КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
Факультет комп'ютерних наук та кібернетики
Кафедра теоретичної кібернетики

Роботу розглянуто й допущено до захисту на
засіданні кафедри теоретичної кібернетики
« » травня 2021 р.,
протокол №
Завідувач кафедри
Ю. В. Крак _____
(підпис)

Випускна кваліфікаційна робота

на здобуття ступеня бакалавра

за спеціальністю 122 Комп'ютерні науки
на тему:

Інтелектуальні методи обробки фінансових даних

Виконав студент 4 курсу
Листопад Дмитро Андрійович

_____
(підпис)

Науковий керівник:
професор, доктор фізико-математичних наук
Пашко Анатолій Олексійович

_____
(підпис)

Засвідчую, що в цій дипломній роботі немає запозичень з праць інших авторів без відповідних посилань.

Студент _____
(підпис)

Київ – 2021

РЕФЕРАТ

Обсяг роботи 42 сторінки, 7 рисунків, 7 таблиць, 16 джерел посилань

АКЦІЇ, МОДЕЛІ ПРОГНОЗУВАННЯ, ЧАСОВІ РЯДИ, ARIMA, МОДЕЛЬ МАРКОВІЦА, СЕРЕДНЬОСРОКОВЕ ІНВЕСТУВАННЯ .

Об’єктом дослідження є акції американського ринку, а саме їх інтерпретація, як часових рядів. Предметом роботи є економетрична модель ARIMA для прогнозування ціни акцій і модель Марковіца для формування середньострокового портфелю.

Метою роботи є написання зручної та гнучкої програми для аналізу часових рядів, прогнозування ходу акцій з якомога більшою точністю, формування портфелю із списку акцій, і тестування стратегій.

Методи розроблення: комп’ютерне моделювання, методи аналізу часових рядів, розробка програмного продукту.

Результати роботи: виконано загальний огляд методів дослідження часових рядів, та окремих моделей прогнозування, проаналізовано переваги та недоліки використання різних моделей, розроблено програмний продукт, який дозволяє дослідити ряд на різні властивості, та зробити прогноз на найближче майбутнє.

Створений програмний продукт може використовуватися професійними трейдерами на американській біржі та окремими інвесторами, які хочуть збільшити свій пасивний дохід та підкріпити свої знання якісним аналізом.

ЗМІСТ

СКОРОЧЕННЯ ТА УМОВНІ ПОЗНАЧЕННЯ

AR — autoregressive, процес авторегресії

MA — moving average, процес ковзаючого середнього

ARIMA — autoregressive integrated moving average, інтегрована модель авторегресії-ковзаючого середнього

ACF(АКФ) — auto corellation function, автокореляційна функція

PACF(ЧАКФ) — partial auto corellation function, часткова автокореляційна функція

ВСТУП

Оцінка сучасного стану об’єкта дослідження. Інтелектуальний аналіз даних – це потужна технологія для аналізу важливої інформації зі сховища даних. Ця технологія аналізу даних використовується для ідентифікації прихованих закономірностей у великому наборі даних. Інтелектуальний аналіз даних успішно використовується в різних областях, включаючи й освітнє середовище, він використовується там, де сьогодні є цифрові дані. Помітні приклади можна знайти в бізнесі, медицині, науці та спостереженні. В даний час методи Data Mining отримали широке поширення в різних сферах діяльності. Конкретні переваги видобутку даних різняться залежно від цілі та галузі. Відділи продажів та маркетингу можуть видобувати дані клієнтів для покращення коефіцієнта конверсії або для створення маркетингових кампаній « один на один» . Інформація про видобуток даних про історичні структури продажів та поведінку клієнтів може використовуватися для побудови моделей прогнозування майбутніх продажів, нових продуктів та послуг. Компанії фінансової галузі використовують інструменти видобутку даних для побудови моделей ризику та виявлення шахрайства. Обробна промисловість використовує інструменти для видобутку даних для підвищення безпеки продукції, виявлення проблем якості, управління ланцюгом поставок та покращення операцій. Дослідженнями в цій області займаються такі вчені, як А.А. Барсегян, М.С. Купріянов, Г. Пятецкій-Шапіро, Х. Ромесбург, Дж. Хан.

Актуальність роботи та підстави для її виконання. У зв'язку із зростаючими обсягами даних в усіх сферах, а час на їх обробку, ясна річ, обмежений, потрібно знати розуміти та використовувати різні методи та алгоритми. Так у даній роботі буде розглянуто основні поняття та алгоритми ІАД, а також використання одного з них на цілком реальному прикладі в ситуації, що може виникнути при вирішенні абсолютно різних задач. Кожен

алгоритм підходить для різних специфікацій задач, та є і ті, що можуть бути універсальними.

Мета й завдання роботи. Метою роботи є проведення аналізу дослідження можливостей застосування методів інтелектуального аналізу даних та є пошук оптимального алгоритму для аналізу фінансової діяльності фірми.

Для досягнення цієї мети поставлено такі завдання:

- Дослідити поняття ІАД.
- Дослідити можливі застосування ІАД.
- Дослідити основні алгоритми аналізу фінансових потоків.
- Розробити елементи програмного забезпечення для автоматизації розрахунків.

Об'єкт, методи й засоби розроблення. Об’єктом роботи є інтелектуальний аналіз даних та тестування алгоритму на прикладі аналізу фінансових потоків

РОЗДІЛ 1. DATA MINING АБО ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ

1.1 Визначення та застосування

Інтелектуальний аналіз даних - це процес виявлення шаблонів у великих наборах даних з використанням методів на перетині машинного навчання , статистики та систем баз даних . [1] Інтелектуальний аналіз даних - це міждисциплінарне підполе інформатики та статистики з загальною метою вилучення інформації (з інтелектуальними методами) з набору даних та перетворення інформації в зрозумілу структуру для подальшого використання. [1] [2] Інтелектуальний аналіз даних - це етап аналізу процесу "знань у базах даних" або KDD. [5]Крім кроку сировинного аналізу, він також включає аспекти управління базами даних і даними , попередню обробку даних , розгляд моделей і висновків , показники цікавості, міркування складності , постобробку виявлених структур, візуалізацію та оновлення в режимі онлайн . [1] Різниця між аналізом даних і інтелектуальний аналіз даних полягає в тому, що аналіз даних використовується для тестування моделей і гіпотез на наборі даних, наприклад, аналізу ефективності маркетингової кампанії, незалежно від кількості даних; навпаки, інтелектуальний аналіз даних використовує машинне навчання та статистичні моделі для розкриття підпільних або прихованих моделей у великому обсязі даних. [6]

Data mining також є модним словом [8] і часто застосовується до будь-якої форми великомасштабної обробки даних або інформації (збір , вилучення , складування , аналіз і статистика), а також будь-яке застосування системи підтримки прийняття рішень , включаючи комп'ютер штучний інтелект (наприклад, машинне навчання) і бізнес-аналітика . Книга Інтелектуальний аналіз даних: Практичні засоби та методи машинного навчання з Java [9] (що охоплює в основному матеріал для

машинного навчання) спочатку називався лише практичним машинним навчанням , а термін інтелектуального аналізу даних був доданий лише з маркетингових причин. [10] Часто більш загальні терміни аналізу даних і аналітика - або, коли мова йде про реальні методів, штучного інтелекту і машинного навчання - є більш придатними.

Фактично завдання інтелектуального аналізу даних - це напівавтоматичний або автоматичний аналіз великої кількості даних для вилучення раніше невідомих, цікавих моделей, таких як групи записів даних (кластерний аналіз), незвичайні записи (виявлення аномалій) та залежності (видобування правил асоціації , послідовне видобування шаблонів). Це зазвичай передбачає використання методів бази даних, таких як просторові індекси . Ці закономірності потім можна розглядати як своєрідне резюме вхідних даних і можуть бути використані в подальшому аналізі або, наприклад, в машинному навчанні та аналітиці прогнозування. Наприклад, етап видобування даних може ідентифікувати декілька груп у даних, які потім можуть бути використані для отримання більш точних результатів прогнозування за допомогою системи підтримки прийняття рішень . [11]

Інтелектуальний аналіз даних може допомогти підприємству точніше оцінити свою роботу. Розглянемо один з методів - аналіз споживчого кошика. Його застосовують, щоб виявити переваги споживачів і, відповідно, краще задовольнити попит і підвищити дохід з клієнтів. Однак характер купівельної поведінки присутня в даних неявно, і для його визначення необхідно використовувати саме Data Mining. І тепер можна з'ясувати, наприклад, що клієнт, який збирається купити товар X, буде не проти придбати заодно і товар Y. Ця інформація ляже в основу наступних рішень: може бути, варто розташовувати ці товари на вітрині магазину поруч або, наприклад, просувати один з них, щоб підвищити продажі обох. Додатки Data Mining застосовуються досить широко в: роздрібної торгівлі, маркетингу, фінансах, охороні здоров'я,

промисловому виробництві та інших областях. [11] У бізнесі інтелектуальний аналіз даних - це аналіз історичної ділової діяльності, що зберігається як статичні дані в базах даних сховища даних. Мета полягає в тому, щоб виявити приховані закономірності та тенденції. Програмне забезпечення інтелектуального аналізу даних використовує розширені алгоритми розпізнавання образів для просіювання великих обсягів даних для допомоги у виявленні раніше невідомої стратегічної бізнес-інформації. Приклади того, як підприємства використовують інтелектуальний аналіз даних, полягає в тому, щоб включити проведення аналізу ринку для виявлення нових пакетів продуктів, пошуку корінних причин виробничих проблем, запобігання виснаженню клієнтів і придбання нових клієнтів, перехресних продажів існуючим клієнтам і більш точним визначенням клієнтів . [12]У сучасному світі незмінені дані збираються компаніями зі швидкістю вибуху. Наприклад, Walmart обробляє понад 20 мільйонів операцій з продажу кожного дня. Ця інформація зберігається в централізованій базі даних, але без будь-якого програмного забезпечення для інтелектуального аналізу даних для її аналізу. Якщо Walmart проаналізував їхні дані про продаж з методами інтелектуального аналізу даних, вони могли б визначити тенденції продажів, розробити маркетингові кампанії та більш точно передбачити лояльність клієнтів. [13] [14] Одним з таких прикладів для Walmart буде продаж підгузків і пива, виявлених за допомогою інтелектуального аналізу даних. [15]

В останні роки інтелектуальний аналіз даних широко використовувався в галузях науки і техніки, таких як біоінформатика, генетика, медицина, освіта і електроенергетика. При вивченні генетики людини послідовність видобутку допомагає вирішувати важливу мету розуміння співвідношення карти між індивідуальними варіаціями в послідовності ДНК людини і мінливістю в сприйнятливості хвороб. Простіше кажучи, він прагне з'ясувати, як зміни в послідовності ДНК людини впливають на ризики розвитку загальних

захворювань, таких як рак , що має велике значення для вдосконалення методів діагностики, профілактики та лікування цих захворювань. Один метод інтелектуального аналізу даних, який використовується для виконання цього завдання, відомий як зменшення багатовимірної розмірності . [12]

1.2 Основні завдання, що вирішуються інтелектуальним аналізом

Розвиток методів запису і зберігання даних привів до бурхливого зростання обсягів інформації, що збирається і аналізується. Обсяги даних настільки значні, що людині просто не під силу проаналізувати їх самостійно, хоча необхідність проведення такого аналізу цілком очевидна, адже в цих "сирих" даних укладені знання, які можуть бути використані при прийнятті рішень. Для того щоб провести автоматичний аналіз даних, використовується Data Mining.

Data Mining - це процес виявлення в "сирих" даних раніше невідомих нетривіальних практично корисних і доступних інтерпретації знань, необхідних для прийняття рішень в різних сферах людської діяльності. Інформація, знайдена в процесі застосування методів Data Mining, повинна бути нетривіальною і раніше невідомої, наприклад, середні продажу не є такими. Знання повинні описувати нові зв'язки між властивостями, передбачати значення одних ознак на основі інших і т.д. Знайдені знання повинні бути застосовні і на нових даних з деякою мірою вірогідності. Корисність полягає в тому, що ці знання можуть приносити певну вигоду при їх застосуванні. Знання повинні бути в зрозумілій для користувача не математика вигляді. Наприклад, найпростіше сприймаються людиною логічні конструкції "якщо ... то ...". Більш того, такі правила можуть бути використані в різних СУБД в якості SQL-запитів.

Алгоритми, що використовуються в Data Mining, вимагають великої кількості обчислень. Раніше це було стримуючим фактором широкого практичного застосування інтелектуального аналізу, проте сьогоднішнє зростання продуктивності сучасних процесорів зняв гостроту цієї проблеми.

Тепер за прийнятний час можна провести якісний аналіз сотень тисяч і мільйонів записів. [11]

Завдання, які вирішуються методами Data Mining:

1. **Класифікація** - це віднесення об'єктів (спостережень, подій) до одного з заздалегідь відомих класів.

2. **Регресія** , в тому числі завдання прогнозування. Встановлення залежності безперервних вихідних від вхідних змінних.

3. **Кластеризація** - це групування об'єктів (спостережень, подій) на основі даних (властивостей), що описують сутність цих об'єктів. Об'єкти усередині кластера повинні бути "схожими" один на одного і відрізнятися від об'єктів, які увійшли в інші кластери. Чим більше схожі об'єкти усередині кластера і чим більше відмінностей між кластерами, тим точніше кластеризація.

4. **Асоціація** - виявлення закономірностей між пов'язаними подіями. Прикладом такої закономірності служить правило, яке вказує, що з події X слід подія Y. Такі правила називаються асоціативними. Вперше ця задача була запропонована для знаходження типових шаблонів покупок, що здійснюються в супермаркетах, тому іноді її ще називають аналізом ринкової корзини (market basket analysis).

5. **Послідовні шаблони** - встановлення закономірностей між пов'язаними в часі подіями, тобто виявлення залежності, що якщо відбудеться подія X, то через заданий час відбудеться подія Y.

6. **Аналіз відхилень** - виявлення найбільш нехарактерних шаблонів.

Проблеми бізнес аналізу формулюються по-іншому, але рішення більшості з них зводиться до тієї чи іншої задачі Data Mining або до їх комбінації. Наприклад, оцінка ризиків - це вирішення завдання регресії або класифікації, сегментація ринку - кластеризація, стимулювання попиту - асоціативні правила.

Фактично, завдання Data Mining є елементами, з яких можна зібрати рішення переважної більшості реальних бізнес завдань. [11]

Для вирішення вищеописаних завдань використовуються різні методи і алгоритми Data Mining. З огляду на те, що Data Mining розвивалася і розвивається на стику таких дисциплін, як статистика, теорія інформації, машинне навчання, теорія баз даних, цілком закономірно, що більшість алгоритмів і методів Data Mining були розроблені на основі різних методів з цих дисциплін. Наприклад, процедура кластеризації k-means була просто запозичена з статистики. Велику популярність отримали такі методи Data Mining: нейронні мережі, дерева рішень, алгоритми кластеризації, в тому числі і масштабовані, алгоритми виявлення асоціативних зв'язків між подіями і т. д.

1.3 Найбільш впливові алгоритми Data Mining

Прагнучи визначити деякі з найбільш впливових алгоритмів, які широко використовуються в ітелектуальному аналізі даних, на міжнародній конференції IEEE International Conference on Data Mining 2006 було визначено 10 кращих алгоритмів для інтелектуального аналізу даних. Розглянемо декілька з них:

1. C4.5 створює класифікатор у вигляді дерева рішень. Для цього C4.5 дається набір даних, який представляє собою вже класифіковані речі. Припустимо, в наборі даних є деяка кількість пацієнтів. Про кожного з них нам відомі різні факти: їх вік, пульс, кров'яний тиск, історія спадкових захворювань і т.д. Їх називають атрибутами. Отже з огляду на ці атрибути, ми хочемо передбачити, захворіє чи пацієнт раком. Пацієнт може потрапити в один з двох класів: «захворіє на рак» і «не захворіє на рак». Ми повідомляємо C4.5 про відповідному класі кожного пацієнта. За допомогою набору атрибутів і відповідного класу пацієнта, C4.5 будує дерево рішень, яке може передбачити клас нових пацієнтів на основі їх атрибутів. [15]

C4.5 відрізняється від інших систем дерев рішень тим, що:

 Для початку, C4.5 використовує відносну ентропію при генеруванні дерев рішень.

 Незважаючи на те, що інші системи теж використовують відсікання гілок, C4.5 використовує однопроходное відсікання гілок , щоб спростити перенавчання.Відсікання гілок істотно покращує роботу алгоритму .

 C4.5 може працювати і з безперервними, і з дискретними даними. Вказавши діапазони або порогові значення безперервних даних, можна їх перетворити в дискретні.

2. Метод опорних векторів (**SVM**) знаходить гіперплощину для класифікації даних в два класи. У двох словах, SVM виконує завдання, аналогічну C4.5, за винятком того, що він не використовує дерева рішень. [15] SVM може спроектувати ваші дані в більш високі виміри. А після цього SVM знаходить найбільш підходящу гіперплощину, що розділяє ваші дані в два класи. Наприклад. Уявімо кілька червоних і синіх куль на столі. Якщо кулі не надто сильно перемішані, ви можете покласти між ними палицю, не рухаючи самих куль. Тепер коли на стіл кладуть новий шар, знаючи при цьому, на якій стороні палиці він тепер перебуває, можна передбачити його колір. Кулі символізують точки даних, а червоний і синій кольори символізують два класи. Палка символізує найпростішу гіперплощину - лінію. Якщо ж кулі перемішані між собою, пряма палиця не спрацює. [15] Ось обхідне рішення: Швидко підійміть стіл - і кулі виявляться в повітрі. У той час, поки кулі все ще в повітрі, ви можете використовувати великий аркуш паперу, щоб розділити їх.

Підняття столу відповідає відображенню ваших даних у вищих вимірах. У цьому випадку, ми переходимо від двовимірної площини столу до тривимірної, де кульки знаходяться в повітрі. З використанням ядра у нас з'являється приємний спосіб роботи в вищих вимірах. Великий аркуш паперу все ще можна назвати гіперплощиною, але тепер це функція площині, а не лінії. Коли ми знаходимося в 3-х вимірах, гіперплощина є площиною, а не лінією. [15] SVM і

C4.5 - два класифікатора, які варто спробувати в першу чергу . Немає класифікатора, ідеального у всіх сенсах через теореми No Free Lunch . Крім того, іншими недоліками алгоритму є вибір ядра і інтерпретованість.

3. AdaBoost - це алгоритм посилення класифікаторів. Класифікатор приймає деяку кількість даних і намагається передбачити або класифікувати, до якого класу належить новий елемент даних. Яке відмінність між сильним і слабким учнем? Слабкий учень класифікує з точністю тільки трохи вище випадковості. Відомим прикладом слабкого учня є однорівневе дерево рішень. У сильного учня точність набагато вище, і часто використовуваним прикладом сильного учня є SVM. Можна навести приклад AdaBoost ? Давайте почнемо з трьох слабких учнів. Ми навчимо їх в 10 раундів на навчальному наборі даних, що містить дані про пацієнтів. Набір даних містить детальну інформацію про медичні записах пацієнта. Як би нам передбачити, чи отримає пацієнт рак? Ось як AdaBoost відповідає на це питання. У першому раунді : AdaBoost бере зразок навчального набору і тестує його, щоб побачити наскільки точний кожен учень. В результаті ми знаходимо кращого учня. Крім того, неправильно класифікованих зразкам присвоюється більшу вагу, щоб у них був вищий шанс бути обраними в наступний раунд. Ще дещо: кращому учневі теж присвоюється вага, покладаючись на його точність, і його включають в агрегацію учнів (зараз є тільки один учень). У другому раунді: AdaBoost знову намагається знайти кращого учня. І сіль ось у чому: Зразок даних пацієнта для тренування тепер знаходиться під впливом більшою мірою неправильно класифікованих зразків. Іншими словами, у раніше неправильно класифікованого пацієнта з'являється більш високий шанс з'явитися в зразку. Це щось на зразок "збереження" у відеоіграх, коли тобі не доводиться починати гру з самого початку, коли твого персонажа вбивають. Замість цього ти можеш зосередити всі свої зусилля на те, щоб перейти на наступний рівень. Крім того, перший учень, швидше за все, класифікував деяких пацієнтів правильно. Замість того, щоб класифікувати їх

знову, всі зусилля зосереджені на класифікації раніше неправильно класифікованих пацієнтів. Кращому учню знову присвоюється вага і його включають в агрегацію; неправильно класифікованих пацієнтам присвоюється вага таким чином, що у них більше шансів бути обраними. І все це повторюється знову. Після закінчення десяти раундів: у нас тепер є агрегація учнів, тренованих і повторно тренованих на неправильно класифікованих даних з попередніх раундів. Чому варто використовувати AdaBoost? AdaBoost простий. Алгоритм досить легко реалізувати. До того ж, він дуже швидкий! Слабкі учні зазвичай набагато простіше сильних. Те, що вони простіше, означає що вони, ймовірно, швидше виконуються. Ще дещо… Це дуже елегантний спосіб автоналаштування класифікатора, так як в кожний наступний раунд AdaBoost уточнює значимість для кожного з кращих учнів. Все, що вам потрібно вказати - це кількість раундів. Нарешті, він гнучкий і універсальний. AdaBoost може включити в себе будь-який алгоритм навчання, і він може працювати з дуже різноманітними даними.

4. Apriori. Що він робить? Алгоритм Apriori вивчає правила асоціації та застосовується до бази даних, що містить велику кількість транзакцій.

Що таке правила асоціації? Навчання правилам асоціації - це метод обміну даними для вивчення кореляцій та співвідношень між змінними в базі даних.

Який приклад Апріорі? Скажімо, у нас є база даних, повна транзакцій із супермаркетами. Можна базувати базу даних як гігантську електронну таблицю, де кожен рядок - це операція з клієнтом, а кожен стовпець являє собою інший продуктовий елемент.

Ось найкраща частина. Застосовуючи алгоритм Apriori, ми можемо вивчити продуктові предмети, які купуються разом, як правила асоціації.

Сила цього. Ви можете знайти ті предмети, які, як правило, купуються разом частіше, ніж інші предмети - кінцева мета - змусити покупців купувати більше. Разом ці елементи називаються наборами елементів.

Наприклад:

Ви, ймовірно, швидко можете побачити, що фішки + занурення і чіпси + сода, здається, часто трапляються разом. Вони називаються 2-наборовими наборами. З достатньо великим набором даних буде набагато важче "побачити" відносини, особливо коли ви маєте справу з наборами 3-х елементів або більше. Саме в цьому допомагає Apriori. Як працює Apriori? Перш ніж потрапляти в алгоритм, що потребує гостроті, вам потрібно буде визначити 3 речі:

 Перший - це розмір вашого набору предметів. Ви хочете бачити візерунки для набору 2-х предметів, 3-х предметів тощо?

 Друга - ваша підтримка або кількість транзакцій, що містять набір предметів, поділену на загальну кількість транзакцій. Набір елементів, який відповідає підтримці, називається частим набором елементів.

 Третє - це ваша впевненість або умовна ймовірність якогось предмета, якщо ви маєте певні інші елементи у своєму наборі предметів. Хороший приклад - це фішки у вашому наборі предметів, є 67% впевненості, що газована продукція також є у наборі предметів.

5. PageRank зазвичай використовується в таких пошукових системах, як Google. Це алгоритм аналізу зв’язку, який визначає відносну важливість об'єкта, пов'язаного всередині мережі об'єктів. Аналіз зв’язків - це тип мережевого аналізу, який досліджує асоціації між об'єктами. Пошук Google використовує цей алгоритм, розуміючи зворотні посилання між веб-сторінками. Це один із методів, якими Google користується для визначення відносної важливості веб-сторінки та підвищення її рівня в пошуковій системі Google. Торгова марка PageRank є власником Google, а алгоритм PageRank запатентований

Університетом Стенфорда. Трактується як підхід без нагляду за навчанням, оскільки він визначає відносну важливість лише враховуючи посилання та не потребує інших вкладів.

1.4 Методи та алгоритми класифікації

Метод К-середніх є ітераційним методом кластерного аналізу. У чому полягає основна відмінність ітераційних методів від ієрархічних? У ієрархічному агломеративному кластерному аналізі, у нас спочатку кожне спостереження являє собою окремий кластер, а потім за допомогою якогось методу ми поступово об'єднуємо їх в групи все більшого і більшого розміру, і в кінці у нас все спостереження вже лежать в одному кластері. І після цього ми намагаємося виявити момент, коли розбиття у нас було оптимальним, і ми почали вже об'єднувати кластери між собою. Іншими словами, в разі ієрархічного кластерного аналізу ми не знаємо заздалегідь, скільки кластерів ми виділимо по нашим даним. І, насправді, в цьому полягає основна відмінність ітераційних методів від ієрархічних, тому що в ітераційних методах ми повинні заздалегідь знати, яка кількість кластерів ми збираємося виділити по нашій вибірці.

Ну звідки ми можемо знати цю інформацію? Ми, в принципі, можемо знати її апріорі. Ну, наприклад, ми намагаємося кластеризувати анкетні дані. У нас були респонденти, які відповідали на якісь питання, і зараз ми намагаємося розбити наші дані на кластер чоловіків і кластер жінок. Ну, відповідно, в даному випадку ми апріорі знаємо кількість груп, які ми намагаємося виділити в своїх даних. Якщо ми не знаємо якийсь апріорної інформації, ми можемо отримати інформацію про кількість кластерів, провівши якийсь попередній аналіз. Ну, наприклад, ми могли провести попередньо ієрархічний кластерний аналіз, визначитися з кількістю кластерів, яке оптимально для наших даних, і після цього перейти вже до ітераційним методам кластерного аналізу.

Основна ідея методу К-середніх полягає в тому, що ми намагаємося виділити такі кластери в наших даних, щоб мінімізувати середньоквадратичне відхилення відстані кожного спостереження від центру кожного кластера. Розглянемо алгоритм більш докладно. На початку ми повинні визначитися з кількістю кластерів, яке ми виділяємо. Після того як ми вибрали кількість кластерів, ми повинні задати деяке початкове наближення для центрів цих кластерів. А тут, відповідно, теж ми можемо спиратися або на якусь апріорну інформацію, або на результати якогось попереднього аналізу, але насправді досить часто початкове наближення використовує просто якісь випадково згенеровані точки. Після того як ми вибрали початкове наближення для центрів наших кластерів, ми вважаємо відстань від кожного елемента в нашій вибірці до центрів цих кластерів. І після цього розподіляємо наші дані по кластерам, виходячи з того, до центру якого кластера ближче у нас виявилося те чи інше спостереження. Після того як ми розподілили всі наші дані по кластерам, ми перераховуємо центри кластерів як центри мас тих спостережень, які в нього потрапили. Відповідно, в даному випадку у нас центри кластерів зрушяться, і ми повторюємо етап з перерозподілом даних. Тобто ми знову вважаємо відстань від кожного спостереження до вже нового центру кластера і знову перерозподіляємо дані, тобто знову вибираємо, до якого кластеру зараз у нас виявилося ближче ту чи іншу спостереження. Ну і після чого знову повинні перерахувати центри кластерів, і ми повторюємо цю операцію до тих пір, поки у нас центри кластерів не перестануть зрушуватися. Тобто поки ми не виділимо досить стійкі кластера, які не змінюються у нас від ітерації до ітерації. Ну або ще завжди задають деякий максимальну кількість ітерацій, при перевищенні якого алгоритм також зупиняється. Ну це робиться для того, щоб ми просто не зациклювалися. Насправді у даного методу є проблеми і вимоги.

Основною, звичайно, проблемною вимогою є той факт, що ми повинні задати кількість кластерів. Це досить серйозне обмеження на можливість

застосування цього методу, тому що ми далеко не завжди володіємо достатньою інформацією, щоб сказати, як багато груп ми хочемо виділити в своїх даних. Також слід пам'ятати, що результат даного методу дуже сильно залежить від того початкового наближення центрів кластерів, яке ми використовуємо на першій ітерації. І в даному випадку ми повинні або більше зусиль витратити на якийсь попередній аналіз, щоб вибрати гарне початкове наближення, або, що дуже часто робиться на практиці, робити кілька запусків, ну тобто, наприклад, вибирати початкове наближення якимось випадковим чином , але робити це не один раз, а кілька, ну і порівнювати результати між собою. Такий підхід, звичайно, буде більш стійким. Також слід пам'ятати, що даний алгоритм не гарантує нам, що ми зійдемося до глобального мінімуму нашого функціоналу, тобто до глобального мінімуму середньоквадратичного відхилення відстані спостережень від центрів кластерів. Тобто, в принципі, ми можемо зійтися до якогось локального мінімуму.

К-засіб кластеризації є методом векторного квантування , спочатку з обробки сигналів , який є популярним для кластерного аналізу в інтелектуальному аналізі даних . k -означає, що кластеризація спрямована на розбиття n спостережень на k кластерів, в яких кожне спостереження належить до кластера з найближчим середнім , що є прототипом кластера. Це призводить до розбиття простору даних на комірки Вороного . [16]

Проблема обчислювальна, однак, ефективні евристичні алгоритми швидко сходяться до локального оптимуму . Вони, як правило , аналогічні алгоритм очікування-максимізація для сумішей з гауссових розподілів з допомогою ітераційного уточнення підходу , використовуваного як K-засобів і моделювання гауссових суміші . Вони обидва використовують центри кластерів для моделювання даних; однак, k - означає, що кластеризація має тенденцію до пошуку кластерів порівнянної просторової протяжності, тоді як механізм максимізації очікування дозволяє кластерам мати різні форми. [17]

Застосування класифікатора 1-найближчого сусіда до центрів кластерів, отриманих за допомогою k-засобів, класифікує нові дані в існуючі кластери. Це відомо як найближчий класифікатор центроїда або алгоритм Роккіо .

Кластеризація K-середніх є одним з найпростіших і найпопулярніших алгоритмів машинного навчання без нагляду.

Як правило, без нагляду алгоритми роблять висновки з наборів даних, використовуючи тільки вхідні вектори, не звертаючись до відомих або маркованих результатів. [16]

AndreyBu, який має більш ніж 5-річний досвід машинного навчання і в даний час навчає людей своїм навичкам, каже, що «мета K-засобів проста: об'єднайте подібні точки даних разом і відкрийте основні закономірності. Для досягнення цієї мети K-засоби шукають фіксоване число (k) кластерів у наборі даних. [16] Кластер відноситься до набору точок даних, об'єднаних разом через певні подібності. Ви визначите цільове число k, яке позначає кількість центроїдів, необхідних у наборі даних. Центроїд - це уявне або реальне розташування, що представляє центр кластера. Кожна точка даних виділяється кожному з кластерів за рахунок зменшення сукупності квадратів у кластері.

Іншими словами, алгоритм K-означає ідентифікувати k число центроїдів, а потім виділяє кожну точку даних до найближчого кластера, зберігаючи при цьому центроїди як можна менше.

"Засіб" у K-засобах відноситься до усереднення даних; тобто знаходження центроїда.

Як працює алгоритм K-засобів. Для обробки даних навчання алгоритм K-means у видобутку даних починається з першої групи випадково вибраних центроїдів, які використовуються як початкові точки для кожного кластера, а потім виконує ітераційні (повторювані) розрахунки для оптимізації позицій центроїдів[18]

Це зупиняє створення та оптимізацію кластерів, коли:

 Центроїди стабілізувалися - їх значення не змінилося, оскільки кластеризація була успішною.

 Досягнуто певну кількість ітерацій.

Варіації:

 k -медіани кластеризації використовують медіану у кожному вимірі замість середнього, і цей спосіб мінімізується норма (геометрія таксакабу).

k -медоїди (також: Розбиття навколо Medoids, PAM) використовує медоїд замість середнього і таким чином мінімізує суму відстаней для функцій довільних відстаней.

Нечіткі C-засоби кластеризації - це м'яка версія k -мінів, де кожна точка даних має нечітку ступінь приналежності до кожного кластеру.

Моделі гауссових сумішей , підготовлені алгоритмом максимізації очікування (алгоритм EM), підтримують імовірнісні призначення кластерам замість детермінованих призначень та багатоваріантні гауссові розподіли замість засобів.

k -means ++ вибирає початкові центри таким чином, що дає виправдану верхню межу на об'єкті WCSS.

Алгоритм фільтрації використовує kd-дерева для пришвидшення кожного кроку k- значень.

Деякі методи намагаються пришвидшити кожен к-значення кроку, використовуючи нерівність трикутника .

Уникайте локальної оптими, міняючи точки між кластерами.

Алгоритм кластеризації сферичних к- значень підходить для текстових даних.

Ієрархічні варіанти, такі як поділ k -менів , кластеризація X-засобів та кластеризація G-засобів, неодноразово розбивали кластери для побудови

ієрархії , а також можуть спробувати автоматично визначити оптимальну кількість кластерів у наборі даних .

Заходи щодо внутрішньої оцінки кластерів, такі як силует кластера, можуть бути корисними при визначенні кількості кластерів .

Мінковський зважений k -мереж автоматично обчислює специфічну вагу кластерної ваги, підтримуючи інтуїтивну думку про те, що функція може мати різний ступінь доречності при різних характеристиках. Ці ваги можуть також використовуватися для зміни масштабу даного набору даних, збільшуючи ймовірність оптимізації індексу дійсності кластерів при очікуваній кількості кластерів.

Міні-пакетні k -міні: k -зміна варіантів, використовуючи зразки "міні-партії" для наборів даних, які не вписуються в пам'ять.

Методи ініціалізації

Зазвичай використовуються методи ініціалізації - Forgy і Random Partition. [19] Спосіб Forgy випадковим чином вибирає K спостереження з набору даних і використовує їх в якості початкових коштів. Метод випадкового розділення спочатку випадково призначає кластер кожному спостереженню, а потім переходить до етапу оновлення, обчислюючи тим самим початкове значення як центроїд випадково розподілених точок кластера. Метод Forgy має тенденцію розповсюджувати початкові засоби, тоді як Random Partition поміщає всі вони ближче до центру набору даних. Згідно з Хамеррі та ін., [19] метод випадкового розподілу переважно є кращим для таких алгоритмів, як k-гармонічні засоби та fuzzy k -засоби. Для максимізації очікувань і стандартності k -означає алгоритми, метод Forgy ініціалізації є кращим. Проте всебічне дослідження, проведене Челебі та ін. [20] , виявило, що популярні методи ініціалізації, такі як Forgy, Random Partition і Maximin, часто погано виконуються, тоді як підхід Бредлі та Файяда[20] виконує "послідовно" у "кращій групі" і k -засоби ++ виконує "загалом добре".

1.5 Постановка задачі

Мета й завдання роботи. Метою є прогнозування ходу акцій з найбільшою точністю. Порядок завдань:

- аналіз існуючих математичних моделей для прогнозування ціни акцій.

- розробка моделі.

- проведення розрахунків на реальних даних.

Об'єкт, методи й засоби розроблення. Об’єктом роботи є процес отримання даних про реальні акції; розгляд їх, як часових рядів, аналіз та прогнозування за допомогою економетричної моделі ARIMA.

РОЗДІЛ 2. ТЕОРІЯ ЧАСОВИХ РЯДІВ

2.1 Теорія

На початковому курсі економетрики при вивченні автокореляції розглядаються відношення виду $x_t = px_{t-1} + \varepsilon_t$. Значення деякої економічної змінної залежать від її значень в попередній момент часу, від її значень з лагом – зсувом по часу на один крок назад. Включення змінних з лагом в економетричні відношення означають істотні зміни в "філософії" моделювання. Якщо в звичайних економетричних моделях значення однієї змінної залежать від одночасних значень інших змінних, тобто від поточного стану економічної системи, то присутність змінних з лагом означає, що поведінка системи визначається не тільки її поточним станом, а і траєкторією, котрою система прийшла в цей стан. З математичної точки зору, економетрична модель такого типу представляє собою не функцію пояснюючих змінних, а функціонал від траєкторії (траєкторій) економічних змінних. Тому елементами таких моделей є траєкторії: множини даних $\{x_t, t \in T\}$, де Т – деяка зліченна або континуальна множина. Моделювання залежностей виду $y_t = f(x_t, \varepsilon_t)$, де х і у є даними типами траєкторій, приводить до ситуації, коли звичайні прийоми регресійного аналізу не дають прийнятних оцінок параметрів.

В курсі економетрики ряд змістовних задач приводить до рівнянь, котрі природно називати авторегресійними (autoregressive). Зокрема, при усуненні автокореляції було отримано рівняння $Y_t = \alpha + \beta Y_{t-1} + \gamma X_t + \varepsilon_t$. В ньому в якості пояснюючої змінної з'явилась змінна Y з запізненням, тобто регресія змінної на саму себе – саме цьому таке рівняння і назвали авторегресійним.

Ми будемо називати часовим рядом сукупність спостережень економічної величини в різні моменти часу. При цьому, спостереження може характеризувати економічну величину в даний момент часу, тобто бути типу

запасу (наприклад ціна, ставка відсотку), чи проміжок часу, тобто бути типу потоку (наприклад ВВП, продукція промисловості, надходження податків).

Ми будемо розглядати часовий ряд, як вибірку із послідовності випадкових величин X_t, де t приймає цілі значення від 1 до T. Сукупність випадкових величин ми будемо називати дискретним випадковим або стохастичним процесом.

Іноді кажуть, що стохастичний процес “для кожного випадку” є деякою функцією часу, що дозволяє розглядати процес як випадкову функцію часу $X(t)$. При кожному фіксованому t значення стохастичного процесу розглядається просто як випадкова величина.

Варто задуматися, як ми далі будемо використовувати цю математичну конструкцію? Як правило, є одна єдина реалізація часового ряду, і ясно, що говорити про оцінювання сукупності всіх функцій розподілу взагалі ніколи не доводиться. Поки, на інтуїтивному рівні, якщо процес веде себе так, що його основні статистичні характеристики з часом змінюються, то ми по короткому шматочку наших спостережень взагалі не зможемо нічого сказати про нього. Або нам треба знати щось ще додатково, понад спостережень. Тому, щоб трохи зняти гостроту цієї проблеми, ми будемо говорити про більш вузький клас випадкових процесів, а саме про стаціонарні процеси з нормальним законом розподілу.

2.2 МОДЕЛЬ ARIMA

Основним інструментом економічної науки в даний час є часові ряди. У роботі розглядається комп'ютерна технологія моделювання часових рядів з вираженими коливаннями з використанням розробленого програмного продукту.

Для моделювання використовуються модель тренду і моделі авторегресії і ковзаючого середнього. Як приклад розглядається часовий ряд ціни акції компанії Boeing(BA) за квітень 2018-го року – квітень 2019-го року. При моделюванні використовувалися статистичні процедури, необхідні для ідентифікації та оцінки параметрів моделі і перевірки її адекватності і точності.

Для аналізу поведінки часових рядів з вираженими коливаннями і побудови математичних моделей, описуючих цю поведінку, широко використовують лінійну стохастичну модель авторегресії і ковзаючого середнього ARMA (або авторегресії і проінтегрованого ковзаючого середнього ARIMA). Ця модель зв’язує поточне значення досліджуваної змінної зі значеннями цієї ж змінної в попередні моменти часу, а також з поточним і попередніми значеннями залишків моделі. Дана модель надзвичайно популярна і практика підтвердила її точність і гнучкість. Однак побудова моделі ARMA - складний процес. Його не так просто реалізувати, і потрібно багато практики, щоб оволодіти ім. З метою застосування цієї моделі в практичній діяльності широким колом економістів, трейдерів та інвесторів, а також викладачами і студентами в навчальному процесі представляється доцільним розглянути процес написання програми для побудови моделі ARIMA. Процес побудови моделі включає три етапи: ідентифікацію моделі, оцінювання її параметрів і діагностику моделі.

Ідентифікація моделі здійснюється шляхом визначення та аналізу автокореляційної і часткової автокореляційної функцій (АКФ і ЧАКФ) ряду. Це можливо тільки для стаціонарних часових рядів. На практиці економічні ряди, як правило, є нестаціонарними. Однак у багатьох випадках їх можна звести до стаціонарних часових рядів шляхом виділення тренду або з допомогою переходу до рядів кінцевих різниць. Якщо після виділення тренду ряд став стаціонарним, то для такого ряду будується ARMA модель виду

$$\overline{Y_t} = \sum_{i=1}^{p} a_i Y_{t-i} - \sum_{j=1}^{q} b_j \varepsilon_{t-j} + \varepsilon_t$$

де Y_t, $\overline{Y_t}$ - фактичне і розрахункове значення рівня стаціонарного ряду в момент часу t, a_i – коефіцієнт рівняння авторегресії і -го порядку, b_j – коефіцієнт рівняння ковзаючого середнього j-го порядку, ε_t - відхилення фактичного рівня Y_t від розрахункового $\overline{Y_t}$ рівня ряду в момент часу t.

На етапі ідентифікації здійснюється вибір деякої приватної моделі з усього класу ARMA-моделей, тобто вибір значень p, q і d. Для визначення виду і порядку процесів, породжуючих стаціонарний часовий ряд, використовують апарат автокореляційних функцій: АКФ і ЧАКФ. При цьому в цілому можна виділити три можливих ситуації:

А. Процес авторегресії 1-3 порядку проявляє себе тим, що АКФ експоненціально спадає або представляє суміш синусоїди і спадної експоненти, а ЧАКФ має 1-3 значущих (ненульових) значень, після чого не відрізняється від ЧАКФ білого шуму. Кількість значущих значень ЧАКФ дає порядок авторегресії p.

Б. Процес ковзаючого середнього 1-3 порядку проявляє себе оберненою ситуацією, коли ЧАКФ експоненціально падає або представляє собою суміш синусоїди і спадної експоненти, а АКФ навпаки має одне або кілька значущих значень, після чого не відрізняється від АКФ білого шуму. У цьому випадку кількість ненульових значень АКФ дає порядок ковзаючого середнього q.

В. Суміш двох процесів призводить до появи експонент і синусоїд в обох функціях, однак якщо є дві чисті синусоїди, то це свідчить про наявність процесу першого порядку і для авторегресії, і для ковзаючого середнього. В цілому кількість ненульових значень АКФ і ЧАКФ дає порядок q і p.

Інші форми АКФ і ЧАКФ свідчать про те, що в часовому ряді є тренд (практично не спадна на відрізку 15-20 значень лагу АКФ), або сезонні коливання (АКФ у вигляді практично непадаючої синусоїди), від яких слід позбутися, для того щоб приступити до ідентифікації внутрішніх параметрів ряду.

Вибір порядку d моделі ARIMA ґрунтується на аналізі поведінки АКФ рядів Y_t , ΔY_t^1 , ΔY_t^2. Потрібний порядок досягнутий, якщо АКФ швидко падає.

Однак розглянутий спосіб ідентифікації моделі може використовуватися, якщо часовий ряд є досить довгим (з числом рівнів більше 50). В іншому випадку він може при подальшому аналізі привести до висновку про непридатність ідентифікованої моделі і необхідності заміни її альтернативної моделлю.

На другому етапі проводиться визначення коефіцієнтів моделі.

На третьому етапі застосовуються різні діагностичні процедури для перевірки адекватності вибраної моделі наявними даними. Неадекватності, виявлені в про- процесі такої перевірки, можуть вказати на необхідне коректування моделі, після чого проводиться новий цикл підбору, і т. д. до тих пір, поки не буде отримана задовільна модель.

2.3 Розрахунки АКФ та ЧАКФ

З алгоритмом дій розібрались, приступимо безпосередньо до підрахунку АКФ та ЧАКФ.

Коефіцієнт автокореляції чи часткової автокореляції порядку n також називають коефіцієнтом з лагом рівним n.

Коефіцієнти автокореляції з лагом lag рахуються за такими формулами

Вибіркові середні

$$\overline{y_{2lag-1}} = \frac{\sum_{i=lag+1}^{number} y_i}{number - 1}$$

$$\overline{y_{2lag}} = \frac{\sum_{i=lag+1}^{number} y_{i-lag}}{number - 1}$$

Коефіцієнт

$$r_{lag} = \frac{\sum_{i=lag+1}^{number} \left(y_i - \overline{y_{2lag-1}}\right)\left(y_{i-lag} - \overline{y_{2lag}}\right)}{\sqrt{\sum_{i=lag+1}^{number} \left(y_i - \overline{y_{2lag-1}}\right)^2 \sum_{i=lag+1}^{number} \left(y_{i-lag} - \overline{y_{2lag}}\right)^2}}$$

Початковий часовий ряд скоріше за все, не є стаціонарним. Тому поняття автокореляційної та часткової автокореляційної функцій для нього не дуже коректне, але все ж, порахувати АКФ та ЧАКФ ми можемо. Для того, щоб зробити ряд стаціонарним, перейдемо до ряду кінцевих різниць, для цього зробимо таку процедуру

$$y_i = x_i - x_{i-1};\ i = \overline{number, 2;}$$

$$x_1 = 0;$$

При великій кількості даних, втрата першого значення не є проблемою.

Будемо робити цю процедуру до отримання стаціонарного ряду, на що вкаже АКФ та ЧАКФ. Зазвичай достатньо перейти до 1-го та 2-го порядків

рядів, та прорахувати значення АКФ та ЧАКФ при значеннях лагу 15-20, та ми для наочності зробимо це для рядів до 5-го порядку включно та прорахуємо при 100 лагах.

Наостанок, значущі коефіцієнти АКФ та ЧАКФ рахуються за формулою

$$coef = \pm \frac{2}{\sqrt{number}}$$

Для ряду ціни акції компанії Boeing за один рік, з проміжком часу рівним один день, отримали $coef = 0.125245$.

2.4 Оцінка параметрів моделі та побудова

Загальний вигляд моделі ARIMA

$$y_t = \sum_{i=1}^{p} a_i y_{t-i} + \sum_{j=1}^{q} b_j \varepsilon_{t-j} + \varepsilon_t$$

Визначивши коефіцієнти p і d рівними одиниці, будуємо модель виду

$$y_t = a y_{t-1} + \mathrm{b}\varepsilon_{t-1}$$

Оцінку параметрів a і b будемо проводити методом найменших квадратів. З рівності

$$\binom{a}{b} = (Q^T Q)^{-1} Q^T y,$$

$$Q = \begin{pmatrix} y_0 & 1 \\ y_1 & 1 \\ \dots & \dots \\ y_{n-2} & 1 \end{pmatrix}$$

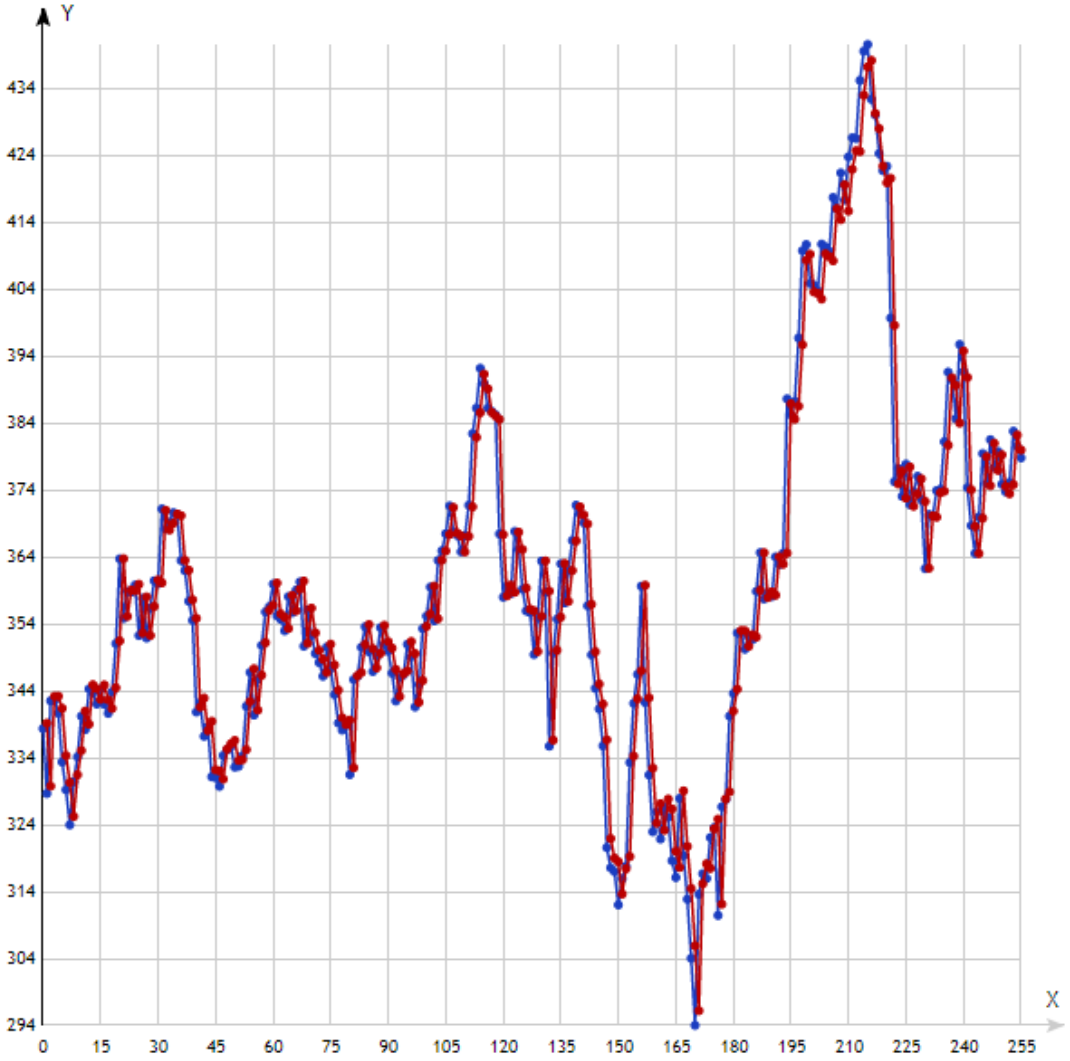
$$y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_{n-1} \end{pmatrix}$$

отримуємо значення параметрів.

В нашому прикладі отримали таку модель:

$$y_t = 0.9689 y_{t-1} + 11.335$$

Побудували графік для того, щоб бути впевненими, що модель працює.



На рисунку синьою лінією позначено справжню ціну, червоним – розраховану. Бачимо що похибка мінімальна, графіки майже накладаються один на одного.

Трішки погравшись з вхідними даними про акцію, шляхом проб було встановлено, що найкращий прогноз робиться на один крок, тобто на один наступний день. З боку трейдингу, прогноз краще всього застосовувати, коли в акції звичайний день, без новин на компанію та без підвищених об’ємів торгів.

Основна проблема моделі ARIMA в її неуніверсальності і необхідності визначення потрібних коефіцієнтів для кожного часового ряду окремо.

РОЗДІЛ 3. ПІДСИСТЕМА ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ФІНАНСОВИХ ДАНИХ

3.1 Статистичний аналіз ціни акцій

Для тестування було взято дані про ціну акцій і кількість проданих акцій на біржі. Дані взято впродовж року, з інтервалом 5 хвилин.

На рис. 1 наведено графік зміни цін акцій

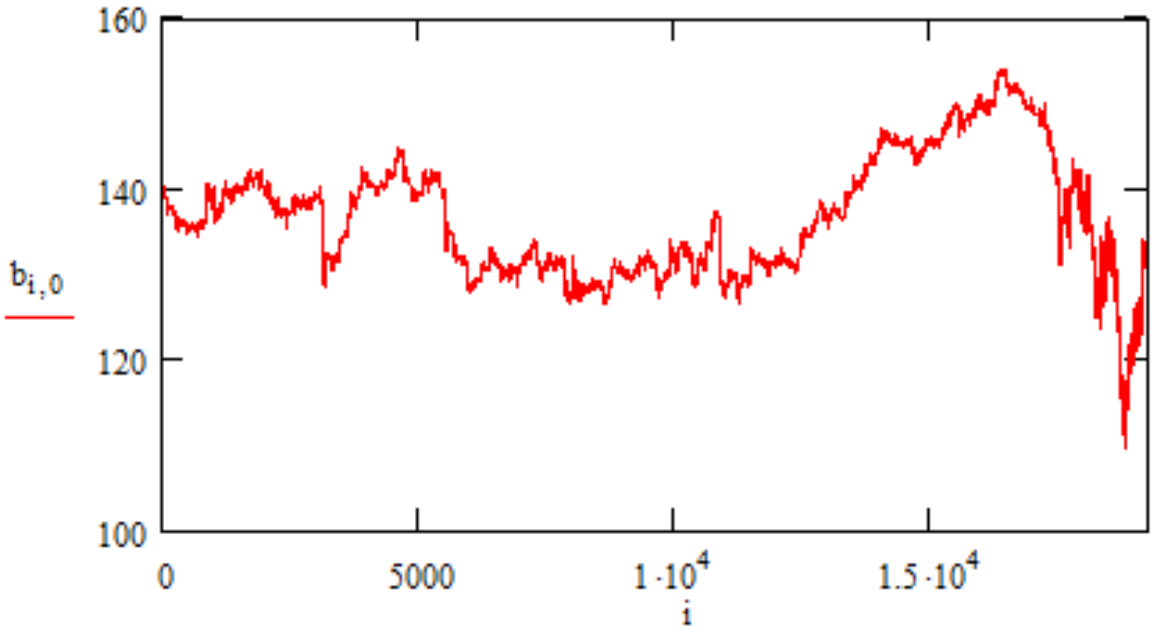


Рис. 1. Ціни акцій

На рис. 2 наведено графік кількості проданих акцій

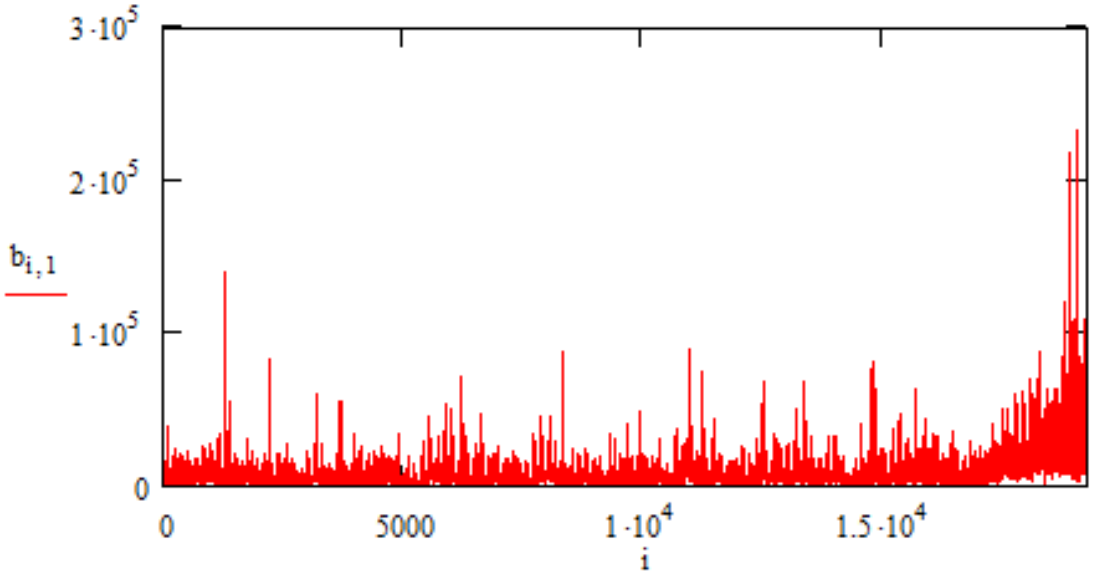


Рис. 2. Кількість проданих акцій

На рис. 3 наведено графік приростів ціни акцій

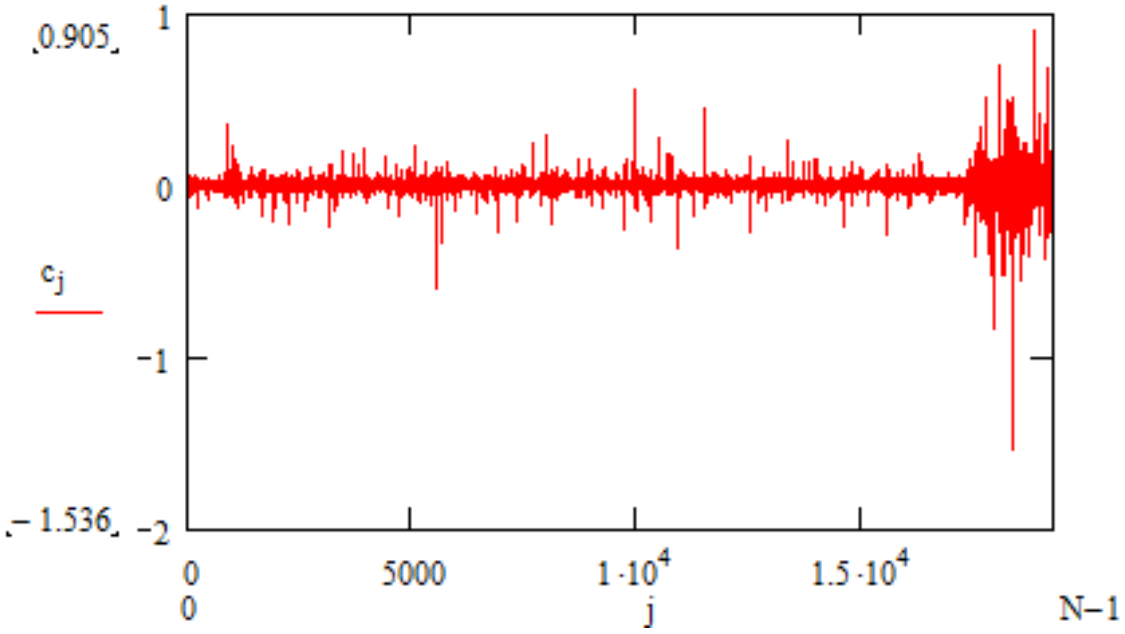


Рис. 3. Прирости ціни акцій

На рис. 4 наведено графік приростів кількості акцій

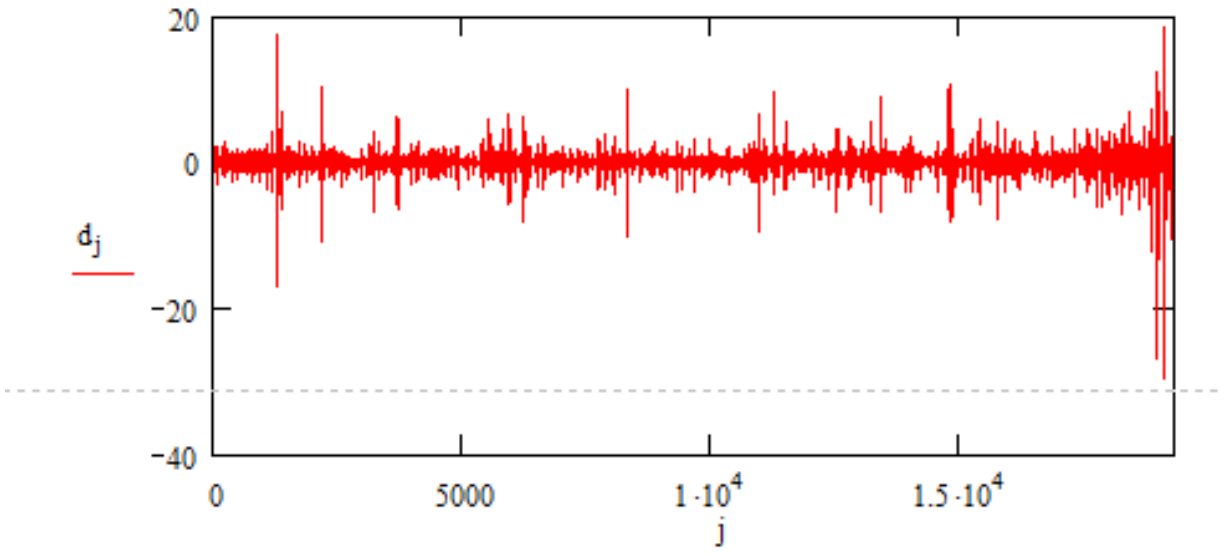


Рис. 4. Прирости кількості акцій

Для оцінювання характеру поведінки та виявлення періодичних складових цін акцій та кількість проданих використовувались оцінки кореляційної функції.

На рис. 5 наведено графік кореляційної функції цін та кількості акцій

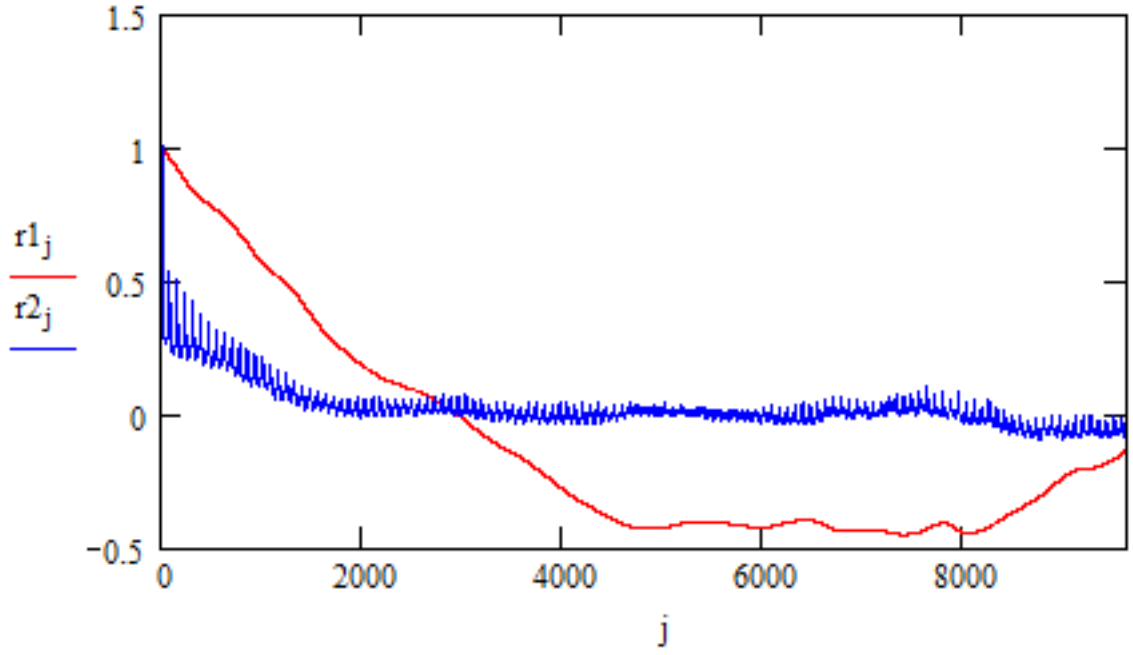


Рис. 5. Оцінка кореляційної функції

На рис. 6 наведено графік кореляційної функції для приростів цін та кількості акцій

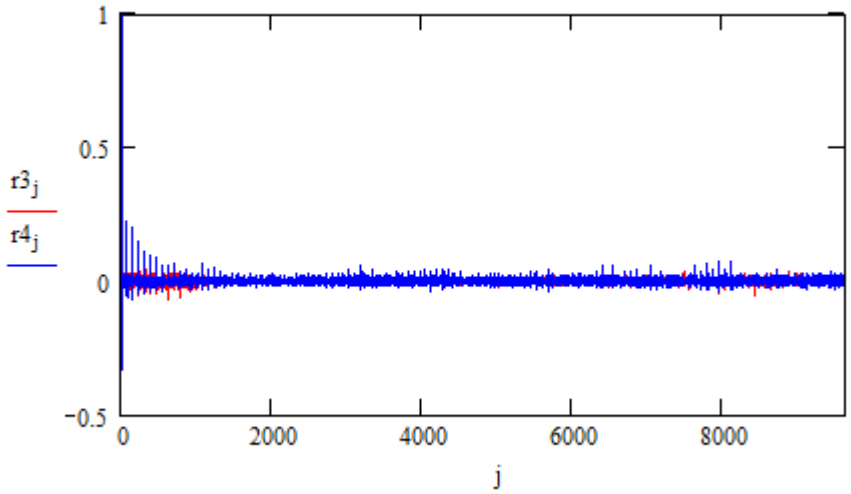


Рис. 6. Кореляційна функція приростів

Аналіз кореляційних функцій показав, що кількість проданих акцій і їх кількість не містять періодичних складових.

А їх прирости є незалежними.

Коефіцієнт кореляції між цінами на акції та кількістю проданих акцій рівний 0.151. Тобто, між цими параметрами немає лінійної залежності. А це свідчить, що ціна акцій не впливає на їх реалізацію.

Результати перетворення Фур'є для приростів ціни акцій та кількості акцій зображені на рис. 7.

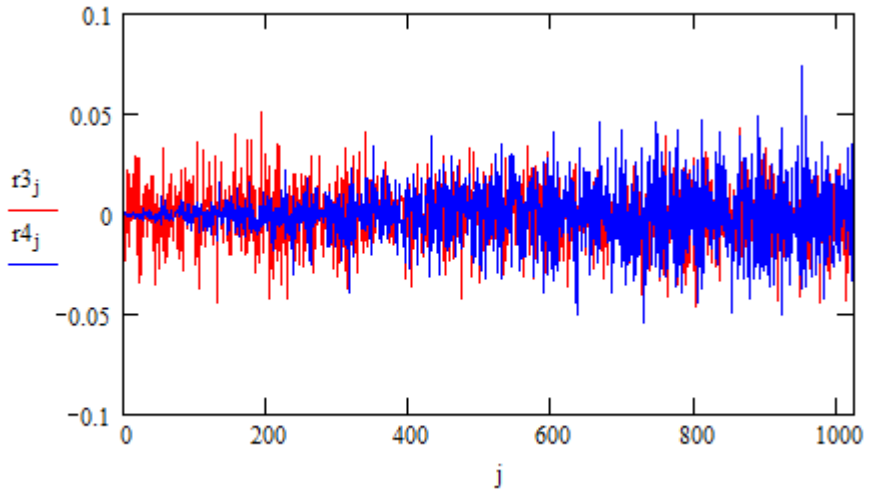


Рис. 7. Спектральний аналіз для приростів ціни акцій (червоний) та кількості акцій (синій).

Результати спектрального аналізу вказують на те, що прирости цін на акції та прирости кількості куплених акцій являють собою «білий шум», а отже, для їх дослідження можна використати стохастичні моделі типу Хестона [14].

3.2 Модель Хестона

При дослідженнях стохастичних моделей доцільним є використання методів статистичного моделювання для побудови реалізацій випадкових величин і випадкових процесів. При використанні сучасних обчислювальних та інформаційних технологій можна будувати обчислювальні експерименти для дослідження поведінки необхідних процесів, як окремих реалізацій, так і в середньому. В роботі продовжуються дослідження [15-16].

Класична модель Блека-Шоулза має вигляд

$$dS(t) = \mu S(t)dt + \sqrt{\sigma} S(t)dW(t),$$

де μ - очікувана дохідність, σ - волатильність, $W(t)$ - стандартний Вінерівський процес, $S(t)$ - ціна акцій.

Щоб врахувати зміни волатильності з плином часу модель Блека-Шоулза була узагальнена.

Один зваріантів – це модель волатильності Хестона

$$\begin{aligned} dS(t) &= \mu S(t)dt + \sqrt{\sigma(t)} S(t)dW_1(t), \quad S(0) = S_0, \\ d\sigma(t) &= k\big(\theta - \sigma(t)\big)dt + \varepsilon\sqrt{\sigma(t)}dW_2(t), \quad \sigma(0) = \sigma_0, \end{aligned} \qquad (1)$$

де $\mu, k, \theta, \varepsilon$ - деякі константи, $\sigma(t)$ - волатильність, $W_1(t), W_2(t)$ - стандартні Вінерівські процеси, $S(t)$ - ціна акцій.

При $k = 0, \varepsilon = 0$ модель Хестона (1) співпадає з моделлю Блека-Шоулза. В моделі Хестона можна розглядати ситуацію, коли Вінерівські процеси $W_1(t)$ та $W_2(t)$ незалежні. Якщо Вінерівські процеси з коефіцієнтом кореляції $\rho \in [-1,1]$, то

процес $W_2(t)$ можна зобразити у вигляді $W_2(t) = \rho W_1(t) + \sqrt{1-\rho^2} W_3(t)$, де $W_1(t)$ та $W_3(t)$ незалежні Вінерівські процеси.

При моделюванні система стохастичних диференціальних рівнянь записується у вигляді системи різницевих рівнянь

$$S(t+\Delta t) = S(t) + \mu S(t)\Delta t + \sqrt{\sigma(t)} S(t)\left(W_1(t+\Delta t) - W_1(t)\right) \quad (3)$$

$$\sigma(t+\Delta t) = \sigma(t) + k\left(\theta - \sigma(t)\right)\Delta t + \varepsilon\sqrt{\sigma(t)}\left(W_2(t+\Delta t) - W_2(t)\right) \quad (3)$$

Алгоритм моделювання.

1. Моделюємо початкові значення σ_0 та S_0.
2. Для заданих точності $\delta > 0$ і надійності $0 < \alpha < 1$ знаходимо кількість доданків M у зображенні
   $W(t, M) = t\eta_0 + \sqrt{2}\sum_{i=1}^{M} \frac{\sin(i\pi t)}{i\pi}\eta_i$.
3. Моделюємо послідовності випадкових величин $\{\eta_i\}_{i=0}^{M}$ з нормальним розподілом $\mathrm{N}(0,1)$.
4. Для заданого кроку Δt моделюємо обчислюємо значення ціни за формулою (2), за формулою (3) обчислюємо значення волатильності.
5. Крок 4 повторюємо необхідне число разів.

Представлена модель дозволяє проводити розрахунки для одного пакету акцій. Але її можна використовувати і для оцінки портфелю акцій. В подальшому дослідження будуть присвячені моделі, коли ціни акцій можуть мати стрибки в короткі проміжки часу.

ВИСНОВКИ

В ході виконання кваліфікаційної роботи було виконано:

1. Проведено аналіз ефективності статистичних моделей інтелектуального аналізу даних.
2. Проведено аналіз методів для моделювання часових рядів та використання моделі ARIMA для вирішення прикладних задач.
3. Створено підсистему інтелектуального аналізу фінансових даних.
4. Поглиблено теоретичні та прикладні знання у напрямку інтелектуальний аналіз даних.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Канторович Г.Г. Анализ временных рядов
2. Юров В.М. Моделирование нестационарных временных рядов с выраженными колебаниями с использованием инструментов Excel
3. Дані про акції https://finance.yahoo.com/
4. Нормальний закон розподілу https://math.semestr.ru/group/prim2.php
5. Онлайн побудова графіків http://yotx.ru
6. Семененко М.Г МОДЕЛЬ МАРКОВИЦА: МАТЕМАТИЧЕСКИЕ АСПЕКТЫ И КОМПЬЮТЕРНАЯ РЕАЛИЗАЦИЯ
7. Инвестиционное дело: учеб./В.М. Аскинадзи, В,Ф. Максимова, В.С. Петров.- М.:Маркет ДС, 2008. С .512
8. Портфельные инвестиции: учебн. пособие/ Максимова В.Ф Московский государственный университет экономики, статистики и информатики.- М.:МЭСИ,2006.-С.54
9. Инвестиции. учеб/Уильям Ф.Шарп, Гордон Дж.Александер, Джефри В Бейли. – М.,2008. С.475
10. Инвестиции :учеб./А.Ю.Андрианов, С.В.Валдацев, П.В.Воробёв.-2-е изд.- М.6ТК Велби,Изд-во Проспект,2008.- С .584
11. Инвестиции: Учеб. пособие./Игонина Л.Л.- М.: Юристъ, 2006. С. 350
12. Банковское дело: Учебник / под ред. Г.Г.Коробовой. – М.: Экономистъ, 2007. С. 420
13. Инвестиции:учебн .пособие/Янковский К.П.-СПБ.: Питер, 2008.-.С. 368
14. Пашко А.О. Аналіз динаміки ціни акцій в моделі Хестона / А.О. Пашко// Матеріали V міжнародної науково-практичної конференції "Інформаційні

технології в культурі, мистецтві, освіті, науці, економіці та бізнесі", 22-23 квітня 2020 р. – Київ: Видавничий центр КНУКіМ, 2020. - С.136-137.

15. Пашко А. Статистичне моделювання стохастичних фінансових процесів. Міжнародна науково-практична конференція "Інформаційні технології в культурі, мистецтві, освіті, науці, економіці та бізнесі", 19-20 квітня 2017р. Київ: Видавничий центр КНУКіМ, 2017. Ч.1. С. 182-184.

16. Pashko A. Methods of statistical modeling of stochastic differential equation solutions. Ukrainian Conference on Applied Mathematics. 28-30 September 2017. Lviv. P. 88-89.