КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА
Міністерство освіти і науки України

Київський національний університет імені Тараса Шевченка

Кваліфікаційна наукова
праця на правах рукопису

Круковець Дмитро Юрійович

УДК 004.8

ДИСЕРТАЦІЯ

Кластеризація компонент інфляції та прогнозування методами машинного навчання

122 Комп’ютерні науки

Подається на здобуття наукового ступеня доктора філософії

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

Д.Ю. Круковець

Науковий керівник
Нікітченко Микола Степанович
доктор фізико-математичних наук, професор

Київ - 2024

АНОТАЦІЯ

Круковець Д.Ю. Кластеризація компонент інфляції та прогнозування методами машинного навчання. Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії в галузі знань інформаційні технології за спеціальністю 12 «Інформаційні технології». Київський національний університет імені Тараса Шевченка, Київ, 2024.

У **вступі** розкрито сутність і стан розробки наукової проблематики, обґрунтовано вибір і актуальність обраної теми розробки багатоступеневого алгоритму на основі алгоритмів кластеризації та нейронних мереж з метою прогнозування часових рядів, зокрема тих що мають властивість дезагрегації на підкомпоненти. Визначено мету, об’єкт, предмет, методи дослідження, розкрито наукову новизну дослідження, теоретичне і практичне значення наукових результатів, особистий внесок здобувача, зазначено інформацію про впровадження і апробацію результатів.

Прогнозування процесів та часових рядів є необхідним, бо це дозволяє створювати стратегії з середньо- та довгостроковим планування, що враховують низку факторів та ризиків, діяти проактивно. Однак створення якісних прогностичних моделей є нетривіальною задачею.

Застосування нейронних мереж у економічному аналізі та прогнозуванні є актуальним та перспективним напрямом. В дисертаційній роботі основна увага приділяється розробці комбінованої багатоступеневої нейромережевої моделі, поєднуючи її з алгоритмами кластеризації дезагрегованого набору даних для одночасного виявлення лінійних та сезонних особливостей у рядах та їх взаємозалежної структури.

Розробка базується на дослідженнях зарубіжних вчених. Зокрема мова про роботу з дезагрегованими наборами даних, запропонованою Huwiler та Kaufmann у 2013, а також про побудову схожих моделей в роботі Almosova

та Andersen у 2023. Також використовувалось експериментальне тестування різноманітних алгоритмів для емпіричного обґрунтування їх ефективності.

Метою роботи є підвищення ефективності алгоритмів прогнозування наборів даних з властивістю дезагрегації та нелінійними взаємозв'язками. Оцінкою досягнення є емпіричне порівняння розроблених алгоритмів з іншими методами машинного навчання та нейромережевими підходами.

Для досягнення мети дослідження в рамках дисертації були вирішені наступні завдання:

- Розроблено методику побудови моделі нейромережевої архітектури та відповідних алгоритмів системи машинного навчання для прогнозування часових рядів з урахуванням нелінійностей між ними
- Побудовано адаптовану до інфляційних компонент модель кластеризації часових рядів за дистанціями між ними
- Використано методологію емпіричної оцінки ефективності моделі для порівняння з моделями, включаючи моделі з літератури

Об’єктом дослідження є методи прогнозування, їх характеристики для наборів даних з особливостями, притаманними рядам з економічної сфери. Предметом дослідження є концепції, методи, алгоритми та архітектури нейромереж для ефективного прогнозування часових рядів, а також алгоритми для пошуку дистанцій та розподілу на групи за схожістю динаміки.

Методи дослідження ґрунтуються на комплексі наукових методів пізнання: метод обробки даних програмними алгоритмами, метод пошуку дистанцій для визначення схожості динаміки, метод побудови двовимірної площини з матриці дистанцій, нейромережеві методи прогнозування.

У першому розділі дисертаційної роботи проведено аналіз здобутків моделювання для створення прогнозів в економічному секторі. Найпоширенішими є моделі часових рядів, через їх простоту та інтерпретованість. Наразі набувають значного поширення моделі машинного навчання та нейромережеві алгоритми. Також проаналізовано моделі кластеризації часових рядів на дистанціях між цими рядами.

Другий розділ описує вибудову алгоритмів пошуку дистанцій між часовими рядами, алгоритму переведення матриці дистанцій у двовимірну площину й, насамкінець, алгоритмів кластеризації, що об'єднуються в модель для оцінки схожості динаміки часових рядів з дезагрегованого набору даних з метою їх подальшого групування й відокремлення власних випадкових шумів окремих компонент початкового набору даних. Ключову увагу приділено адаптації алгоритму DTW. Стандартний DTW враховує розтягнення та зміщення рядів один відносно одного, але допускає існування відповідності між двома точками з різних років, що є економічно-контрінтуітивним. Це вирішується запропонованою адаптацією.

Третій розділ описує створення моделей для прогнозування. Випадкове блукання є найпростішим традиційним алгоритмом, що виступає у ролі орієнтиру з яким порівнюють якість інших моделей. SARIMA є вже класичною економетричною моделлю, що спирається на власну динаміку. Також SARIMA важлива як частина ключового для цієї дисертаційної роботи багатоступеневого алгоритму, де вона використовується як один з етапів. Random Forest та XGBoost є прикладами алгоритмів машинного навчання, що також використовуються в роботі. На основі LSTM та SARIMA моделі будується фінальна модель що дає найкращі результати серед всіх порівняних, де на першому етапі оцінюються моделі SARIMA щоб прибрати власну динаміку, сезонність ряду, а вже потім на залишках оцінюється LSTM, що підхоплює всі нелінійні коливання та взаємозв'язки.

У **четвертому розділі** описуються результати вищезазначених моделей та багатоступеневих алгоритмів. Модель SARIMA+LSTM+UDTW+K-Means дає найкращий результат, проте Random Forest+UDTW+K-Means не набагато гірший в контексті якості прогнозів, але менш ресурсозатратний в контексті як людських ресурсів на побудову моделі, так і обчислювальних ресурсів.

У **висновках** окреслено ключові здобутки моделей, себто їх емпірична якість на основі бази даних компонент інфляції в Україні.

Наукова новизна отриманих результатів полягає в тому, що в поточній

дисертаційній роботі розроблено, скомбіновано та поліпшено алгоритми пошуку дистанцій та моделі прогнозування, а також створено новий набір послідовно-зв'язаних алгоритмів для прогнозування агрегованого показника на основі дезагрегованого набору даних.

Більш прості моделі досить широко представлені в літературі, але саме подібні комбінації алгоритмів є мало дослідженими, тому робота вносить суттєвий внесок у дослідження їх властивостей.

Вперше:

- розроблено комбінацію K-Means+UDTW+LSTM+SARIMA моделей й порівняно її прогностичну здатність з іншими алгоритмами
- алгоритми нейромережевого типу застосовано до дезагрегованих статистичних даних по інфляції в Україні й отримано результати з низьким RMSE
- створено інформаційну технологію на основі методів машинного навчання штучних нейронних мереж, випадкового лісу дерев рішень та інших для точного й швидкого прогнозування рівня базової інфляції України

Удосконалено:

- алгоритм пошуку дистанцій DTW для економічних рядів з щомісячними даними з метою подальшої кластеризації цих рядів

Практичне значення отриманих результатів полягає у створенні моделей прогнозування, що можуть бути застосовані для прогнозування на основі дезагрегованих даних. В рамках роботи демонструються можливості використання цього алгоритму для компонент інфляції, що є актуальним для бізнесу та урядових організацій і вже впроваджено у діяльності Національного Банку України.

Ключові слова: рекурентні нейромережеві алгоритми, LSTM, випадковий ліс, XGBoost, SARIMA, випадкове блукання, евклідова відстань, DTW, матриця дистанцій, метод К середніх, DBSCAN, прогнозування, інфляція, дезагрегація, часові ряди, точність, RMSE.

ABSTRACT

Krukovets D. Clustering of inflation components and forecasting using machine learning methods. Qualification scientific work as a manuscript.

Thesis for the degree of Doctor of Philosophy in the field of Information Technology, speciality 12 "Information Technology". Taras Shevchenko National University of Kyiv, Kyiv, 2024.

The **introduction** reveals the essence and current state of the scientific problem, justifying the choice and relevance of developing a multi-step algorithm based on clustering algorithms and neural networks for time series forecasting, particularly those with the property of disaggregation into subcomponents. The goal, object, subject, and research methods are defined, highlighting the scientific novelty of the research, theoretical and practical significance of the results, the author's personal contribution, and information about the implementation and testing of the results.

Forecasting processes and time series are essential for creating strategies with medium- and long-term planning, taking into account various factors and risks for acting proactively. However, creating high-quality predictive models is a non-trivial task.

The application of neural networks in economic analysis and forecasting is a relevant and promising direction. The dissertation focuses on developing a combined multi-step neural network model, combining it with clustering algorithms for a disaggregated dataset to simultaneously identify linear and seasonal features in the series and their interdependent structure.

The development is based on research by foreign scientists, particularly the work on disaggregated datasets proposed by Huwiler and Kaufmann in 2013, and the construction of similar models in the work of Almosova and Andersen in 2023. Experimental testing of various algorithms was also used for empirical justification of their effectiveness.

The **aim of the work** is to improve the efficiency of forecasting algorithms for datasets with disaggregation properties and non-linear interconnections. The

evaluation of the achievement is an empirical comparison of the developed algorithms with other machine learning methods and neural network approaches.

To achieve the research goal within the thesis, the following tasks were solved:

- Developed a methodology for constructing a neural network architecture model and corresponding machine learning system algorithms for time series forecasting, considering non-linearities between them
- Built a model adapted to inflation components for clustering time series based on distances between them
- Used a methodology for empirical evaluation of the model's effectiveness for comparison with models from the literature

The **object of the study** is forecasting methods and their characteristics for datasets with features inherent in economic series. The subject of the study is concepts, methods, algorithms, and neural network architectures for effective time series forecasting, as well as algorithms for searching distances and grouping based on the similarity of dynamics.

The research methods are based on a complex of scientific cognition methods: data processing methods using software algorithms, distance search methods to determine dynamics similarity, two-dimensional plane construction methods from distance matrices, and neural network forecasting methods.

The **first chapter** of the dissertation analyzes the achievements in modeling for forecasting in the economic sector. The most common are time series models due to their simplicity and interpretability. Currently, machine learning models and neural network algorithms are gaining significant popularity. Clustering models of time series based on distances between these series are also analyzed.

The **second chapter** describes the construction of algorithms for searching distances between time series, translating the distance matrix into a two-dimensional plane, and finally, clustering algorithms that combine into a model for assessing the similarity of time series dynamics from a disaggregated dataset to further group and isolate the individual random noise of the initial dataset

components. Key attention is paid to the adaptation of the DTW algorithm. The standard DTW considers stretching and shifting series relative to each other but allows correspondence between two points from different years, which is economically counterintuitive. This is resolved by the proposed adaptation.

The **third chapter** describes the creation of forecasting models. Random walk is the simplest traditional algorithm serving as a benchmark for comparing the quality of other models. SARIMA is already a classical econometric model relying on its own dynamics. SARIMA is also crucial as part of the multi-step algorithm key to this dissertation, where it is used as one of the stages. Random Forest and XGBoost are examples of machine learning algorithms also used in the work. Based on LSTM and SARIMA models, the final model is built, giving the best results among all compared, where at the first stage, SARIMA models are evaluated to remove the series' own dynamics and seasonality, and then LSTM is evaluated on the residuals to capture all non-linear oscillations and interconnections.

The **fourth chapter** describes the results of the aforementioned models and multi-step algorithms. The SARIMA+LSTM+UDTW+K-Means model gives the best result, but Random Forest+UDTW+K-Means is not much worse in terms of forecast quality, while being less resource-intensive in terms of both human resources for model construction and computational resources.

The **conclusion** outline the key achievements of the models, namely their empirical quality based on the Ukrainian inflation components database.

The scientific novelty of the obtained results lies in the fact that this dissertation developed, combined, and improved distance search algorithms and forecasting models, and also created a new set of sequentially linked algorithms for forecasting an aggregated indicator based on a disaggregated dataset.

Simpler models are widely represented in the literature, but such combinations of algorithms are little studied, thus this work makes a significant contribution to the study of their properties.

For the first time:

- Developed a combination of K-Means+UDTW+LSTM+SARIMA models

and compared their predictive ability with other algorithms

- Neural network-type algorithms were applied to disaggregated statistical data on inflation in Ukraine and obtained results with low RMSE
- Created an information technology based on machine learning methods of artificial neural networks, random forest decision trees, and others for accurate and fast forecasting of Ukraine's core inflation level

Improved:

- The DTW distance search algorithm for economic series with monthly data for further clustering of these series

The practical significance of the obtained results lies in the creation of forecasting models that can be applied for forecasting based on disaggregated data. The work demonstrates the possibilities of using this algorithm for inflation components, which is relevant for businesses and government organizations and has already been implemented in the activities of the National Bank of Ukraine.

Keywords: Recurrent Neural Network, LSTM, Random Forest, XGBoost, SARIMA, Random Walk, Euclidean distance, DTW, Distance matrix, K-means, DBSCAN, forecasting, inflation, disaggregation, time series, accuracy, RMSE.

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ:

А. В яких опубліковані основні наукові результати дисертації:

1. Krukovets, D. (2022). Multi-stage approach with DTW and clustering for forecasting of average deposit rate in Ukraine. Bulletin of Taras Shevchenko. National University of Kyiv. Series Physics & Mathematics, pp.55-65. https://www.doi.org/10.17721/1812-5409.2022/4.7
2. Krukovets, D. (2023). Updated DTW+K-Means approach with LSTM and ARIMA-type models for Core Inflation forecasting. Bulletin of Taras Shevchenko. National University of Kyiv. Series Physics & Mathematics, pp.214-225. https://www.doi.org/10.17721/1812-5409.2023/2.38
3. Krukovets, D. (2024). Exploring an LSTM-SARIMA routine for core inflation forecasting. Technology audit and production reserves, pp. 6-12. https://www.doi.org/10.15587/2706-5448.2024.301209

Б. Конференції за темою, в яких автор приймав участь

1. Krukovets, D. (2019): Non-stationary time-series distance clustering for a similarity analysis. Перша українська конференція «Логіка та її застосування» (UCLA'2019), pp. 117-119.
2. Krukovets, D. (2020). Analysis of similarity between artificially simulated time series with Dynamic Time Warping. In Workshop on Intelligent Information Systems (p. 97).
3. Krukovets, D. (2021): Dynamic Time Warping for uncovering dissimilarity of regional wages in Ukraine. "Proceedings MFOI-2020", pp. 168-185.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

AIC – Інформаційний критерій Акаїке

AR – Авторегресія, Авторегресійна модель

ARIMA – Авторегресійна Інтегрована модель ковзного середнього

ARIMAX – Авторегресійна Інтегрована модель ковзного середнього з екзогенною змінною

BVAR – Баєсова Векторна Авторегресійна модель

CNN – Конволюційна Нейронна Мережа

CPI – Індекс Споживчих Цін

DBSCAN – Модель просторової кластеризації на основі щільності із шумом

DSGE – Модель динамічної стохастичної загальної рівноваги

DTW – Алгоритм динамічного викривлення часу

EDM – Матриця евклідових дистанцій

FAVAR – Факторна Векторна Авторегресійна модель

FastDTW – Швидкий Алгоритм динамічного викривлення часу

HICP – Гармонізований Індекс Споживчих Цін

IQR – Інтерквартильний розмах

K-Means – Метод К Середніх

LSTM – нейронна мережа довгострокової та короткострокової пам’яті

MA – Ковзне середнє

PCA – Метод Головних Компонент

QPM – Квартальна проєктильна модель

RNN – Рекурентна Нейронна Мережа

RMSE – Корінь Середньоквадратичного відхилення

SARIMA – Сезонна Авторегресійна Інтегрована модель ковзного середнього

SVM – Модель опорних векторів

SVAR – Структурна Векторна Авторегресійна модель

VAR – Векторна Авторегресійна модель

XGBoost – Екстремальний Градієнтний Бустинг

ВВП – Внутрішній Валовий Продукт

ВСТУП

Актуальність теми. Прогнозування процесів та часових рядів є невід'ємною частиною сучасного світу. Прогнози дозволяють робити планування, розробляти моделі середньострокового та довгострокового розвитку, що враховують велику кількість факторів та ризиків. Прогнозування використовується як в специфічних контекстах (прогнозування рівня правопорушень в конкретному районі в рамках місяця), так і в дуже загальних (майбутній розвиток економіки, демографії). Розуміння потенціалу економіки дозволяє розробляти політики та стратегії уникнення небажаних наслідків й діяти превентивно. Проте, створення якісних моделей для, власне, прогнозування, є нетривіальною задачею й в рамках цієї дисертаційної роботи буде розглянута побудова однієї з таких моделей.

Прогрес технологій та зростання обсягу інформації мають вирішальне значення для багатьох галузей, зокрема для економічного аналізу. Свідченням цього є поступова відмова від класичних економетричних методів на користь передових інструментів машинного навчання та аналізу даних, зокрема нейромережевих алгоритмів. Хоча нові технології не заміщають повністю простіші методи, які легше застосовувати й інтерпретувати, вони все більше займають вагоме місце в аналізі економічних процесів. Також вони якісно поліпшують можливості підхоплення нелінійних взаємозв'язків у часових рядах на відміну від традиційних алгоритмів.

Використання нейронних мереж у економічному аналізі та прогнозуванні є актуальною та перспективною тенденцією в сучасній науці, розвиваючись паралельно з галуззю комп'ютерних наук, де постійно виникають нові та вдосконалені архітектури, методи оцінки коефіцієнтів та розв'язання різноманітних проблем, починаючи від стандартних проблем перенавчання чи зникаючого градієнту і закінчуючи специфічними проблемами у теорії сигналів [53]. Нейронні мережі, зокрема прості, глибокі та рекурентні, широко застосовуються у вирішенні завдань економічного аналізу та прогнозування.

З огляду на зазначене, в даній дисертаційній роботі фокус буде на розробці комбінованої нейромережевої моделі, разом з використанням алгоритмів кластеризації дезагрегованого набору даних, задля суттєвого покращення можливостей одночасного підхоплення як власних лінійних та сезонних особливостей рядів, так і їх взаємозалежної структури.

Розробка наукових і методологічних основ створення комплексної методології прогнозування з використанням нейромережевих та кластерних алгоритмів базується на дослідженнях зарубіжних вчених через недостатню висвітленість теми в українській науковій літературі. Мова й про використання загальної ідеї роботи з дезагрегованими рядами, описаної Huwiler та Kaufmann у 2013-му році [94], й про спроби побудови схожих моделей, проте не таких пропрацьованих, як в Almosova та Andersen у 2023-му [100], й про експериментальні спроби використання великої кількості алгоритмів [68, 69, 70, 73, 81, 86, 89, 93, 96, 97, 98, 99] задля емпіричного доведення їх ефективності.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконана відповідно до поточних та перспективних планів наукової та науково-технічної діяльності кафедри Теорії та Технології Програмування, факультету комп'ютерних наук та кібернетики Київського національного університету імені Тараса Шевченка. Тема дисертаційної роботи доповнює здобутки кафедри в сфері науки про дані новими алгоритмами та перспективними можливостями їх використання. Здобувач як виконавець брав участь у науково-дослідних роботах та конференціях.

Мета і завдання дослідження. Метою роботи є підвищення ефективності алгоритмів прогнозування часових рядів дезагрегованих баз даних з нелінійними взаємозв'язками між собою. Також метою є емпірична перевірка нових розроблених алгоритмів з іншими методами машинного навчання та більш простими нейромережевими алгоритмами, а також зі стандартними моделями для порівняння та традиційними для сфери застосування та експериментів моделями.

Для досягнення мети дослідження в дисертації вирішені такі завдання:

- розроблено методику побудови моделі нейромережевої архітектури та відповідних алгоритмів системи машинного навчання для прогнозування часових рядів з врахуванням нелінійності взаємозв’язків рядів між собою
- Побудовано адаптовану до інфляційних компонент модель кластеризації часових рядів за дистанціями між ними
- використано методологію емпіричної оцінки ефективності моделі задля порівняння з іншими традиційними або більш простими варіантами розробленої моделі

Об’єкт дослідження – це методи прогнозування та їх властивості в рамках баз даних з певними особливостями, притаманними рядам з економічної сфери.

Предметом дослідження є концепції, методи, алгоритми, архітектури нейронних мереж для ефективного розв’язання задачі прогнозування часових рядів, а також алгоритми для пошуку дистанцій та розбиття на групи за схожістю динаміки часових рядів.

Методи дослідження. Методологічною основою дослідження виступає комплекс наукових методів пізнання: метод обробки даних часових рядів програмними алгоритмами, метод обробки часових рядів статистичними алгоритмами задля підготовки, метод пошуку дистанцій між часовими рядами здаля встановлення міри схожості та взаємозалежності динаміки рядів, метод побудови двовимірної площини з матриці дистанцій, метод прогнозування агрегованих категорій задля подальшої агрегації в загальний ряд економетричними, нейромережевими методами та алгоритмами машинного навчання.

Наукова новизна отриманих результатів полягає в тому, що поточна дисертаційна робота описує розробку, комбінування та поліпшення одразу декількох алгоритмів в різних частинах роботи (мова і про алгоритми пошуку дистанцій, і про моделі прогнозування), а також створення унікальної комбінації з низки алгоритмів задля прогнозування агрегованого показника на основі дезагрегованої бази даних. Більш прості моделі достатньо широко представлені

в літературі, проте саме подібні комбінації алгоритмів розкриті мінімально, тому робота робить суттєвий внесок у дослідження їх властивостей.

Вперше:

- розроблено комбінацію нейромережевого алгоритму з SARIMA моделлю на основі компонент, отриманих алгоритмами пошуку дистанцій між рядами, кластеризацією та агрегацією, й порівняно її прогностичну здатність з іншими моделями
- алгоритми нейромережевого типу застосовано до дезагрегованих статистичних даних по інфляції в Україні й отримано результати з низьким RMSE
- створено інформаційну технологію на основі методів машинного навчання штучних нейронних мереж, випадкового лісу дерев рішень та інших для точного й швидкого прогнозування рівня базової інфляції України

Удосконалено:

- алгоритм пошуку дистанцій DTW для економічних рядів з щомісячними даними з метою подальшої кластеризації цих рядів

Практичне значення отриманих результатів полягає в створенні моделей прогнозування що можуть використовуватися в різноманітних сферах задля прогнозування агрегованих показників з використанням дезагрегованих даних. Також, в рамках дисертаційної роботи показані можливості для використання цього алгоритму на базі даних компонент інфляції, що актуально для бізнесів та урядових організацій й вже використовується у діяльності Національного Банку України, оскільки для цієї інституції прогнози інфляції є критично важливими для розробки монетарної політики.

Особистий внесок здобувача. Дисертація виконана здобувачем самостійно з використанням останніх досягнень галузі інформаційних технологій.

Структура та обсяг дисертації. Дисертація складається з анотації, вступу, чотирьох розділів, висновків до кожного розділу та загальних висновків, а також

списку використаних джерел. Загальний обсяг дисертації становить 137 сторінок, з них основного тексту – 112 сторінок, 5 рисунків та 4 таблиці на 5 сторінках. Список використаних джерел налічує 102 найменувань.

РОЗДІЛ 1. Економетричні моделі прогнозування

Еволюція технологій та збільшення кількості інформації зіграли надзвичайну роль в багатьох галузях, зокрема й у економічному аналізі. Свідченням цього є поступовий перехід від класичних економетричних методів до передових інструментів машинного навчання та науки про дані, нейромережевих алгоритмів. Звісно, ці новітні технології не повністю витісняють більш прості для побудови та інтерпретації алгоритми, проте вони займають все більш значне місце.

Класичні методи економетрики, такі як ARIMA, SARIMA, VAR, QPM та DSGE, довгий час були основою економічного аналізу, надаючи уявлення про економічні тенденції та закономірності [1]. Однак з появою великих обсягів даних та передових обчислювальних технік економісти все більше звертаються до передових алгоритмів машинного навчання та інструментів науки про дані для покращення їх можливостей у прогнозуванні, оскільки традиційні алгоритми не настільки здатні підхоплювати нелінійні зв’язки й неспроможні використовувати великі бази даних з десятками чи, навіть, сотнями часових рядів.

ARIMA (Autoregressive Integrated Moving Average) та SARIMA (Seasonal Autoregressive Integrated Moving Average), використовуються для аналізу та прогнозування часових рядів і є найпростішим прикладом економетричних моделей. Ці методи базуються на математичних моделях для захоплення основних закономірностей та попередньої динаміки у даних задля прогнозування майбутнього [2]. VAR (Vector Autoregression) використовується для аналізу взаємозв'язків між кількома змінними, що дозволяє отримати комплексне уявлення про систему, зокрема й економічну. У VAR моделей є значна кількість аналогів та поглиблених варіантів, такі як структурна VAR модель (SVAR) чи факторна VAR модель (FAVAR), а також Баєсова VAR модель (BVAR). Вони додатково поглиблюють структуру та дозволяють краще імплементувати експертні судження про природу змінних [3].

Моделі QPM (Quarterly Projection Models) [4] та DSGE (Dynamic Stochastic General Equilibrium) [5] представляють більш складні економетричні каркаси, які спрямовані на захоплення взаємозв'язків у економічних системах. По суті це великі моделі з десятками рівнянь (система рівнянь) що взаємопов’язують динаміку різних змінних й об’єднують їх у загальну модель, що досліджує економіку з різних боків відповідно до теоретичних основ. QPM є корисним для аналізу складних економічних явищ та взаємозв’язків між загальними макроекономічними змінними, тоді як DSGE моделює взаємозв'язки між макроекономічними змінними на основі мікроекономічних принципів, на рівні рівнянь для окремих економічних агентів, домогосподарств та бізнесів.

Незважаючи на їх ефективність, класичні методи економетрики мають свої обмеження, зокрема у роботі з великими та складними наборами даних з великою кількістю ознак та нелінійними взаємозв’язками [6]. Це призвело до вивчення можливостей передових алгоритмів машинного навчання та інструментів науки про дані в економічному аналізі [7].

Алгоритми машинного навчання, такі як Random Forest, XGBoost та техніки кластеризації, надають потужні інструменти для прогнозування інфляції та аналізу економічних даних. Random Forest та XGBoost - це ансамблеві методи навчання, які можуть працювати з нелінійними залежностями та взаємозв'язками між змінними, що робить їх ідеальними для захоплення складних економічних динамік [8]. Техніки кластеризації, з іншого боку, дозволяють економістам ідентифікувати відмінні патерни та групи в даних, розкриваючи підґрунтя економічних структур та взаємозв'язків [9]. Ці алгоритми доволі легко використовувати на підготовлених базах даних, що дає непогані результати за мінімальних зусиль, як з точки зору побудови моделей, так і з точки зору обчислювальних потужностей.

Окрім цього, поява нейронних мереж, включаючи прості фідфорвардні мережі, рекурентні нейронні мережі (RNN) та більш складні архітектури, такі як згорткові нейронні мережі (CNN) та моделі трансформери, революціонізувало прогнозування економічних величин й низка дослідницьких агенцій та урядових

установ почали пробувати адаптувати ці структури для власних потреб. Зокрема RNN та LSTM моделі якісно проявили себе в захопленні часових залежностей у рядах даних, що робить їх дуже ефективними для прогнозування інфляції та інших економічних показників [10-12]. З іншого боку, ці моделі є традиційно дуже великими та такими, що довго обчислюються. Тому вони не можуть бути використані як «просте рішення». Проте є якісним варіантом як результат грунтовного дослідження динаміки низки рядів та їх поєднання задля прогнозування певної величини.

Також додатково варто зробити акцент на алгоритмах для пошуку дистанцій між часовими рядами. Цей підхід дозволяє показати схожі часові ряди задля їх потенційного групування без використання ресурсозатратних експертних суджень, а використовуючи чисту автоматизовану статистику та математику [13]. Ці алгоритми використовуються для роботи з великими наборами часових рядів задля зменшення бази даних без значних втрат інформації.

У підсумку, еволюція технологій у економічному аналізі, зокрема в прогнозуванні інфляції, свідчить про перехід від класичних методів економетрики до передових інструментів машинного навчання та науки про дані. Цей перехід відбувається з метою покращення прогностичних здібностей та усвідомлення економічних процесів, що лежать в основі економічних рядів. Огляд цього шляху є основою цього розділу, що поділений на частини, де по черзі будуть розглядатися більш традиційні моделі економетричного аналізу, а потім історія застосування новітніх алгоритмів машинного навчання та науки про дані, також важливого для поточної дисертаційної роботи методу кластеризації часових рядів за їх дистанцією між собою. Це є передумовою та фундаментом на якому вибудовувались алгоритми, що пропонуються та тестуються в рамках цієї дисертаційної роботи.

1.1. Традиційні математичні та статистичні моделі для економічних задач

Впродовж десятків років для аналізу економічних даних та прогнозування широко використовалися різноманітні статистичні та економетричні моделі.

Основні моделі, що застосовуються, включають авторегресійні інтегровані моделі з ковзним середнім (ARIMA), сезонні авторегресійні інтегровані моделі з ковзним середнім (SARIMA), їх версії з екзогеною змінною (ARIMAX та SARIMAX), векторні авторегресійні моделі (VAR), структурні векторні авторегресійні моделі (SVAR), факторно-авторегресійні моделі (FAVAR), квазіпараметричні напівструктурні моделі (QPM) та структурні моделі динамічної стохастичної загальновизначеної рівноваги (DSGE). Впродовж цього розділу ми розглянемо літературу про вищезазначені моделі, як приклади застосування.

ARIMA та SARIMA використовуються для аналізу та прогнозування часових рядів економічних даних, здатні виявляти та моделювати регулярні зміни в даних, а другі ще й можуть виявляти сезонні варіації рядів. Математично, ARIMA модель може бути представлена як:

$$Y_t = c + \beta_1 * Y_{t-1} + \cdots + \beta_n * Y_{t-n} + \gamma_1 * \varepsilon_{t-1} + \cdots + \gamma_m * \varepsilon_{t-m} + \varepsilon_t \quad (1.1)$$

де Y_t – значення часового ряду в момент часу t, c - постійний член або константа, β_1, ... β_n – коефіцієнти авторегресії, γ_1, ... γ_m – коефіцієнти ковзного середнього, ε_t – шумова складова. ARIMA також включає інтегровану складову, яка забезпечує стаціонарність часового ряду.

SARIMA розширює ARIMA, додаючи сезонну складову, де S - період сезонності. Себто ми отримуємо модель з двох складових, де окремо є стандартна ARIMA, а додатково до неї є ARIMA з затримкою (лагом) рівним періодичності ряду (4 для квартальних часових рядів, 12 для місячних і так далі).

Оптимізація параметрів ARIMA та SARIMA моделей зазвичай використовує алгоритм на основі методу максимальної правдоподібності (Maximum Likelihood Estimation, MLE). Цей алгоритм шукає значення параметрів, які максимізують ймовірність отримати спостережені дані при заданих модельних параметрах. Через складність оптимізації параметрів, для розв'язання задачі використовуються різні методи, такі як метод Ньютона-Рафсона, метод квазі-Ньютона та інші ітераційні методи оптимізації.

У статті, написаній Meyler та ін. у 1998 році [14], пропонується комплексна методологія використання моделей авторегресії з інтегрованим рухом середньої (ARIMA) для прогнозування інфляції в Ірландії. Стаття вводить рамки для прогнозування за допомогою ARIMA, акцентуючи увагу на мінімізації помилок прогнозування за межами вибірки ніж на максимізації внутрішньої відповідності. Цей підхід спрямований на оптимізацію результатів прогнозування через процес, який автори називають "видобуванням моделей". Методологія розглядає два підходи до ідентифікації моделей ARIMA: традиційний підхід Бокса-Дженкінса та методи об'єктивних функцій штрафів. Перший включає ітеративні кроки ідентифікації моделі, її оцінки, діагностики та прогнозування, тоді як другий використовує функції штрафу для вибору найбільш підходящої моделі. У статті аргументується акцент на точності прогнозування, враховуючи значення прогнозування інфляції в процесі прийняття монетарних рішень й важливість, власне, точних прогнозів. Практичні аспекти прогнозування часових рядів ARIMA ілюструються на прикладі Гармонізованого індексу споживчих цін (HICP) та його основних підкомпонентів. Досліджуються шість часових рядів, які охоплюють різні аспекти HICP, включаючи непереоблені продукти, оброблені продукти, промислові товари без енергії, енергію та послуги. Це різноманіття дозволяє провести комплексний аналіз проблем і можливостей прогнозування, оскільки окремі компоненти мають, відповідно різні проблеми та виклики для прогнозної моделі. Значний внесок статті в літературу полягає в акценті на якості прогнозування метриками типу RMSE замість акценту на пояснювальній здатності моделі власної динаміки рядів. Також одним з важливих аспектів статті є підхід, що прогнозування відбувається не для інфляції в цілому, а для компонентів, що показує різну природу та динаміку компонент. Обидва ці висновки є вкрай важливими, фундаментальними, для моделей представлених в цій дисертаційній роботі, навіть не кажучи про сферу економічного прогнозування в цілому.

У більш свіжій статті Mendal та ін., написаній у 2014 році [15], досліджується ефективність використання моделі авторегресії з інтегрованим рухом середньої (ARIMA) для прогнозування цін на акції на ринку Індії. Автори обирають ARIMA модель через її простоту та широку прийнятність. Основний внесок статті полягає у тому, що вона охоплює значну кількість акцій на індійському ринку, аналізуючи їх за секторами, і досліджує точність прогнозування в залежності від періоду попередніх даних. Методологічно, автори використовують ARIMA модель для аналізу часових рядів цін на акції. Ця модель конвертує нестаціонарні дані в стаціонарні перед їх аналізом інтегруванням й автори акцентують додаткову увагу на цьому аспекті моделей. Для порівняння та параметризації моделі ARIMA використовується інформаційний критерій Акаїке (AIC). Стаття виходить з того, що прогнозування цін на акції має значення для фінансових рішень та може бути корисним для інвесторів та фінансових професіоналів. Прогнозування цін на акції за допомогою ARIMA моделі дозволяє вирішити ці задачі з високою точністю, сприяючи прийняттю обґрунтованих рішень на ринку. Ця стаття акцентує увагу на методах та особливостях роботи з фінансово-економічними даними у випадку, коли рядів багато і як набір моделей може суттєво скоротити роботу аналітика, котрий замість аналізу великого набору даних покладається на низку моделей.

З іншого боку, ARIMA моделі використовуються не тільки для роботи з інфляцією. У статті, написаній авторами Mohamed Reda Abonazel and Ahmed Ibrahim Abd-Elftah, 2019 [16], досліджується моделювання та прогнозування ВВП Єгипту на основі підходу Бокса-Дженкінса на основі річних даних з 1965 по 2016 рік. Використовуючи цей підхід, автори побудували відповідну модель ARIMA для аналізу даних про ВВП Єгипту. Дослідження показало, що найкращою є ARIMA (1, 2, 1). При цьому, автори порівнюють різні моделі ARIMA за допомогою критеріїв якості підгонки, таких як середньоквадратична помилка (MSE), критерій Акаїке (AIC) та критерій Байєса-Шварца (BIC).

Іншим прикладом є прогнозування обмінного курсу. У статті, написаній Maria, Eva, 2011 [17], досліджується прогнозування обмінних курсів

румунського лея відносно євро, долара США, фунта стерлінгів, японської єни та китайського юаня. Автори порівнюють та аналізують різні методи згладжування, включаючи просте експоненційне згладжування, подвійне експоненційне згладжування, просте та адитивне згладжування Хольта-Вінтерса, а також ARIMA модель. Розглядається питання поведінки обмінних курсів в короткостроковій перспективі та їх вплив на учасників фінансового ринку. Підкреслюється тенденція зростання румунського лея відносно інших валют. Автори вказують на складнощі у оцінці та валідації ARIMA моделей, а також на їхню ефективність у передбаченні середньострокових тенденцій, які, у випадку даного дослідження, становлять 4 місяці. Такі моделі демонструють зміни в тренді, тоді як прогнозні моделі на основі експоненційного згладжування є ефективним інструментом для тих, хто цікавиться динамікою обмінного курсу в найкоротшій перспективі задля врахування найменших коливань ринкових умов.

Розширені моделі ARIMA та SARIMA, такі як ARIMAX (Autoregressive Integrated Moving Average with Exogenous Variables) та SARIMAX (Seasonal Autoregressive Integrated Moving Average with Exogenous Variables), використовуються для аналізу та прогнозування часових рядів з урахуванням впливу зовнішніх факторів або екзогенних змінних. Ці моделі включають додаткові змінні, що не контролюються самим рядом даних, але можуть впливати на нього, що дозволяє покращити точність прогнозів та зробити аналіз більш повним, оскільки враховуються додаткові фактори, які можуть впливати на поведінку ряду. ARIMAX та SARIMAX є корисними інструментами для моделювання та прогнозування складних економічних процесів, де важливо враховувати вплив зовнішніх факторів або контрольовані змінні.

$$\begin{aligned} Y_t = {} & \mathrm{c} + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \\ & \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \\ & \beta_1 X_{1,t} + \beta_2 X_{2,t} + \cdots + \beta_n X_{n,t} + \epsilon_t \end{aligned} \qquad (1.2)$$

де Y_t – значення часового ряду в момент часу t, c - постійний член або константа, ϕ_1, ... ϕ_p – коефіцієнти авторегресії, θ_1, ... θ_q – коефіцієнти ковзного

середнього, ε_t – шумова складова в момент t, β_1, ... β_n – коефіцієнти при екзогенних (зовнішніх) змінних.

У статті, написаній Ulyah, 2019 [18], проведено дослідження впливу президентських виборів в Індонезії 2019 року на ринок капіталу, зокрема на Індонезійський індекс (IDX composite) та акції компанії PT. Saratoga Investama Sedaya Tbk (SRTG). Методологічний підхід дослідження поділяється на дві категорії, проте найцікавішим та основним є саме уніваріатний аналіз, де використовується авторегресійно-інтегрована ковзна середня з екзогенною змінною (ARIMAX). Екзогенною змінною у дослідженні є фіктивна змінна, пов'язана з президентськими виборами в Індонезії. Результати показали, що попередньовиборчий настрій не має значущого впливу на дохідність індексу та акцій. У вступі статті робиться акцент на тому, що індекс Індонезії за гіпотезою має бути під впливом президентських виборів 2019 року, і підтверджується це на основі досвіду попередніх виборів. Себто попередні дослідження показали, що політичний фактор має тенденцію впливати на індекс й його варто включати в аналіз ринку. Крім того, обговорюються попередні дослідження з використанням ARIMA та ARIMAX для аналізу та прогнозу макроекономічних часових рядів, де вказується на перевагу ARIMAX через можливість включення до пояснення динаміки альтернативних змінних. Проте у підсумку виявляється що вплив є незначущим й це показує які можливості для аналізу та прогнозування відкриває включення екзогенних змінних.

В іншому дослідженні Ugoh та ін., 2021 [19], пропонується ARIMAX модель, яка використовується для прогнозування ВВП Нігерії. Прогнозування ВВП ARIMA моделями, взагалі, є не дуже частим явищем через відносно слабку прогностичну здатність, оскільки ВВП пояснюється більшою мірою шоками в економіці (котрих може бути дуже багато), аніж минулою динамікою. Тому для дослідження використовувалися дані за період 1990-2019 років, отримані від Світового банку. Модель ARIMA була підігнана до залишкових даних за допомогою методу Бокса-Дженкінса. Критерій байєсовської інформаційної величини (BIC) було прийнято для оцінки адекватності моделей. Сировинні дані

задовольняли умову мультиколінеарності після вилучення експорту, а залишковий ряд став стаціонарним після першого диференціювання. Дослідження показало, що імпорт є значущою екзогенною змінною для динаміки ВВП, себто відкритість економіки значно впливала й її врахування поліпшувало якість прогнозів й ARIMAX модель дає кращі результати за відповідну ARIMA модель за методом Уайла.

У статті, написаній Anggraeni та ін., 2017 [20], досліджується прогнозування цін на рис, який є одним зі стратегічних товарів і важливим елементом в житті індонезійського суспільства. Одним із пріоритетів урядової політики Індонезії є стабілізація цін на продукти харчування, що може мінімізувати вплив глобальної фінансової кризи на рівень інфляції та покупну спроможність населення. Модель ARIMAX враховує кілька екзогенних змінних що суттєво поліпшують якість прогнозу: роздрібну ціну для споживачів (у одиницях рупій/кг), виробництво рису (у тоннах), внутрішнє та зовнішнє закупівлю рису (у нетто-тоннах), ціну збираного сухого зерна (рупій/кг), площу збирання рису (у гектарах), ціну рису в світі (Bangkok 5%, у доларах США за метричну тонну, FOB).. Результати показали, що модель ARIMAX може прогнозувати ціни на рис для споживачів на 15,27% краще, ніж альтернативні моделі. Для заповнення порожніх значень у даних про роздрібну ціну рису використовується метод інтерполяції. Ця стаття додає до усвідомлення важливості використання екзогених змінних іншої природи, проте схожого характеру, задля прогнозування основної змінної. І це є важливим для поточної дисертаційної роботи з огляду на можливості додавати в базу даних змінні іншого характеру й використовувати їх як екзогенні.

Присутність моделей типу ARIMAX в літературі стосовно України напрочуд низька. Одним з небагатьох прикладів є стаття Kulyk et al., 2023 [21], де досліджується прогнозування тенденцій у зростанні популяцій худоби та корів. Автори використовували різні форми моделі SARIMAX, включаючи простіші варіанти такі як AR та MA, закінчуючи, власне, SARIMAX моделлю. Для аналізу використовуються місячні статистичні дані про кількість худоби та

корів. Для оптимізації даних використовуються методи перетворення Бокса-Кокса та потрійне експоненціальне згладжування. Побудовані часові ряди порівнюються з реальними значеннями, а оцінки квадратичного середнього відхилення та середньої абсолютної відсоткової похибки для різних термінів прогнозування також надаються в статті. Порівняння цих оцінок для різних інтервалів часу дозволяє визначити оптимальний період прогнозування (24 місяці). Робота вказує деталі про важливість тваринництва як ключової галузі сільського господарства, що забезпечує різноманітні сфери економіки та є одним з основних джерел сировини. Підкреслюється значення сталого збереження популяцій тварин для гарантування продовольчої безпеки та забезпечення населення необхідними продуктами харчування. Аналізується тенденція зменшення кількості худоби та корів в Україні, а також обговорюється важливість раціонального управління та стратегій подальшого розвитку галузі. Ці результати є важливими для прийняття управлінських рішень, планування ресурсів, підвищення ефективності та адаптації до змін на ринку. У висновках подаються результати, зазначається, що кількість худоби (включаючи корів) має виражену тенденцію до зменшення. Побудовані моделі SARIMAX показують поступове зменшення кількості худоби до кінця 2024 року. Ці дослідження сприяють формулюванню стратегій та дій, які є критичними для ефективного управління, планування ресурсів, підвищення прибутковості та адаптації до змін у ринкових умовах.

Іншим надпопулярним типом моделей для економічного аналізу є VAR, а також його адаптації типу SVAR (структурна), FAVAR (факторна), BVAR (Баєсова) [22-23]. Вони дозволяють аналізувати взаємозв'язки між декількома змінними та визначати ефекти взаємних змін у великих системах даних. Ці моделі часто застосовуються для вивчення взаємозв'язків між різними економічними показниками, такими як виробництво, споживчі витрати та інвестиції, та визначення їх впливу один на одного в часі. VAR модель дозволяє моделювати взаємозалежності між декількома часовими рядами без чіткої специфікації причинно-наслідкових зв'язків. Вона базується на ідеї, що кожна

змінна в часовому ряді може бути прогнозована на підставі попередніх значень кожної змінної, включаючи саму себе та інші змінні. SVAR модель, натомість, розширює VAR, дозволяючи включати в модель структурні зміни або відповіді на зовнішні шоки. Це дозволяє встановити причинно-наслідкові зв'язки між змінними та оцінити вплив зовнішніх факторів на економічні процеси. FAVAR модель використовується для аналізу великих масивів даних, коли кількість змінних перевищує кількість спостережень. Вона базується на включенні факторних змінних, які дозволяють зменшити кількість змінних у моделі і врахувати загальний вплив групи змінних на цільовий процес. Це часто досягається використанням додаткових алгоритмів на кшталт методу головних компонент (Principal Component Analysis). BVAR модель дозволяє поліпшити оцінку коефіцієнтів експертними судженнями стосовну типу та виду розподілу коефіцієнтів, що задаються експертами на підставі додаткових знань про досліджувану економічну величину.

Звичайні VAR моделі визначаються такою формулою:

$$Y_t = c + \sum_{i=1}^{p} A_i Y_{t-i} + \varepsilon_t$$

де Y_t - k × 1 вектор ендогенних змінних, c - вектор констант, A_i - матриці авторегресії розміром k × k для кожного лагу i, p - кількість лагів, ε_t - k × 1 вектор білих шумів.

Структурні VAR (SVAR) моделі додають до VAR формули систему рівнянь структурних векторних авторегресій:

$$Y_t = B_0 + \sum_{i=1}^{p} B_i Y_{t-i} + \varepsilon_t$$

Де B_0 - матриця констант, B_i - матриці розміром k × k для кожного лагу i, які відображають структурні зв'язки між змінними.

Факторно-векторні авторегресійні моделі (FAVAR) додають до VAR формули факторні змінні:

$$Y_t = c + \Lambda F_t + \sum_{i=1}^{p} A_i Y_{t-i} + \varepsilon_t$$

де F_t- m × 1 вектор факторних змінних, Λ - матриця факторних навантажень, яка відображає вплив факторів на ендогенні змінні.

Одним з перших кроків якісного та глибокого використання багатьох можливостей VAR моделей стала робота Jacobson et al, 1999 [24], де було розглянуто ідеї та новаторські особливості використання VAR-моделі в аналізі питань, що стосуються монетарної політики. Тоді зростала зацікавленість у емпіричних дослідженнях через дерегуляцію фінансових ринків та збільшене використання явних правил та цілей політики, що зробили монетарну політику більш прозорою та цікавою для економічного аналізу. Векторно-авторегресійні (VAR) моделі були широко використані в емпіричних дослідженнях макроекономічних питань з моменту їх створення Sims у 1980 році [25]. Автори застосували підхід VAR може бути корисно застосований в аналізі питань, які є центральними для монетарної політики в невеликій відкритій економіці у режимі інфляційного таргетування наведених далі. Чи допомагає номінальний обмінний курс передбачити інфляцію? Чи коригується номінальний обмінний курс від різниці між внутрішньою та зовнішньою інфляцією, щоб відновити якусь рівновагу реального обмінного курсу? Наскільки корисні різні показники грошових умов та вихідний розрив? Наскільки швидко зміни в монетарній політиці впливають на виробництво і інфляцію? Ці питання стосуються складних відносин між змінними, які всі є ендогенними та одночасно визначаються в економічній системі, й робота частоково відповіла на ці питання саме за допомогою методу одночасної оцінки в системі рівнянь VAR.

Наразі, VAR-підхід є широко дослідженим в економічних роботах й постійно поглиблюється з різних сторін. Наприклад, у статті, написаній Awokuse, 2006 [26], досліджується причинно-наслідковий зв'язок між реальними експортними обсягами та зростанням ВВП в Японії з використанням двох недавно розроблених підходів до каузального моделювання. Застосовуючи

японські часові ряди, стаття використовує розширену VAR методологію, щоб перевірити на відсутність каузальності Гренджера ряди (тести описані, наприклад, Shojaie та Fox у 2022 [27]). Потім також використовується більш недавно розроблена техніка напрямних ациклічних графів (DAG), яка надає додаткові обмеження на інновації з векторної авторегресії (VAR) моделі. На відміну від попередніх робіт, застосування технік DAG дозволяє досліджувати як одночасну, так і динамічну каузальну структуру експортно-продуктивного комплексу. Емпіричні результати показують, що каузальний шлях між експортом та зростанням ВВП в Японії є двонапрямленим. Крім того, інші змінні, такі як капітал та іноземний вивід, також є значущими детермінантами зростання продуктивності в Японії. Природа каузального зв'язку між експортом та економічним зростанням була предметом активної дискусії в літературі з економічного розвитку протягом останніх десятиліть. Оскільки теорія торгівлі не надає чіткої пояснення, дебати, як правило, ґрунтуються на висновках, заснованих на анекдотичному інтуїтивному сприйнятті та емпіричних аналізах, які часто дають двозначні результати. Основне питання у дебатах щодо експортно-орієнтованої політики торгівлі полягає у тому, чи є експортно-орієнтована політика переважнішою за внутрішньо-орієнтовану політику для стимулювання економічного зростання. Деякі дослідники стверджують, що каузальність від експорту до зростання продуктивності, і позначають це як гіпотезу про експортне зростання (ELG), тоді як зворотний потік каузальності від продуктивності до експорту називається зростанням, що обумовлене експортом (GLE). Раніше проведені дослідження по панельним даним різних країн критикувалися за їхню обмежену припущення про постійність параметрів між різними країнами. Однак оскільки з'явилось більше даних, більш пізні аналізи сконцентрувалися на однокраїнних дослідженнях, використовуючи техніки моделювання часових рядів. Більшість цих досліджень використовують концепцію Гренджера для перевірки біваріатних каузальних зв'язків між експортом та економічним зростанням. Часові ряди показали суперечливі докази щодо каузального зв'язку між експортом та зростанням.

У статті, написаній Sznajderska, 2019 [28], досліджується роль Китаю в світовій економіці та вплив шоків, що виникають у китайській економіці, на інші країни. Основними завданнями статті є оцінка впливу негативних шоків на попит та ціни акцій у Китаї на інші економіки, порівняння реакції розвинених країн з реакцією країн з розвиваючоюся економікою, та вивчення трансмісії цих шоків на внутрішню економіку Китаю. Для досягнення цієї мети застосовується модель глобальної векторно-авторегресії (GVAR, методологічно модель є просто VAR, проте з певними особливостями), яка дозволяє моделювати міжнародні зв'язки між країнами. Отримані результати показують, що на короткостроковий період один відсоток негативного шоку ВВП в Китаї зменшує світове зростання на 0,22%. Виявлено, що шок ВВП має сильніший вплив на економіки розвиваючихся країн, ніж на розвинені економіки. Також показано, що шок цін на акції впливає лише на економіки розвиваючихся країн і не впливає на розвинені економіки. Стаття звертає увагу на міжнародні спільні впливи, що виникають внаслідок змін в китайській економіці, та досліджує, як ці негативні шоки поширюються на інші країни, зосереджуючи увагу на реакції розвинених і розвиваючихся економік.

Нарешті перейдемо до адаптації моделі VAR. Ahmel та Wadud, 2011 [29], досліджували вплив нестабільності цін на нафту на макроекономічні активності та монетарні реакції в Малайзії. Для цього використовується вже структурна модель векторної авторегресії (SVAR) на основі щомісячних даних за період 1986-2009 років. Оцінки моделі показують важливий асиметричний ефект шоків цін нафти на умовну волатильність. Динамічні функції імпульсного відгуку, отримані з моделі SVAR, показують тривалий приглушуючий ефект шоку волатильності цін нафти на промислове виробництво Малайзії. Також виявили, що рівні індексу споживчих цін (CPI) знижуються при позитивному шоці нафтової цінової нестабільності. Це є наслідком негативного попитового шоку через відкладення споживання великих покупок фізичними особами, домогосподарствами та іншими секторами економіки. Застосування SVAR моделі дозволяє оцінити динаміку відгуку макроекономічних факторів на шоки,

що виникають від волатильності цін нафти. Результати дослідження вказують на значний вплив умовної волатильності цін нафти на промислове виробництво Малайзії, а також на значний спад рівня цін при шоку від нестабільності цін нафти. З монетарної точки зору, Центральний банк Малайзії здійснює експансивну грошову політику у відповідь на нестабільність цін нафти. Аналіз розкладу варіації підтверджує, що волатильність цін нафти є другим найважливішим фактором, який пояснює зміни у промисловому виробництві після власних шоків.

У статті Sek та Lim, 2016 [30], досліджується вплив шоків цін нафти на визначення внутрішньої інфляції в двох групах країн, а саме: 10 країн, що імпортують нафту, порівняно з 10-ма країнами, що експортують нафту з 1973 по 2015 рік. Зокрема, розглядаються ефекти шоків у виробництві та попиті на нафту на визначення внутрішньої інфляції. В аналізі використовується структурна модель векторної авторегресії (SVAR) для аналізу впливу ортогоналізованих шоків на інфляцію. Ідентифікація Бланчарда-Ква використовується для створення матриці впливу на довгостроковому періоді. Результати виявляють взаємодію між змінними та показують, що шок виробництва нафти має більший вплив на інфляцію, порівняно з шоками в попиті на нафту. В специфікаціях моделі SVAR автори посилаються на критерії Акаїке та Шварца при виборі довжини запізнень. Результати дослідження показують, що інфляція отримує вплив від змінних нафти та обмінного курсу, але ці впливи дуже малі у короткостроковій перспективі. Також є взаємодія між чотирма змінними в короткостроковій перспективі з відносно невеликим впливом один на одного, за винятком зміни в попиті на нафту. Зміни в попиті на нафту більш чутливі до впливу шоків (зміни в обсязі виробництва нафти, зміни в обмінному курсі та інфляція).

Переходячи до факторних моделей, спершу варто зазначити роботи в контексті виділення факторів в цілому. До такої роботи відноситься стаття Dai та ін., 2021 [31], де автори конструюють глобальний індекс економічної невизначеності політики шляхом застосування аналізу головних компонентів

для двадцяти основних економік світу. Встановлено, що PCA-заснований глобальний індекс економічної невизначеності політики є добрим показником економічної невизначеності політики на глобальному рівні, що досить узгоджено з ВВП-зваженим глобальним індексом економічної невизначеності політики. Виявлено, що PCA-заснований індекс економічної невизначеності політики позитивно пов'язаний з волатильністю і кореляцією глобального фінансового ринку, що свідчить про те, що акції стають більш волатильними та корельованими, коли економічна невизначеність політики на глобальному рівні вища. Дослідження також використовує емпіричний аналіз для підтвердження відповідності між побудованим індексом економічної невизначеності політики та показниками волатильності та кореляції глобального фінансового ринку, себто підтвердження результатів більш стандартними статистичними методами. Таким чином, стаття не лише конструює новий індекс, але й досліджує його вплив на глобальні фінансові ринки за допомогою емпіричного аналізу й показує важливість подібного підходу.

Перейдемо до FAVAR моделей. У статті Laine, 2020 [32], досліджується ефект конвенційної грошової політики Європейського центрального банку (ЄЦБ) на реальну економіку. Досліджується, зокрема, як неочікувані зміни у відсотковій ставці політики ЄЦБ впливають на рівень безробіття та промислове виробництво. Виявлено, що ефект грошової політики на безробіття та промислове виробництво є сильним і статистично значущим на основі даних з січня 1999 року по липень 2017 року. Однак після початку кризи реакції виявляються дуже слабкими і іноді статистично незначущими, що свідчить про те, що ефект конвенційної грошової політики ЄЦБ став слабшим після фінансової кризи. Цей висновок надзвичайно цікавий, оскільки можна було б припустити як слабший, так і сильніший ефект на підставі економічної теорії. З метою аналізу можливих змін у ефективності грошової політики застосовуються моделі векторних авторегресій з розширеним факторним навантаженням (FAVAR моделі), запропоновані Bernanke та ін., в 2005 році [33]. Це дозволяє оцінити вплив грошової політики на велику кількість макроекономічних

змінних. Крім того, великий набір інформації забезпечує більш надійну ідентифікацію шоку грошової політики, оскільки центральні банки фактично спостерігають сотні часових рядів в реальності. Результати суперечать ранішній літературі щодо ефектів конвенційної грошової політики ЄЦБ, проте це найімовірніше є результатом широких можливостей для врахування великої кількості інформації та даних використанням факторної версії VAR моделі, аніж суперечливістями з загальною теорією.

Ці методології також зустрічаються й в українській літературі. У статті, написаній Grui, Lysenko, 2017 [34], пропонується підхід для прогнозування поточного значення квартального ВВП України. Запропонований підхід використовує провідні показники з різною частотою оприлюднення. Дані з набору пояснювальних змінних узагальнюються у декілька факторів за допомогою аналізу головних компонентів, і на їх основі оцінюється модель векторних авторегресій з розширеним факторним навантаженням (FAVAR). Система враховує нові дані по мірі їх оприлюднення протягом кварталу для коригування прогнозів ВВП. Крім того, досліджується вплив окремих публікацій даних на точність прогнозів. ВВП оприлюднюється значно пізніше після завершення кварталу, тоді як прийняття рішень щодо економічної політики вимагає інформації в реальному часі про поточний стан економіки. Міжнародна практика показала можливість отримання такої інформації за допомогою так званих моделей "nowcasting", які дозволяють оцінити стан економіки до оприлюднення офіційних даних. У сфері макроекономічного прогнозування може бути велика кількість можливих пояснювальних змінних. Факторний аналіз дозволяє виокремити основні чинники варіації серед набору змінних. Таким чином, менша кількість оцінених факторів може узагальнити значну кількість інформації з великої системи. Факторний аналіз може відкидати власні шоки змінної, які не мають впливу на загальні тенденції у системі і модель не змушена реагувати на шум.

QPM та DSGE є останнім важливим традиційним типом моделей, що надактивно використовуються для економічного аналізу. Вони представляють

більш складні математичні моделі, які враховують взаємозв'язки між різними аспектами економічної системи та їх динаміку в часі. Це велика система рівнянь, іноді більше сотні, що поєднує дані з теорією використовуючи систему залишків в рівняннях. Ключова відмінність від VAR моделей полягає в тому, що через величину моделей та занадто велику кількість коефіцієнтів, оцінювання коефіцієнтів, калібрація, робиться експертними або баєсовими методами. Також модель має більш грунтовні теоретичні основи, на яких будуються системи рівнянь. Насамкінець, часто в QPM та DSGE моделях використовуються методи фільтрів Кальмана [35] для оцінки неспостережних змінних, таких як тренди та розриви величин, оскільки дослідження трендів є важливою задачею в економічному аналізі з теоретичної точки зору, а короткострокові розриви можуть відхилятися від нуля через несподівані та непояснювані шуми в моменті, природа яких може бути як кризовою, так і поведінковою.

У статті, написаній Benes et al., 2017 [36], описано ключові особливості виробничої версії квартальної моделі прогнозування (QPM), яка є перспективною відкритою економічною моделлю прогнозування, налаштованою для представлення індійського випадку, для генерації прогнозів та оцінки ризиків, а також проведення політичного аналізу. QPM враховує кілька важливих особливостей для Індії, таких як важливість сільського господарства та цін на продукти харчування в процесі інфляції; особливості передачі грошової політики та наслідки ендогенного процесу вірогідності для формулювання грошової політики. У статті також описано ключові властивості та історичні декомпозиції деяких важливих макроекономічних змінних. Оскільки існують затримки в ефекті грошової політики та компроміси між досягненням цілі інфляції та стабілізації зростання виробництва, успішне впровадження таргетування інфляції потребує надійних прогнозів середньострокового періоду та деяких знань про те, як дії з політики впливають на цільові змінні інфляції та виробництва. Основний інструмент політики на практиці - це ставка центрального банку, і вона має вплив на виробництво та інфляцію через складний механізм трансмісії, що включає більш довгострокові процентні

ставки, обмінний курс, очікування домогосподарств і ринків. Таким чином, режим інфляційного таргетування ґрунтується на прогнозах та аналізі політики, які належним чином враховують відповідні зв'язки. В такому сценарії макроекономічні моделі QPM допомагають політикам зібрати своє розуміння економіки і структурувати свої думки, обговорення та вправи прогнозування. Вони надають систематичну рамку для характеризації та аналізу ризиків навколо будь-якого умовного базового шляху прогнозування ключових макроекономічних змінних та їхніх політичних наслідків.

В той же час, Grui and Vdovychenko, 2019 [37], описали Квартальну Проекційну Модель (QPM), яку використовує Національний Банк України для здійснення регулярних макроекономічних прогнозів та рекомендацій щодо грошової політики. Модель є напівструктурним відображенням відкритої економіки. Вона захоплює механізм передачі грошової політики в контексті української економіки. Серед ключових особливостей економіки слід відзначити дезінфляційну програму, гетерогенні ціни, недосконалу вірогідність грошової політики, велику відкритість та доларизацію. Основним інструментом політики для досягнення мети у 5% інфляції є короткострокова процентна ставка. Крім того, НБУ здійснює валютні інтервенції з метою пом'якшення ексцесивної волатильності обмінного курсу та накопичення резервів. У той же час, зберігається плаваючий обмінний курс. Під ІТ НБУ прагне забезпечити якість прив'язування інфляції та очікувань щодо неї, покращуючи прозорість своєї політики та надаючи громадськості обгрунтування відносно стану грошової політики. Оскільки існують затримки у передачі грошової політики між прийняттям рішень та їхнім впливом на інфляцію, виникає необхідність у прогнозуванні середньострокового періоду для чіткого визначення того, як конкретні рішення політики впливають на майбутній розвиток економіки. Крім того, прогнозування допомагає оцінити, як різні політичні дії впливають на майбутню інфляцію. Механізм впливу, відомий як механізм передачі, є складним. Це вимагає структурованого мислення з боку політиків та розуміння економіки, в чому сильно допомогає модель QPM.

Переходячи до DSGE моделей, у статті Banerjee та Basu, 2015 [38], вводиться базова модель динамічної стохастичної загальної рівноваги (DSGE), яка використовується для макроекономічного аналізу як у наукових, так і в політичних колах. З метою розширення дослідницької потужності в макроекономіці, автори пропонують базову DSGE-модель для індійської економіки. У цьому документі автори роблять два внески. По-перше, вони досліджують емпіричні регулярності індійського бізнес-циклу та встановлюють кілька стилізованих фактів. По-друге, вони розробляють базову DSGE-модель, яка може слугувати аналітичною рамкою для розуміння цих стилізованих фактів. Модель має ознаки малої відкритої економіки з чітким розрізненням між секторами споживання та інвестиційних товарів. Модель симулюється з вірогідною параметризацією, заснованою на літературі по DSGE моделям. Результати демонструють, що базова модель достатньо точно відтворює стилізовані факти. До того ж, розглядаються особливості індійського бізнес-циклу, такі як зворотно-циклічний рух інфляції, проциклічний рух споживання та інвестицій.

Українська економіка досліджується Antonova, 2018 [39], монетарною DSGE-модель для вивчення впливу підрахунку зарплати в економіці, що характеризується режимом мінімальної заробітної плати, на макроекономічну реакцію на збільшення мінімальної заробітної плати. Основний результат полягає в тому, що при вищому рівні заробітної плати економіка менш реагує на шок мінімальної заробітної плати. Кількісно величина реакції на шок мінімальної заробітної плати залежить від частки нерикардіанських домогосподарств, тобто домогосподарств, які не мають доступу до фінансових ринків і відповідно споживають всі свої доходи кожного періоду. Дослідження присвячене вивченню агрегованих ефектів збільшення мінімальної заробітної плати в економіці, де відбувається підрахунок заробітної плати, та, зокрема, відповіді на таке питання: яку роль відіграє присутність та ступінь підрахунку у формуванні макроекономічної реакції на збільшення мінімальної заробітної плати? Модель розширена у трьох напрямках. По-перше, додана базова

гетерогенність праці: низькокваліфікована праця та висококваліфікована праця. Низькокваліфікована праця ближча до мінімальної заробітної плати. Друге розширення базової моделі дозволяє наявність двох типів домогосподарств: рикардіанських та нерикардіанських. Рикардіанські домогосподарства мають доступ до ринків капіталу та фінансових ринків і, отже, можуть займатися міжчасовим розподілом споживання. Нерикардіанські домогосподарства відсічені від фінансових ринків і, відповідно, кожен період споживають всі свої вільні доходи. Третє розширення полягає у включенні стимулів до підрахунку у модель. Припускається, що висококваліфікована праця може надаватися як формально так і неформально - тобто лише з відомістю про мінімальну заробітну плату перед податковими органами. Усього, в економіці з високим рівнем підрахунку заробітної плати негативний ефект збільшення мінімальної заробітної плати менший порівняно з економікою з низьким рівнем підрахунку.

Усі ці моделі використовуються для аналізу та прогнозування економічних процесів з різних точок зору, спираючись на математичні та комп'ютерні методи для обробки та аналізу великих обсягів даних. Інтеграція комп'ютерних наук та математики у розробці та застосуванні цих моделей дозволяє суттєво поліпшувати результати подібних моделей й вони будуть описані в наступному під-розділі.

1.2. Використання алгоритмів машинного навчання в економічній сфері

В попередньому розділі ми розглянули більш традиційні методи прогнозування та оцінки в економічному секторі. В цьому подивимося на більш нові алгоритми машинного навчання та як саме їх пропонується використовувати, які приклади застосування та особливості цього процесу на прикладі низки статей. Використання алгоритмів машинного навчання, таких як Random Forest, XGBoost, SVM, в економіці відкриває широкі можливості для прогнозування та аналізу економічних явищ [40]. Ці алгоритми, які базуються на ідеях статистики та штучного інтелекту, надають економістам засоби для моделювання складних економічних залежностей та створення точніших прогнозів на основі великих обсягів даних.

Однією з переваг використання алгоритмів машинного навчання є їх здатність автоматично виявляти складні залежності та взаємозв'язки між різними економічними факторами. Незважаючи на всі переваги, алгоритми машинного навчання також мають свої обмеження, що цілком відповідають тим, що з'являються в економетричних моделях. Наприклад, вони можуть виявитися чутливими до шуму в даних. Також, як і в традиційних методах, важливо правильно налаштувати параметри моделей машинного навчання для досягнення оптимальних результатів [41]. Зазвичай це і є одним з найголовніших завдань після вибору моделі або набору моделей. Тому давайте перейдемо до прикладів статей.

Biau and D'Elia, 2010 [42], були одними з перших хто запропонував використання такого нового підходу до прогнозування макроекономічних агрегатів, як Random Forest, спочатку розробленої як засіб класифікації для навчання. Використання алгоритму у звичайних економічних дослідженнях є рідкісним станом на 2010 рік, тому ця стаття досліджує потенційне застосування цієї техніки для моделювання та прогнозування макроекономічних агрегатів за допомогою великих наборів даних з опитувальних змінних (один з найкласичніших варіантів для створення великорозмірних баз даних в економічному секторі, окрім дезагрегації). Автори будують модель саме для прогнозування ВВП на короткостроковий період у Єврозоні з використанням гармонізованого набору даних опитування Європейського союзу про бізнес та споживачів. Цей підхід досліджується з двома цілями: отримання попереднього непараметричного прогнозу зростання ВВП та аналіз низки кандидатних пояснювальних змінних для розрізнення між тими, які значно сприяють поясненню та передбаченню аналізованого явища, та тими, які в основному додають випадковий шум. Індекс важливості змінних має перевагу у виборі важливих змінних незалежно від будь-яких функціональних та розподільчих припущень, що робить його надійним та автоматизованим інструментом для вибору змінних, що покладається виключно на статистику та комп'ютерні науки. Після цього обираються вибрані змінні для побудови лінійної моделі. Результати

порівнюються як з результатами авторегресійної моделі (взятої як базова), так і з квартальними експертними прогнозами єврозони економічного прогнозу, що видаються спільно трьома основними європейськими економічними інститутами й Random Forest показує високу якість.

В роботі Mei et al., 2022 [43], основна увага зосереджена на прогнозуванні цін в реальному часі на ринку електроенергії в Нью-Йорку за допомогою методу випадкового лісу. Точний прогноз вважається найбільш практичним шляхом отримання перемоги в торгівлі електроенергією в, станом на 2022 рік, дуже конкурентному ринку електроенергії США. Модель може адаптуватися до останніх умов прогнозування, тобто останніх кліматичних, сезонних та ринкових умов, шляхом оновлення параметрів випадкового лісу з новими спостереженнями й додатковими даними в режимі, практично, реального часу. Ця адаптивність дозволяє уникнути проблем у випадку оновлення кліматичних умов, зокрема в рамках глобального потепління. Загалом, в літературі поточні методи для цієї задачі можна розділити на кілька груп, включаючи часові ряди (ARIMA та GARCH) та машинне навчання (ANN, SVM). Згадується про різні підходи до прогнозування цін, такі як метод ARIMA, GARCH та методи, засновані на нейромережах, такі як ANN та SVM. Особлива увага приділяється підходам, заснованим на машинному навчанні, зокрема методу випадкового лісу. Зокрема, висвітлюється робота Fan та Mao, які пропонують новий метод прогнозування цін на короткостроковий період на основі гібридної мережі самоорганізації (SOM) та методу опорних векторів (SVM). Загальні результати роботи показують, що підхід має середню абсолютну похибку прогнозування приблизно 10,24% й це краще за інші методи, що підкріплює тезу про високі можливості метода випадкових лісів.

У статті Tyralis and Papacharalampous, 2017 [44], акцентується на оцінці ефективності випадкового лісу в одноступеневому прогнозуванні за допомогою двох великих наборів даних коротких часових рядів з метою запропонувати оптимальний набір змінних для прогнозу. Перший набір даних складається з 16 000 симульованих часових рядів різних моделей авторегресії зі стохастичним

інтегруванням ковзаючого середнього (ARFIMA). Другий набір даних складається з 135 часових рядів середньорічних температур. Найвищу прогностичну ефективність RF спостерігається при використанні невеликої кількості останніх змінних з різними лагами. Цей результат може бути корисним у відповідних майбутніх застосуваннях з можливістю досягнення вищої точності прогнозування. Автори відзначають що використання цього алгоритму в прогнозуванні часових рядів залишалося малодослідженим. Проте автори все одно розглядають цей метод, а також процедури підбору оптимального набору параметрів для алгоритму, що враховують його ефективність залежно від обраних змінних. Зазначається, що у звичайних задачах регресії використовується вибірка спостережень залежної змінної та відповідних змінних-прогнозувальників для навчання моделі. Проте, у прогнозуванні часових рядів змінні-прогнозувальники представлені попередніми значеннями самого ряду, а також «допоміжних» рядів. Результати дослідження вказують на те, що RF добре справляється з одноступеневим прогнозуванням коротких часових рядів, особливо при використанні невеликої кількості останніх лагових змінних. Цей результат є важливим опорним пунктом для подальшої вибудови альтернативних методів, представлених в цій дисертаційній роботі.

У статті Gawthorpe, 2021 [45], проводиться оцінка можливості використання випадкового лісу для прогнозування економіки Чехії. Раніше дослідження показали потенціал випадкового лісу для надання попереджень про рецесію та виявили його конкурентоспроможність порівняно зі старішими моделями прогнозування. Оцінка моделі випадкового лісу ґрунтується на результативності передбачень на один крок вперед на чеських даних та їх перевагах порівняно з експертними прогнозами. Аналіз важливості змінних додатково підкреслює значення "м'яких" змінних як цінних детермінант для прогнозування в Чехії. Хоча цей гнучкий та потужний метод не був відомий в економіці до недавнього часу, ця стаття є черговим підтвердженням перевершення випадковим лісом таких моделей як Баєсівські, загальні лінійні моделі, прості дерева рішень.

Робота Shen та ін., 2021 [46], пропонує новий підхід до прогнозування обмінних курсів, який називається FSPSOSVR. Цей підхід поєднує оптимізацію рою частинок (PSO), відбір ознак випадковим лісом і регресію опорних векторів (SVR). PSO використовується для отримання оптимальних параметрів SVR для прогнозування обмінних курсів. Аналіз включає щомісячні обмінні курси з січня 1971 року по грудень 2017 року семи країн, включаючи Австралію, Канаду, Китай, Європейський союз, Японію, Тайвань і Велику Британію. Результати емпіричних досліджень показують, що алгоритм FSPSOSVR стабільно демонструє відмінну точність прогнозування порівняно з конкуруючими моделями для всіх валют. Результати свідчать про те, що запропонований алгоритм є перспективним методом для емпіричного прогнозування обмінних курсів. Крім того, в роботі демонструється емпірична актуальність прогнозів обмінних курсів, отриманих за допомогою FSPSOSVR, за допомогою здійснення обмінних операцій із заборгованістю в іноземній валюті. За результатами дослідження було виявлено, що запропоновані торгові стратегії можуть забезпечити позитивний додатковий дохід більше ніж 3% на рік для більшості валют, за винятком AUD і NTD. Ця робота є прикладом комбінованого підходу з низки моделей, де комбінація показує кращі результати аніж моделі окремо одна від одної, що також є одним з поінтів дисертаційної роботи.

Ramakrishnan та ін., 2017 [47], досліджують динаміку взаємодії цін на чотири товари та обмінного курсу для Малайзії. Зазначено суперечливі твердження щодо точності передбачення обмінного курсу, тому в цій статті пропонується нова методологія порівняльного аналізу трьох технік машинного навчання: метод опорних векторів, нейронні мережі та випадковий ліс. Експериментальні результати показують, що випадковий ліс є порівняно кращим за метод опорних векторів та нейронні мережі щодо точності та продуктивності. Крім того, стаття виявляє, що ціни на специфічні товари в Малайзії - нафта, пальмова олія, гума та золото, - є сильними динамічними параметрами, які впливають на обмінний курс країни. Отже, ці результати корисні для прийняття

рішень у сфері політики, моделювання інвестицій та корпоративного планування.

Нарешті, перейдемо до методу XGBoost. Hu, Song, 2019 [48], аналізують та оцінюють успішність студентів. Для об'єктивної оцінки академічних досягнень студентів та розробки моделі оцінки використовується алгоритм XGBoost, цікавий аналог методології випадкового лісу з іншим підходом до навчання лісів. Для прогнозування успішності студентів у невідкритих курсах використовується метод XGBoost на основі результатів закритих курсів. Дослідження враховує такі фактори, як різниця між звичайними оцінками та оцінками за тестами, різні спеціальності, можливість відсутності оцінок через порушення дисципліни або зміни в навчальних планах, а також зміни в навчальних планах для студентів різних курсів. Висновки статті показують, що алгоритм XGBoost дозволяє ефективно класифікувати та прогнозувати результати студентів. Це дозволяє встановити зв'язок між курсами та передбачити успішність студентів у майбутньому. Крім того, модель XGBoost має високу швидкість роботи, точність та ефективно використовує ресурси.

Ближче до, власне, економічного застосування алгоритму, Massaro та ін., 2021 [49], досліджують XGBoost для прогнозування продажів товарів у секторі великих роздрібних мереж. Цей алгоритм застосовується як метод навчаного контролю для прогнозування продажів товарів за різних умов акцій та мультипараметричного аналізу. Для підвищення точності прогнозів та у випадках обмежених обсягів даних використовується техніка Augmented Data (AD). Основні методологічні аспекти полягають у використанні XGBoost для прогнозування продажів та використанні техніки AD для підвищення точності прогнозу в умовах обмежених даних. Модель враховує поведінку різних сегментів покупців та дозволяє персоналізовані послуги з урахуванням їхньої покупної поведінки. Автори підкреслюють успішність застосування XGBoost для прогнозування продажів у великих роздрібних мережах, а також важливість техніки Augmented Data для підвищення точності прогнозів у випадках обмежених даних. Аугментація даних є ще одним об’єктом, який дозволяє

поліпшувати прогностичні властивості моделей в цій дисертаційній роботі й стаття непогано вказує на додаткові можливості в цій сфері.

Черговим прикладом використання XGBoost є прогнозування щорічного виробництва рису в Бангладеші, описане Noorunnahar та ін., 2023 [50]. Робота порівнює продуктивність ARIMA та XGBoost моделей. У вступі автори наголошують на важливості точного та своєчасного прогнозування виробництва сільськогосподарської продукції для забезпечення продовольчої безпеки та адміністративного планування. Рис є ключовою культурою в Бангладеші, і точний прогноз виробництва рису має велике значення для економіки країни, тому що він впливає на ВВП, інфляцію, зайнятість, безпеку харчування та боротьбу з бідністю. За висновками авторів, XGBoost зазвичай, і в цьому випадку також, демонструє кращі результати в порівнянні з ARIMA, особливо коли є багато даних та складна структура. Тому, автори використовують модель XGBoost для короткострокового прогнозування виробництва рису на наступні 10 років в Бангладеші. Прогноз показує збільшення обсягу виробництва рису в країні в майбутні роки.

Також хочеться згадати про SVM модель, це більш традиційна модель в сфері фінансів й статті на тему з’являлись ще у 2000-х. Наприклад, Huang та ін., 2005 [51], досліджують прогнозування напрямку фінансових рухів за допомогою методу опорних векторів (SVM). Автори порівнюють продуктивність SVM з іншими методами класифікації. Досліджено прогнозування щотижневого напрямку індексу NIKKEI 225. Фінансовий ринок - це складна, еволюційна та нелінійна динамічна система, характеризується великою кількістю факторів, що взаємодіють, і високим ступенем невизначеності. SVM вирізняється здатністю керування потужністю функції прийняття рішень, використанням ядерних функцій та розрідженістю розв'язку. Алгоритм базується на унікальній теорії мінімізації структурного ризику, що дозволяє йому досягати високої узагальненої точності та уникати проблеми перенавчання. Результати показують, що SVM має найвищу точність серед інших методів класифікації. Вони пояснюють це тим, що SVM мінімізує структурний ризик, що робить його

менш вразливим до перенавчання. Також виявлено, що комбінована модель, яка поєднує SVM з іншими методами класифікації, показує найкращу продуктивність серед усіх методів прогнозування.

Ще однією «старою» статтею є робота Sao, Tau, 2003 [52], де також досліджується застосування SVM у прогнозуванні фінансових часових рядів. Запропоновано адаптивні параметри, враховуючи нестаціонарність фінансових часових рядів, оскільки фінансові часові ряди характеризуються шумом та нестаціонарністю, що вимагає уваги до вибору алгоритму навчання. Автори відзначають, що SVM, розвинутий у 1995 році, використовуючи принцип структурної мінімізації ризику, має кращу узагальнювальну продуктивність, ніж традиційні та надпрості (на той час) нейронні мережі. Ключовою особливістю SVM є унікальний принцип розв'язання квадратичного програмування з лінійно обмеженою проблемою, що дає оптимальне рішення. Однак, недоліком SVM є час навчання, який масштабується між квадратичним і кубічним щодо кількості зразків. Стаття розглядає застосування SVM у прогнозуванні фінансових часових рядів, порівнюючи його з нейронними мережами. Результати показують, що SVM виявляється кращим у прогнозуванні фінансових рядів порівняно. Вплив вільних параметрів на продуктивність SVM і показують, що їх правильний вибір може покращити узагальнювальну продуктивність та зменшити кількість опорних векторів.

Цей розділ узагальнює широке використання алгоритмів машинного навчання, зокрема Random Forest, XGBoost, SVM задля оцінок та прогнозування часових рядів в економічному секторі. Видно тенденцію на використання ближче до поточного часу, проте є й статті за 2000-х роках, більшою мірою про SVM алгоритми та алгоритми що на ньому базуються. Більшість авторів відзначають суттєвіші можливості для підхоплення нелінійних ефектів, а також, більшою мірою, кращі результати цих нових алгоритмів за більш традиційні, описані в минулому розділі.

1.3. Застосування нейромережевих архітектур для вирішення економічних задач

Розвиток використання нейронних мереж у економічному аналізі та прогнозуванні є актуальною та перспективною тенденцією в сучасній науці, що розвивається паралельно до сфери комп’ютерних наук, де постійно з’являються нові та покращені архітектури, методи оцінки коефіцієнтів, вирішення найрізноманітніших проблем, починаючи від стандартних проблем перенавчання або зникаючого градієнту, закінчуючи специфічними проблемами в теорії сигналів [53]. Нейронні мережі, зокрема прості, глибокі та рекурентні, знаходять широке застосування у вирішенні завдань економічного аналізу та прогнозування.

Прості нейронні мережі використовуються для моделювання простих економічних взаємозв'язків, таких як попит та пропозиція на ринку, та можуть бути застосовані для прогнозування тенденцій у розвитку ринку ще з 2000-х років, де архітектури та методи оцінки максимально примітивні відносно сучасних моделей. Глибокі нейронні мережі виявляються ефективними у вирішенні складних економічних задач, таких як прогнозування валютного курсу, моделювання фінансових ринків та аналіз фінансових часових рядів «в моменті». Вони можуть автоматично виявляти складні нелінійні зв'язки та враховувати велику кількість вхідних факторів, що робить їх ефективними інструментами для оцінок в економічній сфері, для наукастів. Рекурентні нейронні мережі дозволяють моделювати динаміку часових рядів, таких як зміни ВВП, інфляція чи безробітність. Вони враховують динаміку даних (як власних, так і допоміжних чи екзогенних) у часі і можуть успішно передбачати майбутні тенденції на основі минулих даних.

Ми розпочнемо з прогнозування інфляції, однієї з найпопулярніших вправ прогнозування. Nakamura, 2005 [54], написав фундаментальну статтю що стала одним з перших кроків використання нейронних мереж для подібних задач. У ті часи нейронні мережі не були популярним інструментом, тому кількість публікацій була обмеженою. Автор використовує щоквартальні дані з 1960 по

2003 рік. Його нейронна мережа є простою, з лише двома парами уніваріатних рівнянь, з'єднаних одне з одним. Метод пошуку найкращих коефіцієнтів суттєво відрізняється від сучасного, використовуючи сто випадкових початкових значень і вибираючи серед них, замість оптимізації з використанням зворотного розповсюдження помилки. Навіть із таким примітивним підходом нейронна мережа показала кращу ефективність, ніж авторегресійна (AR) модель на горизонті прогнозування від 1 до 4 кварталів вперед. Це випливає з здатності нейронних мереж захоплювати нелінійність. Головним висновком є те, що нейронні мережі можуть бути гарним доповненням до пулу вже використовуваних моделей прогнозування. У випадку дисертаційної роботи, ця стаття є першою спробою використання алгоритму, що у підсумку буде ключовим в цій роботі.

Складність мереж і методів зростає разом із популярністю галузі. Choudhary та Haider, 2012 [55], демонструють результати роботи нейронної мережі на різних наборах даних країн і порівнюють їх з AR(1). У цій роботі використані більш складні архітектури, ніж у попередній: дві мережі отримали назви гібридна мережа та динамічна мережа (остання більш схожа на рекурентну нейронну мережу (RNN), але простіша), за якими слідує їх комбінація. У результаті використання бази даних із щомісячною інфляцією з 07.1991 по 06.2008 для 28 країн ОЕСР, нейронна мережа переважно перевершує модель AR(1) в короткостроковому прогнозуванні. Автори стверджують, що постійне порівняння економетричних та інших моделей є бажаною стратегією через нестабільність результатів. Стаття показує динаміку переходу до більш цікавих нейронних мереж в контексті економічного аналізу та прогнозування.

Проте складність моделі - не єдина відмінність від стандартних методик. Світ на шляху до "Світу Великих Даних", де якість та обсяг даних зросли. Це впливає на процес прогнозування та прогностичні моделі. Medeiros та ін., 2018 [56], використовували дуже багатий місячний набір даних, названий FRED-MD, який містить сотні ознак для прогнозування інфляції у США. У роботі описано деякі моделі, від орієнтирів та традиційної економетрики до моделей науки про

дані. Перші алгоритми майже не можуть зафіксувати нелінійності, які можуть моделі машинного навчання, наприклад, взаємозв'язок між інфляцією та зайнятістю. Таким чином, нейромережеві моделі показали найкращі результати на більшості горизонтів прогнозування, тоді як регресії Ridge/Lasso також виступили добре. Більшість моделей згенерували побічний продукт: список ознак, які були обрані як найважливіші для пояснення варіації для кожного горизонту. Результати серед моделей були досить різними. Регресія Лассо виробила виведення та ціни як значущі змінні для пояснення інфляції. З свого боку, нейромережеві алгоритми та Рідж регресія містять зайнятість, ціни та відсоткову ставку. Існує багато простору для аналізу та порівняння результатів різних моделей. Тому завжди корисно розширювати асортимент використовуваних моделей, навіть коли вони використовують один і той самий набір змінних, щоб дослідити питання з різних перспектив.

Jung та ін., 2018 [57], наводять комплексний приклад використання кількох технік машинного навчання - зокрема, Elastic Net, Super Learner та RNN - для прогнозування зростання ВВП у кількох країнах. Основний інтерес полягає в порівнянні ефективності цих моделей з офіційними прогнозами WEO, які ґрунтуються на більш традиційних моделях. Elastic Net та Super Learner мають значно кращі рівні точності (від 35 до 80 відсотків вище, ніж базовий показник) на один квартал вперед. Проте на річній основі було набагато менше впевненості (RNN був кращий для США, Великої Британії та Німеччини, тоді як WEO - для Іспанії, Мексики та В'єтнаму). Ці алгоритми працюють добре для прогнозування на короткий термін та можуть бути корисними у випадках довгострокового прогнозування.

У статті Tkacz, 2001 [58], досліджується можливість покращення точності прогнозування фінансових та грошових показників Канади за допомогою моделей нейронних мереж. Нейронні мережі дозволяють отримати статистично значимо менші помилки прогнозування для річного темпу зростання реального ВВП у порівнянні з лінійними та одновимірними моделями. Однак такі покращення прогнозування менш помітні у випадку прогнозування квартального

зростання реального ВВП. Нейронні мережі не можуть перевершити наївну модель без змін. Більш виражені нелінійності на довших горизонтах прогнозування відповідають можливим асиметричним ефектам грошової політики на реальну економіку.

У статті Longo та ін., 2022 [59], пропонується методика ансамблювання для прогнозування майбутнього зростання ВВП США. Підхід поєднує Рекурентну Нейронну Мережу (RNN) з динамічною факторною моделлю, що враховує зміну часу в середньому з загальним авторегресійним коефіцієнтом (DFM-GAS). Аналіз базується на наборі змінних, які охоплюють широкий спектр факторів, виміряних на різних частотах. Вправа з прогнозування спрямована на оцінку прогностичної здатності кожного компоненту моделі ансамблю, враховуючи зміни в середньому, потенційно викликані рецесіями, що впливають на економіку. Таким чином, ми показуємо, як поєднання RNN і DFM-GAS покращує прогнози зростання ВВП США після глобальної фінансової кризи 2008-2009 років. Автори зазначили, що прогнозування майбутнього стану економіки значно покращилось з моменту фінансової кризи 2008-2009 років завдяки наявності різних та різнорідних джерел даних з різними частотами Серед різних підходів нейронні мережі привертають найбільшу увагу науковців завдяки своєму природному застосуванню в контексті часових рядів. Цікаво, що статистичне навчання є більш корисним для передбачення макроекономічних показників у тих випадках, коли відбуваються структурні зміни економіки. Фактично, це наслідок так званої «Критики Лукаса». Це типовий випадок періодів надзвичайних економічних рецесій, таких як фінансова криза 2008-2009 років та недавня криза Covid-19. У зв'язку з цим більшість останніх досліджень спрямовані на покращення точності прогнозування в умовах рецесії шляхом пошуку альтернативних технік. Автори показують, що DFM-GAS завжди перевершує свого контрагента з фіксованим параметром. Також встановлено, що ансамбль нейронних мереж покращує результати прогнозування в розглянутому вікні, особливо для короткострокового прогнозного горизонту. Причиною великої різниці з короткостроковими прогнозами є те, що зміни в середньому

можна частково пояснити зсувом середнього, що викликає структурні зміни у процесі генерування даних. Саме тому ми використовуємо вікно поза вибіркою, де ми розглядаємо кризу 2008-2009 років, одночасно з використанням тесту Чоу, щоб оцінити, наскільки добре модель прогнозує під час структурних розривів.

В свою чергу, у статті Zhang, Berardi, 2001 [60], досліджується використання методів комбінування нейронних мереж у порівнянні з традиційною моделлю з вибором найкращої. Методи ансамблю застосовуються до проблеми прогнозування обмінного курсу. Пропонуються два загальні підходи до комбінування нейронних мереж: систематичні та послідовні методи розділення для побудови ансамблів нейронних мереж. Виявлено, що базовий підхід ансамблю, створений за допомогою незмінних архітектур мереж, навчаних за допомогою різних початкових випадкових ваг, не є ефективним у покращенні точності прогнозування, тоді як ансамблеві моделі, що складаються з різних структур нейронних мереж, можуть систематично перевершувати прогнози окремих `найкращих' мереж. Результати також показують, що нейронні ансамблі, побудовані на основі різних розділень даних, ефективніші, ніж ті, що розроблені на основі повного навчального набору даних для прогнозування поза вибіркою. Більше того, зменшення кореляції між прогнозами, зробленими членами ансамблю, шляхом використання методів розділення даних є ключем до успіху для нейронних ансамблевих моделей. У статті також розглядаються обмеження традиційного підходу у виборі моделі нейронної мережі. Вказується, що кінцева мережа може не бути оптимальною через велику кількість факторів, які можуть впливати на навчання нейронної мережі та вибір моделі. Таким чином, може відбутися перенавчання на конкретних вибірках даних та модель не матиме здатності до узагальнення. У статті висвітлюється, що застосування нейронних ансамблів може вирішити ці проблеми, оскільки вони не ґрунтуються лише на продуктивності однієї окремої моделі мережі. Це ще раз підкреслює потужність комбінованих методів, зокрема що важливо для даної дисертаційної роботи.

Підсумовуючи розділ, нейромережеві моделі почали використовуватися та розроблятися в сфері макроекономічного аналізу ще на початку 2000-х років, проте в ті часи вони були простими та примітивними як в контексті архітектури, так і в контексті методів оцінок та оптимізації коефіцієнтів. Наразі використовуються більш просунуті структури, включаючи рекурентні нейронні мережі та моделі довгострокової та короткострокової пам’яті.

1.4. Особливості алгоритмів пошуку відстаней та вирішення кластеризаційних задач, їх використання в економічній сфері

Розвиток алгоритмів кластеризації та алгоритмів, які знаходять відстані між часовими рядами у економічному аналізі та прогнозуванні, є цікавими та не дуже розвиненими напрямками досліджень в сучасній науці, оскільки вони не є популярними в цілому. Проте це не через неефективність, а саме через специфічність підходу в контексті часових рядів, оскільки існує ряд проблематичних та не до кінця коректних підходів, що були піддані значній критиці у 2005-ому в роботі Keogh та Jin [61]. Тому література в цій темі відносно обмежена, тим не менш, розглянемо декілька робіт.

Алгоритми кластеризації дозволяють групувати схожі об'єкти в кластери на основі їх характеристик. Це допомагає ідентифікувати залежності та взаємозв'язки між економічними показниками, що в свою чергу сприяє удосконаленню моделей прогнозування та прийняттю обґрунтованих рішень в економічній сфері. Алгоритми, які знаходять відстані між часовими рядами, відіграють роль у роботі з великими базами даних, що налічують десятки чи, навіть, сотні показників однієї природи. Вони дозволяють порівнювати динаміку та визначати схожість. Це допомагає виявляти патерни та тенденції в цілому у системі.

Одна з традиційних робіт в тематиці написана Wolfson et al., 2004 [62], де розглядається використання кластерного аналізу як доповнення до методів регресійного аналізу для отримання подальшого поліпшення систематичного розуміння зв'язку між політикою, економікою та конфліктами. Автори вважають, що ці змінні утворюють частину ще не зрозумілої, нелінійної, часово залежної

інтерактивної системи. Кластерний аналіз використовується для класифікації об'єктів у групи та спрямований на пояснення на основі характеристик, що перетинаються, використовуючи крос-секційні дані за 1967, 1974, 1981, 1988 та 1995 роки. Аналіз ідентифікує кластери держав на основі ряду характеристик. Як і очікувалося у часово залежній системі, існують докази постійного кластеризування країн протягом років, а також докази змін. Кілька кластерів, таких як розвинені держави, є дуже стійкими і вказують на патерни, які слід досліджувати далі за допомогою регресійного аналізу. Себто ця стаття емпірично доводить класичне групування країн за певними ознаками (розвинена економіка чи така, що розвивається).

У статті, написаній на українських даних авторами Rashkovan, Pokidin, 2016 [63], досліджується кластеризація та ідентифікація шести різних банківських бізнес-моделей за допомогою карт Кохонена для самоорганізації. Автори показують, як ці моделі змінювалися під час кризи та приходять до висновку, що деякі з банківських моделей більш схильні до дефолту. Також аналізуються профілі ризиків банківських бізнес-моделей та робиться розрізнення між найбезпечнішими та найризикованішими. Зокрема, використовуються шість типів ризиків (рентабельність, кредитний, ліквідності, концентрації, позикодавства пов'язаних осіб та відмивання грошей), щоб побудувати ризикові карти для кожної бізнес-моделі. Методика, застосована в цій роботі, є ефективним та високоточним інструментом передбачення дефолту. Крім того, для покращення аналізу ризиків розроблена картографія ризиків на основі набору ризикових показників, які підтверджують попередні висновки стосовно ризикових зон кожної бізнес-моделі. Виявлено, що більшість банків, що зазнали дефолту, знаходилися в "ризиковій" зоні карти перед своїм банкрутством. Основна відмінність цієї роботи від попередніх полягає у використанні карт Кохонена для самоорганізації як інструменту кластеризації, що дозволяє не лише розділити дані на однорідні групи, а й мати дуже зручні можливості для візуалізації даних, а також інші функціональні можливості, такі

як аналіз траєкторій. Це дозволяє ефективно виявляти типи банківських бізнес-моделей та їх зміни в часі.

Переходячи від більш простих кластеризацій до технік пошуку дистанцій в економічному секторі, Franses, Wiemann, 2020 [64], використали техніку непараметричного динамічного вирівнювання часу (DTW) для дослідження подібностей у економічних часових рядів. DTW має важливі переваги, оскільки він усуває побоювання щодо заздалегідь визначеного фіксованого часового вирівнювання рядів. Наприклад, на відміну від інших методів, DTW може зафіксувати чергування між веденням і відставанням рядів. Автори проілюстрували можливості DTW на прикладі вивчення бізнес-циклів штатів США навколо Великої рецесії та виявили значні докази динамічного вирівнювання між штатами. Через аналіз кластерів додатково задокументовано відмінності у відновленні після рецесії в різних штатах. Відмінності між стандартним і підходом авторів полягають у здатності модифікованого DTW виявити змінні ведучі і відставані відносини в рамках часового вікна. Цей підхід дозволяє оцінити взаємне відношення між часовими рядами, що змінюється з часом, і може переходити з позитивних на від'ємні значення та навпаки. Підходи з літератури обмежені тим, що припускають, що коли один ряд веде інший, він робить це протягом усього часу. Проте автори вивчають співміщення, що ґрунтується на техніці DTW, де існує можливість динамічної зміни параметра, що визначає ведучий або відставаний зв'язок між рядами з часом.

У статті Rutkowska, Szyszko, 2021 [65], застосовується алгоритм динамічного вирівнювання часу з новим обмеженням вікна для оцінки інформаційного змісту очікувань споживачів щодо поточної та майбутньої інфляції. У емпіричному дослідженні автори охоплюють сім європейських країн і порівнюють результати DTW з результатами попередніх досліджень у цих економіках з використанням стандартної методології. Стандартна процедура оцінює гібридну специфікацію очікувань за економічною теорією та інтуїцією щодо інформаційного змісту прогнозів. Проте, його застосування може бути піддане сумніву через властивості часових рядів (очікування часто є

нестаціонарними) та стійкість результатів (вони залежать від використаного оцінювача та реагують на навіть невеликі зміни в періоді дослідження). DTW частково виправляє ці недоліки й дозволяє покращити результат розділення рядів.

Насамкінець, Smiech, 2014 [66], досліджує класифікацію часових рядів цін на товари у передкризовий та післякризовий періоди. Основна мета полягає в визначенні того, чи спостерігається спільний рух цін на товари у цих двох періодах. Аналіз базується на щомісячних даних в 1990-2014 рр. Методологічний підхід дослідження відрізняється від попередніх робіт за використанням методу динамічного вирівнювання часу. Динамічне вирівнювання часу дозволяє оцінювати схожість форм часових рядів, тобто відстань між ними, і виявляється доречним для аналізу спільного руху цін на товари. Зокрема, застосовуються три методи кластеризації: метод Уорда, метод повної ієрархічної кластеризації та метод розділення. Оцінка результатів проводиться за допомогою середньої ширини силуету, яка вимірює внутрішню кластеризацію. Отримані результати свідчать про те, що спільний рух цін на товари є більш помітним у передкризовий період, коли кластери є більш однорідними і складаються з товарів однієї категорії. В період після кризи кластери менш однорідні. Дослідження також показує, що деякі товари однієї категорії не завжди проявляють подібну динаміку цін.

Підсумовуючи розділ, існує низка статей що досліджують питання кластеризації та пошуку відстаней між часовими рядами в економічному контексті, проте кількість цих статей та глибина не настільки великі відносно використання нейромережевих алгоритмів чи алгоритмів машинного навчання, що вчергове підкреслює важливість теми дисертації для розширення літератури в цьому напрямку й вибудови покращених та адаптованих алгоритмів.

РОЗДІЛ 2. Моделі групування дезагрегованих компонент за схожістю їх динаміки

У сучасному світі обробка та аналіз часових рядів має значну роль для розуміння та прогнозування різних явищ. Одним із важливих завдань є поділ часових рядів на групи з подібними характеристиками за допомогою алгоритмів кластеризації. У цьому дослідженні ми розглянемо процес поділу часових рядів, зокрема індексів інфляції України, за допомогою алгоритмів, що ґрунтуються на визначенні відстаней та кластеризації.

Першим кроком у поділі часових рядів є визначення відстаней між ними. Для цього можна використовувати різні метрики відстаней, такі як евклідова відстань, кореляційно-заснована відстань, метрики Мінковського та динамічне вирівнювання часу.

Евклідова відстань є одним із найпоширеніших методів визначення відстаней між часовими рядами. Вона вимірює прямолінійну відстань між двома точками у просторі. Однак цей метод може бути чутливим до зміщень та масштабування даних.

Кореляційно-заснована відстань використовує коефіцієнт кореляції між двома часовими рядами для визначення їх відносної схожості. Цей підхід може бути корисним для виявлення сезонних або циклічних залежностей в даних.

Метрики Мінковського використовуються для вимірювання відстаней між точками у n-вимірному просторі. Вони включають евклідову відстань як частинний випадок та можуть бути корисними в разі, коли дані мають різний масштаб.

Динамічне вирівнювання часу (DTW) враховує нелинійні зсуви та зміни масштабу між часовими рядами. Цей метод особливо ефективний при аналізі часових рядів з різною швидкістю або фазою.

Одним з ключових аспектів цієї дисертаційної роботи є побудова адаптованого алгоритму динамічного вирівнювання часу, а саме його адаптації для випадку економічних часових рядів. Однією з особливостей таких рядів є періодичність, що є денною, місячною або квартальною. В основному цю

особливість можна підхоплювали низкою способів, і один із тих, що розглядаються в цій дисертаційній роботі, це використання маски, що обмежує матрицю шляху, ряди та їх відповідність одним календарним роком.

Після визначення відстаней між часовими рядами, ми можемо перейти до кластеризації, щоб поділити їх на групи з подібними характеристиками. Для цього можна використовувати різні алгоритми кластеризації, такі як K-Means, DBSCAN та ієрархічна кластеризація.

Алгоритм K-Means є одним із найпоширеніших методів кластеризації. Він розділяє дані на певну кількість кластерів, де кожен кластер представляється своїм центроїдом, а об'єкти призначаються кластерам відповідно до найближчого центроїда.

Алгоритм DBSCAN (Density-Based Spatial Clustering of Applications with Noise) кластеризує дані на основі їх щільності. Він визначає кластери, які складаються з об'єктів, що знаходяться в областях високої щільності, і враховує об'єкти-викиди, які не входять в жоден кластер.

Ієрархічна кластеризація будує дерево поділу, представляючи кластери на різних рівнях дерева. Цей метод може бути корисним для аналізу структури та відношень між кластерами.

У цьому дослідженні ми розглянемо процес поділу часових рядів за допомогою алгоритмів кластеризації. Починаючи з визначення відстаней між часовими рядами за допомогою різних метрик, таких як евклідова відстань, кореляційно-заснована відстань, метрики Мінковського та динамічне вирівнювання часу, ми перейшли до використання алгоритмів кластеризації, таких як K-Means, DBSCAN та ієрархічна кластеризація, для групування часових рядів за їхньою схожістю.

Цей підхід до аналізу часових рядів може мати різноманітні застосування, зокрема у фінансах, економіці, медицині та інших галузях. Наприклад, у фінансовому секторі цей підхід може бути використаний для аналізу динаміки цін на фондовому ринку або для виявлення сезонних змін в обсягах продажів товарів. В цій роботі будуть вказані приклади використання цього методу на базі

даних компонент інфляції, себто зміни цін на продукти харчування, одяг та взуття, послуги та інше. Детальніше це буде описано у розділі 4. Проте важливо пам’ятати що це не єлиний можливий варіант застосування низки цих алгоритмів й поточна робота створює внесок у літературу в цілому з точки зору вибудови алгоритму й дозволяє багато різноманітних застосувань.

В цьому розділі надалі описуватимуться описані раніше алгоритми більш детально, починаючи від алгоритмів пошуку відстаней між часовими рядами й закінчуючи алгоритмами групування.

2.1. Попередня програмна обробка даних з метою їх використання у відповідних алгоритмах

Існує низка методів попередньої підготовки даних до алгоритмів кластеризації та пошуку відстаней в контексті часових рядів. Часто дані необхідно готувати і для використання в подальшому в алгоритмах, нейромережевих алгоритмах, машинного навчання і інших. До цих методів підготовки відносяться і вирівнювання даних, їх нормалізація та стандартизація, і пошук та заповнення пропущених значень, інтерполяція та екстраполяція. Також сюди варто віднести знаходження сезонної компоненти відповідними алгоритмами. Це дозволяє суттєво покращити якість майбутнього аналізу результатів роботи моделей і загалом дослідження результатів. Також важливим може бути пошук аномалій в часових рядах і їх видалення або повне видалення рядів, котрі мають такі аномалії. Це важливо, тому що в економічному сегменті значну роль відіграє природа рядів. Інколи природа є неадекватною або такою, що є шумом за своєю суттю, і вона не додає прогностичних властивостей моделям. Це призводить до потреби видаляти такі шоки вручну, аналізувати дані і робити попередню обробку. В даному дослідженні робота не спирається на економічний аналіз, проте важливо згадати про подібні методи також для повнішої картини. Надалі опишемо можливі кроки підготовки даних, які використовуються в дисертаційній роботі.

Перш ніж застосовувати алгоритми, важливо вирівняти часові ряди за однаковими проміжками часу. Це допоможе уникнути проблем, пов'язаних з

різними часовими інтервалами та відсутніми даними. Важливим етапом підготовки даних для наукової роботи є процес вирівнювання часових рядів, який має вирішити проблеми, пов'язані з нерегулярними інтервалами часу та пропусками даних. Наприклад, при аналізі місячних даних, де записи почалися у 2007 році, але були відсутні в період з 2012 по 2016 рік, необхідно вирівняти ці дані для забезпечення їхньої подальшої аналізу та порівняння.

Математично процес вирівнювання можна виразити як:

Метод середнього:

$$x(t) = \left(\frac{1}{n}\right) * \sum(x_i)$$

де $x(t)$ - вирівняний часовий ряд, x_i - значення у вихідному ряді, n - кількість значень у вихідному ряді.

Метод найближчого сусіда:

$$x(t) = x_i$$

де $x(t)$ - вирівняний часовий ряд, x_i - значення найближчого сусіда у вихідному ряді.

Ці методи дозволяють ефективно підготувати дані для подальшого аналізу, зберігаючи важливу інформацію та мінімізуючи вплив пропусків та нерегулярностей у часових рядах.

Для зменшення шуму та виявлення трендів у часових рядах можна використовувати методи згладжування, такі як ковзне середнє або експоненціальне згладжування.

Експоненціальне згладжування:

$$x(t) = \alpha x_i + (1 - \alpha)x_{i-1}$$

де x(t) - вирівняний часовий ряд, x_i - поточне значення у вихідному ряді, x_{i-1} - попереднє вирівняне значення, α - коефіцієнт згладжування ($0 < \alpha < 1$).

Важливо виявити та видалити аномалії або викиди з даних, які можуть спотворювати результати кластеризації. Проте ці ідеї часто базуються на економічних особливостях часових рядів й є нерелевантними конкретно для цього дослідження. Хоча правило інтерквартильного розмаху має місце бути в

поточному дослідженні для виявлення аномалій, що далі можуть бути замінені на більш згладжені значення.

Правило інтерквартильного розмаху (Interquartile Range, IQR) є важливим інструментом для виявлення викидів у наборі даних [67]. Це статистичне правило використовується для визначення верхньої та нижньої межі нормальних значень у розподілі даних на основі їхнього інтерквартильного розмаху. Інтерквартильний розмах - це різниця між третім квартилем (Q3) та першим квартилем (Q1) у впорядкованому наборі даних.

Правило інтерквартильного розмаху визначає нормальні межі даних як $Q1-1.5\times IQR$ для нижньої межі та $Q3+1.5\times IQR$ для верхньої межі, де $IQR=Q3-Q1$. Це правило дозволяє виявити викиди у наборі даних, які зазвичай є значеннями, що відстоять від нормального розподілу та можуть вказувати на наявність нетипових або аномальних подій.

Наприклад, якщо у наборі даних про вартість товарів першому квартилю відповідає значення \$80, а третьому квартилю - \$100, то інтерквартильний розмах IQR = 100 - 80 = \$20. Тоді нижня межа нормальних значень буде 80 - 1.5*20 = \$50, а верхня межа - 100 + 1.5*20 = \$130. Значення, які виходять за межі \$50 та \$130, можуть бути визнані викидами та вимагати уваги при подальшому аналізі даних.

Для забезпечення порівнянності між різними часовими рядами можна використовувати нормалізацію даних, наприклад, шкалювання до діапазону [0, 1] або стандартизацію [68].

Нормалізація - це процес приведення значень вхідних даних до певного діапазону або стандартного розподілу. У контексті нейронних мереж та алгоритмів машинного навчання, нормалізація відіграє важливу роль у підготовці даних, оскільки допомагає зробити модель більш стійкою до шуму та ефективнішою у навчанні.

Одним із поширених методів нормалізації є мінімаксна нормалізація, де значення кожного ознаки масштабуються до певного діапазону, зазвичай від 0 до 1. Математично, цей процес можна виразити формулою:

$$x_{norm} = \frac{(x - x_{min})}{(x_{max} - x_{min})}$$

де x - вихідне значення, x_{min} та x_{max} - мінімальне та максимальне значення ознаки в наборі даних. Наприклад, якщо ми маємо набір даних з відповідями від 0 до 100, то мінімальне значення x_{min} = 0 і максимальне значення x_{max} = 100, і після мінімаксної нормалізації, значення будуть лежати в діапазоні від 0 до 1.

Іншим поширеним методом нормалізації є стандартизація, де значення ознаки масштабуються так, щоб мати середнє значення 0 та стандартне відхилення 1. Цей метод особливо корисний, коли ознаки мають різні масштаби. Математично, цей процес можна виразити формулою:

$$x_{std} = \frac{(x - \mu)}{\sigma}$$

де x - вихідне значення, μ - середнє значення ознаки, σ - стандартне відхилення. Нормалізація дозволяє алгоритмам машинного навчання швидше та ефективніше знаходити закономірності у даних, що робить їх більш ефективними у вирішенні різних завдань.

Однією з основних переваг нормалізації є те, що вона допомагає збільшити швидкість збіжності нейронних мереж. Нейронні мережі, які мають вхідні дані з різних діапазонів або розподілів, можуть демонструвати повільну збіжність або навіть розходження, оскільки нейрони можуть втрачати градієнт на швидко змінюючихся функціях активації. Нормалізація допомагає зменшити цю проблему, стабілізуючи градієнти та забезпечуючи більш прогнозовану та швидку збіжність моделі.

Крім того, нормалізація зменшує чутливість моделі до масштабування даних, що робить її більш універсальною та загальновикористованою. Це особливо важливо в контексті реальних даних, де вхідні ознаки можуть мати різні масштаби або одиниці вимірювання. Таким чином, нормалізація грає критичну роль у покращенні якості та швидкості навчання нейронних мереж та алгоритмів машинного навчання, роблячи їх більш ефективними та стійкими до впливу різних факторів.

Додавання нових ознак, наприклад екзогенних змінних, може поліпшити результати кластеризації показавши схожість рядів з цими змінними. Наприклад, обмінний курс корелює з інфляцією, можливо з певною затримкою. Додавання курсу в базу даних допомогло б об’єднати деякі ряди й також пояснити якісніше групу, що у підсумку виходить. Проблема цього підходу полягає в тому, що він базується на економічному підгрунті і є неуніверсальним методом покращення та обробки бази даних якщо метод буде використовуватися в інших сферах досліджень.

Сезонне вирівнювання є важливою складовою аналізу часових рядів, особливо у сфері економіки, фінансів та соціальних наук. Цей процес необхідний для виокремлення сезонних компонент з даних та отримання стабільних оцінок базових тенденцій або трендів. Сезонність - це періодичні коливання в даних, які повторюються з певною періодичністю, наприклад, щоквартально, щомісяця або щодня. Видалення сезонності дозволяє краще розуміти динаміку даних і здійснювати більш точний аналіз. Одним з класичних варіантів сезонного коригування, зокрема використовуваного в цій дисертаційній роботі, є алгоритм X12, запропонований й постійно адаптований в подальшому. Проте саме X12 є найбільш елегантним в контексті балансу між простотою та ефективністю [69].

Алгоритм X12 для сезонного вирівнювання є широко використовуваним методом у статистичному аналізі для виділення сезонної компоненти з часових рядів. Його основна мета - визначення трендів, сезонних змін та інших факторів, які впливають на часовий ряд. Алгоритм X12 базується на методі розкладу часового ряду на компоненти, такі як тренд, сезонність, циклічність та інші фактори, і подальшому аналізі цих компонент.

У математичному вигляді, алгоритм X12 можна представити наступним чином:

Розкладання на сезонні та циклічні компоненти:

$$Y_t = T_t + S_t + C_t + I_t + \varepsilon_t$$

де Y_t - значення часового ряду у момент часу t, T_t - трендова компонента, S_t - сезонна компонента, C_t - циклічна компонента, I_t - неперіодичні варіації, ε_t - помилка.

Першим етапом є аналіз і видалення сезонної компоненти. Сезонність визначається за допомогою методу скінченних різниць та коефіцієнтів ковзного середнього. Після видалення сезонної компоненти решта даних аналізується для виявлення інших факторів.

Другим етапом є аналіз і коригування тренду та інших компонент. Знайдений тренд може бути скоригований, щоб врахувати інші впливи, такі як циклічність чи інші фактори. Цей процес може включати застосування різних методів, таких як згладжування або регресійний аналіз.

Алгоритм X12 дозволяє ефективно аналізувати та вирівнювати часові ряди, виокремлюючи їх складові та дозволяючи проводити дальший аналіз даних. Його широке використання свідчить про його ефективність та значення в статистичному аналізі.

Після підготовки та трансформації даних можна переходити до застосування алгоритмів кластеризації. Важливо пам'ятати, що ефективність кластеризації залежить від правильного вибору методів підготовки та трансформації даних, а також від обраного алгоритму кластеризації. Ретельний аналіз та експерименти з різними підходами можуть допомогти досягти оптимальних результатів в поділі часових рядів на групи.

2.2. Математичні методи пошуку відстаней між часовими рядами

Визначення попарної відстані між низкою часових рядів є важливою задачею у багатьох областях, зокрема в аналізі часових рядів, машинному навчанні та прогнозуванні. Ця задача полягає у вимірюванні схожості чи відмінності між двома часовими рядами на основі їхніх значень у різні моменти часу. Незалежно від області застосування, визначення відстані між часовими рядами вимагає вирішення ряду ключових проблем та викликів.

Перш за все, для того щоб визначити відстань між двома часовими рядами, необхідно визначити метрику, яка відображатиме ступінь подібності або

відмінності між ними. Однією з найпоширеніших метрик є Евклідова відстань та подібні методи, що прийшли з геометрії, яка обчислюється як квадратний корінь з суми квадратів різниць між відповідними точками двох рядів.

Необхідно також враховувати особливості самого часового ряду, такі як сезонність, тренд та шум, при визначенні відстані між ними. Наприклад, врахування сезонності може вимагати застосування додаткових методів, які враховують цю характеристику часових рядів. Вищезазначена метрика Евклідової відстані неякісно впорується з цією задачею, тому потребує або використання інших алгоритмів описаних в минулому розділі, себто передобробки даних, або використання більш нішевих алгоритмів як Dynamic Time Warping, що описуватимуться в цьому розділі також.

Однією з основних проблем при визначенні відстані між часовими рядами є різнорідність даних. Часові ряди можуть мати різну частоту спостережень, пропуски в даних або навіть різну довжину. Це ускладнює порівняння рядів та може призводити до неточних результатів. Тому важливо розробляти методи, які можуть ефективно враховувати ці різниці у даних. Більшість цих проблем вирішується на етапі передобробки даних, проте деякі алгоритми (як зазначений вище Dynamic Time Warping) мають здатність незважати на ці пропуски. Для вирішення проблеми різнорідності даних також можна використовувати методи інтерполяції або ресемплінгу, щоб привести всі часові ряди до одного стандартного формату. Наприклад, можна використовувати методи інтерполяції, щоб заповнити пропуски в даних, або методи ресемплінгу, щоб зменшити частоту спостережень до єдиного рівня.

Щодо вибору підходящої метрики, важливо провести аналіз характеристик даних та врахувати конкретний контекст задачі. Наприклад, якщо важливо враховувати форму коливань часових рядів, може бути корисно використовувати метрики, які збільшують вагу для більших відмінностей між точками рядів. Усунення цих викликів дозволить ефективно визначати відстань між часовими рядами та отримувати більш точні результати аналізу, що може бути корисним у багатьох областях, включаючи прогнозування, класифікацію та кластеризацію.

Наприклад, в аналізі фінансових даних важливо визначити ступінь подібності між різними фінансовими інструментами для розуміння їхньої взаємодії та ризиків. У медичній сфері визначення відстані між часовими рядами може допомогти в ранньому виявленні патологій або моніторингу стану пацієнтів.

Надалі описуватимуться ключові методи пошуку відстаней між рядами за їх схожістю, а також буде представлений запропонований автором дисертації алгоритм що заточений для роботи з часовими рядами.

2.2.1. Геометричний метод

Відстань між двома часовими рядами розміру N, X(t) = {x(1), x(2), ..., x(N)} та Y(t) = {y(1), y(2),..., y(N)}, є довжиною шляху, який з'єднує пару точок. Ця відстань є мірою схожості [70]. Більша відстань вказує на меншу схожість і навпаки. Найбільш поширеним і простим методом вимірювання відстані в області класифікації є метод, що походить від відстані Мінковського, представлений в рівнянні, де він описаний як загальне рівняння як для евклідової відстані ($D_{Euclidean}$) так і для мангеттенської відстані ($D_{Manhattan}$):

$$D_{\text{Minkowski}(X(t),Y(t))} = \left(\sum_{t=1}^{N} |x_t - y_t|^p\right)^{\frac{1}{p}}$$

У випадку, коли p=1, рівняння представляє мангеттенську відстань. Якщо p=2, евклідова відстань легко обчислюється для часових рядів однакової довжини, див. наступне рівняння:

$$D_{\text{Euclidean}(X(t),Y(t))} = \sqrt{\sum_{t=1}^{N} |x_t - y_t|^2}$$

Проте, евклідова відстань має обмеження. Вона не дозволяє різну довжину послідовностей, різні частоти дискретизації, зсуви в часовій осі (навіть якщо ці часові ряди схожі між собою). Ці недоліки ускладнюють використання евклідової відстані безпосередньо для аналізу часових рядів, що особливо проявляється в аналізі економічних рядів для яких подібні проблеми є стандартними.

Проблема недооцінки різних частот і зсувів/розтягнень є одним із ключових аспектів чому геометричні відстані є не найкращим вибором, незважаючи на свою простоту. Це може бути не проблемою в певних сферах діяльності, наприклад в медицині, проте для економічних та фінансових рядів така ситуація може відігравати погану роль через значну залежність рядів від сезонності (різного характеру), реакцію на шоки (наприклад на неочікувані зміни обмінного курсу деякі ряди реагують практично моментально, в той час як інші адаптуються з часом).

Один з недоліків полягає у чутливості до великих значень. Оскільки євклідова відстань обчислюється як квадратний корінь з суми квадратів різниць між відповідними компонентами часових рядів, великі значення впливають на результат вимірювання значно більше, ніж малі. Це може призводити до перекосів у вимірюванні, особливо у випадках, коли такі великі значення є результатом випадкових аномалій або шуму в даних. Таким чином, варто уважно розглядати використання евклідової відстані, особливо у випадках, коли дані мають велику дисперсію або потенційно великі викиди.

Крім того, евклідова відстань не враховує контекстуальну інформацію про часові ряди, таку як порядок або частота змін. Вона розглядає кожну точку в просторі як незалежну одиницю, і не враховує можливість, що деякі зміни можуть бути більш важливими або значущими в залежності від контексту. Це може призводити до недооцінки схожості чи відмінності між рядами у випадках, коли певні зміни важливіші, ніж інші. Більше того, евклідова відстань не робить відсутності змін у рядах значущим аспектом. Це означає, що якщо два ряди мають однакові значення в усіх точках, вони будуть між собою абсолютно схожими за євклідовою відстанню, навіть якщо вони мають різний характер або тренд. Це може призводити до неправильного визначення схожості між рядами, оскільки воно не враховує контекстуальної інформації про їхній характер або динаміку. Саме ці проблеми неможливо оцінити стандартними статистичними методами, лише комбінованими з експертними судженнями, тому в даній роботі ми відкинемо врахування цієї проблеми задля більшої універсальності методу.

Значну кількість проблем, з якими стикається цей метод, можна виправити на минулому етапі підготовки даних, зробивши нормалізацію чи чистку від сезонності та трендів. Проте це не гарантує повне очищення даних й все ще не вирішує проблему розтягнутої в часі реакції на однакові шоки.

Отже метод розрахунку дистанції Мінковського між часовими рядами для оцінки їх схожості є одним з найпростіших та універсальних методів, проте він страждає від низки проблем які створюють необхідність для спроби використання інших алгоритмів або для суттєвої адаптації поточної методології. Автор дисертації йтиме кобмінованим шляхом з використанням більш сучасних алгоритмів, але все одно адаптованих задля ще кращого підхоплення схожості рядів незважаючи на всі проблеми.

2.2.2. Кореляційний метод

Одним з інших стандартних підходів є використання кореляційних вимірів для часових рядів [71]. Кореляція дає можливість визначити ступінь схожості між двома рядами даних, що дає змогу оцінити відстань між ними. Себто кореляція є оберненою дистанції між рядами. Тут розглянемо основні поняття та формули, пов'язані із кореляційними вимірами відстані між часовими рядами.

Одним з найпоширеніших методів вимірювання кореляції між часовими рядами є кореляція Пірсона. Цей метод вимірює ступінь лінійної залежності між двома рядами даних. Для двох часових рядів

X та Y, кореляція Пірсона обчислюється наступним чином:

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Однією з переваг кореляційного підходу є його простота та інтуїтивна зрозумілість. Обчислення коефіцієнту кореляції між двома рядами досить просте та легко інтерпретується: значення близьке до 1 вказує на високу позитивну кореляцію, значення близьке до -1 вказує на високу негативну кореляцію, а значення близьке до 0 свідчить про відсутність кореляційного зв'язку між рядами.

Крім того, кореляційний підхід дозволяє враховувати не лише абсолютні значення відповідних точок часових рядів, але й їхні взаємні зміщення. Наприклад, кореляційний аналіз може виявити схожість між двома рядами навіть у випадках, коли їхні значення зміщені у часі.

Однак кореляційний підхід має свої обмеження та недоліки. Один з найбільш важливих недоліків полягає у тому, що він вимірює лише лінійний зв'язок між рядами. Це означає, що кореляція може бути низькою навіть у випадках, коли між рядами існує складний нелінійний зв'язок. Таким чином, кореляційний підхід може недооцінювати схожість між рядами у випадках, коли вони мають складну структуру або нелінійні взаємозв'язки.

Крім того, кореляційний підхід не враховує контекстуальну інформацію про дані. Він не розглядає порядок або частоту змін у рядах, що може призводити до неправильного визначення схожості між рядами у випадках, коли певні зміни є важливішими або значущими в залежності від контексту.

2.2.3. Dynamic Time Warping

Евклідову відстань та кореляційну відстань можна активно використовувати для задачі уточнення схожості часових рядів. Проте, як було зазначено в попередніх пунктах, основним недоліком використання евклідової та кореляційної відстані для часових рядів є те, що їх результати дуже неінтуїтивні якщо два часові ряди ідентичні, але один трохи зміщений вздовж вісі часу або ряди розтягнуті. Евклідова відстань може вважати їх дуже різними одне від одного в даному випадку. Динамічне вирівнювання часу (DTW) було створене групою авторів Berndt та ін. у 1994-му році, щоб подолати це обмеження та надати інтуїтивні вимірювання відстані між часовими рядами, ігноруючи як глобальні, так і локальні зміщення в часовому вимірі. В цій дисертаційній роботі ми зосередимося на DTW, найбільш високорозвиненому і відповідному для подібних часових рядів, що змінюються з різною швидкістю, алгоритмі.

Постановка задачі. Проблема динамічного вирівнювання часу формулюється наступним чином: задані два часових ряди X та Y довжини |X| і |Y|,

X = x_1, x_2, ..., x_|X|

Y = y_1, y_2, ..., y_|Y|

Треба побудувати шлях вирівнювання W = w_1, w_2, ..., w_K

де K - довжина шляху вирівнювання, а k-й елемент шляху вирівнювання є w_k = (i, j)

де i - індекс з часового ряду X, а j - індекс з часового ряду Y. Шлях вирівнювання повинен починатися з початку кожного часового ряду на w_1 = (1, 1) і закінчуватися в кінці обох часових рядів на w_K = (|X|, |Y|). Це забезпечує використання кожного індексу обох часових рядів у шляху вирівнювання. Існує також обмеження на шлях вирівнювання, яке змушує i та j бути монотонно зростаючими у шляху вирівнювання, тому лінії, що представляють шлях вирівнювання на рисунку 1, не перетинаються.

Кожен індекс кожного часового ряду повинен бути використаний. Формально це можна виразити так:

$1 \leq i' - i \leq 1$, $1 \leq j' - j \leq 1$, для w_k = (i, j), w_(k+1) = (i', j')

Оптимальний шлях вирівнювання - це шлях вирівнювання мінімальної відстані, де відстань від шляху вирівнювання W дорівнює

Dist(W) = Сума від k=1 до K Dist(w_ki, w_kj)

Dist(W) - це відстань (зазвичай евклідова відстань) шляху вирівнювання W, а Dist(w_ki, w_kj) - відстань між двома індексами даних (один з X і один з Y) у k-му елементі шляху вирівнювання.

Динамічний програмний підхід використовується для знаходження цього шляху вирівнювання мінімальної відстані. Замість спроби вирішити всю проблему одночасно, знаходяться рішення для підзадач (частин часового ряду) і використовуються для повторного знаходження рішень для трохи більшої проблеми, поки не буде знайдено рішення для всього часового ряду. Конструюється двовимірна матриця вартості |X| на |Y|, де значення в D(i, j) - це

шлях вирівнювання мінімальної відстані, який можна побудувати з двох часових рядів X'=x_1,...,x_i та Y'=y_1,...,y_j. Значення в D(|X|, |Y|) буде містити шлях вирівнювання мінімальної відстані між часовими рядами X та Y. Обидва ряди D представляють час. Вісь x - це час часового ряду X, а вісь y - це час часового ряду Y. На рисунку 2 D показано приклад матриці вартості та шляху вирівнювання мінімальної відстані, прокладеного через неї від D(1, 1) до D(|X|, |Y|).

Щоб знайти шлях вирівнювання мінімальної відстані, потрібно заповнити кожну комірку матриці вартості. Методика застосування динамічного програмування до цієї проблеми полягає в тому, що оскільки значення в D(i, j) - це мінімальна відстань вирівнювання двох часових рядів довжинами i та j, якщо мінімальні відстані вирівнювання вже відомі для всіх трохи менших частин цих часових рядів, які відокремлені на одне дане значення від довжин i та j, то значення в D(i, j) є мінімальною відстанню для всіх можливих шляхів вирівнювання для часових рядів, які на одне дане значення менші, ніж i та j, плюс відстань між двома точками x_i та y_j. Оскільки шлях вирівнювання повинен або збільшуватися на одне значення, або залишатися таким самим вздовж осей i та j, відстані оптимальних шляхів вирівнювання на одне значення менше, ніж довжини i та j, містяться у матриці в D(i-1, j), D(i, j-1) та D(i-1, j-1). Таким чином, значення комірки у матриці вартості є:

$$D(i,j) = Dist(x_i, y_j) + \min[D(i-1,j), D(i,j-1), D(i-1,j-1)]$$

Шлях вирівнювання до D(i, j) повинен проходити через одну з цих трьох клітинок сітки, і оскільки мінімальна можлива відстань вирівнювання вже відома для них, все, що потрібно, - це просто додати відстань між поточними двома точками до найменшої. Оскільки це рівняння визначає значення комірки в матриці вартості, використовуючи значення в інших комірках, важливо, в якому порядку вони оцінюються. Матриця вартості заповнюється по одній колонці знизу вгору, зліва направо.

Після того, як вся матриця заповнена, потрібно знайти шлях вирівнювання від D(1, 1) до D(|X|, |Y|). Шлях вирівнювання фактично обчислюється у зворотному порядку, починаючи з D(|X|, |Y|). Виконується жадібний пошук, який

оцінює комірки ліворуч, вниз та по діагоналі до нижньо-лівої сторони. Та з трьох сусідніх клітинок, яка має найменше значення, додається до початку знайденого шляху вирівнювання, і пошук продовжується з цієї клітини. Пошук завершується, коли досягається D(1, 1).

Час і просторова складність DTW легко визначається. Кожна комірка у матриці вартості |X| на |Y| заповнюється рівно один раз, і кожна комірка заповнюється за постійний час. Це дає як часову, так і просторову складність |X| на |Y|, що дорівнює $O(N^2)$, якщо N=|X|=|Y|. Квадратична просторова складність є особливо обмежувальною, оскільки вимоги до пам'яті становлять терабайт для часових рядів, що містять лише 177 000 вимірів. Лінійна реалізація складності простору алгоритму DTW можлива шляхом збереження лише поточної та попередньої колонок у пам'яті, коли матриця вартості заповнюється зліва направо. Зберігаючи лише дві колонки одночасно, можна визначити оптимальну відстань вирівнювання між двома часовими рядами. Однак неможливо відновити шлях вирівнювання між цими двома часовими рядами, оскільки інформація, необхідна для розрахунку шляху вирівнювання, видаляється разом з викинутими колонками. Це не є проблемою, якщо потрібна лише відстань між двома часовими рядами, але застосунки, які знаходять відповідні області між часовими рядами або об'єднують часові ряди разом, вимагають знаходження шляху вирівнювання.

2.2.4. Запропонована адаптація методу Dynamic Time Warping для економічних часових рядів

Проте навіть в такого потужного алгоритма як Dynamic Time Warping, описаного в минулому пункті, є суттєві недоліки які не дозволяють йому бути максимально ефективним алгоритмом для вирішення задачі подібності часових рядів.

Алгоритм Dynamic Time Warping (DTW) чутливий до довжини і величини подібностей між даними часових рядів. У разі тривалих періодів зі схожими значеннями це може призвести до недоречних вимірювань подібності. Один із способів вирішення цього питання полягає в більшій агрегації даних за

періодичністю (перехід від місячних до квартальних). Такий підхід зменшить вплив тривалих періодів зі схожими значеннями і підвищить чутливість алгоритму DTW до короткострокових змін у даних. Проте це призводить до надзначних втрат інформації. Також варіантом є додавання невеликого шуму, що досліджувався в одній з моїх робіт і є ключовим аспектом, що використовується в даній роботі.

Також, коли мова про економічні часові ряди, з'являються не тільки проблеми сезонності та інші подібні речі, що можна частково виправити в минулих пунктах про підготовку даних. З'являється ще проблема «логічної» відповідності точок. Якщо ми говоримо про місячні дані, то ряди можуть бути подібними, проте в рамках календарного року. Ця ідея полягає в тому, що всі економічні процеси відбуваються в рамках одного бізнес-року. Це може бути рік з точки зору бюджету та податків (зазвичай кінцем є березень). Це може бути рік з точки зору агрокультурних явищ (збирання врожаю, складування та розпродаж, експорт та купівля добрив й т.п.). Себто економіка є циклічною в рамках року, й тому саме в рамках року і необхідно вибудовувати відповідність. Цю ідею можливо реалізувати за допомогою ідей, схожих на одну з вже розроблених адаптацій алгоритму DTW, що називається маскування й що активно використовується в FastDTW [73]. Тому надалі спершу розповімо про алгоритм FastDTW.

FastDTW спрямований на адаптації, які дозволяють алгоритму працювати швидше, ніж оригінальний. Ці адаптації та ідеї, імплементовані в алгоритм, є наслідком багатьох робіт [74-77]. Хоча обидва алгоритми знаходять оптимальний шлях вирівнювання між двома часовими рядами, обчислюючи відстань між кожною парою точок у ряді, FastDTW досягає цієї мети, використовуючи версію ряду з меншою роздільністю, а потім ітеративно вдосконалюючи шлях вирівнювання, використовуючи все вищі роздільності. Це дозволяє FastDTW зменшити кількість порівнянь, необхідних для знаходження оптимального шляху вирівнювання, зробивши його швидшим, ніж DTW, але жертвуючи певною точністю. FastDTW також використовує техніку під назвою

"маскування", щоб додатково зменшити кількість порівнянь, необхідних для знаходження оптимального шляху вирівнювання. Маскування працює, розділяючи часові ряди на підрядки і обчислюючи відстань між лише певними точками в кожному підрядку, ігноруючи інші. Це досягається за допомогою створення бінарної маски, яка вказує, які точки в кожному підрядку слід враховувати при обчисленні відстані. Маска створюється за допомогою знаходження медіанної точки в кожному підрядку і вибору діапазону точок навколо неї, розмір якого залежить від користувацького параметра, що називається "радіусом". За допомогою масок FastDTW може обмежити кількість порівнянь, необхідних для знаходження оптимального шляху вирівнювання, зробивши його ще швидшим, ніж стандартний алгоритм FastDTW. Однак жертвою є те, що точність шляху вирівнювання може бути знижена, особливо якщо параметр радіуса встановлено дуже великим.

Квадратична часова та просторова складність DTW породжує потребу у методах прискорення динамічного вирівнювання часу. Методи, які забезпечують прискорення DTW, можна розподілити на три категорії:

Обмеження - обмеження кількості клітинок, які оцінюються в матриці вартості.

Абстрагування даних - виконання DTW на зменшеному представленні даних.

Індексування - використання функцій нижнього обмеження для зменшення кількості разів, коли потрібно запускати DTW під час класифікації часових рядів або кластеризації.

Обмеження широко використовуються для прискорення DTW. Два найпоширеніші обмеження - це Полоса Сако-Чуби та Паралелограм Ітакури.

Ширина кожної заштрихованої області, або вікно, визначається параметром. Коли використовуються обмеження, алгоритм DTW знаходить оптимальний шлях вирівнювання через вікно обмеження. Однак глобально оптимальний шлях вирівнювання не буде знайдений, якщо він не повністю знаходиться всередині вікна. Використання обмежень прискорює DTW на

константний множник, але алгоритм DTW все ще має складність O(N^2), якщо розмір вхідного вікна є функцією від довжини вхідних часових рядів. Обмеження працюють добре в областях, де очікується, що оптимальний шлях вирівнювання буде близьким до лінійного вирівнювання і проходитиме по діагоналі матриці вартості відносно прямою лінією з незначними коливаннями.

Індексування використовує функції нижнього обмеження, щоб обрізати кількість разів, коли потрібно запускати DTW для певних завдань, таких як кластеризація набору часових рядів або знаходження часового ряду, який найбільше схожий на заданий часовий ряд. Індексування значно прискорює багато застосувань DTW, зменшуючи кількість разів, коли запускається DTW, але не прискорює сам алгоритм DTW.

Алгоритм FastDTW використовує ідеї як з категорії обмежень, так і з категорії абстрагування даних. Використання комбінації обох дозволяє подолати багато обмежень використання кожного методу окремо і призводить до алгоритму, який має складність O(N) як за часом, так і за простором.

Повертаючись до запропонованого алгоритму, використання вищезазначених ідей достатньо активно допомагає вирішити проблему використання алгоритму для економічних рядів. Зокрема мова про маскування. Використання полоси Сако-Чуби з довжиною рівною періодичності ряду дозволяє обмежити відповідність одним календарним роком. Також ідея «пікселізації» теж лягає на цю концепцію, якщо брати пікселізацію періодичності вищого порядку, наприклад для тих же місячних даних брати «пікселізацію» 3*3, себто відповідність між місяцями буде спершу відповідністю між кварталами, а потім буде уточнюватиметься.

2.3. Побудова матриці дистанцій, підготовка до кластеризації

Наступною стадією для новоствореної матриці дистанцій стає її перетворення з, власне, двовимірної матриці у площину з набором точок, де кожна точка відповідає часовому ряду, а відстані між часовими точками відповідають відстаням з матриці дистанцій. Це обов’язковий крок який об’єднує алгоритми пошуку дистанцій та кластеризації і є мостом між ними

задля подальшого ділення рядів на групи за схожістю їх динаміки. Проте задача цього перекладу є неочевидною й потребує більш точного визначення та математичного обгрунтування. Ідеальна стаття на цю тему написана Dokmanic та ін., 2015 [78\, де якісно обгрунтовується побудова дистанційної матриці і вся необхідна математика у випадку евклідових відстаней.

Евклідові відстані - корисний спосіб опису точкових множин та вихідна точка для розробки алгоритмів. Типовим завданням є відновлення вихідної конфігурації точок: на початковому етапі може здатися, що для цього потрібно лише розкладання на власні значення (EVD) симетричної матриці. Фактично, більшість проблем щодо евклідових відстаней потребує відновлення точкової множини, але завжди з одним або декількома з наступних ускладнень:

- Відстані є шумними,
- Деякі відстані відсутні,
- Відстані не мають міток.

Для прикладів застосування, які потребують розв'язання проблем з евклідовими відстанями з різними ускладненнями. Є дві основні проблеми, пов'язані з геометрією відстаней

- Задана матриця - визначте, чи є вона EDM
- Задано можливо неповний набір відстаней - визначте, чи існує конфігурація точок у визначеному вбудованому вимірі - розмірність найменшого афінного простору, що включає точки - що генерує відстані.

Основне завдання, пов'язане з EDM, - відновлення вихідної точкової множини. Це завдання є оберненою задачею до простішої прямої задачі пошуку EDM за даними точками. Таким чином, бажано мати аналітичний вираз для EDM у термінах матриці точок. Ми можемо очікувати, що такий вираз надасть цікаві структурні інсайти.

Одна з важливих теорем є потужна теорема, що стверджує, що ранг EDM не залежить від кількості точок, які його генерують. У багатьох застосуваннях d становить три або менше, тоді як n може бути у тисячах. Доказ цієї теореми

простий, але щоб оцінити, що властивість не є очевидною, ви можете спробувати обчислити ранг матриці неквадратних відстаней.

Те, що дійсно важливо в Теоремі – афінна розмірність точкової множини – розмірність найменшого афінного підпростору, який містить точки, позначена affdim(X). Наприклад, якщо точки лежать на площині (але не на прямій або колі) в R, ранг відповідної EDM складає чотири, а не п'ять. Це буде зрозуміло, оскільки будь-який афінний підпростір - це лише переклад лінійного підпростору.

Найбільш ключовим є, власне, алгоритм, який використовується в цій задачі й виглядає він наступним чином.

1. Створення матриці геометричного центрування J
2. Пошук матриці Грамма $G \leftarrow -\frac{1}{2}JDJ$
3. $U, [\lambda_i]_{i=1}^{n} \leftarrow EVD(G)$
4. Повертаємо вектор коренів елементів діагоналі

Це є ключовим алгоритмом, на базі якого вибудовуються уточнення для певних випадків, які можна детальніше розглянути в [78].

Процес та ідея є абсолютно ідентичними й у випадку кореляційних відстаней та DTW відстаней, що непогано обгрунтовується в низці робіт [79, 80], оскільки побудова власне матриці є однаковою для будь-яких дистанцій, тому ключова математика торкається власне самої матриці і її переведення у площину.

2.4. Методи групування (кластеризації) часових рядів

Після створення площини з набором точок, кожна з яких відповідає певному часовому ряду, а дистанції між ними є дистанціями відповідно до матриці дистанцій з попереднього розділу, їх можна розглядати як множину, для якої можна застосувати певний алгоритм кластеризації. Ідея технік кластеризації базується на групуванні за допомогою схожих ознак. У випадку множини точок ті, які знаходяться близько одна до одної, будуть розглядатися як точки з одного й того ж кластера. Існує велика різноманітність алгоритмів кластеризації, які поділяються на кілька основних категорій: засновані на центроїдах; засновані на

щільності; засновані на з'єднаності; засновані на розподілі та засновані на сітці. Загальний та детальний огляд цих алгоритмів можна прочитати в [81, 82]. Усі ці типи кластеризації відрізняються фундаментальним підходом до створення та обробки кластерів: в той час як алгоритми на основі центроїда переважно ґрунтуються на ідеї того, що існує певний центр для конкретного кластера, який мінімізує загальну відстань від елементів до цього центру, алгоритми на основі щільності визначають множини достатньо близьких точок як кластери, ігноруючи розріджені частини загальної множини точок. Ці підходи фундаментально відрізняються, тому в роботі будуть використані найпопулярніші алгоритми з різних типів.

У даному дослідженні ми використовуємо алгоритми кластеризації з різних категорій для виявлення та аналізу груп точок на площині, отриманій після побудови матриці відстаней та відповідних часових рядів. Це дозволяє нам краще розуміти структуру даних та виявляти схожість між різними частинами досліджуваного простору.

Алгоритми на основі центроїда [83] використовують концепцію центрів кластерів для групування точок. Вони припускають, що кожен кластер має свій центр, який представляє середнє значення всіх точок у цьому кластері в одному чи іншому сенсі та контексті (середнє, зважена медіана тощо). Метою таких алгоритмів є мінімізація відстані між кожною точкою та центром її відповідного кластера. Це може бути корисно для визначення центральних точок у масиві даних або для виявлення груп з однаковими характеристиками.

Алгоритми на основі щільності [84], навпаки, зосереджені на визначенні областей простору, де точки знаходяться відносно близько одна до одної. Ці алгоритми ігнорують розріджені області даних і фокусуються на виявленні компактних кластерів. Це корисно для виявлення груп зі схожими закономірностями у великих наборах даних, де взагалі може бути важко визначити центри кластерів.

Кластерні алгоритми з'єднаності [85] базуються на ідеї визначення кластерів як наборів точок, які з'єднані між собою. Ці алгоритми шукають

найближчі сусіди кожної точки і об'єднують їх у кластери. Це може бути корисно для виявлення груп точок, які мають спільні зв'язки або схожі властивості.

Алгоритми розподілу базуються на припущенні, що дані генеруються з певного ймовірнісного розподілу. Вони намагаються розділити дані на кластери, які відповідають цьому розподілу. Це може бути корисно для моделювання складних структур даних, які не піддаються простим методам кластеризації.

Кластерні алгоритми на основі сітки розділяють простір на рівномірні частини і призначають кожну точку до певного кластера в залежності від її розташування на сітці. Це може бути корисно для швидкого розподілу даних у великих обсягах.

Такий різноманітний набір алгоритмів дозволяє вибирати той, який найкраще відповідає конкретним потребам дослідження та характеристикам даних. Використання різних типів алгоритмів також може допомогти знизити вплив певних обмежень чи асиметрій, які можуть виникнути в результаті конкретної методології чи особливостей набору даних.

При цьому в літературі немає чіткої визначеності які саме алгоритми кластеризації краще використовувати в контексті даних, на яких тестуватиметься алгоритм в поточній дисертаційній роботі. Себто для економічного застосування. Тому метод проб та помилок, а якщо точніше, статистичних оцінок буде найліпшим методом визначення якості моделей, оскільки іншим є, власне, економічний аналіз груп та оцінка відповідності груп загальній економічній логіці що не є частиною та фокусом поточної дисертаційної роботи, на відміну від чисто програмістських та статистичних алгоритмів.

2.4.1. K-Means та інші centroid-based алгоритми

Центроїдні алгоритми кластеризації - це важливий компонент у сучасних методах аналізу даних та машинного навчання. Вони використовуються для групування наборів даних на основі їхньої відстані до центроїдів, що представляють центри кластерів. На відміну від інших методів кластеризації, таких як ієрархічна кластеризація чи методи на основі графів, центроїд-засновані

алгоритми, такі як k-середніх або k-медоїди, створюють чітко визначені кластери, що дозволяє їх легше інтерпретувати.

Однією з головних переваг центроїд-заснованих алгоритмів є їхня простота та ефективність. На практиці це означає, що вони швидко працюють навіть з великими обсягами даних та мають невеликі вимоги до обчислювальних ресурсів. Крім того, їх можна легко налаштувати для вирішення різноманітних задач кластеризації, включаючи виявлення груп у статистичних даних, сегментацію зображень та аналіз текстової інформації.

Проте, ці алгоритми також мають свої недоліки та обмеження. Наприклад, вони чутливі до початкових умов, таких як початкове розташування центроїдів, і можуть давати різні результати при різних запусках. Крім того, вони неефективно працюють з даними, які мають високу вимірність, оскільки обчислювальні витрати зростають експоненціально зі збільшенням кількості вимірів.

Додатковою проблемою є необхідність попереднього визначення кількості кластерів, що може бути нетривіальною задачею, особливо у випадку, коли маємо справу з невідомою структурою даних. В таких випадках використання інших методів, таких як ієрархічна кластеризація, може бути більш підходящим рішенням.

Найкласичнішим прикладом є алгоритм К-середніх [86]. Це некерований алгоритм машинного навчання, що використовується для кластеризації точок даних у К відмінних груп або кластерів. Алгоритм працює шляхом ітеративного призначення точок даних найближчому центру кластера, а потім оновлює центри кластерів на основі ново призначених точок даних. Алгоритм збігається, коли призначення та центри кластерів більше не змінюються або досягнуто максимальну кількість ітерацій. Кількість кластерів визначається радше експериментально або відповідно до певної логіки (економічної), також використовується «метод ліктя» [87].

Дано набір N точок даних $X = \{x_1, x_2, \dots, x_t\}$ і кількість кластерів К, алгоритм прагне мінімізувати суму квадратів відстаней між кожною точкою

даних та її призначеним центром кластера. Це відомо як внутрішній сумарний квадрат відстаней (WCSS) і може бути виражено як:

$$WCSS = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \left|\left|x_j - c_i\right|\right|^2$$

де n_i - кількість точок даних, призначених до і-го кластера, c_i - центр і-го кластера, а ||.||^2 позначає евклідову відстань.

Для знаходження оптимальних центрів кластерів, які мінімізують WCSS, алгоритм K-середніх використовує ітеративний підхід, відомий як алгоритм Ллойда. На кожній ітерації алгоритм спочатку призначає кожну точку даних найближчому центру кластера за допомогою наступної формули:

$$argmin \left|\left|x_j - c_i\right|\right|$$

де argmin - функція, яка повертає індекс центру кластера, який мінімізує відстань до точки даних x_j.

Після призначення всіх точок даних до кластера, алгоритм оновлює центри кластерів, обчислюючи середнє значення всіх точок даних, призначених до цього кластера, за допомогою наступної формули:

$$c_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_j$$

де n_i - кількість точок даних, призначених до і-го кластера, а x_j- j-та точка даних, призначена до і-го кластера.

Алгоритм повторює ці кроки, доки призначення та центри кластерів більше не змінюються або досягнуто максимальну кількість ітерацій.

Адаптації алгоритму k-середніх, такі як k-середні++ (k-means++) [88], є важливими удосконаленнями основного алгоритму k-середніх. Основна відмінність полягає у способі вибору початкових центроїдів для кластерів. У класичному підході, кожен початковий центроїд вибирається випадковим чином серед набору даних. Однак у k-середні++ використовується більш складний метод вибору центроїдів, що сприяє поліпшенню результатів кластеризації.

Ключова ідея методу k-середні++ полягає в тому, щоб збільшити ймовірність вибору початкових центроїдів таким чином, щоб вони були розподілені більш рівномірно по всьому простору даних. Це досягається за допомогою ітеративного процесу вибору центроїдів, де кожен новий центроїд обирається з ймовірністю, пропорційною квадрату відстані до найближчого вже вибраного центроїда. Такий підхід дозволяє уникнути згортання кластерів у випадку, коли початкові центроїди випадковим чином вибрані дуже близько один до одного.

Крім k-середні++, існують інші адаптації алгоритму k-середніх, такі як k-середні|| (k-means||), які спрямовані на оптимізацію швидкості та ефективності алгоритму у випадку великих обсягів даних.

Алгоритм k-медоїдів є важливим удосконаленням алгоритму k-середніх, пропонуючи альтернативний підхід до визначення центроїдів кластерів. Основна відмінність між k-середніми та k-медоїдами полягає в тому, як обираються центральні точки кластерів.

У k-середніх, центроїд кожного кластера розраховується як середнє арифметичне всіх точок у кластері. Однак у k-медоїдах центроїд представлений однією з точок набору даних, яка мінімізує суму відстаней від неї до всіх інших точок у кластері. Цей підхід, відомий як медоїд, є більш робастним у порівнянні з центроїдом, особливо в тих випадках, коли дані мають нелінійну структуру або коли вони містять викиди.

Крім того, вибір медоїда як центральної точки кластера дозволяє уникнути проблем, пов'язаних з відсутністю або невизначеністю середнього значення у випадку категоріальних даних або великих викидів. Також, порівняно з k-середніми, k-медоїди є менш чутливими до початкових умов, оскільки вони базуються на конкретних точках набору даних, а не на випадково обраних центроїдах.

Алгоритм розмитих к-середніх (Fuzzy C-means) є вдосконаленням класичного алгоритму k-середніх, яке дозволяє призначати кожній точці даних членство у всіх кластерах з деякою ймовірністю. Основна відмінність між fuzzy

C-means та k-середніми полягає у способі визначення кластерних центрів та призначенні членства точок у кластерах.

У k-середніх, кожна точка даних належить тільки до одного кластера, а центр кожного кластера розраховується як середнє арифметичне координат усіх точок у кластері. Однак у fuzzy C-means, кожній точці даних призначається членство у кожному кластері з деякою ймовірністю, яка визначається відстанню від точки до центра кластера. Це означає, що кожна точка може мати деяку ступінь належності до кожного кластера.

Цей підхід дозволяє враховувати невизначеність у виборі кластера для кожної точки та дозволяє кожній точці мати вагомий внесок у кожен кластер. Таким чином, fuzzy C-means дозволяє краще моделювати складні структури даних, де точки можуть належати до декількох кластерів одночасно або мати невизначеність у виборі кластера.

У підсумку, всі алгоритмі вкрай схожі між собою за основною ідеєю та математикою, проте вони змінюють або вдосконалюють тим чи іншим методом ідеї, закладені в детально розібраний алгоритм к-середніх.

2.4.2. DBSCAN та інші density-based алгоритми

Алгоритми кластеризації на основі щільності є важливим класом методів аналізу даних, які відрізняються від традиційних методів, таких як k-середні або ієрархічна кластеризація. Основна ідея полягає у визначенні кластерів на основі локальної щільності точок даних. Замість того, щоб визначати кластери на основі глобальних властивостей даних, таких як середнє значення або розмах, ці алгоритми визначають кластери як групи точок, які знаходяться близько одна до одної в просторі.

Однією з головних переваг алгоритмів кластеризації на основі щільності є їхня здатність ефективно виявляти кластери будь-якої форми та розміру, включаючи кластери з високою щільністю та низькою щільністю. Це дозволяє їм працювати ефективно навіть у випадках, коли дані мають складну або нерегулярну структуру.

Однак, алгоритми кластеризації на основі щільності також мають свої недоліки та обмеження. Наприклад, вони чутливі до параметрів, таких як радіус сусідства або мінімальна кількість точок у кластері, які потрібно вручну налаштувати. Крім того, вони можуть мати проблеми з визначенням кластерів у випадках, коли щільність даних варіюється значно або коли вони мають велику розмірність.

Алгоритм кластеризації DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [89] є одним з найкласичніших методів кластеризації на основі щільності. Він особливо ефективний в контексті роботи з великими наборами даних або даними зі складною структурою. Основна ідея полягає у визначенні кластерів на основі щільності точок даних у просторі.

У DBSCAN визначається два основних параметри: радіус ε (epsilon) та мінімальна кількість точок, якою обмежуємося для формування кластера (minPts). Кожна точка класифікується як ядро, границя або шум, залежно від того, чи відповідає вона наступним параметрам. Ядро - це точка, яка має щонайменше minPts сусідів всередині радіусу ε, границя - це точка, яка не є ядром, але має сусіда ядра, та шум - це точка, яка не є ні ядром, ні границею.

Формули DBSCAN включають в себе обчислення відстані між точками у просторі, визначення сусідства та підрахунок кількості сусідів у межах радіусу ε. Формула для визначення відстані між двома точками p та q може бути визначена як евклідова відстань між ними або інші метрики відстані. Сусідство точки p визначається як та точка q, для якої відстань між ними менше або дорівнює ε. Підрахунок кількості сусідів у межах радіусу ε дає можливість класифікувати точки як ядра, границі або шум.

Однією з головних переваг DBSCAN є його здатність ефективно виявляти кластери будь-якої форми та розміру, включаючи кластери зі складною геометрією та кластери з різною щільністю. Крім того, він автоматично визначає кількість кластерів та не вимагає вручну задання кількості кластерів або параметрів.

Проте DBSCAN також має свої недоліки та обмеження. Наприклад, алгоритм може мати проблеми з виявленням кластерів у випадках, коли щільність даних варіюється значно або коли кластери мають різну щільність. Крім того, він може бути чутливим до вибору параметрів, таких як радіус ε та мінімальна кількість точок minPts, і вимагати тщательної настройки для досягнення оптимальних результатів.

Алгоритм OPTICS (Ordering Points To Identify the Clustering Structure) [90] є розширенням алгоритму DBSCAN, яке пропонує альтернативний підхід до виявлення кластерів у наборах даних. Основна відмінність між DBSCAN та OPTICS полягає у способі визначення порядку обробки точок та визначення кластерів.

У DBSCAN кластери визначаються на основі щільності точок, де кожна точка класифікується як ядро, границя або шум в залежності від кількості сусідів у межах заданого радіусу ε та мінімальної кількості сусідів minPts. У порівнянні з цим, OPTICS працює, формуючи послідовність точок на основі їхньої віддаленості до найближчого ядра або границі, а також відстані до найближчого сусіда. Це дозволяє визначити ступінь належності кожної точки до кластера та визначити структуру кластерів у вигляді гіперпараметру "мінливість" (reachability distance).

Однією з головних переваг OPTICS є можливість виявлення кластерів різної щільності та форми, що дозволяє ефективно враховувати особливості різних типів даних. Крім того, він автоматично визначає кількість кластерів та не вимагає вручну задання параметрів.

Проте, OPTICS може бути менш ефективним у порівнянні з DBSCAN у випадках, коли структура даних дуже складна або коли кластери мають нерегулярну форму. Крім того, він може вимагати більше обчислювальних ресурсів через потребу в обробці послідовності точок.

У підсумку, алгоритм OPTICS є потужним інструментом у сфері кластеризації даних, який пропонує альтернативний підхід до виявлення кластерів порівняно з DBSCAN. Він може бути особливо корисним у випадках,

коли кластери мають різну щільність та форму, але вимагає уваги до особливостей даних та параметрів для досягнення оптимальних результатів.

Підсумовуючи розділ, тут наведено два методи, заснованих на щільності, обидва мають свої плюси та недоліки і, найголовніше, однакову ідею й схожу реалізацію, що кардинально відрізняється від методів, описаних в минулому розділі. Ці методи також використовуватимуться в цій дисертаційній роботі для побудови та оцінки моделей.

2.4.3. Hierarchical Clustering та інші ієрархічні алгоритми

Алгоритм ієрархічної кластеризації є одним із важливих методів групування даних, який дозволяє створювати ієрархічну структуру кластерів. Основна ідея полягає в послідовному об'єднанні або розбитті кластерів на основі схожості між їхніми елементами.

Алгоритм ієрархічної кластеризації [91] може бути реалізований у двох основних варіантах: агломеративному та дивізійному. У агломеративному підході кожен об'єкт спочатку розглядається як окремий кластер, а потім поступово об'єднується з іншими кластерами на основі їхньої схожості, поки не буде сформована одна велика ієрархічна структура. У дивізійному підході весь набір даних спочатку розглядається як один кластер, а потім поступово розбивається на менші кластери до досягнення заданих критеріїв.

Формули, що використовуються в алгоритмі ієрархічної кластеризації, зазвичай включають в себе визначення схожості між об'єктами, таке як відстань або подібність, а також способи об'єднання або розбиття кластерів. Наприклад, одним із методів визначення схожості є використання евклідової відстані між точками, а для об'єднання кластерів можуть використовуватися методи, такі як однорідне з'єднання (single linkage), з'єднання середнє (average linkage) або повне з'єднання (complete linkage).

Single Linkage (однорідне з'єднання):

У методі однорідного з'єднання відстань між двома кластерами визначається як відстань між найближчими точками в цих кластерах. Формула

для обчислення відстані між двома кластерами C_1 та C_2 може бути записане як:

$$d(C_1, C_2) = \min_{x \in C_1, y \in C_2} d\,(x, y)$$

де d(x, y) - це відстань між точками x та y, яка може бути обчислена, наприклад, за допомогою евклідової відстані.

Average Linkage (з'єднання середнє):

У методі з'єднання середнього відстань між кластерами визначається як середнє значення відстаней між всіма парами точок з різних кластерів. Формула для обчислення відстані між двома кластерами C_1 та C_2 може бути записане як:

$$d(C_1, C_2) = \frac{1}{|C_1| \cdot |C_2|} \sum_{x \in C_1} \sum_{y \in C_2} d(x, y)$$

Complete Linkage (повне з'єднання):

У методі повного з'єднання відстань між кластерами визначається як максимальна відстань між всіма парами точок з різних кластерів. Формула для обчислення відстані між двома кластерами C_1 та C_2 може бути записане як:

$$d(C_1, C_2) = \max_{x \in C_1, y \in C_2} d\,(x, y)$$

Ці метрики використовуються для визначення того, яким чином кластери будуть об'єднані під час агломеративного процесу ієрархічної кластеризації. Кожен метод має свої властивості та може призводити до різних структур кластерів, що відображається на результаті кластеризації.

Переваги алгоритму ієрархічної кластеризації включають в себе можливість візуалізації ієрархічної структури кластерів, яка дозволяє аналізувати взаємозв'язки між кластерами на різних рівнях деталізації. Крім того, цей метод не вимагає попереднього визначення кількості кластерів та дозволяє отримувати ієрархічну структуру кластерів, яка може бути корисною для подальшого аналізу даних.

Однак алгоритм ієрархічної кластеризації має свої недоліки. Один із найбільш очевидних - це велика обчислювальна складність, особливо при

великих обсягах даних. Крім того, ієрархічна структура кластерів може бути важко інтерпретована, особливо у випадку складних ієрархій або великої кількості кластерів.

У підсумку, алгоритм ієрархічної кластеризації є важливим методом у сфері аналізу даних, який дозволяє створювати ієрархічну структуру кластерів. Він має свої переваги та недоліки, але відповідно до контексту даних та задач аналізу може бути відмінним інструментом для групування та визначення взаємозв'язків між елементами даних.

2.5. Висновки до розділу 2

Підсумовуючи цей розділ, першопочаткова база даних з низки часових рядів проходить шлях від, власне, набору рядів до згрупованого набору рядів за певними ознаками схожості динаміки.

Спершу схожість шукається методами відстаней часових рядів один від одного. Для цього в цій роботі запропоновано декілька ідейно різних методів пошуку відстаней, зокрема евклідова відстань, кореляційна відстань, метод динамічного викривлення часу, а також його адаптована версія, заточена спеціально для роботи з часовими рядами з економічної сфери.

Це дозволяє вибудувати матрицю дистанцій, котра перетворюється на двовимірний простір з точками, кожна з котрих відповідає часовому ряду й дистанції між цими точками є рівними дистанціям між рядами. Можливість та алгоритм такого перетворення наведено в розділі 2.3. Там доведена його унікальність з точністю до повороту, що не є суттєвою проблемою в контексті задачі, що розглядається в даній дисертаційній роботі, адже це не впливає на наступний етап.

А наступний етап є кластеризацією вищезгаданої площини з точками, де точки діляться на групи різноманітними алгоритмами: центроїдними, щільнісними та іншими. Ці алгоритми ідейно дуже різні між собою й немає чіткої відповіді котрий є ідеальним для використання в поточному кейсі на розгляненій базі даних.

В даній роботі ми використаємо різні комбінації алгоритмів пошуку відстаней між часовими рядами та методами кластеризацій, тому пояснення всіх цих алгоритмів було необхідно для вибудови розуміння пайплайну в рамках цієї роботи.

РОЗДІЛ 3. Побудова моделей для прогнозування агрегованих за динамікою показників

Наступним кроком для вибудови моделей є створення статистичних моделей, що прогнозуватимуть загальний ряд виходячи з агрегованих компонент, отриманих як результат алгоритмів минулого розділу.

Класичним методом для порівняння є випадкове блукання (random walk), який використовує останнє спостереження часового ряду для прогнозування майбутніх значень. Випадкове блукання є простим інструментом, що встановлює базовий рівень прогнозування, використовуючи лише попередні дані. Однак його головний недолік полягає у тому, що він не враховує інші важливі фактори або змінні у часовому ряді, що можуть вплинути на майбутні значення. Тому він часто використовується лише як початкова точка порівняння з більш складними моделями прогнозування, які здатні ураховувати більше інформації і складніші закономірності у часових даних.

ARIMA (Autoregressive Integrated Moving Average) є однією з найпоширеніших моделей для прогнозування часових рядів. Ця модель поєднує в собі авторегресійні (AR) і скользячі середні (MA) компоненти з інтегрованими (I) диференціюваннями, що дозволяє моделювати нелінійність та ураховувати нестаціонарність часових рядів. У AR компонент моделі враховує залежність між поточним значенням і попередніми значеннями часового ряду. MA компонент дозволяє враховувати залежність між поточним значенням і статистичними збуреннями у минулому. SARIMA (Seasonal ARIMA) розширює базову модель ARIMA, додавши сезонні компоненти для моделювання сезонних змін у часовому ряді. Ця модель корисна для прогнозування даних з вираженими сезонними коливаннями, наприклад, продажів у різні місяці чи сезонні зміни у фінансових показниках. Переваги ARIMA та SARIMA включають їхню простоту в налаштуванні та інтерпретації результатів. Вони можуть добре працювати з невеликими часовими рядами та надають можливість враховувати як лінійні, так і нелінійні залежності. Однак ці моделі можуть бути менш ефективними для

складних даних зі складними нестаціонарними або нелінійними залежностями, а також вимагають чіткого розуміння вихідних даних та їхніх властивостей.

Обидва вищезазначених методів можна використовувати як на загальному агрегованому ряді, так і на компонентах і агрегувати вже результати з певною вагою. Обидві гіпотези мають місце бути, проте в літературі вважається що агрегація є, в цілому, більш якісним варіантом. В даній дисертаційній роботі у розділі 4 ця гіпотеза буде додатково перевірена.

Випадковий ліс (Random Forest) і XGBoost, хоч і не є традиційними алгоритмами аналізу часових рядів в контексті машинного навчання, проте їх можливо якісно адаптувати та використовувати для таких задач. Випадковий ліс є ансамблем рішень, де кілька дерев рішень об'єднуються для узагальнення прогнозів. Його переваги включають здатність автоматично враховувати важливі функції, що робить його ефективним у роботі з великими обсягами даних. Випадковий ліс може автоматично розпізнавати складні шаблони і взаємозв'язки у часових рядах. XGBoost (eXtreme Gradient Boosting) є іншим популярним методом, який базується на градієнтному бустінгу дерев рішень. Він використовує стратегію побудови послідовних моделей, які коригують помилки попередніх моделей для покращення точності прогнозів. XGBoost має високу швидкодію і дозволяє автоматично розраховувати важливість ознак, що робить його ефективним у роботі з часовими рядами з великою кількістю ознак. Хоча випадковий ліс і XGBoost є потужними інструментами для прогнозування часових рядів, вони мають свої обмеження. Наприклад, вони можуть бути чутливими до перенавчання, особливо при наявності великої кількості даних або надмірній складності моделі. Крім того, ці моделі вимагають уважного налаштування гіперпараметрів для досягнення оптимальної продуктивності.

Рекурентні нейронні мережі (RNN), зокрема довга короткочасна пам'ять (LSTM), є потужними інструментами для прогнозування часових рядів. LSTM вирішує проблему зниклої градієнта в звичайних RNN, що дозволяє зберігати довгострокові залежності у даних. Основна перевага LSTM полягає в здатності

ефективно працювати з послідовними даними, такими як часові ряди, і виявляти складні залежності між попередніми спостереженнями і майбутніми значеннями.

LSTM широко застосовується у прогнозуванні валютних курсів, погоди, продажів та інших областях, де важливо враховувати контекст і залежності у часових даних. Вона може автоматично враховувати сезонні або трендові зміни, що робить її корисною для прогнозування змінливих часових рядів.

Однак LSTM має свої обмеження. Наприклад, вона може бути вимогливою до обчислювальних ресурсів і часу навчання, особливо при роботі з великими обсягами даних. Крім того, її ефективність може залежати від правильного підбору гіперпараметрів і налаштувань моделі. Наприклад, занадто довгий час навчання може спричинити перенавчання, а недостатній час - недонавчання.

Загальна перспектива щодо моделей прогнозування часових рядів включає широкий спектр підходів з різними перевагами і обмеженнями. Класичні методи, такі як ARIMA і SARIMA, добре працюють зі стаціонарними та нелінійними часовими рядами, забезпечуючи ефективність інтерпретації результатів. У той же час, машинне навчання відкриває нові горизонти завдяки своїм здатностям розпізнавати складні шаблони і працювати з великими обсягами даних.

Тим не менше, вибір моделі залежить від конкретного завдання і властивостей даних. Наприклад, для простих часових рядів з невеликою кількістю спостережень можуть підійти класичні методи, тоді як для складних нелінійних залежностей і великих обсягів даних може бути корисним застосування машинного навчання. Крім того, інновації в глибокому навчанні, такі як LSTM, продовжують покращувати точність прогнозів часових рядів.

В рамках поточного розділу детальніше розглянемо всі ці моделі та специфікації, що дозволяють використовувати ці моделі для цілей, означених в цій дисертаційній роботі та таких, що будуть детальніше досліджуватися в наступному розділі.

3.1. Випадкове блукання

Модель випадкового блукання (Random Walk, RW) [91] передбачає, що потенційне значення ряду буде таким самим, як його найближче спостережене

значення. Цей підхід часто використовується як еталон для інших моделей прогнозування через його простоту використання та відсутність припущень щодо процесу генерації даних. Математично RW можна виразити як:

$$y_{t+1} = y_t$$

де y_t- спостережувана змінна у момент часу t, а y_{t+1} - прогнозована у момент часу t+1. Прогноз для подальшого горизонту може бути побудований ітеративно.

Модель випадкового блукання ігнорує будь-які інші змінні, які можуть впливати на серію, такі як сезонність, відгуки від інших змінних або динаміка, яка не є лінійною, оскільки вона передбачає, що майбутнє значення серії буде ідентичним останньому відомому значенню серії. В результаті ця модель часто використовується як примітивна модель прогнозування або як еталон, за яким вимірюється ефективність більш складних моделей.

Ця модель не передбачає використання алгоритмів для оцінки коефіцієнтів, оскільки вони вже є оціненими, що також збільшує її простоту для використання. Але також це є базою для розуміння найпростіших AR моделей, оскільки випадкове блукання є AR(1) моделлю без константи й з коефіцієнтом при лаговій змінній рівному одиниці. Також, випадкове блукання можна переписати як ARIMA(0,1,0) модель, перемістивши y_t в ліву частину рівняння й отримавши:

$$\Delta y_{t+1} = 0$$

що є першою інтеграцією, котра рівна 0.

Часто в цю модель додають нормально-розподілений шок з математичним очікуванням рівним нулю. Зазвичай це робиться для генерації часового ряду що є випадковим блуканням. Це є важливим в контексті генерації ряду, що не викривить ключові характеристики першопочаткового ряду у випадку додавання до нього випадкового блукання, що є важливим для подолання однієї з проблем, описаної у розділі

3.2. SARIMA

Перейдемо до реально використовуваних моделей прогнозування й розпочнемо зі стандартної моделі авторегресії, інтегрованого ковзного середнього (Autoregressive Integrated Moving Average, ARIMA), широко використовуваної методики прогнозування часових рядів у більшості областей, пов'язаних з аналізом часових рядів. Це вибіркова модель для більшості досліджень економетрики та аналізу часових рядів. Її відомою особливістю є здатність захоплювати часові залежності та закономірності в послідовних даних. ARIMA поєднує три ключові компоненти: авторегресію (AR), інтегрування (I) та ковзне середнє (MA). Компонент AR пояснює взаємозв'язок між точкою даних та її запізненими спостереженнями, порядок якого позначається як 'p' і визначає кількість запізнених значень, які беруться до уваги під час прогнозування. Інтегрування, з порядком 'd', використовується для досягнення стаціонарності шляхом віднімання кожної точки даних від її попереднього значення, ефективно зменшуючи тенденції або змінні середні. Компонент MA характеризує взаємозв'язок між поточною точкою даних та минулим білим шумом або помилками, з порядком 'q', що визначає кількість запізнених термінів помилок, які використовуються під час прогнозування. Загальна формула подана нижче:

$$(1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_\mathrm{p} L^\mathrm{p})\,(1 - L)^\mathrm{d} X_\mathrm{t} = (1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_\mathrm{m} L^\varphi) Z_\mathrm{t}$$

Де X_t - це ряд в період t, Z_t - помилка в період t, L - оператор зсуву назад, ϕ_p і θ_m - оцінені коефіцієнти.

Як було зазначено в другому розділі, ряди можуть мати проблеми з сезонністю і є два можливих рішення, які будуть використані в цьому дослідженні. Перше - це сезонна корекція ряду за допомогою підходу X-12. Це традиційний підхід до вирішення проблем сезонності й саме він був частково описаний в другому розділі.

Другий підхід до роботи з проблемами сезонності - це використання моделі Seasonal ARIMA (SARIMA). Ця модель працює як звичайна ARIMA, але має також сезонні параметри p, d і q, які базуються на частоті серії (щомісячно у

нашому випадку, тому сезонна затримка становить 12), і додають до стандартної формули ARIMA сезонну частину (12-та затримка). Це чудовий підхід для моделі, що базується на даних, щоб враховувати сезонність. Загальна формула подана нижче:

$$(1 - \phi^1 L - \phi^2 L^2 - \ldots - \phi_{\mathrm{p}} L^{\mathrm{p}}) * (1 - L) * (1 - \Phi^1 L^s - \Phi^2 L^{2s} - \cdots - \Phi_{\mathrm{p}} L^{ps}) * (1 - L^s)^{\mathrm{D}} X_{\mathrm{t}} = (1 + \theta^1 L + \theta^2 L^2 + \ldots + \theta_{\mathrm{m}} L^{\varphi}) * (1 + \Theta^1 L^s + \Theta^2 L^{2s} + \ldots + \Theta_{\mathrm{m}} L^{qs}) * Z_{\mathrm{t}}$$

Де X_t - це ряд в період t, Z_t - помилка в період t, L - оператор зсуву назад, ϕ_p і θ_m - оцінені коефіцієнти, s - сезонна частота, яка визначає затримку, яка повинна бути взята (12 для щомісячного ряду).

Ця модель менш поширена як еталонна ARIMA. Проте вона демонструє, наскільки корисним може бути такий підхід, коли сезонність є проблемою.

В обох випадках специфікації, а саме p, d, q, P, D, Q, вибираються за допомогою процедури auto.arima() [93]. Процедура базується на критерії інформаційної Акаіке (Akaike Information Criteria, AIC) і перевіряє моделі з усіма комбінаціями p, d, q, P, D, Q у режимі пошуку по сітці. Вона оцінює AIC і вибирає модель з найменшим значенням. AIC, на відміну від SSE (сума квадратів помилок), покарає за занадто велику кількість запізнень у моделі, тому уникаються ті, які мають багато коефіцієнтів без значного покращення ймовірності. Формула для AIC подана нижче:

$$AIC = 2k - 2ln(L)$$

Де k - кількість коефіцієнтів, L - максимальна вирогідність моделі з відповідним набором коефіцієнтів.

Цей підхід є прикладом grid search, себто його складність зростає експоненційно у випадку перебору значної кількості специфікацій. Тому часто рішенням є обмеження, що накладаються на максимальний розмір кожного значення p, d, q, P, D, Q. Зазвичай, d та D беруться зі значенням не більшим за 2, а дуже ймовірно що 1. Оскільки це інтеграція, що контролює стаціонарність

досліджуваних рядів в першу чергу, то треба звертатися до природи досліджуваного ряду. Оскільки досліджуються ряди інфляції, що буде описано в розділі 4, то очікуваним обмеженням для коефіцієнтів d та D є 0, оскільки ряди є природньо стаціонарними. Проте в літературі є порада брати обмеження на 1 більше очікуваного, тому в моделях, побудованих в цій дисертаційній роботі, для цих коефіцієнтів братиметься обмеження 1. Це вже 4 моделі через всі можливі комбінацій (0,0), (1,0), (0,1), (1,1). Це множитметься на обмеження для всіх інших коефіцієнтів. Оскільки сезонність підхоплюється сезонною компонентою, що поширюється на рік назад (від чого очікувані максимальні значення для P та Q мають бути 1, себто беремо 2), то значення для p та q мають бути суттєво менше 12 й, більш того, менше 6, оскільки ми працюємо з місячними даними. Тому беремо обмеження 5 для p та q. Це призводить до загальної кількості оцінюваних моделей в auto.arima() процедурі рівним 5*5*2*2*4 = 400.

Існують деякі обмеження моделей типу ARIMA. Одним із найважливіших обмежень є неможливість захоплення залежностей з іншими змінними. Модель розглядає лише історичні значення ряду і ігнорує будь-які інші змінні, які можуть впливати на ряд. Якщо інші змінні мають сильну взаємозв'язок з рядом, але не включені до моделі, це може бути серйозним недоліком.

Ще одним обмеженням моделей типу ARIMA є їх нездатність виявляти нелинійні патерни даних. Модель передбачає лінійний зв'язок між рядом і його попередніми значеннями та помилками. Якщо в даних присутні нелинійні патерни, такі як експоненційний ріст або зниження, модель може не здати точно зафіксувати ці патерни.

Використовуючи алгоритми, описані у розділах 3.1 та 3.2 в комбінації з алгоритмами, описаними у розділі 2, постає проблема агрегації прогнозів. Отже останнім етапом саме цього пайплайну є модель, яка допоможе загрегувати прогнози з попереднього етапу в одну середню ставку. Для цього ми використаємо простий підхід OLS (Ordinary Least Squares), який оцінюється на історичних агрегованих рядах.

$$y_t = \beta_1 y_{t,1} + \cdots + \beta_k y_{t,k} + \varepsilon_t$$

де y_t – значення у момент t; $y_{t,k}$ – значення для категорії k в момент t.

Надалі прогнози агрегуються з оціненими вище коефіцієнтами для зваженої суми ряду, себто максимально агрегованого оригінального ряду.

Саме ця ідея використовується у праці, написаній Marco Huwiler і Daniel Kaufmann у 2013 році [94], автори представляють та оцінюють модель ARIMA, засновану на дезагрегованих даних індексу споживчих цін (ІСЦ) у Швейцарії, яка використовується для короткострокового прогнозування інфляції. Новизна їхнього підходу полягає в акценті на дезагрегації даних. Замість того, щоб застосовувати методологію ARIMA безпосередньо до загального ІСЦ, вони оцінюють окремі ARIMA моделі для кожної компоненти індексу (тобто категорії товарів та послуг) і потім агрегують прогнози від цих моделей із використанням вагових коефіцієнтів витрат. Згідно з теоретичними передумовами, дезагрегація дозволяє врахувати гетерогенність рухів цін окремих компонентів ІСЦ і тим самим отримати точніші прогнози загальної інфляції. Ще однією новою рисою є врахування у моделі змін частоти збору даних для різних компонентів ІСЦ. Автори розширюють базову ARIMA модель, використовуючи представлення стану-простору та фільтр Калмана, щоб коректно прогнозувати компоненти з нерегулярною частотою збору. Це вдосконалення дозволяє безпосередньо включати додаткові спостереження для тих компонентів, для яких частота збору даних зросла в останні роки. Результати показують, що дане розширення покращує точність прогнозів. Ця стаття також є фундаментальною для дисертаційної роботи через одне з найкращих введень поняття дезагрегації часового ряду на компоненти. Вона надихнула подальше дослідження тематики що призвело до низки статей автора дисертації й дослідження теми з ухилом в алгоритми машинного навчання.

У статті Krukovets and Verchenko, 2019 [95], що є прямим наслідком попередньої роботи, розглядається емпірична ефективність кількох альтернативних моделей прогнозування інфляції в Україні на основі структурних та орієнтованих на дані підходів, а також на основі агрегованих та розгорнутих даних. Автори демонструють, що комбінована ARMA модель з

додатковими даними, яка використовує розгорнуті дані про основну інфляцію для України, значно покращує якість прогнозу інфляції порівняно з більш структурними моделями, заснованими на агрегованих даних. Там підкреслюється, що високоякісний прогноз інфляції є ключовим для центрального банку, оскільки він надає основу для багатьох його рішень і політичних заходів. Досліджуються різні економетричні моделі для прогнозування інфляції, зокрема, невеликі моделі, що працюють з великою кількістю даних, і структурні моделі, які описують складні зв'язки між різними частинами економіки. Основна специфікація моделі базується на комбінованій ARMA (CARMA) моделі, зробленій на базі минулоописаної статті та використовуваній Швейцарським національним банком. Кожна компонента інфляції моделюється окремо, а їх прогнози комбінуються в один загальний прогноз основної інфляції. Даний підхід дозволяє використовувати багату структуру даних про різні компоненти інфляції. Мета полягає у порівнянні його ефективності з іншими альтернативними статистичними моделями, які використовують як агреговані, так і розгорнуті дані, а також з базовими прогнозами Національного банку України, що ґрунтуються на структурній моделі квартального прогнозування (QPM). Дослідження вносить важливий внесок у наукову літературу з декількох позицій. По-перше, досі мало емпіричних доказів щодо відносної ефективності ARMA-моделей для прогнозування інфляції в розвиваючихся економіках. По-друге, пропонуються кілька специфікацій додаткових даних для захоплення періодів ексцесивної волатильності, і показано, що вони можуть значно покращити якість прогнозування моделі. По-третє, це перше дослідження, що емпірично досліджує розгорнуті дані про українську інфляцію з точки зору їх прогностичної сили порівняно з агрегованою інфляційною серією. Таким чином, ця стаття несе вклад у обговорення про корисність розгорнутих моделей в порівнянні з агрегованими моделями, надаючи нові емпіричні дані. Саме ця база даних є основою для дослідження якості прогностичних моделей запропонованих в цій дисертаційній роботі.

Методи, згадані в розділах 3.1 та 3.2, випадкове блукання та SARIMA, можна використовувати як на загальному агрегованому ряді, так і на його компонентах, а результати потім агрегувати з відповідними вагами. Обидва ці підходи є корисними для аналізу як загальних даних, так і окремих компонентів, а результати їх застосування можна об'єднати з використанням ваги для отримання комплексного висновку. Однак у висвітленні цієї проблеми в науковій літературі перевага надається методу агрегації, оскільки він зазвичай забезпечує більш якісні результати. В цьому дослідженні детально розглянемо цю гіпотезу у розділі 4 і проведемо додаткову перевірку її ефективності.

3.3. Випадковий ліс

Випадковий ліс (Random Forest) [96] - потужний алгоритм машинного навчання, який часто використовується для завдань регресії та класифікації. Це метод ансамблю, який поєднує прогнози кількох дерев рішень для покращення точності та надійності моделі. Зростання популярності технік науки про дані призвело до великої кількості робіт, де Random Forest виступає як основна або додаткова (базова) модель для фінансових та навіть макроекономічних часових рядів.

RF особливо корисний для захоплення загальної динаміки та нелінійності в даних. На відміну від лінійних моделей, таких як ARIMA, модель Random Forest може зафіксувати складні нелінійні зв'язки між ознаками та цільовою змінною, що може мати місце для поточного набору даних. Це пов'язано з використанням моделлю дерев рішень, які можуть симулювати нелінійні зв'язки та взаємодії між ознаками. Оскільки RF може обробляти дані паралельно і не потребує такого самого рівня попередньої обробки, як ARIMA, він може ефективно обробляти великі набори даних. Нашою метою є прогнозування середньої ставки усіх банків в Україні, але набір даних включає ставки окремих банків. У той час як ARIMA обмежується одним часовим рядом і може не здати захопити загальну динаміку набору даних, моделі RF можуть використовувати цю інформацію для більш точних прогнозів середньої ставки.

Кожне дерево рішень у моделі RF навчається на випадково вибраному піднаборі атрибутів і стохастичному піднаборі навчальних даних. Це зменшує перенавчання і підвищує різноманітність моделі, покращуючи її узагальнювальні властивості. У кожному етапі процесу навчання для кожного дерева в лісі вибирається випадковий піднабір навчальних даних. Це допомагає уникнути перенавчання та покращує різноманітність дерев. Для кожного дерева в лісі випадковим чином вибирається підмножина ознак для поділу вузлів дерева. Це допомагає зменшити кореляцію між деревами та покращити узагальнювальні властивості моделі. Кожне дерево в лісі будується за допомогою рекурсивного процесу бінарних поділів, де кожен вузол дерева представляє тест на одній з вибраних ознак. Дерево росте до досягнення певного критерію зупинки, такого як мінімальна кількість вибірок у вузлі листка. Кінцевий прогноз моделі формується шляхом об'єднання прогнозів всіх дерев рішень у лісі, в даному випадку - шляхом обчислення середнього значення.

RF навчається за допомогою ансамблю дерев рішень, кожне з яких навчається зменшувати варіацію цільової змінної. Варіація цільової змінної у визначається як:

$$Var(y) = E\left(\left(y - E(y)\right)^2\right)$$

Де E(y) – математичне сподівання. Метою RF є мінімізація варіації цільової змінної шляхом створення ансамблю дерев рішень з низькою кореляцією та високою індивідуальною прогностичною силою.

Критерій поділу, використовуваний для кожного вузла дерева рішень, базується на індексі чистоти Джині, який визначає однорідність вибірок у кожному вузлі і використовується для вибору ознаки та точки поділу, що максимізує інформаційний приріст поділу.

Індекс чистоти Джині вимірює ступінь або ймовірність того, що певна точка даних буде неправильно класифікована, коли вона випадково маркується відповідно до розподілів класів у наборі даних. Індекс чистоти Джині коливається від 0 до 1, де значення 0 означає повністю чистий вузол (всі вибірки належать до одного класу), а значення 1 вказує на повністю нечистий вузол

(вибірки рівномірно розподілені між усіма класами). Він обчислюється за формулою:

$$Gini = 1 - \sum(p_i^2)$$

Де p_i - це ймовірність класу i.

Індекс чистоти Джині використовується для оцінки якості кожного потенційного поділу при побудові дерева рішень. Індекс чистоти Джині обчислюється для кожного можливого поділу, і вибирається поділ з найнижчим індексом чистоти Джині. Цей процес повторюється для кожного послідовного поділу до досягнення критерію зупинки (наприклад, до досягнення максимальної глибини або коли кількість вибірок у вузлі стає менше вказаного порогу).

При прогнозуванні нового зразка кожне дерево в лісі повертає прогноз, а кінцевий прогноз визначається більшістю голосів. Цей підхід допомагає уникнути перенавчання і покращує загальну ефективність узагальнення моделі. Формула може бути записана як:

$$\hat{y} = \frac{1}{N} * \sum(\hat{y_i})$$

де $\hat{y_i}$ - прогнозоване значення i-го дерева рішень у лісі, а N - загальна кількість дерев у лісі.

Ключові аспекти алгоритму RF спрямовані на створення узагальненого та надійного прогнозу, захоплюючи потенційні нелінійності за допомогою агрегації багатьох моделей з різними відповідями на ті ж самі збурення. Крім того, він працює добре з великими наборами даних без перенавчання, що є важливим для даної роботи

Використання алгоритму випадкового лісу (Random Forest) у випадку часових рядів відрізняється від його застосування до звичайних статичних даних через особливості такого типу даних. Основна відмінність полягає в тому, що часові ряди містять залежності в часі, що потребує спеціального підходу при побудові моделей прогнозування.

Однією з основних проблем у використанні випадкового лісу для часових рядів є врахування автокореляції, тобто залежностей між значеннями часового ряду в різний момент часу. У звичайному випадку, коли дані не є часовими

рядами, припущення про незалежність спостережень може бути прийнятним. Однак у випадку часових рядів це припущення не виконується, що може призвести до некоректних результатів, якщо не буде враховано автокореляцію.

Для успішного застосування випадкового лісу до часових рядів необхідно враховувати наступні аспекти:

Лаги (затримки) часових рядів: Під час побудови набору даних для моделі потрібно враховувати попередні значення (лаги) вхідних ознак як додаткові фактори для прогнозування майбутніх значень. Наприклад, для прогнозування значення часового ряду в момент t, можна використовувати значення в момент $t-1$,$t-2$ і т.д. Себто необхідно доповнити вхідну матрицю й додати в неї лаги самостійно, набір цих лагів визначається експериментально.

Обробка автокореляції: Перед побудовою моделі важливо проаналізувати автокореляцію часового ряду і врахувати ці залежності при виборі параметрів моделі. Наприклад, можна використовувати деякі техніки для врахування автокореляції в вихідному наборі даних або налаштувати гіперпараметри моделі для врахування цих залежностей. Також можна використовувати випадковий ліс як додаткову модель після очищення даних від автокореляційних проблем, особливо на великому лаговому періоді, що Random Forest підхопити не здатний.

Оптимізація параметрів: Враховуючи специфіку часових рядів, необхідно оптимізувати параметри моделі, такі як кількість дерев у лісі, глибина кожного дерева, кількість використовуваних ознак у кожному розділі тощо.

Одна з надважливих робіт в контексті використання моделей машинного навчання, Chu and Qureshi, 2022. Саме цей підхід в тому чи іншому форматі є ключовим для інституцій, що роблять короткострокове прогнозування. Автори проводять "змагання" серед популярних методів прогнозування, включаючи методи машинного навчання (ML), які використовуються для прогнозування зростання ВВП США. З огляду на нестабільний характер даних про зростання ВВП, використовується рекурсивна стратегія прогнозування для розрахунку метрик ефективності прогнозів за межами вибірки для кількох підперіодів. Для цього використовуються три набори предикторів: великий набір з 224

предикторів з великої щоквартальної макроекономічної бази даних (FRED-QD), невеликий набір з дев'яти потужних предикторів, вибраних з великого набору даних, та інший невеликий набір, що включає додатково індекс високочастотних бізнес-умов. Висновки такі, що при прогнозуванні за великою кількістю передикторів з різним передбачувальним потенціалом методи ML перевершують інші для прогнозування на короткі строки, але важко відрізнити їхню ефективність для прогнозування на довгі строки. А також методи ML засновані на щільності зазвичай працюють краще з великим набором передикторів, ніж з невеликим піднабором сильних передикторів, особливо коли мова йде про прогнозування на короткий строк. Це показує та зайвий раз підкреслює важливість алгоритмів машинного навчання у роботі з великими наборами даних, непересічні можливості які з'являються від використання й що активно досліджується в даній дисертаційній роботі.

3.4. XGBoost

Алгоритм екстремального градієнтного бустингу (XGBoost) [97] є потужним і досить популярним методом машинного навчання для задач класифікації та регресії. Його основна ідея полягає у послідовному навчанні набору слабких моделей (зазвичай дерев рішень) з метою покращення загальної моделі. XGBoost відрізняється від інших алгоритмів градієнтного бустингу завдяки його особливій оптимізації, яка має на увазі як локальну, так і глобальну структуру моделі.

Математично, XGBoost мінімізує функцію втрат, яка представляє собою суму функцій втрат для кожного індивідуального прикладу в навчальному наборі. Основний принцип полягає в тому, щоб на кожному кроці додавати нову модель (дерево) до ансамблю таким чином, щоб мінімізувати цю функцію втрат. Основними компонентами XGBoost є дерева рішень, що підсилюються (boosted trees), регуляризація для запобігання перенавчанню та оптимізація швидкодії за рахунок використання оптимізованої структури даних.

До важливих формул, що описують алгоритм, варто віднести (і почати) з функції втрат при оптимізаційній задачі:

$$\text{loss} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

де l - функція втрат для індивідуального прикладу, $\hat{y}_i$ - прогнозоване значення, $\Omega(f_k)$ - регуляризаційна функція для дерева f_k.

Рекурсивна оптимізація дерев: Основна ідея полягає у тому, щоб на кожному кроці розділити навчальний набір даних на дві частини відповідно до певного правила, щоб мінімізувати функцію втрат.

Функція приросту

$$\text{gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right]$$

де G_L і G_R - суми градієнтів в лівій і правій дитинах, H_L та H_R - суми гессіанів (других похідних) в лівій і правій дитинах, λ - параметр регуляризації, γ - поріг розбиття.

Переваги XGBoost включають:

- Висока точність: XGBoost відомий своєю високою точністю завдяки послідовному покращенню моделі за рахунок додавання слабких моделей.
- Регуляризація: Має ефективні методи регуляризації, що дозволяють запобігти перенавчанню.
- Швидкодія: Оптимізована структура даних та процес оптимізації дозволяють швидко побудувати складні моделі.

До недоліків алгоритму можна віднести вимогливість до параметрів. Хоча XGBoost має багато параметрів для налаштування, правильне налаштування може бути складним завданням.Також важливою є чутливість до даних. Може бути чутливим до значних коливань у даних.

В цілому поведінка алгоритму у випадку часових рядів дуже схожа до поведінки випадкового лісу, описаного в розділі 3.3. Тому більшість аргументів стосовно адаптації алгоритму відповідає аналогічним, описаних в тому розділі.

3.5. Рекурентна Нейронна Мережа, LSTM

Рекурентні нейронні мережі (RNN) [98] є потужним інструментом у галузі машинного навчання, який використовується для роботи з послідовними даними. Основна ідея полягає в здатності мережі до роботи з вхідними даними в послідовному порядку, враховуючи контекст попередніх даних. Порівняно зі звичайними шарами нейронів у простих нейронних мережах, де кожен вхід оброблюється незалежно від інших, рекурентні нейронні мережі здатні зберігати і використовувати інформацію з попередніх кроків часу.

Основним компонентом RNN є рекурентний зв'язок, який дозволяє передавати інформацію про попередні стани мережі на поточний момент часу. Формально, для одного часового кроку t, вихід h_t обчислюється на основі вхідного вектора x_tі попереднього стану h_{t-1}:

$$[h_t = \sigma(W_h \cdot [h_{t-1}, x_t] + b_h)]$$

Де:

σ - функція активації (наприклад, сигмоїда або гіперболічний тангенс)

W_h - матриця ваг

$[h_{t-1}, x_t]$ - конкатенація попереднього стану та вхідного вектора

b_h - зсув

Ця формула дозволяє мережі використовувати попередній стан h_{t-1} для обробки поточного вхідного вектора x_t у контексті попередніх даних.

У простих нейронних мережах (NN) без рекурентних зв'язків, кожен вхід оброблюється окремо і не враховує попередніх станів мережі. Це робить NN менш ефективними для роботи з послідовними даними, де контекст та порядок даних мають важливе значення.

Рекурентні нейронні мережі мають декілька варіантів, таких як одношарові RNN (проста структура з одним рекурентним шаром) і багатошарові RNN (з багатьма рекурентними шарами, які дозволяють моделі більш складніші залежності). Однак звичайні RNN також мають свої обмеження.

Основним недоліком звичайних RNN є проблема з пам'яттю довгих залежностей. У таких мережах важко зберігати інформацію про попередні вхідні

дані на великій кількості часових кроків. Це призводить до проблеми втрати контексту при обробці довгих послідовностей. LSTM (Long Short-Term Memory) [99] був розроблений для вирішення цієї проблеми.

Ідея LSTM полягає в використанні спеціальних блоків пам'яті, які можуть зберігати і оновлювати інформацію на тривалий термін. Основна структура LSTM включає вхідні гейти, вихідні гейти і ворота забування, які керують потоком інформації в середині мережі.

Давайте розглянемо основні компоненти LSTM і їх формули:

1. **Вхідні гейти (Input Gate)**:

Вхідні гейти визначають, яка частина нових даних буде оновлювати стан пам'яті. Це контролюється сигмоїдною функцією.

Формула вхідних гейтів:

$[i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)]$

Де:

i_t - вектор вхідного гейту на часовому кроці \(t \)

σ - сигмоїдна функція активації

W_i- матриця ваг

h_{t-1} - попередній вихідний стан

x_t - вхідний вектор на часовому кроці \(t \)

b_i - зсув вхідного гейту

2. **Вихідні гейти (Forget Gate)**:

Вихідні гейти визначають, яка частина попереднього стану пам'яті буде забута. Також контролюється сигмоїдною функцією.

Формула вихідних гейтів:

$$[f_t = \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big)]$$

Де:

f_t - вектор вихідного гейту на часовому кроці \(t \)

W_f - матриця ваг

b_f - зсув вихідного гейту

3. **Оновлення стану пам'яті**:

Стан пам'яті c_t оновлюється з урахуванням вхідних гейтів i_t та вихідних гейтів f_t.

Формула оновлення стану пам'яті:

$[c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)]$

Де:

c_t- новий стан пам'яті на часовому кроці \(t \)

W_c- матриця ваг

b_c - зсув

$\odot$ - покомпонентне множення

tanh - гіперболічний тангенс

4. **Вихідний гейт (Output Gate)**:

Вихідний гейт визначає, яка частина оновленого стану пам'яті буде виведена як вихід з LSTM.

Формула вихідного гейту:

$$[o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)]$$

Де:

o_t - вектор вихідного гейту на часовому кроці \(t \)

W_o - матриця ваг

b_o - зсув вихідного гейту

5. **Вихід на часовому кроці**:

Вихід LSTM на часовому кроці t обчислюється з урахуванням вихідного гейту o_t та тангенсу стану пам'яті c_t.

Формула виходу:

$[h_t = o_t \odot \tanh(c_t)]$

Де:

h_t - вихід на часовому кроці t

Переваги LSTM включають їх здатність до моделювання складних залежностей в часових рядах та мовних данних. Вони дозволяють ефективно

працювати з даними довільної довжини та зберігати внутрішні стани для тривалих залежностей. Крім того, LSTM має здатність автоматично визначати, яку інформацію слід зберігати або забувати залежно від поточного контексту.

Проте LSTM має деякі недоліки. Вони вимагають більш складного навчання через більшу кількість параметрів та гіперпараметрів. Також, підбір оптимальних значень може бути витратним за ресурсами обчислень.

У випадку цієї дисертаційної роботи, модель що використовується для прогнозування отримує на вхід набір дезагрегованих часових рядів, які пропускаються через низку шарів, як LSTM шарів так і звичайних, щоб перетворити багатовимірний вектор на вході в одне число на виході з алгоритму.

Оцінка параметрів, кількості та типу шарів, кількості епох визначається одним з двох способів: експериментально, себто на основі якості вихідного результату моделі, або відповідно до літератури. Інші та більш специфічні методи не є метою цієї дисертаційної роботи. Детальніше про специфікацію цієї мережі та її експериментальну оцінку буде в розділі 4.

Almosova та Andersen у 2023 [100] досліджували ефективність прогнозування інфляції за допомогою рекурентних нейронних мереж (RNN). Вони використовують модель довготермінової пам'яті рекурентної нейронної мережі (LSTM), і обчислюють уніваріативні прогнози щомісячної інфляції США. Автори показують, що хоча LSTM трохи перевершує авторегресійну модель (AR), NN та моделі переключення Маркова, його ефективність на рівні з сезонною авторегресійною моделлю SARIMA. Крім того, вони проводять аналіз чутливості з урахуванням гіперпараметрів і надають якісне тлумачення того, як навчаються мережі, застосовуючи нову техніку розповсюдження значимості шар за шаром. Автори виокремлюють три основні переваги використання LSTM для прогнозування інфляції. По-перше, LSTMs є гнучкими та керованими даними, тобто дослідник не повинен визначати точну форму нелінійності, а замість цього LSTM може виводити її з даних самостійно. По-друге, за універсальною теоремою апроксимації (Cybenko, 1989) [101], при деяких м'яких умовах LSTMs та нейронні мережі будь-якого типу можуть довільно точно апроксимувати будь-

яку неперервну функцію. По-третє, LSTMs були розроблені спеціально для аналізу послідовних даних і виявилися дуже успішними в цій задачі. Нарешті, останні розробки оптимізаційних методів для NN та бібліотек, які використовують комп'ютерні GPU, зробили навчання NN та RNN значно більш доступним. Отже, результати показують, що LSTM є ефективною нелінійною моделлю для прогнозування інфляції, і його використання може бути виправданим, за умови, що наявно достатньо спостережень для оцінки моделі. Це також є суттєвим внеском та орієнтиром для роботи через використання алгоритму LSTM для прогнозування інфляції в цій роботі.

3.6. Запропонований комбінований підхід SARIMA+LSTM

У цій дисертаційній роботі ми побудуємо комбінований підхід, що визнає складну динаміку інфляції та спрямовується на захоплення як лінійних, так і нелінійних частин динаміки компонентів індексу споживчих цін (CPI) окремо, використовуючи для цього моделі SARIMA та LSTM разом. Обидві ці моделі описані в попередніх розділах й загальна математична складова особливо не відрізняється, тому в цьому розділі буде загальний опис алгоритму й формалізована версія, проте всі детальніші пояснення математики за алгоритмами можна побачити в минулих розділах.

Перший крок цієї моделі полягає у використанні SARIMA для декомпозиції часового ряду на лінійно-пояснювані, сезонні та залишкові компоненти. Це може не повністю враховувати всі стохастичні флуктуації в даних і нелінійності, тому очікується, що вони будуть у залишковій складовій. Тому ми зосереджуємося на них, які представляють невідому варіацію досліджуваних рядів після видалення сезонності та лінійних авторегресійних ефектів і потребують концентрації на нелінійних ефектах і взаємозв'язках.

Далі ми використовуємо нейронну мережу LSTM, передаючи залишкові компоненти від SARIMA у вигляді вхідних матриць в модель LSTM. Ми дозволяємо нейронній мережі вловлювати ці витончені відносини і нелінійні патерни, присутні в непоясненій частині даних. Вбудовані можливості пам'яті у LSTM дозволяють йому ефективно захоплювати динаміку рядів, включаючи

інформацію з минулих спостережень для точного прогнозування майбутніх значень.

Через цей гібридний підхід ми використовуємо доповнюючі переваги SARIMA і LSTM для підвищення точності прогнозування. Комбінуючи міцну статистичну основу SARIMA з можливостями глибокого навчання LSTM, модель пропонує комплексне рішення для прогнозування часових рядів, що дозволяє приймати більш обґрунтовані рішення аналітикам у відповідь на змінні умови в сфері дослідження.

Переходячи до математичної складової алгоритму, його можливо описати в 5-ти ключових пунктах:

- Використовуємо SARIMA модель щоб отримати оцінені значення ряду $(\widehat{y_t^i})$ та шокову непоясненну складову ε_t^i для групи i (відповідно до поділу, зробленного в рамках алгоритмів, описаних у розділі 2), також цю вправу необхідно буде зробити й для загального агрегованого ряду й отримати результат без приставки i.
- $y_t - \sum(\widehat{y_t^i})$, а саме різниця між агрегованим рядом та сумою спрогнозованих на минулому етапі компонент, себто по суті помилка моделі SARIMA, буде закинута в LSTM модель як ряд, котрий LSTM модель тренуватиметься прогнозувати. В той же час, ε_(t-q)^i та ε_(t-q), себто помилки SARIMA моделі будуть складовими вектору, що згодовується моделі у вхідних даних. Більш того, тут використовуватиметься декілька підходів, коли береться тільки вектор значень на q періодів назад (враховуючи що у вправі робиться прогноз на q періодів вперед), але також й підхід де береться низка значень з різних періодів: q, q-1, q-2,… періодів тому.
- Використати SARIMA для прогнозування на q періодів вперед та отримати $\left(\widehat{y_{t+q}^i}\right)$ в той же час отримуючи ε_t^i з, власне, оцінки моделі

- Використати вектор з елементами ε_t^i та ε_t (в певній специфікації й минулі $\varepsilon_{t-1}^i, \varepsilon_{t-2}^i, \ldots$ як вхідний вектор для пренатренованої LSTM моделі щоб отримати як вихід число, що відповідає значенню ε_{t+q}.
- Це дозволяє побудувати загальний прогноз $\sum(\widehat{y_{t+q}^i}) + \varepsilon_{t+q}$ з результатів пунктів 3 та 4.

Повністю описана методологія в цьому розділі є фундаментом для вибудови пайплайну і його оцінки в Розділі 4. Важливим моментом є те, що архітектура (не оцінені коефіцієнти, а саме кількість та тип шарів) LSTM моделі є таким самим як і в звичайної моделі.

3.7. Методи оцінки якості прогнозної моделі

Оцінка якості моделей часових рядів є ключовим етапом у процесі аналізу даних в часі. Основна мета полягає в тому, щоб визначити, наскільки добре модель прогнозує майбутні значення на основі наявних даних. Існує кілька загальних концепцій оцінки якості моделей часових рядів.

Перш за все, одним із способів оцінки є порівняння прогнозованих значень зі справжніми даними. Це означає розрахунок різниці між прогнозованими і фактичними значеннями для кожного моменту часу. Ця різниця відображає точність моделі у відтворенні динаміки часового ряду. Мова про те, наскільки добре модель пояснює наявні дані, що відповідає й надихнуте ідеями R квадрату для лінійних регресій. Проте цей підхід не достатньо якісно відповідає на питання прогностичних можливостей моделей, а зконцентрований на «пояснювальній здатності». Це є суттєвим недоліком.

Другий підхід до оцінки якості моделі полягає у використанні статистичних метрик, що враховують різні аспекти прогнозу. Такі метрики можуть оцінювати якість прогнозів з точки зору середньої абсолютної помилки, середньої квадратичної помилки, часових рядів тощо. Вони надають числову оцінку ефективності моделі, яка дозволяє порівнювати різні моделі за їх точністю. Саме цей підхід є найбільш стандартним та ключовим в аналізі часових рядів.

Третій метод оцінки моделей часових рядів полягає у використанні методів перехресної перевірки (cross-validation). Це дозволяє оцінити ефективність моделі на основі різних підвибірок даних. Зазвичай використовується метод поділу даних на навчальні та тестові набори для оцінки прогнозів. Цей підхід допомагає визначити стійкість моделі та уникнути перенавчання (overfitting) на конкретних даних. Проте в контексті часових рядів він не дуже поширений через низку обмежень. Зокрема неможливість взяти дійсно випадкові підмножини даних через структурованість за часом. Також це неефективно для часових рядів не з денною періодичністю, адже цей підхід вимогливий до кількості даних. Але ідея різних оцінок на різних семплах, що походить з цього методу, активно використовується в аналізі якості моделей часових рядів.

Важливим аспектом при оцінці якості моделей часових рядів є аналіз різних аспектів прогнозів, таких як точність, стійкість до змін у вхідних даних, швидкість прогнозування тощо. Вибір оптимального методу оцінки залежить від конкретних характеристик даних та задачі прогнозування. Тому ми закцентуємо увагу саме на другому підході.

3.7.1. RMSE

Одним із методів оцінки якості моделей часових рядів є Root Mean Square Error (RMSE) - квадратний корінь з середньої квадратичної помилки [102]. Цей показник використовується для вимірювання середнього рівня відхилення прогнозів моделі від справжніх значень часового ряду.

Формула для RMSE виглядає наступним чином:

$$[RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}]$$

Де:

n - кількість спостережень у часовому ряді

y_i - справжні значення часового ряду в момент часу \(i \)

$\widehat{y_i}$ - прогнозовані значення часового ряду в момент часу \(i \)

RMSE вимірюється у тих самих одиницях, що і вихідні дані (наприклад, одиниці виміру часового ряду). Чим менше значення RMSE, тим краще прогнозує модель.

Цей метод оцінки часто використовується у задачах прогнозування часових рядів, оскільки він дає інтерпретований результат у тих самих одиницях, що й вихідні дані. Через використання квадрату різниці між справжніми і прогнозованими значеннями, RMSE надає більше ваги великим помилкам, що робить його чутливим до викидів (outliers) у даних.

Іншою перевагою RMSE є його інтерпретованість, оскільки він може бути безпосередньо порівняний з одиницями виміру часового ряду. Проте RMSE також має деякі недоліки, зокрема він не є робастим до великих відхилень у даних та може підсилити ефект викидів. Також він може бути більш чутливим до великих значень, оскільки використовує квадрат помилок.

У практичних застосуваннях, RMSE часто використовується як основна метрика для порівняння різних моделей часових рядів. Він дозволяє оцінити точність прогнозів та зробити висновки щодо ефективності моделі на підставі середньої квадратичної помилки.

3.7.2. RMSE з вікном, що розширюється

Метод розширюючого вікна RMSE працює наступним чином: спочатку модель навчається на початковому наборі даних (наприклад, перших t спостережень), і на цій основі робиться прогноз для наступного моменту часу (t+1). Потім модель перенавчається на розширеному наборі даних (перші (t+1) спостережень) і знову робиться прогноз для моменту часу (t+2). Цей процес повторюється до кінця часового ряду. Цей підхід дозволяє отримати більш реалістичну оцінку точності моделі, оскільки враховує динаміку прогнозування на кожному кроці.

Порівняно зі звичайним RMSE, метод розширюючого вікна дозволяє більш точно відобразити реальну точність моделі на нових даних. Він уникне перенавчання моделі на всьому наборі даних і дозволяє періодично оцінювати її ефективність на зростаючій кількості даних. Це особливо корисно в задачах, де

характеристики часового ряду можуть змінюватися з часом, і модель повинна адаптуватися до нових умов.

3.8. Висновки до Розділу 3

В рамках розділу було розглянуто низку методів, котрі використовуватимуться для прогнозування й будуть оціненими у наступному, 4-му, розділі. Зокрема мова про такий класичний бенчмарк як Випадкове блукання, класичну модель економетричного аналізу SARIMA, найпоширеніши приклади алгоритмів машинного навчання Випадковий ліс та XGBoost, нейромережевий алгоритм LSTM, а також розроблений автором пайплайн з алгоритмів SARIMA та LSTM. Також в розділі описуються методи оцінки якості прогнозу, що надалі використовуватимуться для цієї вправи.

З ключових висновків можна вказати те, що вищезазначені моделі можна використовувати різними способами. Наприклад, перші два описані методи є уніваріантними, тому їх можна використовувати на агрегованому ряді, але також можна і використати на зібраних компонентах після моделей з другого розділу, а далі, з певними вагами, зібрати результуючі ряди в агрегований. Самі ж ваги можна визначити аналітично, статистично, або використавши метод лінійної регресії на історичних даних. Методи машинного навчання та нейромережевих алгоритмів здатні приймати не тільки значення в минулий період, але також значення в ще більш минулих періодах одночасно, що дозволяє якісніше використати попередні дані й, власне, структуру наданих часових рядів.

РОЗДІЛ 4. Результати побудованих моделей на базі даних дезагрегованих компонент інфляції

Цей, фінальний, розділ дисертаційної роботи присвячений розгляду моделей, описаних в розділах 2 та 3 та їх прогностичних можливостей на базі даних компонент інфляції в Україні. Важливість цієї задачі, а саме якісних прогнозів інфляції для України, є значною, саме тому розробка все більш глибоких моделей що досліджують та прогнозують інфляцію є суттєвим прогресом для України. Якісні прогнози інфляції відіграють роль для суспільства в цілому, як певний орієнтир що допомагає в плануванні власних дій та витрат. Він важливий для уряду та бізнесів також для короткострокового, середньострокового та довгострокового планування своїх бізнес-стратегій, політик, інвестиційних програм та іншого.

В рамках розділу, ми детальніше розглянемо базу даних та проблеми, що присутні в цій базі даних та які потребують вирішення. Далі ми використаємо моделі, описані в розділах 2 та 3 на цьому датасеті й, спершу, охарактеризуємо результати для моделей кластеризації, а згодом й для моделей прогнозування. Це дозволить зробити всеохоплююче порівняння між великою кількістю моделей й привести емпіричні докази кращих прогностичних властивостей для певних моделей, використовуючи метрику RMSE з вікном, що розширюється.

Ці результати дозволять вибудувати дискусію, що торкнеться як загальних показників ефективності моделей, так і більш грунтовне пояснення чому ті чи інші варіанти є більш продуктивними. У підсумку ця дисертаційна робота відкриє можливість запропонувати найкращі моделі для прогнозування інфляції до використання в організаціях. Більше того, деякі з них вже наразі використовуються Національним Банком України станом на 2024 рік через достатню низку доказів про їх ефективність у прогнозуванні. Звісно, ці моделі не є ідеальним інструментом й додаткові моделі та експертні судження відповідно до інших статистичних показників та екзогенної інформації є необхідними.

4.1. Опис даних

Ця дисертаційна робота для оцінки якості моделей та демонстрації їх можливостей використовує часові ряди даних у галузі економіки, продовжуючи традицію моїх попередніх робіт автора. Ті роботи досліджували схожі, проте менш розвинені техніки та моделі з різних боків, проте теж на економічних або на зсимульованих даних. Конкретно в цій роботі основна увага приділяється базовому індексу споживчих цін (Core CPI) в Україні та розробці відповідних прогностичних моделей. Ці моделі намагаються здійснити найкращий можливий прогноз, враховуючи багато підкомпонентів загального індексу та їх нелінійні взаємозв'язки один з одним в певній частині моделей, описаних в Розділі 3. Таким чином, глибоке дослідження властивостей цих підкомпонентів та підтримуючих змінних є невід'ємною частиною успішних досліджень, і увага, навіть в більш технічній роботі, повинна бути приділена даним.

Індекс споживчих цін (CPI) вимірює загальні витрати на кошик товарів і послуг, що представляють середній споживчий розхід. Державна служба статистики України (Укрстат) збирає дані рівня цін (CPI) за понад 400 категоріями, охоплюючи продукти харчування, одяг, житло, транспорт і інше. Дані CPI дозволяють перевіряти зміни цін та витрат на життя. Це одна з двох ключових змінних, разом з обмінним курсом, за допомогою яких люди відстежують багато сприйняттєвих та очікуваних показників. Наприклад, сприйняття благополуччя, вибір споживання проти заощадження, очікування щодо майбутньої економіки та багато інших показників. До речі, прогнозування також має велике значення для розуміння майбутнього стану речей, таких як мінімальна заробітна плата, вартість життя та інші, що часто використовуються в податкових ініціативах, що є важливим для планування бюджету. Таким чином, прогнозування таких змінних є ключовим для більшості галузей економіки. Чимало робіт зосереджено на прогнозуванні базового індексу споживчих цін з різними методами, починаючи від традиційних до високотехнологічних, як для різних країн, так і для України.

Базовий індекс споживчих цін фокусується на більш стабільних компонентах CPI, які менше піддаються адміністративним рішенням, погодним впливам або глобальним ціновим шокам. Він охоплює оброблені продукти харчування, одяг, послуги та інші категорії. В даній роботі використовуються щомісячні дані базового індексу споживчих цін, а також його компонент, з січня 2007 року по грудень 2023-го року. Деякі компоненти з'являються в даних пізніше (у 2012-му році або навіть у 2017-2022-х роках), оскільки через кожні п'ять років набір рядів даних переробляється. Поріг для появи ряду даних полягає в тому, що товар має складати більше 0,1% річного кошика товарів для середнього українця. Таким чином, автомобіль, навіть якщо купується нечасто через надзвичайно високу вартість, з'являється в наборі даних. Так само, як хліб або ковбаса, які купуються практично щодня.

Для забезпечення ефективного прогнозування дані потребують обробки для досягнення стаціонарності та врахування сезонності. Стаціонарність частково вирішується за допомогою метрики базової інфляції (відсоткова зміна базового індексу споживчих цін) замість абсолютних рівнів. Сезонність — це інша історія. Для деяких моделей вона автоматично враховується, наприклад, для моделі SARIMA, яку ми порівняємо. Однак попередні дослідження показують, що чиста модель LSTM не здатна ідеально вловити сезонність. Крім того, деякі ряди показують зміну сезонності, яка з'являється в певний момент часу через зміни методології.

Останнім питанням, яке потрібно вирішити, є сегмент набору даних, який охоплює 335 компонентів. Як вже зазначалося, деякі компоненти були включені Укрстатом значно після 2007 року, що призвело до розповсюдження пропущених значень (NA). Хоча існують різні підходи до подолання цієї проблеми, обраний і найпростіший метод полягає в видаленні всіх стовпців з пропущеними значеннями і подальшому налаштуванні ваг, зменшуючи загальну кількість компонентів з 335 до 270.

4.2. Оцінка ефективності моделей прогнозування інфляції

В цій дисертаційній роботі результати діляться, в цілому, на дві частини. Перша частина – результати роботи алгоритмів, описаних в Розділі 2, а в другій частині розглянемо результати роботи алгоритмів оцінених на основі результатів першої частини моделями з Розділу 3.

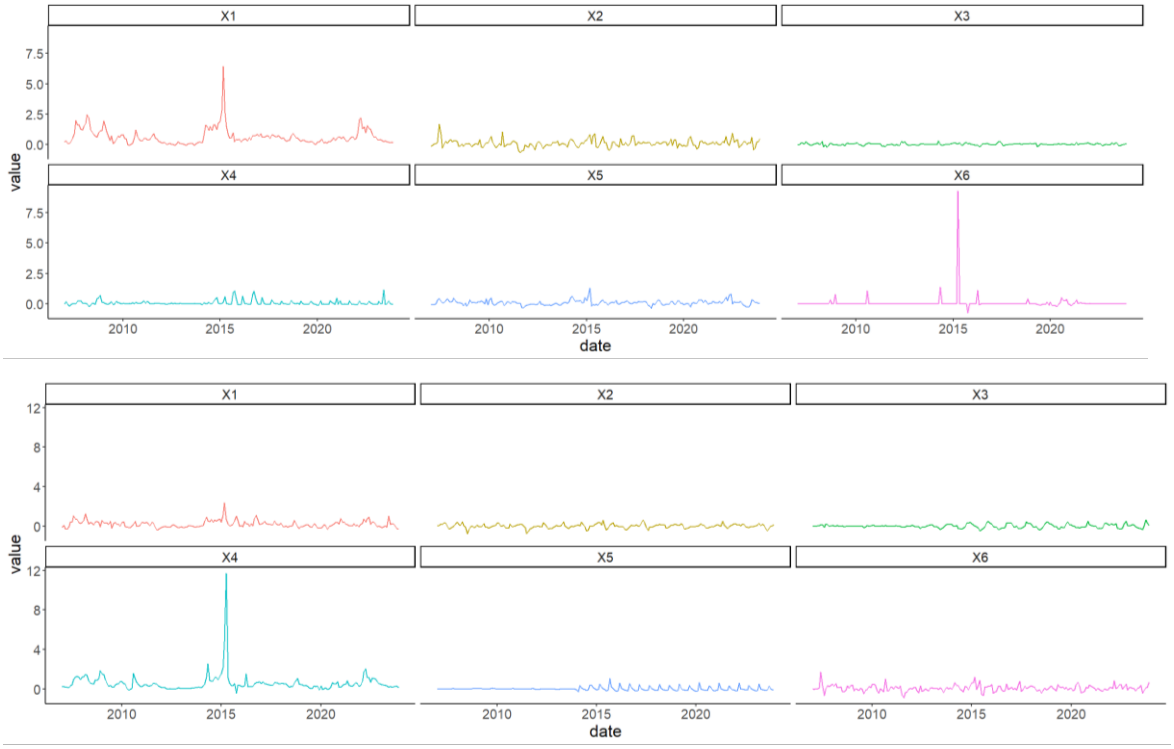
Отже, першим етапом є кластеризація з використанням алгоритмів K-Means, DBSCAN та Hierarchical Clustering на основі набору точок, отриманого з матриці дистанцій, котра в свою чергу отримана методами евклідової відстані, кореляційної відстані та, насамкінець, методом Dynamic Time Warping з адаптаціями, розробленого спеціально для цієї дисертаційної роботи.

Варто нагадати, що в рамках поліпшеного DTW алгоритму, до рядів також додається незначний шок, а саме ряд побудований на основі білого шуму з дуже низькою варіацією, в 10 разів меншою за варіацію основного ряду у відповідному році й з математичним очікуванням рівним нулю. Це зроблено задля того щоб запобігти проблемі, описаній в одній зі статей автора, коли ряд має нульову зміну впродовж довгого періоду часу й DTW алгоритм не може знайти якісно відповідну точку, що призводить до проблем оцінки алгоритму. Це специфічна проблема низки рядів, наприклад інфляції цін на вищу освіту, що змінюється 2 рази на рік (себто має суттєву сезонність, але недостатньо стабільну за розміром, щоб бути викоріненою алгоритмами врахування сезонності). Відповідно до однієї зі статей автора, додавання такого шоку незначно (або взагалі не) впливає на роботу алгоритмів прогнозування, проте поліпшує розділення рядів алгоритмом DTW, не змінюючи результати там, де ця проблема відсутня.

Отже, одразу в фігурі 1 покажемо приклад дії алгоритму й розділення на категорії, що підводить до висновку що краще працювати із сезонно-скорегованими рядами як на нижній половині графіку. Причиною тому є, в першу чергу, наявність явної та різної сезонності в більшості категорій верхнього графіку, що говорить про слабкість алгоритму в контексті підхоплення й розподілення сезонних компонент в одну категорію. Це є одним з двох

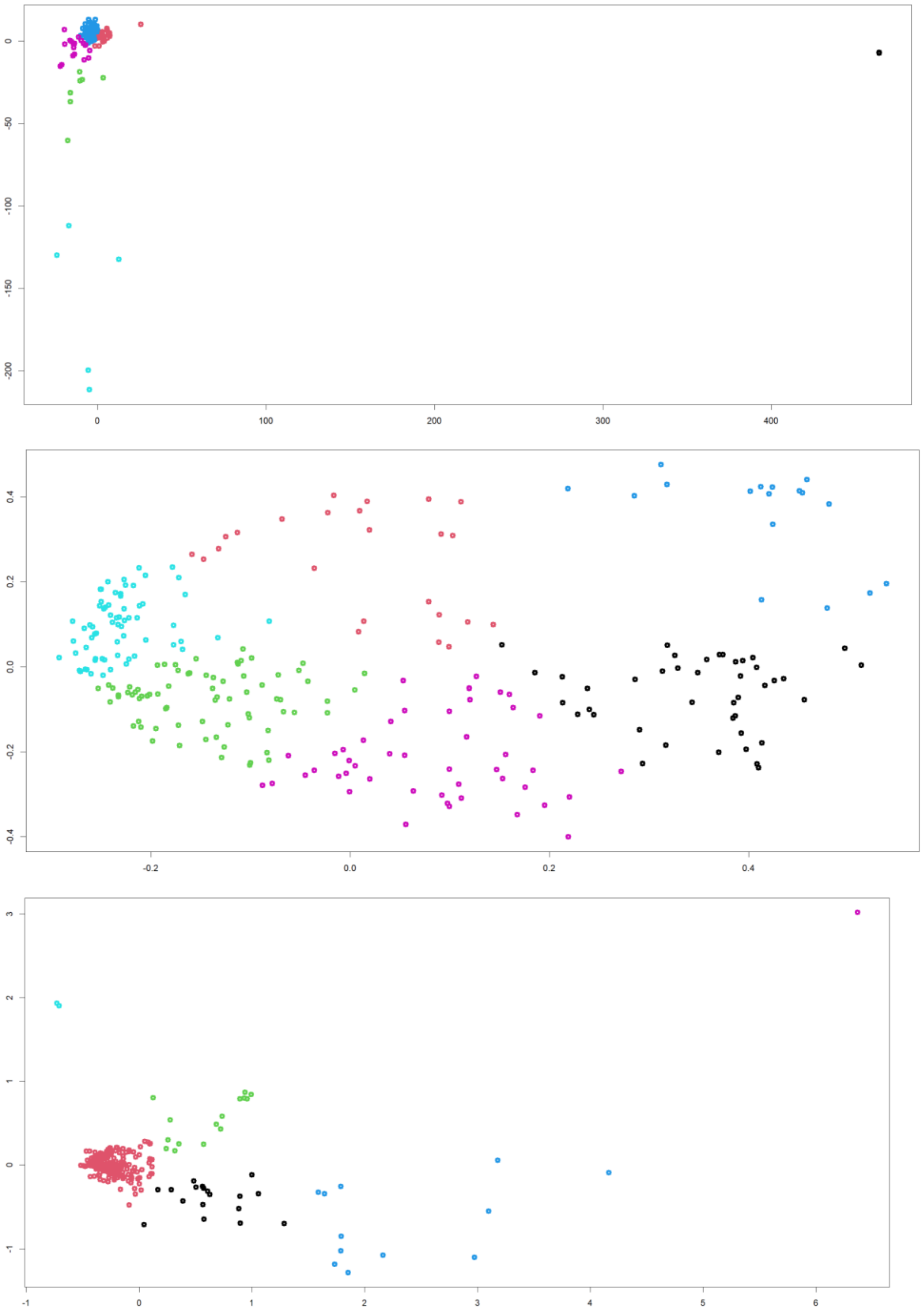
аргументів на користь сезонно-скоригованої бази даних, іншим є якість прогнозу моделей, про що поговоримо детальніше далі в цьому розділі.

Фігура 1. DTW категорії на сезонно-нескоригованих (верхній набір графіків) і сезонно-скоригованих (нижній набір графіків) рядах, розділені алгоритмом K-Means.



Тепер, є сенс порівняти, власне, розділення рядів на компоненти за різними методологіями пошуку відстаней: евклідова відстань, відстань на базі кореляції, адаптований DTW. Ці приклади по черзі наведені в Фігурі 2.

Фігура 2. Двовимірні площини з точками, що відповідають рядам, де відстані є евклідовими (верхній графік), кореляційно-заснованими (середній графік) та побудованими з використанням адаптованого DTW (нижній графік) та розділені за допомогою алгоритму K-Means.



На графіку вище явно видно наскільки «проміжним» варіантом є поліпшений DTW алгоритм й це найбільш відповідає природі даного набору даних відповідно до загальної економічної логіки. Це є одним із аргументів стосовно високої якості алгоритму, а також додатковою причиною акцентувати увагу саме на результатах цього алгоритму в комбінації з моделями прогнозування, розглянутими в розділі 3.

Надалі розглянемо низку таблиць й ознайомимося з результатами певних моделей, оцінених на проміжку з початку з початку 2021-го до кінця 2023-го як розписано в Розділі 4.1.

Таблиця 1. RMSE описаних моделей прогнозування на основі компонент, створених агрегацією схожих рядів за методом евклідових відстаней

RMSE моделей	1m	2m	3m	4m	5m	6m
RW	1.073	1.167	1.117	1.144	1.161	1.181
RW (SA)	0.992	1.092	1.053	1.039	1.089	1.111
SARIMA	0.914	0.961	0.989	1.003	1.008	1.001
SARIMA (категорії)	0.915	1.098	0.990	1.061	1.016	0.826
ARIMA (SA компоненти)	0.945	0.837	0.816	0.815	0.948	0.895
LSTM	0.916	1.009	0.966	1.116	0.916	1.041
LSTM (SA компоненти)	0.731	0.814	0.878	0.865	0.809	0.840
LSTM+SARIMA routine	**0.645**	**0.886**	**0.707**	**0.849**	**0.883**	**0.684**
Random Forest	0.728	0.871	0.819	0.901	0.978	1.060
Random Forest (SA компоненти)	0.649	0.872	0.678	0.825	1.092	0.997
XGBoost	0.744	1.055	0.986	1.130	0.956	1.311
XGBoost (SA компоненти)	0.814	0.733	0.814	0.913	0.999	0.952

В Таблиці 1 наведені результати моделей за виміром RMSE з ковзним вікном, оціненим на проміжку з початку 2021-го року до кінця 2023-го року. Моделі побудовані на базі розподілу за евклідовими відстанями та на базі найкращого для цих результатів методу кластеризації. Саме в цьому випадку виявилось, що для всіх, окрім XGBoost алгоритму, де відіграв DBSCAN, найкращим виявився K-Means. Стосовно позначень моделей, перші 3 є алгоритмами, що були використані на загальному ряді базової інфляції без поділу на компоненти. Далі вже йде поділ на компоненти й пошук «категорій» алгоритмом евклідової відстані вкупі з кластеризацією. Приставка SA компоненти відповідає за ті випадки, коли першопочаткова база даних з 335 компонент була попередньо сезонно скоригована. Аналогічним методом побудуємо таблиці для випадку коли відстані знаходилися кореляційним (Таблиця 2) та адаптованим методом DTW (Таблиця 3).

Таблиця 2. RMSE описаних моделей прогнозування на основі компонент, створених агрегацією схожих рядів за методом кореляційних відстаней.

RMSE моделей	1m	2m	3m	4m	5m	6m
RW	1.073	1.167	1.117	1.144	1.161	1.181
RW (SA)	0.992	1.092	1.053	1.039	1.089	1.111
SARIMA	0.914	0.961	0.989	1.003	1.008	1.001
SARIMA (категорії)	0.983	0.929	1.083	1.080	1.128	1.082
ARIMA (SA компоненти)	0.960	0.906	1.025	0.994	1.005	1.000
LSTM	1.035	1.104	0.956	1.024	1.021	1.004
LSTM (SA компоненти)	0.830	0.948	0.917	0.957	0.886	0.954
LSTM+SARIMA routine	**0.655**	**0.830**	**0.771**	**0.907**	**0.804**	**0.842**
Random Forest	0.746	0.870	0.957	0.944	1.055	0.979
Random Forest (SA компоненти)	0.719	0.967	0.858	0.840	1.023	0.901
XGBoost	0.846	1.210	1.147	1.249	1.173	1.186
XGBoost (SA компоненти)	0.892	0.928	0.938	0.894	0.862	1.007

Таблиця 3. RMSE описаних моделей прогнозування на основі компонент, створених агрегацією схожих рядів за методом адаптованого алгоритму DTW.

RMSE моделей	1m	2m	3m	4m	5m	6m
RW	1.073	1.167	1.117	1.144	1.161	1.181
RW (SA)	0.992	1.092	1.053	1.039	1.089	1.111
SARIMA	0.914	0.961	0.989	1.003	1.008	1.001
SARIMA (категорії)	0.831	0.887	0.899	0.913	0.901	0.954
ARIMA (SA компоненти)	0.796	0.825	0.834	0.826	0.868	0.900
LSTM	0.830	0.939	0.858	0.896	0.931	0.901
LSTM (SA компоненти)	0.651	0.777	0.756	0.785	0.749	0.801
LSTM+SARIMA routine	**0.575**	**0.683**	**0.611**	**0.631**	**0.700**	**0.734**
Random Forest	0.665	0.808	0.800	0.821	0.898	0.902
Random Forest (SA компоненти)	0.604	0.713	0.749	0.754	0.838	0.818
XGBoost	0.674	1.038	0.954	1.043	0.989	1.100
XGBoost (SA компоненти)	0.638	0.744	0.766	0.753	0.791	0.856

В Таблиці 2 та Таблиці 3 знову переважна таблиця результатів була на основі алгоритму K-Means в контексті кластеризації з нечастими виключеннями, більшість яких була в таблиці 2, там частим переможцем був алгоритм DBSCAN. Hierarchical Clustering дав непогані результати з усіх трьох таблиць тільки у випадку ARIMA з сезонно-скорегованим датасетом й кореляційними відстанями.

Також хочеться до прикладу навести Таблицю 4, що показують результати моделей машинного навчання та нейромережевих при використанні їх на всьому датасеті, себто на всіх 335 компонентах інфляції.

Таблиця 4. RMSE описаних моделей прогнозування на основі бази даних без агрегації на компонент методами пошуку відстаней.

RMSE моделей	1m	2m	3m	4m	5m	6m
LSTM (SA 335)	0.815	0.904	0.877	0.888	0.811	0.870
LSTM+SARIMA routine (335)	0.758	0.745	0.683	0.762	0.786	0.832
Random Forest (SA 335)	0.652	0.863	0.841	0.879	0.932	0.926
XGBoost (SA 335)	0.751	0.868	0.917	0.918	0.901	0.964

Більше деталей та дискусії стосовно результатів в Розділі 4.3.

4.3. Дискусія стосовно якості моделей

Найголовнішим висновком цієї дисертаційної роботи, як можна побачити з таблиць, поданих в минулому розділі, є те, що основна модель дає найкращі результати в порівнянні з усіма іншими. Це є далеко не єдиним висновком, що можна зробити з результатів поданих в цій роботі, тому давайте детальніше розглянемо результати й інших моделей, алгоритмів, задля того щоб зробити висновки, які саме аспекти найбільше покращують якість прогнозування.

По перше, варто відзначити, що в більшості таблиць, в різних специфікаціях, нейронні мережі показували дуже непоганий результат, багато в чому кращий за інші моделі, особливо в прогнозуванні на більш короткий горизонт, а саме на 1-3 місяці вперед. Вищезазначені тези застосовуються як моделі, котра є комбінацією LSTM та SARIMA, так і звичайна LSTM моделі на базі сезонно-скоригованих компонент.

Важливим фактом є те, що моделі машинного навчання також показують доволі непогані результати. Незважаючи на те, що вони гірші за нейромережеві моделі, ці моделі будуються набагато швидше та простіше. І все одно вони показують результат кращий за класичні бенчмарки.

До речі, стосовно бенчмарків, моделі ARIMA перевершують модель випадкового блукання в усіх специфікаціях на усіх проміжках часу, і це

відповідає загальному консенсусу в літературі, що моделі ARIMA мають певну прогнозну здатність. Не ідеальну, але тим не менш. В контексті прогнозування інфляції.

Також цікавим фактом є те, наскільки сильно в усіх специфікаціях сезонне коригування на дезагрегованому датасеті поліпшує якість фінальних прогнозів. Цей результат є негативним в контексті нездатності алгоритмів машинного навчання та нейромежевих алгоритмів підхоплювати сезонність на місячних даних, проте з іншого боку рішення проблеми знайдене і є достатньо ефективним.

Стосовно алгоритмів пошуку відстаней, результати одностайно й практично по всім моделям та по всім дальностям прогнозування приводять до наступного висновку: адаптований алгоритм DTW перевершує евклідову відстань, що, в свою чергу, перевершує кореляційну відстань. Це є однозначним і, насправді, цілком логічним висновком, адже поточний датасет доволі сильно страждає від проблем стиснення та лагової залежності один від одного рядів. Наприклад, ціни на консервоване м’ясо набагато повільніше, проте приблизно на тому ж рівні, реагує на шок цін на сире м’ясо як і сосиски, котрі реагують раніше. Й таких прикладів дуже багато, через що адаптований DTW якісніше впорується з задачею.

Більше того, якщо говорити про алгоритми кластеризації, то найякісніший результат показує K-Means (абсолютно ідентичний до K-Means++) алгоритм. Всі інші показували гірші або незначно гірші результати практично по всім моделям. Це є якісним неекономічним обгрунтуванням переваги алгоритму.

Насамкінець, важливо також написати що нейромережеві алгоритми та алгоритми машинного навчання, при використанні їх на повному датасеті, не впоруються в контексті якості прогнозів через занадто високу зашумленість даних, що зайвий раз підкреслює значущість методів, описаних у Розділі 2, задля прогнозування базової інфляції.

В цілому, висновок який можна зробити у підсумку цієї роботи полягає в тому, що комбінація алгоритмів DTW+K-Means+SARIMA+LSTM на сезонно-

скоригованих рядах показує найякісніший результат, проте це не стовідсотково визначає те, що на будь-якому іншому датасеті саме цей алгоритм буде найкращим. Проте є висока потреба розглядати саме цей алгоритм як один з потенційно-найкращих. Також, як алгоритм для «першої спроби» варто також розглядати Випадковий Ліс, оскільки будувати його набагато швидше та простіше, він набагато менш ресурсозатратний й, тим не менш, показує якісні результати що суттєво перевершують моделі випадкового блукання та традиційні типу ARIMA, хоч і не дотягують до вищезазначених нейромережевих.

ВИСНОВКИ

Дисертаційну роботу присвячено методам прогнозування часових рядів, зокрема економічної природи, на основі нейромережевих алгоритмів, кластеризації та інших методів машинного навчання, а саме – розробці моделей та методів підвищення точності та ефективності прогнозування, що було протестовано на базі часових рядів зміни цін на окремі товари споживчого кошику громадян України.

У роботі докладно розглянуто існуючі методи для розв’язання поставлених задач та запропоновано вдосконалення їх характеристик. Тематика та суть розробок лежить на перетині галузей інформаційних технологій та економіки, але саме рішення – спирається на методи програмної інженерії, комп’ютерних наук та математики.

Головний результат дисертаційної роботи – розроблено низку моделей для прогнозування й оцінено їх прогностичні властивості на базі даних компонент базової інфляції. До найбільш якісних методів належать як нейромережеві алгоритми, так і методи машинного навчання. Проте дослідження показало що самі по собі вони дають посередні результати, а найкращі виходять тоді, коли вони об’єднуються з методами кластеризації дезагрегованого датасету, використовуючи методологію пошуку відстаней між часовими рядами. При тому що найкращий результат показує саме нейромережевий алгоритм, інші алгоритми машинного навчання, зокрема Random Forest, також мають доволі високий рівень якості, проте значно простіші для побудови як з точки зору експериментального визначення параметрів, так і з точки зору обчислювальних здатностей, що необхідні для навчання.

В результаті проведеної роботи вирішено такі наукові задачі:

1. Проведена робота з вибудови історичного контексту та передумов популяризації нейромережевих алгоритмів та алгоритмів машинного навчання у економічній сфері. Традиційні методи економетрики, такі як ARIMA, SARIMA, VAR, QPM та DSGE, протягом тривалого часу використовувалися як основа економічного аналізу, що допомагало у

розумінні економічних тенденцій та закономірностей. Проте з появою великих обсягів даних та передових обчислювальних технік економісти все частіше звертаються до передових алгоритмів машинного навчання та інструментів аналізу даних для поліпшення своїх можливостей у прогнозуванні, оскільки традиційні алгоритми не так ефективно враховують нелінійні зв'язки та не можуть ефективно працювати з великими базами даних, що містять десятки або навіть сотні часових рядів. До нових алгоритмів належали нейромережеві, такі як проста нейронна мережа, RNN та LSTM, алгоритми машинного навчання на кшталт Random Forest та XGBoost, методи кластеризації.

2. Детально розкривається яким саме чином проводиться кластеризація часових рядів й які ключові здобутки та доповнення були зроблені в цьому напрямку. Це робиться побудовою матриці дистанцій між рядами, трансформацією матриці спеціальним алгоритмом у двовимірний простір, насамкінець використанням алгоритму кластеризації. Для побудови дистанцій між часовими рядами використано декілька підходів, евклідова відстань, кореляційна відстань, метод динамічного вирівнювання часу. Але, найголовніше, був розроблений адаптований метод динамічного викривлення часу. Однією з характеристик розглянутих в роботі рядів є їх періодичність, яка може бути денною, місячною або квартальною. Цю особливість зазвичай враховують за допомогою різних методів, але в даній роботі є використання спеціальної маски, що обмежує матрицю шляху, ряди та їх відповідність до календарного року, що дозволяє відповідності бути в рамках року. Також адаптований алгоритм вирішує проблему багатьох однакових точок, використовуючи додатковий незначний шум, що не викривлює результатів інших алгоритмів, проте допомагає відійти від вищеназваної проблеми.
3. Побулова моделей прогнозування на основі зібраних кластерів, визначених алгоритмами кластеризації й відстаней між часовими

рядами. Було побудовано різні моделі, від стандартних типу ARIMA, до методів машинного навчання Random Forest та XGBoost й нейромережевих LSTM. Проте було розроблено спеціальний метод що є комбінацією двох вищезазначених SARIMA+LSTM. Цей метод дозволяє одночасно якісно підхоплювати власні сезонні та лінійні коливання компонент, так і нелінійні складові та взаємозалежності між рядами. Це є значним просуванням в області якісного прогнозування.

4. Насамкінець, було використано базу даних з компонентами базової інфляції задля того, щоб оцінити якість прогнозних моделей й прийти до висновку, описаного на початку цього розділу. Також необхідно було зробити низку трансформацій бази даних, зокрема сезонне коригування, нормалізацію в певних випадках й для певних моделей. Це все дозволило отримати високоякісні результати прогнозних моделей.

Вперше:

- розроблено пайплайн нейромережевого алгоритму з ARIMA моделлю й порівняно її прогностичну здатність в порівнянні з іншими моделями
- алгоритми нейромережевого типу застосовано до дезагрегованих статистичних даних по інфляції в Україні й отримано результати з низьким RMSE
- створено інформаційну технологію на основі методів машинного навчання штучних нейронних мереж, випадкового лісу дерев рішень та інших для точного й швидкого прогнозування рівня базової інфляції України

Удосконалено:

- розроблені в процесі проведення дисертаційного дослідження та випробувані раніше моделі для розв’язання поставленої задачі.
- алгоритм пошуку дистанцій DTW для економічних рядів з щомісячними даними з метою подальшої кластеризації цих рядів

Основні положення та висновки дисертаційного дослідження обговорювалися на наукових семінарах кафедр теорії та технології

програмування та інтелектуальних програмних систем факультету комп’ютерних наук та кібернетики Київського національного університету імені Тараса Шевченка і отримали схвальні відгуки.

Результати дисертаційної роботи знайшли застосування у Національному Банку України й були впроваджені задля поліпшення прогнозу інфляції в цілому та базової інфляції зокрема на щоквартальній основі відповідно до проведення раунду прогнозування макроекономічних змінних в рамках підготовки щокварatального Інфляційного Звіту Національного Банку України. Це підтверджує науково-практичну важливість результатів дисертації.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Pescatori, A., & Zaman, S. (2011). Macroeconomic models, forecasting, and policymaking. Economic Commentary, pp.2011-2019. https://doi.org/10.26509/frbc-ec-201119

2. Shumway, R. H., Stoffer, D. S., Shumway, R. H., & Stoffer, D. S. (2017). ARIMA models. Time series analysis and its applications: with R examples, pp.75-163. https://doi.org/10.1007/978-3-319-52452-8_3

3. Pesaran, M. H., & Smith, R. P. (1998). Structural analysis of cointegrating VARs. Journal of economic surveys, 12(5), pp.471-505. https://doi.org/ 10.1111/1467-6419.00065

4. Berg, A., Karam, P. D., & Laxton, D. (2006). A practical model-based approach to monetary policy analysis: Overview. pp.1-45. https://doi.org/10.5089/9781451863406.001

5. An, S., & Schorfheide, F. (2007). Bayesian analysis of DSGE models. Econometric reviews, 26(2-4), pp.113-172. https://doi.org/10.1080/07474930701220071

6. Mizrach, B., Anderson, H., Becker, R. A., Bjørnland, H. C., Ravazzolo, F., Chang, Y., ... & Scheinkman, J. (2000). StudieS in nonlinear dynamicS and econometricS. MIT Press. pp.1-116. https://doi.org/10.1515/snde-2021-frontmatter5

7. Krukovets D., 2020. “Data Science opportunities in Central Banks: Overview.” Visnyk of the National Bank of Ukraine 249: pp.13-24. https://doi.org/ 10.26531/vnbu2020.249.02

8. Woloszko, N. (2020). Adaptive Trees: a new approach to economic forecasting. ECD Economics Department Working Papers, No. 1593, OECD Publishing, Paris, pp.1-43. https://doi.org/10.1787/18151973

9. Windarto, A. P. (2017). Implementation of data mining on rice imports by major country of origin using algorithm using k-means clustering method. International Journal of artificial intelligence research, 1(2), pp.26-33. https://doi.org/10.29099/ijair.v1i2.17

10. Binner, J. M., Gazely, A., & Elger, T. (2004). Dynamic neural network based inflation forecasts for the uk. Global Business & Economics Review-Anthology, pp.546-564.

11. Angelini, E., Di Tollo, G., & Roli, A. (2008). A neural network approach for credit risk evaluation. The quarterly review of economics and finance, 48(4), pp. 733-755. https://doi.org/10.1016/j.qref.2007.04.001

12. Ugurlu, U., Oksuz, I., & Tas, O. (2018). Electricity price forecasting using recurrent neural networks. Energies, 11(5), 1255. https://doi.org/10.3390/en11051255

13. Eduardo Levy Yeyati & Federico Sturzenegger, 2016. "Classifying Exchange Rate Regimes: 15 Years Later," CID Working Papers 319, Center for International Development at Harvard University.

14. Meyler, Aidan and Kenny, Geoff and Quinn, Terry (1998): Forecasting irish inflation using ARIMA models. Published in: Central Bank and Financial Services Authority of Ireland Technical Paper Series , Vol. 1998, No. 3/RT/98 (December 1998): pp. 1-48.

15. Mondal, P., Shit, L., & Goswami, S. (2014). Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices. International Journal of Computer Science, Engineering and Applications, 4(2), 13. https://doi.org/10.5121/ijcsea.2014.4202

16. Abonazel, M. R., & Abd-Elftah, A. I. (2019). Forecasting Egyptian GDP using ARIMA models. Reports on Economics and Finance, 5(1), 35-47. https://doi.org/10.12988/ref.2019.81023

17. Maria, F. C., & Eva, D. (2011). Exchange-Rates Forecasting: Exponential smoothing techniques and ARIMA models. Annals of Faculty of Economics, 1(1), 499-508.

18. Ulyah, S. M. (2019, July). Forecasting index and stock returns by considering the effect of Indonesia pre-presidential election 2019 using ARIMAX and VARX approches. In Journal of Physics: Conference Series (Vol. 1277, No. 1, p. 012053). IOP Publishing. https://doi.org/10.1088/1742-6596/1277/1/012053

19. Ugoh, C. I., Echebiri, U. V., Temisan, G. O., Iwuchukwu, J. K., & Guobadia, E. K. (2022). On Forecasting Nigeria's GDP: A Comparative Performance of Regression with ARIMA Errors and ARIMA Method. International Journal of Mathematics and Statistics Studies, 10(4), 48-64. https://doi.org/10.37745/ijmss.13/vol10n44864

20. Anggraeni, W., Mahananto, F., Sari, A. Q., Zaini, Z., & Andri, K. B. (2019). Forecasting the price of Indonesia's rice using hybrid artificial neural network and autoregressive integrated moving average (Hybrid NNs-ARIMAX) with exogenous variables. Procedia Computer Science, 161, 677-686. https://doi.org/10.1016/j.procs.2019.11.171

21. Kulyk, Anatolii & Fokina-Mezentseva, Katerina & Piankova, Oksana & Sierova, Liudmyla & Slokva, Maryna. (2023). Forecasting husbandry development using time series. Scientific Horizons. 26. 166-174. 10.48077/scihor11.2023.166. https://doi.org/10.48077/scihor11.2023.166

22. Lütkepohl, H. (2013). Vector autoregressive models. In Handbook of research methods and applications in empirical macroeconomics (pp. 139-164). Edward Elgar Publishing.

23. Dungey, M., & Pagan, A. (2000). A structural VAR model of the Australian economy. Economic record, 76(235), 321-342. https://doi.org/10.1111/j.1475-4932.2000.tb00030.x

24. Jacobson, T., Jansson, P., Vredin, A., & Warne, A. (1999). A VAR model for monetary policy analysis in a small open economy (No. 77). Sveriges Riksbank Working Paper Series.

25. Sims, Christopher A, 1980. "Comparison of Interwar and Postwar Business Cycles: Monetarism Reconsidered," American Economic Review, American Economic Association, vol. 70(2), pages 250-257, May.

26. Awokuse, T. O. (2006). Export-led growth and the Japanese economy: evidence from VAR and directed acyclic graphs. Applied Economics, 38(5), 593-602. https://doi.org/10.1080/13504850500358801

27. Shojaie, A., & Fox, E. B. (2022). Granger causality: A review and recent advances. Annual Review of Statistics and Its Application, 9, 289-319. https://doi.org/10.1146/annurev-statistics-040120-010930

28. Sznajderska, A. (2019). The role of China in the world economy: evidence from a global VAR model. Applied Economics, 51(15), 1574-1587. https://doi.org/10.1080/00036846.2018.1527464

29. Ahmed, H. J. A., & Wadud, I. M. (2011). Role of oil price shocks on macroeconomic activities: An SVAR approach to the Malaysian economy and monetary responses. Energy policy, 39(12), 8062-8069. https://doi.org/10.1016/j.enpol.2011.09.067

30. Sek, S. K., & Lim, H. S. (2016, June). An investigation on the impacts of oil price shocks on domestic inflation: A SVAR approach. In AIP Conference Proceedings (Vol. 1750, No. 1). AIP Publishing. https://doi.org/10.1063/1.4954607

31. Dai, P. F., Xiong, X., & Zhou, W. X. (2021). A global economic policy uncertainty index from principal component analysis. Finance Research Letters, 40, 101686. https://doi.org/10.1016/j.frl.2020.101686

32. Laine, O. M. J. (2020). The effect of the ECB's conventional monetary policy on the real economy: FAVAR-approach. Empirical Economics, 59(6), 2899-2924. https://doi.org/10.1007/s00181-019-01739-9

33. Bernanke, B. S., Boivin, J., & Eliasz, P. (2005). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. The Quarterly journal of economics, 120(1), 387-422. https://doi.org/10.1162/qjec.2005.120.1.387

34. Grui, A., & Lysenko, R. (2017). Nowcasting Ukraine's GDP Using a Factor-Augmented VAR (FAVAR) Model. Visnyk of the National Bank of Ukraine, (242), 5-13. https://doi.org/10.26531/vnbu2017.242.005

35. Welch, Greg & Bishop, Gary. (2006). An Introduction to the Kalman Filter. Proc. Siggraph Course. 8.

36. Beneš, Jaromír and Clinton, Kevin and george, asish thomas and Gupta, Pranav and John, Joice and Kamenik, Ondra and Laxton, Douglas and Mitra, Pratik

and Nadhanael, G.V. and Portillo, Rafael and Wang, Hou and Zhang, Fan, Quarterly Projection Model for India: Key Elements and Properties (February 2017). IMF Working Paper No. 17/33, Available at SSRN: https://ssrn.com/abstract=2938332

37. Grui, A., & Vdovychenko, A. (2019). Quarterly projection model for Ukraine (No. 03/2019).

38. Shesadri Banerjee and Parantap Basu, 2015. "A Dynamic Stochastic General Equilibrium Model for India," Macroeconomics Working Papers 24975, East Asian Bureau of Economic Research.

39. Antonova, A. (2018). macroeconomic Effects of minimum Wage Increases in an Economy with Wage Underreporting. Visnyk of the National Bank of Ukraine, (246), 10-33. https://doi.org/10.26531/vnbu2018.246.010

40. Basak, S., Kar, S., Saha, S., Khaidem, L., & Dey, S. R. (2019). Predicting the direction of stock market prices using tree-based classifiers. The North American Journal of Economics and Finance, 47, 552-567. https://doi.org/10.1016/j.najef.2018.06.013

41. T. Batsuuri, S. He, R. Hu, J. Leslie, and F. Lutz. 2024. "Predicting IMF-Supported Programs: A Machine Learning Approach." IMF Working Paper WP/24/54, International Monetary Fund, Washington D.C.

42. Biau, O., & D'Elia, A. (2010). Euro area GDP forecasting using large survey datasets. A random forest approach. Unpublished Paper, European Commission.

43. Mei, J., He, D., Harley, R., Habetler, T., & Qu, G. (2014, July). A random forest method for real-time price forecasting in New York electricity market. In 2014 IEEE PES general meeting| conference & exposition (pp. 1-5). IEEE. https://doi.org/10.1109/PESGM.2014.6939932

44. Tyralis, Hristos, and Georgia Papacharalampous. 2017. "Variable Selection in Time Series Forecasting Using Random Forests" Algorithms 10, no. 4: 114. https://doi.org/10.3390/a10040114

45. Gawthorpe, K. (2021). Random Forest as a Model for Czech Forecasting. Prague Economic Papers, 30(3), 336-357. https://doi.org/10.18267/j.pep.765

46. Mei-Li Shen & Cheng-Feng Lee & Hsiou-Hsiang Liu & Po-Yin Chang & Cheng-Hong Yang, 2021. "An Effective Hybrid Approach for Forecasting Currency Exchange Rates," Sustainability, MDPI, vol. 13(5), pages 1-29, March.

47. S. Ramakrishnan, S. Butt, M. A. Chohan and H. Ahmad, (2017). Forecasting Malaysian exchange rate using machine learning techniques based on commodities prices. International Conference on Research and Innovation in Information Systems (ICRIIS), Langkawi, Malaysia, 2017, pp. 1-5. https://doi.org/10.1109/ICRIIS.2017.8002544

48. Hu, T., & Song, T. (2019, October). Research on XGboost academic forecasting and analysis modelling. In Journal of Physics: Conference Series (Vol. 1324, No. 1, p. 012091). IOP Publishing. https://doi.org/10.1088/1742-6596/1324/1/012091

49. Massaro, A., Panarese, A., Giannone, D., & Galiano, A. (2021). Augmented data and xgboost improvement for sales forecasting in the large-scale retail sector. Applied Sciences, 11(17), 7793. https://doi.org/10.3390/app11177793

50. Noorunnahar, M., Chowdhury, A. H., & Mila, F. A. (2023). A tree based eXtreme Gradient Boosting (XGBoost) machine learning model to forecast the annual rice production in Bangladesh. PloS one, 18(3), e0283452. https://doi.org/10.1371/journal.pone.0283452

51. Huang, W., Nakamori, Y., & Wang, S. Y. (2005). Forecasting stock market movement direction with support vector machine. Computers & operations research, 32(10), 2513-2522. https://doi.org/10.1016/j.cor.2004.03.016

52. Cao, L. J., & Tay, F. E. H. (2003). Support vector machine with adaptive parameters in financial time series forecasting. IEEE Transactions on neural networks, 14(6), 1506-1518. https://doi.org/10.1109/TNN.2003.820556

53. Qin, Y., Xu, Z., Wang, X., & Skare, M. (2023). Artificial intelligence and economic development: An evolutionary investigation and systematic review. Journal of the Knowledge Economy, 1-35. https://doi.org/10.1007/s13132-023-01183-2

54. Nakamura, E. (2005). Inflation forecasting using a neural network. Economics Letters, 86(3), 373-378. https://doi.org/10.1016/j.econlet.2004.09.003

55. Choudhary, A., Haider, A. (2012). Neural network models for inflation forecasting: an appraisal. Applied Economics, 44(20), 2631-2635. https://doi.org/10.1080/00036846.2011.566190

56. Medeiros, M., Vasconcelos, G., Veiga, A., Zilberman, E. (2019). Forecasting inflation in a data-rich environment: the benefits of machine learning methods. Journal of Business & Economic Statistics. https://doi.org/10.1080/07350015.2019.1637745

57. Jung, J., Patnam, M., Ter-Martirosyan, A. (2018). An algorithmic crystal ball: forecasts-based on machine learning. IMF Working Papers, WP/20/7. International Monetary Fund

58. Tkacz, G. (2001). Neural network forecasting of Canadian GDP growth. International Journal of Forecasting, 17(1), 57-69. https://doi.org/10.1016/S0169-2070(00)00063-7

59. Longo, L., Riccaboni, M., & Rungi, A. (2022). A neural network ensemble approach for GDP forecasting. Journal of Economic Dynamics and Control, 134, 104278. https://doi.org/10.1016/j.jedc.2021.104278

60. Zhang, G. P., & Berardi, V. L. (2001). Time series forecasting with neural network ensembles: an application for exchange rate prediction. Journal of the operational research society, 52, 652-664. https://doi.org/10.1057/palgrave.jors.2601133

61. Keogh, E., & Lin, J. (2005). Clustering of time-series subsequences is meaningless: implications for previous and future research. Knowledge and information systems, 8, 154-177.

62. Wolfson, M., Madjd-Sadjadi, Z., & James, P. (2004). Identifying national types: A cluster analysis of politics, economics, and conflict. Journal of Peace Research, 41(5), 607-623.

63. Rashkovan, V., & Pokidin, D. (2016). Ukrainian banks' business models clustering: application of Kohonen neural networks. Visnyk of the National Bank of Ukraine, (238), 13-38.

64. Franses, P. H., & Wiemann, T. (2020). Intertemporal similarity of economic time series: An application of dynamic time warping. Computational Economics, 56, 59-75.

65. Rutkowska, A., & Szyszko, M. (2022). New DTW windows type for forward-and backward-lookingness examination. Application for inflation expectation. Computational Economics, 59(2), 701-718.

66. Śmiech, S. (2015). Co-movement of commodity prices–results from dynamic time warping classification. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie, 940(04), 117-130.

67. Wan, X., Wang, W., Liu, J., & Tong, T. (2014). Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. BMC medical research methodology, 14, 1-13.

68. Ali, P. J. M., Faraj, R. H., Koya, E., Ali, P. J. M., & Faraj, R. H. (2014). Data normalization and standardization: a technical report. Mach Learn Tech Rep, 1(1), 1-6.

69. Sax, C., & Eddelbuettel, D. (2018). Seasonal adjustment by x-13arima-seats in r. Journal of Statistical Software, 87, 1-17.

70. Kianimajd, A., Ruano, M. G., Carvalho, P., Henriques, J., Rocha, T., Paredes, S., & Ruano, A. E. (2017). Comparison of different methods of measuring similarity in physiologic time series. IFAC-PapersOnLine, 50(1), 11005-11010.

71. Zhou, Z. (2012). Measuring nonlinear dependence in time-series, a distance correlation approach. Journal of Time Series Analysis, 33(3), 438-457.

72. Berndt, D. J., & Clifford, J. (1994, July). Using dynamic time warping to find patterns in time series. In Proceedings of the 3rd international conference on knowledge discovery and data mining (pp. 359-370).

73. Salvador, S., & Chan, P. (2007). Toward accurate dynamic time warping in linear time and space. Intelligent Data Analysis, 11(5), 561-580.

74. Gold, O., & Sharir, M. (2018). Dynamic time warping and geometric edit distance: Breaking the quadratic barrier. ACM Transactions on Algorithms (TALG), 14(4), 1-17.

75. Daniel Meliza, C., Keen, S. C., & Rubenstein, D. R. (2013). Pitch-and spectral-based dynamic time warping methods for comparing field recordings of harmonic avian vocalizations. The Journal of the Acoustical Society of America, 134(2), 1407-1415.

76. Lemire, D. (2009). Faster retrieval with a two-pass dynamic-time-warping lower bound. Pattern recognition, 42(9), 2169-2180.

77. Al-Naymat, G., Chawla, S., & Taheri, J. (2012). Sparsedtw: A novel approach to speed up dynamic time warping. arXiv preprint arXiv:1201.2969.

78. Dokmanic, I., Parhizkar, R., Ranieri, J., & Vetterli, M. (2015). Euclidean distance matrices: essential theory, algorithms, and applications. IEEE Signal Processing Magazine, 32(6), 12-30.

79. Cui, S. Y., He, J. X., & Tian, G. X. (2019). The generalized distance matrix. Linear algebra and its applications, 563, 1-23.

80. Narita, H., Sawamura, Y., & Hayashi, A. (2007, November). Learning a kernel matrix for time series data from dtw distances. In International Conference on Neural Information Processing (pp. 336-345). Berlin, Heidelberg: Springer Berlin Heidelberg.

81. Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. Annals of data science, 2, 165-193.

82. Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on neural networks, 16(3), 645-678.

83. Uppada, S. K. (2014). Centroid based clustering algorithms—A clarion study. International Journal of Computer Science and Information Technologies, 5(6), 7309-7313.

84. Kriegel, H. P., Kröger, P., Sander, J., & Zimek, A. (2011). Density-based clustering. Wiley interdisciplinary reviews: data mining and knowledge discovery, 1(3), 231-240.

85. Grygorash, O., Zhou, Y., & Jorgensen, Z. (2006, November). Minimum spanning tree based clustering algorithms. In 2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06) (pp. 73-81). IEEE.

86. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666.

87. Hamerly, G., & Elkan, C. (2003). Learning the k in k-means. Advances in neural information processing systems, 16.

88. Bahmani, B., Moseley, B., Vattani, A., Kumar, R., & Vassilvitskii, S. (2012). Scalable k-means++. arXiv preprint arXiv:1203.6402.

89. Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Transactions on Database Systems (TODS), 42(3), 1-21.

90. Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (1999). Optics-of: Identifying local outliers. In Principles of Data Mining and Knowledge Discovery: Third European Conference, PKDD'99, Prague, Czech Republic, September 15-18, 1999. Proceedings 3 (pp. 262-270). Springer Berlin Heidelberg.

91. Nielsen, F., & Nielsen, F. (2016). Hierarchical clustering. Introduction to HPC with MPI for Data Science, 195-211.

92. Lawler, G. F., & Limic, V. (2010). Random walk: a modern introduction (Vol. 123). Cambridge University Press.

93. Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. Journal of statistical software, 27, 1-22.

94. Huwiler, M., & Kaufmann, D. (2013). Combining disaggregate forecasts for inflation: The SNB's ARIMA model: the SNB's ARIMA model. Swiss National Bank Economic Study, 7.

95. Krukovets, D., & Verchenko, O. (2019). Short-Run Forecasting of Core Inflation in Ukraine: a Combined ARMA Approach. Visnyk of the National Bank of Ukraine, (248), 11-20.

96. Rigatti, S. J. (2017). Random forest. Journal of Insurance Medicine, 47(1), 31-39.

97. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

98. Grossberg, S. (2013). Recurrent neural networks. Scholarpedia, 8(2), 1888.

99. Graves, A., & Graves, A. (2012). Long short-term memory. Supervised sequence labelling with recurrent neural networks, 37-45.

100. Almosova, A., & Andresen, N. (2023). Nonlinear inflation forecasting with recurrent neural networks. Journal of Forecasting, 42(2), 240-259.

101. Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems, 2(4), 303-314.

102. Hodson, T. O. (2022). Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. Geoscientific Model Development Discussions, 2022, 1-10.

Додаток 1

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Наукові праці, в яких опубліковані основні наукові результати дисертації:

Публікації у фахових виданнях України:

1. Krukovets, D. (2022). Multi-stage approach with DTW and clustering for forecasting of average deposit rate in Ukraine. Bulletin of Taras Shevchenko. National University of Kyiv. Series Physics & Mathematics, pp.55-65. https://www.doi.org/10.17721/1812-5409.2022/4.7
2. Krukovets, D. (2023). Updated DTW+K-Means approach with LSTM and ARIMA-type models for Core Inflation forecasting. Bulletin of Taras Shevchenko. National University of Kyiv. Series Physics & Mathematics, pp.214-225. https://www.doi.org/10.17721/1812-5409.2023/2.38
3. Krukovets, D. (2024). Exploring an LSTM-SARIMA routine for core inflation forecasting. Technology audit and production reserves, pp. 6-12. https://www.doi.org/10.15587/2706-5448.2024.301209

Наукові праці, які засвідчують апробацію матеріалів дисертації:

4. Krukovets, D. (2019): Non-stationary time-series distance clustering for a similarity analysis. Перша українська конференція «Логіка та її застосування» (UCLA'2019), pp. 117-119. м.Київ, грудень 2019
5. Krukovets, D. (2020). Analysis of similarity between artificially simulated time series with Dynamic Time Warping. In Workshop on Intelligent Information Systems (p. 97). Chisinau, Republic of Moldova, December 2020
6. Krukovets, D. (2021): Dynamic Time Warping for uncovering dissimilarity of regional wages in Ukraine. "Proceedings MFOI-2020", pp. 168-185. м.Київ, січень 2021