СХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ВОЛОДИМИРА ДАЛЯ

Факультет ___інформаційних технологій та електроніки_____

Кафедра _____інформаційних технологій та програмування_____

Пояснювальна записка

до магістерської дипломної роботи

_____магістр_____
(освітньо-кваліфікаційний рівень)

на тему: _Дослідження процесу прогнозування фінансових часових рядів із зменшенням ознакового простору_____

Виконав: студент групи_ІСТ-22дм_____
126 «Інформаційні системи та технології___
(шифр і назва спеціальності)
_____Єфімов Д.С._____
(прізвище та ініціали)
Керівник ___Іванов В.Г._____
(прізвище та ініціали)
Рецензент_ Меняйленко О.С_____
(прізвище та ініціали)

Київ – 2023 року

СХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ВОЛОДИМИРА ДАЛЯ

Факультет ___інформаційних технологій та електроніки_____
Кафедра _____інформаційних технологій та програмування_____
Освітньо-кваліфікаційний рівень __магістр_____
Спеціальність 126 «Інформаційні системи та технології»_____
(шифр і назва спеціальності)

ЗАТВЕРДЖУЮ
Завідувач кафедри ІТП
_____ д.т.н., доц. Захожай О.І.
(підпис)
« ___ » _____2023 р.

ЗАВДАННЯ

на магістерську дипломну роботу студенту
_____Єфімов Даніїл Сергійович_____
(прізвище, ім'я, по батькові)

1.Тема роботи: Дослідження процесу прогнозування фінансових часових рядів із зменшенням ознакового простору_____
керівник роботи доцент, к.т.н. Іванов Віталій Геннадійович, ___
(вчене звання, науковий ступінь, прізвище, ім'я, по батькові)
затверджені наказом університету від «_ __ »____ 2023року № _____
2. Строк подання студентом роботи: 06 грудня 2023 р._____
3. Вихідні дані до роботи: Матеріали науково-дослідної практики, науково-методична література; дані інтернет-мережі . _____
4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити)
4.1 Вступ
4.2 Аналітичний огляд питання (огляд публічних джерел інформації)
4.3 Основна частина, в якої висвітлити методи, які будуть використовуватися для реалізації проекту.
4.4 Практична частина – огляд технологій, які використовуються під час реалізації проекту.
4.4 Висновки
4.5 Перелік використаних джерел
5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень)
_____
_____

6. Консультанти розділів проєкту (роботи)

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		Завдання видав	Завдання прийняв

7. Дата видачі завдання __20 жовтня 2023р._____

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів дипломної роботи	Строк виконання етапів роботи	Примітка
1.	Одержання завдання на виконання роботи	20.10.2023	
2.	Укладання і погодження з керівником плану і етапів виконання роботи	21.10.2023-24.10.2023	
3.	Узагальнення даних літературних джерел	25.10.2023-28.10.2023	
4.	Аналіз шляхів виконання завдання. Вибір і погодження з керівником оптимального шляху виконання завдання	29.10.2023-01.11.2023	
5.	Аналіз технічних засобів та існуючих систем	02.11.2023-07.11.2023	
6.	Реалізація практичної частини завдання	08.11.2023-24.11.2023	
7.	Укладання, оформлення та погодження пояснювальної записки з керівником	25.11.2023-05.12.2023	
8.	Здача пояснювальної записки на кафедру	06.12.2023	
9.	Підготовка доповіді та презентації	07.12.2023-09.12.2023	

Студент _____ Єфімов Д.С. _
(підпис) (прізвище та ініціали)

Керівник роботи _____ Іванов В.Г.
(підпис) (прізвище та ініціали)

РЕФЕРАТ

Магістерська дипломна работа: стор. 65 с., Рисунок 12, табл. 2, джерел 11.

Тема дипломної работи: "Дослідження процесу прогнозування фінансових часових рядів із зменшенням ознакового простору".

Метою роботи є побудова моделі прогнозування майбутніх значень фінансового індексу S&P500 на основі цін простих акцій компаній, на основі яких розраховується індекс. У ході побудови рішення проведено процедуру зменшення ознакового простору.

Магістерська дипломна работа складається із четирьох розділів. У першому розділі представлений огляд предметної області, який відбиває специфіку фінансового аналізу часових рядів, методи і підходи до вирішення завдання прогнозування. У наступних розділах зроблено постановку завдання та представлено докладний теоретичний аналіз існуючих методів для вирішення завдань зменшення розмірності та прогнозування часових рядів. В останньому розділі представлений алгоритм вирішення поставленого завдання та отримані результати.

Результати дослідження відображені у роботі у вигляді порівняння точності та ефективності використовуваних методів аналізу даних для вирішення задачі прогнозування фінансового індексу S&P 500.

ЛІНІЙНА МОДЕЛЬ РЕГРЕСІЇ, МЕТОД ОПОРНИХ ВЕКТОРІВ, НЕЙРОННІ МЕРЕЖІ, ФІНАНСОВІ РЯДИ, ПРОГНОЗУВАННЯ, КЛАСТЕРНИЙ АНАЛІЗ.

ABSTRACT

Master's thesis: p. 65 p., Figure 12, table. 2, sources 11.

Topic of the thesis: "Investigation of the process of forecasting financial time series with the reduction of the characteristic space".

The purpose of the work is to build a model for predicting future values of the S&P500 financial index based on the prices of common shares of the companies on the basis of which the index is calculated. During the construction of the solution, the procedure for reducing the characteristic space was carried out.

The master's thesis consists of four sections. The first chapter presents an overview of the subject area, which reflects the specifics of financial analysis of time series, methods and approaches to solving the forecasting task. In the following sections, the problem statement is made and a detailed theoretical analysis of existing methods for solving the problems of dimensionality reduction and time series forecasting is presented. The last chapter presents the algorithm for solving the task and the obtained results.

The results of the study are reflected in the work in the form of a comparison of the accuracy and efficiency of the data analysis methods used to solve the problem of forecasting the S&P 500 financial index.

LINEAR REGRESSION MODEL, SUPPORT VECTOR METHOD, NEURAL NETWORKS, FINANCIAL SERIES, FORECASTING, CLUSTER ANALYSIS

ЗМІСТ

ВСТУП

Прогнозування фінансових часових рядів є одним із найважливіших завдань аналізу фондового ринку, яка завжди привертала увагу стейкхолдерів з погляду пошуку нових методів та підходів для знаходження оптимального її вирішення. Вважається, що можливість зрозуміти і передбачити тенденції на фінансових ринках, а також цін акцій, дозволить знизити ризики і в певному сенсі обіграти ринок, отримавши дохід вище за звичайний, особливо у разі високого рівня волатильності.

З цією метою було розроблено теорію фінансового аналізу часових рядів, на основі якої були побудовані математичні моделі, які згодом стали класичними моделями фінансового аналізу.

Разом з тим розвиток нових технологій, у тому числі у фінансовій сфері, призвело до того, що класичні моделі згодом перестали ефективно відображати зміни, що відбуваються на фінансовому ринку.

У зв'язку з цим багато дослідників почали вдаватися до використання методів, які раніше не застосовувалися у фінансовому аналізі часових рядів у класичному його розумінні. Наприклад, стали активно використовуються різні методи аналізу даних (Data mining). Однією з переваг аналізу даних є те що, що його інструментарій дозволяє знаходити приховані і складні закономірності, які здатні відбити прості лінійні моделі.

Застосування інструментарію аналізу даних дозволило не обмежуватися лише класичним лінійним аналізом часових рядів, але й нівелювати їхні недоліки, зокрема «прокляття» розмірності. Оскільки фінансовий ринок схильний до впливу великої кількості макроекономічних, соціальних і політичних факторів, то при побудові ефективних математичних моделей, що передбачають або інші індикатори, потрібно враховувати великий обсяг інформації, отриманих із різних джерел, що має різну природу.

У роботі розглядається завдання прогнозування фінансового індексу S&P500 зі зниженням розмірності ознакового простору при побудові

математичної моделі.

Фінансовий індекс S&P 500 є агрегованим показником і відображає сумарну капіталізацію акцій, що входять до його складу.

При побудові моделі прогнозування майбутнього значення індексу S&P500 розглядалося завдання зниження розмірності у зв'язку з тим, що модель включені як пояснюють змінних такі параметри як значення самого індексу в попередні моменти часу, а також історичні значення цін акцій 500 компаній, що входять до складу індексу .

У ході побудови ефективного вирішення поставленого завдання було досліджено різні існуючі методи скорочення розмірності за умови мінімізації втрати інформативності ознак, а також лінійні та нелінійні моделі прогнозування.

За допомогою методів кластерного аналізу було проведено процедуру зниження розмірності ознакового простору, в результаті якої вдалося досягти необхідного рівня стиснення даних. При цьому, як показали результати прогнозування, вдалося зберегти інформативність ознакового простору, що дало змогу отримати якісний прогноз. Як розглянуті моделі прогнозування були використані лінійна модель регресії, метод опорних векторів і нейронна мережа.

РОЗДІЛ 1 ОГЛЯД ПРЕДМЕТНОЇ ГАЛУЗІ

Фінансовий ринок відіграє велику роль, впливаючи на зростання чи спад як глобальної економіки в цілому, так і економік окремих макро- та мікрорегіонів. Ефективне передбачення поведінки фінансових рядів дають інвесторам можливість нівелювати можливі ризики при управлінні портфелем ризикових активів і побудувати стратегії, що хеджують, а також дозволить у деякому сенсі обіграти ринок, отримавши дохід вище звичайного, особливо у разі високого рівня волатильності цінних паперів.

З цією метою було розроблено теорію фінансового аналізу часових рядів, на основі якої були побудовані математичні моделі, які згодом стали класичними моделями фінансового аналізу. До таких моделей відносять авторегресійна модель (AR), модель ковзного середнього (MA), авторегресійна модель ковзного середнього (ARMA), моделі авторегресійної умовної гетероскедастичності (ARCH, GARCH) та інші складніші моделі. Крім вимог стаціонарності прогнозованого часового ряду зазначені моделі будувалися на основі постулатів, що інвестор приймає рішення раціонально і фінансовий ринок безарбітражний. Існує безліч досліджень щодо оптимальних механізмів оцінки параметрів цих моделей.

Разом з тим розвиток нових технологій, у тому числі у фінансовій сфері, призвело до того, що зазначені постулати не виконуються і відповідно класичні моделі згодом перестали ефективно відображати зміни, що відбуваються на фінансовому ринку. Це викликало необхідність побудови нових моделей, здатних відбивати складну і неоднорідну природу фінансових коливань, викликаних, зокрема зростаючим кількістю чинників, які впливають динаміку часових рядів [2].

У зв'язку з цим багато дослідників почали вдаватися до використання методів, які раніше не застосовувалися у фінансовому аналізі часових рядів у класичному його розумінні. Наприклад, стали активно використовуються різні методи аналізу даних (Data mining). Однією з переваг аналізу даних є те що, що

його інструментарій дозволяє знаходити приховані і складні закономірності, які здатні відбити прості лінійні моделі.

Іншим підходом до вирішення завдання прогнозування часових рядів є побудова гібридних моделей на стику класичної теорії фінансового аналізу та нових методів аналізу даних. Крім того, якщо раніше як пояснюючі змінні розглядалися в основному історичні дані прогнозованого значення, а також різні економічні показники, то тепер як такі змінні можуть виступати дані з різних доменів інформації. Наприклад, деякі дослідницькі роботи присвячені аналізу залежностей поведінки часового ряду від тональності зведень новин, що мають відношення до фінансового ринку. Також деякі моделі включають показники технічного аналізу фінансових рядів. Було показано, що існує залежність між тональністю фінансових новин та динамікою ризикових активів. У багатьох роботах проводиться аналіз тональності текстової інформації як Twitter, LiveJournal та інші соціальні медіа з метою отримання знань, які дозволять як мінімум визначити напрямок динаміки поведінки ризикових активів.

Основними методами аналізу даних, що використовуються для прогнозування фінансових рядів, є методи опорних векторів, нейронні мережі.

У деяких дослідницьких роботах ставиться завдання класифікації, метою якої є визначення трендів зростання чи падіння, де застосовуються такі методи як байєсівський класифікатор, логістична регресія, метод «випадкового лісу». Останнім часом популярністю користуються методи "глибокого" навчання - нейронні мережі різної архітектури, від простої до складної, з великою кількістю шарів.

Крім того, класичні моделі, такі як модель ковзного середнього (MA), авторегресійна модель ковзного середнього (ARMA), моделі авторегресійної умовної гетероскедастичності, також продовжують використовуватися.

Побудова гібридних моделей або включення додаткових змінних у модель має на меті збільшення інформативності пояснюючих змінних про досліджуваний фактор. Оскільки фінансовий ринок схильний до впливу великої кількості макроекономічних, соціальних і політичних факторів, то при побудові

ефективних математичних моделей, що передбачають ті чи інші індикатори, потрібно враховувати великий обсяг інформації, отриманих з різних джерел і різну природу. Це призводить до необхідності розробки нових алгоритмів обробки даних з різних джерел та їх комбінування.

Одним із ключових питань при побудові оптимальної моделі прогнозування часових рядів є розмір навчальної вибірки. Невелика кількість змінних може призвести до того, що модель з досить великою похибкою прогнозуватиме майбутні значення. Надмірна кількість змінних може призвести до перенавчання, яке також призведе до високої похибки моделі. Це питання присвячено достатню кількість досліджень.

Існують аналітичні методи для визначення оптимального розміру вибірки, але на практиці їх застосування не є тривіальним через необхідність урахування великої кількості параметрів, у тому числі розподіл змінних, тип моделі, залежності між змінними.

У роботі розглядається завдання прогнозування фінансового індексу S&P 500 зі зниженням розмірності ознакового простору.

Фінансові індекси, у тому числі індекс S&P 500, є агрегованими показниками та відображають стан розвинених економік світу. Так індекс S&P 500 враховує динаміку зміни цін на ризикові активи 500 компаній, що яскраво характеризують стан того чи іншого сектора економіки. До нього включено в основному акції компаній, зареєстрованих на Нью-Йоркській фондовій біржі, проте присутні також акції деяких корпорацій, які котируються на Американській фондовій біржі та у позабіржовому обороті. Індекс становить близько 80% ринкової вартості всіх випусків, котируються на Нью-Йоркській фондовій біржі. Прогнозування індексу дозволяє інвестору формувати оптимальний ризиковий портфель і будувати беззбиткову стратегію, що хеджує.

Фінансовий індекс S&P 500 має найбільший інтерес з погляду дослідження та прогнозування його майбутніх значень, оскільки він є певним барометром для інвесторів з метою ухвалення рішення. Крім того, існує ряд похідних ризикових активів (ф'ючерси, опціони), вартість яких так чи інакше залежить від динаміки

значень індексу S&P500.

У розглянутій задачі прогнозування значення індексу S&P 500 постає завдання зниження розмірності, оскільки при побудові моделі з метою використання якнайбільше інформації враховуються не тільки історичні значення самого індексу, але історичні значення цін акцій, що входять до його складу, а це означає, що модель включає додатково 500 змінних [4].

У ході побудови ефективного вирішення поставленого завдання було досліджено різні існуючі методи скорочення розмірності за умови мінімізації втрати інформативності ознак, а також лінійні та нелінійні моделі прогнозування.

За допомогою методів кластерного аналізу було проведено процедуру зниження розмірності ознакового простору, внаслідок якої вдалося досягти необхідного рівня стиснення даних. При цьому, як показали результати прогнозування, вдалося зберегти інформативність ознакового простору, що дало змогу отримати якісний прогноз. Як розглянуті моделі прогнозування були використані лінійна модель регресії, метод опорних векторів і нейронна мережа.

Завданню прогнозування фінансового індексу S&P 500 присвячено багато досліджень, але при цьому в роботах зроблено акцент на прогнозуванні індексу на основі залежностей від інших економічних показників або з використанням аналізу даних. Проблему надмірної багатовимірності простору ознак найчастіше вирішується з допомогою методу головних компонент, який завжди дає бажаний результат. У цьому роботі завдання зниження розмірності часових рядів побудований алгоритм з допомогою методів кластерного аналізу, який дозволив отримати адекватні результату прогнозування індексу.

Визначення. Тимчасовий (чи динамічний) ряд – це впорядкована у часі сукупність чи послідовність вимірів однієї з характеристик об'єкта, що досліджується.

Нехай є фінансовий часовий ряд фінансового індексу S&P 500

$SP=\{SPt\}_{t\in T}$, майбутні значення якого необхідно спрогнозувати. Як пояснюючі змінні виступають історичні значення прогнозованого ряду, а також

кілька пояснюючих змінних $S=\{\{Si\}_{t\in T}\}_{i=1..m}$

- Значення цін простих акцій 500 компаній, з урахуванням яких розраховується фінансовий індекс.

Необхідно побудувати модель прогнозування значень фінансового індексу S&P 500, у своїй необхідно скоротити розмірність ознакового простору завдання.

Оскільки моделі як пояснюючих змінних обрані як самі історичні значення індексу, а й історичні значень цін акцій – компонентів індексу, маємо щодо високу розмірність ознакового простору, що підвищує трудомісткість побудови моделі прогнозування і призводить до її надмірності з погляду змінних.

Разом з тим, при скороченні розмірності ознакового простору необхідно мінімізувати втрату інформативності даних.

РОЗДІЛ 2 ЗАДАЧА ЗМЕНШЕННЯ РОЗМІРНОСТІ ОЗНАКОВОГО ПРОСТОРУ

Аналіз предметної області показав, що залежно від типу вихідних даних, галузі застосування, розмірності даних можуть бути застосовані різні методи, спрямовані на скорочення ознакового вектора з мінімальною втратою прихованої інформації [6].

Усі методи можна розділити на 2 класи: скорочення простору без його трансформації та скорочення простору з трансформацією. Одним із найпоширеніших методів зменшення розмірності ознакового простору є метод головних компонентів (the principal component analysis, PCA). Цей метод відноситься до групи методів, які трансформують вихідний простір високої розмірності в новий простір меншої розмірності.

2.1. Метод головних компонентів (PCA)

Нехай є деяка вибірка об'єктів

$$X = \{x_n\}_{n=1}^{N}, x_n \in R^m \qquad (2.1)$$

Завдання зменшення розмірності полягає в отриманні представлення цієї вибірки у просторі меншої розмірності

$$T = \{t_n\}_{n=1}^{N}, x_n \in R^p, \text{де } p \ll m \qquad (2.2)$$

Ідея методу полягає у пошуку у вихідному просторі гіперплощини заданої розмірності з подальшим проектуванням вибірки на дану гіперплощину. Як критерій вибору гіперплощини є максимізація розкиду спроектованих точок вибірки.

Для знаходження нової гіперплощини необхідно побудувати новий ортогональний базис, осі якого орієнтовані за напрямами максимальної дисперсії набору вхідних даних.

Характеристикою розкиду даних в одновимірному просторі є вибіркова

дисперсія:

$$DX = \frac{1}{n}\sum_{i=1}^{n}(x_i - Mx_i)^2 = \frac{1}{n}u^T x x_u^T = u^T S \qquad (2.3)$$

Характеристикою розкиду даних у багатовимірному просторі є вибіркова матриця коваріації:

$$S = Cov(x) = Mxx^T - MxMx^T \qquad (2.4)$$

Як скалярний критерій розкиду виберемо слід вибіркової матриці коваріації, що еквівалентно сумі вибіркових дисперсій по ортогональних напрямках u_1, u_2, … u_p.

Тоді критерій знаходження оптимального рішення:

$$J = u^T S u \to max, \qquad (2.5)$$

$$u^T u = 1$$

Оптимальними векторами u_i, відповідно до вирішення задачі знаходження глобального максимуму, є власні вектори матриці S, що відповідають її p найбільшим власним значенням.

Позначимо

$$U = [u_1, \dots, u_p] \qquad (2.6)$$

Тоді редукція X при проектуванні на оптимальну гіперплощину обчислюється як

$$t_n = (x_n - \mu)^T U \qquad (2.7)$$

а самі точки проекції визначаються як

$$(x_n)_{pr} = t_n^T U + \mu \qquad (2.8)$$

Таким чином, метод головних компонентів передбачає перехід від вихідного базису до базису з власних векторів матриці коваріації S з подальшим

відкиданням проекцій вибірки на власні вектори, що відповідають n - p найменшим власним значенням. У базисі із власних векторів матриця коваріації S має діагональний вигляд

$$\Lambda=\text{diaq}(\lambda_1, \dots, \lambda_p) \qquad (2.9)$$

Отже, ознаки, одержувані з допомогою методу головних компонентів, є некорельованими. Перехід до некорельованих ознак часто є розумним методом попередньої обробки вихідних даних.

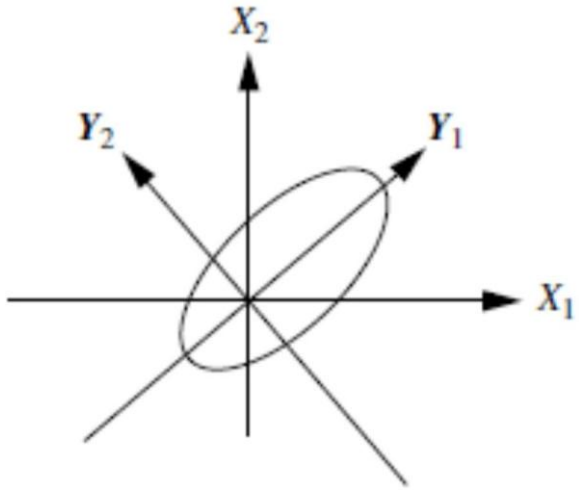


Рисунок 1. Метод головних компонентів. Вектори є першими двома головними компонентами для вихідної вибірки.

В цілому, процедура виділення головних компонент подібна до обертання, що максимізує дисперсію вихідного простору змінних. Обертання називається обертанням, що максимізує дисперсію, оскільки критерій (мета) обертання полягає в максимізації дисперсії (мінливості) "нової" змінної (фактора) та мінімізації розкиду навколо неї.

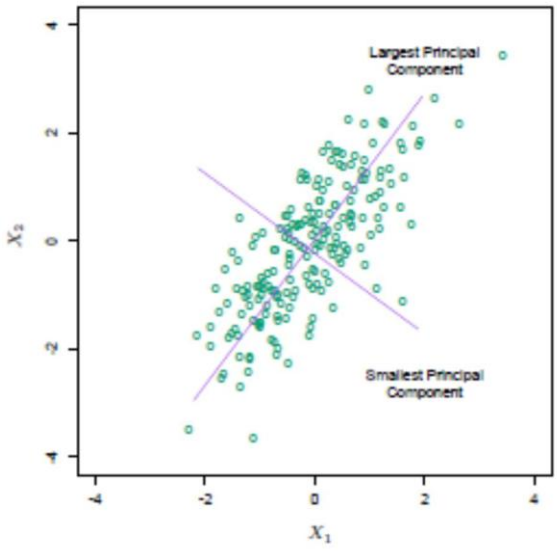


Рисунок 2. Ілюстрація способу головних компонент.

Варто зазначити, що для застосування методу головних компонентів необхідно виконання наступних умов:

– розмірність даних може бути ефективно знижена шляхом лінійного перетворення;

– більше інформації несуть напрямки, в яких дисперсія вихідних даних максимальна.

При цьому на практиці ці умови не завжди виконуються. Наприклад, якщо точки вихідної множини розташовуються на поверхні гіперсфери, то ніяке лінійне перетворення не зможе знизити розмірність (але з цим легко впорається нелінійне перетворення, що спирається на відстань від точки до центру сфери).

Це недолік однаковою мірою властивий усім лінійним алгоритмам і може бути подолано за рахунок використання додаткових фіктивних змінних, які є нелінійними функціями від елементів набору вихідних даних (kernel trick).

Другий недолік методу головних компонентів полягає в тому, що напрямки, що максимізують дисперсію, далеко не завжди максимізують інформативність.

2.2. Непараметричний (ядерний) метод головних компонентів (Kernel PCA)

Даний метод є узагальненою версією методу головних компонентів і застосовується у разі, коли умова зниження розмірності за допомогою лінійного перетворення не виконується.

Вводиться поняття ядерної функції

$$K(x_i, x_j) = \phi(x_i)\phi(x_j)^T \qquad (2.10)$$

Тоді вибіркова матриця коваріації як характеристика розкиду в багатовимірному просторі матиме вигляд:

$$S = Cov(X) = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)\phi(x_i)^T \qquad (2.11)$$

2.3.Кластерний аналіз

Кластерний аналіз включає широкий набір методів для визначення підмножин або кластерів у вихідній вибірці. Мета кластеризації полягає в тому, щоб розділити вихідну вибірку так, щоб спостереження, що потрапили в один кластер, були схожі та відмінні від тих спостережень, які потрапили в інший кластер.

У зв'язку з цим дуже важливо правильно визначити критерії або метрики, що вимірюють відмінність/схожість спостережень між собою, що є одним із головних завдань кластерного аналізу.

Область застосування кластерного аналізу досить широка: розпізнавання образів, інтернет – пошук, біологія, медицина, маркетинг, соціологія тощо. У тому числі кластерний аналіз застосовується в завданнях стиснення даних, коли вихідна вибірка досить велика, можна скоротити її, залишивши по одному найбільш типовому представнику від кожного кластера.

Залежно від сфери застосування кластерного аналізу існує кілька підходів. У рамках завдання розглядається імовірнісний підхід. Найпоширенішим методом даного підходу є метод *к -середніх.*

2.3.1. Метод К – середніх (K-Means)

Метод до-середніх є простим і елегантним методом для поділу вибірки на різні кластери, що не перетинаються [5].

Нехай C_1,. . . , C_K - безлічі індексів спостережень відповідного кластера. Дані множини задовольняють двом умовам:

1. $C_1 \cup C_2 \cup \ldots \cup C_K = \{1,2, \ldots , n\}$. Тобто множини C*k* являють собою повне розбиття вихідної множини індексів, що відповідає вихідній вибірці.

2. $Ck \cap Ck' = \emptyset$ для $k \neq k'$. Тобто виконується умова непересічності множин – кластерів. Кожне спостереження може належати лише одному кластеру.

Отже, якщо і-ое спостереження входить у к-ый кластер, отже

$i \in C_k$. Ідея методу до-середніх, полягає в мінімізації відстані між спостереженнями, що входять до одного кластеру. Відстань для кластера Ck це міра $W(Ck)$ того, наскільки спостереження відмінні один від одного.

Таким чином, критерій виглядає так

$$J = min\{\sum_{k=1}^{K} W(C_k)\} \quad (2.12)$$

Існує велика кількість метрик, що використовуються для визначення відстані всередині кластера (відстань Махаланобіса, відстань Мінковського тощо). Найпоширенішою метрикою є евклідовий простір:

$$W(C_k) = \frac{1}{|C_k|} \sum i, l \in C_k \sum_{j=1}^{p} \left(x_{ij} - x_{l,j}\right)^2 \quad (2.13)$$

де $|Ck|$ - Кількість елементів у к-му кластері. Іншими словами, внутрішньокластерна відстань вимірюється як сума всіх попарних евклідових відстаней між спостережень до кластера, розділена на кількість спостережень до кластера.

Таким чином, критерій набуває вигляду:

$$min\left\{\sum_{k=1}^{K} \frac{1}{|C_k|} \sum i, l \in C_k \sum_{j=1}^{p} \left(x_{i,j} - x_{l,j}\right)^2\right\} \quad (2.14)$$

Сам алгоритм побудовано так:

– На першому кроці випадково вибираються центри майбутніх кластерів (кількість кластерів відома). До кожного спостереження вибірки розраховується відстань кожного з центрів. Спостереження буде віднесено до кластера, від центру якого до розглянутого спостереження відстань буде мінімальною.

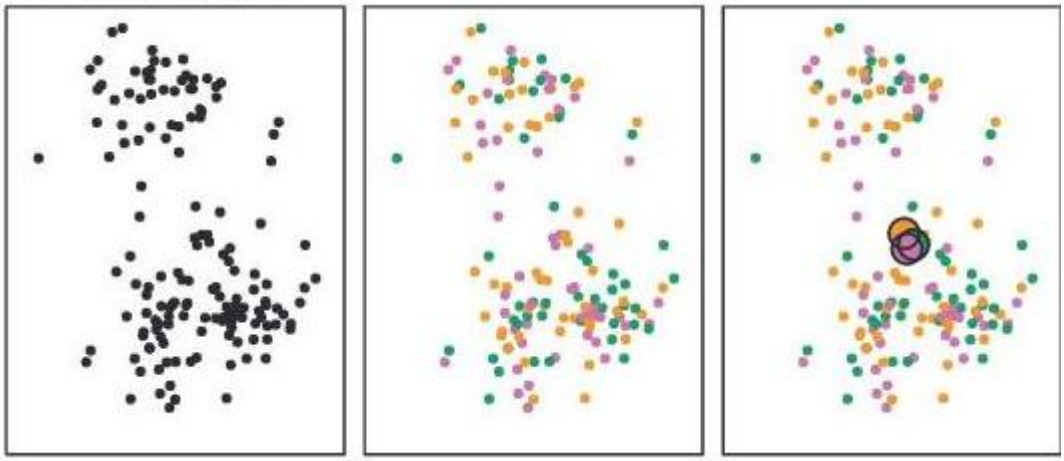
– Потім на кожній ітерації заново для кожного отриманого на попередньому кроці кластера обчислюється центр як середнє вибіркове значення, вектори розбиваються на кластери знову відповідно до того, який з нових центрів виявився ближчим за обраною метрикою.

Алгоритм завершується, коли якоїсь ітерації немає зміни центрів

кластерів. Це відбувається за кінцеве число ітерацій, так як кількість можливих розбиття кінцевої множини звичайно, а на кожному кроці сумарне квадратичне відхилення не збільшується, тому зациклювання неможливо.

Недоліком методу *К – середніх* є те, що алгоритм знаходить рішення задачі оптимізації локальний, а не глобальний мінімум. У зв'язку з цим підсумкове рішення залежить від обраних на першому кроці випадковим чином центрів майбутніх кластерів.

Вихідна модель



Результат роботи алгоритму к-середніх

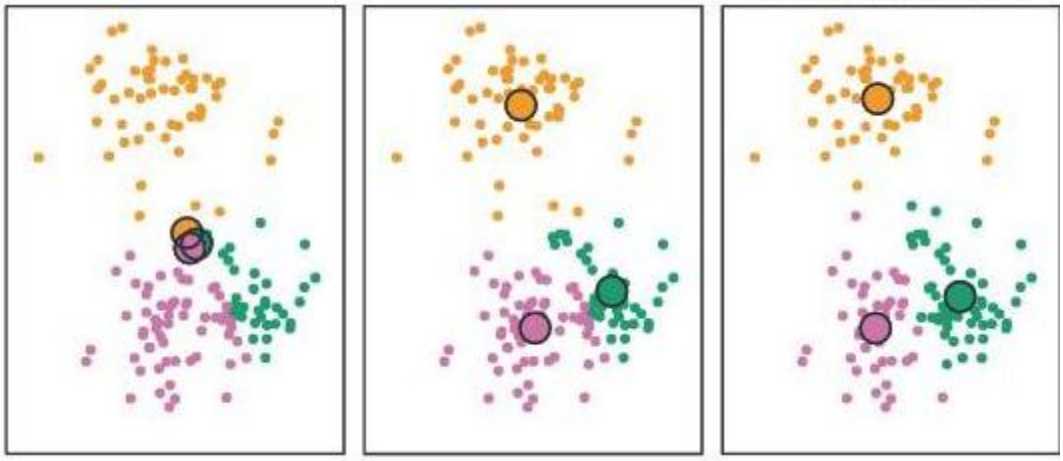


Рисунок 3. Ілюстрація методу К-середніх.

Крім того, застосування даного алгоритму вимагає заздалегідь визначити значення К (кількість кластерів), що є нетривіальним завданням, особливо у випадку багатовимірних даних.

2.3.2. Метод К-медіан (K – medians)

Метод кластерного аналізу *К-медіан* (k-medians) є варіацією методу *К-середніх*. Відмінність у тому, що визначення центру (центроїда) формованого кластера замість середнього обчислюється медіана. Це відповідає мінімізації помилки за всіма кластерами в метриці з 1- нормою, замість метрики з 2-нормою для *К-середніх*

$$min\left\{\sum_{k=1}^{K}\frac{1}{|C_k|}\sum i,l\in C_k\sum_{j=1}^{p}\left(x_{i,j}-x_{l,j}\right)\right\} \quad (2.15)$$

Метод *К-медіан* (k-medians) іноді працює краще, ніж метод *К-середніх* (k-means), де мінімізується сума квадратів відстаней, оскільки є нечутливою до так званих викидів (outliers).

2.3.3. Ядерний метод К-середніх (Kernel K-means)

Метод *К-середніх* здатний визначити оптимальну структуру кластеризації, якщо кластери є лінійно роздільними.

Основна ідея методу: побудувати проекцію вихідної множини в простір вищої розмірності, а потім застосувати метод до- середніх у новому просторі.

Алгоритм методу:

-Для кожної точки вихідного простору ставимо у відповідність деяку точку із простору вищої розмірності за допомогою використання ядерної функції.

- До даних, отриманих в результаті проекції, застосовуємо алгоритм *К-середніх*.

Обчислювальна складність даного методу вища, ніж у класичного методу до-середніх: у ході реалізації алгоритму необхідно обчислювати та зберігати в пам'яті матриці розміром nxn, одержуваних у результаті застосування ядерної функції щодо вихідних даних.

Широко використовується спектральний метод кластеризації може бути розглянутий як варіація ядерного методу до-середніх.

Однією з найбільш широко використовуваних ядерних функцій є

гауссівське ядро, яке має вигляд:

$$K(x_i, x_j) = exp\left\{\frac{-\|x_i - x_j\|^2}{2\sigma^2}\right\} \quad (2.16)$$

Критерій мінімізації може бути сформульований наступним чином:

$$J = \sum_{K=1}^{K} \sum x_i \in C_k \|\phi(x_i) - C_k\|^2 \quad (2.17)$$

Формула для обчислення центру кластера:

$$C_k = \frac{\sum x_i \in C_k \phi(x_i)}{|C_k|} \quad (2.18)$$

2.3.4. Спектральний кластерний аналіз (алгоритм Ng-Jordan-Weiss)

Дано безліч $S = \{s_1, s_2, \dots, s_n\} \in R^m$, яку необхідно розділити на k кластерів. Для цього:

1. Складаємо матрицю відповідності $A \in R^{nxn}$, яка визначається за формулою

$$A_{ij} = \begin{cases} exp\frac{-\|s_i - s_j\|^2}{2\sigma^2}, \text{якщо } i \neq j \\ 0, \\ \text{якщо } i = j \end{cases} \quad (2.19)$$

2. Визначаємо матрицю D як діагональну матрицю, чий діагональний елемент є сума елементів відповідного рядка матриці A та знаходимо матрицю L за формулою:

$$L = D^{-1/2} A D^{-1/2} \quad (2.20)$$

3. Знаходимо власні вектори матриці L і вибираємо перші вектори k з найбільшими значеннями. Формуємо матрицю X:

$X = \{x_1, x_2, \dots, x_k\} \in R^{nxk}$.

4. На цьому кроці формуємо матрицю Y на основі матриці X:

$$Y_{ij} = \frac{x_{ij}}{(\sum_j x_{ij})^{1/2}} \qquad (2.21)$$

5. Розглядаючи кожен рядок матриці Y як координати деякої точки в просторі R_k , отримуємо безліч точок, які ділимо на k кластером з використанням алгоритму k-means.

6. В результаті вихідну точку s_i буде віднесено до кластера j тоді і тільки тоді, якщо і-ий рядок матриці Y віднесений до кластера j.

Таким чином, алгоритм будує проекцію і ставить у відповідність вихідну множину деякій множині точок у просторі меншої розмірності.

2.3.5. Умова опуклості досліджуваної множини

Визначення. Безліч $G \in R^n$ називається опуклим, якщо для x_1, $x_2 \in G$ існують таке значення λ, що лінійна комбінація:

$$\lambda x_1 + (1 - \lambda)x_2 \in G \qquad (2.22)$$

Іншими словами, безліч G є опуклим, якщо дана множина разом з будь-якими двома своїми точками містить у собі відрізок, що з'єднує ці точки.

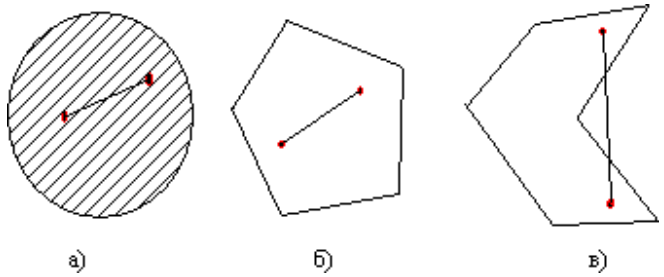


Рисунок 4. Приклади множин: а), б) – опуклі множини, в) – невипуклі безліч.

Більшість класичних методів кластерного аналізу, зокрема до середніх, припускають, що безліч точок, які необхідно розбити на кластери, і відповідно кластери, є опуклими. В іншому випадку ці методи можуть дати неякісний результат. Якщо безліч не є опуклим, застосовують інші методи, в основі яких спектральний аналіз або аналіз щільності даних.

РОЗДІЛ 3 ЗАДАЧА ПРОГНОЗУВАННЯ ЧАСОВИХ РЯДІВ

Фінансовий аналіз часових рядів пов'язаний з теорією та практикою оцінювання зміни (поведінка) акцій у часі.

Ключовою характеристикою фінансового (фондового) ринку, що визначає методи та моделі, що використовуються при аналізі часових рядів, є елемент невизначеності. Математичні статичні методи грають велику роль фінансовому аналізі часових рядів [7].

Основними поняттями аналізу часових рядів є прибутковість та волатильність. Більшість методів оперує замість ціни поняттям прибутковості цінних паперів. Це пов'язано з тим, що дохідність дає повнішу картину, ніж ціна і є безрозмірною відносною величиною, що полегшує застосування статистичних методів.

Визначення. Прибутковість цінного паперу - це кількісна характеристика цінного паперу, що визначається за формулою:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}} \qquad (3.1)$$

де P_t –ціна ризикового активу (акції, облігації чи індексу) на момент часу t.

Також під прибутковістю розуміють величину:

$$r_t = ln(1 + R_t) = ln\frac{P_t}{P_{t-1}} = p_t - p_{t-1} \qquad (3.2)$$

Дані визначення еквіваленти між собою, тому немає різниці яке визначення прибутковості використовувати при аналізі фінансових тимчасових рядів.

Лінійний аналіз фінансових часових рядів дозволяє вивчити їхню динамічну структуру.

Даний аналіз включає наступні ключові характеристики:

– стаціонарність,

– динаміка залежності,

– автокореляційна функція (АКФ),

– моделювання та прогноз майбутніх значень прибутковості.

У класичній теорії аналізу фінансових часових рядів існує низка економетричних моделей, за допомогою яких описують динаміку поведінки прибутковості. Прикладом таких моделей можуть послужити: проста авторегресійна модель (AR(p)), модель ковзного середнього (MA(q)), авторегресійна модель ковзного середнього (ARMA (p,q)), сезонна модель і т.д. Дані моделі дозволяють оцінити лінійну залежність значень часового ряду, оскільки це відіграє у розумінні моделей. Основний фокус у моделях робиться в залежності від поточного значення від його минулих (історичних) значень.

Крім того, теорія фінансового аналізу будуватиметься на наступних припущеннях:

1. Тимчасовий ряд є стаціонарним.

Визначення. Говорять, що ряд є строго стаціонарним, якщо функція розподілу задовольняє умову:

$$F(r_{t_1}, \dots, r_{t_k}) = F(r_{t_1}, \dots, r_{t_{k+1}}) \qquad (3.3)$$

тобто ряд є інваріантним щодо часу.

Виконання умови суворої стаціонарності складно визначити емпірично. Тому на практиці вважається достатнім виконання менш суворої умови: виконання умови, що тимчасовий ряд стаціонарний у широкому розумінні.

Визначення. Кажуть, що тимчасовий ряд є стаціонарним у сенсі, якщо його математичне очікування залежить від часу, а функція ковариации залежить лише від довжини часового інтервалу.

$$\mathrm{M}(\mathrm{r_t}) = \mu = \text{const}, \qquad Cov(r_t, r_{t-1}) = \gamma_l \qquad (3.4)$$

Насправді передбачається, що є перші T спостережень. Стаціонарність у сенсі означає, що у графіці ці значення потраплятимуть у певний інтервал. Обмовляючи виконання цієї умови, неявно передбачається, що значення перших двох моментів часового ряду кінцеві. Також з визначень випливає, що якщо ряд

строго стаціонарний і значення перших двох моментів кінцеві, то ряд стаціонарний і в широкому сенсі. Назад, в загальному випадку, не вірно.

Коли необхідно виявити залежність між значеннями ряду, використовують функцію автокореляції, що визначається за формулою:

$$p_t = \frac{Cov(r_t, r_{t-1})}{\sqrt{D(r_t)D(r_{t-1})}} \frac{Cov(r_t, r_{t-1})}{D(r_t)} = \frac{\gamma_l}{\gamma_0} \qquad (3.5)$$

Значення функції автокореляції змінюються на інтервалі [-1,1]. Стаціонарний у широкому значенні ряд буде некорельованим, якщо $\rho_t = 0$ для всіх $t>0$.

3.1 Лінійна регресійна модель прогнозування

Одним із простих видів даної моделі є модель першого порядку, яка має вигляд:

$$r_t = \phi_0 + \phi_1 r_{t-1} + a_t \qquad (3.6)$$

де ряд $\{a_t\}$ є білим шумом з нульовим математично очікуванням та дисперсією σ_a^2 .Ця модель використовується при моделюванні волатильності стохастичного процесу, де замість r_t виступає значення волатильності.

Властивості моделі:

1. $M(r_t|r_{t\_1}) = \phi_0 + \phi_1 r_{t\_1}, D(r_t|r_{t\_1}) = D(a_t) = \sigma_a^2$ (3.7)

Ця умова означає, що якщо відомо значення прибутковості в минулий момент часу, то поточне значення центроване щодо рівня $\phi_0 + \phi_1 r_{t-1}$ зкидом σ_a^2 . Крім того, це означає, що виконується властивість Маркова: поточне значення залежить лише від попереднього значення і ніяк не залежить від значень у минулому.

Разом з тим очевидно, що умова Маркова не завжди виконується, зокрема, коли використовується модель вищого порядку, яка гнучкіше відображає залежність поточного значення від минулих значень.

Загальний вигляд моделі:

$$r_t = \phi_0 + \phi_1 r_{t-1} + \phi_2 r_{t-2} + \cdots + \phi_p r_{t-p} + a_t, \quad (3.8)$$

де p - порядок моделі.

2. $M(r_t) = \phi_0 + \phi_1 M(r_{t-1})$

З умови стаціонарності випливає, що $M(r_t) = M(r_{t-1}) = \mu$. Звідси випливає, що значення математичного очікування задовольняє умову:

$$\mu = \frac{\phi_0}{1-\phi_1} \quad (3.9)$$

У зв'язку з цим, накладаються обмеження на коефіцієнт при r_t: $\phi 1 \neq 1$. Таким чином, математичне очікування ряду дорівнюватиме нулю тоді і тільки тоді, коли $\phi_0 = 0$.

3. $D(r_t) = \phi_1^2 D(r_{t-1}) + \sigma_a^2$

З умови стаціонарності випливає, що

$$D(r_t) = \frac{\sigma_a^2}{1-\phi_1^2} \quad (3.10)$$

Оскільки дисперсія випадкової величини має бути кінцевою та невід'ємною, то коефіцієнт $\phi_1^2 < 1$.

Одним з найважливіших елементів аналізу часових рядів є прогнозування. Значення прогнозу вибирається за таким критерієм:

$$M\left(r_{k+1} - \hat{r}_k(l)\right)^2 \leq M(r_{k+1} - g)^2 \quad (3.11)$$

де g – деяка функція, l – визначає крок прогнозу.

Розглянемо однокроковий прогноз у разі лінійно регресійної моделі.

Загальний вигляд моделі:

$$r_{k+1} = \phi_0 + \phi_1 r_k + \phi_2 r_{k-1} + \cdots + \phi_p r_{k+1-p} + a_{k+1} \quad (3.12)$$

Тоді однокроковий прогноз визначатиметься за формулою

$$\hat{r}(l) = M(r_{k+1}|r_k, r_{k-1}, \dots) = \phi_0 + \sum_{i=1}^{p} \phi_i \, r_{k+1-i} \quad (3.13)$$

та помилка прогнозу:

$$e_h(l) = r_{h+1} - \hat{r}_h(l) = a_{h+1} \quad (3.14)$$

Варто зазначити, що порядок моделі, яка застосовується для прогнозування часових рядів, заздалегідь не відома і визначається емпірично. Існує два головних підходи до визначення порядку: перший метод використовує приватну функцію автокореляції, а другий метод визначає порядок за допомогою локального критерію.

3.2 Регресійна модель з використанням методу опорних векторів

Метод опорних векторів є потужним та досить гнучким інструментом для вирішення задач машинного навчання.

Одним із важливих етапів побудови регресії з використанням методу опорних векторів є налаштування ряду внутрішніх параметрів. При використанні довільних значень параметрів алгоритму опорних векторів якість роботи алгоритму може суттєво змінюватись. Існують різні методи вибору параметрів алгоритму, проте жоден з них не має універсальності [9].

Завданням побудови регресії є оцінка невідомої речової функції:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \xi \quad (3.15)$$

Для визначення параметрів моделі використовується функція втрати (штрафна функція). Критерій мінімізації для пошуку оптимальних значень коефіцієнтів регресійної моделі визначається як:

$$Cost \sum_{i=1}^{n} L_\epsilon(y_i - \hat{y}_i) + \sum_{\mathrm{i=1}}^{\mathrm{p}} \beta_{\mathrm{j}}^2 \quad (3.16)$$

де $L_\epsilon(*)$ – деяка функція. Параметр Cost – певна величина, яка

визначається емпірично і задає допустиме відхилення для результату.

Проста лінійна регресія визначає прогнозне значення на основі лінійної комбінації даних та параметрів. Для деякої вибірки даних u регресійне рівняння має вигляд:

$$\hat{y} = \beta_0 + \beta_1 u_1 + \beta_2 u_2 + \cdots + \beta_p u_p = \beta_0 + \sum_{j=1}^{p} \beta_j u_j \qquad (3.17)$$

Рівняння у разі методу опорних векторів має схожий на рівняння (3.17) вигляд. Оцінки параметрів у цьому випадку можуть бути записані як функції від невідомих параметрів αi та даних вибірки таким чином, як

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j \, u_j = \beta_0 + \sum_{j=1}^{p} \sum_{i=1}^{n} \alpha_i \, x_{ij} u_j = \beta_0 + \sum_{i=1}^{n} \alpha_i \left(\sum_{j=1}^{n} x_{ij} \, u_j \right) \qquad (3.18)$$

З точки зору класичної регресійної моделі рівняння (3.18) виглядає занадто параметризованим, зазвичай набагато краще, якщо кількість параметрів набагато менше, ніж спостережень у вибірці. Водночас параметр *cost* добре регулює процес оцінки параметрів, знижуючи трудомісткість.

Рівняння можна записати у більш компактному вигляді:

$$f(u) = \beta_0 + \sum_{i=1}^{n} \alpha_i \, K(x_i, u) \qquad (3.19)$$

де $K(*,*)$ – ядерна функція, яка б задовольняла умовами Мерсера. Елементи даних, яким відповідають ненульові значення α_i, називають опорними векторами.

Вибір відповідної ядерної функції - ключове завдання при отриманні якісної моделі регресії. Одні з найширше використовуваних ядерних функцій – RBF-ядра. Це універсальні апроксиматори.

$$RBF \; = \; exp(-\sigma \| x \; - \; u \| 2) \qquad (3.20)$$

3.3 Регресійна модель з використанням нейронної мережі

Нейронні мережі це потужний метод моделювання, в основі якого лежить ідея схожа на принцип роботи мереж нервових клітин людського мозку. Моделі нейронних мереж дозволяють відтворювати складні та нелінійні залежності.

Вихідний параметр моделі нейронних мереж обчислюється за допомогою множини латентних змінних (називаються прихованими шарами або прихованими нейронами). Приховані нейрони обчислюються як лінійні комбінації вихідних змінних. Потім лінійна комбінація перетворюється за допомогою нелінійної функції g(*) (ще її називають функцією активації) такою, як логістична функція:

$$h_k(x) = g\left(\beta_{0k} + \sum_{i=1}^{p} x_j\, \beta_{jk}\right) \qquad (3.21)$$

де $g(u) = \frac{1}{1+e^{-u}}$

Коефіцієнти β відповідають коефіцієнтам у регресійній моделі. Зазвичай модель нейронної мережі включає кілька прихованих шарів для обчислення необхідного вихідного параметра.

Після визначення кількості прихованих шарів кожен нейрон повинен бути співвіднесений з вихідним параметром. Лінійна комбінація, яка визначає зв'язки між прихованими шарами та вихідним параметром, має вигляд:

$$f(x) = \gamma_0 + \sum_{k=1}^{H} \gamma_k\, h_k \qquad (3.22)$$

Модель нейронної мережі може розглядатись як нелінійна модель регресії, параметри якої обчислюються на основі критерію мінімізації суми квадратичних помилок. На першому кроці значення параметрів моделі визначаються випадковим чином, які з використанням спеціальних алгоритмів перераховуються кожному кроці з метою мінімізації помилки.

Недоліком моделі нейронних мереж є тенденція до її «перенавчання» (over-fitting) залежності між пояснювальними змінними та вихідним параметром

через велику кількість параметрів регресії. Одним із способів вирішення даної проблеми - взяти як критерій мінімізації функцію виду:

$$\sum_{i=1}^{n}(y_i - f_i(x)^2) + \lambda \sum_{k=1}^{H}\sum_{j=0}^{P}\beta_{jk}^2 + \lambda \sum_{k=0}^{H}\gamma_k^2 \qquad (3.23)$$

Параметр λ називається регулюючим параметром. Чим більший даний параметр, тим більше поведінка моделі набуває згладженого характеру і модель менш схильна до «перенавчання».

РОЗДІЛ 4 ПОБУДОВА РОЗВ'ЯЗАННЯ ЗАДАЧ ПРОГНОЗУВАННЯ

Фондовий індекс S&P500 є одним із відомих індексів США, другий за популярністю (перше місце – індекс Dow Jones) та розраховується на підставі вартості простих акцій 500 американських компаній з найвищим рівнем капіталізації. Індекс S&P 500 є розробкою міжнародної рейтингової компанії Standard&Poor`s.

До індексу входять акції підприємств різних секторів економіки в наступному співвідношенні:

Промисловість – 400 підприємств; фінанси - 40 компаній; комунальна сфера – 40 компаній; транспорт - 20 компаній.

Цінні папери всіх підприємств, що використовуються для розрахунку цього індексу, вільно торгуються на фондових біржах NASDAQ та NYSE.

Згідно з формулою розрахунку індекс S&P500 є середньозваженим за рівнем вільної капіталізації (часткою акцій, що знаходяться у вільному обігу). З формули розрахунку випливає, що чим вище значення індексу, тим вище середня ціна акцій американських компаній і, відповідно, більшим попитом користуються їхні акції, а це означає, що рівень інвестиційної привабливості зростає. Таким чином, фінансовий індекс S&P500 можна розглядати як індикатор зростання (або падіння) економіки США.

$$Index = \frac{\sum_{i=1}^{500} P_i N_i}{Divisor} \qquad (4.1)$$

де P_i – ціна і-ої акції;

N_i – кількість акцій у вільному обігу;

Divisor – фактор, значення якого встановлюється компанією Standard & Poors.

S&P 500 є одним із найбільш відстежуваних фондових індексів акцій, і багато експертів вважають, що він найкраще представляє ситуацію на ринку. Національне бюро економічних досліджень США (National Bureau of Economic

Research) представило звичайні акції як провідний індикатор бізнес-циклів.

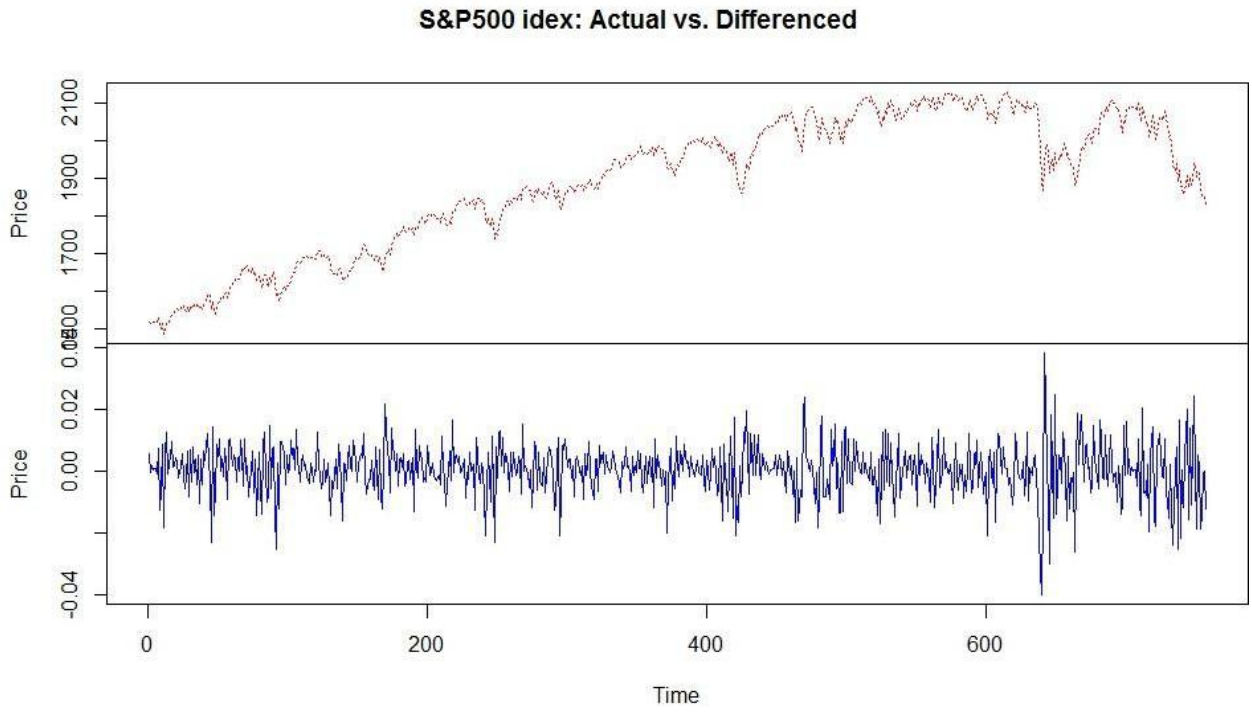


Рисунок 5. Фінансовий індекс S&P500. У верхній половині малюнка представлений графік реальних значень фінансового індексу. У нижній частині малюнка – прибутковість фінансового індексу.

Згідно з представленим графіком на Рисуноку 5 видно, що за період у 600 біржових днів значення індексу зросло на третину, що означає, що зростання капіталізації економіки США становило 30%.

Крім того, індекс S&P 500 також дуже популярний біржовий актив – за допомогою ф'ючерсів їм активно торгують трейдери всього світу. Основні торги проходять на товарній біржі Чикаго. Також існує спеціальний біржовий інвестиційний фонд SPDR S&P 500, що торгується на Нью-Йоркській біржі, який показує один із найвищих ступенів ліквідності серед світових інструментів фондового ринку.

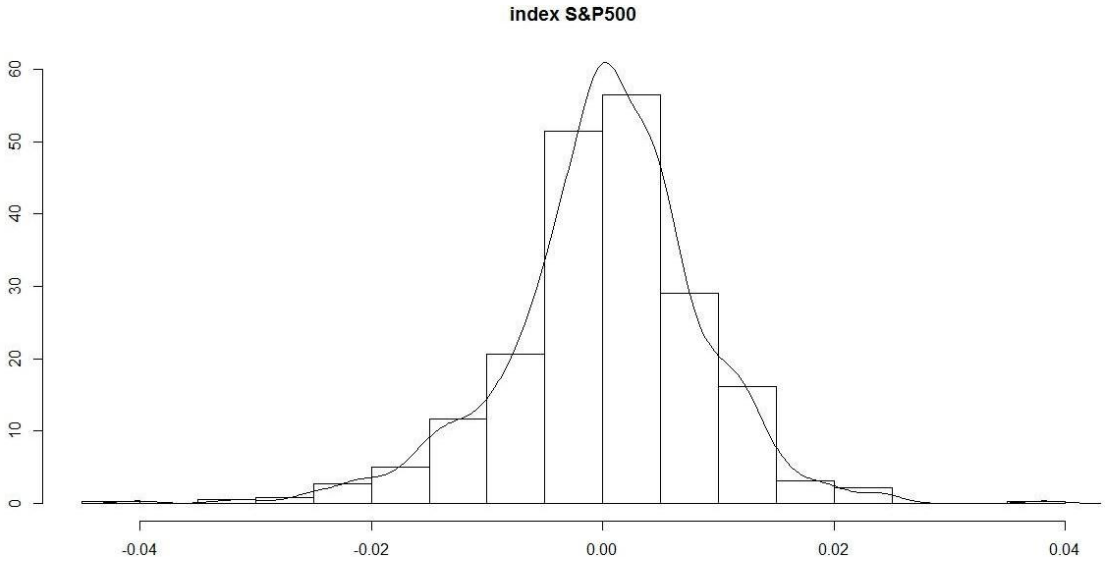


Рисунок 6. Індекс S&P500: розподіл значень у аналізованій вибірці.

Оскільки в моделі як пояснюючі змінні обрані не тільки самі історичні значення індексу, але також і історичні значення цін акцій – компонентів індексу, маємо відносно високу розмірність ознакового простору, що підвищує трудомісткість розв'язання задачі прогнозування індексу. У зв'язку з цим на етапі передобробки вихідних даних було проведено процедуру зниження розмірності даних, критерієм якої мінімізація втрати інформативності даних. Як методи зниження розмірності використовувалися 2 альтернативи: метод головних компонент і кластерний аналіз.

4.1 Аналіз застосування методів зменшення розмірності

Нехай F – ознаковий простір вихідного завдання, яке має вигляд:

$$F = [RS]_{n \times l}, \qquad (4.2)$$

де $R \in F$ – підпростір ознак розмірності 758 х 5, що містить інформацію про минулі значення фінансового індексу S&P500, обсяг торгів та найвищі та найнижчі значення індексу за кожен торговий день.

$S \in F$ – підпростір ознак розмірності 758 х 500, що містить інформацію про

історичні значення цін акцій компаній, на основі яких обчислюється фінансовий індекс S&P500.

У рамках розв'язуваної задачі скорочення розмірності як вихідний простір для зменшення вибрано простір $S \in F$. Далі в ході пошуку оптимального рішення задачі скорочення розмірності простору під вихідним простором розуміється простір компонентів фінансового індексу S.

Метод головних компонент

Як основний спосіб зі скорочення розмірності ознакового простору було розглянуто спосіб головних компонент.

В рамках застосування даного методу було отримано зниження розмірності ознакового простору з 500 до 306 змінних змінних. Даний результат можна оцінити як досить хороший, але для побудови моделі прогнозування часових рядів така кількість змінних може бути надмірною з погляду обчислювальних витрат.

Тому як альтернативний сценарій були розглянуті методи кластерного аналізу.

Кластерний аналіз

У рамках кластерного аналізу розглядалися такі методи: *К – середніх*, *К – медіан*, ядерний *К-середніх* та спектральний аналіз (алгоритм Іна – Джордана-Вейсса).

Перш, ніж проводити порівняльний аналіз отриманих результатів роботи алгоритмів під час кластерного аналізу, необхідно визначити головний параметр аналізу – кількість кластерів. Теоретично цей параметр не може бути занадто малий, оскільки, об'єднуючи в кластери на основі схожості, ми припускаємо, що змінні, які потрапили в кластер, можна замінити на нову змінну з усередненими значеннями. При цьому нам важливо максимально можливо зберегти інформативність вихідних змінних при переході до нових змінних. Усереднення великої кількості вихідних змінних може призвести до спотворення та втрати інформативності.

Разом з тим кількість кластерів не повинна бути надто великою, оскільки

перед нами стоїть завдання зменшення розмірності ознакового простору з метою зниження трудомісткості побудови моделі прогнозування.

Одним із найпоширеніших аналітичних методів визначення параметра кількості кластерів є так званий метод "ліктя" ("elbow" method) у разі використання методу К-середніх.

На Рисуноку 7 представлена графічно залежність усередині кластерної відстані кількості кластера. Можна побачити, що на графіці відзначається різке падіння значень по осі Y (усередині кластерної відстані між елементами) на інтервалі значень параметра кількості кластерів від 1 до 5, при цьому на інтервалі від 6 до 50 падіння відбувається більш монотонно.

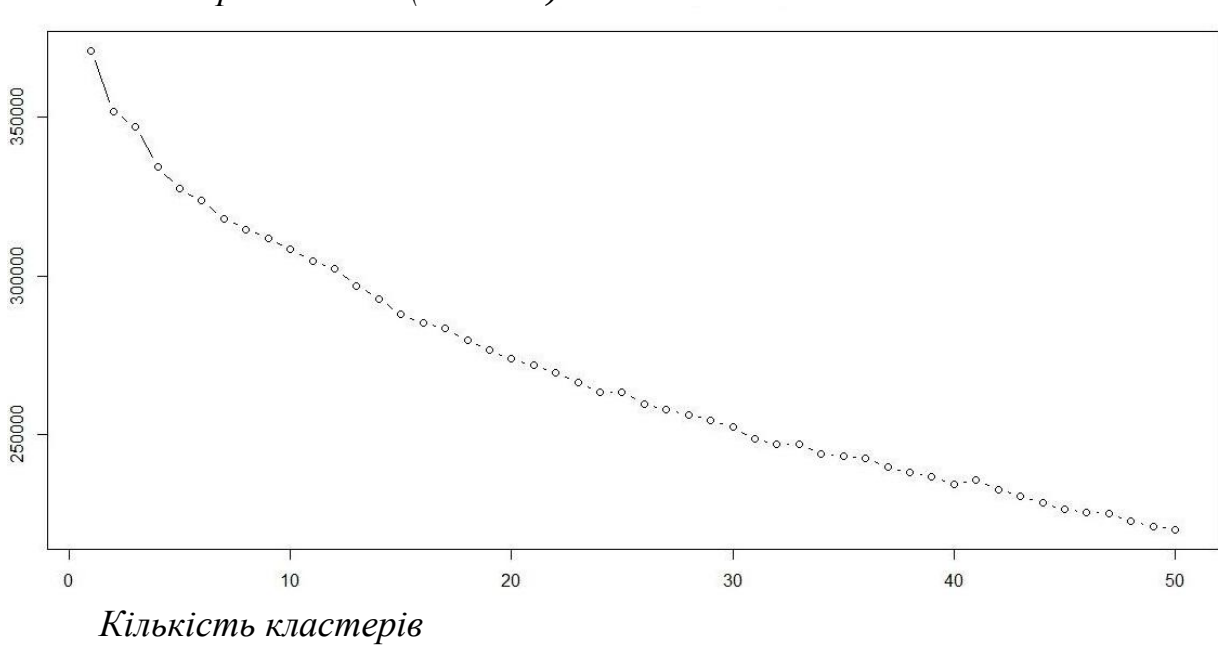


Рисунок 7. Метод локтя для кластеризации финансовых данных.

Разом з тим, аналізуючи графік, зображений на Рисуноку 7, неможливо за допомогою методу ліктя визначити оптимальну кількість кластерів, оскільки характер зменшення монотонний функції без різких стрибків і важко з точністю стверджувати яку кількість кластерів дасть найбільш оптимальний варіант кластеризації.

У зв'язку з цим визначення необхідної кількості кластерів підбір був виконаний вручну. Для різних значень К в інтервалі від 20 до 50 з кроком 5 були

отримані результати з використанням різних методів, визначених раніше, та проведено порівняльний аналіз.

У разі застосування методу К-середніх при значеннях До в інтервалі більше 7 як результат отримуємо сильно нерівномірну кластеризацію і з'являються кластери, кількість елементів яких від 1 до 3 або більше 100. Наявність кластерів з дуже малою кількістю спостережень може вказувати на два факти:

1. кількість кластерів є неоптимальним і надлишковим і, відповідно, необхідно вибрати менше значення;

2. у вибірці є звані викиди (outliers), що означає або як така вибірка є вкрай неоднорідної чи було допущено спотворення даних за її складання. В даному випадку необхідно провести передобробку даних, мінімізувавши можливість спотворення даних.

Крім того, неоднорідність кластеризації у випадку *К – середніх* може бути обумовлена його властивістю, що він не дає оптимального рішення, якщо безліч даних є неопуклою. У зв'язку з цим вирішили відмовитися від методу *К – середніх*.

Метод К-медіан є нечутливим до викидів на відміну від *К-середніх*, але також не дає оптимального рішення, якщо безліч даних не опукло. Від використання методу К-медіан також вирішено було відмовитися через сильну неоднорідність кластеризації.

У свою чергу ядерний метод *К-середніх* та спектральний метод дали набагато кращі результати на відміну від розглянутих раніше методів. Дані методи використовуються у разі, якщо безліч елементів не опукло. Кластеризація більш однорідна, відсутні кластери з кількістю елементів менше 3 або більше 50. У таблиці 4.1 представлені результати кластеризації для різних методів при К=40.

Таблиця 4.1 - Результати кластеризації для різних методів при К=40

Найменування методу	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**….**	**36**	**37**	**38**	**39**	**40**
Ядерний К - середніх (лінійний)	46	13	19	34	25	8	8	9	11	8	….	9	9	7	22	7
Ядерний К - середніх (гаусовий)	33	6	14	16	25	11	11	12	10	22	….	10	13	10	8	9
К - середніх	21	15	4	14	1	1	35	11	2	29	….	91	4	13	32	15
К-медіанних	18	13	3	15	4	1	9	1	1	141	….	7	1	10	76	9
Спектральний метод	4	24	20	8	6	11	23	19	12	7	….	24	16	10	20	7

Параметр К був обраний на основі результатів ядерного методу к-середніх (гаусовий випадок) та спектрального методу, виходячи з наступних критеріїв:

1. Рівномірна кластеризація (більшість кластерів містять від 15 до 30 елементів, кількість малих кластерів мінімальна).
2. Відсутність кластерів, що містять менше 3 елементів.

На основі порівняльного аналізу значення К (кількість кластерів) було обрано рівним 40.

На наступних двох графіках (рисунках. 8-9) представлені гістограми розподілу змінної – кількість елементів у кластері. На першому графіку видно, що більшість кластерів мають кількість елементів в межах від 5 до 15 елементів і зовсім небагато від 15 до 25 один кластер має кількість елементів більше 30.

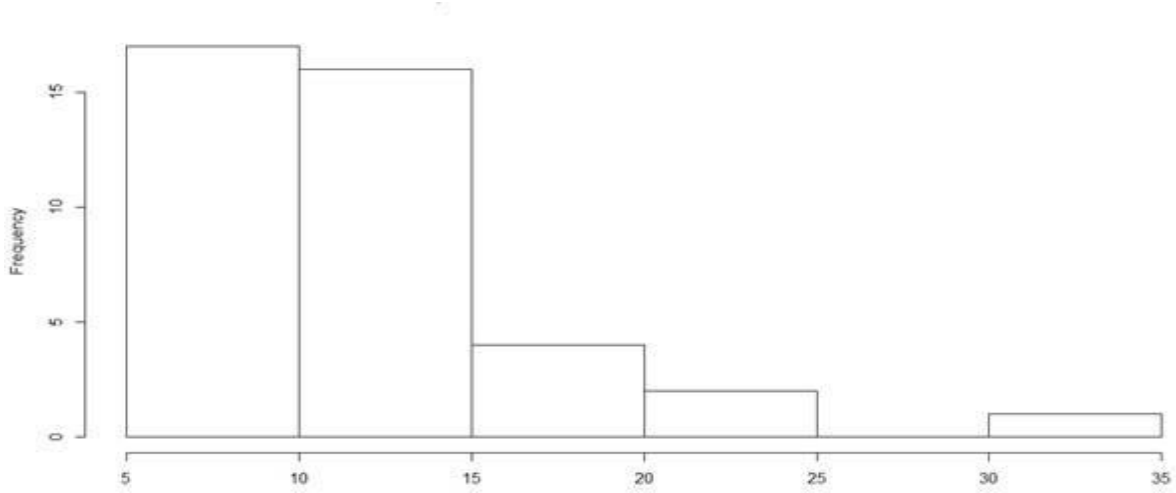


Рисунок 8. Гістограма частот за кількістю елементів у кластері (Ядерний К – середніх).

На другому графіку співвідношення інше: більшість кластерів мають кількість елементів від 10 до 20, 4 кластери - більше 20 елементів, 7 кластерів містять менше 5 елементів, але більше 3 (умова 2 виконується).

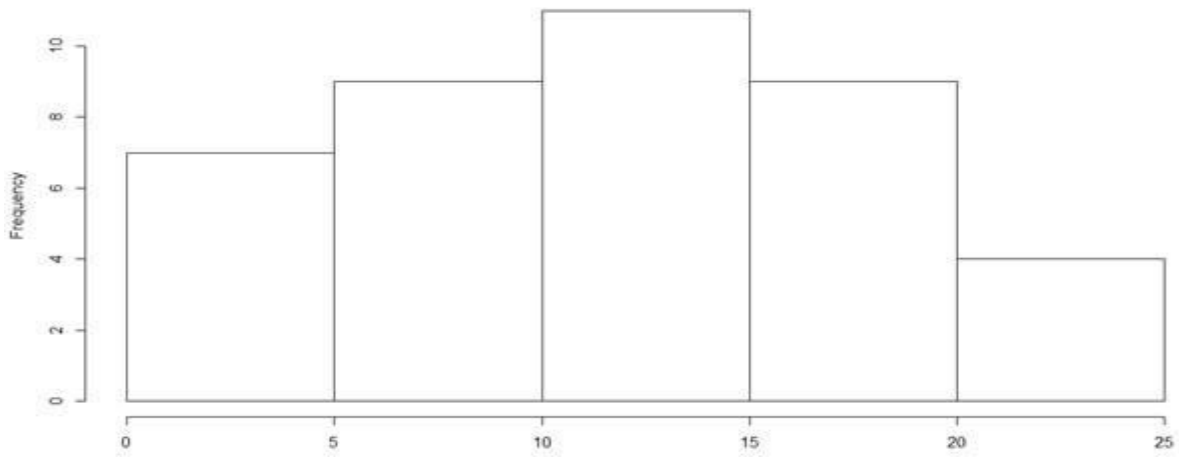


Рисунок 9. Гістограма частот за кількістю елементів у кластері (спектральний метод).

На основі отриманих результатів як оптимальне рішення було обрано рішення, отримане за допомогою спектрального методу.

На наступному кроці для кожного кластера було проведено наступну трансформацію.

Нехай S_k – матриця змінних розміром n x m, де кожен рядок є вектором значень змінної, що потрапила в k – ий кластер, а кількість рядків дорівнює кількості елементів у j-му кластері. Іншими словами, матриця S_k є підпростір вихідного простору ознак S. Матриця має вигляд:

$$S_k = \begin{bmatrix} S_{11}^k & \cdots & S_{1m}^k \\ \vdots & \ddots & \vdots \\ S_{n1}^k & \cdots & S_{nm}^k \end{bmatrix} \quad (4.3)$$

Для кожного i - го стовпця матриці Sk обчислюємо наступне значення:

$$T_i^k = \frac{1}{n}\sum_{j=1}^{n} S_{ji}^k \quad (4.4)$$

В результаті отримуємо наступний вектор:

$$T^k = \begin{pmatrix} T_1^k \\ \vdots \\ T_m^k \end{pmatrix} \quad (4.5)$$

Вектор T^k є вектором, кожен елемент якого є усередненим значенням шпальт вихідної матриці Sk. Тоді:

$$T = \begin{pmatrix} T_1^1 & \cdots & T_1^k \\ \vdots & \ddots & \vdots \\ T_m^1 & \cdots & T_m^k \end{pmatrix} \quad (4.6)$$

Для прикладу на наступних двох рисунках (Рисунок 10-11) представлений результат усереднення елементів для 2 із 40 отриманих кластерів. Змінна, отримана за рахунок усереднення, зображена червоною лінією.

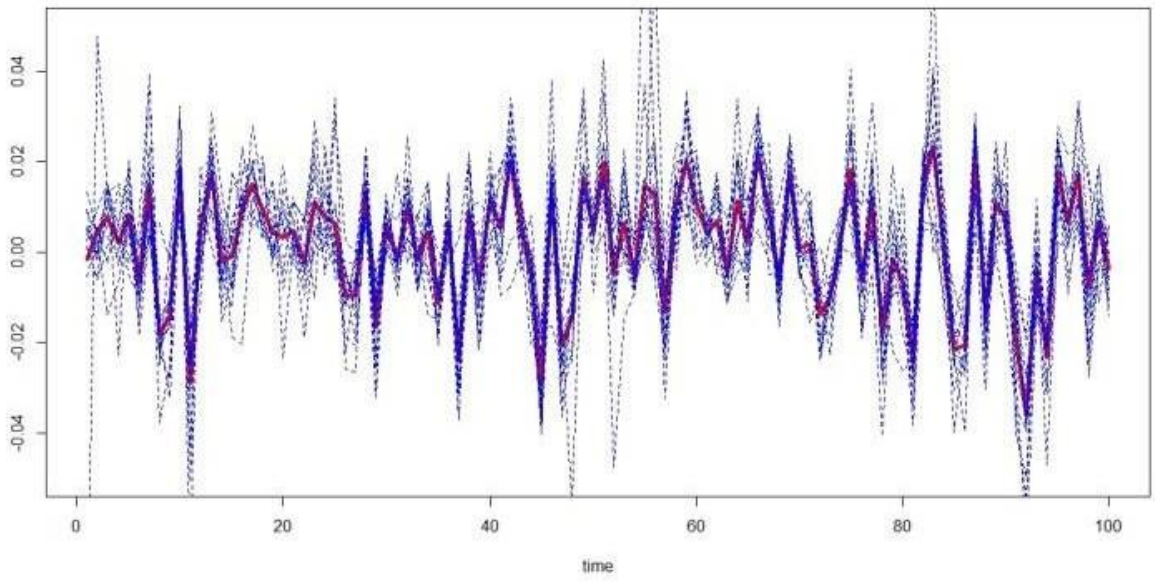


Рисунок 10. Усереднення всередині кластера.

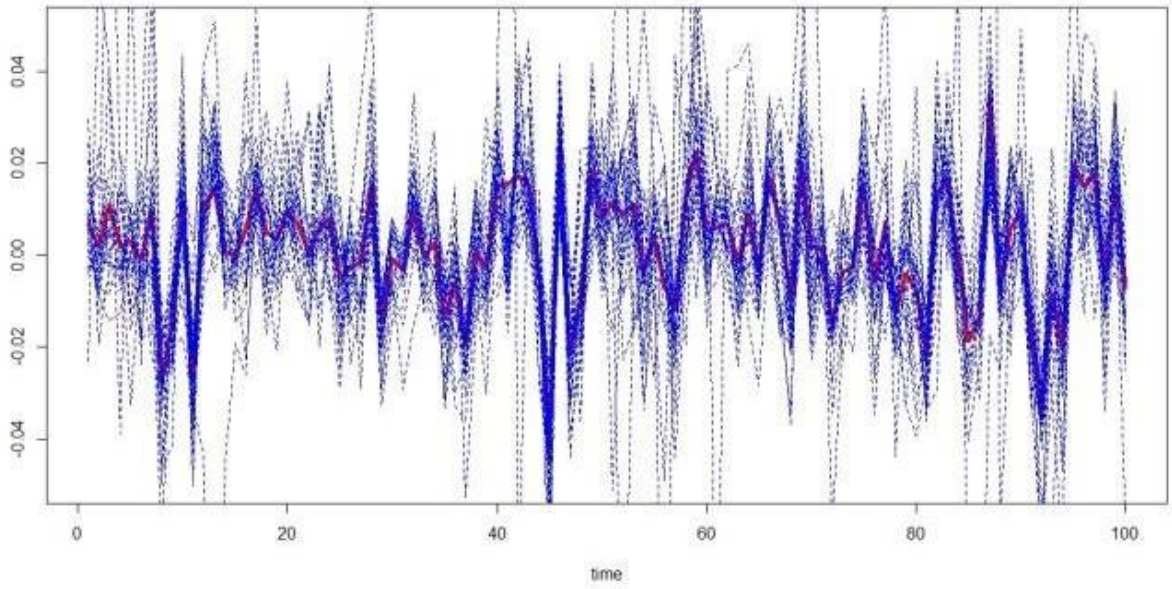


Рисунок 11. Усереднення всередині кластера.

Аналізуючи Рисунок 10-11, бачимо, що нова змінна відбиває тренди зміни значень змінних, цін акцій, які у кластер. Зокрема, на Рисунок 11, видно, що у моменти падіння цін акцій компаній, у поступовій динаміці нової змінної також відзначається стрибок вниз – падіння значень. Аналогічна ситуація у моменти зростання. Отже, можна дійти невтішного висновку, що нова змінна загалом відбиває напрям руху змінних кластера.

Матриця T є новим ознаковим простором, що відповідає вихідному простору S. Таким чином, в ході кластерного аналізу вихідний ознаковий простір розмірністю 758x500 було трансформовано у простір розмірністю

758x40. Якість мінімізації ознакового простору буде оцінена за результатами прогнозування фінансових тимчасових лав.

4.2 Аналіз результатів прогнозування часового ряду

Для побудови моделі прогнозування фінансових часових рядів були використані такі методи: регресійна модель, метод опорних векторів та нейронна мережа.

При побудові прогнозу часового ряду в якості базової регресійної моделі було взято таку модель:

$$SP_t = \beta_0 + \sum_{j=1}^{7} \beta_j SP_{t-j} + \sum_{k=1}^{7} \gamma_k S_{t-k} + a_t \qquad (4.7)$$

Ключовим параметром для оцінки отриманих результатів взято коефіцієнт детермінації R^2. Коефіцієнт детермінації характеризує точність досліджуваної регресійної моделі і є мірою її адекватності.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \qquad (4.8)$$

Коефіцієнт детермінації для моделі з константою приймає значення від 0 до 1. Чим ближче значення коефіцієнта до 1, тим сильніша залежність. Для прийнятних моделей передбачається, що коефіцієнт детермінації повинен бути хоча б не менше 50% (у цьому випадку коефіцієнт множинної кореляції перевищує модуля 70%). Рівність коефіцієнта детермінації одиниці означає, що змінна, що пояснюється, в точності описується аналізованою моделлю.

Результати даних моделей представлені в таблиці 4.2.

Таблиця 4.2 - Результати даних моделей

Модель	RMSE	RSE	R^2	Adj. R^2
Linear regression	0.00672	0.00025	0.42278	0.42690
SVM	0.01102	0.00166	0.73846	0.72461
Neural Net	0.01938	0.01248	0.50657	0.50543

З таблиці видно, що найкращий результат дає модель, побудована з допомогою методу опорних векторів (коефіцієнт детермінації дорівнює 0.73). Найгірший результат дала лінійна модель (коефіцієнт детермінації дорівнює 0.42).

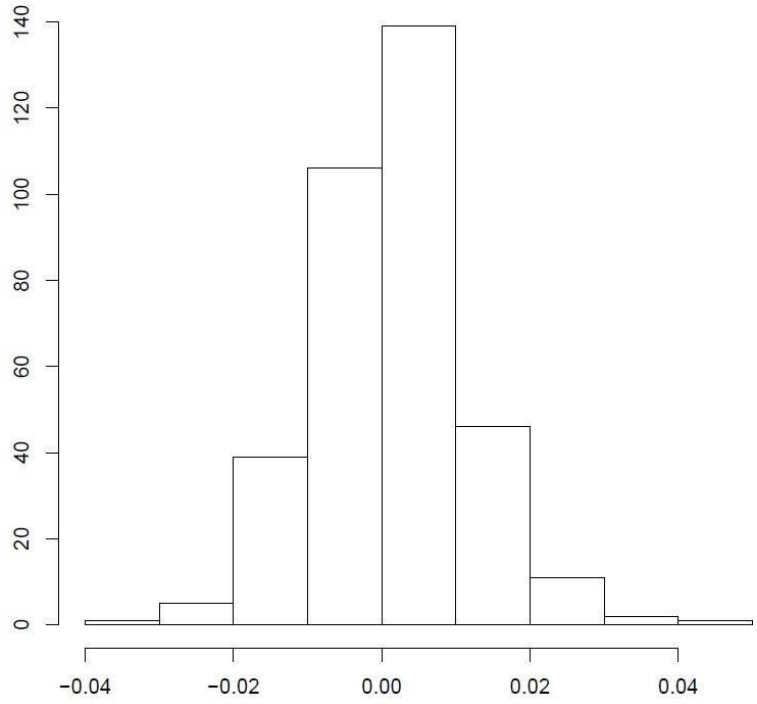


Рисунок 12. Розподіл помилки моделі прогнозування (випадок – метод опорних векторів).

Таким чином, на основі даних, отриманих в результаті трансформації ознакового простору, було побудовано моделі прогнозування фінансового індексу S&P500, значення помилок яких досить малі. Крім того, модель, побудована з використанням методу опорних векторів, з точки зору оцінки точності дає найоптимальніші результати прогнозу У зв'язку з цим, можна дійти невтішного висновку, що реалізована процедура скорочення розмірності простору зберігає інформативність, що у вихідних ознаках.

ВИСНОВКИ

У межах цієї роботи було здійснено аналіз існуючих методів скорочення розмірності ознакового простору, і навіть проведено моделювання реальних даних. При моделюванні були використані різні методи кластерного аналізу, серед яких *К-середніх, К-медіан*, ядерний *К-середніх*, спектральний метод. Ці методи незалежно один від одного призвели до достатнього скорочення розмірності ознакового простору. У свою чергу, один з розглянутих методів може бути визнаний ефективним, якщо скорочення розмірності призвело до мінімальної втрати інформативності вихідних даних.

Як вимір рівня втрати інформативності даних може виступати середньоквадратична помилка математичної моделі прогнозування часових рядів і значення коефіцієнта детермінації. Якщо середньоквадратична помилка моделі значно зростає після застосування процедури скорочення розмірності ознакового простору, це означає, що метод не може бути визнаний ефективним і варто або скоригувати параметри, або використовувати інший метод.

На основі даних, отриманих в результаті трансформації ознакового простору, було побудовано три моделі прогнозування (лінійна модель регресії, метод опорних векторів та нейронна мережа). Отримані результати прогнозування дозволяють стверджувати, що проведена трансформація ознакового простору не призводить до критичного спотворення даних, яке б дозволяло робити адекватний прогноз. Крім того, за допомогою методів кластерного аналізу процедура скорочення ознакового простору дозволила досягти більш високого рівня стиснення даних, ніж метод головних компонентів.

ПЕРЕЛІК ПОСИЛАНЬ

1. Jiawei Han, Micheline Kamber, Jian Pei.Data mining : concepts and techniques – 3rd ed. p. cm.

2. Hastie Trevor, Tibshirani Robert, Friedman Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. – 2nd Edition.

3. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning with Applications in R . - Springer Science+Business Media New York 2013.

4. Max Kuhn, Kjell Johnson. Applied Predictive Modeling. - Springer Science+Business Media New York 2013.

5. A.T. Chen and M.T. Leung, Regression neural network for error correction in foreign exchange forecasting and trading, Computers Operations Research,31(7)1049{1068, 2004}

6. S. Ghoshal, S. Roberts, Extracting Predictive Information from Heterogeneous Data Streams using Gaussian Processes, e-print arXiv:1603.06202v1, 2016

7. Y. Wang and R. Gu, Analysis of efficiency for Shenzhen stock market based on multifractal detrended fluctuation analysis, International Review of Financial Analysis, 18, 271{276,2009}

8. K.R. Murphy and B. Myors, Statistical power analysis: A simple and general model for traditional and modern hypothesis tests, Mahwah, NJ: Erlbaum, 1998

9. R. Breatley and S. Meyers, Principles of corporate finance, Irwin, McGraw{Hill, 2000}

10. R.J. Shiller, Irrational Exuberance Princeton: Princeton University press, 2000.

11. T.O. Sprenger, A. Tumasjan, P.G. Sandner, and I.M. Welpe, Tweets and trades: the information content of stock microblogs, European Financial Management, 20(5), 926{957, 2014}.

Додаток А

Структура фінансового індексу S&P500

Name	Close price	Min/max	Name	Close price	Min/max
3M	200,67	199,05	Anadarko Petroleum	51,84	50,7
Abbott Laboratories	44,71	43,83	Analog Devices	82,81	81,94
AbbVie	66,06	65,86	Anthem	183,08	180,6
Accenture	122,89	122,53	Aon	130,01	129,89
Activision Blizzard	58,28	58,02	Apache	48,19	47,34
Acuity Brands	164,87	164,04	Apple	153,61	153,31
Adobe	141,89	141,68	Applied Materials	45,5	44,75
Advance Auto Parts	134,16	131,44	Arthur J Gallagher &	57,12	57,11
AES	11,79	11,68	S&P Global	140,21	139,95
Aetna	145,57	145,25	Apartment Investment & Management	43,18	43,07
Affiliated Managers Group	153,92	153,2	Archer Daniels Midland	42,45	42,18
Aflac	74,84	74,45	Assurant	99,28	98,95
Agilent Technologies	59,51	59,12	AT&T	38,12	38,01
Air Products and Chemicals	143,91	143,58	Autodesk	113,03	112,87
Akamai	47,39	47,14	Automatic Data Processing	101,35	100,99
Alaska Air Group	86,62	85,86	AutoNation	39,27	38,2
Albemarle	112,86	112,24	AutoZone	615,62	605
Alexion Pharmaceuticals	97,7	96,18	AvalonBay Communities	191,32	190,73
Allegion	79,38	79,11	Avery Dennison	84,29	83,55
Allergan	223,12	222,97	Baker Hughes	55,45	54,76
Alliance Data Systems	244,04	241,39	Ball	40,69	40,54
Alliant Energy	41,36	41,23	Bank of America	23,24	23,17
Allstate	86,5	85,98	Bank of New York Mellon	47,41	47,11
Alphabet A (ex Google)	993,27	987,36	Baxter International	58,64	58,46
Alphabet C (ex Google)	971,47	965,07	BB&T	42,56	42,46
Altria	74,41	73,9	Becton, Dickinson &	186,48	185,65
Amazoncom	995,78	989,33	Bed Bath & Beyond	34,85	34,42
Ameren	56,15	55,89	Berkshire Hathaway	165,69	165,02
American Airlines	48,74	47,75	Best Buy	58,97	58,89
American Electric Power	70,87	70,71	Biogen	250,8	250,52
American Express	77,46	77,05	BlackRock	407,02	403,12
American International Group (AIG)	63,55	62,95	Boeing	186,59	186,36

Name	Close price	Min/max	Name	Close price	Min/max
American Tower	131,11	130,49	BorgWarner	40,99	40,4
American Water Works	77,53	77,34	Boston Properties	121,35	120,79
Ameriprise Financial	122,3	122,03	Boston Scientific	27,32	27,19
AmerisourceBergen	91	90,91	Bristol-Myers Squibb	53,97	53,9
Ametek	60,87	60,87	Broadcom	241,21	238,38
Amgen	155,01	154,79	Brown-Forman b	51,73	51,36
Amphenol	74,8	74,33	CH Robinson Worldwide	67,04	66,97
CA	31,82	31,81	Comerica	69,22	68,91
Cabot Oil & Gas	22,96	22,69	ConAgra Foods	39,03	38,76
Campbell Soup	58,96	58,63	Concho Resources	131,28	128,77
Capital One Financial	79,79	79,09	ConocoPhillips	45,35	44,69
Cardinal Health	73,1	72,95	Consolidated Edison	82,1	81,86
CarMax	64,72	64,05	Constellation Brand a	180,93	179,6
Carnival	63,36	62,84	Corning	29,3	29,28
Caterpillar	105,66	104,65	Costco Wholesale	177,86	177,62
CBRE Grou a	34,39	34,3	Crown Castle	102,81	101,95
CBS	61,57	61,21	CSRA	30,46	30,39
Celgene	116,78	116,35	CSX	53,96	53,48
Centene	74,57	74,33	Cummins	156,64	155,73
CenterPoint Energy	28,33	28,16	CVS Health	76,64	75,94
CenturyLink	24,8	24,76	DR Horton	33,11	33,02
Cerner	64,95	64,73	Danaher	84,4	84,05
CF Industries	28,13	27,63	Darden Restaurants	87,95	87,66
Charles Schwab	39,52	39,12	DaVita HealthCare Partners	64,75	64,31
CR Bard	308,7	307,97	Deere &	122,79	122,37
Chesapeake Energy	5,29	5,15	Delphi Automotive	87,5	86,64
Chevron	104,72	104,61	Delta Air Lines	50,8	50,4
Chipotle Mexican Grill	480,15	478,95	DENTSPLY SIRONA	62,72	62,59
Chubb	142,24	142,24	Devon Energy	35,82	35,22
Church & Dwight	51,1	50,97	Digital Realty Trust	117,81	116,98
Cigna	161,32	159,93	Discover Financial Services	59,8	58,88
Cimarex Energy	110,95	110,07	Discovery Communication a	25,87	25,55
Cincinnati Financial	69,91	69,85	Discovery Communications	25,25	24,98
Cintas	125,16	124,97	Dollar General	72,32	71,71
Cisco	31,5	31,36	Dollar Tree	78,28	77,71
Citigroup	62,07	61,65	Dominion Resources	81,03	80,76
Citizens Financial Group	34,64	34,34	Dover	82,5	81,66

Name	Close price	Min/max	Name	Close price	Min/max
Citrix Systems	82,49	82,26	Dow Chemical	61,06	60,45
Clorox	134,99	134,48	Dr Pepper Snapple Group	93,45	93,23
CME Grou a	117,98	117,58	DTE Energy	108,88	108,49
CMS Energy	47,41	47,24	Duke Energy	85,07	84,96
Coach	46,33	46,04	DuPont (E I DuPont de Nemours and	77,85	77,27
Coca-Cola	45,39	45,34	E*TRADE FINANCIAL	35,08	34,86
Cognizant	66,79	66,7	Eastman Chemical Company	79,91	79,41
Colgate-Palmolive	75,45	75,14	Eaton	77,56	77,14
Comcast	40,91	40,37	eBay	34,9	34,66
Ecolab	131,3	130,76	Gap	22,42	21,86
Edison International	79,96	79,77	Garmin	52,4	52,2
Edwards Lifesciences	114,25	114	General Dynamics	202,24	200,94
Electronic Arts	112,13	111,79	General Electric	27,45	27,29
Eli Lilly and	78,05	77,72	General Mills	57,32	56,78
Emerson Electric	58,92	58,58	General Motors	33,07	32,35
Entergy	77,65	77,53	Genuine Parts	93,28	91,41
EOG Resources	91,22	90,14	Gilead Sciences	64,5	64,32
EQT	56,69	55,97	Global Payments	91,85	91,76
Equifax	136,43	136,06	Goldman Sachs	223,53	221,18
Equinix	441,22	440,1	Goodyear Tire & Rubber	32,6	31,94
Equity Residential	65,17	64,85	H&R Block	26,11	25,92
Essex Property Trust	256,3	255,15	Halliburton	45,76	45,26
Estée Lauder Companies	93,7	93,35	Hanesbrands	20,7	20,35
Exelon	35,82	35,74	Harley-Davidson	52,32	51,68
Expedia	144,58	143,58	Harris	110,56	110,17
Expeditors International of Washington	52,95	52,95	Hartford Financial Services Group	49,23	49,03
Express Scripts	59,76	59,58	Hasbro	104,21	103,72
Extra Space Storage	78	77,34	HCA	82,7	81,97
ExxonMobil	81,55	81,41	HCP	31,23	31,04
F5 Networks	127,16	126,4	Helmerich & Payne	53,67	53,63
Facebook	152,13	151,16	Henry Schein	182,68	182,09
Fastenal	43,47	43,07	Hess	48,14	47,08
Federal Realty Investment Trust	126,32	125,53	Hewlett Packard Enterprise	18,83	18,74
FedEx	194,26	193,64	Hewlett-Packard (HP)	18,47	18,22
Fidelity National Information Services	85,28	85,05	Hologic	43,35	42,86
Fifth Third Bancorp	24,26	24,24	Home Depot	154,9	154,64

Name	Close price	Min/max	Name	Close price	Min/max
FirstEnergy	28,92	28,66	Honeywell International	133,25	132,84
Fiserv	124,09	123,88	Hormel Foods	33,12	33,07
FLIR Systems	37,05	36,69	Host Hotels & Resorts	18,07	18,03
Flowserve	48,1	47,51	Humana	232,25	231,26
Fluor	44,98	44,94	Huntington Bancshares	12,86	12,8
FMC	75,25	75,19	IBM	152,49	152,08
Northeast Utilities (Doing business as Eversource Energy)	61,4	61,24	Illinois Tool Works	140,37	140,06
Foot Locker	59,82	59,07	Illumina	175,91	173,85
Ford Motor	10,93	10,81	Ingersoll-Rand	89,35	89,08
Fortune Brands Home & Security	62,85	62,76	Intel	36,26	36,14
Franklin Resources	41,7	41,51	IntercontinentalExchange Group	60,35	60,28
Freeport-McMoRan	11,67	11,52	Interpublic Group of Cos	24,69	24,6
The Hershey	115,96	114,96	Mallinckrodt	43,28	41,1
International Flavors & Fragrances	138,44	138,37	Marathon Oil	13,52	13,3
International Paper	52,42	51,86	Marathon Petroleum	52,82	52,24
Intuit	138,56	136,83	Marriott	106,9	106,64
Intuitive Surgical	910,36	908,25	Marsh & McLennan Cos	76,37	76,02
Invesco	32	31,64	Martin Marietta Materials	227,76	226,8
Iron Mountain	35,17	34,99	Masco	37	36,9
J M Smucker	128,17	128,01	MasterCard	121,63	121,33
JB Hunt Transportation Services	84,54	84,32	Mattel	22,53	22,02
Jacobs Engineering Group	52,66	52,4	McCormick &	104,48	104,01
Johnson & Johnson	126,92	126,84	McDonalds	149,86	149,25
Johnson Controls International	42,03	41,94	McKesson	161,2	161,13
JPMorgan Chase &	85,36	85,04	Mead Johnson	89,29	89,14
			Nutrition		
Juniper Networks	29,32	29,29	Medtronic	85,09	84,6
Kansas City Southern	95,95	92,75	Merck	64,92	64,82
Kellogg	72,62	72,29	MetLife	51,05	50,96
KeyCorp	17,99	17,92	Michael Kors	36,76	35,8
Kimberly-Clark	129,51	129,11	Microchip Technology	81,7	81,07
Kimco Realty	18,03	17,93	Micron Technology	29,76	28,88

Name	Close price	Min/max	Name	Close price	Min/max
Kinder Morgan	19,25	19,2	Microsoft	69,96	69,52
KLA-Tencor	104,74	103,91	Mohawk Industries	235,89	235,18
Kohl`s	38,73	38,4	Molson Coors Brewing Company (MCBC)	95,78	93,86
Kroger	29,45	29,25	Mondelez	46,44	46,41
L Brands	50,51	49,62	Monsanto	116,78	116,55
Laboratory	140,1	139,99	Monster Beverage	50,9	50,7
Lam Research	155,12	153,1	Moodys	117,11	116,25
Leggett & Platt	52,08	51,91	Morgan Stanley	42,84	42,45
Lennar	51,53	51,36	Motorola Solutions	81,86	81,67
Leucadia National	24,62	24,55	Murphy Oil	25,31	24,94
Level 3 Communications	59,47	59,43	Mylan	39,72	39,66
Lincoln National	64,95	64,59	Nasdaq	67,58	67,55
LKQ	31,55	31,33	National-Oilwell Varco	32,77	32,5
Lockheed Martin	283,65	281,66	Navient	14,48	14,27
Loews	46,88	46,6	NetApp	40,24	40,1
Lowes Companies	80,91	80,57	Netflix	162,43	161,12
Lyondellbasell Industries	80,73	80,03	Newell Brands	53,19	52,79
M&T Bank	160,2	159,99	Newfield Exploration	33,08	32,4
Macerich	58,77	58,44	Newmont Mining	34,11	33,63
Macys	23,44	23,13	News b	13,8	13,6
News	13,42	13,17	Pricelinecom	1 863,90	1 854,02
NextEra Energy	140,71	140,42	Principal Financial Group	62,75	62,48
Nielsen	38,54	37,95	Procter & Gamble	87,25	86,84
Nike	52,59	52,24	Progressive	42,3	42,15
Nisource	25,56	25,45	Prologis	55,79	55,44
Noble Energy	29,45	29,33	Prudential Financial	105,07	104,92
Nordstrom	42,47	41,64	Public Service Enterprise Group	44,43	44,09
Norfolk Southern	122,33	119,2	Public Storage	216,39	215,92
Northern Trust	88,69	88,22	PulteGroup	22,84	22,8
Northrop Grumman	256,99	256,33	PVH	105,29	105,28
NRG Energy	16,47	16,28	Qorvo	77,84	77,76
Nucor	58,15	57,97	QUALCOMM	57,52	57,37
NVIDIA	141,84	137,12	Quanta Services	31,07	30,82
O Reilly Automotive	248,48	243,74	Quest Diagnostics	107,98	107,77
Occidental Petroleum	61	60,71	Ralph Lauren a	66,11	66,06
Omnicom Group	83,26	83,13	Range Resources	24,2	23,65
ONEOK	51,69	51,13	Raytheon	163,3	162,72

Name	Close price	Min/max	Name	Close price	Min/max
Oracle	45,26	45,06	Realty	55,18	54,99
Paccar	62,63	62,48	Red Hat	88,58	87,71
The Mosaic	23,16	22,78	Regeneron Pharmaceuticals	455,17	453,46
Parker Hannifin	159,26	158,55	Regions Financial	14,13	14,12
Patterson Companies	43,94	43,91	Republic Services	63,07	62,71
Paychex	59,28	58,96	Reynolds American	66,98	66,64
PayPal	51,19	50,93	Robert Half	45,9	45,68
Pentair	66,25	65,92	Rockwell Automation	159,28	158,34
Peoples United Financial	16,87	16,75	Rockwell Collins	107,3	106,86
PepsiCo	117,91	117,2	Roper Industries	227,08	225,97
PerkinElmer	62,8	62,76	Ross Stores	63,52	62,79
Perrigo Company	69,95	69,85	Ryder System	65,81	65,28
Pfizer	32,14	32,11	21st Century Fox (A)	27,03	26,96
PG&E	67,37	67,27	Royal Caribbean Cruises	111,08	110
Philip Morris	119,83	118,97	Salesforce	90,83	90,46
Phillips 66	77,44	77,17	SCANA	67,86	67,72
Pinnacle West Capital	87,4	87,27	Schlumberger	70,09	69
Pioneer Natural Resources	170,24	168,12	Scripps Networks Interactiv a	66,7	66,4
PNC Financial Services Group	121,32	120,64	Seagate	42,53	42,2
PPG Industries	107,24	106,72	Sealed Air	44,55	44,2
PPL	39,86	39,8	Sempra Energy	114,48	114,22
Praxair	131,97	131,39	Sherwin-Williams	334,95	333,52
Signet Jewelers	49,31	48,12	21st Century Fox (B)	26,95	26,9
Simon Property Group	157,26	157,01	Grainger	173,58	171,89
Skyworks Solutions	105,89	104,58	Tyson Foods	58,34	57,95
SL Green Realty	101,84	101,48	US Bancorp	51,44	51,39
Snap-On	160,71	158,51	UDR	38,33	38,28
Southern	50,45	50,35	Ulta Salon Cosmetics & Fragrance	302,4	301,44
Southwest Airlines	60,67	60,5	Under Armour	18,07	17,86
Stanley Black & Decker	136,79	136	Under Armour	19,69	19,31
Staples	8,98	8,96	Union Pacific	110,57	108,55
Starbucks	63,3	62,99	United Continental	81,25	80,74
State Street	81,99	81,65	United Parcel Service	105,87	105,54
Stericycle	82,4	82,34	United Rentals	111,17	109,95
Stryker	141,27	140,85	United Technologies	121,85	121,59
SunTrust Banks	54,32	54,14	UnitedHealth	177,5	177

Name	Close price	Min/max	Name	Close price	Min/max
Symantec	29,61	29,51	Universal Health Services	116,07	115,44
Synchrony Financial	27,37	27,08	Unum Group	45,05	44,97
Sysco	54,43	54,17	VF	53,71	53,22
T Rowe Price Group	69,7	69,39	Valero Energy	62,45	62,13
Target	54,4	53,93	Varian Medical Systems	97,87	97,48
TE Connectivity	78,01	77,68	Ventas	66,58	66,27
TechnipFMC	29,07	28,8	VeriSign	90,14	89,96
TEGNA	24,07	23,73	Verisk Analytic a	80,55	80,52
Teradata	28,6	28,57	Verizon	45,32	45,25
Tesoro	83,46	83,09	Vertex Pharmaceuticals	119,92	118,1
Texas Instruments	81,12	80,35	Viacom	35,13	34,4
Textron	48,16	47,95	Visa	94,67	94,4
The Kraft Heinz Company	92,99	92,52	Vornado Realty Trust	92,52	92,04
The Western Union Company	19,07	19,05	Vulcan Materials	126,43	125,26
Thermo Fisher Scientific	173,42	172,89	Walgreens Boots Alliance	81,25	80,65
Tiffany &	86,18	85,15	Walmart	78,13	77,79
Time Warner	99,07	98,7	Walt Disney	108,41	107,5
TJX Cos	75,53	75,24	Waste Management	71,84	71,32
Torchmark	75,23	75,19	Waters	178,07	177,76
Total System Services	59,58	59,54	WEC Energy Group	62,45	62,4
Tractor Supply	54,88	54,02	Wells Fargo &	52,41	52,37
TransDigm Group	264,41	262,8	Welltower	72,53	72,37
Transocean	9,56	9,45	Western Digital	72,53	73,4
Travelers	123,66	123,39	WestRock	55,15	54,97
TripAdvisor	39,07	38,33	Weyerhaeuser	33,01	32,93
Whirlpool	179,13	178,89			
Whole Foods Market	35,12	35,03			
Williams Companies	29,69	29,41			
Willis Towers Watson	145	144,03			
Wyndham Worldwide	99,63	98,95			
Wynn Resorts	125,73	124,45			
Xcel Energy	47,3	47,2			
Xerox	7	6,85			
Xilinx	65,85	64,99			
XL Group	43,3	43,29			
Xylem	51,41	51,35			

Додаток Б

Результати трансформації ознакового простору на основі проведеного кластерного аналізу незалежних змінних, що входять до моделі прогнозування

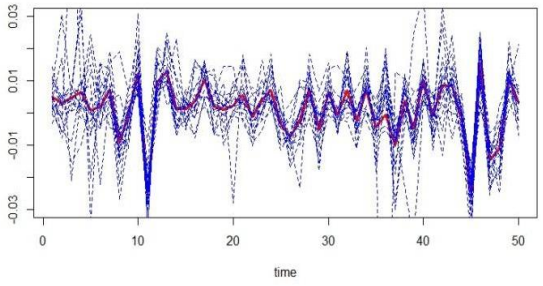


Рисунок 1. 1-ий кластер

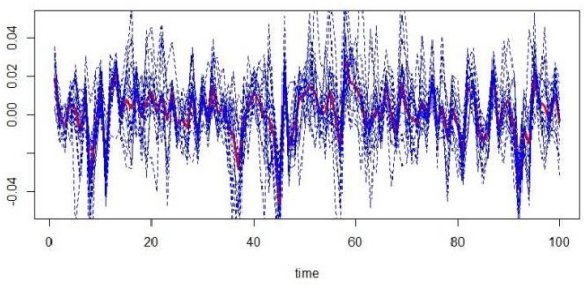


Рисунок2. 2-ий кластер

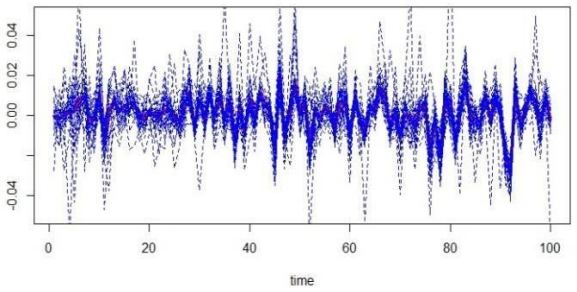


Рисунок 3. 3-ий кластер

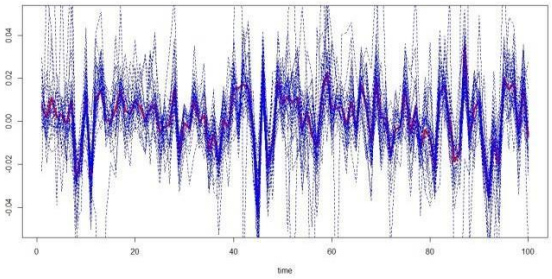


Рисунок4. 4-ий кластер

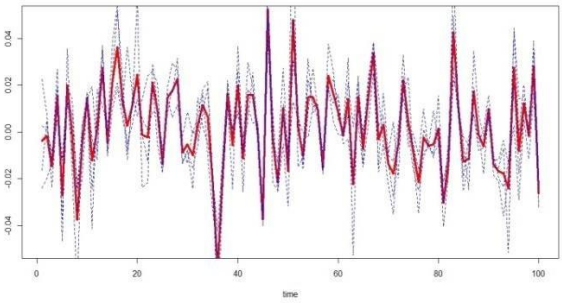


Рисунок 5. 5-ий кластер

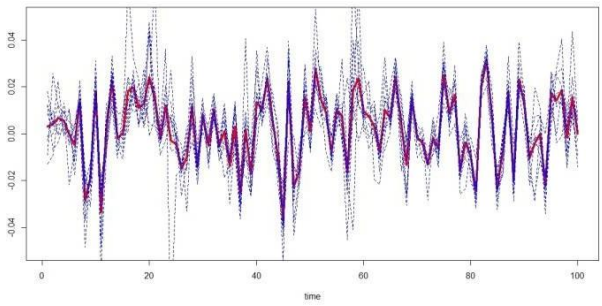


Рисунок6. 6-ий кластер

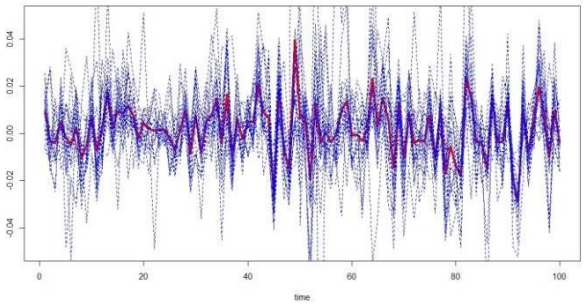


Рисунок 7. 7-ий кластер

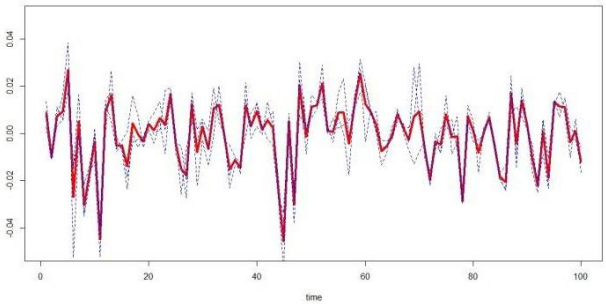


Рисунок8. 8-ий кластер

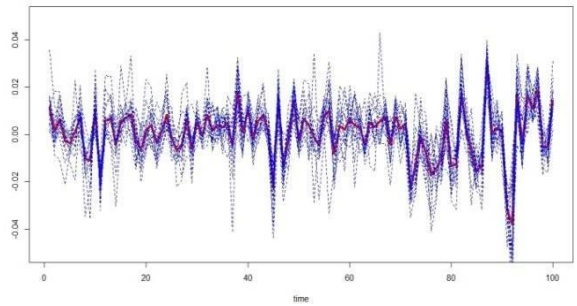


Рисунок 9. 9-ий кластер

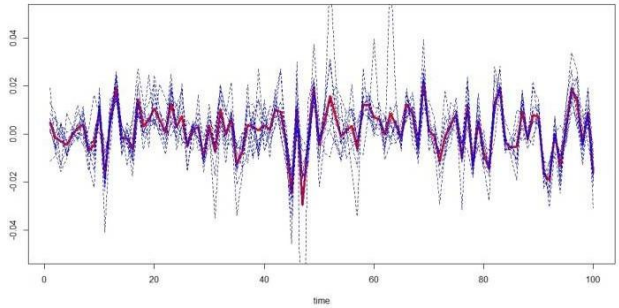


Рисунок10. 10-ий кластер

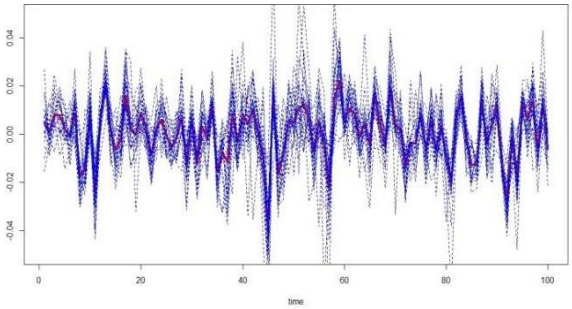


Рисунок 11. 11-ий кластер

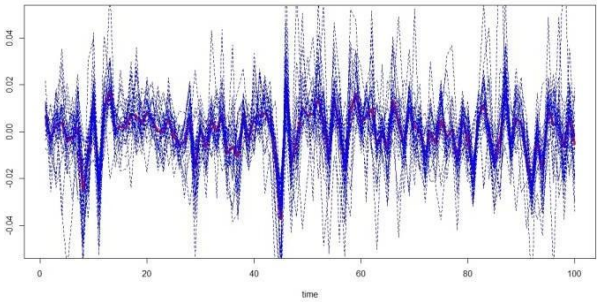


Рисунок12. 12-ий кластер

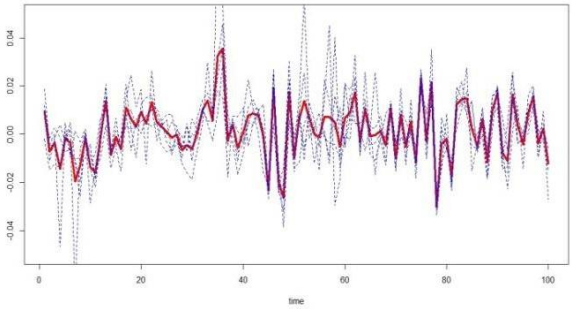


Рисунок 13. 13-ий кластер

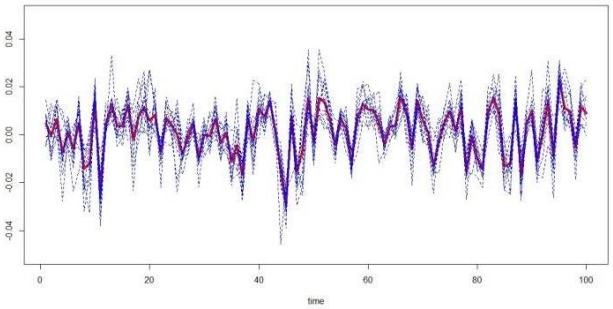


Рисунок14. 14-ий кластер

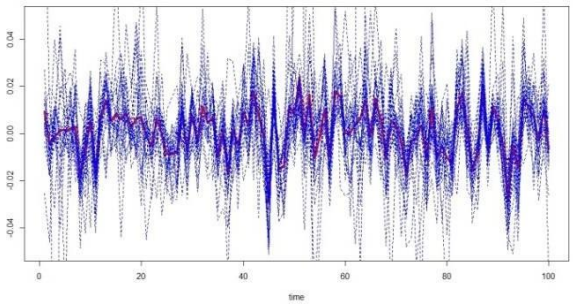


Рисунок 15. 15-ий кластер

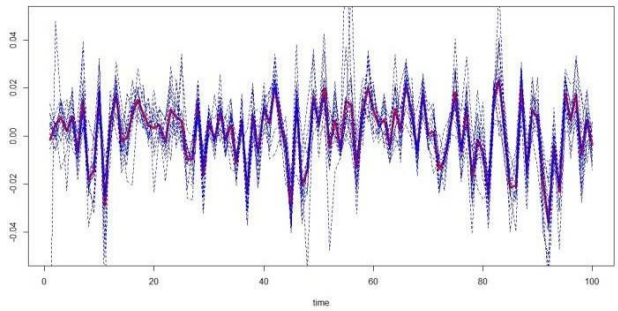
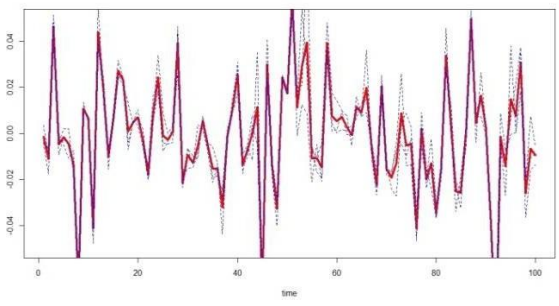
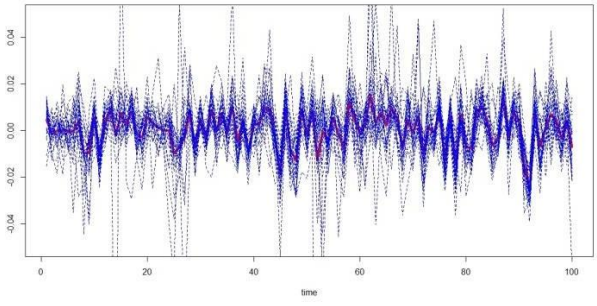


Рисунок16. 16-ий кластер





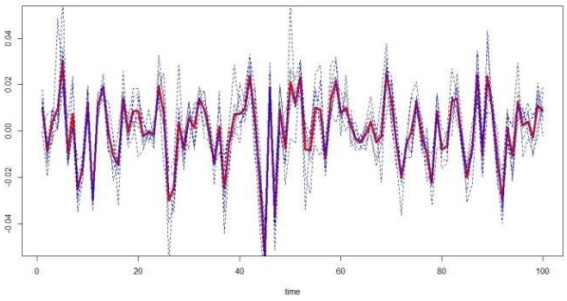


Рисунок 17. 17-ий кластер

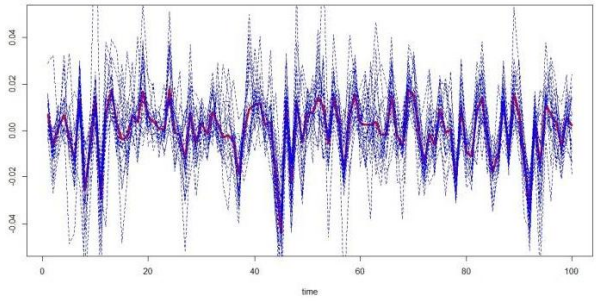


Рисунок18. 18-ий кластер

Рисунок 19. 19-ий кластер

Рисунок20. 20-ий кластер

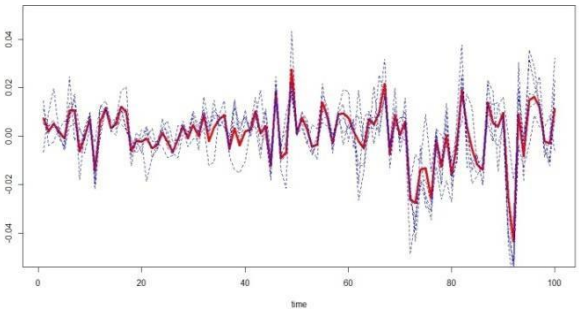


Рисунок 21. 21-ий кластер

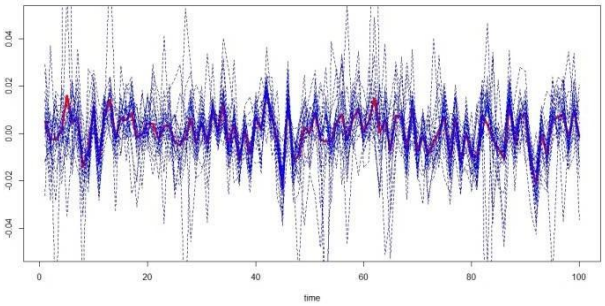


Рисунок22. 22-ий кластер

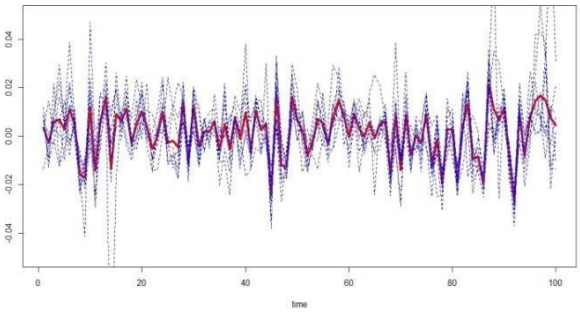


Рисунок 23. 23-ій кластер

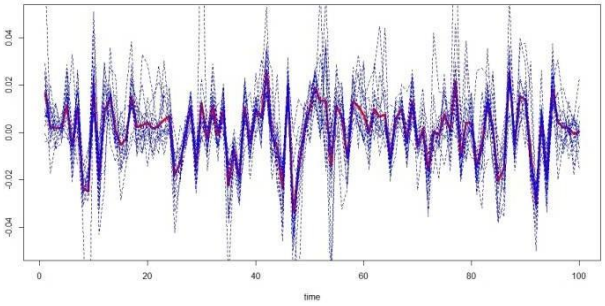


Рисунок24. 24-ий кластер

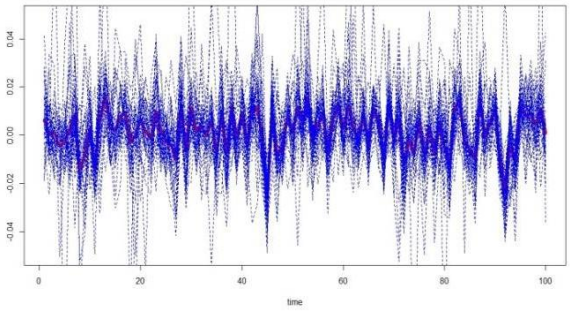


Рисунок 25. 25-ий кластер

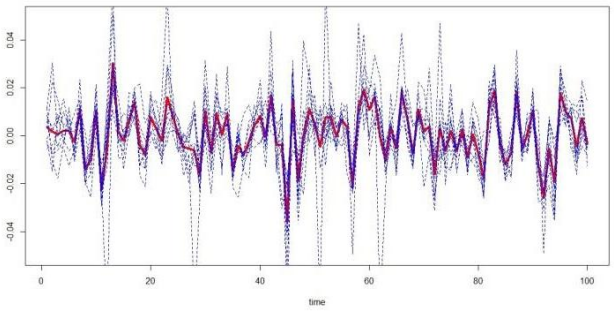


Рисунок26. 26-ий кластер

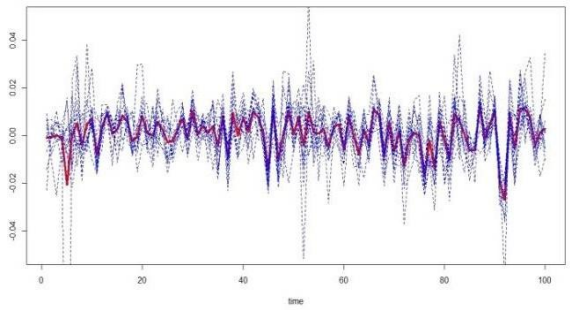


Рисунок 27. 27-ий кластер

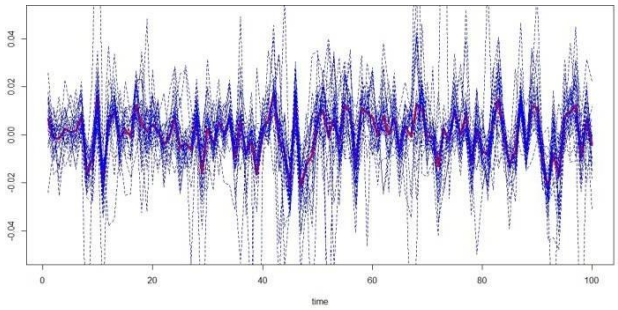


Рисунок28. 28-ий кластер

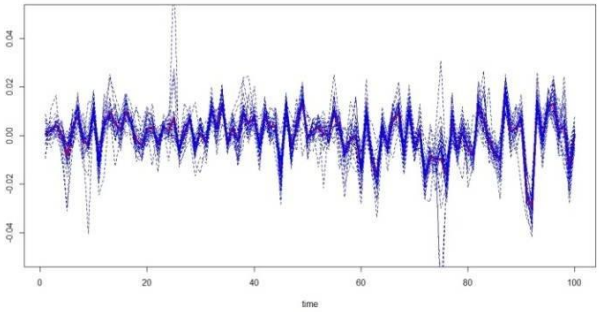


Рисунок 29. 29-ий кластер

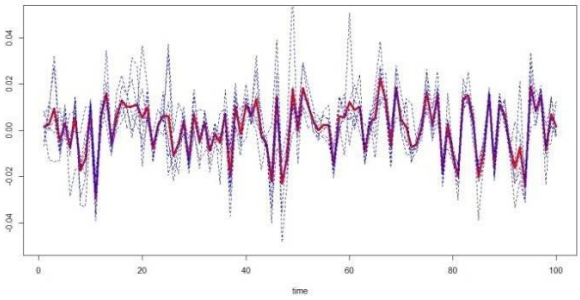


Рисунок30. 30-ий кластер

Додаток В

Реалізація кластерного аналізу серед RStudio

```
setwd("D:/ Forecasting")

## libraries
library(kernlab)
library(cluster)
library(flexclust)

##loading data
SPdata <- read.table("S&P 500.csv", header = TRUE, sep = ',') #
S&P500 data
SP <- read.table("S&P 500.csv", header = TRUE, sep = ',') #
constituents data
Sent <- data.frame(read.table("S&P 500.csv", header = TRUE, sep
= ',')) #sentiments data

##

## Functions
######## Concatenating function
p <- function(..., sep='') {
  paste(..., sep=sep, collapse=sep)
}
##

#Part 1: extraction informative variables from SPdata

Del = c("Close","Vol","Open","High","Low")

delno = 1:5;
for(i in 1:5){
  delno[i] = which(names(SPdata)==Del[i]) }

SPdaa = SPdata[,delno]

#Part 2: Cluster analysis (constituents and sentiment data)

dim(SP)
SPS<-data.frame(SP[,-1])

SPS3<-data.frame(t(SPS)) #transposed  data
SPS3_scaled<-scale(SPS3) #scaled data
```

```
#Block 1: "Elbow" method for defining number of clusters

#1 Plot of the total within-groups sums of squares
#against the number of clusters in K-means solution

wssplot<-function (data,nc=100,seed=1234) {

  wss<-(nrow(data)-1)*sum(apply(data,2,var))
  for (i in 2:nc) {
    set.seed(seed)
    wss[i]<-sum(kmeans(data,centers=i)$withinss)}

  plot(1:nc,wss,type="b", main="Кластерный анализ (k-
means)",xlab="количество кластеров", ylab="")}

wssplot(SPS3)

#Block 2: Cluster analysis of constituents data

#2 Comparing different methods of clustering

k=40 #number of clusters

#spectral clustering (Ng - Jordann - Weiss algorithm)
set.seed(2134)
output<-specc(SPS3_scaled, centers=k, kernel = "rbfdot", kpar =
"automatic",nystrom.red = FALSE, nystrom.sample =
dim(SPS3)[1]/6, iterations = 1000, mod.sample = 0.75)

class_out = as.data.frame(factor(output))
clusters(output)
size(output)
class_out[[1]]

class_sps3<-cbind(SPS3,as.data.frame(class_out))


v<-c("spec_radial",size(output))
output_total<-t(data.frame(v))
nm<-1:k
nnm<-c("method_name",nm)
colnames(output_total)<-nnm

# kernel k-means (radial kernel)
set.seed(2134)
output2<-kkmeans(SPS3_scaled, centers=k, kernel = "rbfdot", kpar
= "automatic",alg="kkmeans", p=1)
output2
```

```
v<-c("kkmeans_radial",size(output2))
newr<-t(data.frame(v))
output_total<-rbind(output_total,newr)

#kernel k-means (linear kernel)
set.seed(2134)
output3<-kkmeans(SPS3_scaled, centers=k, kernel = "vanilladot",
kpar = "automatic",alg="kkmeans", p=1)
output3
size(output3)

v<-c("kkmeans_linear",size(output3))
newr<-t(data.frame(v))
output_total<-rbind(output_total,newr)


#k-means (euclidian measure)
set.seed(2134)
output4 = kmeans(SPS3_scaled, centers=k)
output4$size

v<-c("kmeans",output4$size)
newr<-t(data.frame(v))
output_total<-rbind(output_total,newr)

# kernel k-means (polynomial kernel)
set.seed(2134)
output5<-kkmeans(SPS3_scaled, centers=k, kernel = "polydot",
kpar = "automatic",alg="kkmeans", p=1)
output5

v<-c("kkmeans_polynomial",size(output5))
newr<-t(data.frame(v))
output_total<-rbind(output_total,newr)

# spectral clustering (polynomial kernel)
set.seed(2134)
output6<-specc(SPS3_scaled, centers=k, kernel = "polydot", kpar
= "automatic",nystrom.red = FALSE, nystrom.sample =
dim(SPS3)[1]/6, iterations = 1000, mod.sample = 0.75)
output6

v<-c("spec_polynomial",size(output6))
newr<-t(data.frame(v))
output_total<-rbind(output_total,newr)

#spectral clustering (linear kernel)
```

```
set.seed(2134)
output7<-specc(SPS3_scaled, k, kernel = "vanilladot", kpar =
"automatic",nystrom.red = FALSE, nystrom.sample =
dim(SPS3)[1]/6, iterations = 1000, mod.sample = 0.75)
output7

v<-c("spec_linear",size(output7))
newr<-t(data.frame(v))
output_total<-rbind(output_total,newr)

#K-medians
set.seed(2134)
output8<-kcca(SPS3_scaled,k, family=kccaFamily("kmedians"),
weights=NULL, group=NULL,control=NULL, simple=FALSE,
save.data=FALSE)
output8

v<-c("k-medians",output8@clusinfo[[1]])
newr<-t(data.frame(v))
output_total<-rbind(output_total,newr)

# Spectral clustering (ANOVA kernel)
set.seed(2134)
output9<-specc(SPS3_scaled,centers=k, kernel = "anovadot", kpar
= "automatic",nystrom.red = FALSE, nystrom.sample =
dim(SPS3)[1]/6, iterations = 1000, mod.sample = 0.75)
output9

v<-c("spec_anova",size(output9))
newr<-t(data.frame(v))
output_total<-rbind(output_total,newr)

output_total
```

Додаток Г

Реалізація моделювання прогнозу за допомогою методу опорних векторів у середовищі RStudio

```
##############
library(caret)
require(kernlab)

data1=read.csv("S&P 500.csv",header=TRUE, sep=",",na="")
data=data1[,-1]

#############################################################
######################Preprocessing##########################

y<-data$Close
x<-data[,-which(names(data)=="Close")]

n<-ncol(x)
yhat<-c()
test<-c()

#1. Create the partition of the data into train and test sets
#set.seed(43)

nk<-400
tk<-5
for (i in seq(1,nrow(x)-(nk+tk), by=tk)){

   inTrain<-i+nk-1
   inTest<-inTrain+tk

   trainX<-x[i:inTrain,]
   testX<-x[(inTrain+1):inTest,]
   trainY<-y[i:inTrain]
   testY<-y[(inTrain+1):inTest]

############################################################
#2. Setting model
############################################################

#option: using package caret
##############################

# Start Parallel process
   library(doParallel)

   cl <- makeCluster(detectCores())
```

```
   registerDoParallel(cl)

   ctrl<-trainControl(method = "timeslice",
                      initialWindow = 360,
                      horizon = 40,
                      fixedWindow = FALSE)

   set.seed(2364)

   SVMmodel<-train(trainX, trainY,
                   method="svmPoly",
                   #tuneGrid=svmGrid,
                   trControl=ctrl,
                   preProc=c("center","scale"),
                   tuneLength = 20) #use the default grid search
of parameters values

   pred<-predict(SVMmodel,testX)

   outputY<-postResample(pred,testY)
   print(outputY)


   yhat<-c(yhat,pred)
   test<-c(test,testY)

   stopCluster(cl)

#stop the parallel process

}

head(SVMmodel$pred)

SVMmodel
SVMmodel$finalmodel

alphaindex(SVMmodel$finalModel)
coef(SVMmodel$finalModel)

result<-postResample(yhat,test)
rse<-(mean(yhat-test)^2)/(mean(yhat-mean(test)^2))

result
rse

result1<- resamples(yhat,test)
```

```
summary(result1)

#quality<-cbind(as.data.frame(result), as.data.frame(rse))
#quality

edelta<-yhat-test

pdf("Histogram for residuals(SVM).pdf")
hist(edelta, main="Histogram of residuals")
dev.off()
pdf("Output for SVM (lag3).pdf")

plot(test, type="l", ylim=range(c(-0.02,0.02)))
lines(yhat, type="l", col="red")
dev.off()

summary(edelta)
shapiro.test(edelta)
output<-data.frame(cbind(yhat,test,edelta))
output
```

Додаток Д

Реалізація моделювання прогнозу за допомогою нейронної мережі в середовищі RStudio

```
###################Neural network modeling#####################
library(caret)
library(nnet)

data1=read.csv("S&P 500.csv",header=TRUE, sep=",",na="")
data=data1[,-1]

#######################Preprocessing#########################

y<-data$Close
x<-data[,-which(names(data)=="Close")]
n<-ncol(x)
yhat<-c()
test<-c()

nk<-400
tk<-10

#1. Create the partition of the data into train and test sets
#set.seed(43)

for (step in seq(1,(nrow(x)-(tk+nk)+1), by=tk)){
   inTrain<-step+nk
   inTest<-inTrain+tk

   trainX<-x[step:inTrain,]
   testX<-x[(inTrain+1):inTest,]
   trainY<-y[step:inTrain]
   testY<-y[(inTrain+1):inTest]

#2. Creation and optimization  neural nets

# Start Parallel process

   library(doParallel)
   cl <- makeCluster(detectCores())
   registerDoParallel(cl)

#option 2: using package caret

   nnetGrid<-expand.grid(.size=5:30,
.decay=c(0.01,0.1,0.2,0.5,1,1.5,2))
```

```
   set.seed(100)

   ctrl<-trainControl(method = "timeslice",
                      initialWindow = 390,
                      horizon = 10,
                      fixedWindow = TRUE)

   nnetFit<-train(trainX, trainY,
                  method="nnet",
                  tuneGrid=nnetGrid,
                  trControl=ctrl,
                  preProc=c("center","scale"),
                  MaxNWts=30*(ncol(trainX)+1)+30+1,
                  maxit=500)

   predNNt<-predict(nnetFit,testX)

   yhat<-c(yhat,predNNt)
   test<-c(test,testY)

   stopCluster(cl)
#stop the parallel process

}

nnetFit

#Quality of the model
result<-postResample(yhat,test)
rse<-(mean(yhat-test)^2)/(mean(yhat-mean(test)^2))

result
rse

#Output residuals

edelta<-yhat-test

#Histogram for residuals

pdf("Histogram for residuals(nnet).pdf")
hist(edelta, main="Histogram of residuals")
dev.off()

summary(edelta)
shapiro.test(edelta)
output<-data.frame(cbind(yhat,test,edelta))
output
```