яких представити як нескінченно спадну геометричну прогресію, тобто

нескінченний операторний поліном. При множенні на скінченний поліном знову отримується нескінченний поліном. Вираз має сенс, коли всі характеристичні корені полінома $\alpha_p(L)$ по модулю менші одиниці. Тоді отриманий розклад є розкладом Вольда (1) і процес є стаціонарним. Все вищесказане приводить до того, що стаціонарність процесу $ARMA$ визначається тільки його AR частиною. Тому умови стаціонарності процесу $ARMA$ ті ж самі, що і у процесу AR. Процес $ARMA$ стаціонарний, якщо корені характеристичного рівняння AR частини по модулю менші одиниці. Аналогічно, умови оборотності процесу, тобто можливість виразити ε_t через X_t, повністю визначаються умовами оберненості MA частини:

$$\varepsilon_t = [\beta_q(L)]^{-1}\alpha_p(L)X_t.$$

Якщо MA частина процесу $ARMA$ оборотна, то і весь процес оборотний.

Отже, якщо процес $ARMA$ стаціонарний, то він обов'язково має $MA(\infty)$ представлення, як і будь-який стаціонарний процес згідно з теоремою Вольда. У той же час процес $ARMA$ має ще і скінченне представлення $ARMA(p, q)$. Процес $ARMA$ може мати ще і нескінченне $AR(\infty)$ представлення. Тим самим, якщо деякий процес можна привести до вигляду $ARMA$, то він визначатиметься $p + q$ параметрами.

Очевидно, що математичне сподівання стаціонарного процесу $ARMA$ рівне нулю: $EX_t = 0$.

Для того, щоб з'ясувати як веде себе автокореляційна і частинна автокореляційна функції процесу $ARMA$, розглянемо найпростіший випадок – процес $ARMA(1, 1)$:

$$X_t = \alpha X_{t-1} + \varepsilon_t + \beta\varepsilon_{t-1}.$$

Використовуючи лаговий оператор, процес $ARMA$ має вигляд:

$$(1 - \alpha L)X_t = (1 + \beta L)\varepsilon_t.$$

Умова стаціонарності тоді має вигляд: $|\alpha| < 1$, умова оборотності – $|\beta| < 1$. Для обчислення дисперсії даного процесу зручно використовувати його $MA(\infty)$ представлення, яке відповідає теоремі Вольда – так зване представлення лінійного фільтра.

$$X_t = \frac{1+\beta L}{1-\alpha L}\varepsilon_t = [1 + (\alpha+\beta)L + \alpha(\alpha+\beta)L^2 + ...]\varepsilon_t.$$

Тоді

$$Var(X_t) = (1 + (\alpha+\beta)^2 + \alpha^2(\alpha+\beta)^2 + ...)\sigma_\varepsilon^2 =$$
$$= \left[1 + \frac{(\alpha+\beta)^2}{1-\alpha^2}\right]\sigma_\varepsilon^2 = \frac{1+\beta^2+2\alpha\beta}{1-\alpha^2}\sigma_\varepsilon^2.$$

Автокореляційна функція процесу авторегресії ковзного середнього має вигляд:

$$\rho_1 = \frac{(\alpha+\beta)(1+\alpha\beta)}{1+\beta^2+2\alpha\beta}.$$

Підсумовуючи вищесказане, отримуємо, що якщо процес відноситься до типу $ARMA(p,q)$, то, починаючи з деякого номера (причому цей номер є важливим, так як він вказує на величину p і q), і автокореляційна, і частинна автокореляційна функції ведуть себе як сума затухаючих експонент, якщо ряд стаціонарний.

Процеси $ARMA$, які розглядалися до цих пір, є строго недетермінованими з нуьовим математичним сподіванням. Якщо у рівняння процесу $ARMA$ $\alpha_p(L)x_t = \beta_q(L)\varepsilon_t$ підставити вираз $x_t = X_t - \mu$, то отримаємо: $\alpha_p(L)(X_t - \mu) = \beta_q(L)\varepsilon_t$. Відкриваючи дужки і використовуючи очевидну рівність $\alpha_p(L)\mu = \mu(1 - \alpha_1 - \alpha_2 - ... - \alpha_p) = \alpha_p(1)\mu$, отримуємо $\alpha_p(L)Y_t = \alpha_p(1)\mu + \beta_q(L)\varepsilon_t$, де $\theta = \alpha_p(1)\mu$ – деяка константа. За допомогою введення вільного члена у рівняння береться до уваги ненульове математичне сподівання. Тоді математичне сподівання процесу буде рівне $\mu = \frac{\theta}{\alpha_p(1)}$. Тому без втрати загальності можа розглядати процес з ненульовим, але сталим математичним сподіванням.

2.3. Процеси $ARIMA$

Розглянемо процес випадкового блукання (Random walk). Інколи цей процес називають процесом броунівського руху. Процес випадкового блукання задається наступним чином:

$$X_t = X_{t-1} + \varepsilon_t, \tag{2}$$

де ε_t – білий шум. Даний процес можна розглядати як $AR(1)$ процес. Обчислимо математичне сподівання, дисперсію та автокореляційну функцію даного процесу при умові, що X_0 – початкова точка. Рівняння (2) є звичайним різницевим рівнянням. Його загальний розв'язок має вигляд:

$$X_t = X_{t-2} + \varepsilon_t + \varepsilon_{t-1} =$$

$$= X_{t-3} + \varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2} + \cdots =$$

$$= X_0 + \sum_{\tau=0}^{t} \varepsilon_{t-\tau}.$$

Математичне сподівання даного процесу задовольняє умову стаціонарності:

$$EX_t = E(X_0) + E\left(\sum_{\tau=0}^{t} \varepsilon_{t-\tau}\right) = X_0 + \sum E(\varepsilon_{t-\tau}) = X_0 + 0 = const.$$

Дисперсія процесу випадкового блукання розсте пропорційно часу t:

$$V(X_t) = t \cdot \sigma_\varepsilon^2.$$

Отже, процес випадкового блукання не є стаціонарним.

Якщо взяти першу різницю даного процесу

$$\Delta X_t = X_t - X_{t-1} = (1 - L)X_t,$$

то рівняння зведеться до наступного вигляду:

$$\Delta X_t = y_t = \varepsilon_t.$$

Тобто, після взяття першої різниці часовий ряд стане стаціонарним. Варто зауважити, що даний підхід приводить до стаціонарності не тільки процес випадкового блукання.

Розглянемо ряд вигляду $X_t = \alpha + \beta t + \gamma t^2 + \varepsilon_t$. Його повністю детермінована частина – тренд є параболічною функцією часу. Очевидно, що даний ряд нестаціонарний. Математичне сподівання даного процесу залежить від часу. З'ясуємо, чи приводиться такий ряд за допомогою взяття послідовних різниць до стаціонарного. Візьмемо першу різницю:

$$\Delta X_t = \alpha + \beta t + \gamma t^2 + \varepsilon_t - \alpha - \beta(t-1) - \gamma(t-1)^2 - \varepsilon_{t-1} =$$

$$= \beta + 2\gamma t - \gamma + (\varepsilon_t - \varepsilon_{t-1}).$$

Степінь полінома, який описує тренд, понизилась на одиницю. Якщо провести взяття другої різниці, то залишиться

$$\Delta^2 X_t = \Delta X_t - \Delta X_{t-1} = 2\gamma + (\varepsilon_t - 2\varepsilon_{t-1} + \varepsilon_{t-2}),$$

тобто отримаємо стаціонарний процес. Варто відмітити, як у рівняння починає "проникати" ковзне середнє. Отриманий дворазовим взяттям різниць стаціонарний процес є процесом $MA(2)$. Проте, у крайньому випадку, взяттям послідовних різниць вихідний ряд з квадратичним трендом приводиться до стаціонарного вигляду. Підмітивши цю властивість, Бокс і Дженкінс запропонували виділити клас нестаціонарних часових рядів, які взяттям послідовних різниць можна привести до стаціонарного вигляду, а саме до рядів типу $ARMA$.

Означення 8 *Якщо ряд після взяття d послідовних різниць приводиться до стаціонарного, то він називається $ARIMA(p, d, q)$ процесом.*

Так прийнято, що d – кількість взятих різниць вставляється всередину. Скорочення I (з англ. – Integrated) означає інтегрований. Операція,

обернена до взяття послідовних різниць, – сумування. Нескінченно мала різниця називається диференціалом. Обернений підхід до диференціала називається інтегруванням. Саме даний термін використовується у абревіатурі процесу, хоча на увазі мається сумування.

$ARIMA$ – процес авторегресії інтегровного ковзного середнього. При цьому p – параметр AR-частини, d – степінь інтеграції, q – параметр MA-частини. В операторному вигляді процес $ARIMA(p,d,q)$ записується так:

$$\alpha_p(L)\Delta^d x_t = \beta_q(L)\varepsilon_t.$$

Підхід Бокса-Дженкінса.

Нехай дана реалізація деякого часового ряду. Необхідно підібрати модель до даної реалізації часового ряду, яка б описувала дану реалізацію часового ряду. Бокс та Дженкінс запропонували наступний підхід до вибору моделі типу $ARIMA$.

I етап

Ідентифікація моделі $ARIMA(p,d,q)$:

1. Встновити порядок інтеграції d, тобто зробити ряд стаціонарним, взявши достатню кількість послідовних різниць.

2. До отриманого часового ряду Y_t необхідно підібрати процес $ARMA(p,q)$. Виходячи з поведінки автокореляційної (Sample Autocorrelation Function, $SACF$) і частинної автокореляційної функції (Sample Partial Autocorrelation Function, $SPACF$), встановити параметри p та q.

II етап

Оцінка коефіцієнтиів $\alpha_1, \alpha_2, ..., \alpha_p, \beta_1, \beta_2, ..., \beta_q$ при умові, що p та q – відомі.

III етап

Тестування та діагностика побудованої моделі за залишками.

IV етап

Використання побудованої моделі. Найчастіше побудовані моделі часових рядів використовують для прогнозування майбутніх значень часового ряду.

2.3.1. Нестаціонарні часові ряди

Із теореми Вольда випливає, що моделі типу $ARMA$ включають всі стаціонарні процеси. З нестаціоарними часовими рядами ситуація інша – фактично ми будемо розглядати тільки частинні випадки нестаціонарних часових рядів. Згідно з означенням процесу $ARIMA(p, d, q)$, d – це степінь інтеграції ряду, тобто ряд стає стаціонарним після застосування d разів операції взяття послідовних різниць. Виявляється, два різних за властивостями нестаціонарні часові ряди приводяться до стаціонарного вигляду за допомогою взяття послідовних різниць.

1 тип: процес з детермінованим поліноміальним трендом. $X_t = P_k(t)$, де $P_k(t)$ – поліном степені k від t, а ε_t – стаціонарний процес, не обов'язково білий шум. Якщо обмежитися розглядом тільки лінійного тренду $X_t = \alpha + \beta t + \varepsilon_t$, то можна записати:

$$\Delta X_t = X_t - X_{t-1} = (1 - L)X_t = \beta + (\varepsilon_t - \varepsilon_{t-1}).$$

Оскільки ε_t – стаціонарний процес, то його перша різниця є також стаціонарним процесом, хоча, якщо ε_t – білий шум, то з'являється MA частина. У випадку поліноміального тренду для приведення до стаціонарного виду потрібно взяти послідовні різниці декілька разів.

2 тип: процес випадкового блукання

$$X_t = \mu + X_{t-1} + \varepsilon_t.$$

У цьому випадку $\Delta X_t = \mu + \varepsilon_t$, а процес X_t називається випадковим блуканням з дрейфом. Розв'язок даного різницевого рівняння можна записати у наступному вигляді:

$$X_t = \mu t + \sum_{j=0}^{t-1} \varepsilon_{t-j}.$$

Застосувавши підхід Бокса-Дженкінса та перейшовши до різниць, можна оцінити параметри моделі $ARMA(p, q)$. Для того, щоб повернутися назад до процесу X_t, необхідно вибрати схему, по якій повертатися.

У обох розглянутих процесів є лінійний тренд, але дані процеси відрізняються випадковою частиною. У першому випадку випадкова частина – це поточний шок, поточні збурення, а в другому випадку – це накопичення збурень від попередніх шоків.

Якщо процес випадкового блукання можна привести до стаціонарного вигляду тільки методом взяття першої різниці, то ряд першого типу можна привести до стаціонарного вигляду також за допомогою виділення лінійного тренду, наприклад, побудувавши лінійну регресію та розглянувши стаціонарний залишок.

Отже, слід розглядати два типи нестаціонарних процесів:

1 тип: процес, який зводиться до стаціонарного методом виділення лінійного тренду – TSP (trend stationary process). Даний процес має вигляд:

$$X_t = \alpha + \beta t + \varepsilon_t.$$

Даний процес приводиться до стаціонарного процесу методом включення в регресію лінійного тренду. Тобто, це процес з детермінованим трендом. Інколи даний процес називається TS.

2 тип: процес, який приводиться до стаціонарного методом взяття першої різниці – DSP (diferencing stationary process). Даний процес має наступний вигляд:

$$X_t = X_{t-1} + \varepsilon_t.$$

Інколи процес такого типу називають DS.

Властивості процесу TSP:

- Процес TSP – нестаціонарний через змінний тренд.
- Скінченна пам'ять про шоки – процес забуває про помилку одразу на наступному кроці. Якщо замість білого шуму буде стояти більш загальний процес $ARMA(p, q)$, то, звичайно, шоки здійснюють вплив протягом деякого часу, проте їхній вплив з часом слабшає.

Властивості процесу DSP:

- Процес DSP – нестаціонарний через змінну дисперсію
- Так як у явному розв'язку стоїть сума всіх попередніх ε, то шоки пам'ятаються увесь час. Даний процес є процесом з нескінченною пам'яттю. З точки зору економіки це не зовсім зрозуміло – шоки не повинні з'являтися постійно.

Процеси TSP та DSP описуються наступними рівняннями:

TSP	DSP
$X_t = \alpha + \beta t + \varepsilon_t$	$X_t = \mu + X_{t-1} + \varepsilon_t$

Табл. 1: Існуючі шляхи оцінки C_S для різних систем

Розглянемо модель:

$$X_t = \alpha + \rho X_{t-1} + \beta t + \varepsilon_t.$$

Дана модель увібрала риси обох моделей TSP та DSP. Гіпотези про характер ряду можна записати наступним чином:

$$H_0: \text{ ряд типу } DS => \rho = 1, \beta = 0.$$

$$H_1: \text{ ряд типу } TS => |\rho| < 1.$$

Тоді для гіпотези H_1 ε буде не просто білим шумом, а деяким стаціонарним рядом. Нульова гіпотеза відноситься до класу загальних лінійних гіпотез. При традиційному підході для її перевірки потрібно оцінити 2 регресії:

$$X_t = \alpha + \rho X_{t-1} + \beta t + \varepsilon_t$$

і

$$X_t = \alpha + X_{t-1} + \varepsilon_t.$$

А потім перевірити значимість різниці сум квадратів залишків, використовуючи F-статистику. Розглянемо для початку більш просту версію даної моделі:

$$X_t = \alpha + \rho X_{t-1} + \varepsilon,$$

тобто без включення лінійного тренду. У цьому випадку гіпотези матимуть наступний вигляд:

$$H_0 : \rho = 1 => DS,$$

$$H_1 : |\rho| < 1 => TS.$$

Тепер можна перевірити гіпотезу про те, що $\rho = 1$ за допомогою t-статистики. Рівняння можна переписати в іншому вигляді. Після віднімання із обох частин X_{t-1} отримуємо:

$$\Delta X_t = \alpha + (\rho - 1) X_{t-1} + \varepsilon_t.$$

Нехай $\rho - 1 = \gamma$, тоді гіпотези матимуть вигляд:

$$H_0 : \gamma = 0,$$

$$H_1 : \gamma < 0.$$

У класичній лінійній регресії для перевірки такої гіпотези застосовується одностороння t-статистика. Але у випадку виконання нульової гіпотези, ряд X_t є випадковим блуканням, його дисперсія прямує до безмежності

при збільшенні t і розподіл статистики $\frac{\hat{\gamma}}{s.e.(\hat{\gamma})}$ не є розподілом Стьюдента. Звідси випливає, що асимптотичний розподіл даної статистики не є нормальним. Причиною цього є невиконання умов Центральної граничної теореми у даному випадку. Аналітичний вираз для асимптотичного розподілу статистики $\frac{\hat{\gamma}}{s.e.(\hat{\gamma})}$ можна виразити через стохастичні інтеграли Вінерового випадкового процесу. Критичні точки даного розподілу приходиться розраховувати у числовій формі, використовуючи симуляцію процесу методом Монте-Карло. Вперше даний розподіл був розглянутий у роботі Діккі і Фуллера. Тест, який використовує для перевірки типу нестаціонарності даний розподіл, при умові, що $\gamma = 0$, тобто, коли процес належить типу DS, називається тестом Діккі-Фуллера і позначається як DF-тест. При умові, що нульова гіпотеза $\gamma = 0$ виконується, маємо процес випадкового блукання (DSP). Саме для даного випадку не працює t-статистика.

2.3.2. Тест Дікі-Фуллера

Тест Діккі-Фуллера призначений для того, щоб розрізняти часові ряди типу TS та DS. Відповідно до нульової гіпотези H_0 досліджуваний ряд належить до типу DS. Згідно з альтернативною гіпотезою він може бути типу TS і, одночасно, бути нестаціонарним, тобто містити детермінований тренд або не мати тренду – бути стаціонарним. Специфікація моделі у прямій та альтернативній гіпотезі, тобто включення або не включення у модель вільного члена і/або детермінованого тренду, впливає на розподіл t-співвідношення.

Діккі і Фуллер [93] почали з дослідження рівняння: $\Delta X_t = \gamma X_{t-1} + \varepsilon_t$. У даному випадку при умові, що виконується гіпотеза H_0, тобто, що $\gamma = 1$, розподіл t-відношення (відношення оцінки коефіцієнта γ, отриманої МНК, до оцінки його середньоквадратичного відхилення) називається DF-

розподілом. Включивши у модель вільний член, отримаємо модель $\Delta X_t = \alpha+\gamma X_{t-1}+\varepsilon$. А, додатково, включивши у модель лінійний тренд, отримаємо модель $\Delta X_t = \alpha + \beta t + \gamma X_{t-1} + \varepsilon$.

Виявляється, що у всіх трьох випадках розподіл t-співвідношення $\frac{\hat{\gamma}}{s.e.(\hat{\gamma})}$ виражається через інтеграли від Вінерового процесу, але по-різному. Всі три розподіли прийнято пов'язувати з іменами Діккі і Фуллера. Проте ці розподіли різні і залежать від того, які додаткові регресори входять у рівняння. Позначимо критичні величини для першого розподілу τ_0, для другого розподілу – τ_μ, для третього розподілу – τ_τ. Всі ці значення від'ємні, причому $\tau_\tau < \tau_\mu < \tau_0 < t_{St}$, де t_{St} – критичне значення розподілу Стьюдента.

Розглянемо ситуацію, коли ряд типу TS приймають за ряд типу DS та з'ясуємо, чи має це негативні наслідки. Іншими словами, перевіримо, чи має негативні наслідки зайва кількість взятих різниць. З точки зору підходу Бокса-Дженкінса, до чого приводить переоцінка параметра d?

Розглянемо стаціонарний процес $MA(1)$:

$$x_t = \varepsilon_t + \alpha\varepsilon_{t-1} = (1 + \alpha L)\varepsilon_t.$$

Взявши першу різницю, отримаємо:

$$y_t = \Delta x_t = (1 - L)x_t = (1 - L)(1 + \alpha L)\varepsilon_t = (1 - (1 - \alpha)L - \alpha L^2)\varepsilon_t.$$

Отримана модель тепер відноситься до типу $MA(2)$, оцінювати потрібно уже два параметри, а не один. Оцінимо звичайним способом коефіцієнти моделі $y_t = u_t+\theta_1 u_{t-1}+\theta_2 u_{t-2}$. Залишкова дисперсія процесу x_t: $Var(x_t) = (1 + \alpha^2)\sigma_\varepsilon^2$. Для процесу y_t один із коренів відповідного характеристичного рівняння рівний одиниці. Це означає, що процес y_t – необоротний, для нього не існує $AR(\infty)$ представлення. Зрозуміло, що це приведе до ускладнень при оцінюванні коефіцієнтів. Дисперсія процесу y_t: $Var(y_t) = (1+\theta_1^2+\theta_2^2)\sigma_u^2$ або можна записати по-іншому: $Var(y_t) = \left[1 + (1 - \alpha)^2 + \alpha^2\right]\sigma_\varepsilon^2$. Очевидно, що $Var(y_t) > Var(x_t)$, тобто взяття зайвих різниць приводить до

збільшення дисперсії у тому числі і дисперсії залишків після застосування методу найменших квадратів [94]. Інколи даний ефект поряд із поведінкою автокореляційної функції може служити деяким неформальним критерієм для відповіді на питання – потрібно брати послідовні різниці чи ні. Якщо при переході до наступної різниці дисперсія зростає, то, швидше за все, різницю брати не потрібно.

2.5. Знаходження відстані між часовими рядами

Переважна більшість задач, які грунтуються на часових рядах як математичних моделях, розв'язується шляхом підбору оптимальних моделей для часових рядів [95, 96, 97]. Поряд із цією проблемою, внаслідок розвитку машинного навчання та теорії випадкових процесів, все більше уваги приділяється кластеризації часових рядів [27]. Для розбиття даних на кластери необхідно вказати метрику, за допомогою якої знаходять відстані між вимірюваннями часових рядів. До знаходження відстані між часовими рядами найчастіше застосовують такі підходи:

1. Наївний. Розглядають часовий ряд як точку в R^T, де T – довжина часового ряду. При цьому кластеризація часового ряду перетворюється в кластеризацію точок в R^T [23].

2. Кластеризація на основі параметрів моделі. Тут для всіх часових рядів вибирається одна «базова» модель. Наприклад, [25, 62, 66]. Наступним кроком є припущення про розподіл основних параметрів моделі і обчислення апостеріорних оцінок параметрів часових рядів. На основі даних параметрів і здійснюється кластеризація.

Недолік першого підходу – очевидний, оскільки наївний спосіб не враховує особливості моделі часового ряду: наявність кореляції між часовими рядами, присутність детермінованих компонент (тренда, сезонної компоненти), вид автокореляційної функції, спектральну щільність тощо. Основними недоліками другого способу, як ймовірнісного методу, є те, що модель часового ряду є фіксованою і вибір початкового розподілу параметрів обирається з точки зору простоти обчислення апостеріорного розподілу. Хоча друга проблема для другого алгоритму досить легко вирішується за допомогою алгоритмів Метрополіса-Гастінга [27], Гіббса [60], однак для невеликих часових рядів невдалий підбір початкового розподілу параметрів призводить до поганого наближення апостеріорного розподілу.

Крім того, методика даного способу ґрунтується на розбитті часового ряду на три основні частини:

$$x_t = p_t + s_t + c_t, \tag{3}$$

де x_t – значення часового ряду, p_t – поліноміальний тренд, s_t – значення сезонної компоненти, c_t – випадкова складова часового ряду, для якої визначається модель часового ряду. Аналогічне розбиття нами буде використано далі у дисертаційному дослідженні. Слід зазначити, що розбиття ряду на компоненти (3) простежується також і в багатьох інших моделях випадкових процесів – семимартингали, півмартингали тощо.

У роботі розглядаються стаціонарні часові ряди: математичне сподівання і дисперсія не змінюються з часом – постійні величини; коваріація залежить тільки від відстані між вимірами часового ряду і не залежить від їх значень. Якщо вхідні часові ряди є нестаціонарними, за допомогою взяття диференціальних різниць їх приводять до стаціонарних часових рядів [68].

2.5.1. Метрики для знаходження відстані між часовими рядами

Способи оцінки подібності даних часових рядів вже давно викликають інтерес фахівців з аналізу даних, оскільки часові ряди є однією з математичних моделей, які найчастіше застосовуються для опису реальних явищ. Критичною дослідною проблемою з аналізу часових рядів є вибір функції відстані для визначення поняття подібності між двома часовими рядами.

Розглядаючи безперервні функції, Евклідова відстань є найбільш поширеною метрикою для часових рядів:

$$d_E = \sqrt{(\vec{x} - \vec{y})(\vec{x} - \vec{y})'}.$$

Зауважимо, що Евклідова відстань є інваріантною при розгляді змін в порядку уявлення часових рядів. Отже, Евклідова відстань непридатна для знаходження відстаней у багатовимірних часових рядах і не враховує кореляцію між спостереженнями, що є істотним недоліком. Для порівняння даних часових рядів, де тренди і еволюції повинні бути враховані, або, якщо форма часового ряду описується послідовністю функцій, актуальним є співвідношення Пірсона:

$$d_C(\vec{x} - \vec{y}) = 1 - \frac{(\vec{x} - \vec{y})(\vec{x} - \vec{y})'}{\sqrt{(\vec{x} - \vec{x})(\vec{x} - \vec{x})'}\sqrt{(\vec{y} - \vec{y})(\vec{y} - \vec{y})'}}.$$

яке також широко використовують, хоча воно недосконале [66].
Відстань Махаланобіса:

$$d_M(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})C^{-1}(\vec{x} - \vec{y})'}$$

можна вважати еволюційною Евклідовою відстанню, яка враховує кореляцію даних. Ця відстань використовує матрицю коваріації вхідних векторів C для «зважування» особливостей часового ряду. Відстань Махаланобіса

[86] зазвичай успішно працює з великими наборами даних, де зменшено характеристики. Більш детальний опис відстаней можна знайти в [21].

В останніх дослідженнях з цієї області розроблено серію методів для знаходження відстаней між багатовимірними часовими рядами, які можна розділити на два класи.

Перший клас включає функції, засновані на нормах L_1 и L_2 [21]. Прикладами функцій цього класу є динамічна трансформація часу (Dynamic Time Warping distance, DTW) [69] і коректування відстані зі штрафом (Edit Distance with Real Penalty, ERP) [66].

Другий клас функцій відстаней включає методи, які обчислюють оцінку подібності за порогом відповідностіε. Прикладами цього класу функцій є найдовша спільна підпослідовність (Longest Common Subsequence, LCSS) [50] і коректування відстані в реальній послідовності (Edit Distance on Real Sequence, EDR) [70]. Попередні дослідження [23] показали, що цей другий клас методів є кращим при наявності шумів і зміщення за часом.

Динамічна трансформація часу (DTW) спеціально призначена для порівняння часових рядів [65]. Застосування DTW дозволяє провести нелінійне відображення двох векторів, мінімізуючи відстань між ними. Відстань DTW може бути використано для векторів різної довжини:$X = [x_1, x_2, ..., x_n]$ і $Y = [y_1, y_2, ..., y_n]$. За допомогою DTW визначається матриця витрат C розмірності $n \times m$, яка містить відстані (зазвичай Евклідову) між двома точками x_i і y_i. Тимчасова трансформація $W = w_1, w_2, ..., w_K$, де $\max(m, n) \leq K < m + n - 1$ формуються набором матричних компонентів на наступних умовах:

1. Граничні умови: $w_1 = C(1, 1)$ і $w_K = C(n, m)$;

2. Умови монотонності: дано $w_k = C(a, b)$ і $w_{k-1} = C(a', b')$, $a \geq a'$ и $b \geq b'$;

3. Умова величини кроку: дано $w_k = C(a, b)$ і $w_{k-1} = C(a', b')$, $a - a' \leq 1$ і $b - b' \leq 1$.

Існує багато способів, що забезпечують виконання зазначених умов. Одним з таких способів є відстань DTW, за допомогою якого мінімізується значення часової трансформації в часовому ряду:

$$d_w(\vec{x}, \vec{y}) = \min \sqrt{\sum_{k=1}^{K} w_K}. \tag{4}$$

Основним недоліком (4) залишається розрахунок шляху мінімальної вартості. До того ж відстань DTW не може розглядатися як метрика, тому що воно не задовольняє нерівність трикутника.

Огляд показників метрик відстаней для кластеризації часових рядів можна знайти в [67]. Іншими важливими способами знаходження відстаней є косинус вимірювання [62], що добре підходить для моделей з різною або змінною довжиною або методи подібності Жаккара і Танімото [26], які також можна інтуїтивно розуміти як комбінацію Евклідових відстаней і кореляцій.

Отже, розглянемо часовий ряд X, який визначається послідовністю реальних значень (вимірювань), де кожне значення x_i взято в конкретний момент часу, тобто $X = [x_1, x_2, ..., x_n]$. Маючи послідовність X, яка описує часовий ряд, можна нормувати часовий ряд, використовуючи середнє μ і стандартне відхилення σ:

$$Norm(X) = \left[\frac{x_1 - \mu}{\sigma}, \frac{x_2 - \mu}{\sigma}, ..., \frac{x_n - \mu}{\sigma}\right].$$

Нормалізація рекомендується для того, щоб відстань між двома часовими рядами була інваріантною для масштабування амплітуди і зміщення часових рядів.

Проблема використання L_1 – норми для часових рядів, як зазначено вище, полягає в тому, що вона потребує, щоб часові ряди були однакової

довжини і не мали локального зміщення за часом. Якщо необхідно знайти відстань між двома часовими рядами різної довжини, можна додати вимірювання в якості пропущених значень (gaps) в часовому ряду з меншою довжиною або видалити значення в довшому часовому ряді. Тим самим зникне проблема рівності довжин і локального зміщення за часом.

Відстань коригування рядка

$$dist(y_i, x_i) = \begin{cases} 0, \text{ якщо } y_i = x_i, \\ 1, \text{ якщо } y_i \text{ або } x_i \text{ пропущені}, \\ 1, \text{ в інших випадках.} \end{cases} \tag{5}$$

де y_i, x_i – елементи рядків.

При переході від послідовностей до часових рядів складність полягає в тому, що елементи y_i і x_i є не символами, а реальними значеннями.

Для аналізу часових рядів необхідно враховувати кореляції між значеннями ряду. Відстань EDR дозволяє врахувати реальні значення за допомогою введення параметра δ, який будемо називати *порогом.*

$$dist_{EDR}(y_i, x_i) = \begin{cases} 0, \text{ якщо } |y_i - x_i| \leq \delta, \\ 1, \text{ в інших випадках.} \end{cases} \tag{6}$$

Недоліки відстані EDR: кожна різниця між елементами часових рядів дорівнює 1 (другий випадок формули (6)); використання *порогу* δ. Відстань DTW не має цієї проблеми, так як воно використовує L_1– норму між двома непропущеними елементами.

Відстань DTW відрізняється від EDR двома основними моментами, які описуються наступною формулою:

$$dist_{DTW}(y_i, x_i) = \begin{cases} |y_i - x_i|, \text{ якщо } y_i, x_i \text{ не пропущені}, \\ |y_i - x_{i-1}|, \text{ якщо } x_i \text{ пропущено}, \\ |x_i - y_{i-1}|, \text{ якщо } y_i \text{ пропущено}. \end{cases} \tag{7}$$

Основна причина, чому DTW не задовольняє нерівність трикутника, полягає в тому, що коли необхідно інтерполювати пропущений вимір, він повторює (копіює) попередній елемент. Таким чином, як показано в другому і третьому випадках формули (7), різниця між елементом і пропущеним значенням залежить від y_{i-1} або x_{i-1}.

Розглянемо відстань ERP, так як воно використовує «штраф» між двома не пропущеними значеннями і постійною величиною обчислення відстаней для пропущених значень. Таким чином, метод ERP використовує наступну формулу для обчислення відстані:

$$dist_{ERP}(y_i, x_i) = \begin{cases} |y_i - x_i|, \text{ якщо } y_i, x_i \text{ не пропущені,} \\ |y_i - g|, \text{ якщо } x_i \text{ пропущено,} \\ |x_i - g|, \text{ якщо } y_i \text{ пропущено.} \end{cases} \quad (8)$$

де g– деяка константа, «штраф» за пропущене значення або усереднене значення сусідніх значень часового ряду. Константа g може бути обрана, як середнє двох або кількох сусідніх значень часового ряду. Спеціаліст з аналізу даних за допомогою аналізу графіка часового ряду може зробити висновок стосовно того, скільки значень необхідно розглянути для заповнення пропущених вимірювань.

Ґрунтуючись на формулі (8), відстань ERP між двома часовими рядами визначається формулою (**??**) з таблиці 1 і позначається $ERP\,(Y, X)$. Ретельне порівняння формул з таблиці 1 показує, що відстань ERP можна розглядати як комбінацію L_1 – норми і EDR.

$$DTW(Y, X) = \begin{cases} 0, \text{ якщо } m = n = 0, \\ \infty, \text{ якщо } n = 0 \text{ або } m = 0, \\ dist_{dtw}(y_1, x_1) + A, \text{ в інших випадках.} \end{cases} \quad (9)$$

$$A = \min\left\{DTW(Rest(Y), Rest(X)), DTW(Rest(Y), X), DTW(Y, Rest(X))\right\}.$$

$$EDR(Y,X) = \begin{cases} n, \text{ якщо } m = 0, \\ m, \text{ якщо } n = 0, \\ EDR(Rest(Y), Rest(X)), dist_{edr}(y_1, x_1) = 0, \\ B, \text{ в інших випадках.} \end{cases} \tag{10}$$

Складність алгоритмів DTW і ERP становить $O(NM)$, де N і M – довжини двох вхідних послідовностей. Без втрати спільності, припускаючи, що $N \geq M$, складність алгоритму становить $O\left(N^2\right)$. Алгоритми мають ту ж складність при розгляді багатовимірних часових рядів в просторі.

2.5.2. Алгоритм пошуку оптимальних моделей

Як було зазначено вище, важливу роль в дослідженні часових рядів відіграє кластеризация часових рядів. В якості ступеню схожості при кластеризації є наступна міра [98]:

$$HM(G_1, ..., G_m) = \sum_{k=1}^{m} \frac{2}{n_k - 1} \sum_{i<j \in G_k} d(M_i, M_j) \ , \tag{11}$$

де $d(M_i, M_j)$ – відстань між моделями часових рядів i і j. Дану відстань, як основний об'єкт роботи, ми обчислимо нижче. Отже, головною проблемою для визначення ступеню схожості $HM(G_1, ..., G_m)$ є визначення відстані між часовими рядами $d(M_i, M_j)$.

Розглянемо N часових рядів, функціонування яких відстежується протягом T періодів. Тобто дані мають такий вигляд:

$$\begin{pmatrix} x_{11} & ... & x_{1T} \\ x_{12} & ... & x_{2T} \\ ... & ... & ... \\ x_{N1} & ... & x_{NT} \end{pmatrix}. \tag{12}$$

Базовою моделлю часового ряду, як було зазначено вище, є $ARMA\,(p,q)$ модель:

$$\Phi\,(L) = \phi_0 + ... + \phi_p L^p,$$

$$\Theta\,(L) = \theta_0 + ... + \theta_q L^q,$$

причому $\phi_0 = \theta_0 = 1$; L – лаговий оператор.

Таким чином, поряд з роботою [98], де визначалася одна модель для всіх часових рядів, в нашому випадку для перевірки потрібно переглянути:

$$N_{total} = ((P+1)\,(Q+1))^N \tag{13}$$

моделей, що зазвичай є неможливим при великій кількості моделей, тобто при великих P, Q і N. З огляду на обчислювальну складність поставленого завдання, в роботі запропоновано алгоритм пошуку відстані між двома часовими рядами, грунтуючись на їх моделях, з використанням тільки $ARMA\,(P,Q)$ для всіх часових рядів.

У більшості робіт, «відстань» між моделями, яка характеризує часові ряди x_1 та x_2, визначаються в такий спосіб:

$$d\,(M_1, M_2) = \sqrt{\sum_{t=1}^{T} (x_{1t} - x_{2t})^2},$$

тобто розглядається L_2– норма.

Цей метод взятий з наївного підходу і не враховує моделей, згідно з якими функціонують часові ряди. Нами для визначення відстані між моделями буде використано поняття $MA\,(\infty)$ уявлення [68]. Відповідно до твердження розділу 2 [32], довільний стаціонарний часовий ряд x_t, що заданий за допомогою $ARMA\,(p,q)$ моделі (14) однозначно можна записати у вигляді $MA\,(\infty)$ уявлення:

$$x\,(t) = \tilde{\Theta}\,(L)\,\varepsilon_t, \tag{14}$$

де

$$\tilde{\Theta}(L) = L + \tilde{\theta}_1 L + ... + \tilde{\theta}_k L^k + ..., \tag{15}$$

причому для моделі, яка задає стаціонарний часовий ряд,

$$\sum_{k=1}^{\infty} \left|\tilde{\theta}_k\right|^2 < \infty.$$

Використовуючи уявлення (14) для $ARMA(p,q)$ моделі, визначимо відстань між моделями M_1 и M_2 наступним чином:

$$d(M_1, M_2) = \sum_{k=1}^{\infty} \left|\tilde{\theta}_k^1 - \tilde{\theta}_k^2\right|. \tag{16}$$

Для моделювання попередньої формули на прикладі слід зазначити, що ряд $\sum_{k=1}^{\infty} \left|\tilde{\theta}_k^1 - \tilde{\theta}_k^2\right|$є збіжним. Тому можна знехтувати елементами ряду після деякого K. Дана відстань буде інтерпретуватися саме як відстань між моделями, а не відстань між даними.

2.5.3. Порівняння відстаней

Для порівняння відстаней між часовими рядами нами буде використаний метод Монте-Карло [61]. В якості моделей для порівняння розглянемо L_1 – відстань, DTW і ERP методи. Основними характеристиками, які обумовлюють зміну відстаней і правильність визначення моделей часового ряду, є наявність викидів, структурна зміна часового ряду [18] і довжина часового ряду. У наступних обчисленнях нами буде досліджено вплив наявності викидів і довжини часового ряду на похибку в обчисленні між часовими рядами.

Розглянемо вплив наявності аутлайерів на відстань між самими часовими рядами і відстанями між їх моделями. Для цього розглянемо генерацію двох часових рядів з $T = 150$ дослідженнями, перший з яких моделюється $ARMA(p,q)$ моделлю, другий – це той же часовий ряд, в якому $\alpha\%$ значень замінені викидами.

В якості оригінального часового ряду взято $ARMA\,(2,1)$ модель:

$$x_t = 0.3x_{t-1} - 0.4x_{t-2} + \varepsilon_t + 0.1\varepsilon_{t-1}.$$

З наступного Рис. 1 видно, що при наявності великої кількості викидів, класичні методи лінійно збільшують відстані між часовими рядами, в той же час відстань за моделями поводиться як логарифмічна функція. При цьому слід зауважити, що при моделюванні розглянуто невеликі викиди, які дещо виходять за межі інтервалу $(Q_1 - 1.5IQR, Q_1 + 1.5IQR)$. У разі ж наявності великих викидів, різниця між тимчасовими рядами зростатиме лінійно з коефіцієнтом, рівним середній величині викиду.

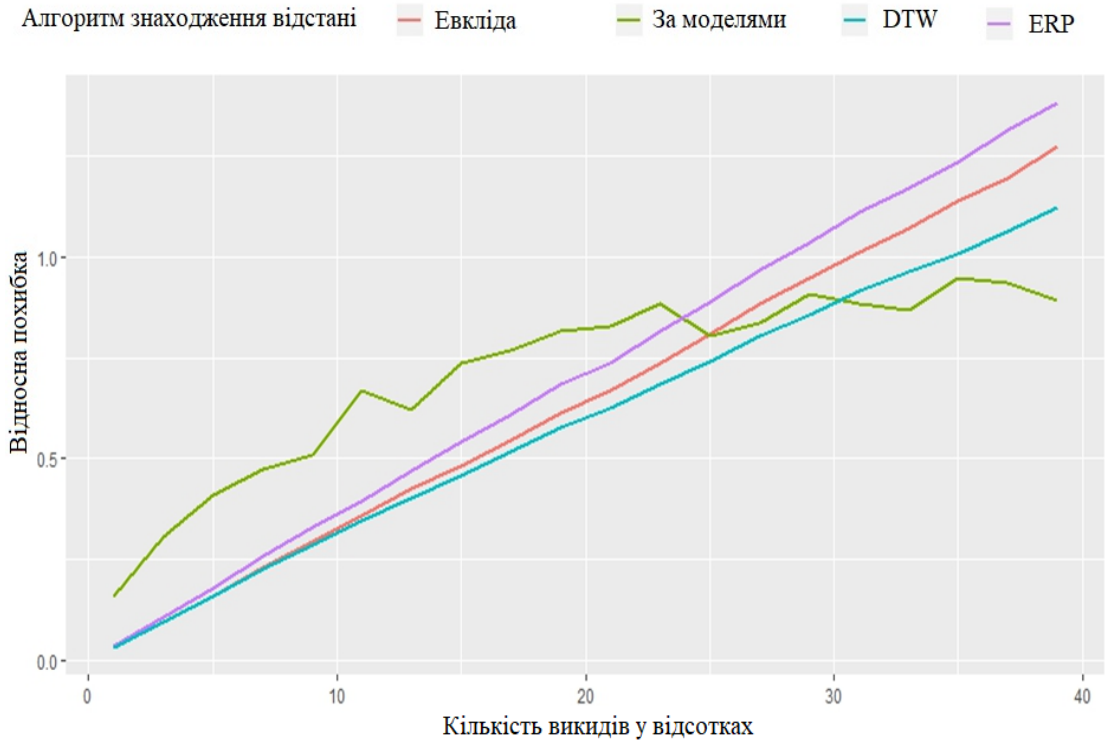

Рис. 1: Відносна похибка як функція від величини викидів в часовому ряду для Евклідової відстані, відстані по моделях DTW і ERP методів

Поряд з цією залежністю, розглянемо також залежність відстаней від величини часового ряду (Рис. 2), тобто від величини T.

Таким чином, з Рис. 2 зрозуміло, що великих T відстані між часовими рядами для однакових стаціонарних моделей є практично постійними для

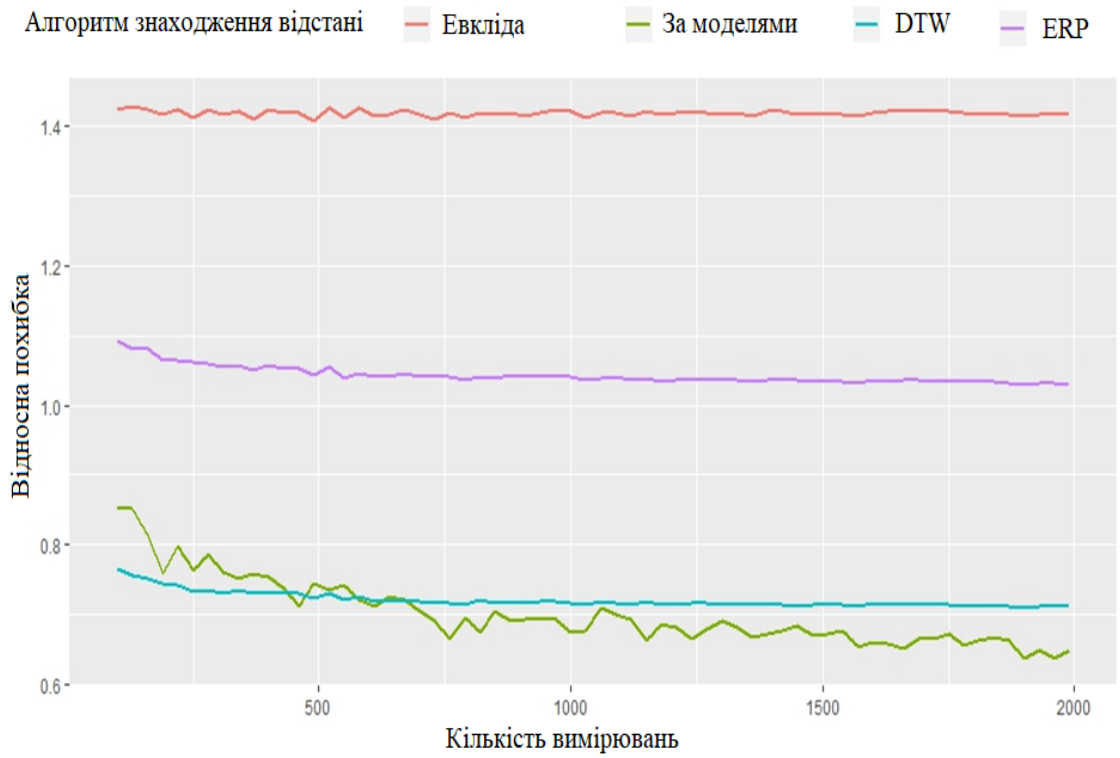

Рис. 2: Відносна похибка як функція від кількості вимірювань у часовому ряду для Евклідової відстані, відстані по моделях DTW і ERP методів

всіх методів, однак для оцінки відстані за моделями ця константа значно менша, ніж відповідна константа для всіх інших методів.

Висновки до розділу II

У даному розділі здійснено огляд стаціонарних ARMA та нестаціонарних ARIMA моделей часових рядів, які використано для визначення відстані між часовими рядами. ARMA та ARIMA процеси є популярними та потужними інструментами для моделювання та прогнозування часових рядів і використовуються в різних галузях, включаючи фінанси, економіку, погоду та інші.

ARMA модель включає процеси авторегресії та ковзного середнього. Ця модель дозволяє аналізувати залежності між значеннями ряду в різний

момент часу. ARIMA модель доповнює ARMA модель компонентом інтегрування. Цей компонент дозволяє моделювати нестаціонарні ряди, тобто ряди, у яких статистичні характеристики змінюються з часом. ARIMA модель робить ряд стаціонарним шляхом віднімання початкових значень від кожного значення ряду. Потім ARMA модель застосовується до змодельованого стаціонарного ряду. ARIMA може бути корисною для аналізу та прогнозування рядів, де присутні тренди та сезонні компоненти. Основне завдання при роботі з даними полягає в обранні підходящої моделі та правильній оцінці параметрів у залежності від характеристик досліджуваного часового ряду.

В розділі II розглянуто відстань між часовими рядами, що ґрунтується не на даних, а на моделях часових рядів. Використовуючи метод Монте-Карло, показано, що дана відстань є більш стійкою до викидів і дає точніші результати для більш довгих часових рядів (при великих T). Крім того, складність алгоритму обчислення даної відстані для N часових рядів становить $O\left(T * N^2\right)$, в той же час аналогічна складність для алгоритмів DTW, ERP, EDR становить $O\left(T^2 * N^2\right)$. Немає сумнівів в тому, що використання моделей є одним з найбільш зручних інструментів дослідження подібності процесів. Крім того, для аналізу зручно використовувати усереднені еволюції і граничні еволюції в схемі дифузійної апроксимації [28]. Також за рахунок стійкості до викидів, дану відстань можна використовувати при кластеризації для побудови більш стійких до шумів кластерів.

РОЗДІЛ ІІІ. КЛАСТЕРИЗАЦІЯ З ВИКОРИСТАННЯМ МАРКОВСЬКОГО АЛГОРИТМУ

Кластеризація об'єктів різної природи є на даний час однією з найбільш важливих задач машинного навчання та глибинного навчання [99]. Всі задачі кластеризації (Unsupervised learning) об'єднує декілька спільних проблем: визначення метрики в просторі об'єктів [100, 101, 102, 103], для яких проводиться кластеризація, та визначення оптимальної кількості кластерів для даних об'єктів [104, 105, 106, 107]. У випадку числових даних (структурованих даних) перша проблема вирішується за допомогою розгляду відстані в R^n як метрики. Проте для нечислових (неструктурованих даних) вибір потрібної метрики відіграє важливу роль при подальшій кластеризації. Неструктурованих об'єкти повністю (або частково) можуть визначатися часовими рядами та графами. Кластеризація на графах та кластеризація часових рядів є одними із найважливіших розділів машинного навчання [108, 109], оскільки багато об'єктів задаються саме за допомогою даних неструктурованих об'єктів [73]. Тому розгляд нових алгоритмів кластеризації для даних об'єктів є одним із основних напрямків машинного навчання в даний час.

Дослідження в області вибору оптимальної кількості кластерів є ключовим кроком у задачах кластеризації часових рядів. Кластеризація дозволяє групувати схожі часові ряди разом і відділяти різні групи часових рядів одна від одної. Це важливо для багатьох прикладних задач, таких як

аналіз споживчої поведінки, класифікація медичних даних або виявлення аномалій в часових рядах [110, 111, 112]. Однак визначення оптимальної кількості кластерів – це нетривіальна задача.

Існують різні методи визначення оптимальної кількості кластерів. Один з найпоширеніших підходів – це використання "ліктя"(elbow method) [113, 114, 115]. Він полягає в аналізі варіації між кластерами при різній кількості кластерів. Зазвичай, зі збільшенням кількості кластерів варіація між кластерами буде зменшуватися, але на певному етапі цей приріст варіації зменшиться і навіть стане практично нульовим. Точку, де ця зміна стає меншою, ніж величина "ліктя" можна вважати оптимальною кількістю кластерів. Однак цей метод не завжди є достатньо точним, і він може давати неправильні результати, особливо коли структура даних складна та нерівномірна.

Ще однією поширеною технікою для визначення кількості кластерів є силуетний аналіз (silhouette analysis) [116, 117, 118]. Він оцінює, наскільки кожен об'єкт даних схожий на інші об'єкти у своєму кластері порівняно з іншими кластерами. Ця метрика надає значення від -1 до 1, де більші значення вказують на кращу роздільність між кластерами. Оптимальна кількість кластерів визначається як та, яка має найвище середнє значення силуету. Цей підхід може бути більш точним, ніж метод "ліктя але він також має свої обмеження, особливо коли дані мають складну структуру.

Іншим підходом до визначення кількості кластерів є використання зовнішніх метрик, таких як Adjusted Rand Index (ARI) [119, 120, 121] або Normalized Mutual Information (NMI), для порівняння різних кількостей кластерів. Ці метрики вимірюють подібність між дійсними мітками та мітками, призначеними алгоритмом кластеризації, і можуть допомогти визначити оптимальну кількість кластерів. Однак цей підхід також має свої обмеження, і він вимагає наявності дійсних міток в навчальних даних, що

не завжди можливо у реальних завданнях.

Алгоритми кластеризації для часових рядів можна умовно розділи на дві категорії:

- ***Наївні алгоритми***. Дана категорія алгоритмів базується на припущенні, що часовий ряд є вектором у Евклідовому просторі R^n. Недолік даних алгоритмів очевидний, оскільки не враховується структура моделей часових рядів [7].

- ***Модель – орієнтовані алгоритми.*** Дані алгоритми враховують припущення щодо моделей часових рядів, які потрібно кластеризувати.

Кожна із запропонованих категорій володіє своїми перевагами та недоліками, проте спільною проблемою для кожного алгоритму є визначення оптимального числа кластерів k_{opt}. Найбільш вживаними методами для визначення k_{opt} є метод ліктя, метод силуета, k-Core decomposition.

В даній роботі основна увага буде приділена саме визначенню оптимальної кількість кластерів k_{opt} в незалежності від категорії, до якої належить метод для побудови метрики. На відміну від багатьох аналогічних робіт, в даному дослідженні буде використано марковський алгоритм кластеризації, який грунтується на використанні стохастичної матриці графа.

Так як власні значення стохастичних матриць є визначальними при кластеризації даних, варто відзначити роботу [75]. Biely та Thurner вивчали форму спектрів власних значень корреляційних матриць часових рядів, отриманих із наборів броунівських випадкових блукань зі зсувом по часу. Дані матриці можна розглядати як випадкові, асиметричні матриці із спеціальною структурою внаслідок зсуву по часу. У роботі показано, що спектр власних значень є круговим для стохастичних матриць. Щільність власних значень отримано на основі некорельованих гауссовських часових

рядів. Теоретичні результати порівнюються зі щільностями власних значень, отриманими з високочастотних (5 хв) даних S&P500. У роботі також ідентифіковано різні невипадкові закономірності та знайдено асиметричні залежності, пов'язані з власними знаннями.

3.1. Основні позначення та припущення

У цьому розділі описано алгоритм кластеризації на графах з використанням дискретних ланцюгів Маркова. Будь-який реальний процес людського життя, такий як функціонування транспортних та комп'ютерних мереж, проектування конструкцій та молекулярне моделювання, можна описати за допомогою графів. Крім того, алгоритм кластеризації з використанням дискретних ланцюгів Маркова є швидким і масштабованим та може бути застосований до будь-якого об'єкта, для якого задано матрицю подібності. Використовуючи цей алгоритм, вдається визначити оптимальну кількість кластерів k_opt, які неможливо отримати за допомогою деяких класичних методів (метод ліктя та метод розкладання $k - core$). Перевірка роботи алгоритму проводиться за допомогою симуляції методом Монте-Карло. Крім того, теоретичні спостереження, отримані в дослідженні, пов'язані з деякими властивостями стохастичних випадкових матриць, які є визначальними у випадку застосування кластеризації на графах.

Кластеризація на графах є одним із найважливіших підрозділів машинного навчання, так як більшість реальних процесів задаються графами [1]. Тому для початку важливо розглянути найновіші дослідження та алгоритми кластеризації на графах. Алгоритми кластеризації на графах можна умовно поділити на дві категорії:

- ***детерміновані алгоритми кластеризації***, які грунтуються на властивостях графів (матриці суміжності вершин графа). Прикла-

дом детермінованих алгоритмів кластеризації на графах є алгоритм Hirvan–Newman [2]. Алгоритм Hirvan–Newman визначає спільноти шляхом послідовного видалення ребер з вихідної мережі. Зв'язані компоненти мережі, яка залишається, виступають спільнотами. Замість спроб побудови міри, яка показує, яке ребро є центральним в сукупностях, алгоритм Hirvan–Newman фокусується на ребрах, які найбільш ймовірно знаходяться між спільнотами. Даний алгоритм є представником ієрархічної кластеризації на графах.

- ***Марковські алгоритми кластеризації***, які грунтуються на припущенні про те, що граф разом із матрицею суміжності A визначають однорідний ланцюг Маркова (ОЛМ) з матрицею переходу P, яка може бути визначена наступним чином:

$$P_{ij} = \frac{A_{ij}}{\sum_{j=1}^{N} A_{ij}}. \quad (1)$$

.

Алгоритм кластеризації на графах, розглянутий в роботі [3], є хорошим прикладом Марковського алгоритму кластеризації. Даний алгоритм базується на властивостях дискретного ланцюга Маркова, які визначаються матрицею переходів (1).

Обидві з розглянутих категорій мають свої переваги та недоліки. Тим не менше спільним завданням для обох категорій є визначення оптимальної кількості кластерів k_{opt}. Одним із найбільш ефективних способів отримати k_{opt} є розв'язок оптимізаційної задачі для функції $f(k)$, де k – кількість кластерів, f – функція, яка використовується для визначення якості кластеризації [4]. Сучасні оптимізаційні задачі, які використовуються для визначення k_{opt}, включають метод ліктя, метод силуету та $k - core$ декомпозицію. Розглянемо детальніше згадані методи.

Метод ліктя. Ключовим моментом кожного алгоритму кластеризації є визначення найкращої кількості кластерів k_{opt}. Метод ліктя є найбільш популярним методом для цього завдання. Він передбачає багаторазовий запуск алгоритму через цикл зі збільшенням кількості обраних кластерів; згодом будується індикатор кластеризації у залежності від кількості кластерів. Метод ліктя включає наступні кроки [6]:

1. Кластеризація виконується за допомогою методу k-середніх. Поряд із роботою k-середніх обчислюється та записується функція штрафів.

2. Будується графік залежності функції штрафів від відомої кількості кластерів.

3. Вибирається та кількість кластерів, яка відповідає кількості перегинів графіка. Це число і є оптимальною кількістю кластерів.

Метод декомпозиції $k - core$. Оптимальна кількість кластерів також може бути визначена за допомогою методу $k - core$ декомпозиції. Цей метод використовується для ідентифікації найбільшої зв'язаності графу і дозволяє швидко знайти $k-core$ – максимально зв'язаний підграф, в якому кожна вершина пов'язана принаймні з k сусідніми вершинами в підграфі.

Метод $k - core$ декомпозиції використовується зазвичай для аналізу великих нейронних мереж. Метод $k-core$ декомпозиції є алгоритмои $O(m)$, де m позначає кількість потоків у непаралельних обчисленнях [8, 9]. Його мета – отримати підгрупу, члени якої представляють комунікаторів на графіку. Кожен вузол у підграфі повинен мати принаймні k градусів.

Метод $k - core$ декомпозиції має наступні властивості :

$$\forall u \in V : \; k - core\,(u) \; = \; k \leftrightarrow$$

$$\leftrightarrow \begin{cases} A\ maximum\ subgraph\ V_k\ exists\ such\ that\ \forall v \in V_k : deg\,(v) \geq k \\ and \\ a\ subgraph\ V_{k+1}\ does\ not\ exist\ such\ \ that\ \forall\ v \in V_{k+1} : deg\,(v) \geq k+1. \end{cases}$$

Цей алгоритм був розроблений для отримання підграфа, який демонструє найсильніші зв'язки k. Остання властивість вказує на те, що кожен учасник цього підграфа має принаймні k сусідів і, що більший підграф, де кожен учасник має більше k сусідів, не існує. Тому вершина з найвищим ступенем у цьому підграфі є гарним кандидатом стати центром кластера.

Як було зазначено вище, часові ряди та графи є неструктурованими об'єктами, що робить їх подібними в контексті задачі кластеризації та визначення оптимальної кількості кластерів k_{opt}. Тому далі розглядатимемо лише графи, розуміючи, що часові ряди мають аналогічну структуру.

Розглянемо основні позначення графів, які будуть використовуватися у даній роботі. Граф позначатимемо через $G = (V, E)$, де $V = \{1, 2, \ldots, N\}$ – множина вершин графа (окремі часові ряди), $E = \{e_1, \ldots, e_m\}$ – множина ребер графа, що з'єднують вершини із V (визначаються модулями коваріацій для часових рядів). Припустимо, що граф задається матрицею суміжності A:

$$A = A_{N \times N},$$

де елемент A_{ij} рівний ваговому коефіцієнту між вершинами i та j. Припустимо, що $A_{ij} \in R_+$, причому для графів A_{ij} утотожнюють з кількістю ребер між вершинами i та j, для часових рядів A_{ij} будемо утотожнювати із модулем коваріації між часовими рядами, що визначають i-й та j-й об'єкти. Стохастична матриця P, що відповідає графу G та представлена матрицею суміжності A, задається співвідношенням

$$P_{ij} = \frac{A_{ij}}{\sum\limits_{j=1}^{N} A_{ij}}. \tag{17}$$

Дана матриця є стохастичною матрицею, яка і буде основним об'єктом дослідження в даній роботі.

Розглянемо декілька основних тверджень, які будуть далі використані при оцінці кількості кластерів k_{opt}. Основне твердження щодо визначення оптимальної кількості кластерів k_{opt} тісно пов'язане із алгебраїчною кратністю власного значення $\lambda = 1$ для стохастичної матриці P [76].

Теорема 2 *Алгебраїчна кратність власного значення $\lambda = 1$ для стохастичної матриці P рівна кількості незвідних класів для ланцюга Маркова, що описується матрицею P.*

Отже, дана теорема стверджує, що кількість неперетинних класів рівна кратності власного значення $\lambda = 1$ для матриці P, що в свою чергу співпадає з кількістю кластерів у кластерному аналізі. Даний результат є теоретичним, оскільки в реальних системах всі стани є сполучними. Використовуючи неперервну залежність між елементами матриці та її власними значеннями, буде показано, що кількість кластерів в стохастичній матриці характеризує кількість власних значень, близьких до 1 . Дану залежність можна описати наступним співвідношенням, яке буде використано при доведенні основного твердження роботи:

$$\sum_{i=1}^{N} \left(\lambda_i\left(A\right) - \lambda_i\left(B\right)\right)^2 \ \leq Tr(A-B)^2, \tag{18}$$

де $A,\ B$ – квадратні матриці розмірності $N \times N$; $\lambda_i\left(A\right),\ \lambda_i\left(B\right),\ i = 1, .., N$ – власні значення матриць $A,\ B$, відповідно, упорядковані за спаданням чи зростанням.

3.2. Випадкові та стохастичні матриці

Теорія випадкових матриць є однією із математичних моделей автоматизованих розумних енергосистем [80], що описують функціонування ве-

ликих незалежних систем. Одним із основних завдань, пов'язаних із побудовою розумної системи, є розподіл функцій системи між основними "центрами" системи. Ще одним прикладом застосування випадкових матриць є системи зв'язку та фінансова математика, розглянуті в роботі [72]. Таким чином, випадкові матриці є однією із математичних моделей для опису Big Data. Основні результати теорії випадкових матриць по'вязані із спектральним аналізом даних матриць, тобто із властивостями власних значень. Одним із найбільш важливих результатів, отриманих в даному напрямку, є результати українських математиків Марченко та Пастура [81], якими було знайдено асимптотичний розподіл власних значень симетричної випадкової матриці

$$Z = \frac{1}{m} X X',$$

де X – симетрична матриця розмірності $n \times m$, елементами якої є незалежні випадкові величини із нульовим середнім та скінченною дисперсією σ^2. Основний результат роботи [81] стверджує, що випадкова величина Λ, що визначається наступним розподілом

$$P\left(\Lambda \in B\right) = \frac{1}{m} \#\left\{\lambda_i\left(Z\right) \in B\right\},$$

має граничним розподілом розподіл Марченко – Пастура з параметрами (μ, a, b), який задається щільністю

$$f\left(x\right) = \frac{\sqrt{(b-a)(x-a)}}{2\pi\sigma^2\mu x},\ x \in [a, b],$$

де

$$\mu = \lim_{m\to\infty} \frac{n(m)}{m};$$
$$a = \sigma^2(1-\sqrt{\mu})^2;$$
$$b = \sigma^2(1-\sqrt{\mu})^2.$$

Розглянемо поняття стохастичних випадкових матриць, яке буде використане при дослідженні основного результату роботи.

Означення 9 *Стохастичною випадковою матрицею P називається квадратна матриця, що володіє наступними властивостями:*

1. *Елементи матриці P_{ij} є невід'ємними випадковими величинами.*

2. *Елементи матриці P_{ij} є нормованими, тобто*

$$\sum_{j=1}^{N} P_{ij} = 1.$$

3. *Рядки матриці P є незалежними випадковими векторами.*

Розглянемо приклад стохастичної випадкової матриці та розглянемо основні особливості даної матриці. Припустимо, що матриця суміжності на графі A розмірності $N \times N$ є випадковою матрицею, елементи якої мають рівномірний розподіл на довільному інтервалі $[a, b]$, де

$$0 \leq a < b < \infty.$$

Тобто в даному прикладі елементи матриці мають рівномірний розподіл на $[a, b]$:

$$A_{ij} \sim U(a, b),$$

де $U(a, b)$ – рівномірний розподіл на $[a, b]$. Використовуючи матрицю A, утворимо стохастичну матрицю за формулою (17). Розглянемо деякі властивості власних значень матриці P, які безпосередньо випливають із означення стохастичної випадкової матриці.

1. З ймовірністю 1, власним значенням матриці P є $\lambda = 1$.

2. Власні значення матриці P задовольняють співвідношення

$$0 \leq \sum_{i=1}^{N} \lambda_i(P) \leq N.$$

3. Математичне сподівання власних значень рівне 1:

$$\sum_{i=1}^{N} E(\lambda_i(P)) = 1.$$

Остання властивість є наслідком рівності суми власних значень матриці та сліду матриці:

$$\sum_{i=1}^{N} \lambda_i(P) = \sum_{i=1}^{N} P_{ii}$$

та того факту, що всі елементи матриці P мають однаковий розподіл. Враховуючи дане зауваження, отримаємо наступні співвідношення для елементів матриці P:

$$EP_{ij} = \frac{1}{N}.$$

Таким чином для власних значень матриці P справедлива наступна рівність:

$$\sum_{i=2}^{N} E(\lambda_i(P)) = 0,$$

де власні значення матриці P впорядковані за спаданням дійсних частин:

$$1 = Re(\lambda_1) \geq Re(\lambda_2) \geq ... \geq Re(\lambda_N).$$

Розглянемо результати симуляції власного значення λ_2, що найближче до 1, тобто

$$\lambda_2 = argmin_{\lambda_i \neq 1} \{|\lambda_i - 1|\}.$$

Із наведеного нижче Рис. 3 видно, що розподіл власного значення $\lambda_2(N)$ "стягується" до 0 у випадку рівномірного розподілу всіх елементів матриці суміжності A. Даний результат може бути інтерпретований як узагальнення класичного результату в класичному алгоритмі Маркова [**?**], який можна сформулювати наступним чином:

Лема 3 *[122] Нехай всі елементи стохастичної матриці P рівні між собою, тобто*

$$P_{ij} = \frac{1}{N}.$$

Тоді кількість кластерів, обчислена згідно класичного алгоритму рівна 1 *для довільних значень параметрів* r *та* s*.*

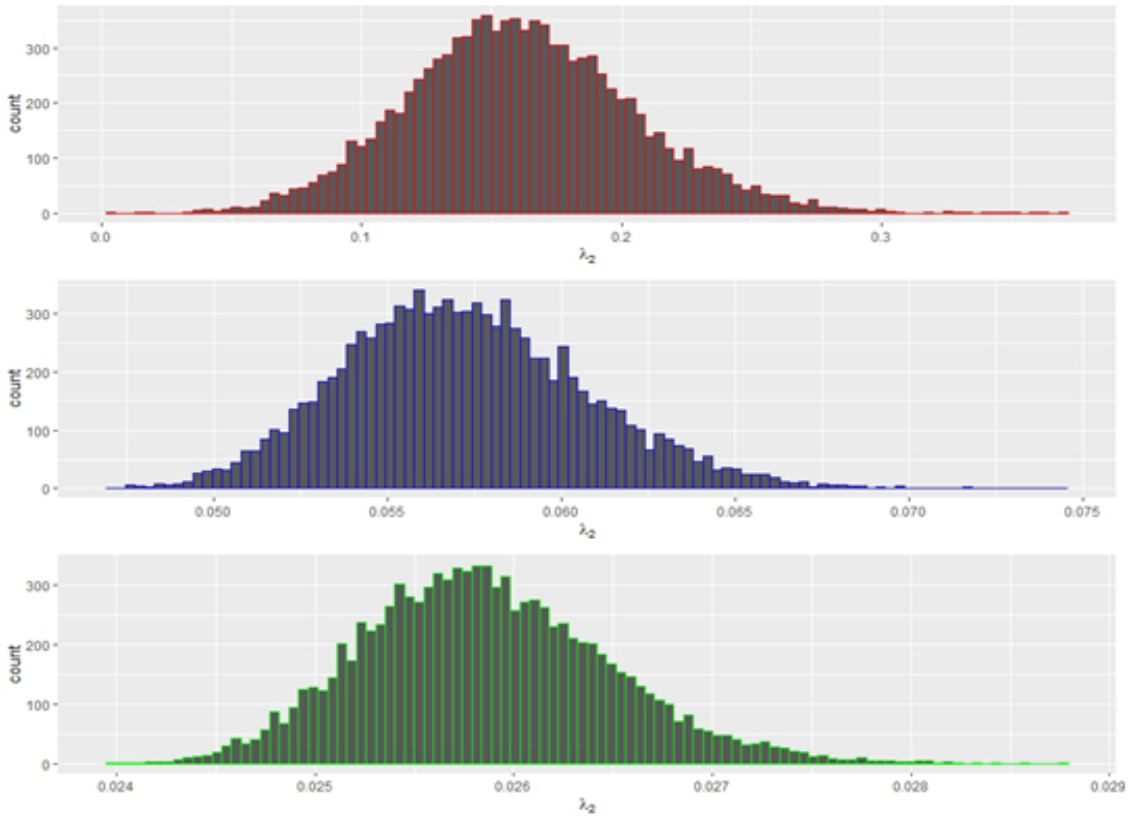

Рис. 3: Розподіл λ_2 матриці P за умови $A_{ij} \sim Unif\,(0,1)$ з розмірністю $N = 10,\ 100,\ 500$ (зверху вниз)

Даний результат є тривіальним наслідком з використанням двох операцій у класичному методі:

$$P = P^r,$$

$$P = L\,(P; s)\,,$$

де оператор $L\,(P; s)$ визначає операцію інфляції для матриці P з параметром s. Результат леми 3 є частинним випадком навіть для графів невеликої розмірності, оскільки при найменших відхиленнях значень P від свого середнього значення, результати використання класичного Марковського алгоритму змінюються в залежності від параметрів алгоритму (r, s). Нижче буде доведено більш сильний результат для довільного розподілу

на $(0, \infty)$, що узагальнює лему 3. Даний результат буде використано для відслідковування наявності одного кластера в графі, що задається стохастичною випадковою матрицею P.

Теорема 4 *Нехай виконуються наступні умови:*

1. *Всі елементи матриці A є незалежними та мають одинаковий розподіл*

$$A_{ij} \sim Distr,$$

де розподіл $Distr$ має носієм підмножину множини $(0, \infty)$ та має скінченний момент порядку $2 + \delta, \delta > 0$.

2. *Елементи матриці переходу за 1 крок P_{ij} визначаються із співвідношення (17).*

Тоді за ймовірністю має місце збіжність

$$\lambda_2 (P; N) \to 0,$$

де $\lambda_2 (P; N) \neq 1$ – найближче до 1 власне значення матриці P, $N \times N$ – розмірність матриці P.

Доведення. Для доведення даного факту відзначимо, що значення $\lambda_2 (N)$, яке відповідає значенням матриці

$$P_{ij} = E\left(\frac{A_{ij}}{\sum_{j=1}^{N} A_{ij}}\right)$$

рівне 0. Для доведення цього факту розглянемо власні значення матриці P у наступному вигляді

$$\det (\lambda I - P) \ = \lambda^n + c_1 \lambda^{n-1} + \ldots c_{n-1}\lambda + c_n,$$

де коефіцієнти c_i визначаються за наступною формулою

$$c_i = (-1)^i \sum_k \det (M_k (i)) \ ,$$

де $M_k(i)$ – головні мінори матриці P розмірності $i \times i$. Згідно побудови,

$$c_2 = ... = c_N = 0.$$

Використовуючи даний факт, знайдемо, що характеристичний поліном для матриці $E(P)$ рівний

$$E\left(\det\left(\lambda I - P\right)\right) = E\left(\lambda^n + c_1\lambda^{n-1} + \cdots + c_{n-1}\lambda + c_n\right) =$$

$$= \lambda^n + E\left(c_1\right)\lambda^{n-1} + \cdots + E\left(c_{n-1}\right)\lambda + E\left(c_n\right) =$$

$$= \lambda^n - \lambda^{n-1}.$$

Таким чином, розв'язками характеристичного рівняння є

$$\lambda_1(E(P)) = 1,\ \lambda_i(E(P)) = 0,\ i = 2, \ldots, N.$$

Отже, справедливим є той факт, що математичне сподівання власного значення λ_2(P) рівне 0 за умов теореми.

Для доведення збіжності до 0 випадкової величини λ_2 розглянемо дисперсії діагональних елементів

$$D\left(\sum_{i=1}^{N} P_{ii}\right) = \sum_{i=1}^{N} D\left(P_{ii}\right) = ND\left(P_{11}\right),$$

оскільки всі значення P_{ii} є незалежними та однаково розподіленими. Обчислимо дисперсію випадкової величини

$$D\left(P_{11}\right) = E\left(P_{11}^2\right) - \frac{1}{N^2}.$$

Використовуючи означення матриці P, можна показати, що

$$E\left(P_{11}^2\right) = \frac{1}{N^2} + O\left(\frac{1}{N^2}\right).$$

Таким чином

$$D\left(P_{11}\right) = O\left(\frac{1}{N^2}\right)$$

та

$$D\left(\sum_{i=1}^{N} P_{ii}\right) = O\left(\frac{1}{N}\right).$$

Для доведення того факту, що $\lambda_2(P;N)$ прямує до 0 за ймовірністю, використаємо нерівність (18). Використовуючи дане співвідношення, отримаємо наступну оцінку для дисперсії $\lambda_2(P;N)$:

$$E(\lambda_2(P;N))^2 = D(\lambda_2(P;N))^2 \leq$$

$$\leq \sum_{i=1}^{N} D(\lambda_i(N)) \leq E\left(\sum_{i=1}^{N}\sum_{k=1}^{N}\left(P_{ik}-\frac{1}{N}\right)\left(P_{ki}-\frac{1}{N}\right)\right) =$$

$$= E\left(\sum_{i=1}^{N}\left(P_{ii}-\frac{1}{N}\right)^k\right) = \sum_{i=1}^{N} E\left(P_{ii}-\frac{1}{N}\right)^2 =$$

$$= \sum_{i=1}^{N} D(P_{ii}) = D\left(\sum_{i=1}^{N} P_{ii}\right).$$

Отже, справедлива нерівність

$$D(\lambda_2(N)) \leq O\left(\frac{1}{N}\right).$$

Використовуючи дану нерівність та нерівності Чебишова, отримаємо твердження теореми.

Результати моделювання власних значень для рівномірного розподілу $A_{ij} \sim U(a,b)$ наведені на Рис.4. З даного рисунка видно, що власні значення $\lambda_2(P;N), \ldots, \lambda_N(P;N)$ "стягуються" до 0 з ростом N. Цей факт описує поведінку власних значень стохастичної матриці графа у випадку наявності тільки одного кластера.

Загальний вигляд розподілу власних значень $\lambda_2(N), \ldots, \lambda_N(N)$ при довільній кількості кластерів можна знайти, використовуючи методику, запропоновану Марченком та Пастуром [81]. Зауважимо, що на відміну від класичного випадку випадкових матриць, стохастичні випадкові матриці володіють трьома особливостями: по-перше, середнє значення кожного елемента стохастичної випадкової матриці має математичне сподівання

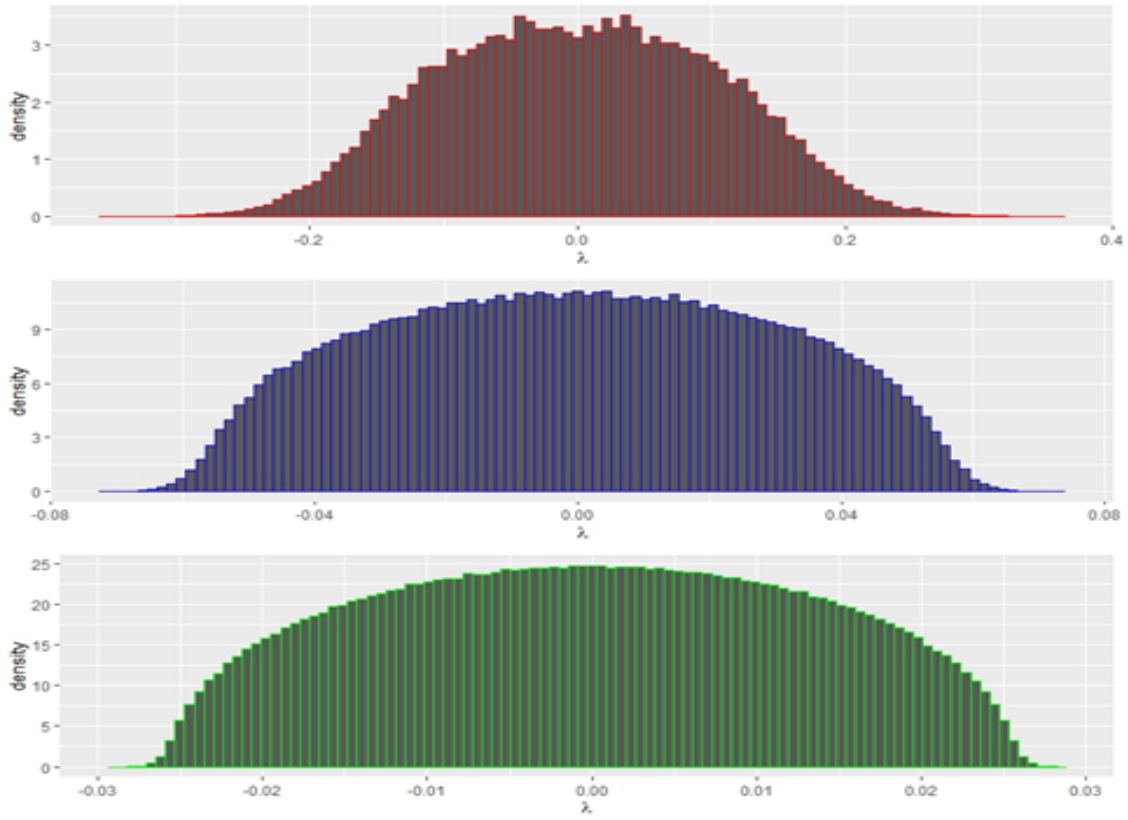

Рис. 4: Розподіл λ_2 матриці P за умови $A_{ij} \sim Unif\,(0,1)$ з розмірністю $N = 10,\ 100,\ 500$ (зверху вниз)

не менше за 0; по-друге, сума елементів в кожному рядку повинна бути рівною 1 з ймовірністю 1; по-третє, елементи стохастичної випадкової матриці повинні бути невід'ємними. Дані додаткові обмеження на елементи матриці змінюють розподіл власних значень стохастичної випадкової матриці P. Проте загальна асимптотична поведінка власних значень має ту ж поведінку, що і для загального випадку випадкових матриць – Рис. 5. Використовуючи отриману вище теорему, нам вдалося показати, що власні значення матриці P, які не несуть інформації про кластерність графу, визначаються співвідношенням

$$|\lambda_i(P;N)| \leq \frac{1}{\sqrt{N}}. \tag{19}$$

Використовуючи означення власних значень (19), будемо визначати опти-

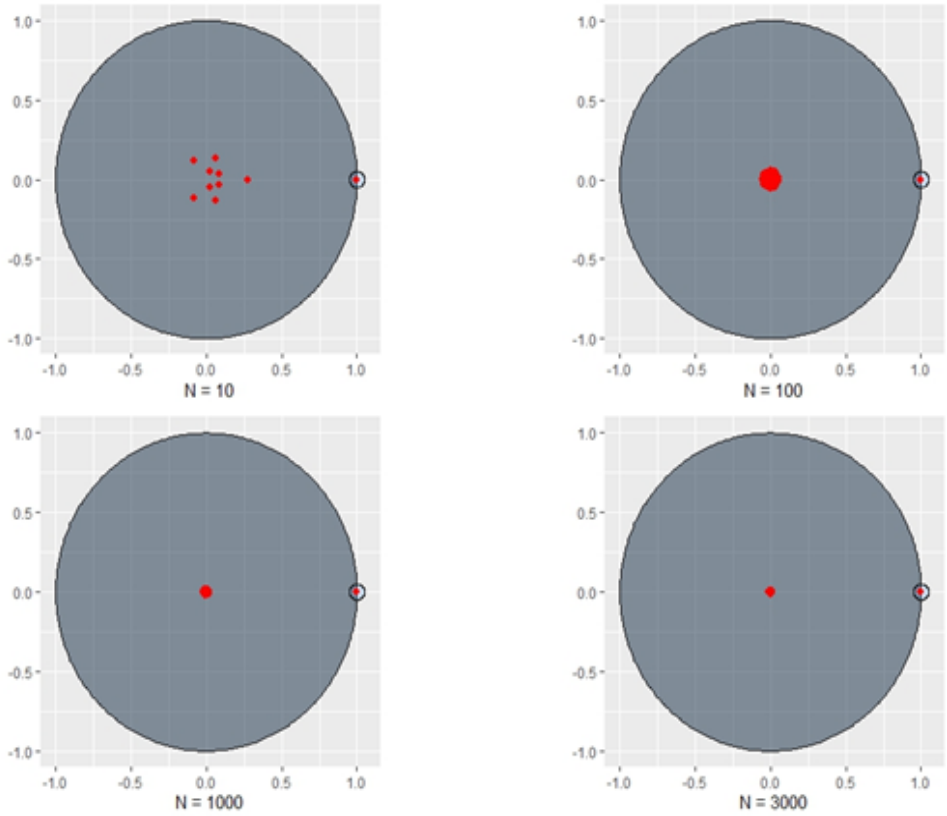

Рис. 5: Власні значення матриці P за умови $A_{ij} \sim Unif\,(0,1)$ з розмірністю $N = 10,\ 100,\ 1000,\ 3000$

мальну кількість кластерів в графі за допомогою співвідношення

$$k_{opt} = \#\left\{i = 1, ..., N : |\lambda_i(P) - 1| < 1 - \frac{1}{\sqrt{N}}\right\}. \tag{20}$$

3.3. Аналіз методом Монте-Карло та аналіз S&P500

3.3.1. Метод Монте-Карло

Перевіримо точність визначення кластерів за допомогою класичного методу Монте – Карло [77]. Для перевірки визначення точності кластеризації, розглянемо наступну процедуру утворення матриці суміжності для графа

$$A = D + Error,$$

де D – діагональна матриця

$$D = \begin{pmatrix} D_1 & 0 & 0 & \dots & 0 \\ 0 & D_2 & 0 & \dots & 0 \\ 0 & 0 & D_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & D_k \end{pmatrix}.$$

Матриці D_i є матрицями розмірності $n_i \times n_i$, що визначають кластери в моделі. Матриця $Error$ – матриця розмірності $N \times N$, що задає міжкластерні зв'язки. Зрозуміло, що при умові $Error = 0$, кількість власних значень, які рівні 1 з ймовірністю близькою до 1, близька до k. Розглянемо випадок, коли матриця помилок $Error$ відмінна від 0. Для моделювання матриць D та $Error$ будемо використовувати наступні параметри:

1. Кількість кластерів – k.
2. Рівномірний дискретний розподіл $Ud(n_a, n_b)$ для моделювання розмірності кластерів, а саме для моделювання n_i.
3. Рівномірний дискретний розподіл $Ud(d_a, d_b)$ для моделювання ненульових елементів матриці D.
4. Рівномірний дискретний розподіл $Ud(e_a, e_b)$ для моделювання ненульових елементів матриці D.

Таким чином, модель залежить від 7 гіперпараметрів

$$par = (k,\, n_a, n_b, d_a, d_b, e_a, e_b)\,.$$

Для порівняння отриманих у даній статті результатів (наш метод) використаємо Марковський алгоритм, описаний в роботі [78], та метод ліктя. Для моделювання у методі Монте – Карло будемо використовувати $N = 10^5$ ітерацій.

Використовуючи наведений вище алгоритм для кластеризації на графах, кількість кластерів будемо визначати наступним чином

$$k_{opt} = \# \left\{ i : |\lambda_i (P) - 1| < 1 - \frac{2}{\sqrt{N}} \right\}.$$

Результати оцінки кластерів за трьома алгоритмами відображені в таблиці 1.

Гіперпараметри $(k,\ n_a, n_b, d_a, d_b, e_a, e_b)$	Марковський алгоритм [78]		Метод ліктя	Наш метод
	$r = 2$ $s = 2$	$r = 5$ $s = 2$		
$(15, 10, 20, 4, 5, 0, 1)$	40	19	12	15
$(15, 10, 20, 10, 20, 2, 5)$	43	22	11	15
$(25, 30, 50, 10, 20, 2, 5)$	67	34	24	25
$(40, 4, 20, 10, 20, 2, 5)$	100	51	34	38
$(15, 10, 20, 10, 20, 2, 5)$	42	28	14	15
$(15, 10, 20, 10, 20, 8, 15)$	36	23	13	15
$(50, 10, 20, 10, 20, 8, 15)$	115	54	53	46
$(20, 10, 20, 10, 20, 8, 15)$	53	35	23	20
$(20, 10, 20, 10, 20, 1, 15)$	32	39	17	20

Таб. 1. Результати визначення оптимального числа кластерів k_{opt}

Як ми бачимо із таблиці 1, запропонований в роботі метод дозволяє найбільш точно визначати оптимальну кількість кластерів, базуючись на стохастичних матрицях графів. Даний факт не є дивним, оскільки дійсні частини власних значень матриці P розбиваються на три групи:

1. Власні значення, дійсні частини яких близькі до 0 та не рівні 1. Дані власні значення не є визначальними для кластеризиції, оскільки відображають лише однорідність елементів матриці. Однорідність еле-

ментів в свою чергу згідно Теореми 4 свідчить про наявність одного кластера.

2. Власні значення, дійсні частини яких "відмінні" від 0. Дані власні значення є визначальними для кластеризації, оскільки саме на їх кількості і ґрунтується визначення оптимальної кількості кластерів k_{opt}.

3. Власне значення $\lambda = 1$.

На Рис. 6-8 зображено групування власних значень в методі Монте – Карло для різних значень параметрів. Як можна бачити із даних рисунків, при помірній величині шуму власні значення із другої групи чітко відокремлені від власних значень із першої групи. Це й дозволяє точно визначати кількість кластерів в даних випадках. Крім того, власні значення із першої групи розподілені згідно з напівколовим розподілом Вінгера з параметром $\frac{1}{\sqrt{N}}$, що узгоджується із результатами теореми 4. Зауважимо, що матриця шумів $Error$ визначається величиною шуму, тобто дисперсією шумів. Використовуючи дану аналогію, аналогічно до теорії сигналів з гауссівськими шумами [122], можна визначити показник

$$SNR = \frac{\max\{\sigma_i^2\}}{\sigma^2},$$

де σ_i^2 – дисперсія елементів в матриці D_i, σ^2 – дисперсія елементів матриці шумів $Error$. Назва показника SNR (Signal to Noise Ratio) має аналогіну назву і при дослідженні сигналів.

У випадку сильного зашумлення даних (Рис. 8) власні значення із груп 1 та 2 об'єднуються. У цьому випадку практично неможливо визначити точне значення k_{opt}. Проте слід відмітити, що розподіл власних значень із групи 1 у даному випадку має статистично додатні непарні моменти. У випадках, коли вплив зашумлення є меншим, асиметрія дійсних частин

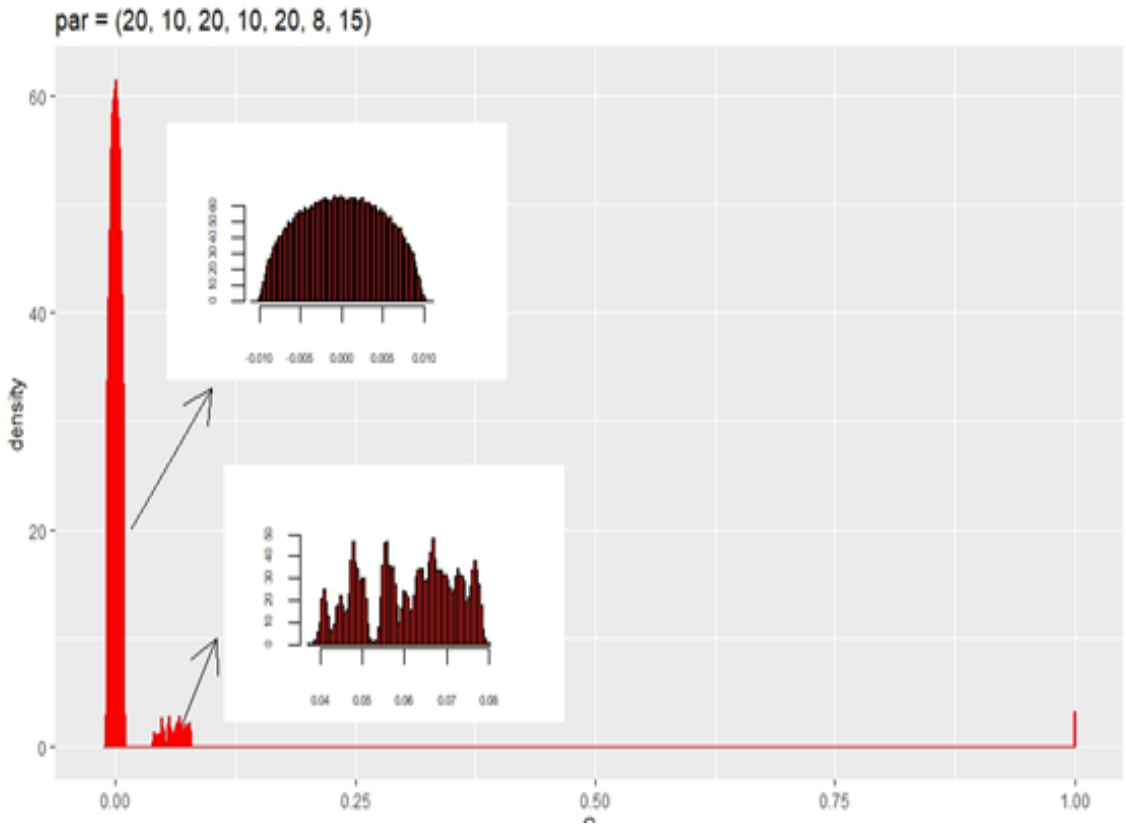

Рис. 6: Розподіл дійсних частин власних значень матриці P з параметрами $par = (20, 10, 20, 10, 8, 15)$

власних значень із групи 1 рівна 0. Даний факт може слугувати відправною точкою для подальшого аналізу значення k_{opt} у випадку сильної зашумленості даних.

3.3.2. Аналіз акцій S&P500

Розглянемо використання методики, описаної в даній роботі, для кластеризації акцій $N = 470$ компаній S&P500 даних, фіксованих в період 2013 – 2018 років. Для аналізу розглянуто $M = 1249$ денних показників вартості акцій при закритті бірж (close price). Аналогічно роботі [75], розглянемо акції із 10 секторів економіки, тобто реальна кількість кластерів рівна 10 для даного набору даних. Аналогічно більшості робіт з аналізу часових рядів вартості, використаємо відсоткові ставки акцій, що визнача-

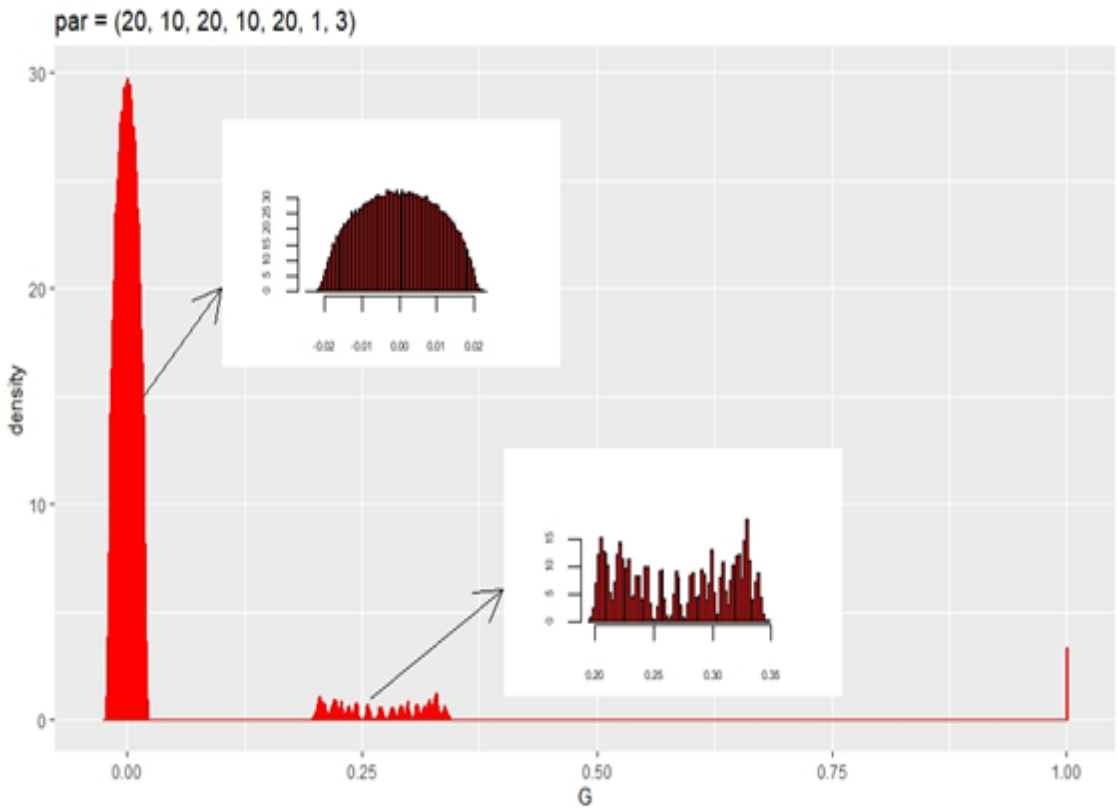

Рис. 7: Розподіл дійсних частин власних значень матриці P з параметрами $par = (20, 10, 20, 10, 1, 3)$

ються наступним чином

$$r_{i,t} = \log\left(\frac{S_{i,t}}{S_{i,t-1}}\right),$$

де $r_{i,t}$ – відсоткова ставка для i-ї акції в період t, $S_{i,t}$ – вартість i-ї акції в період t, $i = 1, ..., N$, $t = 2, ..., M$. При аналізі кількості кластерів на основі методу, запропонованого в даній роботі, було виявлено, що оптимальна кількість кластерів для даного набору даних рівна

$$k_{opt} = 5.$$

Таким чином, оптимальна кількість кластерів майже вдвічі менша ніж кількість галузей економіки. Аналогічний факт для $N = 400$ акцій S&P500 в період 2002 – 2004 років з $M = 44720$ відмічений також авторами роботи [75], де зазначено, що реальна кількість груп (кластерів) із різною поведінкою часових рядів приблизно вдвічі менша за відповідну кількість галузей

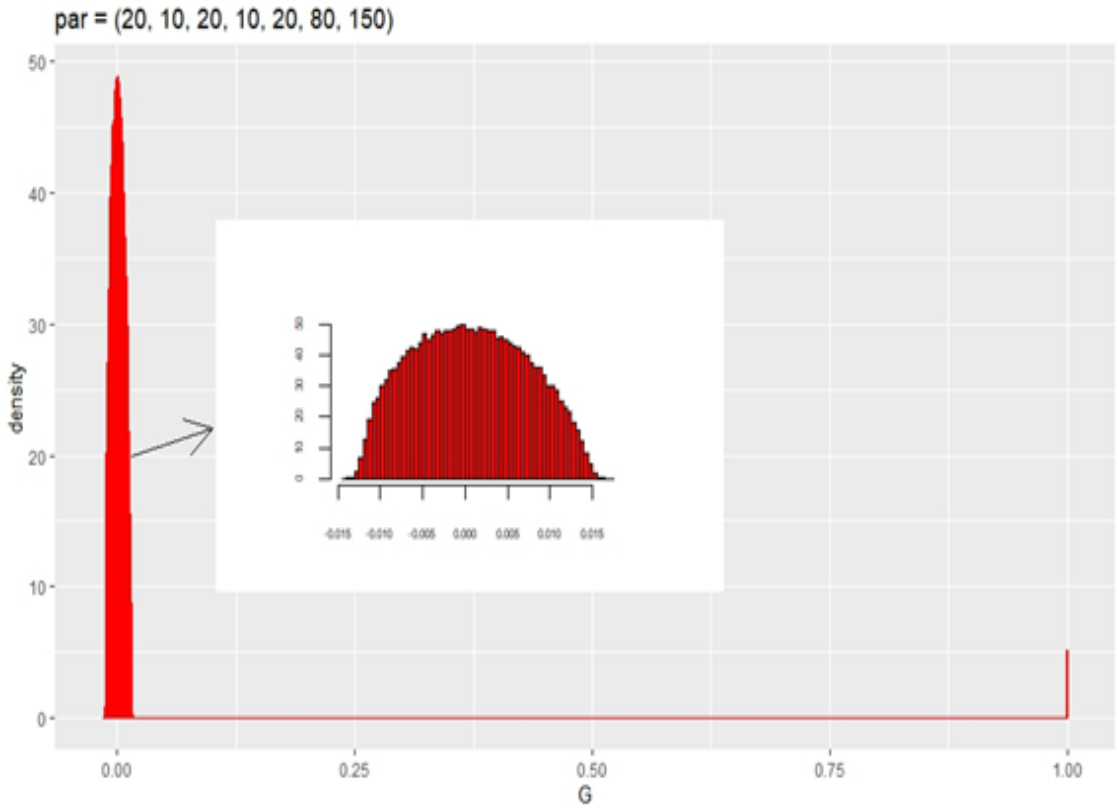

Рис. 8: Розподіл дійсних частин власних значень матриці P з параметрами $par = (20, 10, 20, 10, 80, 150)$

економіки. Проте в роботі [75] не вказано точного критерію визначення k_{opt}, оскільки проводиться спектральне дослідження графа на основі матриці суміжності A, що не дозволяє зробити висновків щодо кількості кластерів.

Значення $k_{opt} = 5$ означає, що деякі сектори економіки є дуже пов'язані і не можуть бути відслідковані за допомогою вартостей акцій компаній із даних галузей. Крім того, вплив "міжгалузевих" компаній[1] також призводить до зменшення числа кластерів k_{opt} для даного набору даних.

[1]Наприклад, компанія Apple Inc є міжгалузевою компанією, оскільки реалізує свої товари на ринках різних галузей.

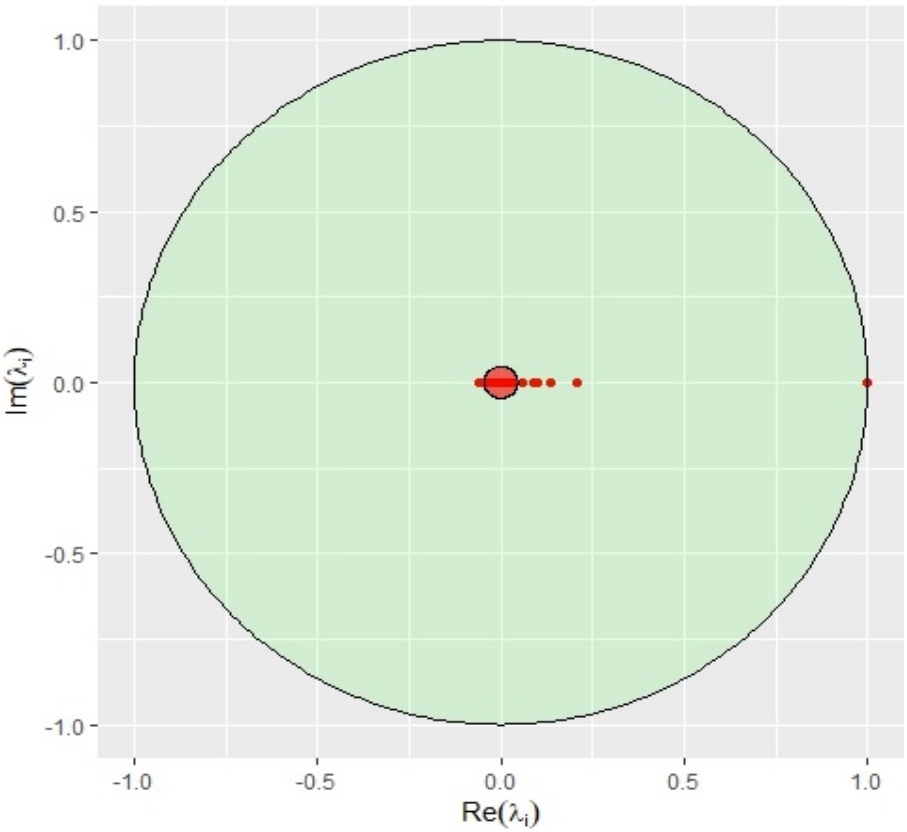

Рис. 9: Власні значення стохастичної матриці, побудованої для $N = 470$ акцій S&P500 в період 2013 – 2018 років з $M = 1249$ показниками.

Висновки до розділу III

У розділі III розглянуто проблему кластеризації на графах на основі власних значень стохастичної матриці графа. Доведено, що власні значення стохастичної матриці длявеликих графів ($N > 100$) поділяються на три групи, одна із яких є визначальною для чи-сла кластерів у графі. Використовуючи теорію випадкових матриць, вдалося показати, щоасимптотичний розподіл підгрупи дійсних частин власних значень стохастичної матриці графу описується напівколовим розподілом Вігнера. Використання стохастичних матриць дало змогу точно локалізувати власнізначення, що відповідають за кількість кластерів, чого не вдавалося зробити для матрицьсуміжності. Основні припущення моделі пов'язані з властивостями дискретних ланцюгів Маркова, що дозволяє розширити область застосу-

вання отриманих результатів на більшширший клас об'єктів. Теоретичні результати перевірені на кластеризації часових рядів,що описують вартості $N = 470$ акцій S&P500 в період 2013 –2018 рр. Результати кластеризації даних часових рядів показали наявність5чітко виражених груп, які узгоджується із використаними даними.

ОСНОВНІ РЕЗУЛЬТАТИ І ВИСНОВКИ

У дисертаційному дослідженні запропоновано нову метрику для знаходження міри подібності між в просторі стаціонарних часових рядів, представлених моделями $ARMA(p, q)$ Описана метрика грунтується на знаходженні відстані між параметрами моделей часових рядів, а не між самими вимірюваннями часового ряду. У роботі наведено порівняння запропонованої метрики з класичними моделями метрик в просторі часових рядів за допомогою симуляції даних методом Монте-Карло у середовищі R Programming. Показано, що отримана метрика є більш стійкою до викидів і дає більш точні результати для часових рядів з великою кількістю вимірювань. Встановлено, що обчислювальна складність алгоритму знаходження відстані з використанням запропонованої метрики для N часових рядів складає $O(T * N^2)$, в той же час аналогічна складність алгоритмів DTW, ERP становить $O(T^2N^2)$. За рахунок стійкості до викидів дана метрика дозволяє отримувати більш стійкі до шумів кластери.

У роботі також запропоновано новий метод визначення оптимальної кількості кластерів при розгляді задач кластеризації об'єктів, що задаються неструктурованими даними (графами, часовими рядами тощо) на основі спектрального аналізу стохастичної матриці даних. Показано, що дійсні частини власних значень стохастичної матриці графу можна розділити на три групи. До першої групи відноситься 1, так як вона завжди присутня серед власних значень. До другої групи відносяться власні значення стохастичної матриці, які є близькими до нуля, але не є нулями. До

третьої групи відносяться ті власні значення, які знаходяться між 0 та 1. Саме кількість власних значень у третій групі і відповідає оптимальній кількості кластерів для вхідних даних.

Використовуючи симуляцію методом Монте-Карло, показано, що запропонований метод підбору кількості кластерів дає кращі результати для визначення оптимальної кількості кластерів у порівнянні з рядом класичних методів (Марковський алгоритм з двома типами параметрів та метод ліктя). Симуляція Монте-Карло використана для утворення багатовимірних даних — графу з фіксованою кількістю сукупностей (кластерів). Таким чином для проведення порівняння запропонованого методу вибору оптимальної кількості кластерів з Марковським алгоритмом та методом ліктя оптимальна кількість кластерів була наперед відомою.

За допомогою методу Монте-Карло та пакету RStudio для мови R Programming згенерувано часові ряди порядку $ARMA(2, 1)$. У одному із часових рядів замінено % даних на викиди. Таким чином вдалося показати, що відносна похибка вимірювань для запропонованого методу зростає логарифмічно, у той час як відносна похибка для класичних методів – зростає лінійно. Крім того, перевірено роботу запропонованого алгоритму при розгляді довгих часових рядів, тобто при збільшенні кількості вимірювань, та проведено порівняння результатів роботи з класичними підходами. У дослідженні показано, що відносна похибка вимірювань для запропонованого методу спадає з ростом кількості вимірювань часового ряду, у той же час відносна похибка вимірювань для класичних методів не змінюється з ростом кількості вимірювань у часовому ряді. Встановлено, що розроблений алгоритм знаходження оптимальної кількості кластерів є менш чутливим до наявності кластерів різного розміру.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Oliver Vetter, Timo Sturm, Mariska Fecho, and Peter Buxmann. Machine learning developments as stimuli for organizational learning. 12 2023.

[2] Paul Banfield, Rebecca Kay, and Dean Royles. *Learning, Training, and Development*. 08 2023.

[3] Claire Gubbins, Thomas Garavan, and Elisabeth Bennett. *Digital Learning: A Bright New Dawn for Learning and Development*, pages 127–149. 07 2023.

[4] Md Ashraful Alam, Hammed Esa, Kazi Nazrul Islam, Anwar Hossain, Graduate Assistant, and Graduate Student. Data clustering: Prospects challenges. 11:2320–2882, 09 2023.

[5] O.M. Fedoseienko. Challenges and design aspects of microgrid clustering. *Праці Інституту електродинаміки Національної академії наук України*, 2023:68–73, 08 2023.

[6] Yue Liu, Jun Xia, Sihang Zhou, Siwei Wang, Xifeng Guo, Xihong Yang, Ke Liang, Wenxuan Tu, Stan Li, and Xinwang Liu. A survey of deep graph clustering: Taxonomy, challenge, and application, 11 2022.

[7] T.V. Knignitska. Estimate of time series similarity based on models. *Automation and Information Sciences*, 51 (8), 2019.

[8] N. Pavlyukovich, O.V. Pavlyukovich, O.V. Dubolazov, Yu.A. Ushenko, Yu. Ya. Tomka, N.I. Zabolotna, I.V. Soltys, Ya.M. Drin, T.V. Knigni-

tska, M.V. Talakh, A.Ya. Dovgun, A. Kotyra, and A. Kozbakova. Methods and means of "single-point" phasometry of microscopic images of optical-anisotropic biological objects. *Proceedings of SPIE - The International Society for Optical Engineering*, 11176, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments, 2019.

[9] Т.В. Кнігніцька, І.В. Малик, and М.Ю. Горбатенко. Кластеризація: марковський алгоритм. *Буковинський математичний журнал*, pages 59–75, 2020.

[10] I. Doroshenko, T. Knihnitska, and T. Deretorska. Comparison of machine learning algorithms for predicting mortality from covid-19 virus. *SWorld Journal*, 2(11-02):72–77, 2022.

[11] М.А. Іванчук, І.В. Малик, Т.В. Кнігніцька, and Т.О. Лукашів. Статистичний аналіз відносних величин у медицині. *Клінічна та експериментальна патологія*, 18, 4(70):109–114, 2019.

[12] І.В. Малик and Кнігніцька Т.В. Методи машинного навчання для статистичної обробки медичних даних. *Науковий вісник Чернівецького національного університету. Серія: Комп'ютерні системи та компоненти.*, 8(2):77–85, 2017.

[13] T. Knignitska. "from the practice to theory" or how to interest the students by mathematics. *Physical and Mathematical Education*, 4(14):199–204, 2017.

[14] Muhammet Munir, Dadi ¹a, Özer Yıldız, Futbol Hakemlerinin, Var Sistemi, Var Eğitimi, and Hakkındaki Görüşleri. Opinions of football referees on the var system and var training. 10 2022.

[15] Lei Chen and Raymond Ng. On the marriage of lp-norms and edit distance. *Proceedings of the Thirtieth International Conference on Very Large Data Bases*, 10 2004.

[16] Rohaifa Khaldi, Abdellatif El Afia, Raddouane Chiheb, and Siham Tabik. What is the best rnn-cell structure to forecast each time series behavior? *Expert Systems with Applications*, 215:119140, 11 2022.

[17] Jianya Lu, Yingjun Mo, Zhijie Xiao, Lihu Xu, and Qiuran Yao. Distribution estimation and change-point detection for time series via dnn-based gans, 11 2022.

[18] Roy De Maesschalck, D. Jouan-Rimbaud, and D.L. Massart. The mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50:1–18, 01 2000.

[19] Félix Iglesias Vázquez and Wolfgang Kastner. Analysis of similarity measures in times series clustering for the discovery of building energy patterns. *Energies*, 6:579–597, 02 2013.

[20] Y.K. Tor. L1, l2, kalman filter and time series analysis in deformation analysis [c]. *Proceeding of the FIG XXII International Congress*, pages 1–18, 01 2002.

[21] Donald Berndt and James Clifford. Using dynamic time warping to find patterns in time series. volume 10/16, pages 359–370, 01 1994.

[22] Lei Chen, M. Tamer Özsu, and Vincent Oria. Robust and fast similarity search for moving object trajectories. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 491–502, 06 2005.

[23] Alan Lipkus. A proof of the triangle inequality for the tanimoto distance. *Journal of Mathematical Chemistry*, 26:263–265, 10 1999.

[24] Usue Mori, Alexander Mendiburu, and Jose Lozano. Distance measures for time series in r: The tsdist package. *The R Journal*, 8, 08 2016.

[25] Alexander Aue and Lajos Horvath. Structural breaks in time series. *Journal of Time Series Analysis*, 34, 01 2013.

[26] Alessandro Casini and Pierre Perron. Structural breaks in time series. 05 2018.

[27] Hirotugu Akaike. Time series analysis and control through parametric models in applied time series analysis. 12 1978.

[28] Joseph Cavanaugh and Andrew Neath. The akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *Wiley Interdisciplinary Reviews: Computational Statistics*, 11, 03 2019.

[29] Anayochukwu Anyasodor, Ezekiel Nwose, Phillip Bwititi, and Ross Richards. Cost-effectiveness of community diabetes screening: Application of akaike information criterion in rural communities of nigeria. *Frontiers in Public Health*, 10, 07 2022.

[30] S.S. Shaphiro and M.B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52:591–611, 01 1965.

[31] Samuel Shapiro and R. Francia. An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, 67:215–216, 03 1972.

[32] L. Rodgers and Alan Nicewander. Thirteen ways to look at the correlation coefficient. *Am. Stat.*, 42:59–66, 01 1988.

[33] Michalis Vlachos, George Kollios, and Dimitrios Gunopulos. Discovering similar multidimensional trajectories. pages 673–684, 02 2002.

[34] J.B. Breymann, E.T. Hollis, and J.T. Lynch. A continuous countercurrent resin-in-pulp process. 07 1958.

[35] Frank Smith. Partial hydrogenation of linseed oil by a continuous process. *Journal of The American Oil Chemists Society - J AMER OIL CHEM SOC*, 25:328–334, 01 1948.

[36] Will Gersch. Estimation of the autoregressive parameters of a mixed autoregressive moving-average time series. *Automatic Control, IEEE Transactions on*, 15:583 – 588, 11 1970.

[37] Geoffrey Pegram. A continuous streamflow model. *Journal of Hydrology - J HYDROL*, 47:65–89, 05 1980.

[38] Donald Gaver. Imbedded markov chain analysis of a waiting-line process in continuous time. *Annals of Mathematical Statistics*, 30, 09 1959.

[39] James Cadzow. High performance spectral estimation–a new arma method. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, ASSP-28:524 – 529, 11 1980.

[40] Ørnulf Borgan. On the theory of moving average graduation. *Scandinavian Actuarial Journal*, 1979:83–105, 07 1979.

[41] Turkan Gardenier. Moving averages for environmental standards. *Transactions of The Society for Modeling and Simulation International - SIMULATION*, 39:49–58, 08 1982.

[42] Robert Haining. The moving average model for spatial interaction. *Transactions of the Institute of British Geographers*, 3, 01 1978.

[43] Nola Windirah and Ridha Novanda. Price volatility analysis on indonesian palm oil commodities by model arch/garch. *Jurnal Sosial Ekonomi Pertanian*, 19:101–114, 06 2023.

[44] Lidija Madzar, Dušica Karić, and Borjana Mirjanić. Modelling the volatility of the global gold price by applying the arch/garch models. *Ekonomika*, 69:23–34, 07 2023.

[45] Manfred Deistler and Wolfgang Scherrer. *ARCH and GARCH Models*, pages 191–198. 10 2022.

[46] Wahidah Alwi and Ilham Syata. Forcasting stock price pt. indonesian telecomunication with arch-garch model. *Jurnal Varian*, 5:125–136, 04 2022.

[47] J. FRANKE. A levinson-durbin recursion for autoregressive-moving average processes. *Biometrika*, 72, 12 1985.

[48] Marc Nerlove and Francis Diebold. *Autoregressive and Moving-average Time-series Processes*, pages 25–35. 01 1990.

[49] Peter Brockwell and Rob Hyndman. On continuous-time threshold autoregression. *International Journal of Forecasting*, 8:157–173, 02 1992.

[50] Peter Brockwell. On continuous-time threshold arma processes. *Journal of Statistical Planning and Inference*, 39:291–303, 04 1994.

[51] Uffe Thygesen. *Brownian Motion*, pages 63–88. 04 2023.

[52] Aritra Ghosh, Malay Bandyopadhyay, Sushanta Dattagupta, and Shamik Gupta. Quantum brownian motion: A review, 06 2023.

[53] Purba Das, Rafał Łochowski, Toyomu Matsuda, and Nicolas Perkowski. Level crossings of fractional brownian motion, 08 2023.

[54] Marly Silva and Wellington Mazer. Diffusion coefficient and tortuosity: Brownian motion. *CONTRIBUCIONES A LAS CIENCIAS SOCIALES*, 16:18281–18302, 09 2023.

[55] Haitao Li, Guo Yu, Yizhu Fang, Yanru Chen, Chongyang Wang, and Dongming Zhang. Gas production prediction and risk quantification of shale gas reservoirs in sichuan basin based on gauss prediction model and monte carlo probability method. *Frontiers in Earth Science*, 10:977200, 08 2022.

[56] Peter Brockwell. Continuous-time arma processes. *Handbook of Statistics*, 19:249–276, 12 2001.

[57] Osnat Stramer, R. Tweedie, and Peter Brockwell. Existence and stability of continuous time threshold arma processes. *Statistica Sinica*, 6, 11 1995.

[58] Mike West, P. Harrison, and Andy Pole. *Applied Bayesian Forecasting and Time Series Analysis*. 05 2018.

[59] Ilker Yildirim. Bayesian inference: Gibbs sampling. *Technical Note, Department of Brain and Cognitive Sciences, University of Rochester*, pages 1–6, 08 2012.

[60] Peter J. Brockwell and Richard A. Davis. *Introduction to Time Series and Forecasting*. 01 2016.

[61] Josef Dick, Frances Kuo, Gareth Peters, and Ian Sloan. *Monte Carlo and Quasi-Monte Carlo Methods 2012*, volume 65. 01 2013.

[62] Raffaele Argiento, Andrea Cremaschi, and Guglielmi Alessandra. A bayesian nonparametric mixture model for cluster analysis. *Technical report Quaderno Imati CNR, 2012 3-MI, Milano*, 2013.

[63] Richard Davis and William Dunsmuir. Maximum likelihood estimation for ma(1) processes with a root on or near the unit circle. *Econometric Theory*, 12:1–29, 03 1996.

[64] Emily Fox, Erik Sudderth, Michael Jordan, and Alan Willsky. Bayesian nonparametric inference of switching dynamic linear models. *IEEE Transactions on Signal Processing*, 59:1569–1585, 04 2011.

[65] Anurag Ghosh, Soumalya Mukhopadhyay, Sandipan Roy, and Sourabh Bhattacharya. Bayesian inference in nonparametric dynamic state-space models. *Statistical Methodology*, 21, 08 2011.

[66] Gary Grunwald, Kais Hamza, and Rob Hyndman. Some properties and generalizations of nonnegative bayesian time series models. *Journal of The Royal Statistical Society Series B-statistical Methodology - J ROY STAT SOC SER B-STAT MET*, 59:615–626, 08 1997.

[67] Antonio Lijoi, Ramsés Mena, and Igor Prünster. Controlling the reinforcement in bayesian non-parametric mixture models. *Journal of the Royal Statistical Society Series B*, 69:715–740, 09 2007.

[68] Chuan Zhou and Jon Wakefield. A bayesian mixture model for partitioning gene expression data. *Biometrics*, 62:515–25, 07 2006.

[69] Jay Magidson and Jeroen Vermunt. Latent class factor and cluster models, bi-plots, and related graphical displays. *Sociological Methodology*, 31:223 – 264, 01 2001.

[70] Anna Huang. Similarity measures for text document clustering. *Proceedings of the 6th New Zealand Computer Science Research Student Conference*, 01 2008.

[71] Luis Nieto-Barajas and Alberto Contreras. A bayesian nonparametric approach for time series clustering. *Bayesian Analysis*, 9:147–170, 03 2014.

[72] Z. Bai, Zhaoben Fang, and Ying-Chang Liang. *Spectral Theory of Large Dimensional Random Matrices and Its Applications to Wireless Communications and Finance Statistics: Random Matrix Theory and its Applications*. 01 2014.

[73] E.A. Bender and S.G. Williamson. *Lists, Decisions and Graphs – With an Introduction to Probability*. University of California at San Diego, 2010.

[74] T. Liao. Clustering time series data — a survey. *Pattern Recognition*, 38:1857–1874, 11 2005.

[75] Christoly Biely and Stefan Thurner. Random matrix ensembles of time-lagged correlation matrices: Derivation of eigenvalue spectra and analysis of financial time-series. *Quantitative Finance*, 8:705–722, 10 2008.

[76] Bolch Gunter, Greiner Stefan, Meer Hermann, and Trivedi Kishor. *Queueing Networks and Markov Chains*. John Wiley and Sons, 2006.

[77] Dan Crisan, Pierre Del Moral, and Terry Lyons. Discrete filtering using branching and interacting particle systems. *Markov Process. Related Fields*, 5, 03 1999.

[78] Dongen S. *Graph clustering by flow simulation*. PhD thesis, Centre for Mathematics and Computer Science (CWI) in Amsterdam, 2000.

[79] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[80] Robert Qiu and Paul Antonik. *Smart Grid using Big Data Analytics: A Random Matrix Theory Approach*. 02 2017.

[81] V.A. Marchenko and Leonid Pastur. Distribution of eigenvalues for some sets of random matrices. *Math USSR Sb*, 1:457–483, 01 1967.

[82] Juan Pena and Tomas Sauer. Svd update methods for large matrices and applications, 09 2018.

[83] Ewa Gudowska-Nowak, Romuald Janik, Jerzy Jurkiewicz, Maciej Nowak, and Waldemar Wieczorek. Random walkers versus random crowds: Diffusion of large matrices. *Chemical Physics*, 375:380–385, 10 2010.

[84] Versaci Mario, Alessandra Jannelli, Giovanni Angiulli, and Francesco Morabito. Krylov's subspaces iterative methods to evaluate electrostatic parameters. *American Journal of Applied Sciences*, 11:396–405, 03 2014.

[85] Alexander Malyshev and M. Sadkane. On the stability of large matrices. *Journal of Computational and Applied Mathematics - J COMPUT APPL MATH*, 102:303–313, 02 1999.

[86] Achiya Dax. Computing the smallest singular triplets of a large matrix. *Results in Applied Mathematics*, 3:100006, 05 2019.

[87] Pier Poier, Louis Lagardère, and Jean-Philip Piquemal. Smooth particle mesh ewald-integrated stochastic lanczos many-body dispersion algorithm, 07 2023.

[88] Nobuaki Obata. *Spectral Analysis of Growing Graphs*, volume 20. 01 2017.

[89] Ahmed Zaiou. *Quantum machine learning approaches for graphs and sequences : application to nuclear safety assessment.* PhD thesis, 12 2022.

[90] Prasanna Date, Catherine Schuman, Robert Patton, and Thomas Potok. A classical-quantum hybrid approach for unsupervised probabilistic machine learning. pages 98–117, 01 2020.

[91] Yue Meng, Hongli Zhang, and Wenhui Fan. Analysis of the network structure characteristics of virtual power plants based on a complex network. *Electric Power Systems Research*, 204:107717, 03 2022.

[92] Ramadan Kuridan. *Perturbation Theory*, pages 125–134. 09 2023.

[93] Mohamed Boussiala. Implementing the dickey-fuller test in r simulation, 07 2023.

[94] Walaa Hussein, Kamil Audah, Nor Noordin, Habib Kraiem, Aymen Flah, Mohd Fadlee, and Alyani Ismail. Least square estimation-based different fast fading channel models in mimo-ofdm systems. *International Transactions on Electrical Energy Systems*, 2023:1–23, 08 2023.

[95] Vadim Romanuke. Arima model optimal selection for time series forecasting. *Maritime Technical Journal*, 224:28–40, 03 2022.

[96] Zichao Xu, Hongying Zheng, and Jianyong Chen. *Opemod: An Optimal Performance Selection Model for Prediction of Non-stationary Financial Time Series*, pages 304–315. 09 2022.

[97] Xuecai Yin. The optimal selection of ideological and political issues in business courses based on swarm intelligence algorithm. *Scalable Computing: Practice and Experience*, 24:409–417, 09 2023.

[98] Ye Tsarkov, Volodymyr Yasynskyy, and Igor Malyk. Stability in impulsive systems with markov perturbations in averaging scheme. 2. averaging principle for impulsive markov systems and stability analysis based on averaged equations. *Cybernetics and Systems Analysis - CYBERN SYST ANAL-ENGL TR*, 47:44–54, 01 2011.

[99] Ziyi Liu, Rakshitha Godahewa, Kasun Bandara, and Christoph Bergmeir. *Handling Concept Drift in Global Time Series Forecasting*, pages 163–189. 09 2023.

[100] Ahmed Shifaz, Charlotte Pelletier, François Petitjean, and Geoffrey Webb. Elastic similarity and distance measures for multivariate time series. *Knowledge and Information Systems*, 65:1–34, 02 2023.

[101] Ahmed Shifaz, Charlotte Pelletier, François Petitjean, and Geoffrey Webb. Elastic similarity measures for multivariate time series classification, 02 2021.

[102] Izaskun Oregi, Aritz Pérez, Javier Del Ser, and Jose Lozano. On-line elastic similarity measures for time series. *Pattern Recognition*, 88, 12 2018.

[103] Weiqiang He, Lei Ma, Ziyun Yan, and Heng Lu. Evaluation of advanced time series similarity measures for object-based cropland mapping. *International Journal of Remote Sensing*, 44:3777–3800, 07 2023.

[104] Zhen Wang, Jin Duan, Haobo Xu, Xue Song, and Yang Yang. Enhanced pelican optimization algorithm for cluster head selection in heterogeneous wireless sensor networks. *Sensors*, 23:7711, 09 2023.

[105] Kuntal Chowdhury, D. Chaudhuri, and Arup Kumar Pal. Seed selection algorithm through k-means on optimal number of clusters. *Multimedia Tools and Applications*, 78, 07 2019.

[106] Min Li, Huiping Gu, and Qing Li. Optimal number of cluster heads for selection cooperation in clustering wireless sensor networks. *Journal of Physics: Conference Series*, 1754:012220, 02 2021.

[107] Tung-Jung Chan, Ching-Mu Chen, Yung-Fa Huang, Jen-Yung Lin, and Tair-Rong Chen. Optimal cluster number selection in ad-hoc wireless sensor networks. *WSEAS TRANSACTIONS on COMMUNICATIONS*, 7:837–846, 08 2008.

[108] Rachit Nimavat. Graph theory and its uses in graph algorithms and beyond, 08 2023.

[109] Maya Bechler-Speicher, Ido Amos, Ran Gilad-Bachrach, and Amir Globerson. Graph neural networks use graphs when they shouldn't, 09 2023.

[110] Genetic Association, Clinical Association, Clinical Association, Xin Chen, Zhuo Li, Desheng Liang, and Lingqian Wu. Expert consensus on the detection of genome-wide copy number variations in abortive tissues and family reproductive consultation. *Zhonghua yi xue yi chuan xue za zhi = Zhonghua yixue yichuanxue zazhi = Chinese journal of medical genetics*, 40:129–134, 02 2023.

[111] Committee Emergency, Diseases Respiratory, Committee Disease, Chapter Disease, Association Medical, Association Doctor, Association Medical, Medicine Chinese, and Medicine Care. Expert consensus on traditional chinese medicine health management in adults with sars-cov-2 variant infection at home. *Zhonghua wei zhong bing ji jiu yi xue*, 34:1233–1237, 12 2022.

[112] Prenatal Association, Prenatal Association, and Juntao Liu. Guidelines for the application of chromosomal microarray analysis in prenatal diagnosis (2023). 40:1051–1061, 09 2023.

[113] Shi Congming, Bingtao Wei, Shoulin Wei, Wen Wang, Hai Liu, and Jialei Liu. A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP Journal on Wireless Communications and Networking*, 2021, 02 2021.

[114] Rido Trimanto, Eryka Yustari, Zulfatin Nafisah, Nona Carolina, Nursyiva Irsalinda, and Arifah Setyorini. Indonesian provincial clustering using elbow method for the national food security during pandemic. *Bulletin of Applied Mathematics and Mathematics Education*, 2:51–58, 12 2022.

[115] Congming Shi, Bingtao Wei, Shoulin Wei, Wen Wang, Hai Liu, and Jialei Liu. A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm, 08 2020.

[116] Alessandro Bosisio, Alberto Berizzi, Andrea Morotti, Bartolomeo Greco, Gaetano Iannarelli Phd, Cristina Moscatiello, Chiara Boccaletti, and Holguer Noriega. Performance assessment of load profiles clustering methods based on silhouette analysis. pages 1–6, 09 2021.

[117] Meshal Shutaywi and Nezamoddin Kachouie. Silhouette analysis for performance evaluation in machine learning with applications to clustering. *Entropy*, 23:759, 06 2021.

[118] Fei Wang, Hector-Hugo Franco-Penya, John Kelleher, John Pugh, and Robert Ross. An analysis of the application of simplified silhouette to the evaluation of k-means clustering validity. 07 2017.

[119] Valerie Robert and Yann Vasseur. Comparing high-dimensional partitions with the co-clustering adjusted rand index. *Journal of Classification*, 38, 05 2017.

[120] Ka Yee Yeung and Walter Ruzzo. Details of the adjusted rand index and clustering algorithms supplement to the paper "an empirical study on principal component analysis for clustering gene expression data"(to appear in bioinformatics). *Science*, 17, 01 2001.

[121] José Chacón. A close-up comparison of the misclassification error distance and the adjusted rand index for external clustering evaluation. *British Journal of Mathematical and Statistical Psychology*, 74, 10 2020.

[122] Ghadamali Bagherikaram, Abolfazl Motahari, and Amir Khandani. The secrecy capacity region of the gaussian mimo broadcast channel. *Information Theory, IEEE Transactions on*, 59:2673–2682, 05 2013.

ДОДАТОК

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Наукові праці у періодичних наукових виданнях, проіндексованих у наукометричній базі даних Scopus:

1. Knignitskaya T. V. Estimate of time series similarity based on models. Journal of Automation and Information Sciences. 2019. Vol. 51 (№8).

2. Pavlyukovich N., Pavlyukovich O.V., Dubolazov O.V., Ushenko Yu.A., Tomka Yu. Ya., Zabolotna N.I., Soltys I.V., Drin Ya.M., Knignitska T.V., Talakh M.V., Dovgun A.Ya., Kotyra A., and Kozbakova A. Methods and means of "single-point"phasometry of microscopic images of optical-anisotropic biological objects. Proceedings of SPIE – The International Society for Optical Engineering. Vol. 11176, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physic Experiments 2019, 1117630 (6 November 2019).

Наукові праці у виданнях, внесених до переліку наукових фахових видань України:

3. Кнігніцька Т.В., Малик І.В., Горбатенко М.Ю. Кластеризація: марковський алгоритм Буковинський математичний журнал. 2020. 7(2). С. 59-75.

Наукові праці, які додатково відображають наукові результати дисертації:

4. Doroshenko I., Knihnitska T., Deretorska T. Comparison of machine learning algorithms for predicting mortality from COVID-19 virus. SWorld Journal. 2022. 2(11-02). С. 72–77.

5. Іванчук М.А., Малик І.В., Кнігніцька Т.В., Лукашів Т.О. Статистичний аналіз відносних величин у медицині // Клінічна та експериментальна

патологія. – Чернівці, 2019. – Том 18, 4(70), 109 – 114.

6. Малик І.В., Кнігніцька Т.В. Методи машинного навчання для статистичної обробки медичних даних. Науковий вісник Чернівецького національного університету. Серія: Комп'ютерні системи та компоненти. 2017. Том 8, випуск 2. – С. 77 – 85.

7. Knignitska T. «From The Practice To Theory» Or How To Interest The Students By Mathematics. Physical and Mathematical Education ISSN 2413-1571 (print), ISSN 2413-158X (online): scientific journal. 2017. Issue 4(14). – P. 199-204.

Наукові праці, які засвідчують апробацію матеріалів дисертації:

1. Кнігніцька Т.В. Підбір оптимальних параметрів для однієї задачі кластеризації. Міжвузівський науковий семінар "Прикладні задачі та ІТ-технології", присвячений 100-річчю з дня народження професора В.П. Рубаника (1917-1993) і 55-річчю кафедри прикладної математики та інформаційних технологій: матеріали семінару, 9 – 10 червня 2017 р. Чернівці: 2017. С. 64 – 65.

2. Knignitska T. Cluster analysis in data mining. Scientific Conference of Doctoral Students C̈ontemporary trends in the development of science: visions of young researchers̈: Materials of the Scientific Conference of Doctoral Students, 6th Edition, June 15, 2017. Volume 1. Universities of the Academy of Sciences of Moldova. P. 30 – 35.

3. Кнігніцька Т.В., Малик І.В. Вибір оптимальної моделі для аналізу часових рядів. Праці VI-ї Міжнародної науково-практичної конференції «Проблеми інформатики та комп'ютерної техніки» (ПІКТ – 2017) (Чернівці, 5-8 жовтня 2017 року). Чернівці: Видавничий дім «Родовід», 2017. С. 32-33.

4. Knignitska T., Malyk I.V. Method for Evaluating Time Series Similarity. Сучасні проблеми математики та її застосування в природничих науках

і інформаційних технологіях: матеріали міжнародної наукової конференції, присвяченої 50-річчю факультету математики та інформатики Чернівецького національного університету імені Юрія Федьковича, 17 – 18 вересня 2018 р. Чернівці: 2018. С. 128.

5. Кнігніцька Т.В., Малик І.В., Лукашів Т.О. Алгоритми знаходження відстаней між часовими рядами. Праці VII-ї Міжнародної науково-практичної конференції «Проблеми інформатики та комп'ютерної техніки» (ПІКТ – 2018) (Чернівці, 11-14 жовтня 2018 року). Чернівці: Видавничий дім «Родовід», 2018. С. 27-29.

6. Кнігніцька Т.В., Малик І.В., Лукашів Т.О. Порівняння алгоритмів знаходження відстаней між часовими рядами. Праці VIII-ї Міжнародної науково-практичної конференції «Проблеми інформатики та комп'ютерної техніки» (ПІКТ – 2019) (Чернівці, 3-6 жовтня 2019 року). Чернівці: Видавничий дім «Родовід», 2019. С. 30-31.

7. Kyrychenko O.L., Knignitska T.V., Ostapov S.E. Stochastic models in artificial intelligence development. International conference "Modern stochastics: theory and applications V". Kyiv, June 1–4, 2021. P. 35.

8. Кнігніцька Т.В., Малик І.В. Оптимальна комбінація прогнозів для ієрархічних часових рядів. Міжнародна наукова конференція "Диференціально-функціональні рівняння та їх застосування" присвячена 80-річчю від дня народження професора В.І. Фодчука (1936–1992): матеріали конференції, 28 – 30 вересня, 2016. Чернівці: 2016. – С. 56.