УДК 004.89 **doi:** https://doi.org/10.20998/2522-9052.2023.2.02

Г. М. Хорошун, О. І. Рязанцев, М. В. Черпіцький

Східноукраїнський національний університет імені Володимира Даля, Київ, Україна

КЛАСТЕРИЗАЦІЯ ТА АНОМАЛЬНІСТЬ ДАНИХ ІНДЕКСУ ВОЛАТИЛЬНОСТІ ФОНДОВОГО РИНКУ США

Анотація. Актуальність. Інвестування грошей це важливий спосіб покращення фінансового стану, як окремої людини, так і суспільства в цілому. Актуальною є проблема розуміння фінансових даних та прийняття рішень щодо інвестицій грошей в певний проект на даний момент часу. **Об'єктом** дослідження є процес встановлення залежність вартості активу від часу. **Предметом** дослідження є математичні моделі обробки даних, проведення кластеризації даних та пошуку аномалій. **Метою** даної роботи є розробка методу для ефективного інвестуванню грошей з використанням методів обробки та аналізу даних до значень індексу волатильності фондового ринку США, визначення кластерів по діям з активами, а також перевірка наявності аномальних даних. **Результати досліджень.** Обрані офіційні дані значень індексу волатильності та підготовлені для подальшого аналізу шляхом видалення неповних наборів та подальшої нормалізації. Проведена кластеризація часових рядів та розділено масив на п'ять однорідних груп. Кластери визначають діапазони індексу волатильності, що відображають різні настрої інвесторів на ринку та спонукають до відповідних дій з активами: продавати, очікувати, купляти, виводити гроші з проектів, що розвиваються і вкладити в стабільні, перечекати. Програмно проведено сегментацію даних, застосування віконної функції, визначено центроїди для сегментів та проведено реконструкцію сигналу. Визначено точки аномалій даних. Проведено порівняльний аналіз за результатами побудованих початкових даних, реконструйованих та похибки реконструювання.

Ключові слова: кластеризація; аномальність;, індекс волатильності.

Постановка та аналіз проблеми

В сучасному світі надлишок грошей в бюджеті підприємства або сім'ї прийнято інвестувати в бізнес, нерухомість та себе. Інвестиціями є всі види майнових та інтелектуальних цінностей, що вкладаються в об'єкти підприємницької та інших видів діяльності, в результаті якої створюється прибуток (доход) та/або досягається соціальний та екологічний ефект [1].

Існує багато показників в управлінні фінансовими ризиками, які використовуються для аналізу ринку. Одним з таких статистичних фінансових показників є індекс волатильності (VIX), що вказує на очікувані коливання вартості активу протягом наступних 30 днів [2]. Такий часовий період передбачення звужує світогляд до близького терміну. Індекс волатильності є показником настроїв інвесторів та їх прогнозів щодо волатильності основних котирувань активів з часом.

Отже, **метою** даної роботи є розробка методу для ефективного інвестуванню грошей з використанням методів обробки та аналізу даних для значень індексу волатильності.

Для цього пропонується розраховувати індекс волатильності, встановлюючи залежність вартості активу від часу. Для визначення трендів в однорідних групах даних корисно провести кластеризацію даних, а також визначити точки або області аномалій, що можуть бути помилковими даними або презентувати зародження нового кластеру.

Для роботи в галузі дослідження даних [3-7] нам необхідно обрати достовірні джерела інформації та обробити дані таким чином, щоб мати достатню кількість однотипних даних за обраний час з певною періодичністю та зберегти у зручному форматі для обробки. В цій роботі планується обробка економічних даних, тому ми звернулися до FRED — онлайн бази даних, що складається з сотень тисяч економічних часових рядів даних з оцінок національних, міжнародних, державних та приватних джерел [8]. FRED створений і підтримується дослідним відділом при Федеральному резервному банку Св. Луіса. Корисною є можливість використання бази ALFRED (ArchivaL Federal Reserve Economic Data), яка пропонує користувачам можливість доступу до старих даних для багатьох доступних серій FRED. По суті, ALFRED та FRED допомагають користувачам розповідати свої історії даних.

Теоретичні відомості

Індекс волатильності. Волатильність це статистичний показник змінювання цін, який є мірою ризику використання фінансового інструменту в даний період часу. Для розрахунку волатильності використовується статистична стандартна статистика відхилення, що дозволяє інвесторам визначити ризик придбання фінансового інструменту.

Для розрахунку індексу волатильності нам необхідно попередньо розрахувати стандартне відхилення σ_{sd} для вибірки значень ціни:

$$\sigma_{sd} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \underline{x})^2}{n-1}}. \quad (1)$$

В формулі (1) n – розмір вибірки; x_i – розмір індивідуального значення вибірки; $\underline{x}$ – середнє арифметичне вибірки.

Найбільш часто розраховується середньорічна волатильність σ за формулою:

$$\sigma = \frac{\sigma_{sd}}{\sqrt{P}}, \quad (2)$$

де P – часовий проміжок у роках.

Волатильність σ_T на інтервал часу T (обчислюється в днях) розраховується на підставі річної середньої волатильності таким чином:

$$\sigma_T = \sigma\sqrt{T}. \quad (3)$$

Підготовка даних. Для аналізу даних необхідно перевірити їх на відповідність до загальних вимог, забезпечити їх повноту та логічність. Трансформація даних, наприклад нормалізація, може покращити точність та ефективність аналізу даних за допомогою алгоритмів машинного навчання. Такі методи дають кращі результати, якщо дані будуть нормалізовані, тобто масштабовані до певних діапазонів. Атрибут нормалізується шляхом масштабування його значень таким чином, щоб вони зменшувалися у невеликому заданому діапазоні, наприклад, від -1 до 1. Оберемо метод мін-макс нормалізації, яка виконує лінійне перетворення вхідних даних x у вихідні дані x_{new} та може бути обчислена за формулою:

$$V_{new} = \frac{(V - V_{min})}{(V_{max} - V_{min})}. \quad (4)$$

У виразі (4) V_{min} та V_{max} є відповідне мінімальне та максимальне значення змінної x.

Кластеризація даних часових рядів. За допомогою кластерного аналізу [9] можна зробити вибірку за ознакою, що вивчатиметься. Його основна задача полягає в розділенні багатовимірного масиву на однорідні групи. Парний коефіцієнт кореляції або Евклідова відстань між об'єктами за заданим параметром використовується як критерій групування. Найближчі значення групуються разом в єдиний кластер.

Кластерний аналіз включає наступні кроки. Вибір даних або їх перетворення для створення корисних і нових функцій. Розробка або вибір алгоритму кластеризації. Валідація кластера необхідна завдяки тому, що різні підходи зазвичай призводять до формування різних груп кластерів і навіть для одного алгоритму, параметри ідентифікації або послідовність введення шаблонів може вплинути на фінальні результати. Інтерпретація результатів дозволяє надати значущу інформацію з вихідних даних завдяки проведеній кластеризації.

Метод, який ми будемо використовувати, називається кластеризацією середніх або k-means. Алгоритм k- means кластеризує дані, намагаючись розділити вибірку на n груп однакової дисперсії, мінімізуючи критерій, відомий як інерція або сума квадратів всередині кластера. Цей алгоритм вимагає завдання кількості кластерів, які потрібно вказати. Він добре масштабується до великої кількості зразків і використовувався в багатьох областях.

Алгоритм k-means ділить набір з N зразків X на K непересічних кластерів C, кожен з яких описується середнім μ_i зразків у кластері, $\mu_i \in C$. Середнє зазвичай називають «центроїдами» кластера, і вони зазвичай не є точками з X, хоча вони знаходяться в одному просторі. Алгоритм k-means спрямований на вибір центроїдів, які мінімізуються критерій інерції або суми квадратів кластера:

$$\sum_{i=0}^{n} min\left(\|x_j - \mu_i\|^2\right).$$

Інерція може сприйматися, як міра того, наскільки внутрішньо когерентними є кластери.

Методи пошуку аномалій. Загальне завдання виявлення аномалій у часових рядах часто поділяється на дві окремі задачі: виявлення викидів (Outlier Detection) та «нової поведінки» (Novelty Detection). Викиди є наслідком: помилок у даних (неточності вимірювання, округлення, неправильного запису тощо), наявності шумових об'єктів (неправильно класифікованих об'єктів), присутності об'єктів «інших» вибірок.

Для пошуку аномалій застосовуються статистичні методи для вивчення окремих ознак та відокремлення екстремальних та унікальних значень у виборці, а саме: Z-оцінка та надлишковий ексцес. Також в роботі з визначення аномалій використовують модельний підхід, метричні, ітераційні методи та методи машинного навчання.

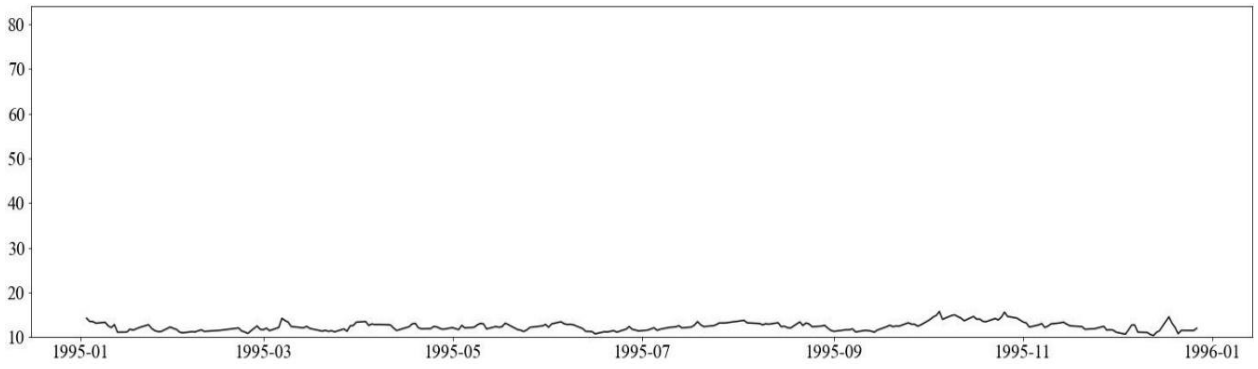
Виявлення аномалії включає: виявлення аномалій шляхом пошуку будь-яких значень за певним порогом; виявлення аномалій за структурою сигналу; більш тонкі похибки - зміна форми періодичної хвилі.

Результати роботи з кластеризації даних та їх аномальності. Використовуються дані у вигляді часових рядів для індексу волатильності, отриманих з бази FRED з 02.01.1990 по 29.09.2022 з періодичністю в 1 день. Вивчаються дані волатильності фондового індексу США S&P 500 [10], що містить інформацію про 500 найпотужніших компаній. Графічне представлення даних індексу волатильності VIX в періоди стабільності та панічних настроїв інвесторів наведений на рис. 1.

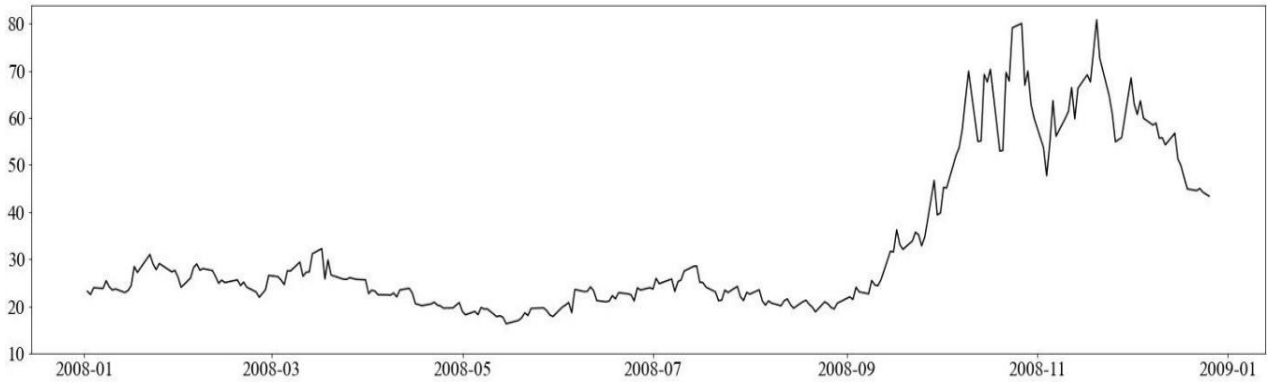
Подальший аналіз даних проведемо за допомогою статистичних методів. Гістограма індексу волатильності початкових даних (рис. 2а) вказує на значну наявність нульових або пропущених даних. Також, ми бачимо, що після індексу волатильності 45 кількість таких подій значно зменшується. Отже, зменшемо кількість випадків подій до 20 та видалимо пропущені дані (рис. 2б). Згідно до графіку, мінімальний індекс волатильності становить 9 % і буде прийнятий нами за нижню межу при проведенні кластеризації, а 83% - за вищу межу.

Кластеризація. Згідно з проведеного статистичного аналізу та існуючого тлумачення значень індексу волатильності VIX можна створити для інтерпретації результатів обчислень наступних 5 кластерів. Введемо такі рекомендації до кластерів, що визначають діапазони індексу волатильності та відображають різні настрої інвесторів на ринку та спонукають до відповідних дій з активами: перший – продавати, другий – очікувати, третій – купляти, четвертий – виводити гроші з проектів, що розвиваються і вкладати в стабільні, п’ятий – перечекати на зменшення індексу.

1. 9-20% - низька волатильність. Це свідчить про оптимістичний настрій учасників ринку. Чим менше значення показника падає, тим більша ймовірність швидкого реверсивного руху тренду і його реверсивного. Це часто є сигналом для інвестора продавати цінні папери і близькі позиції. У цьому випадку важливо зафіксувати прибуток до початку неминучого перелому.

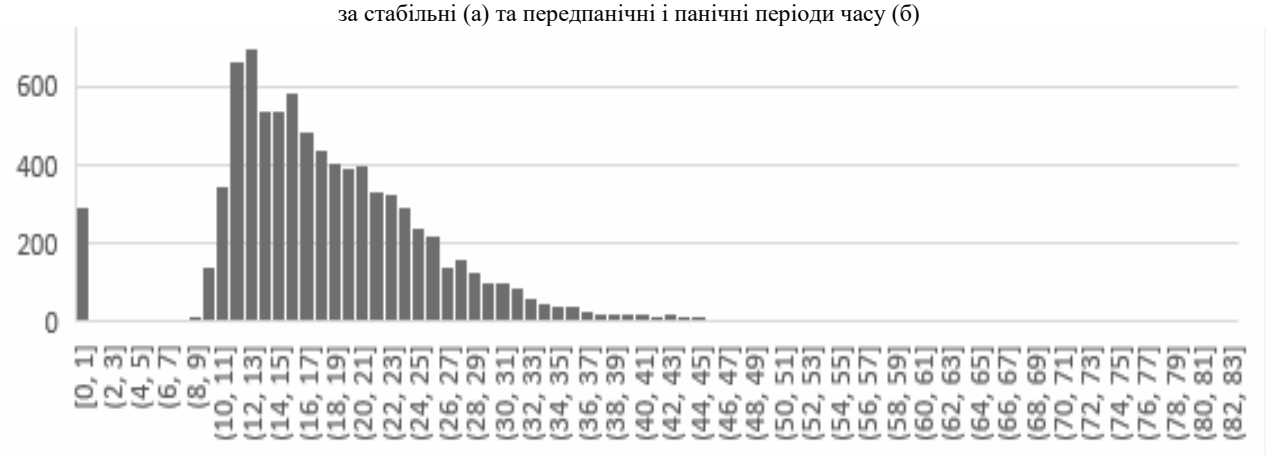


а

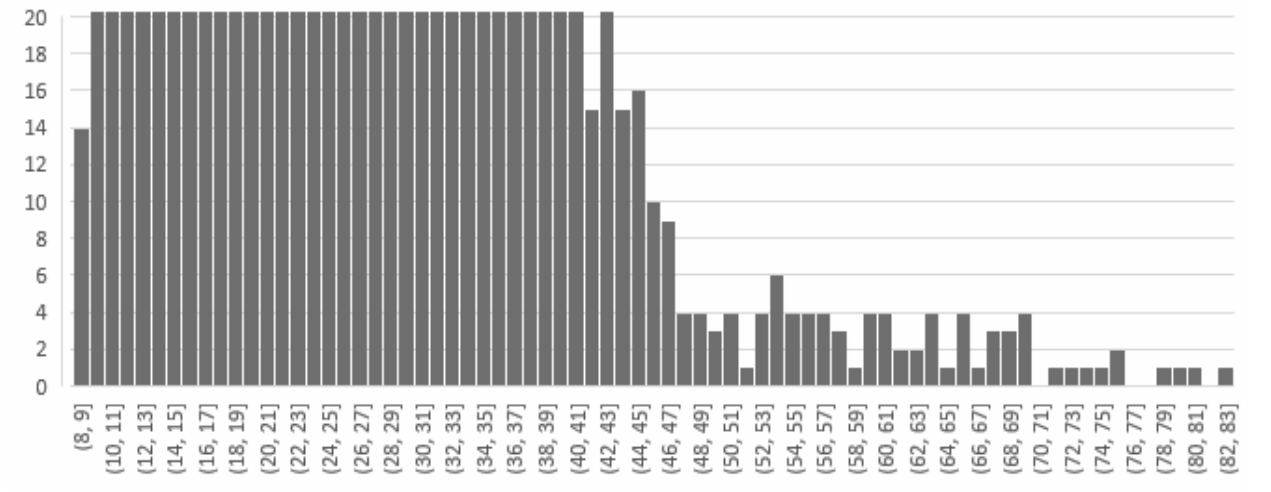


б

Рис. 1. Графічне представлення даних індексу волатильності VIX
за стабільні (а) та передпанічні і панічні періоди часу (б)



а



б

Рис. 2. Гістограми індексу волатильності початкових даних (а) та без пропущених даних з максимальним значенням (б)

2. 21-35% - середня волатильність. Коливання в межах цього діапазону не можуть дати інвестору будь-яких ознак дії.

3. 36-45% - ознака паніки на ринку. Такий стан справ зазвичай супроводжується швидким крахом цін на акції. Це сигнал для інвестора для пошуку точки входу на ринок. Після того, як температура знижується, ціна акцій знову підніметься. Тому це найкращий час для купівлі цінних паперів.

4. 46-59% - ознака великої паніки на ринку.

5. 60-83% – ознака кризи на ринку. Як правило, індекс стрімко збільшується вгору.

Аномальність даних. Для подальшого аналізу та пошуку аномалії в даних оберемо дані за роки відносної стабільності ринку. Кластерний аналіз використовується для визначення аномалії як певної моделі хвилі, яка раніше не спостерігалася. Створена бібліотека нормальних форм сигналу використовується для спроби реконструювати форму сигналу для тестування. Якщо реконструкція погана, то форма хвилі, ймовірно, містить щось ненормальне і, отже, є аномалією. Виявлення аномалій проводиться за наступним алгоритмом.

1. Розділити форму хвилі на сегменти по n зразків.
2. Застосувати віконну функції до даних, яка приводить початок і кінець сигналу до нуля. Сформувати простір в n вимірі, де кожен сегмент представляє одну точку.
3. Провести кластеризацію точок сегментів і визначити центроїди кластерів. Центроїди кластерів забезпечують бібліотеку нормальних форм хвилі.
4. Реконструювати форму хвилі для тестування з використанням кластерних центроїдів, отриманих під час навчання.
5. Визначити області аномалій даних.

Проведемо сегментацію даних за 2012-2014 роки на 235 частин, з яких 7 сегментів наведені у якості прикладу на рис. 3.

Сегментовані фрагменти даних необхідно обробити за допомогою віконної функції. В якості віконної функції обрано частину синусоїди, яка наведена на рис. 4. Подальша обробка даних призводить до трансформації даних у вигляді на рис. 5.

Програмно визначено центроїди для сегментів та проведено їх реконструкцію. Результат реконструйованого сигналу наведений суцільним сірим кольором на рис. 6. Також на ньому легко провести порівняльний аналіз з початковими даними та похибкою реконструкції.

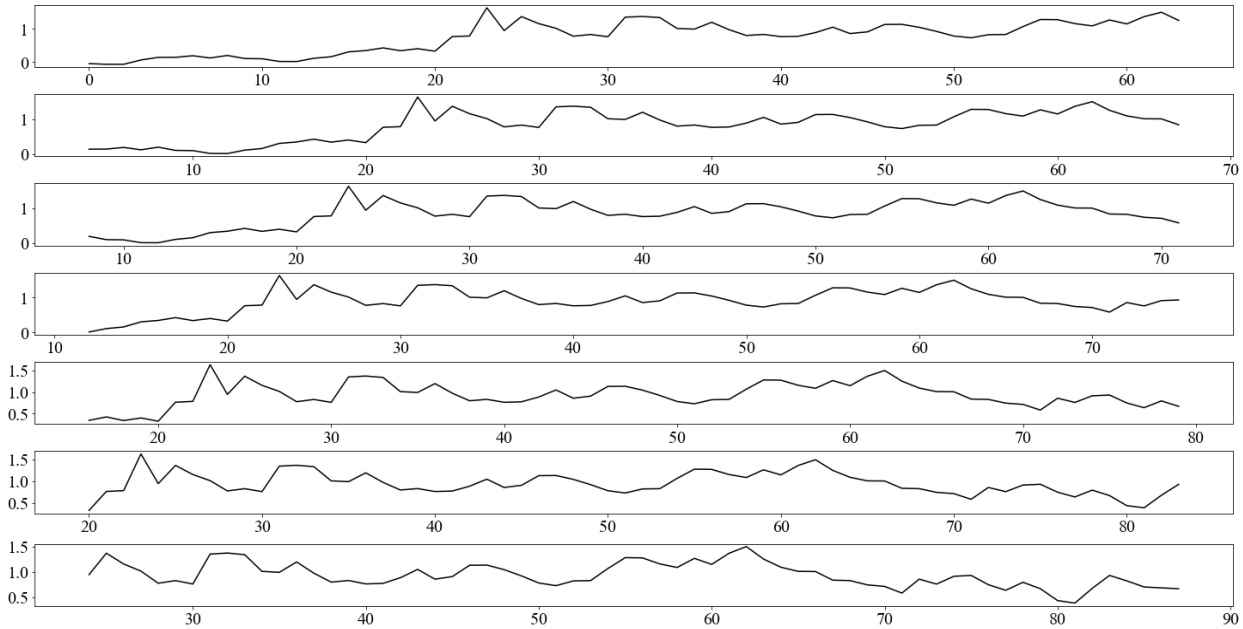


Рис. 3. Сегментовані фрагменти даних

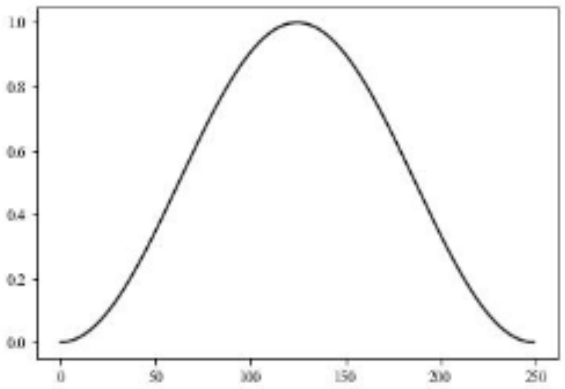


Рис. 4. Віконна функція для обробки даних

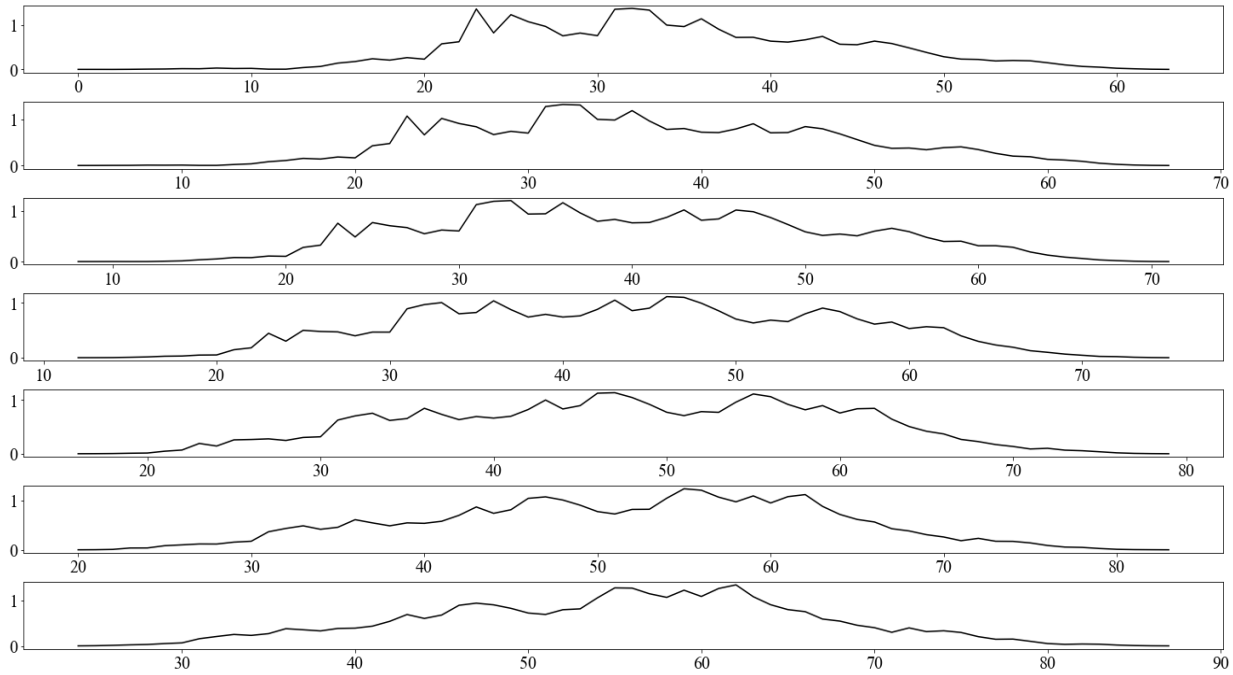


Рис. 5. Сегментовані дані після проходження віконної функції

Візуалізації початкових та оброблених даних наведена у графічному форматі з використанням різних типів та кольорів ліній. Як ми бачимо з графіків сигнали співпадають доволі добре на всьому протязі осі X.

Можна визначити області аномалій для даних, які в нашому випадку є точками з аномальними значеннями. Ці точки змінюють форму сигналу настільки що різниця між початковими даними та реконструйованими більше ніж 0.1 по модулю. Різниця є похибкою реконструкції і позначена на рис. 6 та 7 суцільним сірим кольором.

Форма основного сигналу може значно збільшуватися або зменшуватися у порівнянні з реконструйованим сигналом.

Аномальні точки позначені колами на рис.6 можуть вказувати на непередбачуваність фінансових процесів, а також на появу нових трендів.

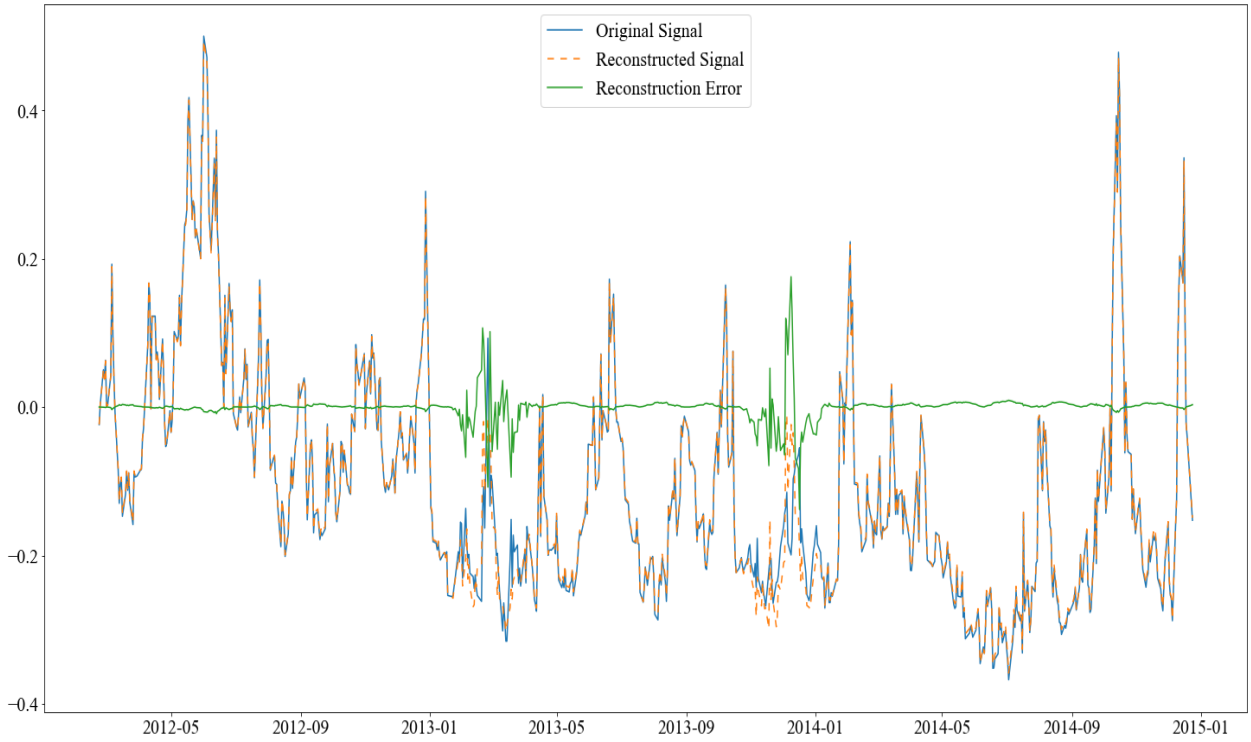
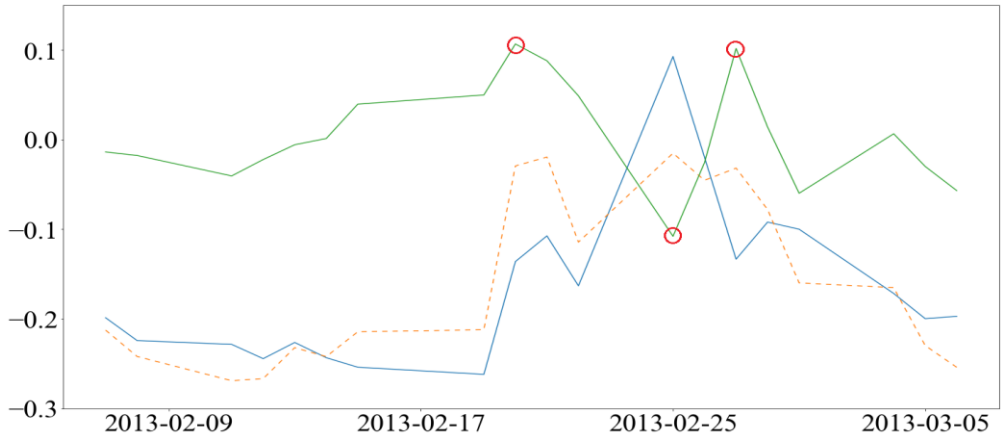
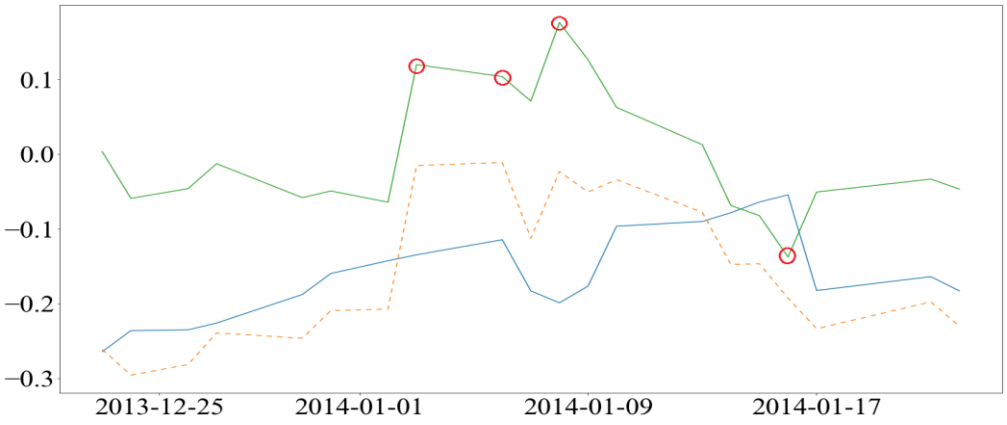


Рис. 6. Візуалізація фрагменту початкових даних (original signal) наведено суцільним чорним кольором, реконструйовані дані (reconstructed signal) пунктирною лінією, похибка реконструкції (reconstruction error) показана суцільним сірим кольором



а



б

Рис. 7. Визначення аномальних даних, позначених колом,
за перевищенням порогового значення похибки реконструкції

Висновки

Роботу присвячено аналізу даних по темі інвестування грошей з використанням індексу волатильності, який встановлює залежність вартості активу від часу. За допомогою статистичних методів проведено первісну обробку даних, проведено кластеризацію даних на п'ять кластерів, які відображають різні настрої інвесторів на ринку та спонукають до відповідних дій з активами. Також визначено точки аномалій даних завдяки кластерному аналізу та алгоритму реконструкції сигналу. Програмно визначено центроїди для сегментів та проведено їх реконструкцію. Проведено порівняльний аналіз за результатами побудованих початкових даних, реконструйованих та похибки реконструювання.

References

1. (2022), *On investment activity*, Law of Ukraine dated September 18, 1991 No. 1560-XII: as of October 10, 2022, available at: https://zakon.rada.gov.ua/laws/show/1560-12#Text
2. (2023), *SPX | S&P 500 Index Overview | MarketWatch*, URL: https://www.marketwatch.com/investing/index/spx
3. Wolf, Andrew (2022), *Machine Learning Simplified:A gentle introduction to supervised learning*, 199 p., available at: https://themlsbook.com
4. Maheshwari, A. (2014), "Business intelligence and data mining", *Business Expert Press*, available at: https://www.amazon.com/Business-Intelligence-Data-Mining-Analytics/dp/1631571206
5. Yang, X. S. (2019), *Introduction to Algorithms for Data Mining and Machine Learning*, Academic Press, doi: https://doi.org/10.1016/C2018-0-02034-4
6. Fernandes, M. (2008), *Statistics for business and economics*, Bookboon, available at: https://bookboon.com/en/statistics-for-business-and-economics-ebook?mediaType=ebook
7. Vercellis, C. (2009), *Business intelligence: data mining and optimization for decision making*, Wiley, New York, 420 p., available at: https://www.amazon.com/Business-Intelligence-Mining-Optimization-Decision/dp/0470511397
8. (2023), *What is FRED? | Getting To Know FRED*, available at: https://fredhelp.stlouisfed.org/fred/about/about-fred/what-is-fred/
9. Smith, L. I. (2002), *A tutorial on Principal Components Analysis*, Computer Science Technical Report No. OUCS-2002-12, available at: http://hdl.handle.net/10523/7534
10. (2023), Chicago Board Options Exchange, CBOE Volatility Index: VIX [VIXCLS], retrieved from FRED, Federal Reserve Bank of St. Louis; *Economic Research*, February 3, 2023, available at: https://fred.stlouisfed.org/series/VIXCLS

СПИСОК ЛІТЕРАТУРИ

1. Про інвестиційну діяльність: Закон України від 18.09.1991 № 1560-XII : станом на 10 жовт. 2022 р. URL: https://zakon.rada.gov.ua/laws/show/1560-12#Text
2. SPX | S&P 500 Index Overview | MarketWatch. URL: https://www.marketwatch.com/investing/index/spx
3. Wolf Andrew. Machine Learning Simplified:A gentle introduction to supervised learning. 2022. 199 p. URL: https://themlsbook.com
4. Maheshwari A. Business intelligence and data mining. *Business Expert Press*, 2014. URL: https://www.amazon.com/Business-Intelligence-Data-Mining-Analytics/dp/1631571206
5. Yang X. S. Introduction to Algorithms for Data Mining and Machine Learning. Academic Press, 2019. Doi: https://doi.org/10.1016/C2018-0-02034-4
6. Fernandes M. Statistics for business and economics. Bookboon, 2008. URL: https://bookboon.com/en/statistics-for-business-and-economics-ebook?mediaType=ebook
7. Vercellis C. Business intelligence: data mining and optimization for decision making. New York : Wiley, 2009. 420 p. URL: https://www.amazon.com/Business-Intelligence-Mining-Optimization-Decision/dp/0470511397
8. What is FRED? | Getting To Know FRED. URL: https://fredhelp.stlouisfed.org/fred/about/about-fred/what-is-fred/
9. Smith L. I. A tutorial on Principal Components Analysis (Computer Science Technical Report No. OUCS-2002-12). 2002. Retrieved from http://hdl.handle.net/10523/7534
10. Chicago Board Options Exchange, CBOE Volatility Index: VIX [VIXCLS], retrieved from FRED, Federal Reserve Bank of St. Louis. *Economic Research*. February 3, 2023. URL: https://fred.stlouisfed.org/series/VIXCLS

Надійшла (received) 10.03.2023
Прийнята до друку (accepted for publication) 17.05.2023

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Хорошун Ганна Миколаївна – кандидат фізико-математичних наук, доцент, доцент кафедри комп'ютерних наук та інженерії, Східноукраїнський національний університет імені Володимира Даля, Київ, Україна;
Ganna Khoroshun – candidate of physical and mathematical sciences, Associated Professor, Associated Professor of Computer Science and Engineering Department, Volodymyr Dahl East Ukrainian National University, Kyiv, Ukraine;
e-mail: an_khor@i.ua; ORCID ID: http://orcid.org/0000-0002-1272-1222.

Рязанцев Олександр Іванович – доктор технічних наук, професор, завідувач кафедри комп'ютерних наук та інженерії, Східноукраїнський національний університет імені Володимира Даля, Київ, Україна;
Oleksandr Ryazantsev – doctor of technical sciences, professor, head of Computer Science and Engineering Department, Volodymyr Dahl East Ukrainian National University, Kyiv, Ukraine;
e-mail: ryazantsev@ukr.net; ORCID ID: http://orcid.org/0000-0002-3740-3132.

Черпіцький Максим Вікторович – магістрант кафедри комп'ютерних наук та інженерії, Східноукраїнський національний університет імені Володимира Даля, Київ, Україна;
Maksym Cherpitskyi – master's student of Computer Science and Engineering Department, Volodymyr Dahl East Ukrainian National University, Kyiv, Ukraine;
e-mail: m_cherpitskiy@ukr.net; ORCID ID: http://orcid.org/0009-0001-8778-7576.

Clustering and anomalies of USA stock market volatility index data

Ganna Khoroshun, Oleksandr Ryazantsev, Maksym Cherpitskyi

Abstract. Actuality. Investing money is an important way to improve the financial condition of both an individual and society as a whole. The problem of understanding financial data and making decisions regarding the investment of money in a certain project at this moment in time is relevant. **The object** of the study is the process of establishing the dependence of the value of the asset on time. **The subject** of research is mathematical models of data processing, data clustering and anomaly detection**. The purpose** of this work is to develop a method for effective investment of money using data processing and analysis methods for CVOE volatility index values in the USA, determination of clusters based on actions with assets, as well as checking the presence of anomalous data. **Research results.** Official data of volatility index values were selected and prepared for further analysis by removing incomplete sets and further normalization. Clustering of time series was carried out and the array was divided into five homogeneous groups. Clusters determine the ranges of the volatility index, which reflect the different sentiments of investors in the market and encourage appropriate actions with assets: to sell, to wait of index increasing, to buy, to remove money from developing projects and invest in stable ones, to wait of index decreasing. Segmentation of data, application of window function, centroids for segments were determined and signal reconstruction was carried out. Data anomaly points were identified. A comparative analysis was carried out based on the results of constructed initial data, reconstructed data and reconstruction error.

Keywords: clustering; anomaly; volatility index.