Міністерство освіти і науки України
Національний університет «Києво-Могилянська академія»
Факультет інформатики
Кафедра математики

Курсова робота
освітній ступінь – бакалавр
на тему: «Застосування методу Грейнджера для визначення причинно-наслідкових зв’язків»

Керівник курсової роботи
Старший викладач,
кандидат фізико-математичних наук
Дрінь Світлана Сергіївна
_____ (підпис)
Виконала студентка 3-го року навчання,
напрямку підготовки
Прикладна математика
Яковенко Катерина Віталіївна
Рецензент _____
Кваліфікаційна робота захищена
з оцінкою: _____
Секретар ЕК_____
«____» _____2021р.

Київ – 2022

Тема: Застосування методу Грейнджера для визначення причинно-наслідкових зв’язків

Календарний план виконання роботи:

№ п/п	Назва етапу курсового проекту (роботи)	Термін виконання етапу	Примітка
1.	Отримання завдання на курсову роботу.	13.10.2021	
2.	Огляд технічної літератури за темою роботи.	29.01.2022	
3.	Опрацювання матеріалів.	05.02.2022	
4.	Погодження плану робіт з керівником.	17.02.2022	
5.	Аналіз знайденої літератури.	25.02.2022	
6.	Реалізація методу Грейнджера та аналіз отриманих результатів.	01.05.2022	
7.	Остаточне оформлення роботи.	29.05.2022	
8.	Захист роботи.	14.06.2022	

Студент _____

Керівник _____

“_____”_____

ЗМІСТ

АНОТАЦІЯ **4**

ВСТУП **5**

1. КЛАЙВ ГРЕЙНДЖЕР **7**

2. ЗАГАЛЬНІ ВІДОМОСТІ **9**

3. ФАЛЬШИВА РЕГРЕСІЯ **11**

4. ПРИЧИННІСТЬ ГРЕЙНДЖЕРА **14**

5. ВИЗНАЧЕННЯ СТАТИСТИЧНОЇ НЕСТАЦІОНАРНОСТІ **15**

6. АЛГОРИТМ ПЕРЕТВОРЕННЯ НЕСТАЦІОНАРНИХ ДАНИХ У СТАЦІОНАРНІ **18**

7. МАТЕТИЧНЕ ФОРМУЛЮВАННЯ **22**

8. ВИЗНАЧЕННЯ ПРИЧИННОСТІ ЗА ГРЕЙНДЖЕРОМ **24**

9. ВИСНОВКИ **32**

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ **33**

АНОТАЦІЯ

У даній курсовій роботі розглядається можливість прогнозування значення однієї змінної на основі лагових значень іншої змінної часового ряду за допомогою тесту Грейнджера. А також визначення стаціонарності даних часових рядів, шляхом застосування критерію Дікі-Фуллера; з’ясування методів видалення нестаціонарності.

ВСТУП

Актуальність теми. Тема визначення причинно-наслідкового зв’язку займала філософів понад дві тисячі років і досі не розв’язана. Це глибоке заплутане питання з багатьма можливими відповідями, які не всіх задовольняють. Дослідження причин тих чи інших явищ має певне значення для розуміння їхньої сутності, закономірностей розвитку. А визначення причини – означає, отримати відповідь на запитання: чому це явище відбулось, чи воно було спровоковане та викликане або ж, що лежить в основі отриманого наслідку. Адже неможливо пізнати явище, не з’ясувавши його причину. Але ж встановити причину явища – нелегка справа, причина зазвичай не лежить на поверхні. Саме тому дуже багато науковців різних галузей займаються питанням встановлення причинно-наслідкового зв’язу. Так й професор Клайв Вільям Джон Грейнджер, лауреат Нобелівської премії з економіки 2003 року, розробив концепцію причинно-наслідкового зв’язку для покращення ефективності прогнозування, названу на його честь “Тест на визначення причинності за Грейнджером”. А також і Урс Шрайбер у 2000 році ввів концепцію ентропії переносу, яка є версією взаємної інформації, що оперує умовними ймовірностями. Дана концепція призначена для виявлення спрямованого обміну інформацією між двома змінними, зумовленими загальною історією та вхідними параметрами.

Останніми роками зростає інтерес до використання причинності Грейнджера для виявлення причинно-наслідкових взаємодій у нейронних даних. Наприклад, Бернасконі та Коніг у 1999 році застосували спектральні заходи Гевеке для опису причинно-наслідкових взаємодій між різними областями зорової кори кішки. У сфері функціональності МРТ у 2005 році застосували причинність Грейнджера до даних, отриманих під час

складного зорово-моторного завдання. Причинність Грейнджера також була застосована до змодельованих нейронних систем, щоб дослідити зв’язок між нейроанатомією, мережевою динамікою та поведінкою.

Мета і завдання дослідження. Метою є визначення максимальної кількості лагових значень однієї змінної, які впливають на прогнозування іншої змінної часового ряду. Завданнями є визначити стаціонарність даних, у випадку нестаціонарності, видалити її та привести дані до стаціонарного виду, шляхом застосування різницевого методу та логарифмування (за потребою); перевірити стаціонарність, застосувавши критерій Дікі-Фуллера; застосувати тест Грейнджера; з’ясувати кількість лагових змінних для прогнозування.

Об’єкт дослідження. Дані, які є часовим рядом.

Методи дослідження. Логарфмування даних; різницевий метод; критерій Дікі-Фуллера; тест Грейнджера.

1. КЛАЙВ ГРЕЙНДЖЕР

Клайв Вільям Джон Грейнджер – англійський економіст, лауреат Нобелівської премії з економіки у 2003 році «за розробку методів аналізу економічних часових рядів з загальними трендами»

Нижче наведено особистий звіт про розвиток методу причинно-наслідкового зв'язку за Грейнджером, люб'язно наданий професором Клайвом Грейнджером.

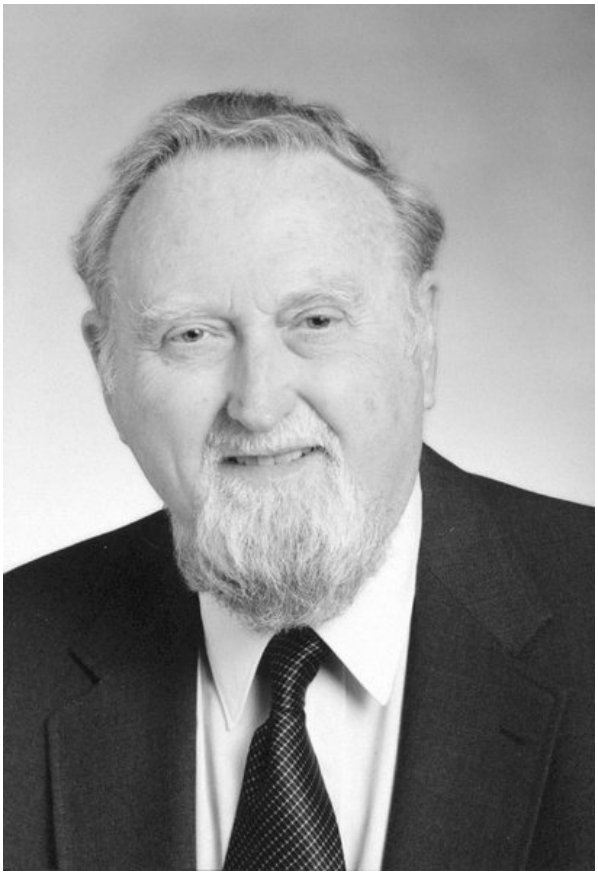

Рисунок 1 "Клайв Вільям Джон Грейнджер"

"На початку 1960-х років я розглядав пару пов'язаних стохастичних процесів, які були чітко взаємопов'язані, і я хотів знати, чи можна розбити цей зв'язок на пару односторонніх зв'язків. Мені запропонували

подивитися на визначення причинності, запропоноване дуже відомим математиком Норбертом Вінером, тому я адаптував це визначення в практичну форму та обговорив його”.

Прикладні економісти визначили, що причинність Грейнджера є досить зрозумілою та має право на існування, саме тому зовсім скоро почали з’являтись роботи із застосуванням даного методу.

Проте декілька письменників заявили, що “звісно, що даний метод визначення зв’язку це не є справжньою причинністю, а лише причинність Грейнджера”. Таким чином, було введено термін “причинність Грейнджера” для використання даного методу, щоб відрізняти його від інших способів (методів) визначення причинно-наслідкового зв’язку між змінними.

2. ЗАГАЛЬНІ ВІДОМОСТІ

Метод Грейнджера для визначення причинно-наслідкових зв’язків застосовується для часових рядів.

Часовий ряд — це набір спостережень, елементами якого є значення, які зібрані в однаковий проміжок часу. Інтервал часу, за який збираються дані, може бути погодинним, щоденним, щотижневим, місячним, квартальним, щорічним тощо.

Дані часових рядів можна розділити на такі категорії:

- Однофакторний часовий ряд – складається з однієї змінної, яка змінюється з часом. Наприклад, баланс рахунку протягом десяти років.
- Багатоваріативний часовий ряд – складається з більш ніж однієї змінної, яка змінюється з часом, а також може існувати залежність між змінними. Наприклад, різні витрати сім'ї протягом десяти років.

Аналіз часових рядів дає уявлення про закономірність або характеристики даних часових рядів. Дані часових рядів можна розкласти декілька компанентів:

- Тренд – показує тенденцію даних протягом тривалого періоду часу; тренд може бути зростаючим або спадним.
- Сезонність – це коливання, яке відбувається періодично і повторюється з деяким часовим періодом.
- Випадковість – це фактор, який визначає результат експерименту з множини можливих результатів, відомих заздалегідь. Тобто вважається, що експеримент має випадковий результат, якщо його кінцевий результат не може бути визначений наперед з множини можливих варіантів.

У режимі реального часу більшість даних складається з кількох змінних, в яких незалежні змінні можуть залежати від інших незалежних змінних, ці відносини впливатимуть на передбачення або прогнозування.

3. ФАЛЬШИВА РЕГРЕСІЯ

Лінійна регресія може вказувати на сильний зв’язок між двома чи більше змінними, але насправді ці змінні можуть бути абсолютно не пов’язаними між собою. Коли справа доходить до знання предметної області, передбачення не вдаються, цей сценарій відомий як помилкова регресія.

На наведеному нижче графіку можна спостерігати сильний зв’язок між двома процесами: споживанням курятини (червона лінія) та експортом сирої олії (чорна лінія). Хоча досить логічним є те, що дані процеси між собою ніяк не пов’язані.

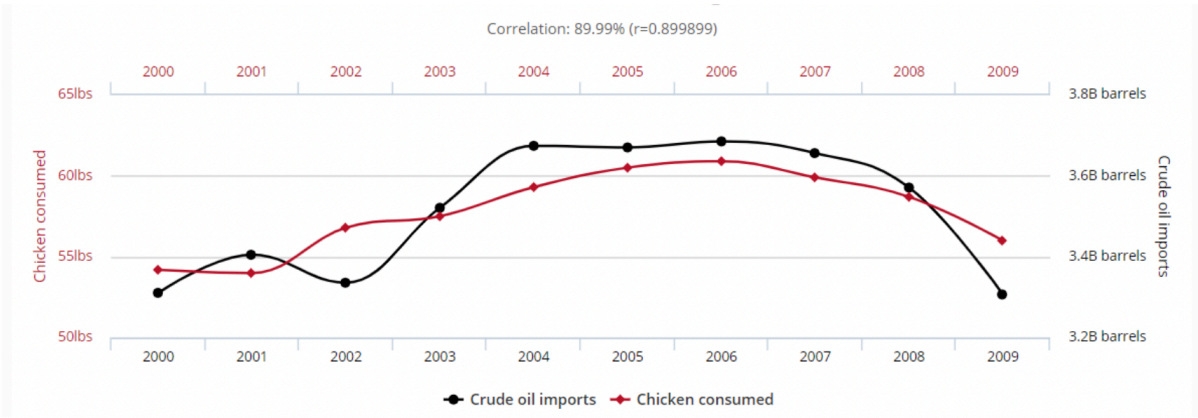

Рисунок 2 "Зв'язок між двома процесами"

На даному графіку можна помітити сильний тренд або нестаціонарність та вищий R^2 (коефіцієнт кореляції, що дорівнює 0,899899), які спостерігаються при помилковій регресії. Під час побудови моделі необхідно усунути помилкову регресію, тобто потрібно перевірити, чи взагалі пов’язані між собою дані, з якими необхідно буде надалі працювати. Це необхідно зробити, оскільки дані можуть бути абсолютно не пов’язані між собою, а звідси і не матимуть причинно-наслідкового зв’язку.

Багатофакторна регресія використовує кореляційну матрицю для перевірки залежності між усіма незалежними змінними. Якщо значення коефіцієнта кореляції між двома змінними високе, будь-яка одна змінна зберігається, а інша відкидається, щоб усунути залежність. У наборі даних, коли час теж є фактором, багатофакторна регресія не вдається, оскільки коефіцієнт кореляції, який використовується для усунення змінної, не прив'язаний до часу, а лише дає кореляцію між двома змінними.

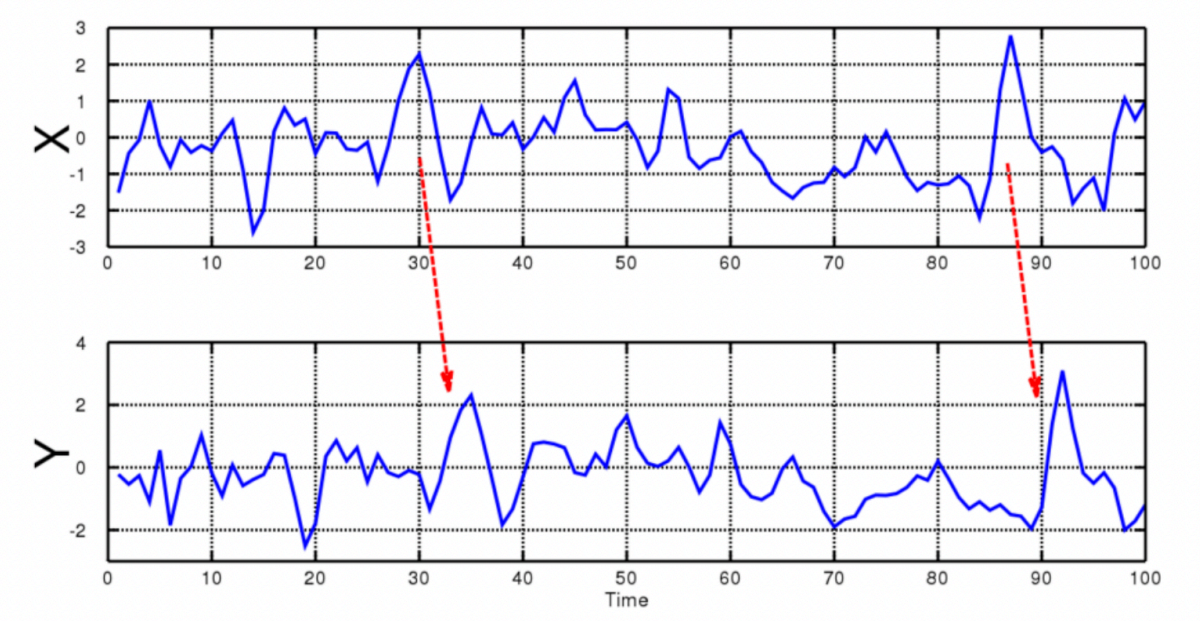

Рисунок 3 "Графіки часових рядів"

Враховуючи наведений вище графік часових рядів, змінна X має безпосередній вплив на змінну Y, але між X і Y є відставання в 5, а це означає, що в даному випадку час є фактором, що плутає, де коефіцієнт кореляції не прив'язаний до часу, а лише дає кореляцію між двома змінними, тому в цьому випадку ми не можемо використовувати кореляційну матрицю.

Розглянемо також випадок для двох змінних, які є залежні один від одного. До прикладу візьмемо за першу змінну – збільшення кількості хворих на коронавірус у місті, а за другу – збільшення кількості госпіталізованих. Для кращого прогнозування необхідно знати, чи існує причинно-наслідковий зв’язок. Саме в таких випадках причинність за Грейнджером приходить на допомогу.

4. ПРИЧИННІСТЬ ГРЕЙНДЖЕРА

Професор Клайв Вільям Джон Грейнджер, лауреат Нобелівської премії з економіки 2003 року, розробив концепцію причинно-наслідкового зв’язку для покращення ефективності прогнозування.

Причинність за Грейнджером – це статистичне поняття причинності, яке ґрунтується на передбаченні. Відповідно до причинно-наслідкового зв’язку за Грейнджером - фактор X_1 “спричиняє” вплив на фактор X_2. Це означає, що значення фактора X_1 за попередній період (тобто значення $X_{1_{i-\tau}}$) повинна містити інформацію, яка допомагає передбачити значення фактора X_2. (тобто значення X_{2_i}).

Його математичне формулювання засноване на лінійному регресійному моделюванні випадкових процесів.

Причинність Грейнджера або причинність за Грейнджером була розроблена в 1960-х роках і широко використовується в економіці з 1960-х років.

По суті, це економетричний гіпотетичний тест для перевірки використання однієї змінної для прогнозування іншої в багатовимірних часових рядах із певним лагом.

Лаг — показник, що відображає відставання в часі одного явища (наприклад, економічного) порівняно з іншим, пов'язаним з ним явищем.

Необхідною умовою для виконання тесту на причинність Грейнджера є те, що дані повинні бути стаціонарними (не залежати від часу), тобто вони повинні мати стале середнє значення, сталу дисперсію і не мати сезонного компонента. Якщо ж дані є нестаціонарними, тоді їх необхідно перетворити в стаціонарні.

5. ВИЗНАЧЕННЯ СТАТИСТИЧНОЇ НЕСТАЦІОНАРНОСТІ

У світі статистики існує безліч способів, щоб перевірити стаціонарність даних. Доповнений тест Дікі-Фулера є, певно, одним з найпопулярніших серед тих, які визначають, чи є дані стаціонарні чи ні. Далі розглянемо застосування даного тесту на прикладі.

Для початку необхідно визначити нульову та альтернативну гіпотези:

Нульова гіпотеза: розподіл нестаціонарний;

Альтернативна гіпотеза: розподіл стаціонарний.

Результатом даного тесту є значення "p". Якщо p менше за критичний поріг 0,05, тоді необхідно відхилити нульову гіпотезу і прийняти альтернативну, а саме, що ряд є стаціонарним. Якщо ж p більше за критичний поріг 0,05, тоді неможливо відхилити нульову гіпотезу, а, отже, прийняти її значення, що ряд є нестаціонарним.

Повний тест досить зручно реалізований як функція "*adfuller*" у пакеті statsmodel, який надає класи та функції для оцінки багатьох різних статистичних моделей мовою програмування Python.

Для початку, застосуємо дану функцію на розподілі, про який точно відомо, що його дані стаціонарні. Це необхідно для того, аби запевнитись, що все працює вірно, а також для ознайомлення з результатами тесту. Для визначення результату застосуємо функцію "*adfuller*" до набору даних "diamonds", колонки "price" та виведемо результат.

Функція:

$$test_results = adfuller(diamonds["price"])$$

Результат:

$ADF\ test\ statistic: -8.1149306683156$

$p - value: 1.1980457313376738e - 12$

Результат даного тесту показує, що значення p майже нульове, а це означає, що з легкістю можна відкинути нульову гіпотезу, та прийняти, що дані є стаціонарними. Що, власне, і потрібно було довести.

Далі, застосуємо тест до деяких фінансових даних, які швидше за все, будуть нестаціонарними. Для початку виведемо графік даних, які маємо.

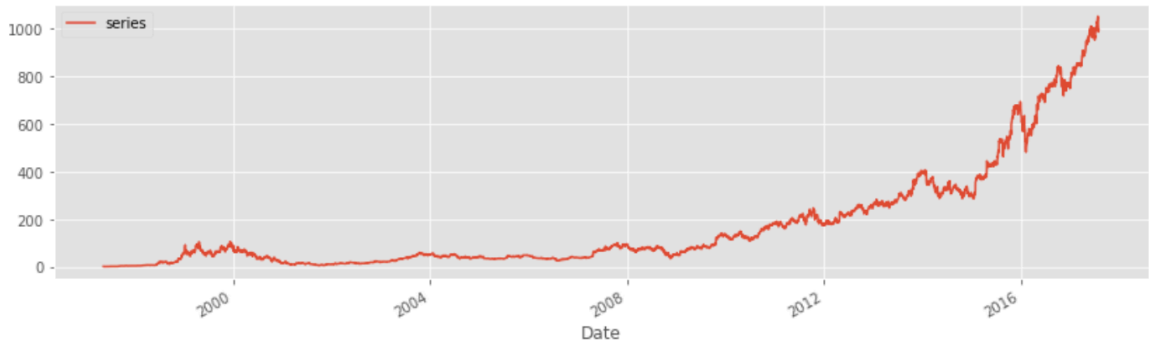

Рисунок 4 "Графік фінансових даних"

На наведеному вище графіку видно, серія демонструє чітку тенденцію до зростання, що свідчить про нестаціонарність даних. Виконаємо тест Дікі-Фуллера:

Функція:

$ts1_results = adfuller(ts1["series"])$

Результат:

1.0

Після виконання результати тесту показують, що значення $p = 1$, тобто дані точно стаціонарні.

6. АЛГОРИТМ ПЕРЕТВОРЕННЯ НЕСТАЦІОНАРНИХ ДАНИХ У СТАЦІОНАРНІ

Одним із методів перетворення найпростіших нестаціонарних даних є різницевий метод. Суть даного методу полягає в тому, щоб отримати різницю послідовних спостережень.

Для більш детального ознайомлення з різницевим методом, опишемо код алгоритму мовою програмування Python із застосуванням бібліотеки Pandas. Pandas - програмна бібліотека, написана для мови програмування Python для маніпулювання даними та їхнього аналізу. Дана бібліотека містить функцію віднімання "*diff*", якою і необхідно скористуватись.

Візьмемо декілька значень часового ряду:

Таблиця 1 "Значення часового ряду"

Date	Series
2012-08-06	26.107651
2012-08-07	26.377876
2012-08-08	26.438896
2012-08-09	26.587088
2012-08-10	26.517351
2012-08-13	26.491201

Для вище зазначених значень застосуємо функцію "diff":

```
ts2["diff_1"] = ts2["series"].diff(periods=1)
ts2["diff_2"] = ts2["series"].diff(periods=2)
ts2["diff_3"] = ts2["series"].diff(periods=3)
```

де $diff_1_{i+1}=series_{i+1}- series_i$ - різниця 1-го порядку;

$diff_2_{i+2}=series_{i+2}- series_i$ - різниця 2-го порядку;

$diff_3_{i+3}=series_{i+3}- series_i$ - різниця 3-го порядку;

Після чого отримаємо наступну таблицю для шести значень:

Таблиця 2 "Значення після застосування різницевого методу"

Date	Series	diff_1	diff_2	diff_3
2012-08-06	26.107651	NaN	NaN	NaN
2012-08-07	26.377876	0.270225	NaN	NaN
2012-08-08	26.438896	0.061020	0.331245	NaN
2012-08-09	26.587088	0.148192	0.209212	0.479437
2012-08-10	26.517351	-0.069737	0.078455	0.139475
2012-08-13	26.491201	-0.026150	-0.095887	0.052305

Таким чином, а саме різницевим методом, дані були перетворені в стаціонарні. Для простих розподілів достатньо взяти різницю першого порядку, щоб досягнути стаціонарності даних.

Аби запевнитись, що отримані дані є дійсно стаціонарними, застосуємо функцію Дікі-Фулера для перевірки на стаціонарність даних. Функція "*adfuller*" для *diff_1* (різниця першого порядку *ts2*).

Перед проведенням тесту, необхідно визначити нульову та альтернативну гіпотези:

Нульова гіпотеза: розподіл стаціонарний, не залежить від часу.

Альтернативна гіпотеза: розподіл нестаціонарний, залежить від часу.

Функція:

$results = adfuller(ts2["diff_1"].dropna())$

Результат:

0.0

де *dropna()* – метод, який видаляє рядки, які містять порожні значення

Після виконання функції “*adfuller*” для перевірки, чи отримані дані після застосування різницевого методу стали стаціонарними, у вихідному розподілі *ts2*, значення p дорівнює 0. А це означає, необхідно відхилити нульову гіпотезу, а саме зробити висновок, що ряд є стаціонарним.

Однак з деякими розподілами справа йде не так просто, як з простими, що були наведені вище. Розглянемо наступний приклад з нелінійним зростанням.

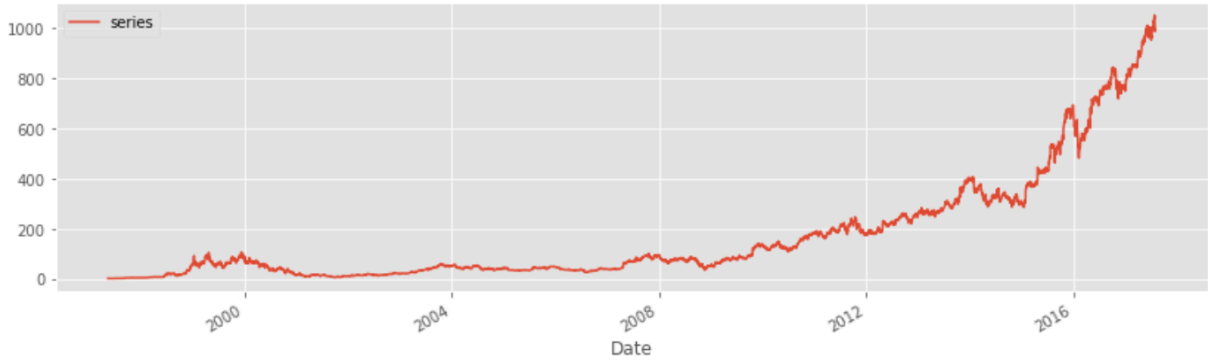

Рисунок 5 "Графік з нелінійним зростанням"

На вище наведеному графіку очевидно помітити нелінійний тренд, що свідчить про те, що дані точно є нестаціонарними. Для того, щоб перетворити нестаціонарні дані в стаціонарні в такому випадку, необхідно прологарифмувати дані, а вже затим використати різницевий метод. Різницевий метод можна використовувати декілька разів, якщо в цьому є потреба, а ось логарифмувати дані потрібно лише раз. Щоб усунути нелінійність, використаємо логарифмічну функцію “*np.log*” і після чого застосуємо різницевий метод першого порядку для класу:

Функція:

$$transformed_ts1 = pd.Series(np.log(ts1["series"])).diff().dropna()$$

де pd – бібліотека Pandas, Series – клас.

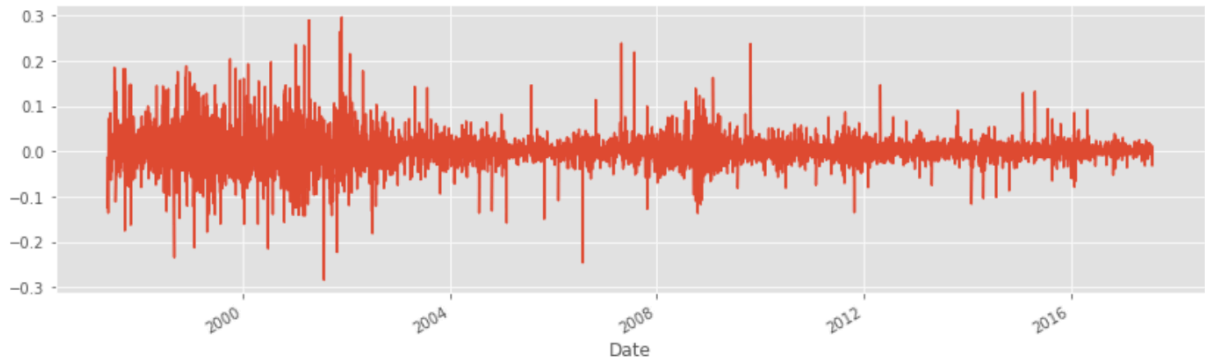

Рисунок 6 "Графік даних після логарифмування та різницевого методу"

На отриманому графіку можна помітити, що дані ряду було перетворено в стаціонарні, оскільки помітно відсутність тренду. Але для того, щоб точно в цьому запевнитись, бажано перевірити стаціонарність даних за допомогою функції Дікі-Фулера:

Функція:

$results = adfuller(transformed_ts1)$

Результат:

$1.9177281093251045e - 22$

Як можна побачити, розподіл, який до перетворення повертав ідеальне значення ($p \approx 0$), наразі повністю стаціонарний, про що свідчить знову ж таки значення p, яке є більшим за 0,05.

7. МАТЕТИЧНЕ ФОРМУЛЮВАННЯ

Причинність Грейнджера зазвичай перевіряється в контексті моделей лінійної регресії. Для ілюстрації розглянемо двовимірну лінійну авторегресивну модель двох змінних X_1 і X_2:

$$X_1(t) = \sum_{j=1}^{p} A_{11,j} X_1(t-j) + \sum_{j=1}^{p} A_{12,j} X_2(t-j) + E_1(t)$$
$$X_2(t) = \sum_{j=1}^{p} A_{21,j} X_1(t-j) + \sum_{j=1}^{p} A_{22,j} X_2(t-j) + E_2(t)$$

де p — максимальна кількість лагових спостережень, A_{ij} – елементи матриці A, яка містить коефіцієнти моделі (тобто внески кожного лагового спостереження передбачені значенням $X_1(t)$ і $X_2(t)$, а E_1 і E_2 є залишками (помилками передбачення) для кожного часового ряду. Якщо дисперсію E_1 зменшити шляхом включення доданків X_2 у перше рівняння, то кажуть, що X_2 впливає на X_1 за Грейнджером. Іншими словами, X_2 за Грейнджером прогнозує X_1, якщо коефіцієнти A_{12} значно відрізняються від нуля. Це можна перевірити, виконавши F-тест нульової гіпотези, де $A_{12} = 0$, враховуючи припущення стаціонарності для X_1 і X_2.

Причинність за Грейнджером можна легко поширити на випадок, коли змінних n, де $n > 2$, шляхом оцінки n-мірної авторегресійної моделі. У цьому випадку X_2 прогнозує за Грейнджером X_1, якщо лагові спостереження X_2 допомагають передбачити X_1, якщо також враховуються лагові спостереження всіх інших факторів $X_3 \dots X_n$. Це багатоваріативне розширення, яке іноді називають «умовна причинність за Грейнджером» є надзвичайно корисною, оскільки повторний парний аналіз кількох змінних іноді може давати оманливі (неточні) результати. Наприклад, повторний

двовимірний аналіз не зможе однозначно показати зв’язок між двома моделями (**Error! Reference source not found.**). Це означає, що потрібно розглядати не окремо пари X та Y, X та Z, Y та Z, а розглядати всі три змінні разом за допомогою авторегресійної моделі за Грейнджером. Саме тоді можна буде дізнатись, як на кожну змінну моделі впливають лагові значення інших змінних.

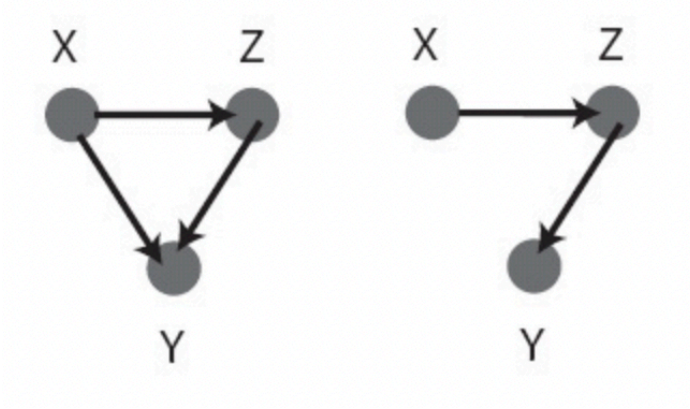

Рисунок 7 " Можливі зв'язки, які важко розрізнити за допомогою лише парного аналізу"

Отже, щоб застосувати вище наведене формулювання причинності за Грейнджером необхідно враховувати два важливих припущення:

- тест Грейнджера застосовується виключно для лінійних моделей;
- тест Грейнджера застосовується за умови, що середнє значення та дисперсія є сталими.

8. ВИЗНАЧЕННЯ ПРИЧИННОСТІ ЗА ГРЕЙНДЖЕРОМ

Тест причинно-наслідкового зв'язку між даними часових рядів за Грейнджером використовується для визначення факту, чи буде один часовий ряд "корисним" для прогнозування іншої змінної шляхом дослідження причинно-наслідкового зв'язку між двома змінними в часовому ряді. Метод є імовірнісним обрахуванням причинності. Він використовує спостережувані набори даних, аби знайти закономірності кореляції між ними. Однією з переваг векторної авторегресії часових рядів (VAR) є те, що є можливість, в якомусь сенсі, перевірити "причинність". Цей тест вперше запропонував Грейнджер у 1969 році, і тому його стали називати причинністю Грейнджера.

Векторна авторегресія (VAR) є економетричною моделлю, що використовуються для описання еволюції і взаємозалежності між кількома часовими рядами, узагальнюючи одномірні AR моделі. Кожна змінна пояснюється її еволюцію на основі власних лагів (значень за попередні періоди) і лагів всіх інших змінних у моделі.

У статистиці, економетриці та обробці сигналів модель авторегресії (AR) є представленням типу випадкового процесу; як такий, він використовується для опису певних змінних у часі процесів природи, економіки тощо. Авторегресійна модель визначає, що вихідна змінна лінійно залежить від її власних попередніх значень та від стохастичного терміну; таким чином, модель має форму стохастичного різницевого рівняння (або рекурентного співвідношення, яке не слід плутати з диференціальним рівнянням).

Далі описано механізм визначення причинності Грейнджера. Він заснований на ідеї, що якщо фактор X спричиняє фактор Y, то прогноз Y,

заснований на попередніх значеннях Y (де попередні – значення $Y_{i-\tau}$) та попередні значення X (де попередні – значення $X_{j-\tau}$).

Тест Грейнджера не слід використовувати для перевірки лагових факторів Y та самого фактору Y. Простіше кажучи, фактор X називають причиною за Грейнджером фактору Y, якщо фактор Y визначається лаговою змінною X та Y, а не лише лаговою змінною Y.

Виконуючи тест на причинність Грейнджера, треба враховувати припущення, що майбутні значення не можуть прогнозувати минулі значення;

Покажемо детальну реалізацію на прикладі пакета statsmodel, який надає класи та функції для оцінки багатьох різних статистичних моделей на Python для тесту Грейнджера. Для реалізації тесту необхідно імпортувати наступні бібліотеки: *pandas, numpy, matplotlib, seaborn, random, matplotlib.pyplot, parse, signal, interp1d, statsfrom statsmodels.tsa.stattools*. Далі вигрузимо декілька значень з набору даних. У наступний таблиці представлено поведінку часових рядів.

Таблиця 3 "Значення, які описують поведінку часових рядів"

Num	Value_date	IxD	RxB
0	2020-07-20	109.50	424.0
1	2020-07-17	110.45	420.0
2	2020-07-10	107.05	405.0
3	2020-07-03	101.20	409.5
4	2020-06-26	104.35	412.0

Зроблено певний аналіз даних. Тут представлено, як дані поводяться в часових рядах.

Рисунок 8 "Графік розподілу часових рядів RxB та IxD"

З наведеного вище рисунка (Рисунок 8) бачимо, що обидва часових ряди рухаються більш менш одночасно, і призводять до взаємозв'язку, який можна використовувати для прогнозування будь-якої змінної.

Тепер використаємо тест Грейнджера для ряду даних.

Тест Грейнджера потребує наступні аргументи: набір даних, максимальний можливий лаг.

Функція:

$$res = grangercausalitytests(df[['IxD','RxB']], maxlag = 15)$$

де maxlag – максимально можливий лаг, який розглядається.

Результат:

Granger Causality

number of lags (no zero) 1

ssr based F test: F=13.3271 , p=0.0003 , df_denom=519, df_num=1

ssr based chi2 test: chi2=13.4041 , p=0.0003 , df=1

likelihood ratio test: chi2=13.2349 , p=0.0003 , df=1

Granger Causality

number of lags (no zero) 2

ssr based F test: F=13.4540 , p=0.0000 , df_denom=516, df_num=2

ssr based chi2 test: chi2=27.1687 , p=0.0000 , df=2

likelihood ratio test: chi2=26.4840 , p=0.0000 , df=2

Granger Causality

number of lags (no zero) 3

ssr based F test: F=8.9401 , p=0.0000 , df_denom=513, df_num=3

ssr based chi2 test: chi2=27.1863 , p=0.0000 , df=3

likelihood ratio test: chi2=26.4994 , p=0.0000 , df=3

Granger Causality

number of lags (no zero) 4

ssr based F test: F=6.7321 , p=0.0000 , df_denom=510, df_num=4

ssr based chi2 test: chi2=27.4037 , p=0.0000 , df=4

likelihood ratio test: chi2=26.7047 , p=0.0000 , df=4

Granger Causality

number of lags (no zero) 5

ssr based F test: F=5.8029 , p=0.0000 , df_denom=507, df_num=5

ssr based chi2 test: chi2=29.6440 , p=0.0000 , df=5

likelihood ratio test: chi2=28.8268 , p=0.0000 , df=5

Granger Causality

number of lags (no zero) 6

ssr based F test: F=5.0143 , p=0.0001 , df_denom=504, df_num=6

ssr based chi2 test: chi2=30.8620 , p=0.0000 , df=6

likelihood ratio test: chi2=29.9760 , p=0.0000 , df=6

Granger Causality
number of lags (no zero) 7
ssr based F test: F=4.3764 , p=0.0001 , df_denom=501, df_num=7
ssr based chi2 test: chi2=31.5520 , p=0.0000 , df=7
likelihood ratio test: chi2=30.6250 , p=0.0001 , df=7

Granger Causality
number of lags (no zero) 8
ssr based F test: F=3.8112 , p=0.0002 , df_denom=498, df_num=8
ssr based chi2 test: chi2=31.5303 , p=0.0001 , df=8
likelihood ratio test: chi2=30.6027 , p=0.0002 , df=8

Granger Causality
number of lags (no zero) 9
ssr based F test: F=3.4022 , p=0.0005 , df_denom=495, df_num=9
ssr based chi2 test: chi2=31.7947 , p=0.0002 , df=9
likelihood ratio test: chi2=30.8501 , p=0.0003 , df=9

Granger Causality
number of lags (no zero) 10
ssr based F test: F=3.0085 , p=0.0011 , df_denom=492, df_num=10
ssr based chi2 test: chi2=31.3696 , p=0.0005 , df=10
likelihood ratio test: chi2=30.4478 , p=0.0007 , df=10

Granger Causality
number of lags (no zero) 11
ssr based F test: F=2.9947 , p=0.0007 , df_denom=489, df_num=11

ssr based chi2 test: chi2=34.4912 , p=0.0003 , df=11
likelihood ratio test: chi2=33.3791 , p=0.0005 , df=11

Granger Causality
number of lags (no zero) 12
ssr based F test: F=2.7394 , p=0.0013 , df_denom=486, df_num=12
ssr based chi2 test: chi2=34.5641 , p=0.0005 , df=12
likelihood ratio test: chi2=33.4453 , p=0.0008 , df=12

Granger Causality
number of lags (no zero) 13
ssr based F test: F=2.6577 , p=0.0013 , df_denom=483, df_num=13
ssr based chi2 test: chi2=36.4818 , p=0.0005 , df=13
likelihood ratio test: chi2=35.2360 , p=0.0008 , df=13
Granger Causality
number of lags (no zero) 14
ssr based F test: F=2.4680 , p=0.0022 , df_denom=480, df_num=14
ssr based chi2 test: chi2=36.6399 , p=0.0008 , df=14
likelihood ratio test: chi2=35.3812 , p=0.0013 , df=14

Granger Causality
number of lags (no zero) 15
ssr based F test: F=2.3488 , p=0.0029 , df_denom=477, df_num=15
ssr based chi2 test: chi2=37.5225 , p=0.0011 , df=15
likelihood ratio test: chi2=36.2014 , p=0.0017 , df=15

Після того, як тест був виконаний, проаналізуємо результати.

Тест був виконаний з 15 лагами!

number of lags (no zero) [xx] - показує кількість лагів, які використовуються для пошуку причинно-наслідкових зв'язків.

Granger Causality
number of lags (no zero) 2
ssr based F test: F=13.4540

Аналіз результатів: На першому кроці значення F-тесту: $F = 13.3271$ при $p = 0{,}003$, На другому кроці значення F-тесту: $F = 13.4540$ при $p = 0.000$, на третьому кроці значення F-тесту: $F = 8.9401$ при $p = 0.000$, на четвертому кроці значення F-тесту: $F = 6.7321$ при $p = 0.000$. Як можна зауважити лаг в 2 показує найвище значення F-тесту з усіх лагів ($F = 13.4540$), тобто вплив другої змінної дуже високий, вплив першої майже такий самий, як і другої. Але вплив третьої та решти змінних вдвічі, а подекуди і втричі менший за другу, що означає, що на даному етапі вже можна не включати дані змінні в прогнозування. F-тест перевіряє, чи лагові змінні X разом з X покращують прогноз Y. Отже, прогнозування Y можна зробити за допомогою двох змінних X_{i-1} і X_i.

І отримаємо модель: $Y_i = b_0 + b_1 * X_{1_i} + b_2 * X_{1_{i-1}}$.

Тобто суть тесту полягає в тому, що користувачем вказується максимальний лаг, а сам тест відповідає наскільки довгий цей лаг, аналізуючи його.

Також слід звернути увагу:

- Тестування причинно-наслідкового зв'язку за Грейнджером за допомогою F-тесту, коли значення одного або обох часових рядів нестаціонарні, може призвести до помилкової причинності. Якщо значення обох часових рядів не є стаціонарними, то перед використанням тесту на причинність

Грейнджера необхідно спочатку застосувати різницевий метод, щоб прибрати тренд.

- F-тест можна використовувати тільки тоді, коли змінні коінтегровані.
- Говориться, що X прогнозує за Грейнджером Y, коли нульову гіпотезу відхиляють.
- Нульова гіпотеза для тесту полягає в тому, що лагові значення фактору X не прогнозує фактор Y.

9. ВИСНОВКИ

Курсову роботу присвячено вивченню теми визначення причинно-наслідкового зв’язку за допомогою тесту Грейнджера. Тест Грейнджера – тест, що визначає вплив лагових значень однієї змінної на прогнозування іншої зміннох. Тобто користувачем вказується максимальний лаг, а сам тест відповідає наскільки довгий цей лаг, аналізуючи його.

У даній роботі розглядається застосування методу Грейнджера для стаціонарних даних, а також методи, які допомагають зробити дані стаціонарними, й тест, що визначає результат стаціонарності даних. У п’ятому розділі цієї роботи було розглянуто тест Дікі-Фуллера, який визначає стаціонарні дані чи ні. Після чого, у шостому розділі, описано алгоритм приведення даних до стаціонарного виду, у випадку їхньої нестаціонарності, за допомогою різницевого методу, а також логарифмування даних. А у восьмому розділі було успішно проведено тест Грейнджера, за результатами якого було визначено, які лагові значення змінної впливають на прогнозування іншої змінної.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. https://www.analyticsvidhya.com/blog/2021/08/granger-causality-in-time-series-explained-using-chicken-and-egg-problem/
2. https://medium.com/swlh/using-granger-causality-test-to-know-if-one-time-series-is-impacting-in-predicting-another-6285b9fd2d1c
3. https://en.wikipedia.org/wiki/Granger_causality
4. http://www.scholarpedia.org/article/Granger_causality
5. https://towardsdatascience.com/how-to-remove-non-stationarity-in-time-series-forecasting-563c05c4bfc7
6. https://uk.wikipedia.org/wiki/Клайв_Грейнджер