

Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет _____ Комп'ютерних наук _____
(повна назва)

Кафедра _____ Системотехніки _____
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

_____ другий (магістерський) _____
(рівень вищої освіти)

_____ ГЮИК.591620.003 _____

ПЗ _____
(позначення документа)
Розробка та дослідження методу прогнозування в центрах надання
_____ гуманітарної допомоги _____
(тема)

Виконав: здобувач групи ІТПм-21-1 _____
спеціальності 122 Комп'ютерні науки _____
(код і повна назва спеціальності)

освітньої програми _____ ОПІ Інформаційні
_____ технології проектування _____
(повна назва освітньої програми)

_____ Голик Є.Є. _____
(прізвище, ініціали)

Керівник _____ проф. Ситніков Д.Е _____
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри системотехніки _____
(підпис)

_____ Гребеннік І.В. _____
(прізвище, ініціали)

2022 р.

Харківський національний університет радіоелектроніки

Факультет _____ Комп'ютерних наук _____

Кафедра _____ Системотехніки _____

Рівень вищої освіти _____ другий (магістерський) _____

Спеціальність _____ 122 Комп'ютерні науки _____
(код і повна назва)

Освітня програма _____ ОПІ Інформаційні технології проектування _____
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. Кафедри

_____ (підпис)
« ____ » _____ 20__ р.

ЗАВДАННЯ НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачові _____ Голику Єгору Євгеновичу _____
(прізвище, ім'я, по батькові)

1. Тема роботи Розробка та дослідження методу прогнозування в центрах надання гуманітарної допомоги
затверджена наказом по університету від «21» листопада 2021р. № 1504 Ст
2. Термін подання студентом роботи до екзаменаційної комісії 24 грудня 2022 р.
3. Вихідні дані до роботи Функції методу: обробка вхідних даних, прогнозування даних на основі моделі авторегресивної інтегрованої ковзаючої середньої, візуалізація прогнозованих даних. Вибірки даних для навчання та тестування методу прогнозування. Використовувати середовище розробки Google Colab, СУБД Microsoft SQL Server, середовище Google Drive, мову програмування Python.
4. Перелік питань, що потрібно опрацювати в роботі 4.1 Вступ, 4.2 Аналіз предметної області організацій що надають гуманітарну допомогу, 4.2.1 Аналіз об'єкту дослідження, 4.2.2 Організація гуманітарної допомоги в Україні, 4.2.3 Аналіз сучасного стану інформатизації гуманітарної допомоги, 4.2.4 Висновки за розділом 1, 4.3 Огляд методів та технологій, які застосовуються в задачах прогнозування даних, 4.3.1 Огляд методів прогнозування, 4.3.2 Огляд методів диференціювання часових рядів, 4.4 Постановка задачі дослідження, 4.4.1 Змістовна постановка задачі, 4.4.2 Постановка задачі, 4.5 Математичний опис задачі, 4.5.1 Порівняння методів прогнозування, 4.5.2 Математичний опис задачі прогнозування в центрах надання гуманітарної допомоги, 4.5.3 Розробка методу прогнозування даних в центрах надання гуманітарної допомоги, 4.5.4 Опис потрібного інструментарію та його застосування., 4.6 Експериментальні дослідження, 4.6.1 Розробка вимог до програмного засобу, 4.6.2 Розробка вимог до моделі бази даних, 4.6.3 Розробка запитів до бази даних, 4.6.4 Програмна реалізація методу прогнозування даних в центрах надання гуманітарної допомоги, 4.6.6 Проведення експериментальних досліджень методу прогнозування даних в центрах надання гуманітарної допомоги, 4.6.7 Аналіз результатів дослідження, 4.7 Висновки

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів) Слайди презентації: титул, мета роботи, опис об'єкту досліджень, постановка задачі, вхідні дані до алгоритму, блок-схема методу прогнозування, опис моделі авторегресивної інтегрованої ковзаючої середньої, розробка вимог до моделі даних, розробка запитів до бази даних, програмна реалізація методу, проведення експериментальних досліджень, аналіз результатів досліджень, висновки.

КАЛЕНДАРНИЙ ПЛАН

Пор. №	Назва етапів атестаційної роботи	Термін виконання етапів роботи	Примітка
1.	Отримання завдання кваліфікаційної роботи	08.11.2022	
2.	Аналіз завдання, літератури та аналогів з теми кваліфікаційної роботи	09.11 — 10.11.2022	
3.	Дослідження методів прогнозування	11.11 — 14.11.2022	
4.	Дослідження методів диференціювання часових рядів	15.11 – 17.11.2022	
5.	Вибір засобів для розробки технічних вимог до програми	17.11 – 20.11.2022	
6.	Розробка методу прогнозування даних	20.11 — 01.12.2022	
7.	Проведення експериментальних досліджень	02.12 — 04.12.2022	
8.	Аналіз результатів досліджень	05.12 — 09.12.2022	
9.	Оформлення пояснювальної записки та програмної документації	10.12 — 15.12.2022	
10.	Оформлення графічної частини та презентаційних матеріалів комп'ютерного захисту	15.12 – 20.12.2022	
11.	Представлення на рецензування	21.12.2022	
12.	Представлення атестаційної роботи в ДЕК	22.12.2022	

Дата видачі завдання 08 листопада 2022 р.

Студент _____ Голик Є.Є.
(підпис)

Керівник роботи _____ професор Ситніков Д.Е.
(підпис)

РЕФЕРАТ

Пояснювальна записка: 64 с., 18 рис., 3 табл., 3 додатки, 25 джерел інформації

ЦЕНТРИ НАДАННЯ ГУМАНІТАРНОЇ ДОПОМОГИ, ПРОГНОЗУВАННЯ ДАНИХ, СТАЦІОНАРНІ ЧАСОВІ РЯДИ, ВЕЛИКІ ДАНІ, МОДЕЛЬ АВТОРЕГРЕСИВНОЇ ІНТЕГРОВАНОЇ КОВЗАЮЧОЇ СЕРЕДНЬОЇ

Об'єктом дослідження є дані центрів надання гуманітарної допомоги.

Предметом дослідження є інформаційні технології й методи прогнозування даних.

Метою досліджень є розробка та застосування методу прогнозування даних в центрах надання гуманітарної допомоги з використанням моделі авторегресивної інтегрованої ковзаючої середньої.

Методи дослідження – аналіз об'єкту дослідження, методів прогнозування даних, розробка штучного набору даних, програмна реалізація методу, проведення експериментальних досліджень.

Результати роботи – програмна реалізація методу прогнозування даних, оцінка точності прогнозування розробленого методу.

Результати роботи можуть бути використані для проведення прогнозування даних довільного набору даних та оцінки орієнтованих прогнозів.

Галузь застосування – надання гуманітарної допомоги.

ABSTRACT

Explanatory note: 64 p., 18 pic., 3 tables., 3 ann., 25 sources

HUMANITARIAN ASSISTANCE CENTERS, DATA FORECASTING,
STATIONARY TIME SERIES, BIG DATA, AUTOREGRESSIVE
INTEGRATED MOVING AVERAGE MODEL

The object of the research is the data of humanitarian aid centers.

The subject of research is information technology and data forecasting methods.

The purpose of the research is the development and application of a data forecasting method in humanitarian aid centers using the autoregressive integrated moving average model.

Research methods – analysis of the research object, data forecasting methods, development of an artificial data set, software implementation of the method, conducting experimental research.

The results of the work are software implementation of the data forecasting method, assessment of forecasting accuracy of the developed method.

The results of the work can be used to forecast the data of an arbitrary data set and evaluate oriented forecasts.

The field of application is the provision of humanitarian aid.

ЗМІСТ

Перелік умовних позначень, символів, одиниць, скорочень і термінів	7
Вступ.....	8
1 Аналіз предметної області організацій що надають гуманітарну допомогу	10
1.1 Аналіз об'єкту дослідження.....	10
1.2 Організація гуманітарної допомоги в Україні	14
1.3 Аналіз сучасного стану інформатизації гуманітарної допомоги.....	15
1.4 Висновки за розділом 1	17
2 Огляд методів та технологій, які застосовуються в задачах прогнозування даних.....	19
2.1 Огляд методів прогнозування.....	19
2.2 Огляд методів диференціювання часових рядів.....	25
3 Постановка задачі дослідження.....	29
3.1 Змістовна постановка задачі	29
3.2 Постановка задачі	32
4 Математичний опис задачі.....	33
4.1 Порівняння методів прогнозування	34
4.2 Математичний опис задачі прогнозування в центрах надання гуманітарної допомоги	35
4.3 Розробка методу прогнозування даних в центрах надання гуманітарної допомоги	36
4.4 Опис потрібного інструментарію та його застосування.....	42
5 Експериментальні дослідження.....	44
5.1 Розробка вимог до програмного засобу.....	44
5.2 Розробка вимог до моделі бази даних.....	44
5.3 Розробка запитів до бази даних	47
5.4 Програмна реалізація методу прогнозування даних в центрах надання гуманітарної допомоги	49
5.6 Проведення експериментальних досліджень методу прогнозування даних в центрах надання гуманітарної допомоги.....	50
5.7 Аналіз результатів дослідження	58
Висновки	60
Перелік джерел	63
Додаток А. Текст програми.....	Ошибка! Закладка не определена.
Додаток Б. Графічний матеріал	Ошибка! Закладка не определена.
Додаток В. Відомість кваліфікаційної роботи.....	Ошибка! Закладка не определена.

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

ЕОМ – електронно-обчислювальна машина;

ANN – Artificial Neural Network – штучна нейронна мережа;

ARIMA – Autoregressive Integrated Moving Average Model – авторегресивна інтегрована ковзаюча середня;

ML – Machine Learning – машинне навчання;

ACF – Autocorrelation Function – автокорреляційна функція;

PACF – Partial autocorrelation Function – часткова автокорреляційна функція;

БД – база даних.

ВСТУП

Важливість науки про дані об'єднує досвід у галузі програмування, математики та статистики, щоб створювати ідеї та осмислювати дані. Коли ми думаємо про те, чому наука про дані стає все більш важливою, відповідь полягає в тому факті, що цінність даних стрімко зростає. Наука про дані – це функція, яка швидко розвивається, але експерти галузі кажуть, що вона все ще перебуває в зародковому стані. Вона користується високим попитом і пояснює, як цифрові дані трансформують бізнес і допомагають їм приймати чіткіші та важливіші рішення.

В контексті прийняття важливих рішень не винятком стала і сфера надання гуманітарної допомоги. На жаль, з кожним днем росте кількість людей, що постраждали внаслідок техногенних катастроф, пандемій або збройних конфліктів. Центри надання гуманітарної допомоги разом із різними волонтерськими фондами отримують безліч запитів на допомогу в їжі, медикаментах, притулку або матеріальної допомоги, що породжує великий обсяг даних, який треба обробляти. Неправильний підхід до обробки цих даних може призвести до масштабних логістичних проблем, внаслідок чого безліч людей може залишитися без надання належної допомоги.

Зараз сфера надання гуманітарної допомоги стрімко розвивається в напрямку інформаційних технологій. Зокрема, з'являються додатки, що дозволяють зручно та швидко створювати запити на отримання гуманітарної допомоги, а також є значні результати в розвитку вирішення проблеми розподілу ресурсів.

В ході суттєвого зростання обсягу цих даних, також з'явилась проблема аналізу майбутніх, ще невідомих даних. Дану проблему відносять до задач прогнозування, і саме цей напрямок вирішення задач потребує сфера надання гуманітарної допомоги.

В ході виконання даної роботи пропонується розглянути існуючі методи та підходи до прогнозування даних, а також на їх основі розробити комплексний метод прогнозування, який буде ефективно аналізувати вхідні набори даних, знаходити закономірності, та створювати найбільш ефективну модель прогнозування даних. Також пропонується дослідити даний метод прогнозування на тестових даних та виконати розрахунки його точності та ефективності.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ОРГАНІЗАЦІЙ ЩО НАДАЮТЬ ГУМАНІТАРНУ ДОПОМОГУ

1.1 Аналіз об'єкту дослідження

Гуманітарна допомога - це матеріальна та матеріально-технічна допомога людям, які потребують допомоги [1]. Зазвичай це короткострокова допомога, поки її не замінить довгострокова допомога від уряду та інших установ. Серед нужденних є бездомні, біженці, постраждалі від стихійних лих, воєн і голоду. Гуманітарна допомога здійснюється в гуманітарних цілях і включає стихійні лиха та техногенні катастрофи. Основною метою гуманітарної допомоги є порятунок життів, полегшення страждань і збереження людської гідності. Тому її можна відрізнити від допомоги на розвиток, яка спрямована на усунення глибинних соціально-економічних факторів, які могли призвести до кризи чи надзвичайної ситуації.

Гуманітарна допомога призначена для порятунку життів і полегшення страждань під час і відразу після надзвичайних ситуацій, тоді як допомога розвитку реагує на поточні структурні проблеми, зокрема системну бідність, яка може перешкоджати економічному, інституційному та соціальному розвитку в будь-якому конкретному суспільстві, і допомагає в розбудові спроможності забезпечити стійкі громади та стійкі засоби до існування [1]. І гуманітарна допомога, і допомога розвитку пов'язані між собою, і різні форми допомоги часто мають як гуманітарний компонент, так і компонент розвитку.

Допомога розвитку може існувати без гуманітарної допомоги (у країнах, що розвиваються, де немає гуманітарних криз), але вона також часто існує на додаток до гуманітарної допомоги під час та після криз. Наприклад, якщо в країні стався землетрус, потрібна короткострокова гуманітарна

допомога [1]. Це включає доставку їжі та води, надання тимчасового притулку, а також медичне обслуговування.

Існують різні види гуманітарної допомоги. Серед основних виділяються наступні види:

а) допомога при стихійних лихах. Стихійні лиха є основною причиною смерті та страждань у світі. Гуманітарна допомога такого типу кризи включає притулок, харчування, чисту воду тощо. Червоний Хрест реагує на багато катастроф у всьому світі;

б) притулок – це один із найважливіших аспектів гуманітарної допомоги. Це часто потрібно, коли надається допомога людям, переміщеним через катастрофи, війни чи інші конфлікти. Надання притулку також може включати допомогу, наприклад оплату готельних номерів;

в) харчування – найчастіше до даного типу гуманітарної допомоги відносять продукти харчування з великим терміном придатності: консервовані продукти, зернові вироби, горіхи, галети або їжа швидкого приготування;

г) санітарія – доступ до безпечної води та санітарії. До цього виду гуманітарної допомоги належать належні санітарні умови, зокрема туалети та засоби гігієни, які попереджають поширення хвороб. WASH (water, sanitation, and hygiene), що означає «вода, санітарія та гігієна», є важливою проблемою громадського здоров'я в цілях сталого розвитку;

г) медична допомога – це допомога, що включає широкий спектр заходів, таких як надання коштів, матеріалів і персоналу;

д) освіта – вид гуманітарної допомоги, що включає надання можливості навчання в закладах початкової, середньої або вищої освіти, доступ до навчальних матеріалів та освітніх платформ;

е) фінансова допомога.

Гуманітарна допомога розглядається як основне вираження універсальної цінності солідарності між людьми та моральний імператив. Гуманітарна допомога може надходити від місцевих або міжнародних

спільнот. Звертаючись до міжнародної спільноти, Управління з координації гуманітарних питань (УКГП) Організації Об'єднаних Націй (ООН) відповідає за координацію реагування на надзвичайні ситуації [1]. Він стосується різних членів міжвідомчого постійного комітету, члени якого відповідають за надання екстреної допомоги. Чотири організації ООН, які відіграють основну роль у наданні гуманітарної допомоги, це Програма розвитку ООН (ПРООН), Агентство ООН у справах біженців (УВКБ), Дитячий фонд ООН (ЮНІСЕФ) і Всесвітня продовольча програма (ВПП).

Гуманітарна допомога охоплює широкий спектр діяльності, включаючи надання продовольчої допомоги, притулку, освіти, охорони здоров'я чи захисту. Більшість допомоги надається у формі товарів або допомоги в натуральній формі, а готівка та ваучери становлять лише 6% загальних гуманітарних витрат [1]. Однак докази показали, що грошові перекази можуть бути кращими для одержувачів, оскільки вони дають їм вибір і контроль, вони можуть бути економічно ефективнішими та кращими для місцевих ринків і економік.

Важливо зазначити, що гуманітарна допомога надається не лише гуманітарними працівниками, яких надсилають двосторонні, багатосторонні чи міжурядові організації, такі як ООН. Такі суб'єкти, як самі постраждалі люди, громадянське суспільство, місцеві неформальні служби першої допомоги, громадянське суспільство, діаспора, підприємства, органи місцевого самоврядування, військові, місцеві та міжнародні неурядові організації – усі вони відіграють вирішальну роль у своєчасній доставці гуманітарної допомоги.

Спосіб доставки допомоги може вплинути на якість і кількість допомоги. Часто в ситуаціях лиха міжнародні агенції з надання допомоги співпрацюють з місцевими агенціями. Можуть існувати різні домовленості щодо ролі, яку відіграють ці агентства, і така домовленість впливає на якість жорсткої та м'якої допомоги, що надається.

Організаційна структура центрів надання гуманітарної допомоги за стандартами ООН складається з декількох взаємопов'язаних відділів, серед яких основними є [2]:

- координаційний відділ;
- відділ мобілізації гуманітарного фінансування та ресурсів;
- виконавчий офіс;
- відділ управління інформацією;
- відділ операцій та адвокацій;
- офіс заступника Генерального секретаря;
- відділ стратегічних комунікацій;
- офіс помічника Генерального секретаря;
- відділ стратегії, планування, оцінки та керівництва;
- гендерний блок.

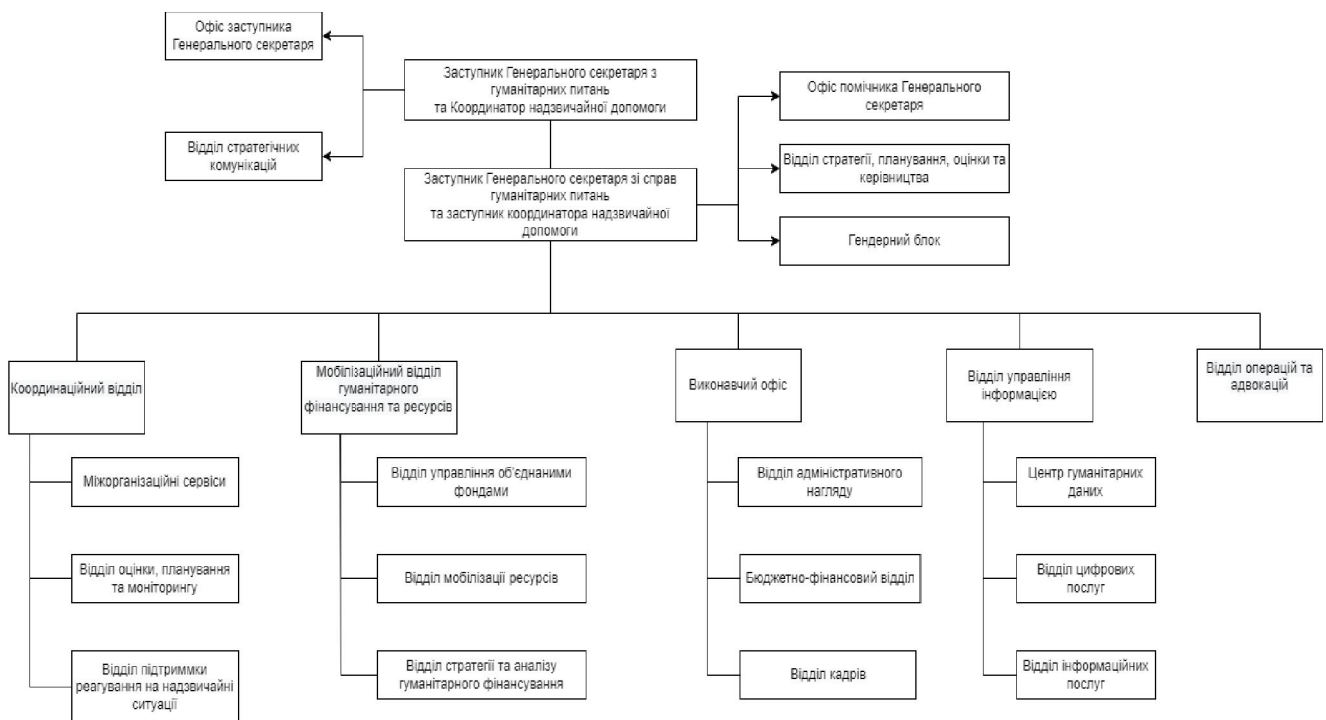


Рисунок 1.1 – Організаційна структура центру надання гуманітарної допомоги

1.2 Організація гуманітарної допомоги в Україні

В Україні існують аналоги центрів надання гуманітарної допомоги, однак відомостей про те, яка в них організаційна структура та алгоритм роботи – немає. Відомо, що наразі гуманітарну допомогу можна отримати в центрах надання адміністративних послуг, маючи статус внутрішньо переміщеної особи. Центр надання гуманітарної допомоги не є окремою організацією, а є підрозділом державної установи. Серед систем, які використовують центри надання гуманітарної допомоги є застосунок “Дія”, за допомогою якого кожен громадянин може отримати статус внутрішньо переміщеної особи, за умови якщо він залишив своє місце проживання внаслідок бойових дій [3]. Маючи статус внутрішньо переміщеної особи, громадянин може звернутися до найближчого центру надання адміністративних послуг за гуманітарною допомогою. Відомий також випадок отримання фінансової допомоги, що є також частиною гуманітарної допомоги, для фізичних осіб-підприємців. Кожен громадянин, що був фізичною особою-підприємцем та сплачував податки протягом минулого кварталу, міг отримати фінансову допомогу, створивши запит через додаток “Дія”. Оскільки додаток “Дія” має доступ до даних банків, було можливо перевірити дані фізичної особи-підприємця, і в разі успішної перевірки надіслати одноразову фінансову допомогу [3].

Також відомі випадки, коли гуманітарна допомога надається не державними організаціями, найчастіше фондами, що збирають кошти або продукти від волонтерів, та видає людям, які постраждали внаслідок певних обставин. Але також не має інформації, якими технологіями вони користуються. Найчастіше, особа, що хоче отримати гуманітарну допомогу, звертається особисто до певного центру надання гуманітарної допомоги, та залишає заяву на отримання продуктів, одягу, житла тощо.

«Сьогодні створено чимало засобів для забезпечення гуманітарної допомоги постраждалим від війни в Україні. Це різні сторінки в соціальних

мережах, форуми, Telegram канали, сайти та додатки. Усі вони є дуже різними та покривають різні види потреб. Кожна платформа має свої плюси та мінуси. Більшість таких організацій, які здійснюють допомогу, розглядають запити внутрішньо, без публікації їх на загал, та самі приймають рішення стосовно них. Це надзвичайно затягує процес зібрання коштів чи необхідної допомоги з декількох причин. По-перше, це знижує ймовірність отримання допомоги, адже доки черга дійде до обробки того чи іншого оголошення, то не можна бути певними, що її задовільнять. По-друге, слідуючи такому підходу, людина, що потребує термінової допомоги, втрачає дорогоцінний час на очікування. По-третє, такого роду організації здебільшого займаються великими запитами для багатьох людей одночасно. Отже, поодинокі оголошення про допомогу можуть часто залишатись без уваги. Таким чином люди, які потребують допомоги та не належать до загального запиту організації, не зможуть отримати необхідне» [4].

1.3 Аналіз сучасного стану інформатизації гуманітарної допомоги

Однією з проблем, з якою імовірно зустрічаються організації, що надають гуманітарну допомогу у великих обсягах, є проблема розподілу ресурсів. Цю проблему відносять до задачі оптимізації за багатьма критеріями. Рішення цієї задачі дозволило організаціям мати основу для побудови системи, або інструменту, за допомогою якого можна ефективно робити розподіл ресурсів при наданні гуманітарної допомоги.

У 2016 році групою науковців із відділу статистики та дослідження операцій Віденського університету, а саме Уолтером Гаджаром та Памелою Нольц, було проведене дослідження та видано статтю за темою “Багатокритеріальна оптимізація у гуманітарній допомозі” [5]. У цьому дослідженні вказано, що використовується велика різноманітність критеріїв оптимізації, серед яких виділяють три основні групи критеріїв, а саме: критерії ефективності, критерії дієвості та критерії справедливості. Було

вдосконалено цю схему, розділяючи групову «ефективність» на підгрупи: вартість, час відгуку, відстань подорожі, покриття, надійність і безпека. Ці критерії згодом використовуватимуться у вирішенні задачі багатокритеріальної оптимізації. Серед методів багатокритеріальної оптимізації, було досліджено наступні методи: оптимізація за методом Парето, лексикографічна оптимізація, вектора оптимізація або скаляризація, цільове програмування, компромісне програмування та метод аналітичного ієрархічного процесу.

Стосовно задачі багатокритеріальної оптимізації є також наукова стаття під назвою «Багатокритеріальна модель оптимізації розподілу гуманітарної допомоги» [6]. У цій статті запропоновано декілька критеріїв для проблеми розподілу допомоги та розроблено багатокритеріальну модель оптимізації, що стосується всіх цих аспектів. Ця модель є ядром системи підтримки прийняття рішень, що розробляється для допомоги організаціям, відповідальним за розподіл гуманітарної допомоги. Після опису запропонованих критеріїв і моделі представлено ілюстративне прикладне дослідження на основі катастрофічного землетрусу на Гаїті 2010 року, яке демонструє корисність пропозиції.

Також однією з проблем є недостатня кількість інформаційних систем, які могли б полегшити процеси надання допомоги як людям, що потребують на допомогу, так і організаціям, що надають її. Є одиничні випадки, де частково використовуються системи в цілях надання гуманітарної допомоги, наприклад в українському додатку “Дія”.

У 2022 науковцями з Національного університету «Львівська Політехніка» було видано статтю “Формування вимог для імплементації мобільного додатку для гуманітарної допомоги”. У науковій статті описано актуальність проблеми малої кількості інформаційних систем, проведено порівняльний аналіз існуючих систем, їх переваги та недоліки. Було розглянуто наступні організації, які діють зараз на території України: Gate to Ukraine, Товариство Червоного Хреста України, СпівДія, 1k Ukraine,

Благодійний фонд «Щаслива дитина», «Допомога, Solidarity, Germany4Ukraine. Було досліджено, що серед восьми організацій тільки у двох використовується мобільний додаток як інструмент для забезпечення обліку заявок на отримання допомоги, у чотирьох – інформаційні веб сторінки, та в жодній організації не має в наявності інформаційної системи. Акцент у даній статті було зроблено на тому, що неможливо забезпечити процес надання гуманітарної допомоги чітко та прозоро, якщо організація веде процеси обліку, реєстрації, прийняття рішень без спеціалізованих систем, веде їх внутрішньо, не оголошуючи результатів [4]. Окрім того, в умовах поточної повномасштабної війни в Україні, коли кількість нужденних в гуманітарній допомозі людей перевищує за мільйон осіб, коли потреба в допомозі розповсюджується не тільки за кордоном, а й на території держави, на якій перебуває понад пів мільйона осіб, вимушених переселитися із небезпечних регіонів, ці недоліки спричиняють масштабні проблеми в контексті швидкості надання гуманітарної допомоги, та покриття потреб громадян. На основі цього дослідження було запропоновано вимоги до системи, яка полегшить діяльність центрів надання гуманітарної допомоги. Задля зручності користувачів пропонується створити у вигляді мобільного додатку. Система також повинна підтримувати різні типи допомоги: не лише фінансову, проте і матеріальну, адже не так рідко люди можуть надати не тільки гроші, але і речі. Також для зручності використання додатку необхідно створити зручний користувацький інтерфейс, який забезпечить легку та швидку навігацію, для задоволення потреб користувачів. Зважаючи на різні типи та сфери оголошень, важливим є також створення рекомендаційної системи, яка дозволить покращити досвід користування додатком, та допоможе користувачам швидко здійснити пожертву на те, чи інше оголошення, яке їх безпосередньо цікавить [4].

1.4 Висновки за розділом 1

З оглядом на вже існуючі дослідження, можна зробити певні висновки. По-перше, однією з критичних задач діяльності центрів надання гуманітарної допомоги є задача розподілу ресурсів, задля якої існують рішення з використанням методів багатокритеріальної оптимізації. По-друге, ще однією з проблем є недостатня кількість централізованих інформаційних систем, які могли б дати основу для розвитку та пошуку нових рішень у сфері надання гуманітарних послуг.

Методи багатокритеріальної оптимізації використовуються при наявності певних вхідних даних, на базі яких ведеться пошук рішення. Якщо абстрактно уявити, що є певна кількість центрів видачі гуманітарної допомоги а також ресурси, які розподіляються по них, виникає питання того, в які центри і скільки потрібно виділити ресурсів, щоб задовольнити потреби людей, що звертаються по допомогу.

Окрім того, з огляду на сучасний стан центрів надання гуманітарної допомоги, спостерігається дефіцит в спеціалізованих інструментів, які могли б покращити результати роботи, зробити процеси більш чіткими та прозорими, а також забезпечити необхідною інформацією людей, що звертаються по допомогу.

Отже, є доцільним дослідити та розробити метод прогнозування в центрах надання гуманітарної допомоги, та розробити застосунок, в якому буде використовуватись цей метод для вирішення задач, поставлених центрам надання гуманітарної допомоги. Зокрема, при вирішенні задачі прогнозування кількості осіб, що будуть звертатися по допомогу, а також кількості ресурсів гуманітарної допомоги, що будуть задіяні.

2 ОГЛЯД МЕТОДІВ ТА ТЕХНОЛОГІЙ, ЯКІ ЗАСТОСОВУЮТЬСЯ В ЗАДАЧАХ ПРОГНОЗУВАННЯ ДАНИХ

2.1 Огляд методів прогнозування

Прогнозування – це передбачення, зроблене шляхом вивчення статистичних даних та минулих закономірностей. Компанії використовують програмні інструменти та системи для аналізу великих обсягів даних, зібраних протягом тривалого періоду. Потім програмне забезпечення прогнозує майбутній попит та тенденції, допомагаючи приймати більш точні фінансові, маркетингові або операційні рішення [7].

За основу класифікації методів прогнозування взято об'єкт, вихідні дані чи модель якого дозволяє побудувати прогноз (рис.2.1) [8]. В першій групі методів (експертні методи) таким об'єктом є експерти, які опитуються чи анкетуються. Основна ідея даних методів полягає у проведенні опитування та статистичному опрацюванні суб'єктивних прогнозів експертів. Оскільки формально джерелами знань є експерти, які володіють власними, здебільшо інтуїтивними правилами прийняття рішень, а експертні методи не будують ніяких моделей міркування експертів, а лише збирають дані, то експертні методи не є предметом дослідження.

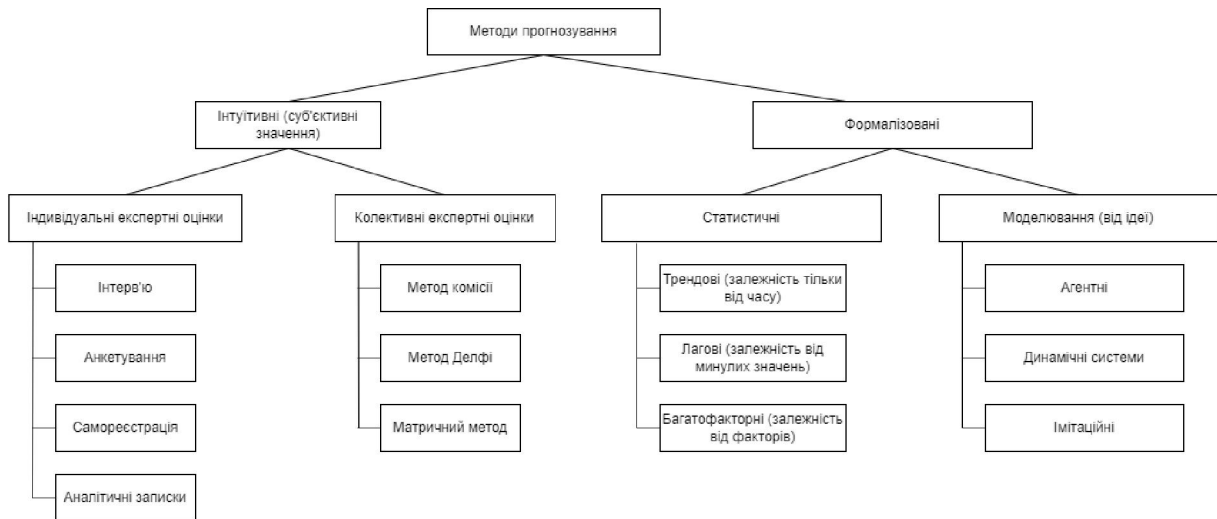


Рисунок 2.1 – Класифікація методів прогнозування

Формалізовані методи прогнозування поділяються на статистичні методи та моделювання. Варто зазначити, що термін «моделювання» тут розуміється у вузькому сенсі, як побудова моделей, які описують безпосередньо функціонування досліджуваного об'єкту [8]. Статистичні методи – методи, які досліджують тільки похідні показники досліджуваного об'єкту, представлені у вигляді часових рядів. Часові ряди — це послідовність деяких значень вимірювань, упорядкованих за їхнім виглядом в інтервалі часу. У часовому ряді є два обов'язкові елементи: відмітка часу і значення, пов'язане з ним [9]. Часовим рядом називається послідовність рівнів ряду:

$$\{y_i\} = y_1, y_2, \dots, y_n, \quad (2.1)$$

де i – натуральні числа, які відповідають моментам чи проміжкам вимірювання досліджуваної величини. Вважаємо, що вимірювання проходить через рівні проміжки часу, y_i – рівні ряду [8]. Усі методи прогнозування можуть бути подані у вигляді наступної математичної моделі:

$$\{y_{t+1}\} = f\left(t, y_t, \dots, y_{t-m}, x_t^{(1)}, \dots, x_{t-m_1}^{(1)}, x_t^{(2)}, \dots, x_{t-m_2}^{(2)}, x_t^{(k)}, \dots, x_{t-m_k}^{(k)}\right), \quad (2.2)$$

де y_t – величина показника, що прогнозується в момент часу t ; $x_t^{(i)}$ – величина i -го фактора в момент часу t , який впливає на y ; m, m_1, \dots, m_k – довжини «пам'яті» ряду, які використовуються при прогнозуванні. Відмінності різних методів будуть стосуватися структури вхідних даних x та виду функції $f(\dots)$ [8].

Таким чином, можна виділити такі види залежностей, які використовуються у сучасних методах прогнозування:

- лінійна;
- поліноміальна;
- нейромережева;
- нечітка;
- випадкова.

Говорячи про часові ряди, слід розрізнити концепцію методу прогнозування та модель прогнозування, які є подібними через їх співвідношення. Метод прогнозування описує порядок дій, що необхідно зробити, щоб отримати модель прогнозування часових рядів. Також метод має виконувати операції, щоб оцінити якість прогнозних значень. У той же час модель прогнозування є функціональним представленням часових рядів. Це основа для прогнозування значень майбутнього процесу [9]. Загальний ітеративний спосіб побудови моделі прогнозування складається з наступних кроків:

а) Вибір класу загальних моделей для прогнозування часових рядів для вибраного часового горизонту. Зазвичай цей клас вибирається на підставі минулого досвіду;

б) Вибір моделі підкласу. Для більшості моделей прогнозування існує безліч варіантів підмоделей. Цей крок вимагає більш глибокого аналізу часових рядів та вимог прогнозування;

в) Аналіз параметрів моделі. Для цього кроку потрібні дані, які дозволяють вказати атрибути моделі прогнозування. Зазвичай це найбільш трудомістка частина роботи, тому що цей крок враховує багато історичних параметрів часових рядів;

г) Верифікація отриманої моделі. Часто, потрібно взяти одну або декілька областей, які будуть використовуватися для тестування прогнозу та оцінки точності. Елементи, які використовуються для перевірки, називаються контрольними-пропускними пунктами;

г) Аналіз результатів. Якщо точність отриманих результатів на попередньому кроці є достатньою, то вважається, що модель готова до використання. В іншому випадку, можна послідовно повторити попередні кроки, починаючи з першого.

Статистичні моделі описують взаємозв'язок між минулим та майбутнім значенням за допомогою рівнянь. До них відносяться: моделі регресії (лінійна та нелінійна регресія), авторегресійні моделі (ARIMA, GARCH та ін.), модель експоненціального згладжування тощо. Структурні моделі визначають однакові залежності у вигляді певної структури та правил переходу на неї [10]. До них відносяться: моделі штучних нейронних мереж, моделі, засновані на ланцюгах Маркова, моделі моделей класифікації та регресії тощо. На сьогоднішній день найбільш поширеними моделями прогнозування є авторегресійні моделі (ARIMA), а також нейромережеві моделі (ANN) [10].

Методи прогнозування з групи «моделювання» базуються на моделях об'єкту дослідження. До них можна віднести імітаційне моделювання, яке традиційно розуміється як генерування послідовності випадкових подій чи реалізації випадкового процесу із заданими статистичними властивостями (розподілом ймовірності, автоковаріаційною функцією тощо) та оцінювання ризику за допомогою імітування критичних ситуацій в досліджуваній системі.

Імітаційне моделювання – це метод дослідження, при якому досліджувана система замінюється моделлю, яка з достатньою точністю описує реальну систему, і з нею проводяться експерименти з метою отримання інформації про цю систему [10]. Експериментування з моделлю називають імітацією (імітація – це збагнення суті явища, не вдаючись до експериментів на реальному об'єкті). Імітаційне моделювання – це окремий випадок математичного моделювання. Існує клас об'єктів, для яких з різних причин не розроблені аналітичні моделі, або не розроблені методи розрахунку отриманої моделі. В цьому випадку математична модель замінюється імітатором або імітаційною моделлю. Імітаційне статистичне моделювання являє собою чисельний метод проведення на ЕОМ обчислювальних експериментів з математичними моделями, що імітують поведінку реальних об'єктів, процесів і систем у часі протягом заданого періоду [10].

Імітаційне моделювання – це сукупність методів алгоритмізації функціонування об'єктів досліджень, програмної реалізації алгоритмічних описів, організації, планування та виконання на ЕОМ обчислювальних експериментів з математичними моделями [10]. Переваги імітаційного моделювання:

- висока адекватність між фізичною суттю описуваного процесу і його моделлю;
- можливість описати складну систему на досить високому рівні деталізації;
- значно більше областей дослідження, ніж аналітичне моделювання;
- відсутність обмежень відображення в моделі залежностей між параметрами моделі;
- можливість оцінки функціонування системи не тільки в стаціонарних станах, але і в перехідних режимах (процесах);

- одержання значної кількості даних про досліджуваний об'єкт (закон розподілу випадкових величин, числові значення абсолютні та відносні, і багато іншого);

- найбільш раціональне ставлення «результат - витрати» по відношенню до аналітичного і фізичного моделювання.

Недоліки імітаційного моделювання:

- складно оцінити ступінь точності моделі, її адекватність досліджуваному процесу,

- відносно високі вимоги до кваліфікації дослідника для написання моделі,

- спільність застосування та індивідуальність реалізації,

- розробка хорошої моделі часто обходиться дорожче, ніж аналітична і вимагає більше часу на створення і налагодження.

Динамічні системи, в яких закономірності розвитку системи моделюються на основі диференціального чи різницевого рівняння. Якщо розуміти поняття «динамічні системи» тільки як розв'язки диференціальних рівнянь (не звертаючи уваги на закономірності, які описують динамічні системи), то формально дані моделі можна віднести до статистичних, зокрема трендових [8]. У цьому розумінні, розв'язок диференціального рівняння апроксимує досліджуваний ряд. Але в даному випадку акцент робиться на системні закономірності, на яких основана побудова диференціального рівняння, що описує динаміку системи. Тому динамічні системи доцільно вважати моделями самої системи.

Потужними можливостями для прогнозування складних систем, зокрема, фінансово-економічних, є агентні моделі. Ідея їх побудови аналогічна моделям молекулярної динаміки у фізиці. Агентна модель є програмною системою, яка включає модель агентів, як окремих відносно простих підсистем досліджуваної системи, та моделі середовища їх взаємодії. Прикладом може бути система підприємств певної галузі, розташованих на певній території, які певним чином взаємодіють між собою, множина

трейдерів, які торгують на фінансовому ринку тощо. При взаємодії великої кількості агентів, з'являються стрибкоподібні властивості побудованої системи, які безпосередньо не впливають із властивостей окремих агентів. Тому навіть прості моделі поведінки агента можуть пояснювати стрибкоподібні явища в сучасних складних системах. На шляху ефективної побудови та використання агентних моделей існують певні проблеми, такі як недостатня складність у поведінці агентів, неможливість точного копіювання усіх учасників системи, проблеми калібровки, налаштування невизначених параметрів алгоритму поведінки агента та інші.

2.2 Огляд методів диференціювання часових рядів

Стаціонарний часовий ряд – це ряд, властивості якого не залежать від часу, в який цей ряд спостерігається. Таким чином, часові ряди з тенденціями або сезонністю не є стаціонарними — тренд і сезонність впливатимуть на значення часових рядів у різний час. З іншого боку, серія білого шуму є нерухомою — не має значення, коли ви її спостерігаєте, вона має виглядати майже однаково в будь-який момент часу [17].

Деякі випадки можуть вводити в оману: часовий ряд із циклічною поведінкою (але без тенденції чи сезонності) є стаціонарним. Це пояснюється тим, що цикли не мають фіксованої довжини, тому до того, як ми почнемо спостерігати за серіями, ми не можемо бути впевнені, де будуть піки та спади циклів.

Загалом, стаціонарний часовий ряд не матиме передбачуваних моделей у довгостроковій перспективі. Часові графіки покажуть, що ряд буде приблизно горизонтальним (хоча можлива певна циклічна поведінка) із постійною дисперсією.

На рисунку 2.2 зображено приклади стаціонарних і нестаціонарних часових рядів. Виходячи з цих даних, можна зробити висновки, що

очевидна сезонність виключає ряди (d), (h) та (i). Тенденції та зміни рівнів виключають ряди (a), (c), (e), (f) та (i). Збільшення дисперсії також виключає (i). Лише ряди (b) і (g) є стаціонарними [17].

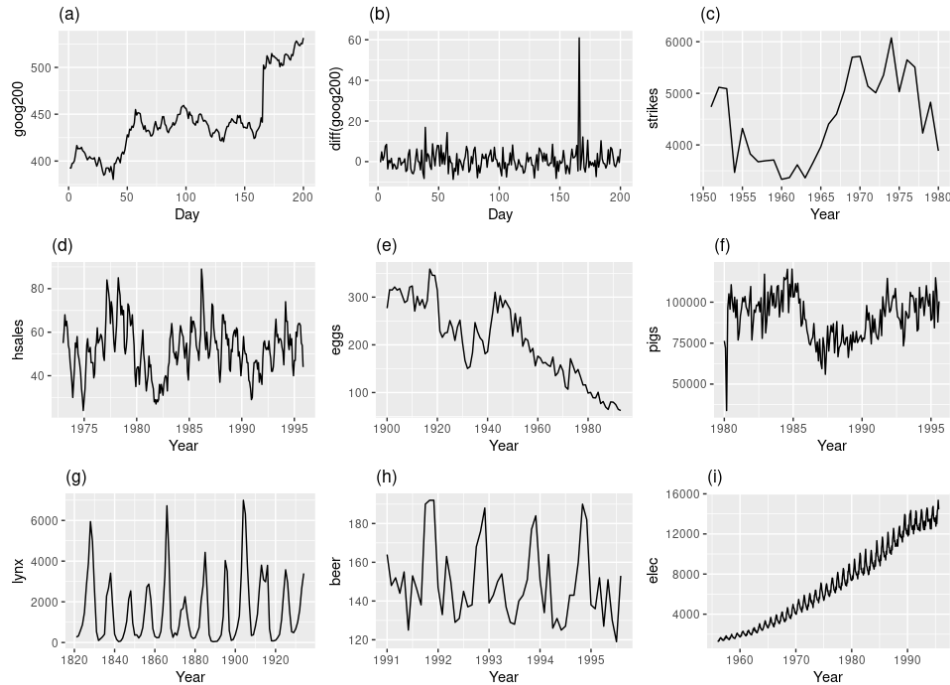


Рисунок 2.2 – Графіки стаціонарних і нестаціонарних часових рядів

Незважаючи на те, що сильні цикли в серії (g) роблять її нестаціонарною, ці цикли є аперіодичними — вони виникають, коли популяція рисі стає занадто великою для доступного корму, так що вони припиняють розмноження, а популяція падає до низького рівня, тоді відновлення їхніх джерел їжі дозволяє популяції знову зростати, і тому на довгостроковій перспективі терміни цих циклів неможливо передбачити. Отже, ряд стаціонарний.

Один із способів зробити нестаціонарний часовий ряд стаціонарним — обчислити різницю між послідовними спостереженнями. Це відомо як диференціювання. Такі перетворення, як логарифми, можуть допомогти стабілізувати дисперсію часового ряду. Розбіжність може допомогти стабілізувати середнє значення часового ряду, усуваючи зміни в рівні часового ряду, а отже, усуваючи (або зменшуючи) тенденцію та сезонність [17].

Одним з найпоширенішим методів диференціювання є модель випадкового блукання, що описується через різницевий ряд. Різницевий ряд — це зміна між послідовними спостереженнями у вихідному ряді, і його можна записати як:

$$y'_t = y_t - y_{t-1}. \quad (2.3)$$

Різні ряди матимуть лише $T - 1$ значення, оскільки неможливо обчислити різницю y'_1 для першого спостереження. Коли різницевий ряд є білим шумом, модель для вихідного ряду можна записати як:

$$y_t - y_{t-1} = \varepsilon_t, \quad (2.4)$$

де ε_t позначає білий шум. Перегрупування призводить до моделі «випадкового блукання».

$$y_t = y_{t-1} + \varepsilon_t. \quad (2.5)$$

Моделі випадкового блукання широко використовуються для нестационарних даних, зокрема фінансових та економічних даних. Випадкові блукання зазвичай мають:

- тривалі періоди очевидних тенденцій до зростання або зниження;
- раптові та непередбачувані зміни напрямку.

Прогнози моделі випадкового блукання дорівнюють останньому спостереженню, оскільки майбутні рухи є непередбачуваними та з однаковою ймовірністю будуть як вгору, так і вниз.

Також існує модель диференціювання другого порядку, яку використовують за умовою, якщо дані, що розрізняються, не

виглядатимуть стаціонарними, і може знадобитися розрізнити дані вдруге, щоб отримати стаціонарний ряд:

$$y''_t = y'_t - y'_{t-1} = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2}. \quad (2.6)$$

В цьому випадку, y''_t буде мати $T - 2$ значення. На практиці майже ніколи не потрібно виходити за межі різниць другого порядку.

Ще одним з методів диференціювання рядів є сезонне диференціювання. Сезонна різниця – це різниця між спостереженням і попереднім спостереженням того самого сезону:

$$y'_t = y_t - y_{t-m}, \quad (2.7)$$

де m – кількість пір року. Вони також називаються «lag- m різниця», оскільки ми віднімаємо спостереження після затримки m періодів. Якщо дані з сезонними відмінностями виглядають як білий шум, це означає, що відповідна модель для вихідних даних буде:

$$y_t = y_{t-m} + \varepsilon_t. \quad (2.8)$$

Прогнози цієї моделі дорівнюють останнім спостереженням відповідного сезону. Тобто ця модель дає сезонні наївні прогнози.

Одним із способів більш об'єктивно визначити, чи потрібне диференціювання часового є використання розширеного тесту Дікі-Фулера [18], який оцінює за допомогою методу найменших квадратів наступну модель:

$$y_t = \alpha + \beta_t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \dots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t, \quad (2.9)$$

де:

- α – константа,
- δ_p – часовий коефіцієнт,
- p – порядок відставання процесу авторегресії.

Нульова гіпотеза про те, що ряд належить класу нестационарних рядів, еквівалентна тому, що $\delta = 0$. Перша гіпотеза, в свою чергу, стверджує що ряд належить до класу стаціонарних рядів, якщо $\delta < 0$.

При практичному використанні тесту Дікі-Фулера важливо правильно специфікувати регресійну модель. Для досліджуваних рядів можливі випадки відсутності тренду, вільного члену чи вибору різної довжини лагу δ . Неправильна специфікація тесту може привести до неадекватних результатів [19].

3 ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

3.1 Змістовна постановка задачі

В таблиці 3.1 наведено вхідний набір даних, який був отриманий в результаті обліку заявок громадян в центрах надання гуманітарної допомоги. Таких центрів може бути як один, так і декілька. Набір даних зображено в короткому вигляді для того, щоб інші дані не заважали при вирішенні задачі. Спочатку необхідно розбити цей набір даних на декілька маленьких наборів в залежності від кількості видів допомоги (на кожен вид допомоги окремий набір). Також, необхідно зробити дані агрегованими. Групування необхідно робити за колонками «Датою створення заявки» та «ID центру», а всередині повинна бути агрегація підрахунку рядку «ID користувача», а також суму рядку «Витрати». Також, необхідно зауважити, що працювати можна тільки

із заявками зі статусом «Оброблено». Наприклад, для виду допомоги «Продукти харчування» вибірка даних зображена в таблиці 3.2, а агрегування даних зображено в таблиці 3.3.

Таблиця 3.1 – приклад набору даних, зібраного центром надання гуманітарної допомоги

ID	ID центру	Дата створення заявки	ID користувача	Статус заявки	Вид допомоги	Витрати (грн)
1	0001	01.09.2022	1101	Оброблено	Продукти харчування	625
2	0001	01.09.2022	1101	Оброблено	Одяг	235
3	0001	01.09.2022	1101	Оброблено	Фінансова допомога	500
4	0001	01.09.2022	1102	Оброблено	Одяг	473
5	0001	01.09.2022	1103	Оброблено	Житло	23232
6	0001	01.09.2022	1104	Оброблено	Продукти харчування	944
7	0001	01.09.2022	1104	Оброблено	Одяг	600
8	0001	02.09.2022	1105	Оброблено	Одяг	786
9	0001	02.09.2022	1106	Оброблено	Житло	18123

Продовження таблиці 3.1

10	0001	02.09.2022	1107	Оброблено	Фінансова допомога	2000
11	0001	02.09.2022	1108	Оброблено	Фінансова допомога	2000
12	0001	02.09.2022	1110	Оброблено	Продукти харчування	640
13	0001	02.09.2022	1111	Оброблено	Продукти харчування	700

99999	0002	01.09.2022	1109	Не оброблено	Продукти харчування	Не відомо

Таблиця 3.2 – Вибірка даних за видом допомоги «Продукти харчування»

ID	ID центру	Дата створення заявки	ID користувача	Статус заявки	Вид допомоги	Витрати (грн)
1	0001	01.09.2022	1101	Оброблено	Продукти харчування	625

4	0001	01.09.2022	1104	Оброблено	Продукти харчування	944
8	0001	02.09.2022	1110	Оброблено	Продукти харчування	640
9	0001	02.09.2022	1111	Оброблено	Продукти харчування	700

Таблиця 3.3 – Вибірка даних з агрегацією за видом допомоги «Продукти харчування»

Дата	ID центру	Кількість користувачів	Сума витрат
01.09.2022	0001	2	1569
02.09.2022	0001	2	1340

Після того, як будуть отримані набори даних з агрегацією, з ними вже можна виконувати певні дії. Можна як зобразити поточні дані у вигляді графіку, так і почати процес прогнозування. В даній роботі акцент робиться на дослідженні методів прогнозування. Тому необхідно розробити технологію, що виконує перетворення вхідних даних на декілька наборів з агрегацією, та застосовувати на даних наборах метод прогнозування, який в якості вихідних даних «продовжить» вхідний набір новими даними, але за наступний день або тиждень, а також зробить графік для того, щоб візуалізувати дані. Приклад візуалізації прогнозованих даних наведено на рисунку 3.4 [15].

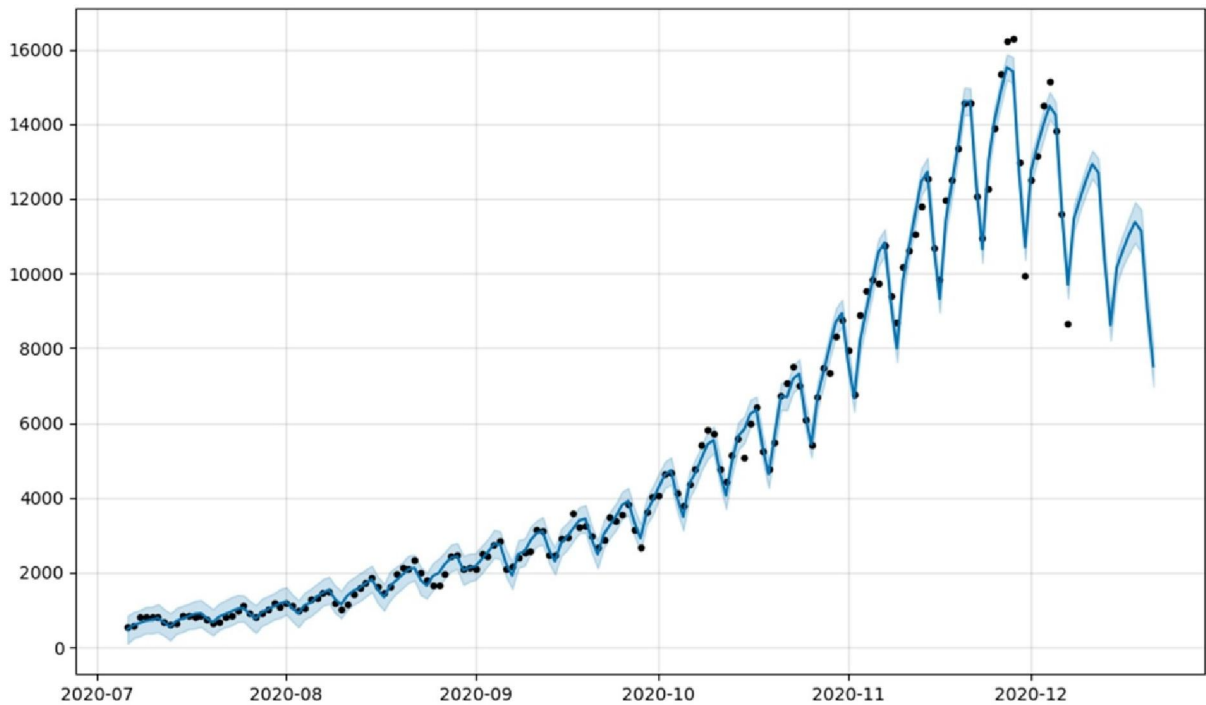


Рисунок 3.4 – Приклад графіку прогнозування даних

3.2 Постановка задачі

Виходячи з цього, можна сформулювати задачу дослідження – розробити та здійснити програмну реалізацію методу прогнозування даних центрів надання гуманітарної допомоги, провести прогнозування даних для отримання нових, ще невідомих наборів даних.

Розроблений метод має відповідати наступним вимогам:

- а) має отримувати вхідний набір даних, робити вибірку за декількома умовами, проводити агрегацію даних, перетворювати набір даних;
- б) має виконувати процес прогнозування даних за певним методом;
- в) має повертати результуючі набори даних;
- г) має створювати графіки на основі вихідних даних.

Для досягнення зазначеної мети необхідно:

- а) обрати метод прогнозування, який найбільш підходить для вирішення задач виходячи із вхідних даних та очікуваних вихідних даних;

- б) зробити математичний опис задачі, опис методу прогнозування, алгоритми обробки вхідних даних;
- в) обґрунтувати вибір визначеного для розробки методу прогнозування;
- г) виконати розробку методу прогнозування;
- г) виконати розробку програмного застосунку для проведення експериментальних досліджень;
- д) провести експериментальні дослідження для визначення ефективності розробленого методу.

4.1 Порівняння методів прогнозування

Виходячи з вище оглянутих методів прогнозування даних та постановки задачі, можна зробити певні висновки. По-перше, інтуїтивні методи прогнозування не підходять для дослідження, тому що вони мають істотний недолік, який полягає в тому, що експерти заздалегідь зорієнтовані на компромісне вирішення проблеми, а це у свою чергу припускають достатнє перекручування результатів прогнозування [8]. По-друге, динамічні системи та імітаційні моделі підходять більше до прогнозування даних в складних системах, для яких притаманні великі обсяги даних, багатомірність, стрибкоподібність та динаміка. В контексті даного дослідження методів прогнозування в центрах надання гуманітарної допомоги в якості вхідних даних виступатиме невеликий за обсягом набір даних, в якому дані змінюються за певним часом. Окрім того, немає достатньої кількості інформації щодо поведінки роботи центрів надання гуманітарної допомоги, тобто немає імовірнісних оцінок настання певних подій, які можуть виникнути під час роботи. Отже, динамічні та імітаційні моделі не підходять.

Тому доцільним є дослідити один із статистичних методів прогнозування, тому що в якості вхідних даних буде виступати часовий ряд – набір даних за певний проміжок часу, в якому буде інформація щодо кількості людей, що звернулись в певні центри надання гуманітарної допомоги, дату звернення та кількість витрачених ресурсів на одну людину. Для даного дослідження найбільш підходить авторегресивний метод ARIMA.

Основна перевага прогнозування ARIMA полягає в тому, що для нього потрібні дані лише про часові ряди, про які йдеться. По-перше, ця функція вигідна, якщо прогнозується велика кількість часових рядів. По-друге, це дозволяє уникнути проблеми, яка іноді виникає з багатоваріантними моделями. Наприклад, модель, що включає заробітну плату, ціни та гроші

[14]. Можливо, послідовний грошовий ряд доступний лише протягом коротшого періоду часу, ніж два інших ряду, що обмежує період часу, за який можна оцінити модель. По-третє, з багатовимірними моделями своєчасність даних може бути проблемою. Якщо побудувати велику структурну модель, яка містить змінні, які публікуються лише з великим відставанням, наприклад дані про заробітну плату, тоді прогнози, що використовують цю модель, є умовними прогнозами, заснованими на прогнозах недоступних спостережень, що додає додаткове джерело невизначеності прогнозу.

4.2 Математичний опис задачі прогнозування в центрах надання гуманітарної допомоги

Дана задача відноситься до класу часових рядів. Особливістю прогнозування часових рядів є те, що аналізуються лише дані спостережень без додаткової інформації, без аналізу впливу зовнішніх сил. Звичайно, такий аналіз виглядає досить неповним, але доволі часто прогнози часових рядів є більш точними. Нехай y_1, y_2, \dots, y_T – значення спостережень за деяким процесом протягом T періодів. Ця послідовність є числовими значеннями, кожне з яких має відповідний індекс, який залежить від номера періоду, в який він спостерігався. Така послідовність, записана у порядку зростання індексу, називається часовим рядом. Будемо позначати часовий ряд з T елементами $\{y_T\}$ [16].

Класичний підхід до побудови моделі часового ряду полягає у розкладі його на декілька компонент, природа кожної з яких принципово відмінна від інших. Далі кожна складова аналізується специфічними для неї методами. Найчастіше на практиці розглядають розклад часового ряду $y_t: y_t = tr_t + s_t + r_t$. Компонента tr_t відповідає складовій, яка змінюється повільно і називається трендом. Це може бути як просто стала, так і многочлен,

експонента чи інша подібна функція. Тренд звичайно описує довготривалі тенденції явища. Компонента s_t відповідає скалодовій, яка змінюється періодично і називається сезонною або циклічною компонентою. Ця компонента часто описує сезонні явища у виробництві протягом року або економічні циклічності. Звичайно складові tr_t та s_t детерміністичні функції. За випадковість у спостереженнях відповідає r_t . Основний напрямок розвитку теорії часових рядів якраз є моделювання і аналіз процесу r_t [16].

4.3 Розробка методу прогнозування даних в центрах надання гуманітарної допомоги

На рисунку 4.1 представлена блок-схема методу прогнозування даних в центрах надання гуманітарної допомоги. Умовно даний метод поділяється на три частини:

- а) Задання початкових даних для прогнозування;
- б) Дослідження часового ряду на стаціонарність;
- в) Проведення прогнозування.

Для задання початкових даних для прогнозування необхідний набір даних, за яким буде проводитися процес прогнозування. Перш за все, необхідно, щоб даний набір даних мав характеристику часового ряду, а саме атрибути, які визначають час, дату або дату і час. Також необхідно мати вимірний атрибут, який показує зміну стану за часом. Отже, необхідний набір даних повинен мати два атрибути: часовий та числовий. Не виключається ситуація, коли подається набір даних, який потребує додаткової фільтрації, наприклад, виключення записів із пустими значеннями. Тому перший етап роботи методу складається із задання всіх необхідних змінних, даних та атрибутів для проведення дослідження. Також перший етап включає додаткову фільтрацію, а також сортування даних.

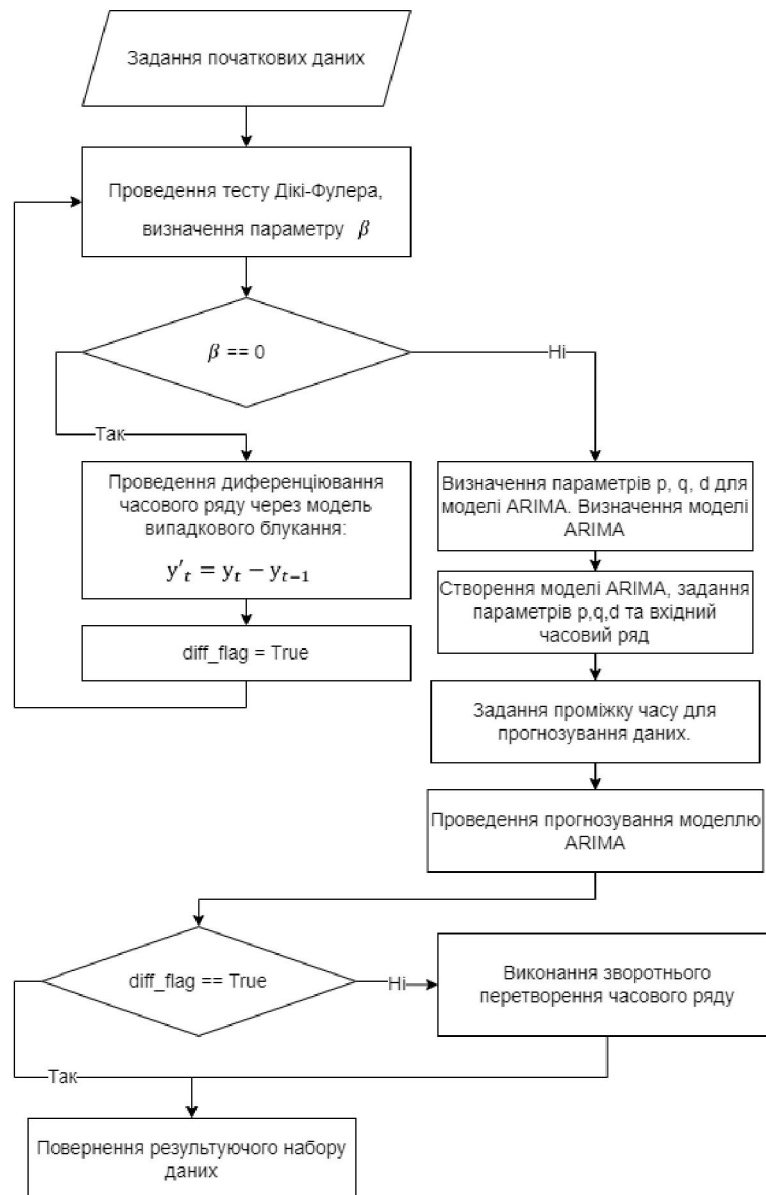


Рисунок 4.1 – Блок схема методу прогнозування

Другий етап методу складається з перевірки часового ряду (в контексті даного методу – набору даних) на стаціонарність. Для цього, буде використовуватися тест Дікі Фулера, описаний у формулі 2.9. Кінцевим результатом тесту є отримання змінної β . Далі, в залежності від значення змінної β , проходить перевірка нульової та першої гіпотези. В разі, якщо підтверджується нульова гіпотеза, це означає, що досліджуваний набір даних не є стаціонарним, а отже потребує диференціювання.

В якості методу диференціювання часового ряду обрано модель випадкового блукання, що наведена в формулі 2.3. Диференціювання

виконується шляхом віднімання попереднього спостереження від поточного спостереження. Необхідно звернути увагу на те, що після проведення прогнозування даних, є необхідним провести зворотнє перетворення. Інвертування процесу необхідне, коли передбачення потрібно перетворити назад у вихідний масштаб [20]. Цей процес можна скасувати, додавши спостереження на попередньому кроці часу до значення різниці:

$$y'_t = y'_t + y'_{t-1}. \quad (4.1)$$

За основу третього етапу методу прогнозування обрано популярний і широко застосовуваний статистичний метод прогнозування часових рядів – модель ARIMA – авторегресивна інтегрована ковзаюча середня. Це клас моделі, який фіксує набір різних стандартних часових структур у даних часових рядів. Кожен з цих компонентів явно вказаний у моделі як параметр. Для ARIMA(p, d, q) використовується стандартне позначення, де параметри заміщені цілими значеннями для швидкої вказівки конкретної моделі ARIMA, що використовується. Разом ці три параметри враховують сезонність, тенденцію та шум у наборах даних [10–11].

Параметри моделі ARIMA визначаються таким чином:

– p – порядок авторегресії (AR), який дозволяє додати попередні значення часового ряду. Цей параметр можна проілюструвати твердженням «завтра, ймовірно, буде тепло, якщо в останні три дні було тепло»,

– d – порядок інтегрування (I; порядок різниць вихідного часового ряду). Він додає в модель поняття різниці часових рядів (визначає кількість минулих часових точок, які потрібно вилучити з поточного значення). Цей параметр ілюструє таке твердження: «завтра, ймовірно, буде така сама температура, якщо різниця в температурі за останні три дні була дуже мала» [11],

– q – порядок змінного середнього (МА), який дозволяє встановити похибку моделі як лінійну комбінацію значень помилок, що спостерігалися раніше. Для відстеження сезонності використовується сезонна модель ARIMA – ARIMA(p, d, q) (P, D, Q) s . Тут (p, d, q) – несезонні параметри, описані вище, а (P, D, Q) слідує тим самим визначенням, але застосовуються до сезонної складової часового ряду. Параметр s визначає періодичність часового ряду (4 – квартальні періоди, 12 – річні періоди і т.д.) [11]. AR(p) – авторегресивна модель по ряду p .

Модель має вид:

$$Y(t) = f_0 + f_1Y(t-1) + f_2Y(t-2) + \dots + f_pY(t-p) + E(t), \quad (4.2)$$

де $Y(t)$ – залежна змінна у момент часу t ; $f_0, f_1, f_2, \dots, f_p$ – оцінювані параметри; $E(t)$ – помилка від впливу змінних, які не враховуються в даній моделі.

Завдання полягає в тому, щоб визначити $f_0, f_1, f_2, \dots, f_p$. Їх можна оцінити різними способами. Найправильніше шукати їх через систему рівнянь Юла-Уолкера, для складання цієї системи буде потрібен розрахунок значень автокореляційної функції. Можна застосувати простіший спосіб – порахувати їх методом найменших квадратів [11–13].

Розглянемо авторегресійне ковзне середнє ARMA(p, q). Під позначенням ARMA(p, q) розуміється модель p авторегресійних складових, що містить q ковзаючих середніх. Точніше, модель ARMA(p, q) включає моделі AR(p) і MA(q):

$$X_t = c + e_t + \sum_{i=1}^q \theta_i e_{t-i} + \sum_{i=1}^p \varphi_i X_{t-i}. \quad (4.3)$$

Зазвичай, значення помилки e_t вважають незалежними однаково розподіленими випадковими величинами, узятими з нормального розподілу з нульовим середнім:

$$e_t \sim N(0, \sigma^2), \quad (4.4)$$

де σ^2 – дисперсія [12].

Припущення можна ослабити, але це може привести до зміни властивостей моделі. Наприклад, якщо не припускати незалежності і однакового розподілу помилок, поведінка моделі суттєво зміниться. Лінійна модель регресії будується, включаючи вказану кількість та тип термінів, і дані готуються за ступенем диференціації для того, щоб зробити її нерухомою, тобто видалити трендові та сезонні структури, що негативно впливають на регресійну модель. Для параметра можна застосовувати значення 0, яке вказує на те, що не використовувати цей елемент моделі. Таким чином, модель ARIMA може бути налаштована для виконання функції моделі ARMA і навіть простої моделі AR, I або MA. Прийняття моделі ARIMA для часового ряду передбачає, що базовий процес, що породив спостереження, є процесом ARIMA. Це може здатися очевидним, але допомагає мотивувати необхідність підтвердження припущень моделі в необґрунтованих спостереженнях та залишкових помилках прогнозів з моделі.

На третьому етапі методу параметри моделі ARIMA будуть визначатися автоматично. Зазвичай у базовій моделі ARIMA нам потрібно надати значення p , d і q , які є важливими. Ми використовуємо статистичні методи для створення цих значень, виконуючи різницю для усунення нестационарності та будуючи графіки ACF і PACF. У Auto ARIMA сама модель генеруватиме оптимальні значення p , d і q , які підходять для набору даних для кращого прогнозування. Для цього буде використана стороння

функція, що буде приймати на вхід набір даних, та в результаті повертати статистичні дані. Приклад роботи функції, що буде автоматично визначати параметри наведено на рисунку 4.2. Дана статистика надає всю інформацію про нашу модель, зокрема параметри моделі p , d , q , а також значення AIC і BIC, які надають інформацію про те, чи хороша модель.

Dep. Variable:	y	No. Observations:	61
Model:	SARIMAX(4, 1, 0)	Log Likelihood	343.479
Date:	Mon, 21 Dec 2020	AIC	-674.957
Time:	19:48:50	BIC	-662.391
Sample:	0	HQIC	-670.042
	- 61		
Covariance Type:	opg		

	coef	std err	z	P> z	[0.025	0.975]
intercept	0.0013	0.000	3.365	0.001	0.001	0.002
ar.L1	1.7404	0.136	12.836	0.000	1.475	2.006
ar.L2	-1.5144	0.256	-5.927	0.000	-2.015	-1.014
ar.L3	0.9586	0.258	3.717	0.000	0.453	1.464
ar.L4	-0.3943	0.129	-3.054	0.002	-0.647	-0.141
sigma2	5.86e-07	1.37e-07	4.289	0.000	3.18e-07	8.54e-07

Ljung-Box (L1) (Q):	0.68	Jarque-Bera (JB):	1.67
Prob(Q):	0.41	Prob(JB):	0.43
Heteroskedasticity (H):	1.15	Skew:	-0.37
Prob(H) (two-sided):	0.76	Kurtosis:	2.67

Рисунок 4.2 – Результат роботи функції автоматичного визначення параметрів для моделі ARIMA

Останнім кроком третього етапу буде задання проміжку часу, за яким модель ARIMA далі виконуватиме прогнозування. У разі, якщо часовий ряд буф диференційований на другому етапі, буде виконуватися інвертування часового ряду за допомогою формули 4.1. Інакше, якщо часовий ряд не був диференційований, метод повертає прогнозований часовий ряд без зворотніх перетворень.

Отже, перевагою даного методу є автоматизовані процеси дослідження часового ряду на стаціонарність, диференціювання нестаціонарного ряду, вибір найкращих параметрів p , q , d для моделі ARIMA.

4.4 Опис потрібного інструментарію та його застосування.

Для реалізації методу прогнозування даних, описаного в розділі 4.3, будуть використані наступні інструменти:

а) Мова програмування Python. Дана мова програмування досить часто використовується при аналізі даних та машинному навчанні, оскільки має широкий спектр спеціалізованих бібліотек, які мають достатній набір функцій та класів. Однією з переваг Python є простота. Легко зрозумілий і читабельний код і короткий синтаксис дуже корисні для процесу Data Science і Machine Learning. Основна мета ML – скоротити людські зусилля, а Python зменшує людські зусилля на етапі кодування. Простота та послідовність Python спрощують роботу навіть для початківців розробників. У машинному навчанні використання Python допомагає зосередитися на вирішенні логічних проблем, а не витратити час на дрібниці мови програмування. Це одна з основних причин, чому розробники погоджуються, що наука про дані та машинне навчання за допомогою Python є найкращим вибором. Конкретно для вирішення задачі прогнозування даних перш за все будуть використані бібліотеки pandas та numpy, основний функціонал яких є обробка масивів або наборів даних. Також буде використана бібліотека matplotlib, основною функцією якої є візуалізація даних у вигляді діаграм, гістограм, таблиць і т.д. Для реалізації методу прогнозування даних будуть використані наступні спеціалізовані бібліотеки: statsmodels.api та pmdarima. А також допоміжні бібліотеки re та datetime;

б) Реляційна СУБД Microsoft SQL Server. Microsoft SQL Server є однією з основних систем керування реляційними базами даних на ринку, яка обслуговує широкий спектр програмних додатків для бізнес-аналітики та аналізу в корпоративних середовищах. Заснований на мові Transact-SQL, він містить набір стандартних мовних розширень програмування, а його додаток доступний для використання як локально, так і в хмарі. Необхідність

використання бази даних обумовлена тим, що мова запитів SQL дозволяє швидко створювати вибірку та фільтрацію даних, а наявність додаткового інструменту Data Generator for SQL Server дозволяє згенерувати набір даних довільного об'єму та розміру;

в) Google Colab. Colab – це безкоштовне середовище для блокнотів Jupyter, написаних на мові програмування Python, яке повністю працює в хмарі. Необхідність використання даної IDE обумовлена тим, що Colab не потребує налаштування, а також підтримує більшість популярних бібліотек машинного навчання, які можна легко завантажити у свій блокнот [21];

г) Google Drive. Даний сервіс є сховищем даних, що дозволяє користувачам зберігати свої дані на серверах у хмарі і ділитися ними з іншими користувачами в Інтернеті [22]. Оскільки Google Colab має інтеграцію з Google Drive, можна використовувати хмарний диск як сховище для файлів з даними, а потім використовувати шлях для цих файлів напряму в Google Colab;

Отже, для вирішення задачі прогнозування даних в центрах надання гуманітарної допомоги SQL Server буде застосовуватися для генерації даних, зберігання наборів даних а також для експорту кінцевого набору даних в Google Drive, на якому буде проведено дослідження. Мова програмування Python буде використана для реалізації методу прогнозування та проведення експериментальних досліджень в середовищі Google Colab.

5 ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ

5.1 Розробка вимог до програмного засобу

Для проведення дослідження методу прогнозування даних в центрах надання гуманітарної допомоги, буде створений програмний застосунок із методом прогнозування, що задовольнятиме наступним вимогам:

а) розроблений метод має отримувати вхідний набір даних, робити вибірку за декількома умовами, проводити агрегацію даних, перетворювати набір даних;

б) розроблений метод має виконувати процес прогнозування даних за певним методом;

в) розроблений метод має повертати результуючі набори даних;

г) розроблений метод має створювати графіки на основі вихідних даних [23].

Для цього буде створено окрема функція в Jupyter Notebook, яка буде приймати всі необхідні вхідні параметри, та виконувати описані вище дії. Окрім того, буде розроблено окремий програмний код із покроковим запуском команд для проведення експериментальних досліджень. Перевага даного застосунку буде в тому, що програмний код можна розташовувати в довільному порядку, а також запускати його, не запускаючи інші команди. Особливо важливим цей програмний код буде в разі, коли необхідно поділити вхідний набір даних на тренований та тестовий набори, змінити вхідні параметри для моделі ARIMA або зобразити проміжні результати та графіки.

5.2 Розробка вимог до моделі бази даних

Для дослідження розробленого методу прогнозування даних в центрах надання гуманітарної допомоги, необхідно створити базу даних, в якій буде зберігатися інформація стосовно користувачів, запитів, гуманітарних центрів та видів гуманітарної допомоги. Далі ці дані необхідно буде зібрати в один набір даних із інформацією, необхідною для проведення експериментальних досліджень. На рисунку 5.1 наведено ER-діаграму бази даних центрів надання гуманітарної допомоги.

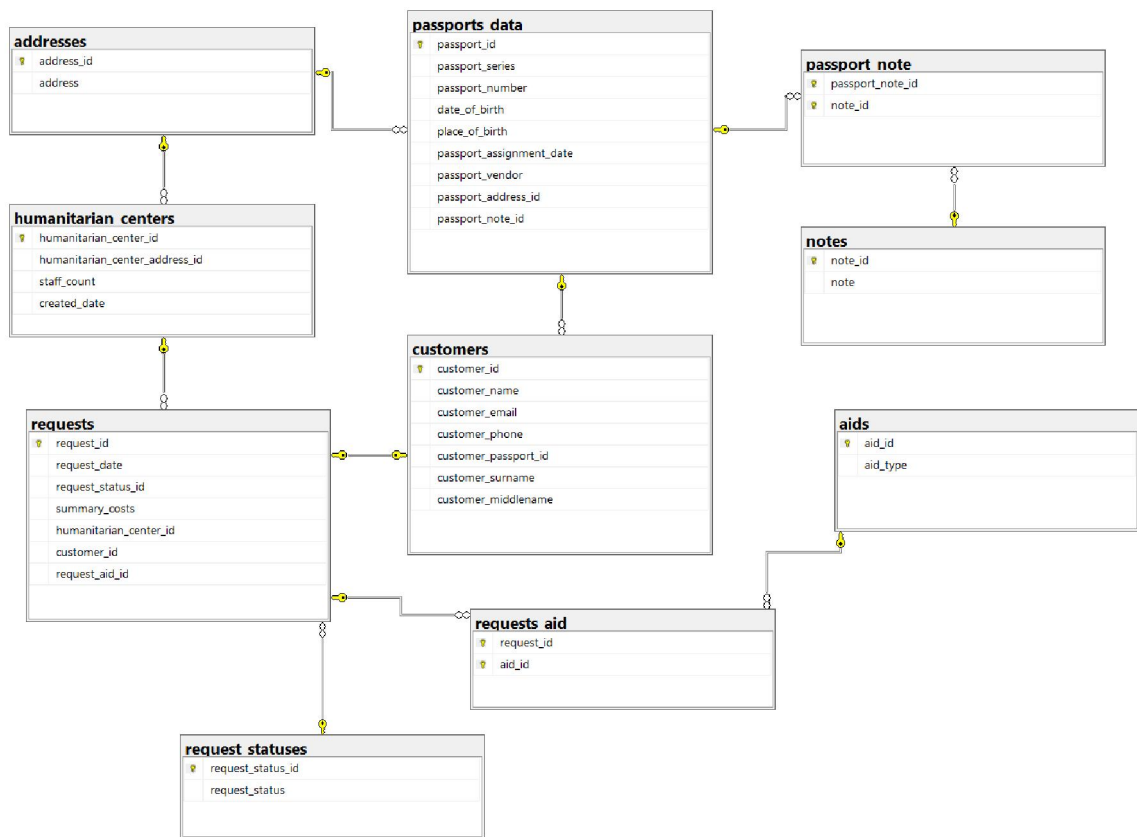


Рисунок 5.1 – ER діаграма бази даних центрів надання гуманітарної
ДОПОМОГИ

Дана модель складається з наступних таблиць:

а) **customers** – дана таблиця зберігає дані про користувачів, які створювали запити на отримання гуманітарної допомоги. Дана таблиця має такі атрибути, як `customer_name`, `customer_email`, `customer phone`,

customer_passport_id, customer_surname, customer_middlename, що описують загальну інформацію про користувача;

б) passport_data – дана таблиця має зв'язок з таблицею customers, та зберігає паспортні дані користувачів, що ідентифікують їх як громадян держави. Необхідність створення відношення між таблицею customers обумовлено тим, що один користувач може мати декілька паспортів різного типу. Дана таблиця має такі атрибути, як passport_id, passport_series, passport_number, date_of_birth, place_of_birth, passport_assignment_date, passport_vendor, passport_address_id, passport_typem passport_address_id, passport_note_id;

в) requests – дана таблиця зберігає дані запитів користувачів на отримання гуманітарної допомоги. Вона має такі атрибути, як request_id, request_date, request_status_id, summary_costs, humanitarian_center_id, customer_id, request_aid_id, summary_costs;

г) request_aid – дана таблиця є розв'язуванням відношення M-N між таблицями requests та aids, та має атрибути request_aid_id та aid_id;

г) ais – дана таблиця зберігає дані про різні види гуманітарної допомоги. Вона має наступні атрибути: aid_id, aid_type;

д) request_statuses – дана таблиця зберігає дані про статус обробки запитів користувачів на отримання гуманітарної допомоги. Вона має такі атрибути, як request_status_id, request_status;

е) humanitarian_centers – зберігає дані про центри надання гуманітарної допомоги. Дана таблиця має наступні атрибути: humanitarian_center_id, humanitarian_center_address_id, created_date, staff_count;

е) passport_note – дана таблиця є розв'язуванням відношення M-N між таблицями passport_data та notes, та має наступні атрибути: request_aid_id, aid_id;

ж) notes – дана таблиця зберігає дані особливих приміток в паспортних даних, та має атрибути note_id, note;

з) addresses – дана таблиця зберігає дані адрес, та має атрибути address_id та address.

5.3 Розробка запитів до бази даних

Оскільки вимоги до вхідних даних описані в розділі 3.1, необхідно розробити запит, для отримання набору даних, який в подальшому буде переданий як вхідний параметр до методу прогнозування. SQL запит для отримання вхідних даних наведено у лістингу 5.2.

Лістинг 5.2 – SQL-запит для отримання вхідного набору даних

```
select
req.request_id as ID,
hc.humanitarian_center_id as humanitarian_center_id,
req.request_date as request_date,
c.customer_id as customer_id,
rs.request_status as request_status,
a.aid_type as humanitarian_aid_type,
req.summary_costs as summary_costs
from dbo.requests as req
inner join dbo.humanitarian_centers as hc on hc.humanitarian_center_id
= req.humanitarian_center_id
inner join dbo.customers as c on c.customer_id=req.customer_id
inner      join      dbo.request_statuses      as      rs      on
rs.request_status_id=req.request_status_id
inner join dbo.requests_aid as ra on ra.request_id = req.request_id
inner join dbo.aids as a on a.aid_id = ra.aid_id
where rs.request_status != 'Not Processed' and
a.aid_type is not null and
req.summary_costs is not null and
```

```
req.request_date >= '2022-01-01'
```

Даний запит ілюструє, що необхідно зробити вибірку із таблиці requests для отримання ідентифікатору запиту (request_id), дату створення запиту (request_date) та суму витрачених коштів (summary_costs). Також дана таблиця поєднується з наступними таблицями:

- humanitarian_centers – для отримання ідентифікатору центру надання гуманітарної допомоги (humanitarian_center_id),
- customers – для отримання ідентифікатору користувача (customer_id),
- request_statuses – для отримання статусу обробки запиту (request_status)
- requests_aid, aids – для отримання типу гуманітарної допомоги (aid_type).

Також описані умови для фільтрації записів, а саме: статус обробки запиту не має бути рівним значенню «Not Processed», тип гуманітарної допомоги не повинен бути порожнім, витрачені кошти не повинні бути порожніми а також дата створення запиту повинна бути більше ніж 2022-01-01. Результат виконання запиту наведено на рисунку 5.3.

Results		Messages					
	ID	humanitarian_center_id	request_date	customer_id	request_status	humanitarian_aid_type	summary_costs
1	1	1001	2022-12-24	133756	Not processed	Food	5315.60
2	2	1001	2022-05-17	174204	Not processed	Financial support	7221.30
3	3	1002	2022-05-06	142017	Processed	Cloth	3283.00
4	4	1004	2022-06-06	134755	Processed	Shelter	12741.00
5	5	1001	2022-03-15	187904	Not processed	Food	3966.40
6	6	1001	2022-02-13	169410	Not processed	Financial support	7957.30
7	7	1001	2022-05-10	184948	Processed	Cloth	2466.10
8	8	1004	2022-07-09	174059	Processed	Shelter	12291.00
9	9	1003	2022-12-14	164218	Processed	Shelter	11551.00
10	10	1002	2022-05-31	129316	Not processed	Financial support	7620.60
11	11	1002	2022-06-24	134327	Processed	Cloth	1900.50
12	12	1004	2022-07-07	183827	Not processed	Food	1121.50
13	13	1003	2022-05-07	166964	Processed	Cloth	1368.80
14	14	1003	2022-09-01	125651	Not processed	Financial support	7981.90
15	15	1002	2022-05-12	168197	Not processed	Food	2246.60
16	16	1002	2022-07-30	160941	Processed	Shelter	10600.00
17	17	1001	2022-11-21	110486	Processed	Shelter	11756.00
18	18	1004	2022-10-06	157943	Not processed	Food	3890.00
19	19	1003	2022-11-03	183179	Processed	Financial support	6296.40

Рисунок 5.3 – Результат виконання SQL запиту для отримання вхідних даних

Отже, результат даного запиту повністю задовольняє вимогам, наведеними в розділі 3.1 і можуть бути використані як вхідний набір даних для дослідження методу прогнозування в центрах надання гуманітарної допомоги.

5.4 Програмна реалізація методу прогнозування даних в центрах надання гуманітарної допомоги

Відповідно до розділу 4.4 розроблений метод прогнозування складається з трьох етапів:

Розроблено етап задання початкових даних. Передумовою цього етапу є те, що набір даних, сформований у розділі 5.3, був експортований у форматі .csv (comma separated values) на сховище у Google Drive. Спочатку даний набір даних зчитується за допомогою бібліотеки pandas, а саме метод `read_csv()`. У функцію задається шлях до файлу у сховищі Google Drive, а також параметр `dtype`, який відповідає за задання схеми типів даних.

Також задаються дані щодо атрибутів, які будуть обчислюватися. По-перше, задається параметр для фільтрації набору даних за атрибутом `humanitarian_center_id` та `humanitarian_aid_type`. По-друге, задаються параметри агрегації. Оскільки ми повинні отримати часовий ряд для проведення прогнозування, необхідно зробити групування всіх значень за атрибутом `request_date` (часовий параметр), а далі зробити агрегацію суми, максимуму, мінімуму або будь яку іншу функцію агрегації на атрибут `summary_costs`, тому метод також приймає параметри `groupby_column`, `agg_column`, `agg_type`.

В ході розробки другого етапу методу прогнозування було виконано програмну реалізацію формул (2.3)(2.4)(2.9). В ході їх використання проведено тест Дікі-Фулера на перевірку вхідного часового ряду на стаціонарність. Було додано розгалуження програмного коду, при цьому в залежності від значення параметру β буде або проведено диференціювання

часового ряду, або ні. Далі будуються графіки автокореляційної функції (ACF) та часткової автокореляційної функції (PACF).

В ході розробки третього етапу методу прогнозування було виконано програмну реалізацію формул (4.2)(4.3)(4.4). В ході їх використання виконано автоматичний підбір параметрів p , d , q для моделі ARIMA. Дані параметри використовуються під час створення моделі ARIMA, окрім цього, також задаються сезонні параметри за замовчуванням а також вхідний набір даних. Далі виконується тренування моделі, після чого в модель подається функція `predict()`, в яку задаються параметри проміжку часу, на який необхідно зробити прогноз.

Далі, в залежності від того, чи було проведення диференціювання вхідного набору даних, виконується інвертування результуючого набору даних. Потім результуючий набір даних відображається у вигляді графіку даних, і в якості результату метод повертає прогнозований набір даних.

5.6 Проведення експериментальних досліджень методу прогнозування даних в центрах надання гуманітарної допомоги

Для початку проведення експериментальних досліджень було завантажено файл з даними `export_hum_aid.csv` у сховище Google Drive. Після зчитування файлу, було проведено фільтрацію вхідного набору даних за наступними критеріями:

- ідентифікатор центру надання гуманітарної допомоги (`humanitarian_center_id`) дорівнює «1001»,
- тип гуманітарної допомоги (`humanitarian_aid_type`) дорівнює «Food»,

Окрім того проведено групування даних за атрибутом `request_date` та агрегацію суми за атрибутом `summary_costs`. Результат вибірки даних наведено у вигляді графіку на рисунку 5.4. Даний результат відображає середню суму витрат на одну людину по датам.

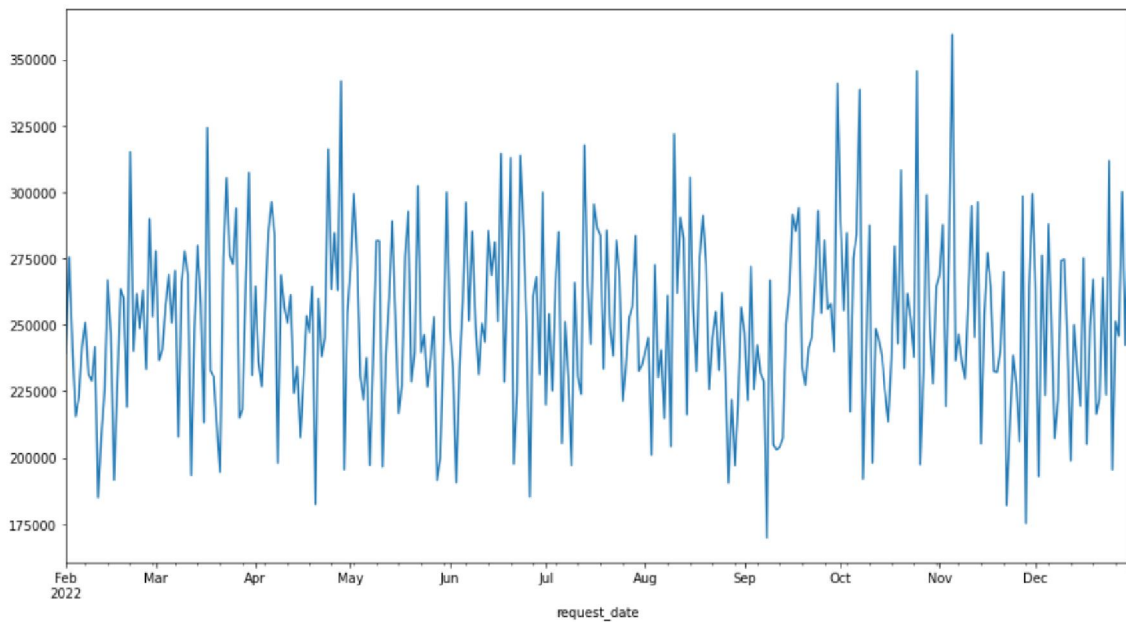


Рисунок 5.4 – Графік середньої суми витрат на день в центрі надання гуманітарної допомоги

Перш за все, необхідно дати оцінку даному часовому ряду на стаціонарність, тому було використано функцію, що виконує тест Дікі-Фулера. Результат тесту наведено на рисунку 5.5

```
adfuller_test(humanitarian_aid_df_filtered_1_food_grouped)
```

```
ADF Test Statistic : -13.205946585473797
```

```
p-value : 1.0708056273671644e-24
```

```
#Lags Used : 0
```

```
Number of Observations Used : 212
```

```
strong evidence against the null hypothesis(Ho), reject the null hypothesis. Data has no unit root and is stationary
```

```
True
```

Рисунок 5.5 – Результат тесту Дікі-Фулера на часовому ряді

Виходячи із наведеного результату тесту, параметр ADF Test Statistic менший за нуль, а отже підтверджується перша гіпотеза. Високе значення p-value дає можливість стверджувати, що нульова гіпотеза про рівність середніх вірна, що свідчить про стаціонарності низки. Отже даний часовий ряд є стаціонарним та не потребує додаткового диференціювання.

Також було створено графіки автокореляційної функції (ACF) та часткової автокореляційної функції (PACF) для заданого часового ряду. Дані функції допомагають зрозуміти, які найкращі параметри p та q слід передати в модель ARIMA. Результат створення графіків наведено на рисунку 5.5.

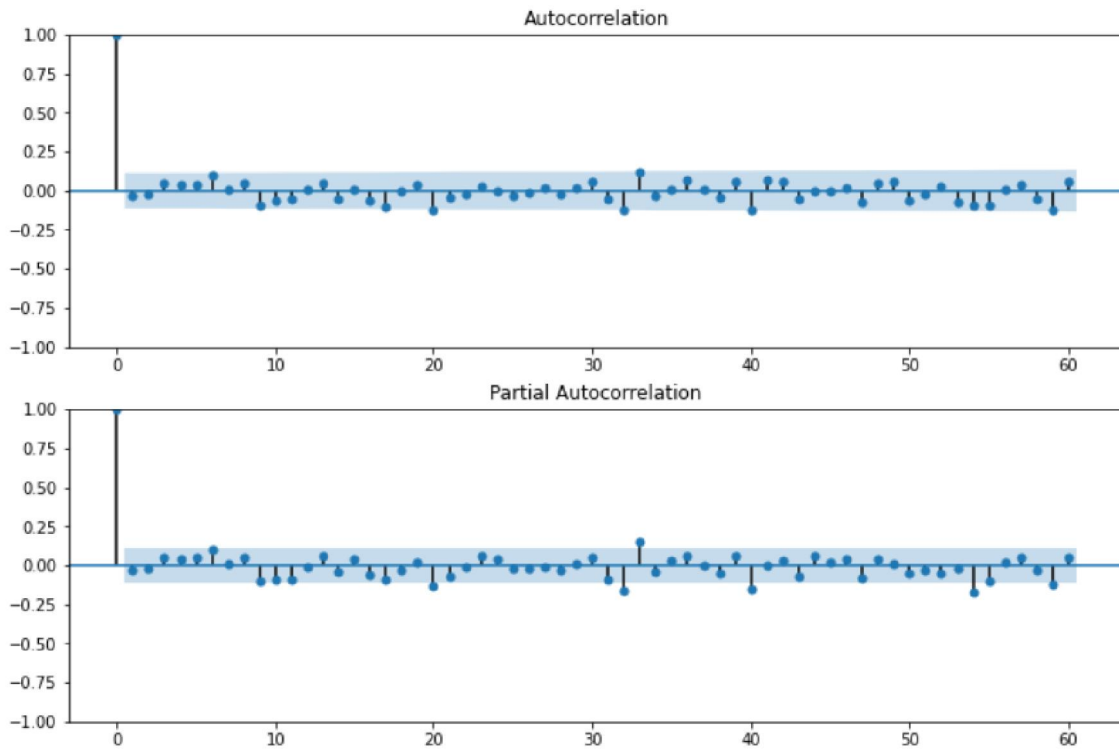


Рисунок 5.5 – Графіки автокореляційної та часткової автокореляційної функції для заданого часового ряду

Функція автокореляції (ACF) — це статистичний метод, який ми можемо використовувати, щоб визначити, наскільки значення в часовому рядуі корельовані одне з одним. ACF будує графік коефіцієнта кореляції з відставанням, яке вимірюється в термінах кількості періодів або одиниць. Лаг відповідає певному моменту часу, після якого ми спостерігаємо перше значення в часовому рядуі. Коефіцієнт кореляції може коливатися від -1 (ідеальна негативна залежність) до +1 (ідеальна позитивна залежність). Коефіцієнт 0 означає, що між змінними немає зв'язку [24-25].

Сині смужки на графіку ACF, наведеному на рисунку 5.5 – це діапазони похибок, і будь-що в цих смугах не є статистично значущим. Це означає, що кореляційні значення за межами цієї області, швидше за все, не є кореляцією. За замовчуванням довірчий інтервал встановлено на 95%. Для нульової затримки ACF завжди дорівнює одиниці, що має сенс, оскільки сигнал завжди ідеально корельований сам із собою.

Часткова автокореляція (PACF) — це статистичний показник, який фіксує кореляцію між двома змінними після контролю впливу інших змінних. Згідно з графіком, що наведений на рисунку 5.5, функція часткової автокореляції фіксує «пряму» кореляцію між часовим рядом і його версією з відставанням [24-25].

Ми можемо вибрати порядок p для моделі $AR(p)$ на основі значних стрибків на графіку PACF. Ще одним свідченням процесу AR є те, що графік ACF затухає повільніше. Модель $MA(q)$ розраховує своє прогнозоване значення, беручи середньозважене минулих помилок. На відміну від моделі AR, ми можемо вибрати порядок q для моделі $MA(q)$ з ACF, якщо цей графік має різке відсічення після затримки q . Ще одним свідченням процесу MA є те, що графік PACF спадає повільніше.

Отже, виходячи з графіку 5.5, найкращими параметрами p та q моделі ARIMA є $p=0$, $q=1$. Для підтвердження даного висновку було використано функцію автоматичного підбору параметрів для моделі ARIMA `auto_arima()`. Результат роботи функції `auto_arima()` наведено на рисунку 5.6.

SARIMAX Results

Dep. Variable: y No. Observations: 213

Model: SARIMAX(0, 0, 1) Log Likelihood: -1476.581

Date: Sun, 18 Dec 2022 AIC: 2959.162

Time: 04:05:06 BIC: 2969.246

Sample: 06-01-2022 HQIC: 2963.237
- 12-30-2022

Covariance Type: opg

	coef	std err	z	P> z	[0.025	0.975]
intercept	6390.0262	19.719	324.061	0.000	6351.379	6428.674
ma.L1	0.1031	0.081	1.278	0.201	-0.055	0.261
sigma2	6.243e+04	5985.815	10.429	0.000	5.07e+04	7.42e+04

Ljung-Box (L1) (Q): 0.00 Jarque-Bera (JB): 3.96
Prob(Q): 0.95 Prob(JB): 0.14

Heteroskedasticity (H): 0.51 Skew: -0.32
Prob(H) (two-sided): 0.01 Kurtosis: 3.21

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Рисунок 5.6 – Результат роботи функції auto_arima()

Виходячи з вище наведеного результату, найкращими параметрами для моделі ARIMA(p,d,q) є $p=0$, $d=0$, $q=1$. Для тренування моделі ARIMA вхідний набір даних було розділено на дві вибірки: тренувана та тестова. Тренувана вибірка була використана для тренування моделі, а тестова була використана для перевірки результатів.

Було проведено тренування моделі ARIMA, та, за допомогою методу predict(), виконано прогноз даних на місяць вперед. Результат прогнозування даних в центрах надання гуманітарної допомоги наведено на рисунку 5.7.

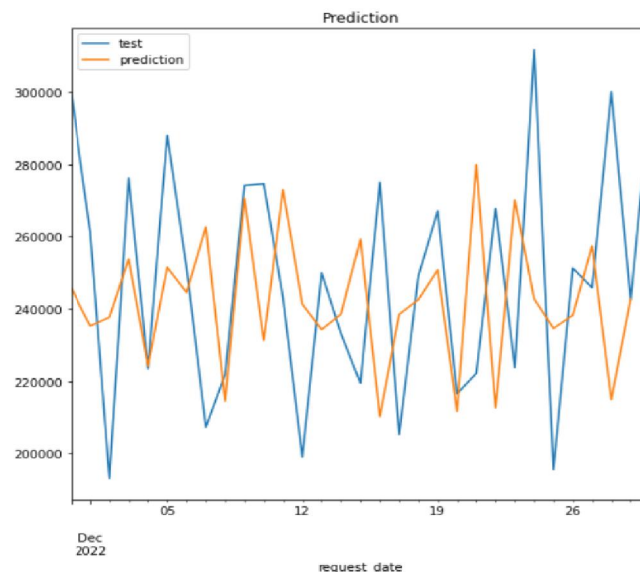


Рисунок 5.7 – Результат прогнозування даних

За даними графіку, результат прогнозування наближений до реальних значень, але спостерігається багато помилок в проведенні прогнозу. Тому, для більш детальної оцінки було пораховано відсоткову різницю між прогнозованими значеннями та тестовою вибіркою, та зроблено запит на отримання мінімальної, максимальної та середньої розбіжності у відсотках. Результати розрахунків наведено на рисунку 5.8.

```
[88] accuracy_df = np.absolute(predict - test_data)/test_data * 100

[90] accuracy_df.aggregate({'min','max','mean'})

max      28.434298
mean     12.720044
min       0.177660
dtype: float64
```

Рисунок 5.8 – Результат розрахунку розбіжності у відсотках між тестовою вибіркою та прогнозованими значеннями

Виходячи з вище наведеного результату, можна зробити висновки, що максимальна похибка моделі дорівнює 28.4%, мінімальна помилка моделі дорівнює 0.17%, а середня помилка моделі дорівнює 12.7%.

Було проведене повторне дослідження на іншому наборі даних з метою уточнити похибки методу прогнозування. Для цього, вхідні дані до методу прогнозування були змінені наступним чином:

- ідентифікатор центру надання гуманітарної допомоги (`humanitarian_center_id`) дорівнює «1002»,
- тип гуманітарної допомоги (`humanitarian_aid_type`) дорівнює «Shelter»,
- групування даних проводити за атрибутом `request_date` та агрегацію підрахунку (`count`) за атрибутом `summary_costs`,
- прогнозування робити на майбутні 10 днів.

Таким чином, було отримано дані кількості користувачів, що надсилали запити на отримання гуманітарної допомоги у вигляді житла. Результат вибірки даних у вигляді графіку наведено на рисунку 5.8,

результат тесту Дікі-Фулера наведено на рисунку 5.9, графіки автокореляційної та часткової автокореляційної функції зображено на рисунку 5.10, результат підбору параметрів p, d, q моделі ARIMA наведено на рисунку 5.11, результат прогнозування даних наведено на рисунку 5.12 та результат розрахунку розбіжності у відсотках між тестовою вибіркою та прогнозованими значеннями наведено на рисунку 5.13.

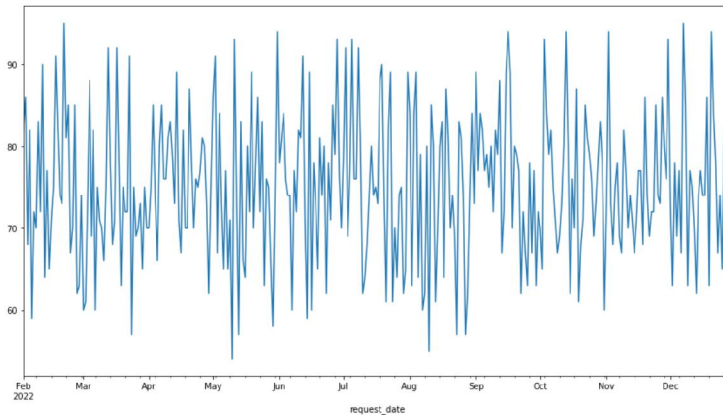


Рисунок 5.8 – Графік кількості запитів користувачів в день на отримання житла в центрі надання гуманітарної допомоги

```
adfuller_test(humanitarian_aid_df_filtered_2_shelter_grouped)
```

```
ADF Test Statistic : -10.648874045609965
```

```
p-value : 4.717479209942513e-19
```

```
#Lags Used : 2
```

```
Number of Observations Used : 330
```

```
strong evidence against the null hypothesis(H0), reject the null hypothesis. Data has no unit root and is stationary
```

```
True
```

Рисунок 5.9 – Результат тесту Дікі-Фулера на часовому ряді

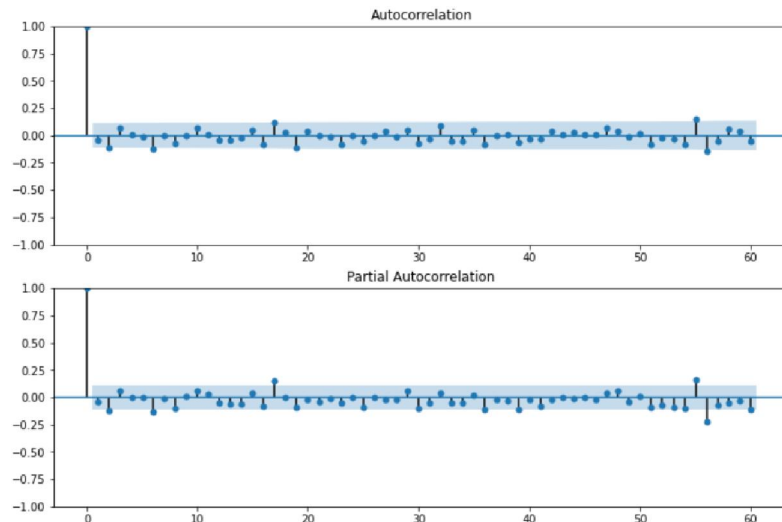


Рисунок 5.10 – Графіки автокореляційної та часткової автокореляційної функції для заданого часового ряду

SARIMAX Results

Dep. Variable: y	No. Observations: 303
Model: SARIMAX	Log Likelihood: -3585.358
Date: Sun, 18 Dec 2022	AIC: 7174.717
Time: 05:21:27	BIC: 7182.144
Sample: 02-01-2022	HQIC: 7177.688
	- 11-30-2022

Covariance Type: opg

	coef	std err	z	P> z	[0.025	0.975]
intercept	2.508e+05	1948.226	128.750	0.000	2.47e+05	2.55e+05
sigma2	1.11e+09	9.03e+07	12.299	0.000	9.33e+08	1.29e+09

Ljung-Box (L1) (Q): 0.02 **Jarque-Bera (JB):** 3.68
Prob(Q): 0.90 **Prob(JB):** 0.16
Heteroskedasticity (H): 1.27 **Skew:** 0.27
Prob(H) (two-sided): 0.23 **Kurtosis:** 3.07

Warnings:

[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Рисунок 5.11 – Результат роботи функції auto_arima()

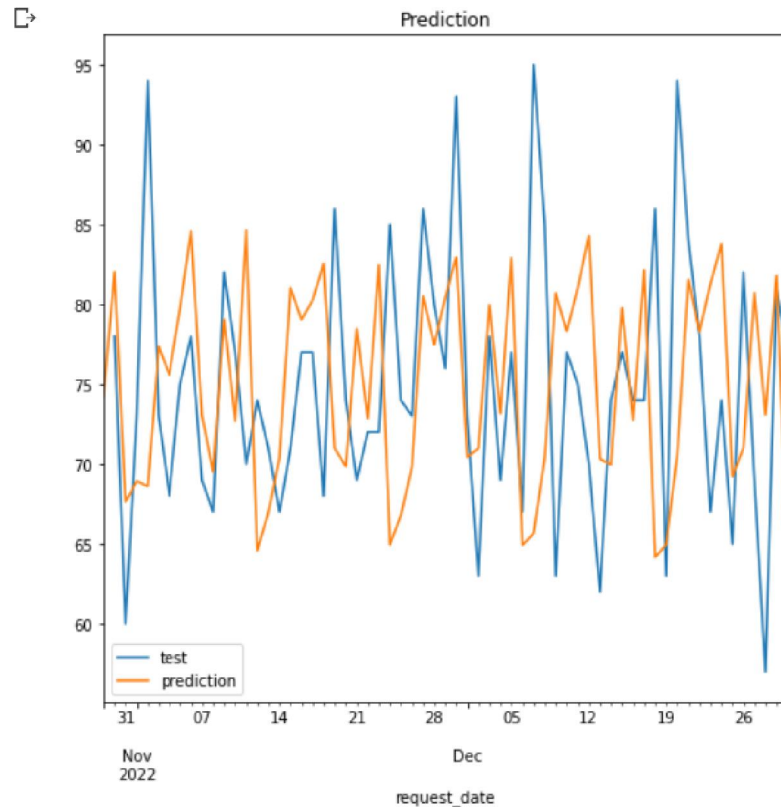


Рисунок 5.12 – Результат прогнозування даних

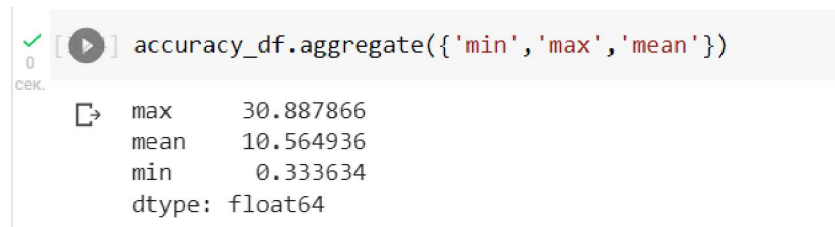


Рисунок 5.13 – Результат розрахунку розбіжності у відсотках між тестовою вибіркою та прогнозованими значеннями

Виходячи з результату, зображеного на рисунку 5.13, максимальна похибка моделі дорівнює 30.8%, мінімальна помилка моделі дорівнює 0.33%, а середня помилка моделі дорівнює 10.5%.

5.7 Аналіз результатів дослідження

Отже, під час проведення експериментальних досліджень методу прогнозування даних в центрах надання гуманітарної допомоги, було

виявлено, що середня похибка методу дорівнює $\sim 10\%$, при максимальній похибці в $\sim 29\%$. Даний результат свідчить про помірну точність розробленого методу прогнозування. Це обумовлено тим, що в якості вхідних даних використовувався штучно згенерований набір даних, що може свідчити за статистичну випадковість даних, аніж їх кореляцією. Також це обумовлено роботою функції, що автоматично визначає параметри p, d, q для моделі ARIMA.

Для більш точних результатів рекомендовано використовувати існуючий набір даних центрів надання гуманітарної допомоги, в якому з більшою імовірністю розроблений алгоритм визначить кореляцію між даними та знайде оптимальні параметри p, d, q для моделі ARIMA.

Отримані результати експериментальних досліджень необхідно враховувати під час застосування методу прогнозування на реальних даних. Наприклад, під час оцінки результатів прогнозованих даних необхідно додатково розрахувати похибку $\pm 30\%$ (максимальна похибка, отримана в результаті експериментальних досліджень) для отримання орієнтованих значень, наближених до реальних.

ВИСНОВКИ

Під час виконання кваліфікаційної роботи було проведено аналіз центрів надання гуманітарної допомоги, проведено аналіз організації гуманітарної допомоги в Україні, визначено проблеми і задачі в діяльності центрів надання гуманітарної допомоги, проведено аналіз існуючих систем аналогів. Було проведено аналіз сучасного стану інформатизації гуманітарної допомоги а також вказано на актуальні проблеми. Було обґрунтовано важливість вирішення задач прогнозування даних в центрах надання гуманітарної допомоги.

Було проведено огляд методів і технологій прогнозування даних, визначено сильні і слабкі сторони кожного з методу, проведено порівняльний аналіз методів прогнозування, було проведено огляд методів і технологій диференціювання часових рядів, та проведено огляд методів дослідження часового ряду на стаціонарність.

Було розроблено змістовну постановку задачі, наведено приклади вхідних даних у вигляді таблиць, було описано загальні вимоги до обробки даних, та наведено приклад очікуваного результату роботи методу прогнозування даних.

Було обґрунтовано вибір методу прогнозування на основі моделі авторегресивної інтегрованої ковзаючої середньої (ARIMA) для вирішення задач прогнозування даних в центрах надання гуманітарної допомоги.

Було розроблено математичний опис задачі прогнозування даних в центрах надання гуманітарної допомоги, а також розроблено комплексний метод прогнозування даних, що включає в себе: перевірку вхідного часового ряду на стаціонарність за допомогою тесту Дікі-Фулера; трансформацію часового ряду в стаціонарний вигляд за умови, якщо ряд не стаціонарний; вибір найкращих вхідних параметрів для моделі ARIMA; проведення тренування моделі та прогнозування даних; відображення результатів методу прогнозування.

Було обґрунтовано вибір інструментів та технологій, що будуть застосовуватися при програмній реалізації методу прогнозування даних в центрах надання гуманітарної допомоги.

Під час проведення експериментальних досліджень було розроблено вимоги до моделі даних, яка буде використовуватися для формування вхідного набору даних до методу прогнозування, розроблено SQL-запит для отримання результуючого набору даних.

Було проведено програмну реалізацію методу прогнозування даних в центрах надання гуманітарної допомоги, що задовольняє потребам, описаними в постановці задачі. Було розроблено функцію, що приймає вхідний набір даних з параметрами прогнозування, та виконує алгоритм описаного вище комплексного методу прогнозування даних. Окрім цього, було проведено програмну реалізацію методу прогнозування даних у вигляді окремих команд для проведення експериментальних досліджень.

Було обрано два вхідних набори даних, та кожний з наборів даних був проаналізований на стаціонарність. Було зображено вхідні дані у вигляді звичайного графіку, а також графіку автокореляційної (ACF) та часткової автокореляційної функції (PACF). За допомогою графіків ACF та PACF було проведено аналіз найкращих параметрів p, d, q для моделі ARIMA та порівняно результат із роботою функції автоматичного підбору даних параметрів. Було проведено прогнозування даних та розраховано похибку результатів прогнозування відносно реальних значень.

В ході виконання експериментальних досліджень було виявлено, що алгоритм має помірну точність, адже середня похибка дорівнює 10%. Було обґрунтовано причину виникання великої похибки під час прогнозування даних, та наведено рекомендації до застосування розробленого методу прогнозування на реальних даних.

Результати виконаного дослідження пройшли апробацію на 41-й Міжнародній інтернет-конференції «Priority areas of modern science» (м.Ліверпуль, Британія) [23].

Завдання на кваліфікаційну роботу було виконано в повному обсязі.
Було виконано всі поставлені завдання.

ПЕРЕЛІК ДЖЕРЕЛ

1. Thornberry C. Peacekeepers, Humanitarian Aid and Civil Conflicts / J. Whitman, D. Pocock. – London: Palgrave Macmillan, 1996. – 244 с.
2. New OCHA organizational chart [Електронний ресурс] // misiura@un.org. – 2018. – Режим доступу до ресурсу: <https://cerf.un.org/document/new-ocha-organizational-chart>.
3. ХАНЧИЧ В. В. ДОДАТОК «ДІЯ»: ПОЗИТИВНІ ТА НЕГАТИВНІ ПОКАЗНИКИ В РОБОТІ //П68 Правове життя: сучасний стан та перспективи розвитку. – С. 133.
4. Романів Я., Думин І. ФОРМУВАННЯ ВИМОГ ДЛЯ ІМПЛЕМЕНТАЦІЇ МОБІЛЬНОГО ДОДАТКУ ДЛЯ ГУМАНІТАРНОЇ ДОПОМОГИ //EDITORIAL BOARD. – 2022. – С. 499.
5. Gutjahr W. J., Nolz P. C. Multicriteria optimization in humanitarian aid //European Journal of Operational Research. – 2016. – Т. 252. – №. 2. – С. 351-366.
6. Vitoriano B. Et al. A multi-criteria optimization model for humanitarian aid distribution //Journal of Global optimization. – 2011. – Т. 51. – №. 2. – С. 189-208.
7. Abraham B., Ledolter J. Statistical methods for forecasting. – New York : Wiley, 1983. – Т. 179.
8. Ганчук А. А., Соловійов В. М., Чабаненко Д. М. Методи прогнозування. – 2012.
9. Широкопетлева М. С., Пономаренко О. А., Дудар З. В. Порівняння методів прогнозування часових рядів //Біоніка інтелекту. – 2018. – Т. 2. – №. 91. – С. 41-47.
10. Неруш В. Б., Курдеча В. В. Імітаційне моделювання систем та процесів. – 2016.
11. Peter Ď., Silvia P. ARIMA vs. ARIMAX–which approach is better to analyze and forecast macroeconomic time series //Proceedings of 30th

international conference mathematical methods in economics. – 2012. – Т. 2. – С. 136-140.

12. Vakaliuk T. A., Lukianchuk Y. V. Теоретичні аспекти розробки мобільного додатка для прогнозування ймовірної вартості земельної ділянки //Актуальні проблеми автоматизації та інформаційних технологій. – 2020. – Т. 24.

13. Макарчук О. П., Петрикова І. І. ПРОГНОЗУВАННЯ КУРСУ НА ОСНОВІ МОДЕЛІ ARIMA //Наукові записки молодих учених. – 2021. – №. 8.

14. Meyler A., Kenny G., Quinn T. Forecasting Irish inflation using ARIMA models. – 1998.

15. Прогноз розвитку епідемії COVID-19 в Україні на 7–21 грудня 2020 року («Прогноз РГ-31») [Електронний ресурс]. – 2020. – Режим доступу до ресурсу:

<https://www.nas.gov.ua/UA/Messages/news/Pages/View.aspx?MessageID=7238>

16. Осипчук М. М. Часові ряди. – 2009.

17. Forecasting: Principles and Practice [Електронний ресурс] // Monash University, Australia. – 2016. – Режим доступу до ресурсу: <https://otexts.com/fpp2/>.

18. Hamilton J. D. Analysis of time series subject to changes in regime //Journal of econometrics. – 1990. – Т. 45. – №. 1-2. – С. 39-70.

19. Назаренко О. М., Карпуша М. В. ІМІТАЦІЯ ТА ПРОГНОЗУВАННЯ НЕСТАЦІОНАРНИХ ФІНАНСОВИХ ЧАСОВИХ РЯДІВ© //Економічна кібернетика. – 2011. – №. 4-6. – С. 74-82.

20. Hyndman R. J., Athanasopoulos G. Forecasting: principles and practice. – OTexts, 2018.

21. What is Google Colab? [Електронний ресурс] – Режим доступу до ресурсу: <https://ledutokens.medium.com/what-is-google-colab-281c5b59638f>.

22. Quick D., Choo K. K. R. Google Drive: Forensic analysis of data remnants //Journal of Network and Computer Applications. – 2014. – Т. 40. – С. 179-193.

23. Holyk Y., Development of a forecast method for humanitarian aid centers // Priority Areas of Modern Science. Abstracts of the 41st International scientific and practical conference. Nika Publishin. Liverpool, Great Britain. 2022. Pp. 39- 43. URL: <http://el-conf.com.ua/>.

24. Mondal P., Shit L., Goswami S. Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices //International Journal of Computer Science, Engineering and Applications. – 2014. – T. 4. – №. 2. – C. 13.

25. Demir V., Zontul M., Yelmen İ. Drug Sales Prediction with ACF and PACF Supported ARIMA Method //2020 5th International Conference on Computer Science and Engineering (UBMK). – IEEE, 2020. – C. 243-247

